
Searching for Microbial Nucleic Acids in Cancer Sequence Data

ABRAHAM FARAG SALAH GIHAWI



Norwich Medical School
UNIVERSITY OF EAST ANGLIA

A thesis submitted to Norwich Medical School at the University of East Anglia in partial fulfilment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

© This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived therefrom must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

JULY 2021

Word Count: 47,621

Table of Contents

Contributorship Statements	ix
Acknowledgements	1
Declaration	3
Chapter 1: Background	5
1.1 Biological Background	5
1.1.1 Introduction to Cancer	5
1.1.2 The Hallmarks of Cancer	7
1.1.3 Cancer Therapeutics	8
1.1.4 Microbes and Cancer Introduction	9
1.1.5 Molecular Aetiology of Oncomicrobes and Cancer	11
1.1.6 Emerging Mechanisms	12
1.1.7 The Human Microbiome	12
1.1.8 Microbes and Therapeutic Approaches	13
1.1.9 High Throughput Sequencing Technology	14
1.2 Computational Background	17
1.2.1 An Overview of Computational Methods of Taxonomic Classification	17
1.2.2 Differential Testing and False Discovery Control	19
1.2.3 Diversity Measures	19
1.2.4 Ordination	20
1.2.5 Clustering	22
1.2.6 Supervised Learning	22
1.2.7 Functional Taxonomic Composition	23
1.2.8 Handling Sequence Quality, Noise and Contamination	24
1.3 Thesis Outline	26
1.3.1 Aims and Objectives	26
1.3.2 Chapter Outline	26
Chapter 2: Developing SEPATH - Template Pipelines to Search for Micro- bial DNA in Host Dominated Sequence Data	29
2.1 Summary	29
2.2 Background	30
2.3 Results	32
2.3.1 Simulated Datasets	32
2.3.2 Human Sequence Depletion	33

2.3.3	Taxonomic Classification - Bacterial Datasets	35
2.3.4	Bacterial Proportion Estimation	36
2.3.5	Bacterial Classification Following Metagenomic Assembly	37
2.3.6	Taxonomic Classification - Viral Datasets	39
2.3.7	Bacterial Consensus Classification	40
2.3.8	Real Cancer Whole Genome Sequence Data	40
2.3.9	SEPATH Template Pipelines	43
2.4	Discussion	44
2.5	Conclusions and Future Directions	46
2.6	Methods	47
2.6.1	Metagenome Simulations	47
2.6.2	Tool Performance Benchmarking	47
2.6.3	Real Cancer Whole Genome Sequence Analysis	48
2.6.4	Computational Tools and Settings	48

Chapter 3: Microbes of Prostate Cancer Exosomal Urine and Potential Microbe-Host Interactions 49

3.1	Prostate Cancer Background	49
3.1.1	Anatomy of the Prostate	51
3.1.2	The History of Microbes and Prostate Cancer	51
3.1.3	Urine as a Source of Biomarkers	53
3.2	Results	55
3.2.1	Evidence for a Range of Genera in Post-DRE Urine Exosome Samples	55
3.2.2	Interrogating a Prognostic Panel of Bacteria	57
3.2.3	ABBS Isolates from Patients Appear Functionally Related	58
3.2.4	Ordination Suggests Community Differences	60
3.2.5	Differential Expression of Human RNA in Exosome Urine	62
3.2.6	Pathway Enrichment Analysis of Differentially Expressed Genes Between Clusters Suggests Immunological Differences	63
3.2.7	Survival Analysis Between Clusters	65
3.3	Discussion	66
3.3.1	Support for Findings in Published Literature	66
3.3.2	Microbial Presence in the Urine Impacts Immune Host Gene Expression	67
3.3.3	Isolate Functional Analysis	68
3.4	Conclusions and Future Directions	69
3.4.1	Potential Future Experiments	70
3.4.2	Potential for Clinical Exploitation	71
3.5	Methods	72
3.5.1	Urine RNA Exosome Sequencing	72
3.5.2	RNA Sequencing Bioinformatics	73
3.5.3	Isolate Sequencing and Assembly	74
3.5.4	Prediction of Functional Pathways	74

Chapter 4: Applying SEPATH to Cancer Whole Genome Sequences of Genomics England's 100,000 Genomes Project 75

4.1	100,000 Genomes Project Overview	75
4.1.1	100,000 Genomes Project Ethical Considerations	77
4.2	Comparable Published Literature	78

4.3	Chapter Overview	78
4.4	Results	78
4.4.1	mOTUS2 is Insensitive on Genomics England Data	78
4.4.2	An Overview of Kraken Results	79
4.4.3	Most Prevalent Kraken Genera	80
4.4.4	Sparsity	92
4.4.5	HPV Diagnosis - Comparison of Kraken with Clinical Detection Status	94
4.4.6	Reporting Clinically Actionable Results	95
4.4.7	Investigating ABBS Genera in Genomics England	97
4.4.8	Kraken Reported Tumour Viriome Structure	98
4.4.9	The Search for Microbial Community Structure	101
4.4.10	Investigating the Poore <i>et al.</i> Manuscript	104
4.5	Discussion	110
4.5.1	Kraken vs mOTUS2 and Sparsity	110
4.5.2	HPV Detection in Genomics England	111
4.5.3	Implications of HTLV Detection	111
4.5.4	Cancer Viriome	112
4.5.5	Contaminants and Microbial Structure	112
4.5.6	Modelling Cancer Type - Comparing with Poore <i>et al.</i>	113
4.5.7	Searching for Clinical Associations	113
4.6	Conclusions	113
4.7	Future Directions	114
4.8	Methods	115
4.8.1	100,000 Genomes Sample Collection and Processing	115
4.8.2	Identification of Mutational Status	115
4.8.3	HTLV Classification	115
4.8.4	Viral Classifications	116
4.8.5	Dimensionality Reduction	116
4.8.6	Re-analysing the Poore <i>et al.</i> Manuscript	118
4.8.7	Supervised learners	119
Chapter 5: Discussions, Future Work and Concluding Remarks		121
5.1	Chapter Summaries	121
5.1.1	Thesis Aims	123
5.2	Discussions	123
5.2.1	The Applications of SEPATH	123
5.2.2	The Search for Pan-Cancer Microbial Structure	124
5.2.3	The Search for Clinical Associations	125
5.2.4	The Search for Functional Pathways	127
5.2.5	Limitations	128
5.2.6	Potential Impact	129
5.3	Future Work	129
5.3.1	Future Work on Genomics England	129
5.3.2	Pan-Prostate Cancer Group Investigations	130
5.3.3	Additional Future Work	131
5.3.4	Potential for Clinical Impact	131
Appendix A: System Information and Background		133

A.1	System Information	133
A.2	Background	136
A.2.1	Beta Diversity Measurements	136
A.2.2	Computational Methods for taxonomic classification	138
A.2.3	Metagenomic Assembly Tools	150
A.2.4	Common Contaminants in Negative Controls	151
Appendix B: Developing SEPATH		153
B.0.1	Species used for Metagenome Simulation	153
B.0.2	BBDuK Parameter Investigation	155
B.0.3	Bacterial Read Retention	157
B.0.4	Contig Coverage in Simulations	158
B.0.5	Classification Performance by Contig Coverage	159
B.0.6	KrakenUniq Filtering Parameters	159
B.0.7	Viral Dataset Classification	160
Appendix C: Prostate Cancer		161
C.0.1	D’Amico Classification system	161
C.1	Urine RNA NMDS Scree and Stress Plot	161
C.2	Urine RNA differences in diversity appear independent of sequencing depth	163
C.3	Differentially Expressed Human Genes between NMDS Clusters	164
C.4	Discovering the Anaerobic bacterial biomarker set (ABBS)	170
C.5	Investigating ABBS in a Separate Dataset	172
C.6	Investigating differentially Expressed human genes in urine	174
C.7	Differentially Expressed Human Genes between ABBS vs non-ABBS Samples	174
C.8	Differential Expression Pathway Enrichment Analysis - ABBS group differential	175
C.8.1	Pathways, ABBS group differential - Gene Ontology Biological Process	175
C.8.2	Pathways, ABBS group differential - Gene Ontology Cellular Component	176
C.8.3	Pathways, ABBS group differential - Gene Ontology Molecular Function	176
C.8.4	Pathways, ABBS group differential - Kegg	176
C.9	Differential expression pathway Enrichment Analysis- NMDS Cluster differential	177
C.9.1	Pathways, cluster group differential - gene ontology, biological process	177
C.9.2	Pathways, cluster group differential - Gene Ontology Cellular Component	177
C.9.3	Pathways, cluster group differential - Gene Ontology Molecular Function	177
C.9.4	Pathways, cluster group differential - Kegg	178
C.10	List of Isolates Interrogated	178
C.11	Isolate Sequencing Statistics	179
C.12	Building Kallisto Index	180
C.13	SEPATHWAY Pipeline	181
Appendix D: Genomics England and SEPATH		187
D.1	Taxa Reported by mOTUs2 and not Kraken	187
D.2	Investigating Structure in mOTUs2 data	187
D.3	Read Complexity	190
D.4	Aligning Reads to HTLV genome	191

D.5	Searching for Contaminant Genera	192
D.6	Comparing TCGA Kraken genera with SEPATH Kraken Genera	200
D.7	Searching for Correlations	209
D.8	Repeating Models with Different Genera	210
D.8.1	Repeating TCGA models with different genera	210
D.9	GBM and Boruta Tumour Type - Feature Importance	213
D.10	Kraken <i>t</i> -SNE analysis	221
D.11	Kraken Community Structure	225
D.12	Searching for Contaminants	225
D.12.1	CramersV results	227
D.13	100,000 Genomes and SEPATH - Negative Findings	229
D.14	ABBS Genera in Genomics England	229
References		245

Abstract

Microbes can play a prominent role in cancer: *Helicobacter pylori* is involved in over 90% of gastric non-cardia adenocarcinoma and Human Papillomavirus plays a major role in cervical tumorigenesis.

In this thesis, I evaluate computational approaches to identify the constituent taxa within human high-throughput sequence data. I have produced the top performing approaches into a pipeline (SEPATH) which is applied to a variety of datasets: RNA sequencing of urine from patients under clinical investigation for prostate cancer and over 10,000 whole genome sequences from Genomics England's 100,000 Genomes Project.

The results are sparse and rife with environmental and sequencing contaminants. Despite this, SEPATH has revealed a range of interesting bacterial and viral genera associated with tumour samples. A preliminary association is observed between the identification of specific taxa and the development of aggressive prostate cancer. Many of the genera identified have been previously suggested for association with tumours such as *Bacteroides* and *Fusobacterium* in colorectal cancer. Alphapapillomavirus was noted in oral cancer which aligns with the expected genomic features (a lack of mutations in tumour suppressor gene *TP53*). Also, evidence has been detected for infectious disease which will be subject to independent validation and followed up appropriately.

Analysing the microbial composition of tumours could provide an additional tool to aid in therapeutic stratification of cancer patients with little added cost following sequencing.

Access Condition and Agreement

Each deposit in UEA Digital Repository is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form. You must obtain permission from the copyright holder, usually the author, for any other use. Exceptions only apply where a deposit may be explicitly provided under a stated licence, such as a Creative Commons licence or Open Government licence.

Electronic or print copies may not be offered, whether for sale or otherwise to anyone, unless explicitly stated under a Creative Commons or Open Government license. Unauthorised reproduction, editing or reformatting for resale purposes is explicitly prohibited (except where approved by the copyright holder themselves) and UEA reserves the right to take immediate 'take down' action on behalf of the copyright and/or rights holder if this Access condition of the UEA Digital Repository is breached. Any material in this database has been supplied on the understanding that it is copyright material and that no quotation from the material may be published without proper acknowledgement.

Contributorship Statements

Gihawi, A., Rallapalli, G., Hurst, R. et al. *SEPATH: benchmarking the search for pathogens in human tissue whole genome sequence data leads to template pipelines*. Genome Biol 20, 208 (2019). <https://doi.org/10.1186/s13059-019-1819-8>. AG developed the manuscript and SEPATH, is responsible for the metagenome simulation and tool benchmarking, and produced all graphical presentations. GR alongside AG modified BEAR for metagenomic simulation. GR developed the Kraken database and supervised the early development of SEPATH. RH advised on the metagenomic content of the simulated datasets. DB, RL, CC, and RH contributed towards the development of the final manuscript. RL and DB supervised the production and development of SEPATH. DB obtained and processed the cancer whole genome sequencing files prior to AG running SEPATH. DB and CC developed the original concept of SEPATH. All authors read and approved the final manuscript.

List of Tables

1.1	Physiological impacts of tumours. Adapted from Ruddon 2017 ⁷	6
1.2	Pathogens with strong evidence of association with cancer. Adapted from Pappas 2009 ⁴¹ and Zhu <i>et al.</i> 2017 ⁴²	10
1.3	Microbes with emerging evidence of association with cancer. Adapted from Pappas 2009 ⁴¹ and Zhu <i>et al.</i> 2017 ⁴²	11
3.1	Gene Ontology - Biological Process - Pathway analysis of differentially expressed human genes in urine RNA sequencing. Samples split by NMDS cluster. Gene IDs and q -values have been removed for display purposes (0.05 maximum threshold as determined by the clusterProfiler package).	64
3.2	Clinical demographics of the Urine RNA sequencing cohort and split by ABBS detection status using Kraken reads. IQR Interquartile range, NEC No evidence of cancer.	72
4.1	A sample community matrix with taxonomic count data.	105
4.2	Voom normalisation of a sample community matrix with taxonomic count data.	105
B.1	Species and Filenames used as part of metagenome simulation	153
C.1	D'amico classification system for prostate cancer	161
C.2	Differentially Expressed human genes between NMDS cluster. Maximum FDR adjusted Q -value for differential expression = 0.1	164
C.3	Differentially Expressed human genes between ABBS groups cluster. Maximum FDR adjusted Q -value for differential expression = 0.1	174
C.4	Gene Ontology - Biological Process - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by ABBS group.	175
C.5	Gene Ontology - Cellular Component - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by ABBS group.	176
C.6	Gene Ontology - Cellular Component - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by ABBS group.	176
C.7	Kegg - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by ABBS group.	176

C.8	Gene Ontology - Biological Process - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by NMDS cluster.	177
C.9	Gene Ontology - Cellular Component - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by NMDS cluster.	177
C.10	Gene Ontology - Cellular Component - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by NMDS Cluster.	177
C.11	Kegg - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by NMDS cluster.	178
C.12	Isolate Names	178
C.13	Isolate sequencing statistics in terms of isolate ID, taxonomy, number of sequenced reads and whether additional minION sequencing was undertaken, average coverage, genome size (Mb) and total Number of contigs assembled. nd not determined.	179

List of Figures

1.1	The Hallmarks of Cancer. Adapted from Hanahan and Weinberg 2011 ¹¹ . . .	7
1.2	Overview of the Human Microbiome Project results. A) A bar plot demonstrating relative abundance at genus level for each sample partitioned by sampling site. B) A principal coordinates plot demonstrating that samples from the same body sites tend to share similar constituent micro-organisms. Retroauricular crease (RC). Image adapted from The Human Microbiome Project Consortium 2012 ⁷²	13
1.3	The process of Illumina sequencing. A) Library Preparation. Sample DNA is fragmented and adapters added to the ends. B) Cluster Amplification. Fragments are attached to the flow cell by binding adapters to oligonucleotides on the flow cell. DNA fragments are bent and amplified by PCR to form a clonal cluster of DNA fragments in a specific region on the flow cell. C) Sequencing. Illumina uses a ‘sequencing by synthesis’ approach whereby fluorescent dNTPs are added sequentially to a growing DNA fragment. Software records the colour from each cluster and assigns a nucleotide forming a ‘read’ with quality (Phred/ <i>Q</i>) scores that estimate the accuracy of base calling. D) Alignment and Data Analysis. The reads for all the fragments can then be subject to downstream analysis including assembly and/or alignment to a reference genome. The number of times a nucleotide in a single position is measured is known as the coverage. Image adapted from Illumina 2018 ⁸⁸ . .	15
1.4	The output of different sequencing technologies in megabases. The number inside each point indicates typical read length. This diagram shows that Illumina has a very high throughput validating their use in a large-scale sequencing project. Image adapted from Reuter <i>et al.</i> 2015 ⁹³	16
1.5	Use of the LCA (Lowest Mommon Ancestor) algorithm by Kraken. The classification of an input sequence is determined by the highest weighted root-to-leaf path of of its classified <i>k</i> -mers. Adapted from Wood and Salzberg 2014 ¹⁰³	18
1.6	Examples of PCoA (Principal Coordinates Analysis) in ecological data exploration. A) PCoA can be used to identify differences between healthy and diseased states. Confounders can also be detected by identifying a separation based on confounding variables (B-C). B) A separation is linked with mouse litter indicating a potentially confounded study design. C) shows a separation of data points by mutational status. All figures adapted from Goodrich <i>et al.</i> 2014. ¹⁴⁰	21

2.1	An overview of the process used to develop the SEPATH computational pipelines. Firstly, metagenomic datasets were simulated containing either bacterial or viral reads. A variety of methods were tested on these datasets before being applied to some real cancer sequencing data from cervical cancer, stomach cancer and medulloblastoma. Positive Predictive Value (PPV), Sensitivity (SSV).	32
2.2	Scatter plot summarizing the constituents of all 100 simulated bacterial metagenomes. The y -axis demonstrates the number of bacterial reads in the datasets, whereas the number of human reads is shown on the x -axis. The number of species in each dataset is indicated by the color, darker points having less species. The distribution of each axis is shown in red.	33
2.3	Human read depletion performance. A) Human read removal using BBDuk, BWA-MEM, and Kontaminant. The remaining numbers of human reads were near identical for BBDuk and Kontaminant (median values of 15,399,252 and 15,399,928 for BBDuk and Kontaminant, respectively.) All conditions retained bacterial reads with near-identical performance. BBDuk was selected for parameter optimization (B , C). This analysis was performed on raw untrimmed reads of $n = 11$ simulated datasets. B-C) BBDuk parameter optimization in terms of the remaining human reads (B) and remaining bacterial reads (C). Default BBDuk settings were used along with alterations of MKF and MCF parameters. The default parameters of BBDuk remove a sequencing read in the event of a single k -mer match, whereas MCF50 requires 50 percent of the bases in a read to be covered by reference k -mers for removal and MKF50 requires 50 percent of k -mers in a read to match the reference for removal. MCF50-Cancer indicates that BBDuk was ran with a database consisting of GRCh38 and mutations identified in human cancer from COSMIC. MCF50-Cancer-A denotes a database consisting of GRCh38, COSMIC cancer genes, and additional African “pan-genome” sequences. Default and both MCF50 parameters (with and without cancer sequences) showed the highest removal of human reads.	34
2.4	Performance estimates for taxonomic classification tools on 100 metagenome simulations following quality filtering and human sequence depletion. Performance is summarized at the genus level in terms of sensitivity (A), positive predictive value (B), and F1 score (C). Computational resources, CPU time and RAM is also shown for the top two performing tools: Kraken and mOTUs2 (D). Kraken utilised 20 threads for most datasets whereas mOTUs2 utilised 17. mOTUs2 output was unfiltered, whereas Kraken had a confidence threshold of 0.2 and a minimum read threshold of 500 in determining positive classifications.	35
2.5	Violin plots demonstrating performance in terms of F1-score, PPV and SSV for taxonomic ranks between Phylum and Species level on $n=100$ simulated datasets. A) Performance of Kraken when run on raw reads with no read threshold. B) Performance following a minimum threshold of 500 reads for each classification.	36

2.6	Quantitative ability for mOTUs2 and Kraken. mOTUs2 output reads vs true reads (A) and Kraken output reads vs true reads (B) . For all true-positive genera classifications (Spearman's rank correlation coefficients 0.91 and 0.69, for $n = 2,084$ and 2,021 true-positive classifications for mOTUs2 and Kraken, respectively). mOTUs2 classifications were left unfiltered whereas Kraken had a confidence threshold of 0.2 and a minimum read threshold of 500.	37
2.7	Genus-level performance of Kraken on contigs following metagenomic assembly with MetaSPAdes. Performance is summarized by genus-level F1 score (A) , sensitivity (B) , and PPV (C) . A single dataset failed metagenomic assembly, and so, data shown is for 99 of 100 simulated datasets. Performance is shown on raw Kraken classifications with no threshold applied (unfiltered) in dark blue. Light blue denotes the performance when a minimum of 5 contigs assigning to a genus was used. Median values for unfiltered performance were 0.83, 0.88, and 0.81, and for filtered performance were 0.89, 0.85, and 0.94 for F1 score, sensitivity, and PPV, respectively. (D) KrakenUniq filtering parameters in relation to detection status. The y -axis indicates the number of unique k -mers assigned to a particular taxon, the x -axis represents the number of contigs assigned to a particular taxon ($\log_{10}$), and the colour gradient shows the coverage of the clade in the database ($\log_{10}$). True-positive results are larger circles, whereas false-positive results are smaller triangles. The scatter plot shows 10,450 contigs classified at genus level as data points. The k -mer size was 31.	38
2.8	Kraken performance on a single dataset containing both bacterial (A) and viral (B) reads. Performance from metagenomic assembly approach is shown on both unfiltered contigs and results filtered by a minimum of 5 contigs required for classification. Kraken performance on raw reads is shown both unfiltered and filtered by a minimum of 100 reads for classification. Bacterial performance is classified at genus level whereas viral performance is regarding species level due to peculiarities in taxonomy.	39
2.9	mOTUs2, Kraken, and Pathseq form a consensus with near-perfect genus-level classification performance. Box plots with individual data points for $n = 11$ simulated bacterial metagenomes showing genus-level F1 score (A) , PPV (B) , and SSV (C) for single tools, an intersection of classification between two tools, and a consensus of all three tools. PPV obtained perfect values in the result of an intersection between two tools or a consensus. Sensitivity generally decreases in the event of combining two tools with an intersection but increases to 0.905 (median) in the result of an intersection. This raise in sensitivity resulted in a genus level F1 score in the consensus approach of 0.95. mOTUs2 output files were unfiltered, whereas Kraken had a filter of >4 contigs and PathSeq >1 reads.	40
2.10	The application of SEPATH pipelines on a range of cancer types. Output genera from Kraken (left) and mOTUs2 (right) human-depleted, quality-trimmed reads from whole genome sequencing files. $n = 10$ for each of cervical cancer (A, B) , stomach cancer (C, D) , and medulloblastoma (E, F) . For display purposes, mOTUs2 results were filtered to show taxa that occurred in at least 3 samples. Kraken results were filtered for taxa that were in a minimum of 5 samples, or had a mean read count of over 5,000.	42

2.11	SEPATH template computational pipeline. The top-performing pipelines from this benchmark are provided as a template for users to adjust according to their own job scheduling systems and resource availability. SEPATH provides two main pathways: a bacterial pipeline using mOTUs2 classifications on raw sequencing reads and a bacterial and viral pipeline employing Kraken on metagenomic contigs assembled using non-human reads with MetaSPAdes.	44
3.1	The patient pathway for prostate cancer in the UK. Patients are stratified based on the D'Amico classification system ²⁴² which uses Digital Rectal Exam (DRE), Magnetic Resonance Imaging (MRI) and histopathology of a template or transrectal ultrasound guided biopsy (TRUS). Patients can also be offered focal therapy, particularly those with low or intermediate risk ²⁴⁹ .	50
3.2	The anatomy of the prostate. Adapted from the urology care foundation ²⁵⁸ .	51
3.3	Bacteria detected in prostate tissue whole genome sequencing. A) A heatmap of the top 15 abundant genera identified by Feng et al 2019. B) An NMDS (non-metric multidimensional scaling) plot of bacterial communities found in prostate tissue, coloured by geographical location and clinical risk. Image adapted from Feng <i>et al.</i> 2019 ²⁶⁶ .	52
3.4	The biogenesis and secretion of small vesicles. A) The release of signal proteins from the rough endoplasmic reticulum and golgi apparatus. B) Merocrine secretion pathway. Cell membranes fold inwards to form early endosomes. The cytoplasmic content is processed into smaller vesicles called exosomes. The endosome then fuses with the plasma cell membrane to release exosomes from the cell. C) larger vesicles (50-1000nm) can also be expelled by the cell by the budding outwards as a form of apocrine signalling. Image adapted from Duijviesz <i>et al.</i> 2011 ²⁸⁴ .	54
3.5	An overview of this chapter. Most of this chapter focuses on RNA sequencing from the urine of 40 prostate cancer patients. This was subject to microbial analysis investigating two partitions of the samples, NMDS (Non-Metric Multidimensional Scaling cluster) and ABBS (Anaerobic Bacterial Biomarker Set). These partitions were used to investigate differential expression of human genes and survival analysis. Differentially expressed genes and enriched pathways were determined for these partitions. Additionally, the functional pathways present in in-house isolates were examined and compared. These isolates were obtained from culturing urine samples performed by Dr Rachel Hurst and Dr Emma Meader. Assembly and taxonomic classification was carried out by Dr Ghanasyam Rallapalli.	55
3.6	The top genera detected in a prostate cancer urine RNA sequencing dataset. Suspected contaminants were removed and taxa present in fewer than three samples are not displayed. Full results are available on the thesis repository. Genera are ordered by median number of reads assigned using Kraken. The number of samples with any evidence for the taxa is indicated on the right.	56

3.7	Kaplan-Meier plots of time to progression based on whether patients had evidence for ABBS genera in their urine. Samples in AB group contained ABBS genera (blue) and the NA group did not (orange). Vertical lines indicate limits of patient follow-up. (A) All samples and (B) only cancer samples. Log-rank test was used to compare survival distributions. The ABBS group constitutes urine samples where at least one of the genera (<i>Fenollaria</i> , <i>Ezakiella</i> , <i>Peptoniphilus</i> , <i>Porphyromonas</i> , <i>Anaerococcus</i> , <i>Fusobacterium</i>) were detected in RNAseq data. The following were used as markers of progression: metastasis, an increase in histological staging, prostate cancer associated deaths, biochemical recurrence or an increase in PSA sufficient to alter the course of clinical treatment.	58
3.8	Principal components analysis of pathways associated with strains isolated from prostate cancer participants ($N=51$). Isolates belonging to ABBS genera ($n=21$) are highlighted in red and non-ABBS isolates ($n=30$) are blue.	59
3.9	Pathways enriched in isolates from ABBS (Anaerobic Bacterial Biomarker Set) genera (right) and from non-ABBS genera (left). Pathways were identified in multiple isolates using InterProScan. The x -axis demonstrates the difference in the percentage of positive samples. A value of 1 indicates that all ABBS bacteria had the pathway and none of the non-ABBS group did. The y -axis indicates the q -value. None of these pathways are significantly differential at the 0.05 level after false discovery adjustment. Gene Ontology (GO) Biological Process (BP), GO Molecular Function (MF), GO Cellular Component (CC).	60
3.10	Dimensionality reduction of microbial community of urine RNA sequencing Kraken classifications. A) Non-metric multidimensional scaling (NMDS) plot obtained using the Jaccard distance. The data points can be optimally partitioned using k -means clustering into two clusters, A (16 samples) and B (24 samples). B) Principal Components Analysis PCA plot of the filtered community matrix to remove taxa present in only one sample. Counts were log2 normalised using the decostand function and the first two principal components are displayed. The first three principal components capture 16.9, 13.5 and 11 percent of the total variation in the dataset. Symbols indicate whether the sample has at least one ABBS genera present (circles) or not (triangles).	61
3.11	Heatmap of log transformed Kraken community data. Blue indicates no evidence for a taxon in a sample, whereas red indicates stronger evidence (legend on log2 of Kraken counts). Samples on the x -axis are ordered by NMDS axis 1. Taxa on the y -axis are ordered by complete-linkage hierarchical clustering and the dendrogram hidden for display purposes. Samples in NMDS cluster A are shown in green and cluster B samples in purple. ABBS genera are highlighted with a yellow asterisk. Sparse taxa in less than two samples were removed for display purposes.	62
3.12	Human genes differentially expressed in urine between NMDS cluster B vs cluster A. Samples with positive log2fold change values indicate an increased expression in cluster B samples compared to cluster A samples. The horizontal line indicates $p=0.01$, the vertical lines indicate an absolute log2fold change greater than 1.	63

3.13	Kaplan-Meier plots of time to progression divided by NMDS cluster assignment. For (A) all samples (B) . Time to progression was defined as: metastasis, an increase in histological staging, prostate cancer associated deaths, biochemical recurrence or an increase in PSA sufficient to alter the course of clinical treatment. Log-rank test was used to compare survival distributions between groups. NMDS cluster determined from taxa count identified by Kraken on Urine RNAseq data.	66
4.1	Summary of cancer types in the 100,000 Genomes Project. These numbers are based on data release version 9 and resemble only tumour samples and not germline.	76
4.2	An overview of the analyses performed in this chapter. SEPATH was applied on Genomics England 100,000 Genomes Project cancer whole genome sequencing (WGS) data. The following are investigated: ABBS genera (Anaerobic Bacterial Biomarker Set), HPV diagnostics, clinically actionable microbes and pancancer microbial structure.	77
4.3	Total number of reads classified at any taxonomic level ('root') across whole genome sequences of cancer types within Genomics England's 100,000 Genomes Project. Samples are ordered by the median number of non-human unclassified reads in order from most (colorectal) to least (ovarian) as shown in appendix figure D.3. This analysis includes 361 FFPE samples and 603 PCR samples.	80
4.4	The 20 most frequently reported microbial genera across all cancer types in the 100,000 Genomes Project whole genome sequence data after applying SEPATH with Kraken. A minimum of 20 reads was required to determine a positive result. Taxa are ordered with the most prevalent genus (<i>Achromobacter</i>) first.	81
4.5	Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for A) colorectal cancer and B) ovarian cancer. The <i>y</i> -axis denotes the number of positive samples.	83
4.6	Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for A) lung cancer and B) prostate cancer. The <i>y</i> -axis denotes the number of positive samples.	84
4.7	Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for A) breast cancer and B) sarcoma. The <i>y</i> -axis denotes the number of positive samples.	85
4.8	Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for A) renal cancer and B) endometrial cancer. The <i>y</i> -axis denotes the number of positive samples.	86
4.9	Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for A) bladder cancer and B) testicular cancer. The <i>y</i> -axis denotes the number of positive samples.	87

4.10	Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for A) adult glioma and B) hepatopancreatobiliary cancer. The <i>y</i> -axis denotes the number of positive samples.	88
4.11	Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for A) malignant melanoma and B) upper gastrointestinal. The <i>y</i> -axis denotes the number of positive samples.	89
4.12	Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for A) unknown primary and B) childhood cancer. The <i>y</i> -axis denotes the number of positive samples.	90
4.13	Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for A) haematological malignancies cancer and B) oral and oropharyngeal cancer. The <i>y</i> -axis denotes the number of positive samples.	91
4.14	Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for sinonasal cancer. The <i>y</i> -axis denotes the number of positive samples.	92
4.15	Sparsity by cancer type from applying SEPATH with Kraken to Genomics England's 100,000 Genomes project whole genome sequence data. The <i>y</i> axis notes the percentage of cells in the taxa-sample community matrix that are occupied by zero values in each cancer type. This is considering all 1,532 genera detected in the entire dataset after a read minimum read threshold of 20 was applied.	93
4.16	Sparsity in Kraken data. Each point represents a genus found in at least one sample. The percentage of positive samples for each taxa in each cancer type, considering 1,532 genera in the entire dataset (minimum read threshold of 20). The number of taxa considered in each cancer type is demonstrated on the <i>y</i> -axis. The line indicates taxa with 50 percent of samples positive. It is important to note that taxa with no samples positive were removed from this analysis. Tumour types are in no particular order.	94
4.17	SEPATH Alphapapillomavirus reads using Kraken and clinical HPV test status. Samples from metastatic tumours are shown in yellow and primary tumours are purple. 38 samples were negative for HPV, 10 were positive. 101 oral tumour samples had no mention of HPV status for which SEPATH identified eight positive samples (minimum number of reads = 115).	95
4.18	Alignment of HTLV-1 classified reads (Kraken) from tumour and germline samples from one participant. The image shows the alignment viewed with IGV. The top track denotes coverage for particular regions (maximum coverage = 13). Coloured regions indicate single nucleotide differences present in the reads and not the reference genomes (orange=G, blue=C, red=T, green=A). In total 172 quality-trimmed, human-depleted reads were subject to alignment (66 and 106 reads from the tumour and germline sample respectively).	97

4.19	The percentage of participants with a positive match for at least one of the ABBS genera at each clinical stage and within each cancer type. Kraken results had a minimum threshold of 10 reads. Please note the <i>y</i> -axis ranges differs between graphs.	98
4.20	A <i>t</i> -SNE plot of viral genera detected by Kraken in the 100,000 genomes project. Detection was limited to nine viral genera (<i>Lymphocryptovirus</i> , <i>Orthohepadnavirus</i> , <i>Roseolovirus</i> , <i>Alphapapillomavirus</i> , <i>Cytomegalovirus</i> , <i>Gammaretrovirus</i> , <i>Alphatorquevirus</i> , <i>Lentivirus</i> and <i>Alphapolyomavirus</i>) and 14 cancer types (glioma, colorectal, lung, ovarian, sarcoma, bladder, endometrial, melanoma, prostate, upper gastrointestinal, breast, hepatopancreatobiliary, oral and renal). Samples with no evidence of these genera were removed from analysis. The data for the plot was a distance matrix of Spearman's correlation values with parameters: random seed=8, dimensions=2, theta=0.2, perplexity=50, maximum iterations=5000. Clusters were determined by a density base clustering algorithm from the dbscan package (parameters: eps=3, minimum points=5).	99
4.21	The distribution of tumour types amongst <i>t</i> -SNE viriome clusters (as in figure 4.20). Numbers at the top of each bar indicate the number of samples within individual tumour types	100
4.22	The distribution of viral genera amongst <i>t</i> -SNE viriome clusters (as in figure 4.20). Numbers at the top of each bar indicates the number of total number of reads assigning to each viral genus.	101
4.23	Principal Variance Component Analysis (PVCA) to estimate the variance explained by biological and technical variables in the dataset of Kraken classifications of tumour samples. Residual variance that is not explained by the variables investigated is denoted 'resid'. GMC = Genomic Medicine Centre	102
4.24	Performance of one-vs-all GBM tumour models on Genomics England Voom normalised Kraken data. Contaminants were removed according to Salter <i>et al.</i> (NPV = negative predictive value, AUC = area under the ROC curve, PPV = positive predictive value).	103
4.25	<i>t</i> -SNE plot of Kraken results coloured by tumour type for the microbial genera that most informative of tumour type. <i>t</i> -SNE was carried out on a matrix of Spearman's correlation values between samples. Informative taxa are defined as those taxa that that were in the top 1500 most important taxa in one-vs-all GBM models. 495 unique taxa met this criteria in 8103 non-FFPE, PCR free, primary tumour samples further details available in 4.8.5.	104
4.26	Supervised normalisation introduces variance into taxa with no reads where variance should not exist. These samples originally had zero reads for Hep-adenosovirus. Disease type is on the <i>x</i> -axis, the <i>y</i> -axis represents Voom-SNM normalised values and points are coloured by sequencing centre. The data used was Voom-SNM normalised Kraken classification data as published by Poore <i>et al.</i> 2020 ¹⁵⁴	107

4.27	Models to predict one tumour type vs all other samples from Poore <i>et al.</i> 2020 were reproduced using different sampling techniques. Performance is summarised here in terms of sensitivity, specificity, AUC, positive predictive value, negative predictive value and F1-score. Sampling techniques are for downsampling (red), ROSE (green), SMOTE (blue) and standard upsampling (purple).	108
4.28	TCGA one vs all tumour models were repeated using only nine nonsensical taxa (Leucothrix, Velarivirus, Tritimovirus, Dinovernavirus, Bacillarnavirus, Rymovirus, <i>Ignicoccus</i> , <i>Salinimicrobium</i> and <i>Thalassomonas</i>) that were deemed the most informative in predicting between tumour types. This data was reanalysed from Voom-SNM normalised data with contaminants removed from Poore <i>et al.</i> 2020.	110
B.1	The choice of k -mer size can influence bacterial and human read depletion. A single simulated dataset was subject to quality trimming and then human depletion using various BB DuK k -mer settings. Additionally, utilising information from single end reads can aid with bacterial read retention. $k=30-32$ appears to be an optimal setting for retaining bacterial reads and removing human reads.	156
B.2	Retention of bacterial reads using different depletion software.	157
B.3	Coverage of contigs following metagenomic assembly on 99 simulated metagenomes. Higher values not shown in density plot.	158
B.4	Violin plot shows genus level performance with increasing minimum contig coverage filter but not to a large degree. Tool used was Kraken on MetaSPAdes contigs.	159
B.5	An in depth look into Krakenuniq filtering parameters vs bacterial classification status for one simulated bacterial dataset.	159
B.6	A more in-depth look into contig parameters vs classification status for one of the viral datasets assembled using MetaSPAdes and classified using Kraken.	160
C.1	Urine RNA sequencing scree plot used to determine the optimal number of clusters.	162
C.2	Analysing differentially present taxa between clusters. Differences in proportions of positive samples between clusters was carried out (Fishers exact test) and is indicated on the y -axis. To investigate directionality, the difference in percentage of positive samples was calculated whereby a value of +1 indicates cluster 'A' samples all contain a taxa and none of the samples in cluster 'B' contain the taxa. The horizontal line indicates $\alpha=0.05$	163
C.3	Urine RNA seq genera diversity appears independent of sequencing depth	164
C.4	Urine 16S PCoA (Manhattan distance). k -means clustering separated the data points into three groups. One of these groups (cluster 1, black) contained an extraordinarily large proportion of metastases as indicated by diamonds on the plot. The first three principal coordinates are shown through A-C.	171

C.5	Box plots to demonstrate differential abundance between the poor prognosis cluster (cluster 1) vs the rest of the dataset (clusters 2 and 3). p -values indicate Wilcoxon signed rank-sum test and Q -values indicate FDR adjusted p -values using Benjamini-Hochberg. Several taxa appear to be higher in abundance in the poor prognosis cluster. These taxa were selected to display the largest differences between clusters.	172
C.6	Targeted alignment survival curve for ABBS bacteria	173
C.7	A volcano plot of human differentially expressed genes in exosome urine RNA sequencing. Samples were partitioned based on the presence of at least one of the ABBS prognostic bacteria. A few genes seem significantly different between samples: <i>PCSK5</i> and <i>ANK1</i> . The horizontal line on the y -axis denotes $p=0.01$. The directionality is such that a positive log fold change indicates a higher expression in the ABBS group.	174
D.1	PCoA of mOTUs2 data using the jaccard distance.	188
D.2	PCoA of mOTUs2 data using the bray-curtis distance.	189
D.3	Total number of reads in each file following quality trimming and human read depletion prior to classification in sequencing files. The bi-modal distribution in total reads appears to be associated with gender.	190
D.4	Low entropy filtering of sarcoma total reads (A) vs high entropy filtering of sarcoma total reads (B). Reads were quality trimmed and human depleted from SEPATH.	191
D.5	A repeat of the TCGA one-vs-all tumour classification models with with stringent filters for contamination removed.	210
D.6	A repeat of the TCGA one-vs-all tumour classification models with with stringent filters for contamination removed. These only include genera found in both SEPATH on and TCGA. Values in the cells denote the change from the models when ran on all genera in the original TCGA publication	211
D.7	A repeat of the TCGA one-vs-all tumour classification models with with stringent filters for contamination removed. These only include genera found in both SEPATH on and TCGA. Values in the cells denote the change from the models when ran on all genera in the original TCGA publication	212
D.8	A repeat of the TCGA one-vs-all tumour classification models with with stringent filters for contamination removed. These only include genera found in TCGA but not in SEPATH. Values in the cells denote the change from the original model	213
D.9	Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	214
D.10	Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	214
D.11	Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	215
D.12	Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	215
D.13	Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	216
D.14	Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	216

D.15 Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	217
D.16 Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	217
D.17 Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	218
D.18 Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	218
D.19 Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	219
D.20 Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	219
D.21 Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	220
D.22 Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	220
D.23 Feature Importance for GBM and Boruta of one-vs-all tumour classification models.	221
D.24 <i>t</i> -SNE plot of kraken results coloured by tumour type for only the microbial genera that are most informative of tumour type. <i>t</i> -SNE was carried out on a matrix of Spearman's correlation values between samples.	225
D.25 The number of microbial genera determined as contaminants after filtering ranked eisenhofer and boruta lists for their top quantiles of contaminants. The colour of the points resembles the number of unique genera as indicated in the legend.	227
D.26 A sample of the results from searching for structure by disease type, sex and ethnicity. This particular result includes the top 30 percent of estimated contaminants removed and all of contaminants in Eisenhofer et al were removed. Different kraken read thresholds were applied. This shows the parameters with the highest estimated performance for disease type.	228
D.27 Results from running Kraken on Tumour Samples from the 100,000 Genomes Project. The <i>y</i> -axis denotes the total number of classified reads and the <i>x</i> -axis denotes the mutational burden in terms of somatic coding variants per megabase (mb).	229
D.28 The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	230
D.29 The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	231
D.30 The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	232

D.31	The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	233
D.32	The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	234
D.33	The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	235
D.34	The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	236
D.35	The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	237
D.36	The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	238
D.37	The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	239
D.38	The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	240
D.39	The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	241
D.40	The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	242
D.41	The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible. . . .	243

Acronyms

5-FU - 5-Fluorouracil

A - Adenine

ABBS - Anaerobic Bacteria Biomarker Set

AF - Attributable Fraction

ATP - Adenosine Triphosphate

AUC - Area Under a receiver operating Characteristic (ROC) curve

BCG - Bacillus of Calmette and Guerin

BCR-ABL - Breakpoint Cluster Region, Abelson Fusion Protein

BEAR - Better Emulation for Artificial Reads

BP - Biological Process

BRCA - Breast Cancer Gene

BSS - Between-Cluster Sum of Squares

C - Cytosine

CAMI - Critical Assessment of Metagenomic Interpretation

CAR T-Cells - Chimeric Antigen Receptor T-Cells

CC - Cellular Component

CML - Chronic Myeloid Leukaemia

CNV - Copy Number Variation

COSMIC - Catalogue of Somatic Mutations in Cancer

CPU - Central Processing Unit

dCMP - deoxycytidine monophosphate

ddNTP - dideoxynucleotide triphosphate

DNA - Deoxyribonucleic Acid

DRE - Digital Rectal Exam

EBV - Epstein-Barr Virus

ER - Oestrogen Receptor

FAD - Flavin Adenine Dinucleotide

FAD-TS - Flavin Adenine Dinucleotide Dependent Thymidylate Synthase

FDR - False Discovery Rate

fdUMP - 5-Fluoro-2-deoxyuridine 5-monophosphate

FFPE - Formalin-Fixed Paraffin-Embedded

G - Guanine

GB - Gigabytes

GeCIP - Genomics England Clinical Interpretation Partnership

GI - Gastro-Intestinal

GIST - Gastrointestinal Stromal Tumours

GMC - Genomic Medicine Center

GR - Genomic Rearrangement

HBV - Hepatitis B Virus

HCV - Hepatitis C Virus

HER2 - Human Epidermal Growth Factor Receptor 2

HERV - Human Endogenous Retrovirus

HHV8 - Human Herpes Virus-8

HIV - Human Immunodeficiency Virus

HMP - Human Microbiome Project

HPC - High Performance Computing Cluster

HPV - Human Papillomavirus

HTLV - Human T-cell Lymphotropic virus

ICGC - International Cancer Genome Consortium

IQR - Inter-Quartile Range

LCA - Lowest Common Ancestor

LHRH - leutenizing hormone releasing hormone

MAG - Metagenome Assembled Genome

MCF - Mean Covered Fraction

MCV - Merkel Cell Polyoma Virus

MDS - Classical Multi-Dimensional Scaling

MDSC - Myeloid-Derived Suppressor Cells

MF - Molecular Function

MHC - Major Histocompatibility Complex

MIF - Macrophage Migration Inhibitory Factor

MKF - Mean k-mer fraction

MRI - Magnetic Resonance Imaging

NCBI - National Center for Biotechnology Information

NGS - Next Generation Sequencing

NHS - National Health Service

NICE - National Institute for Clinical Excellence

NMDS - Non-Metric Multidimensional Scaling

NNUH - Norfolk and Norwich University Hospitals

OTU - Operational Taxonomic Unit

PARP - Poly ADP-Ribose Polymerase

PCA - Principal Components Analysis

PCoA - Principal Coordinates Analysis

PCR - Polymerase Chain Reaction

PDGFRA - Platelet-Derived Growth Factor Receptor A

PPV - Positive Predictive Value (precision)

PR - Progesterone Receptor

PRIMA-1 - P53 Reinduction of massive Apoptosis

PSA - Prostate Specific Antigen

RAM - Random Access Memory

RNA - Ribonucleic Acid

ROC - Receiver Operating Characteristic

SCNA - Somatic Copy-Number Alteration

SEPATH - SEarching for PATHogens computational pipeline

SGR - Somatic Genome Rearrangement

SHP2 - Src homology region 2-containing protein tyrosine phosphatase 2

SMOTE - Synthetic Minority Oversampling TEchnique

SNM - Supervised Normalisation
SNP - Single Nucleotide Polymorphism
SSV - Sensitivity (recall)
SV - Structural Variant
T - Thymine
TCGA - The Cancer Genome Atlas Program
TRUSS - Transrectal Ultrasound Biopsy
TS - Thymidylate Synthase
TSG - Tumour Suppressor Gene
UTI - Urinary Tract Infection
UTR - Untranslated Region
VCF - Variant Call Format
WGS - Whole Genome Sequencing
WSS - Within-Cluster Sum of Squares

Acknowledgements

I am fortunate to have been supported by an excellent supervisory team throughout the course of this project. Professor Daniel Brewer, thank you for being the ‘*unicorn*’ supervisor, for all of your guidance, feedback, support and for treating me as a colleague from the start. Dr Rachel Hurst, thank you for your insights and discussions. Your passion has been nothing short of inspirational. Professor Colin Cooper, your research wisdom is astounding. Thank you for your guidance at some key moments in this project. Dr Richard Leggett, thank you for stepping in as secondary supervisor, for welcoming me into your group meetings and for all of your valuable feedback and expertise. Dr Ghanasyam Rallapalli, thank you for being there in the beginning and ‘showing me the ropes’ as I adjusted to the world of bioinformatics.

Thank you to the Big C cancer charity, their donors and to Genomics England for making this research possible. Thanks to the system administrators from the various computing clusters, some of which I’m sure I have driven insane at certain points. Thank you to the tens of thousands of patients that have selflessly donated biological samples to scientific research.

I would additionally like to mention my appreciation for my loving partner Jessica Purcell for her support throughout this project. Thank you for being there and for being so understanding throughout these years.

Finally, I would like to dedicate this thesis to two people that have always been there and have offered their unparalleled support, my parents. Firstly thanks to my father. A man whom I am honoured to share the name of, Farag Salah Gihawi. Thank you for always showing an interest in my work and for being the original muse behind my gravitation towards the empirical sciences. Last but certainly not least, thanks to my mother Deborah Gihawi for instilling a curious nature within me and being one of the reasons behind my interest in cancer.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other university or other institute of learning.

Abraham Farag Salah Gihawi

November 2020

Chapter 1

Background

Microorganisms including bacteria and viruses have been linked with the development of cancers such as cervical, gastric, hepatocellular and bladder cancer¹⁻⁴. It is hypothesised that there are a range of other microbes that are associated with cancer. The overarching theme of this thesis is to investigate and search for novel associations between microbes and cancer by analysing sequence data. Computational pipelines are applied to various sources of sequencing data including Genomics England's 100,000 Genomes Project and urine from participants under investigation for prostate cancer. This chapter will provide the biological and computational background relevant to this thesis.

1.1 Biological Background

1.1.1 Introduction to Cancer

Cancer is a condition whereby cells are uncontrollably proliferative compared to the surrounding cells. This rampant growth results in a physiological effects detrimental to the hosts quality of life and often proves fatal (table 1.1). Cancer is a disease of international concern with 14 million new cases and 8.8 million deaths annually. These figures are projected to rise as life expectancy increases as a consequence of economic development⁵. The broadest possible partition of tumours is that of benign vs (pre-)malignant. Tumours occur when uncontrolled cellular growth arises in solid tissue. Benign tumours are not considered cancerous, are not invasive, do not spawn metastases and are composed of hyperproliferative yet comparatively functional cells. Pre-malignant cells are predominantly benign in nature yet are morphologically abnormal and possess the ability to transition into a malignant phenotype which can spawn metastases. Metastasis occurs when malignant cells spread to one or more different anatomical sites accounting for 90% of cancer related deaths⁶. The current mortality rates necessitate improvements in therapeutics. To do so, we must understand more about the fundamentals of tumour biology.

Table 1.1: Physiological impacts of tumours. Adapted from Ruddon 2017⁷.

Effect	Manifestations	Causes
Anaemia	Pallor, weakness, fatigue	Bone marrow invasion affecting haematopoiesis
Hypercoagulability	Thrombophlebitis, disseminated intravascular coagulation, haemorrhage, renal failure	Release of thromboplastin-like substances
Cachexia	Loss of fat/muscle, general weakness	Anorexia, nausea, malabsorption, involvement of digestive organs, metabolic demands of tumour and host immune system
Fever	Sweating, malaise, confusion	Infection or pyrogenic substances released by tumour
Hormone release	Various. e.g., Cushing's, water retention, gynecomastia, precocious puberty, hypercalcaemia, hypoglycaemia	Tumour hormone production
Hypercalcaemia	Lethargy, weakness, nausea, confusion, coma, decreased gastrointestinal motility, cardiac arrhythmias, kidney stones	Tumour invasion/metastasis to bone, parathyroid hormone related protein release, tumour production of osteolytic substances (prostaglandins and osteoclast stimulation)
Leukopenia	Infection	Crowding out normal marrow cells, chemo/radiotherapy, obstructing cells in spleen
Neurological symptoms	Muscle weakness, decreased reflexes, atrophy	Cachexia, chemotherapy, metastases to or involvement with neurological tissue
Pain	Pain dependent on tumour location	Tissue destruction, obstruction, pressure on organs, infection
Thrombocytopenia	Petechia, purpura, internal haemorrhage	Crowding of marrow, chemotherapy, radiotherapy, immunosuppression, hypercoagulability

1.1.2 The Hallmarks of Cancer

Cancer is a disease of genomics with tumours arising when one cell (monoclonal) or more than one cell (polyclonal) gain a survival advantage and growth exceeds the surrounding cells of the same type. For a tumour to grow, this excess proliferation is insufficiently modulated by the host immune system⁶. Each of our cells contains approximately 6×10^9 nucleotide base pairs of deoxyribonucleic acid (DNA), arranged into long strands wrapped around histones to form chromosomes⁸. These base pairs can be one of four nucleotides: Adenine (A), Guanine (G), Cytosine (C), Thymine (T). The central dogma of molecular biology is that DNA is transcribed into RNA, which is translated into proteins that carry out cellular functions. Therefore, changes in these nucleotides alter protein function and cellular characteristics. Modifications to the genome (DNA within a cell or population of cells), or epigenome (non-base changes to DNA that impact expression) of a cell is inherited by subsequent generations.

Cancer genetics includes germline features (parental in origin) and somatic alterations (features that arise throughout the life of a cell or organism). Somatic features can be acquired cumulatively with age, through misreplication or through environmental factors such as radiation damage (*i.e.*, ultraviolet light), pathogens or smoking. Most mutations negatively impact cellular survival but occasionally a mutation can confer a selective advantage and evade correction by host repair mechanisms. Mutations providing cells with a survival advantage are termed ‘*driver*’ mutations⁹. ‘*Passenger*’ mutations can also occur which are carried through generations of cells but are not considered to be a driving force behind tumour characteristics. ‘*Driver*’ mutations typically occur in either tumour suppressor genes that ordinarily act to suppress excessive cellular growth, or in oncogenes that when mutated often propel cellular growth. In 2000 and 2011, Hanahan and Weinberg summarised the different characteristics by which tumours are adapted to survival in their environment^{10,11} (figure 1.1).

	Resisting Cell Death		Enabling Replicative Immortality
	Genome Instability		Avoiding Immune Destruction
	Inducing Angiogenesis		Evading Growth Suppressors
	Invasion and Metastases		Sustaining Proliferative Signaling
	Tumour Promoting Inflammation		Deregulating Cellular Energetics

Figure 1.1: The Hallmarks of Cancer. Adapted from Hanahan and Weinberg 2011¹¹.

The hallmarks of increased angiogenesis and dysregulated cellular energetics provide aggressive tumours with the nutrients and biochemical pathways to grow uninhibited. Tumours are able to circumvent the mechanisms that ordinarily suppress excess growth, including immune system modulation, evading apoptosis and direct methods of growth suppression. This is exacerbated by tumour promoting inflammation and tumours becoming increasingly genomically unstable. Unfortunately, genome instability combined with rapid division leads

to tumours becoming more aggressive with time.

It is possible to obtain representations of tumour DNA with whole genome sequencing. Efforts to comprehend the pan-cancer genome and how to use this information are still underway. A recent report on pan-cancer genomics has highlighted that on average, tumours tend to have 4.6 distinct driver mutations and that some driver mutations remain undiscovered¹².

Tumour suppressor genes, *TP53*, *CDK2NA/B*, *SMAD4*, *PTEN* and *RB1* are among the most frequently altered genes in cancer. The most common types of change are single nucleotide variations, small insertions & deletions, and copy number deletions. Alterations in genes that result in a gain of function (oncogenes *i.e.*, *PIK3CA*, *BRAF*, *CTNNB1*, *ERG* and *MYC*) are less common^{12,13}.

1.1.3 Cancer Therapeutics

Cancer has been treated surgically for thousands of years¹⁴ until the advent of radiotherapy in the early 20th century¹⁵, followed by chemotherapy circa 1960¹⁶. Surgery is still considered the mainstay of cancer therapeutics, with radiotherapy and chemotherapy considered supplementary to surgery, either before surgery (neoadjuvant), after surgery (adjuvant) or for inoperable cancers.

Tying together molecular biology with pharmacology has yielded some clinical benefit in recent years. Therapeutically targeting loss-of-function mutations has a limited number of examples. One such example is PRIMA-1 (P53 Reinduction of Massive Apoptosis) re-activates mutant P53 *in vivo* and is the subject of ongoing trials¹⁷. Another approach is the concept of synthetic lethality which exploits the mutational state of tumours to over-stress a particular pathway, leading to cell death. This is the mechanism by which PARP (poly ADP-ribose polymerase) inhibitors function in some *BRCA* mutant cancers¹⁸.

Pharmacologically targeting gain of function mutations has been more successful in comparison. Imatinib, an orally administered drug with an affinity for multiple overactive tyrosine kinases has contributed to complete haematological responses of >95% in CML¹⁹ and gastro-intestinal stromal tumours have seen some benefit in the neo-/adjuvant setting²⁰. *PI3K* and *BRAF* are frequently mutated in cancer and have efficacious drugs but have not boasted such high response rates. Vemurafenib has been trialled in metastatic *BRAF*-mutant melanoma with incredible initial responses but these were short lived due to the rapid acquisition of resistance²¹. To target both acquired and intrinsic resistance, trials are currently turning to combination therapy to either achieve a synergistic effect or to target common resistance pathways²².

Immunotherapy is an approach with some promising results such as PD-1/PD-L1 (programmed death-ligand) therapeutics entering clinical practice which stimulate anti-tumour immune responses²³. Another modern immunotherapy is CAR T-cells (Chimeric Antigen Receptor) whereby a patient's T-cells are genetically altered *ex vivo* to provide an affinity for a particular tumour marker²⁴. These T-cells are propagated and transplanted into the patient with the hope of stimulating a sustained immunological response against the tumour. This technique has demonstrated 90% response rates in relapsed/refractory Acute Lymphocytic Leukaemia (ALL)²⁴.

Perhaps the longest marriage between molecular biology and clinical practice is shown

by breast cancer immunohistochemical markers including ER (Oestrogen Receptor), PR (Progesterone Receptor) and HER2 (Human Epidermal Growth Factor Receptor 2). ER and HER2 are frequently overamplified proteins with successful therapeutics (*e.g.* tamoxifen/fulvestrant and trastuzumab)^{25,26}. These markers coincide with molecular subtypes investigated out by Sørlie *et al.* 2003²⁷. This stratification alongside developments such as improved imaging has led to significant improvements in breast cancer survival²⁸. Despite a more complete picture of breast cancer, there still exist cases with a degree of clinical equipoise such as identifying women at risk of recurrence and those that are over-diagnosed^{28,29}.

Even with recent developments, understanding more about tumour biology remains critical for the development of novel therapeutics and biomarkers. This could lead to the ultimate goal of medicine, which is eradication or prevention of disease before it occurs. Human papillomavirus (HPV) vaccinations have achieved this by preventing the majority of cervical cancers^{30,31}. Identifying additional associations between microbes and cancer could contribute to improved patient care.

1.1.4 Microbes and Cancer Introduction

The association between microbes and cancer has long been a subject of scientific interest. In 1776, Bernard Peyrilhe was convinced that a virus caused cancer and investigated injecting emulsions of human breast cancer into dogs to reproduce tumours³². In 1911, Rous' transmissibility experiments would subsequently result in the discovery of the Rous sarcoma virus³³. Despite this, it would take until around the year 2000 and significant advances in microbiology to fully establish associations between cancer and microbes³⁴. The most significant microbe that causes cancer (oncomicrobe) is the bacterium *Helicobacter pylori*, attributed with 770,000 cases of gastric noncardia carcinoma in 2012^{2,35}. The next four most prominent oncomicrobes as estimated by Plummer *et al.*³⁵ for 2012 were: HPV (640,000 cancers), hepatitis B virus (HBV, 420,000 cancers), hepatitis C virus (HCV, 170,000 cancers) and Epstein-Barr Virus (EBV, 120,000 cancers)³⁵. Established oncomicrobes are shown in table 1.2 and some emerging causes are shown in table 1.3.

15.4% of all cancers globally (2.2 million) in 2012 could be attributed to infection with an attributable fraction (AF) of <5% for more economically developed countries, and an AF >50% in some lesser economically developed countries³⁵. Therefore, if infection were listed as a single type of cancer, it would equate to the most common type of cancer, exceeding lung cancer by 400,000 annual cases. These incidence rates may be underestimates, as they retain a focus on only include the most common oncomicrobes. It is loosely estimated that less than 1% of microbial species have been discovered and so there is the potential that undiscovered microbes might be associated with cancer³⁶.

Microbes have been associated with tumorigenesis by a variety of methods, including epidemiological studies, molecular level studies, animal models, cell culture experiments and, in the case of Merkel Cell Polyoma Virus (MCV), a mixture between laboratory and computational techniques³⁷. Moving from associations to causality is a difficult step. Historically, causality has been established with Koch's postulates whereby an infectious agent hypothesised to cause a disease must be isolated from a diseased host, propagated *in vitro* and re-inoculated into a disease-free host to give rise to the same condition³⁸. This presents insurmountable limitations to establishing some oncomicrobes as some species are impossible

to propagate *ex vivo*³⁸. Additionally, some oncomicrobes may require co-occurring microbes, environmental co-factors or certain host factors (*i.e.*, immune status or genetic predispositions) to give rise to tumours^{37–39}. The International Agency for Research on Cancer (IARC) has formally recognised ten infectious agents as group 1 carcinogens: *H. pylori*, HBV, HCV, HPV, EBV, HHV-8, HTLV-1 as well as three parasitic species: *Opisthorchis viverrini*, *Clonorchis sinensis*, and *Schistosoma haematobium*⁴⁰.

Table 1.2: Pathogens with strong evidence of association with cancer.
Adapted from Pappas 2009⁴¹ and Zhu *et al.* 2017⁴².

Pathogen	Cancer	Annual.Cases
<i>Epstein–Barr virus (EBV)</i>	Lymphoma, Nasopharyngeal,	100,000
<i>Helicobacter pylori</i>	Gastric Gastric, Gastric mucosal associated lymphoid tissue lymphoma	490,000
<i>Hepatitis B virus (HBV)</i>	Hepatocellular	280,000
<i>Hepatitis C virus (HCV)</i>	Hepatocellular	110,000
<i>Human herpes virus-8 (HHV8)</i>	Kaposi’s sarcoma, primary effusion lymphoma	55,000
<i>Human papilloma virus (HPV)</i>	Cervical, head & neck, anal, vulvar, vaginal, penile	550,000
<i>Human T-cell Lymphotropic virus (HTLV)</i>	Acute T-cell leukaemia	2,700
<i>Opisthorchis viverrini</i>	Cholangiocarcinoma	800
<i>Schistosoma haematobium</i>	Bladder	9,000

Table 1.3: Microbes with emerging evidence of association with cancer.
Adapted from Pappas 2009⁴¹ and Zhu *et al.* 2017⁴².

Microbe	Cancer
<i>Clonorchis sinensis</i> / <i>Opisthorchis felinus</i>	Cholangiocarcinoma
<i>Human Immunodeficiency virus (HIV)</i>	Lymphomas, Kaposi's sarcoma
<i>Polyomaviruses</i>	Brain, Osteosarcoma, mesothelioma, Merkel cell
<i>Schistosoma japonicum</i>	Colorectal, hepatocellular
<i>Fusobacterium nucleatum</i>	Colorectal, Oral, Pancreatic
<i>Porphyromonas ssp.</i>	Colorectal, Oral, Pancreatic
<i>Streptococcus</i>	Pediatric acute myeloid leukemia, breast, Leukemia, lymphoma, colorectal
<i>Enterococcus faecium</i>	Haematological malignancies
<i>Salmonella typhimurium</i>	Gallbladder
<i>Human Endogenous Retroviruses (HERV)</i>	Breast, choriocarcinoma, melanoma, myeloproliferative disorders, ovarian, prostate, small cell lung, teratocarcinoma

1.1.5 Molecular Aetiology of Oncomicrobes and Cancer

Oncomicrobes can contribute to carcinogenesis using various mechanisms. This section describes the main mechanisms using two examples *H. pylori* and HPV. These well studied microbes are associated with the most number of cases of pathogen associated cancer at over one million annually between them (table 1.2).

Human Papillomavirus

Of the hundreds of HPV subtypes, only a few are associated with cervical carcinoma. Most notably, subtypes 16 and 18 are estimated to cause 70% of all cervical cancer³¹. These high-risk subtypes have developed solutions that increase the local population of epithelial cells in which they reside in by two viral proteins, E6 and E7^{31,43}. E6 binds to P53, targeting it for ubiquitin mediated proteosomal degradation and E7 inactivates the retinoblastoma protein; thereby hampering two potent tumour suppressors⁴⁴. It is noteworthy that HPV⁺ oropharyngeal tumours demonstrate little evidence of somatic *TP53* mutations^{45,46}. Due to the shared characteristics of HPV⁺ oral cancers, HPV has been suggested as a biomarker for treatment stratification⁴⁷.

Other effects of HPV on the host cell relating to E6/7 include accumulation of mutations, inhibition of differentiation and inhibition of the hosts immune response⁴⁴. Another sophisticated mechanism of HPV is that its genome is often integrated into the host genome, rendering it difficult to cure pharmacologically⁴⁴.

Helicobacter pylori

Inflammation is a non-specific mechanism of *H. pylori* that has been causally related to carcinogenesis via replication stress - a cycle of DNA injury and repair with inhibited DNA replication^{48,49}. This leads to increased genomic instability in human cells which

eventually causes aberrant proliferation and increased invasion and metastasis^{48,49}. More specific mechanisms pertain to particular oncoproteins. The most notorious of which is CagA which is directly injected into gastric epithelial cells by *H. pylori*⁵⁰. CagA is then phosphorylated by the host which then binds and activates SHP2 (Src homology region 2-containing protein tyrosine phosphatase 2). This increases the half-life of activated RAS (GTP bound RAS) which in turn leads to elevated ERK activation and subsequently cellular proliferation^{50,51}. CagA also interacts with host E-cadherin, leading to an accumulation of β -catenin in the cytoplasm and nucleus⁵². This accumulation means that E-cadherin is free to activate Wnt responsive genes, increasing proliferation and differentiation⁵³.

1.1.6 Emerging Mechanisms

Other mechanisms exerted by oncomicrobes are less specific such as HIV, which is able to permit tumours purely by suppressing the host immune system⁵⁴. Like *H. pylori*, *Fusobacterium nucleatum* activates β -catenin and Wnt signaling, providing evidence for its association with cancer⁵⁵. Additionally, *F. nucleatum* activates TLR4 leading to an increased expression of an oncogenic microRNA (Ribonucleic Acid) providing another strand of evidence⁵⁵. It is clear that microbes influence a number of cancer hallmarks. A more complete review is provided by Fulbright *et al.* 2017⁵⁶.

1.1.7 The Human Microbiome

The term ‘microbiome’ describes the collective microbial constituents of a particular species or sampling site. The gastro-intestinal (GI) tract has long been a site of microbiome research and demonstrates microbiological communities along its entirety⁵⁷. Microbes are ubiquitous in humans and this has led to interest in what constitutes a ‘normal’ microbiome and investigating dysbiosis with disease.

Historically, scientists have investigated the microbiome by culturing which can skew observations towards the species that are easier to culture. Additionally, obtaining pure isolates can be difficult and is not free from the risk of contamination. It wasn’t until 1996 that a culture independent method was trialled to characterise the faecal microbiome, showing a good concurrence with traditional culture-based methodology⁵⁸. One culture independent method focuses on analysing amplicons of the gene encoding the small sub-unit (16S) of prokaryotic ribosomal ribonucleic acid (rRNA). This was selected because generally rRNA is relatively highly conserved between bacterial species, is usually present as multiple copies of the gene in bacterial species and is readily isolated from samples with eukaryotic contamination such as diagnostic samples.

One of the earliest landmark papers utilising 16S gene sequencing (or 16S sequencing in short) was by Costello *et al.* 2009⁵⁹. They found that most sequences (92.3%) were related to *Actinobacteria*, *Firmicutes*, *Proteobacteria* and *Bacteroidetes*. Some body sites such as the gut contained highly distinctive microbial communities that were highly individual and stable over months⁵⁹.

As costs have declined, Whole Genome Sequencing (WGS) has emerged as an alternative sequencing approach and has some useful advantages over 16S sequencing. WGS can obtain the sequence from an entire microbiome (including viruses, fungi *etc.*), rather than specific regions of bacteria. This provides a wider variety of sequence data for classification and so a higher degree of differentiation can be achieved (higher genomic resolution), potentially

providing more confidence in a particular classification. Despite reducing costs, WGS is also significantly more expensive compared with 16S sequencing and has fewer established computational pipelines^{60,61}.

Since then, sequencing methods have been expanded to larger cohorts such as the Human Microbiome Project (HMP). The HMP was established in 2008 to characterise the landscape of commensal micro-organisms associated with humans by examining samples from 300 healthy individuals over multiple anatomical sites^{62,63}. The project analysed over 5,000 16S rRNA samples using Illumina 454 reads (V1-3 and V3-5 hypervariable regions) and over 700 Illumina WGS samples⁶³. The latter were subject to metabolic reconstruction and metagenomic assembly. The second phase in 2014 aimed to characterise the role of the microbiome in health and disease.

The human body is host to an incredibly diverse array of microbes with each anatomical site displaying differences in microbial community (figure 1.2). The most abundant genera across the dataset include *Lactobacillus*, *Propionibacterium*, *Streptococcus*, *Bacteroides*, *Corynebacterium*, *Staphylococcus*, *Moraxella*, *Haemophilus*, *Prevotella* and *Veillonella*. The skin and anterior nares demonstrated a high abundance of *Propionibacterium*, *Corynebacterium* and *Staphylococcus*. Oral samples contained a high abundance of *Streptococcus*, *Haemophilus*, *Prevotella* and *Veillonella*. Stool samples were dominated by *Bacteroides* and vaginal samples by *Lactobacillus*.

The microbiome stabilises within the first year of life⁶⁴ but varies with many factors such as: age, gender, diet, pharmaceuticals, geographical location/ethnicity, mental health status^{65–71}.

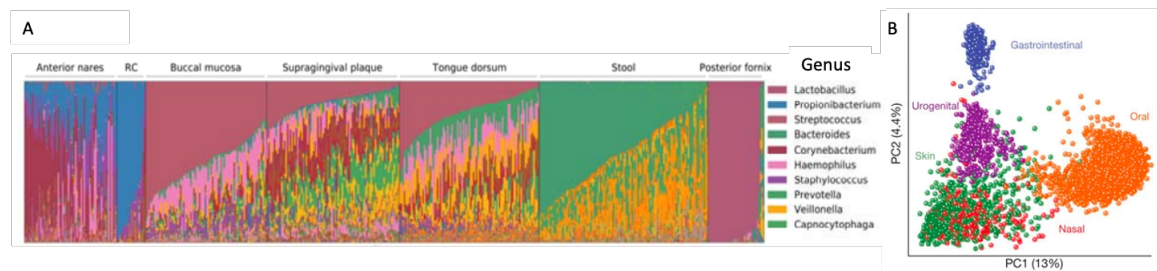


Figure 1.2: Overview of the Human Microbiome Project results. **A)** A bar plot demonstrating relative abundance at genus level for each sample partitioned by sampling site. **B)** A principal coordinates plot demonstrating that samples from the same body sites tend to share similar constituent micro-organisms. Retroauricular crease (RC). Image adapted from The Human Microbiome Project Consortium 2012⁷².

1.1.8 Microbes and Therapeutic Approaches

Two vaccines (Cervarix and Gardasil) were introduced into the UK population in 2008 which immunise against HPV capsid proteins and have proven successful⁷³. It is estimated that vaccination will prevent up to 80% of cervical cancer^{30,31}. A recent study with 11 years of follow-up of 1.5 million females in Sweden has shown that the risk of cervical cancer by age 30 was 63% lower in vaccinated women⁷⁴. Therefore, offering the vaccine to all females as is the current clinical standard and potentially expanding to males is expected to have a

huge effect on saving lives.

Apart from the successful example of HPV, modulating the microbiome for therapeutic benefit is still largely considered experimental. Some species of bacteria commonly found in the intestine can reactivate metabolites of the topoisomerase inhibitor Irinotecan: a chemotherapy used to treat a range of cancers⁷⁵. To avoid dose-limiting toxicities, patients are prescribed antibiotics prior to treatment⁷⁵. This might be the strongest example of microbiome modulation in clinical practice with regards to cancer. Although clearly associated with tumorigenesis, prophylactic eradication of *H. pylori* remains largely controversial because of concerns over effectiveness in the face of growing antibiotic resistance and the impact on the gastrointestinal microbiome overall^{76,77}.

An emerging point of view is that modulating the gut microbiome might affect cellular characteristics at distant sites⁷⁸. Some studies have suggested that taxa in the gut alter systemic levels of metabolites such as oestrogens and therefore may influence breast cancer development⁷⁸. Modulating the gut microbiome through faecal transplants is another possibility under consideration^{79,80}.

Discovering additional microbes associated with cancer is likely to reveal many more mechanisms that are primed for therapeutic exploitation. One of the difficulties in doing so is picking apart the commensal organisms and the microbes that are non-causally associated with different disease states from the pathogens that have profound pathophysiological effects on host cells^{81,82}.

1.1.9 High Throughput Sequencing Technology

In 1953 Watson and Crick proposed the alpha helix structure of DNA⁸³. Techniques to identify the sequences of the four nitrogenous bases of DNA would not follow until the mid 1970's when two were developed: Sanger and Maxam-Gilbert sequencing⁸⁴. Maxam-Gilbert sequencing uses radioactive isotopes and hazardous chemicals that cleave single stranded DNA at specific points, depending on their nucleotide sequences. Sanger sequencing utilises radiolabelled dideoxynucleotide triphosphate (ddNTP) molecules lacking a 3'-hydroxyl group. The logic being that ddNTPs are incorporated randomly into a developing strand, preventing elongation at each position, thereby producing fragments that can be sorted by size to identify the base at each position. Sanger sequencing gained popularity and has since been refined. Fluorescently labelled bases reduced the number of reaction vessels from four to one⁸⁶ and capillary electrophoresis is used rather than gel⁸⁶. Additionally, thermostable DNA polymerase I is used from *Thermus aquaticus*, preventing repeated replenishing after heating steps⁸⁷.

In 2005, Roche ushered in the second generation of sequencing technology by utilising pyrosequencing⁸⁶. Since then Illumina (which acquired Solexa) have become a dominant force in today's sequencing market with a different approach to second generation sequencing (figure 1.3). The process involves three main steps: library preparation where sample DNA is fragmented and bound to adapter sequences; cluster/bridge amplification whereby the fragments are bound to a flow cell and amplified into clonal clusters of sequences; followed by sequencing whereby the individual nucleotides of each cluster is determined. Figure 1.4 reflects the throughput around the time that Illumina was selected as the partner to sequence samples of the 100,000 Genomes Project analysed later in this thesis.

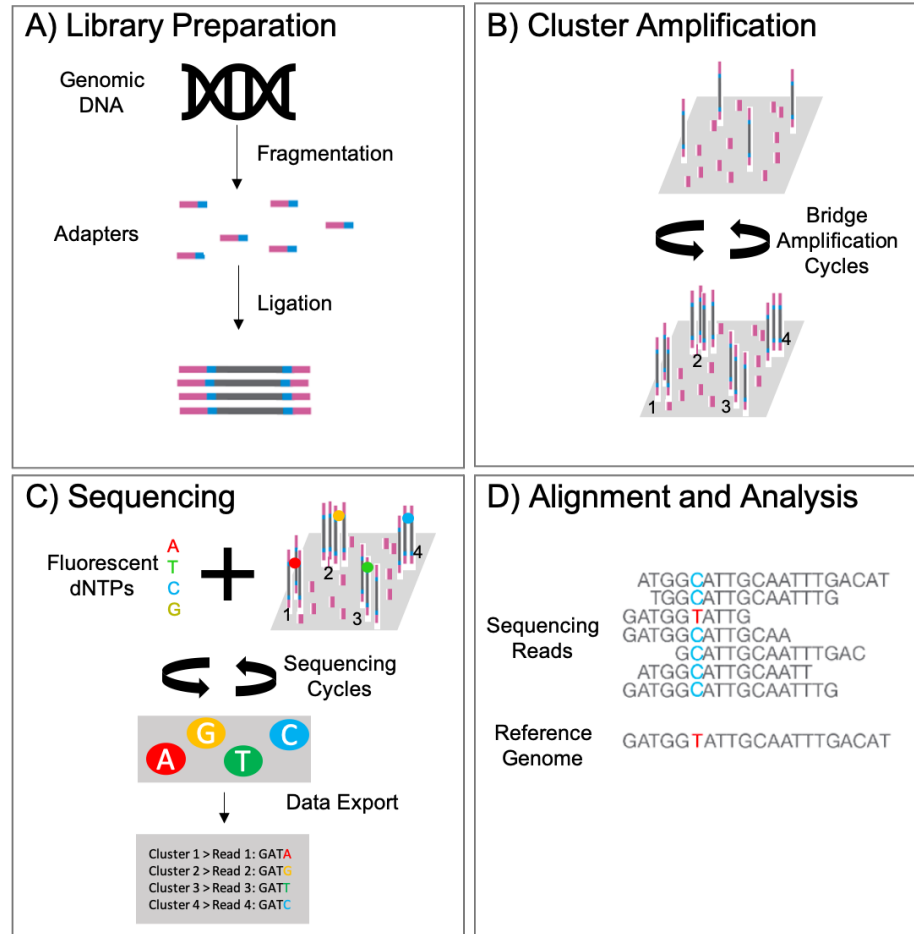


Figure 1.3: The process of Illumina sequencing. **A)** Library Preparation. Sample DNA is fragmented and adapters added to the ends. **B)** Cluster Amplification. Fragments are attached to the flow cell by binding adapters to oligonucleotides on the flow cell. DNA fragments are bent and amplified by PCR to form a clonal cluster of DNA fragments in a specific region on the flow cell. **C)** Sequencing. Illumina uses a 'sequencing by synthesis' approach whereby fluorescent dNTPs are added sequentially to a growing DNA fragment. Software records the colour from each cluster and assigns a nucleotide forming a 'read' with quality (Phred/ Q) scores that estimate the accuracy of base calling. **D)** Alignment and Data Analysis. The reads for all the fragments can then be subject to downstream analysis including assembly and/or alignment to a reference genome. The number of times a nucleotide in a single position is measured is known as the coverage. Image adapted from Illumina 2018⁸⁸.

Third generation sequencing is defined by sequencing single molecules without the need for amplification. Two popular third generation sequencing technologies include the Single Molecule Real Time (SMRT) platform, from Pacific Biosciences (PacBio) and nanopore sequencing, employed by Oxford Nanopore Technologies (ONT)⁸⁹. ONT's MinION sequencer has potential field applications due to its handheld size and potential for megabase reads

(longest reported so far is 2Mb), which can massively improve the contiguity of *de novo* assembly compared to short-read sequencing technology⁸⁹. Typically, MinION sequencing is more expensive per gigabase to sequence than Illumina and has a lower throughput. These characteristics make Illumina a good candidate for bulk sequencing projects; although these drawbacks are diminishing^{90,91}. In today's climate, strengths of third generation sequencing technologies have emerged over Illumina technologies⁹² (figure 1.4).

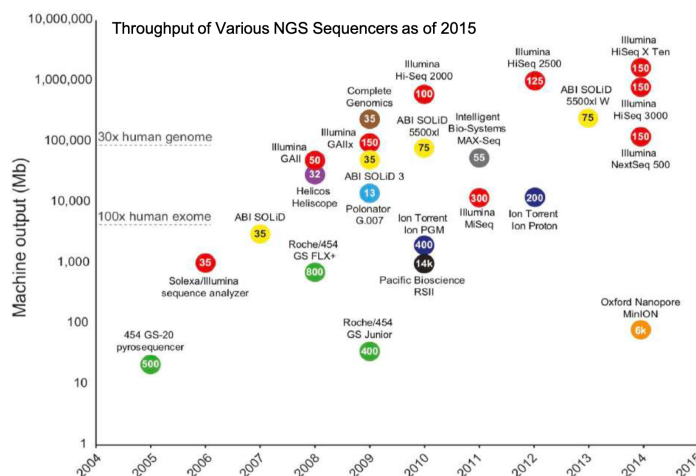


Figure 1.4: The output of different sequencing technologies in megabases. The number inside each point indicates typical read length. This diagram shows that Illumina has a very high throughput validating their use in a large-scale sequencing project. Image adapted from Reuter *et al.* 2015⁹³.

Cancer Tissue Sequencing

Having a deep coverage and high quality sequences is particularly important for cancer sequence data where mutations and a mixture of clones are expected and so deep sequencing is required for good sequence confidence¹¹. In addition to the issue caused by multiple heterogenous cancer subclones, all cells in the immediate cancer microenvironment will also be sequenced. This is sometimes considered to be a hindrance in obtaining good quality sequence data. These cells surrounding the tumour are diverse and include cells such as “*normal*” endothelial cells and immune cells⁹⁴. Some techniques such as laser microdissection can help increase the purity of cancer cells in the sample, but overall, it is expected that there is unavoidable interference from other cell types. One benefit of this is that any microbes close to the cancer cells may be sequenced along with the human cells in the sample.

The whole genome sequencing of cancer samples has been expanded in recent years. This has been aided by the exponentially decreasing cost of sequencing (from \$300 million to \$1000)⁹⁵ and large scale sequencing projects being established such as TCGA, ICGC, and Genomics England (section 4)⁹⁶. There is therefore a unique opportunity to investigate further any evidence for microbial DNA captured along with tumour samples.

1.2 Computational Background

This section provides an overview of the computational approaches used in microbial ecology. There are many methods of investigating the taxonomic breakdown of microbial communities; which to use depends on a number of factors including the nature of the data to be analysed, the research questions and the computing resources available.

The differential abundance of specific taxa between diseased and normal states can lead to insights into disease. Differences in the total number of species can indicate a dysbiosis which could be linked to disease⁴⁸. Lastly, overall microbial compositions can differ between diseased and normal states⁹⁷.

1.2.1 An Overview of Computational Methods of Taxonomic Classification

The term ‘metagenome’ describes sequence data obtained from multiple organisms and was first coined by Handelsman *et al.* 1998 to refer to the collective genomes present in soil microflora. Since then, the study of metagenomics has increased to become its own specialism. In recent years there has been an explosion of computational tools to classify metagenomic sequence data. A non-exhaustive list of computational approaches is provided in Appendix section A.2.2. These tools can be organised into metagenome assemblers (with or without associated classification), genome binners, taxonomic binners and taxonomic profilers⁹⁸. Metagenomic assembly tools attempt to reconstruct whole genomes from DNA fragments into as long and contiguous sequences as possible (contigs) for each individual taxa present in a metagenome. Binner assign (bin) contigs or reads according to genomic or taxonomic origin into groups called OTUs (Operational Taxonomic Units). Genome binners group sequences into unlabelled bins, whereas taxonomic binners group sequences into bins that have an associated taxonomy⁹⁸. Taxonomic profilers attempt to gain an overall picture of what is present in a sample, typically by searching for sequences that have been pre-determined as taxonomically informative (such as MetaPhlAn2⁹⁹). Whilst metagenomic assembly may remove some data from the context of the original sequencing reads, it can present as an attractive approach to species or MAG (Metagenome Assembled Genome) discovery. Long contiguous sequences that do not assign well to any known species can warrant further investigation as a previously unidentified species. This is an advantage of computational methods of identification over conventional culture-based methods¹⁰⁰.

Numerous classification approaches rely on the concept of k -mers: a sequence substring of length k . Binner that use exact matches usually break down reads/contigs into k -mers in order overcome mismatch issues. Typically, the shorter the k -mer length, the higher the sensitivity, but the lower the specificity in making classifications¹⁰¹.

Selecting an appropriate and high performing tool is difficult as each tool claims superiority over its competitors¹⁰². Different ways of assessing tool performance will be discussed in chapter 2. For now, the functionality of two tools for metagenomic classification will be outlined Kraken¹⁰³ and mOTUS2¹⁰⁴. These tools are expanded upon as they will be used extensively throughout this thesis.

Kraken and mOTUs2

Kraken

Kraken is a taxonomic binner that uses a k -mer based approach to classification. A database is built from user defined reference sequences, which are broken down into distinct k -mers using Jellyfish¹⁰⁵. Kraken takes this information and assigns a taxonomy to each stored k -mer as the LCA (Lowest Common Ancestor) of sequences containing the k -mer.

Similar k -mers from reference genomes are grouped and assigned a ‘*minimizer*’. When testing a new input sequence, rather than look up a match against every single k -mer, the previous minimizer group is searched. If there is no match, Kraken will compute the new minimizer for the k -mer and search within this group to help Kraken ‘know where to look’. This greatly increases efficiency.

When classifying query sequences, the taxonomy for each k -mer within the query sequence is recorded. Overall taxonomy for a given sequence is determined by analysing the counts of each k -mer classification and the LCA is calculated by the highest weighted ‘*root-to-leaf*’ path in a taxonomic classification tree (figure 1.5). In the event that there are multiple equally weighted paths, the lowest common ancestor for both options is selected.

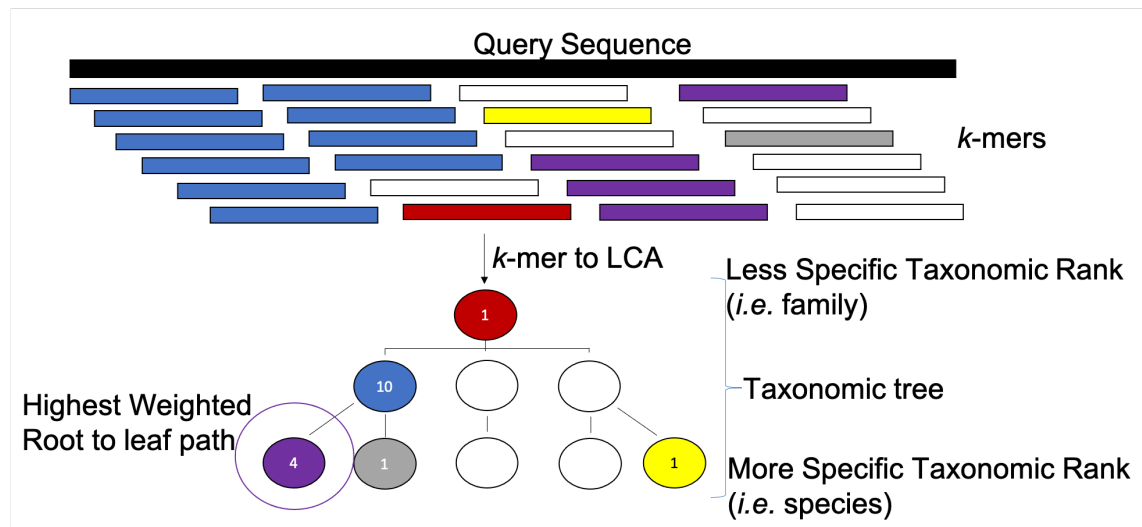


Figure 1.5: Use of the LCA (Lowest Common Ancestor) algorithm by Kraken. The classification of an input sequence is determined by the highest weighted root-to-leaf path of its classified k -mers. Adapted from Wood and Salzberg 2014¹⁰³.

mOTUs2

mOTUS2 is a taxonomic profiler that takes a very different approach to Kraken. It is based on a set of 40 widely expressed single-copy marker gene families that are bacterial ‘*house-keeper*’ genes identified by Mende *et al.* 2017¹⁰⁶. These 40 genes accurately reconstruct the bacterial ‘tree of life’ at species level by clustering the sequences associated with these genes into 5,232 marker gene based operational taxonomic units (OTU). mOTUs2 expanded on this by including information from these marker genes from > 3100 MAGs. mOTUs2 iden-

tified some marker gene clusters from unknown species. mOTUs2 aligns input sequences to this database of OTUs using BWA¹⁰⁷ and normalises the number of hits to each marker gene, by gene length to achieve a relative abundance.

1.2.2 Differential Testing and False Discovery Control

Microbial community data is typically presented as a large matrix of samples by taxa, with the values in each cell resembling either the counts of reads associated with that taxon in each sample or another continuous measure resembling the amount. These can be scaled, transformed or left as raw data¹⁰⁸. It is possible to evaluate differences in the abundance of organisms between two or more groups. Relative abundance is a commonly used measure and denotes the proportion of one taxon relative to the others in the sample. Relative abundance is typically normalised by genome size of the classified taxa to account for the issue of larger genomes naturally obtaining more classified read counts^{109,110}. Differences in abundance measures or read counts between conditions can be subject to differential testing with non-parametric tests, such as Wilcoxon rank sum test, which tests the null hypothesis that there is no difference between two conditions¹⁰⁸. Traditional two-sample tests provide a P -value that suggests the probability that the two populations being tested share identical distributions. Differential presence can also be investigated between two (or more) groups with tests such as Fisher's exact test or the Chi-squared test¹¹¹.

The threshold of significance (alpha level) is disputed in modern science, yet $\alpha = 0.05$ is a widely accepted standard¹¹². Take a small example community matrix containing 100 taxa from healthy and diseased samples. Testing every taxon, by chance we would expect to find five taxa as significantly different ($P < 0.05$) even if the null hypothesis is true (type I error). Microbiome research often contains > 100 taxa, which illustrates the importance of adjusting P -values to limit the FDR (False Discovery Rate).

One *post-hoc* method of adjusting P -values is the Bonferroni correction¹¹³. This approach multiplies P -value by the number of comparisons m ¹¹⁴ but is criticised as overly conservative, resulting in no significance when true differences exist (type II error)¹¹³. In our example community of 100 comparisons, a P -value of 0.0005 ($\alpha/m = 0.05/100$) would be required for statistical significance ($\alpha = 0.05$). A less stringent FDR adjustment and one used later in this thesis is that of the Benjamini-Hochberg adjustment^{114,115}.

A few multivariate statistical frameworks packages exist in the R programming language that attempt to automate statistical testing within community matrix data. Some of the more prominent tools include: LEfSe¹¹⁶ (Linear Discriminant Analysis Effect Size), MaAsLin^{117,117-119} (Multivariate Association with Linear Models), metagenomeSeq¹²⁰ and SIAMCAT¹²¹ (Statistical Inference of Associations between Microbial Communities And host phenoTypes). These tools often use false discovery control but are not particularly suited for the low biomass investigations conducted in this thesis.

1.2.3 Diversity Measures

Measuring diversity to find out if communities differ can often be a good starting point for analysis¹²². Diversity measures can broadly be classified into three subgroups, α -diversity, β -diversity and γ -diversity. The latter measures diversity over a region or a landscape¹²³ and won't be considered in this thesis.

α -diversity is the measure of diversity *within* a sample¹²³. Measures of α -diversity can either be based on phylogenetic relatedness (such as the Faith-PD¹²⁴ measure) or based on data summarising a single taxonomic level¹²⁵. Calculating phylogenetic relatedness can be computationally intensive and so this text illustrates non-phylogenetic diversity. Probably the most widely used measure of α -diversity is that of the Shannon diversity index^{126,127}. Richness is also a measurement of diversity similar to α -diversity that resembles the number of taxa in a community normalised by the number of sampled taxa¹²³.

β -diversity measures differences in composition and abundance *between* samples or sites. Like α -diversity, β -diversity can be calculated based on phylogenetic information or at a discrete taxonomic level. Of β -diversity measurements, there is also the choice of using abundance/count data or binary presence/absence data^{128–130}. Phylogenetic β -diversity measurements can be highly sensitive to rarefaction¹²⁸, or to the method of generating the phylogenetic tree¹³¹.

Calculating β -diversity provides a single numerical value indicating the (dis)similarity between two samples and is therefore also a method of dimensionality reduction; reducing two datasets of N taxa to a single value. The Jaccard distance first published in 1912¹³² is a distance metric that utilises purely presence/absence data. The Jaccard distance is simple to calculate and can effectively highlight differences between samples with different taxa. However, given that it is based purely on the intersection and union of taxa between samples, it weighs all taxa equally. Therefore, two samples containing identical taxa would appear identical even if the distribution of taxa within the two samples was completely different. The Bray-Curtis dissimilarity measure is a commonly used example of a non-metric β -diversity measure that takes into consideration proportional or count data¹³³. A much more complete but non-exhaustive list of β -diversity measures can be found in Appendix, section A.2.1. The real strength of this technique comes when comparing multiple samples and constructing a matrix of distances between individual samples which can be viewed graphically using ordination techniques.

1.2.4 Ordination

The overall aim of ordination (dimensionality reduction) is to graphically represent samples in low dimensional space to see if samples will group together (cluster). In the event that clusters associate with a particular variable (*i.e.*, disease vs control), then there exists some evidence for a community difference. This could be responsible for, or as a result of, differences in the variable of interest. Ordination requires a distance matrix containing β -diversity measurements.

Three ordination methods predominate in microbial ecology: Principal coordinates analysis (PCoA, also known as metric multidimensional scaling), non-metric multidimensional scaling (NMDS) and t -distributed stochastic neighbour embedding (t -SNE)¹³⁴. The choice of distance measure has been suggested to have a more prominent effect on analysis quality than the method of ordination¹³⁵.

Principal Coordinates Analysis - PCoA

PCoA performs eigenvalue decomposition on a distance matrix^{136,137}. The output is a set of axes (or components) that summarise variation within a dataset. PCoA attempts to represent as much of the variation as possible in as few dimensions as possible. To do

this, each axis has an associated eigenvalue, the magnitude of which indicates the amount of variation captured within that axis/component, with the first axis capturing the most variance, the second the second most and so on. Typically, a graphical representation is considered adequate when $> 50\%$ of the variation in the data can be shown within the two or three largest components, with all other component eigenvalues being small^{138,139}. When used with standard Euclidean distances (*i.e.*, geographical distances), PCoA is equivalent to the more widely used principal components analysis (PCA). PCoA is considered more flexible as it can utilise a variety of ecological distances. PCoA can be effectively used to investigate differences in samples as well as identifying confounding variables in study design (figure 1.6). Of the three ordination methods described in this section, PCoA graphically preserves dissimilarity measures the best¹³⁸.

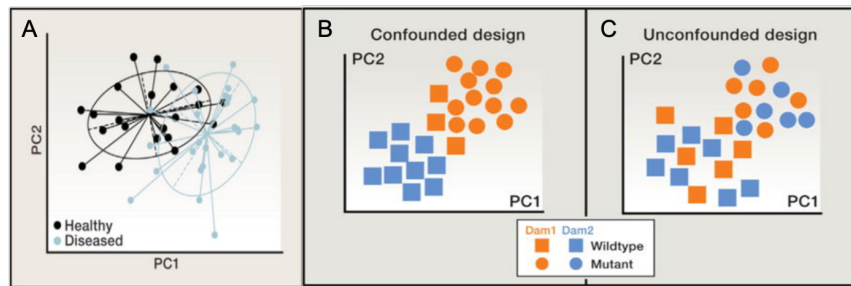


Figure 1.6: Examples of PCoA (Principal Coordinates Analysis) in ecological data exploration. **A)** PCoA can be used to identify differences between healthy and diseased states. Confounders can also be detected by identifying a separation based on confounding variables (**B-C**). **B)** A separation is linked with mouse litter indicating a potentially confounded study design. **C)** shows a separation of data points by mutational status. All figures adapted from Goodrich *et al.* 2014.¹⁴⁰

Non-Metric Multidimensional Scaling - NMDS

NMDS is an iterative solution to ordination pioneered by Shepard in 1962¹⁴¹ and Kruskal in 1964^{142,143}. NMDS is widely used in microbial ecology because it is a tolerant approach that can use a range of metric and non-metric distance measures^{138,139}. NMDS provides a rank-based solution to ordination insensitive to outliers. Briefly, NMDS randomly positions points, adjusts and estimates fit by calculating *stress*. This is repeated until an optimal solution is found that preserves the overall order of distances. A stress value > 0.2 indicates a poor fit to the original distance measures, 0.1 is considered fair, 0.05 good, 0.025 excellent and 0 is perfect^{137,142,143}.

The iterative nature of NMDS can render it intense computationally, particularly on large datasets and there is little room for parallelising the processes as each iteration must be carried out sequentially. Axes of NMDS plots do not hold as much information as PCA (Principal Components Analysis) plots, although it is usually a convention that the axes are rotated such that the first axis demonstrates the longest distances^{137,144}.

***t*-Distributed Stochastic Neighbour Embedding - *t*-SNE**

PCoA does not always reduce all the variation within a dataset to a graphically representable number of axes. NMDS iteratively and intentionally modifies distances to retain the rank order of samples. The goal of both techniques is to identify significant separation or clustering of samples. *t*-distributed stochastic neighbor embedding (*t*-SNE) is a more conceptually abstract method of ordination that iteratively mutates distances towards or away from clusters. In other words, *t*-SNE iteratively, intentionally and nonlinearly modifies distances in order to retain the grouping of samples in low-dimensional space.

t-SNE comes with a few caveats. Firstly, little can be inferred about the distance between points, between clusters or the size of clusters in low dimensional space. Secondly, the *t*-SNE outcome can vary between runs. Thirdly, most *t*-SNE implementations come with many parameters (*i.e.*, perplexity and the number of iterations).

1.2.5 Clustering

Clustering is the process of identifying similar groups of data within a dataset. When this is carried out without the inclusion of any prior information, this process is known as unsupervised clustering.

Clustering approaches can be broadly categorised into two groups: hard and soft clustering. Hard clustering (for example *k*-means clustering^{145,146}) categorises each observation into one cluster. Soft clustering provides each sample with proportional measurements or probabilities of each data point belonging to each cluster. Soft clustering increases the complexity of analysis. Another sub-division of clustering is that of hierarchical vs non-hierarchical clustering. As the name suggests, hierarchical clustering orders samples into a hierarchy of similarity which is useful for interpreting relationships between clusters.

1.2.6 Supervised Learning

Supervised learners use prior information in a subset (training set) of the data. This can also be described as machine learning. They fit a mathematical model on input data to determine a set of parameters to predict a response variable which is then tested on a separate set of data. The simplest form of supervised learning is linear regression whereby a response variable y is fitted to an explanatory variable(s) x ($y \sim x$), a regression line is produced that fits the trend and new data points are interpolated using this line.

A model can fail to ‘follow’ training data closely enough (underfitting) or follow training data too closely, fitting to the noise, meaning that it will struggle in classifying new data (overfitting) and is therefore not generalizable. There are techniques to limit overfitting such as training the model on multiple partitions of the training data (cross-validation)¹⁴⁷. Another important consideration is that of sampling. Consider an example machine learning classifier that was built with the aim of predicting participants that have a mutant gene, *PTEN*. In the training set there may be a very small number of participants that have mutant *PTEN* compared with the number of participants with wildtype *PTEN*. With these proportions of sample sizes (class sizes) left as they are, a typical machine learning algorithm will obtain very little information from those samples that are *PTEN* mutant because there are not many samples (the minority class). This is an unbalanced dataset; the model would benefit from equal sample sizes to equally learn from both groups. One

option is to *downsample* by randomly removing patients with wildtype *PTEN* which risks discarding useful information about wildtype *PTEN* (the majority class). Another option is to artificially increase the low number of participants with mutant *PTEN*. This can be done by duplicating data, although this can lead to overfitting. Other techniques such as SMOTE (Synthetic Minority Oversampling TEchnique)^{148,149} synthesise new data points based on characteristics observed in the minority class.

It is useful to interpret ‘*black-box*’ models produced by machine learning algorithms, particularly for models used in important decisions¹⁵⁰ *i.e.*, by quantifying the variables or ‘*features*’ that have been deemed informative by the algorithm. Doing so can shed light on the underlying relationships between features and outcome variables. Algorithms that provide important features are known as feature selection algorithms such as Boruta¹⁵¹. Boruta is a ‘*wrapper*’ of random forest models that select robust features by using subsets of variables and comparing feature importance to “shadow” features which are randomly shuffled versions of variables. If a feature consistently outperforms the best performing shadow variables, then it is ‘*confirmed*’ as important in its predictive power. Features are classed as tentative if their performance falls between the best and the worst performing shadow counterparts and rejected if they perform worse than the worst performing shadow features.

Difficulties of Supervised Learning in Metagenomics

Supervised learning can be a useful tool for metagenomics. It can identify informative features from multiple variables and patterns that more manual approaches may not detect¹⁵². However, there are several limiting factors that have prohibited the widespread use of supervised learning in metagenomics. Firstly, metagenomics communities are often sparse. As a result, when a classification approach partitions samples into training and test sets, some taxa are completely absent in one partition. Also, microbiome studies typically have many features (taxa) which limits the scope of what is computationally possible¹⁵³. Additionally, the community matrix can be heteroskedastic in nature and may require adequate normalisation, particularly when developing linear models¹⁵⁴.

1.2.7 Functional Taxonomic Composition

Estimating the biochemical pathways present in samples can provide additional biological information beyond taxonomy. Typically functional analysis uses protein searches against a database of orthologues with documented functionality (EggNOG¹⁵⁵ or COG¹⁵⁶), enzymes (KEGG^{157–159}), or protein domains (PFAM¹⁶⁰, TIGRFAM¹⁶¹, INTERPRO¹⁶²)¹⁶³. InterProScan¹⁶⁴ is a tool that can combine many of these databases to provide unique insights across multiple domains for the same sample. It does so with relative computational efficiency by checking for pre-computed results for each protein with unique identifiers (MD5 hash¹⁶⁵). These approaches are all dependent on accurate databases that relay functionality to sequences which are routinely updated¹⁶⁶.

A few approaches have attempted to reduce the computational complexity by pre-running functional analysis by taxa that have clade specific marker genes (PICRUSt 1/2^{166,167} based on 16S identification and HUMAnN¹⁶⁸ based on MetaPhlAn2 output). These tools make assumptions about pathways based on potentially inaccurate taxonomy which may miss important functionality.

There are numerous curated nomenclatures for pathways and metabolic analysis that each provide different insights into the potential functionality exerted by sequences including: MetaCyc¹⁶⁹, Reactome, KEGG, and several levels of Gene Ontology^{170,171} (Biological Process (BP), Cellular Component (CC) and Molecular Function (MF)). Later in this thesis we see how these various pathways can be brought together by utilising InterProScan¹⁶⁴ to elucidate the pathways that may play a role in prostate cancer. It is important to note that predicting these pathways computationally is far from rigorous evidence. The use of pathway analysis in this thesis should be considered purely for hypothesis generation and any findings should be confirmed by experimentation.

1.2.8 Handling Sequence Quality, Noise and Contamination

Contamination in the context of this thesis is the reported presence of any entity that did not originate from within the sample at the time of collection. Contamination is one of the single largest challenges in microbiome research¹⁷². This is particularly pertinent in low biomass studies, where the noise represented by contaminants is large relative to the signal from taxa of interest¹⁷³. Next generation sequencing is so sensitive, that it can often sequence stray pieces of DNA from kit reagents collectively known as the “*kitome*”^{172,174}. A recent report has shown that the microbial DNA of placental samples is not distinguishable from microbial DNA from negative control samples: refuting the existence of a placental microbiome¹⁷⁵.

The kitome is not the only source of contamination. Sample collection can result in contaminants either from non-sterile methods^{176,177}. Sample processing in busy histopathology laboratories can lead to contamination, potentially via non-sterile mincing equipment, blades or the processing of samples in open air as opposed to in a biological safety cabinet. Bacteria have even been reported on surfaces in the international space station, highlighting the ubiquitous nature of environmental microbes¹⁷⁸. Cross-sample contamination is another possible source and although more a false positive than a contaminant, inaccurate base calling could also lead to the erroneous taxonomic classification^{179,180}.

The analysis pipeline can also lead to errors and inaccuracies. A recent study has estimated that 2,767 species in NCBI RefSeq and 6,795 in GenBank suffer from contamination, totalling 2,161,746 entries¹⁸¹. The most common source of contamination was from human sequences¹⁸¹. These genomes are frequently used to build reference databases and therefore may have erroneously classified human variation as microbial taxa. The cleaning of publicly available genomes is an active area of research.

There are various recommendations in the literature that can be followed to minimise the impact of contamination^{172,180,182,183}. Concurrent sequencing of negative controls may elude the constituents of the kitome and environment¹⁸². Various publications have reported common contaminants found in negative controls, which have been summarised by Eisenhofer *et al.* 2019, and a reprinted list of these genera is available in appendix section A.2.4.

Eisenhofer *et al.* 2019¹⁸² suggest a checklist of minimum standards for performing/reviewing low microbial biomass microbiome studies (*‘RIDE’*), which is as follows:

- Report the experimental design and approaches used to reduce and assess the contributions of contamination.
- Include controls to assess contaminant DNA. One of each type of negative control

(sampling blanks, DNA extraction blanks, and no-template amplification) must be included per sampling, extraction, or amplification batch.

- Determine the level of contamination by comparing biological samples to controls.
- Explore contaminant taxa within each study and report their impact on the interpretation of biological samples.

A taxon identified in samples but not in negative controls does not necessarily substantiate the claim that that taxa is a non-contaminant. Furthermore, following the RIDE criteria on large-scale human WGS studies that are not designed for microbe detection is problematic. There are frequently no negative controls and the experimental design is beyond the control of the researcher. Due to these limitations, the analysis presented in this thesis should be considered for hypothesis generating purposes and any results would require additional validation.

There are a limited number of approaches to handle contamination computationally. One approach would be to remove all common contaminant genera from analysis - a ‘black-list’. This is used throughout this thesis, but it does present with some drawbacks. It will remove potentially biologically relevant genera such as *Escherichia* or *Fusobacterium* (which are common contaminants and commensals^{184,185}). Additionally, there will be many reported taxa that are contaminants despite not being on the black-list. The other approach is to use a ‘white-list’ of taxa that are not common contaminants and are potentially biological relevant. Forming such a white-list is challenging and may limit discovery by excluding taxa. Another approach could be to weight taxa based on their probability of being a contaminant, but such an approach does not yet exist. Another approach is to keep all taxa in and evaluate the potential for contamination upon identifying a biologically meaningful result.

Some key considerations of each taxa include the likelihood that a taxon is a contaminant (by examining literature regarding contamination), the likelihood of a false positive classification, the significance of detecting that taxa and the strength of computational evidence. There are some clearly nonsensical results that have been published and refuted, *i.e.*, *Raphus cucullatus* (the extinct dodo bird)¹⁸⁶. This is clearly impossible and therefore should be excluded from any analysis or the taxa correctly identified. Afshinnekoo *et al.* 2015¹⁸⁷ obtained classifications for *Yersinia pestis* and *Bacillus anthracis* on the New York subway. Reanalysing this data using different approaches did not reproduce these findings highlighting the requirement for stringent and thorough identification of highly significant pathogens^{186,188,189}. This can be carried out by investigating the strength of computational evidence for these taxa by replicating the findings using a different approach and by investigating the level of evidence such as the number of classified sequencing reads or the coverage of the clade.

There are some measures that can be carried out computationally to limit contamination. Sequencing reads with an average Phred quality of 30 translates to a $\frac{1}{1000}$ chance ($P = 10^{-\frac{Q}{10}} = 10^{-\frac{30}{10}} = 0.001$) that a base is called incorrectly and therefore a 15% (0.001×150) chance that at least one base has been incorrectly called in a 150bp read. Removing low quality reads such as those beneath this threshold would limit misclassification of poor-quality sequences.

1.3 Thesis Outline

1.3.1 Aims and Objectives

Aims

The main aim of this thesis is to develop computational pipelines to search for microbial nucleic acids in cancer sequence data and apply them to different human 'omics datasets. It is hoped that this will provide preliminary *in silico* evidence for associations between microbes and cancer. I aim to generate hypotheses regarding the cancer metagenome that will lead to further experimentation and liaising with cancer type specific scientists to develop experiments beyond the confines of purely computational methods.

Objectives

- Develop computational pipelines to classify microbes in cancer whole genome sequence data.
- Apply these pipelines to a simulated benchmarking dataset.
- Carry out pilot studies in a range of cancer types including cancers with a known pathogenic involvement such as cervical and gastric cancer.
- Apply these pipelines to whole genome cancer sequences from Genomics England's 100,000 Genomes Project and RNA sequencing of urine samples collected from participants under investigation for prostate cancer.
- Determine the landscape of microbes detectable in tumour sequence data.
- Search for associations between microbes and metadata.
- Generate hypotheses regarding associations between microbes and cancer.
- Liaise with cancer-type specific scientists to design experiments to verify findings.

1.3.2 Chapter Outline

The total word count for this thesis is 47768 (35062 excluding the bibliography), including 8503 words for the current chapter.

Chapter 2 (4920 words) describes a benchmark of computational approaches to find an optimal approach to detecting microbial sequences in sequencing data produced from cancer samples. This is investigated by running a variety of computational approaches on simulated datasets that have been created to resemble cancer sequence data. The important computational steps to process such data is outlined before benchmarking tools capable of these features. This chapter covers the performance of: Bacterial taxonomic classification, bacterial proportion estimation, viral classification and classification of bacteria and viruses following metagenomic assembly. Then, data is presented from two of the best performing approaches on real sequencing data from cervical cancer, stomach cancer and medulloblastoma.

Chapter 3 (7194 words) outlines the application of SEPATH to RNA sequences obtained from urine of participants under clinical investigation for prostate cancer. Evidence for a wide variety of urogenital bacteria was identified. This dataset was investigated by two grouping approaches: *k*-means clustering of non-metric multidimensional scaling (NMDS) and by focusing on samples with presence/absence of bacteria in an anaerobic bacterial biomarker set, termed ABBS. The differential gene expression of human genes between

these groups was determined, including functional analysis of the differentially expressed genes. Functional pathways were predicted from assembled isolates obtained from urine samples of prostate cancer patients and investigated.

Chapter 4 (8627 words) describes the application of SEPATH to tumour samples from the 100,000 Genomes project. The microbial constituents of over 10,000 tumour samples are determined, revealing potentially significant taxa, some of which have previously been associated with cancer. The identification of clinically actionable communicable diseases is reported on including an in-depth investigation of the potential diagnosis of Human T-cell lymphotropic virus in a single participant. This chapter also describes an approach towards identifying contamination by using machine learning approaches. Most of this chapter focuses on the search for differences in microbial community structure between cancer types. The results presented in this chapter suggest that the tumour microbiome is not linked with the anatomical site of a particular cancer apart from gastrointestinal and oral cancers. This is contradictory to some recent publications which is further investigated.

The concluding chapter (chapter 5, 5451 words) provides a comprehensive overview of the analyses chapters in this thesis and discusses topics such as taxonomic classification, the issue of contamination and outlines potential avenues of future research.

Additional Material

This thesis has been written in RStudio using the `thesisdown` package. The GitHub repository for this thesis contains additional files that are too large to include in the appendices (chapter A). Where possible, code used for any analyses and figures is directly integrated into the code that produced the text for this thesis. The associated GitHub and additional instructions is available from: https://github.com/Agihawi/Gihawi_Thesis.

Chapter 2

Developing SEPATH - Template Pipelines to Search for Microbial DNA in Host Dominated Sequence Data

This chapter is based on an open access publication¹⁹⁰ produced during this PhD project. Much of the content of this chapter is as it appears in this publication.

A collection of resources and scripts to reproduce this analysis can be located at: <https://doi.org/10.5281/zenodo.3387205>¹⁹¹

2.1 Summary

As we transition into an era of genomic medicine, human tissue is increasingly being whole genome sequenced. With this, arises the potential to detect sequences originating from microorganisms, including pathogens amid the plethora of human sequencing reads. In cancer research, the tumorigenic ability of pathogens is being recognized, for example *H. pylori* and human papillomavirus in the cases of gastric non-cardia and cervical carcinomas respectively. As of yet, no benchmarking exercise has been carried out on the performance of computational approaches for bacterial and viral detection within host-dominated sequence data.

The results are shown from benchmarking over 70 distinct combinations of tools and parameters on 100 simulated cancer datasets spiked with realistic proportions of bacteria. mOTUs2 and Kraken are the highest performing individual tools achieving median genus level F1-scores of 0.90 and 0.91 respectively. mOTUs2 demonstrates a high performance in estimating bacterial proportions. Employing Kraken on unassembled sequencing reads produces a good but variable performance depending on post-classification filtering parameters. These approaches are investigated on a selection of cervical and gastric cancer whole genome sequences where Alphapapillomavirus and *Helicobacter* are detected in addition to a variety of other interesting genera.

The top performing pipelines from this benchmarking process are provided in a unifying tool called SEPATH, which is amenable to high throughput sequencing studies across a range of high-performance computing clusters. SEPATH provides a benchmarked and convenient approach to detect microbes in tissue sequence data helping to determine the relationship between pathogens identified in metagenomes and disease.

2.2 Background

There have been some previous attempts at comparing metagenomics tools and even tools to combine the output profiles of multiple classifiers such as MetaMeta¹⁹². Usually, each release of a new metagenomic classification tool is benchmarked against a small selection of other tools using simulated metagenomes that have been chosen by the authors. Because the number of tools chosen is typically small and the choice of tools varies by publication, direct comparison between tools is extremely difficult without user testing each approach.

The Critical Assessment of Metagenomic Interpretation (CAMI) was established to address this issue⁹⁸. CAMI is a community driven approach to benchmarking tools for metagenomic classification on simulated datasets of ‘unprecedented complexity’. The challenge offers a level platform for each metagenomics tool and offers the ability for bioinformaticians and developers to use their expert knowledge to submit a best effort at classifying a metagenome, the composition of which is blinded to researchers.

The first round of CAMI contained an assembly challenge, a genome binning challenge, a taxonomic profiling challenge and a taxonomic binning challenge. Overall, 19 teams took part in wrangling the synthetic datasets built from ~700 newly sequenced microorganisms and ~600 novel viruses and plasmids. Overall, CAMI reported a few interesting findings. Firstly, multi- k -mer assemblers outperformed single k -mer assemblers although assembly overall was hindered by closely related genomes. Secondly, they noted a high performance in taxonomic bidders and profilers down to family level, but that genus and species level performance could be improved. The second phase of the CAMI challenge offers teams new challenges with a shift in focus to different environments¹⁹³. This includes a marine dataset and a dataset with high strain diversity. Additionally, the challenge marks a transition towards evaluating performance of metagenomics in a clinical setting with a case study involving a sequenced nasal swab from a 32 year old female with vomiting and abdominal pain. At the time of writing, the results of this challenge were unpublished. CAMI represents perhaps the most comprehensive attempt at benchmarking tool performance to date. However, CAMI doesn’t include the specific case of classifying microbes within tumour sequence data. Therefore, it is important that a benchmark on such datasets is carried out in order to provide confidence in the results of analysing real tumour samples.

This chapter presents and discusses the development of SEPATH (SEarching for PATHogens). SEPATH is a group of template computational pipelines designed to obtain genus level classifications from within human tissue sequence data and optimized for large WGS studies. This section provides rationale for the constituent tools of SEPATH by analysing the performance of tools for quality trimming, human sequence depletion, metagenomic assembly and classification (figure 2.1). The utility of these pipelines is further assessed by running them on a selection of cancer WGS datasets. A small selection of samples from cervical cancer is analysed, where it is expected that Alphapapillomavirus will be frequently identified, and gastric cancer where it is expected that *Helicobacter* will

be identified. A selection of ten paediatric medulloblastoma samples is also analysed for which it is expected that not many, if any, taxa will be identified due to the historically noted sterility of the brain. Recent research has suggested that finding microbial DNA in the brain has been exaggerated by off-target amplification and poorly controlled studies^{194,195}.

The process of obtaining microbial classifications from host tissue reads can be broken down into three key computational steps:

- **Sequence quality control** - Good quality data is fundamental to any downstream sequence analysis including metagenomics. The effect of sequence quality control has not been thoroughly investigated here as it varies greatly between datasets and sequencing technologies. Therefore, Trimmomatic¹⁹⁶ was employed as it was designed for Illumina data which is examined in the remainder of this thesis.
- **Host sequence depletion** - Including Host sequence depletion offers two advantages. Firstly, the volume of data inputted to the classification stage is vastly reduced by discarding the majority of sequencing reads. This makes processing the files more manageable and reduces the use of computational resources. Secondly, this step reduces the number of non-microbial reads that are subject to taxonomic classification thereby limiting false positive results
- **Taxonomic classification** - Many tools exist to classify the taxonomic origin of nucleotide sequences. The choice of tool and approach can greatly impact the taxonomic results obtained. Here we evaluate candidate tools for their classification performance.

For host sequence depletion and taxonomic classification, a series of tools and parameters were benchmarked on simulated metagenomes (see methods section 2.6).

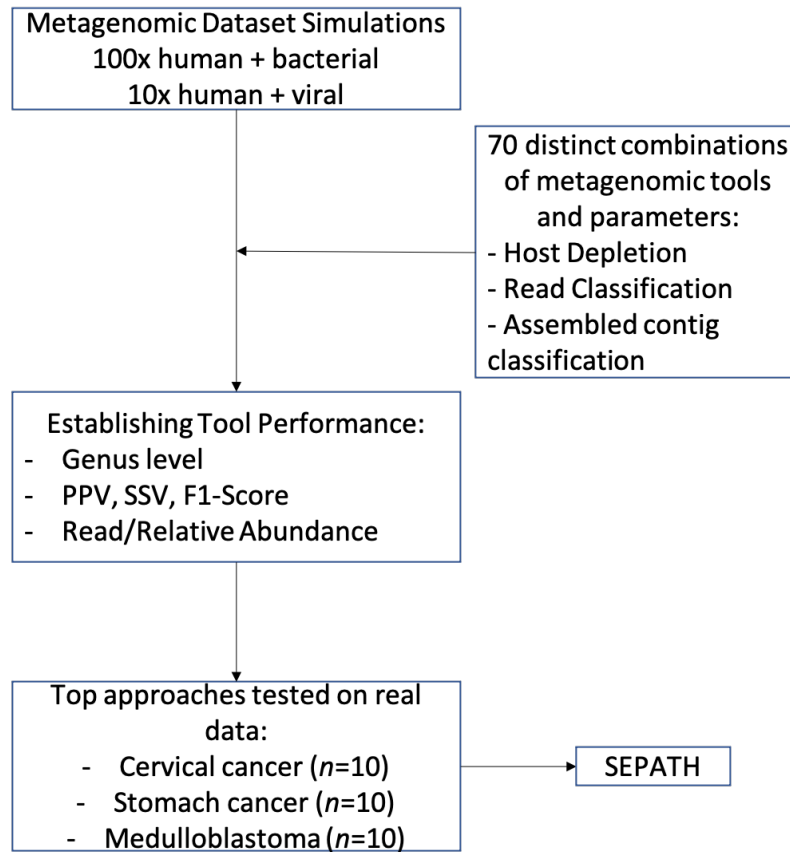


Figure 2.1: An overview of the process used to develop the SEPATH computational pipelines. Firstly, metagenomic datasets were simulated containing either bacterial or viral reads. A variety of methods were tested on these datasets before being applied to some real cancer sequencing data from cervical cancer, stomach cancer and medulloblastoma. Positive Predictive Value (PPV), Sensitivity (SSV).

2.3 Results

2.3.1 Simulated Datasets

The simulated metagenomes emulate empirical observations from other cancer tissue sequence data⁹⁶, with the percentage of human reads ranging from 87% to > 99.99%. Genomes from 77 species were selected as constituents for the metagenomes¹⁹⁷. These species were identified from Kraal et al. 2014¹⁹⁸ with additional suggestions from Dr Rachel Hurst and bacterial species associated with cancer such as *H. pylori*¹⁹⁹.

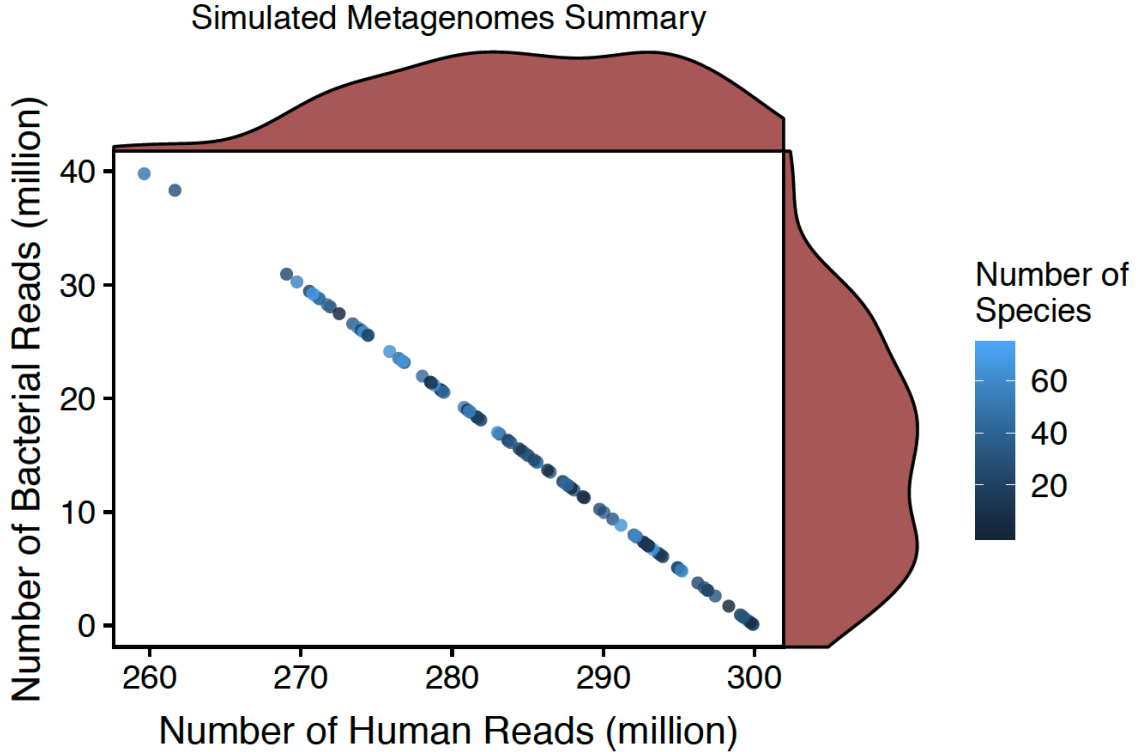


Figure 2.2: Scatter plot summarizing the constituents of all 100 simulated bacterial metagenomes. The y -axis demonstrates the number of bacterial reads in the datasets, whereas the number of human reads is shown on the x -axis. The number of species in each dataset is indicated by the color, darker points having less species. The distribution of each axis is shown in red.

2.3.2 Human Sequence Depletion

Three methods of host depletion were investigated on eleven simulated datasets (2x150bp Illumina reads, see section 2.6). Two of these methods were k -mer based methods: Kontaminant^{200,201}, and BBDuk²⁰². The third method involved extracting unmapped reads following BWA-MEM¹⁰⁷ alignment, an approach that is facilitated by the likelihood that data will be available as host-aligned BAM files in large scale genomic studies. BWA-MEM is used as a baseline and parameters were set to preserve any potential non-human reads.

All methods retained the majority of bacterial reads (median of $> 99.9\%$ bacterial reads retained for all conditions; appendix figure B.2), but the number of human reads remaining in each dataset varied (figure 2.3). Using default parameters BBDuk and Kontaminant retained a median of 15.4 million human reads, compared to 259 million from BWA-MEM with intentionally lenient filtering parameters. BBDuk was investigated further, establishing default BBDuk performance following BWA-MEM depletion which demonstrated no tangible difference in human read removal (figure 2.3A). BBDuk parameters were also adjusted from the default setting of a single k -mer match to the reference database (figure 2.3B-C). It was found that removing a read when 50% or more of the bases have k -mer matches to the

human reference (MCF50) provided an approach that removed near identical proportions of human and bacterial sequences to the default parameters (figure 2.3).

To capture k -mers specific of cancer sequences, a BBDuk database was generated containing human reference genome 38 concatenated with coding sequences of all cancer genes in the COSMIC database²⁰³. With the additional cancer sequences, a near identical performance (in terms of the number of reads filtered) was obtained when compared with just the human reference database (figure 2.3B-C). Therefore, including extra cancer sequences did not alter the retention of microbe derived reads, providing an opportunity for increased human sequence removal on real data without sacrificing bacterial sensitivity. To investigate using a BBDuk database capturing a higher degree of human sequence variation, the inclusion of additional human sequences from a recent analysis into the African ‘*pan-genome*’²⁰⁴ was additionally investigated. Including these extra sequences removed slightly more bacterial reads but had a very minor effect (figure 2.3C).

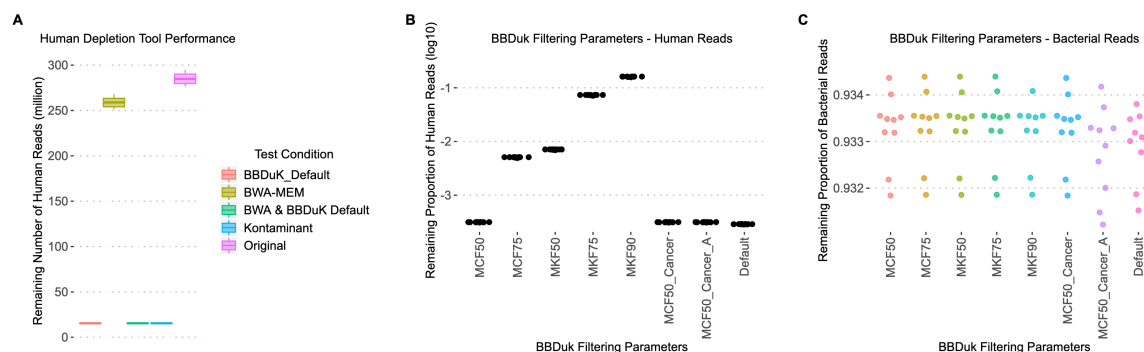


Figure 2.3: Human read depletion performance. **A)** Human read removal using BBDuk, BWA-MEM, and Kontaminant. The remaining numbers of human reads were near identical for BBDuk and Kontaminant (median values of 15,399,252 and 15,399,928 for BBDuk and Kontaminant, respectively.) All conditions retained bacterial reads with near-identical performance. BBDuk was selected for parameter optimization (**B**, **C**). This analysis was performed on raw untrimmed reads of $n = 11$ simulated datasets. **B-C)** BBDuk parameter optimization in terms of the remaining human reads (**B**) and remaining bacterial reads (**C**). Default BBDuk settings were used along with alterations of MKF and MCF parameters. The default parameters of BBDuk remove a sequencing read in the event of a single k -mer match, whereas MCF50 requires 50 percent of the bases in a read to be covered by reference k -mers for removal and MKF50 requires 50 percent of k -mers in a read to match the reference for removal. MCF50-Cancer indicates that BBDuk was ran with a database consisting of GRCh38 and mutations identified in human cancer from COSMIC. MCF50-Cancer-A denotes a database consisting of GRCh38, COSMIC cancer genes, and additional African “pan-genome” sequences. Default and both MCF50 parameters (with and without cancer sequences) showed the highest removal of human reads.

2.3.3 Taxonomic Classification - Bacterial Datasets

The performance of three taxonomic profilers (mOTUs¹⁰⁴, MetaPhlAn2^{99,110} and Gottcha²⁰⁵) and three taxonomic binner (Kraken¹⁰³, Centrifuge²⁰⁶ and Kaiju¹⁰¹) was examined by applying them after filtering and host-depletion on 100 simulated datasets (figure 2.6 A-C). Performance was measured in terms of presence/absence metrics at the genus level: positive predictive value (PPV/precision), sensitivity (SSV/recall) and F1 score (the harmonic mean of precision and recall). These tools were selected for a few reasons: performance in CAMI, prevalence in recently published literature and ease of use/installation. In our analysis, Kraken and mOTUs2 delivered the best median genus F1 of 0.90 (IQR=0.083) and 0.91 (IQR=0.10) respectively. With median genus PPV scores of 0.97 (IQR=0.084), 0.95 (IQR=0.080) and median genus sensitivity scores of 0.86 (IQR=0.123), 0.88 (IQR=0.126) for Kraken and mOTUs2 respectively.

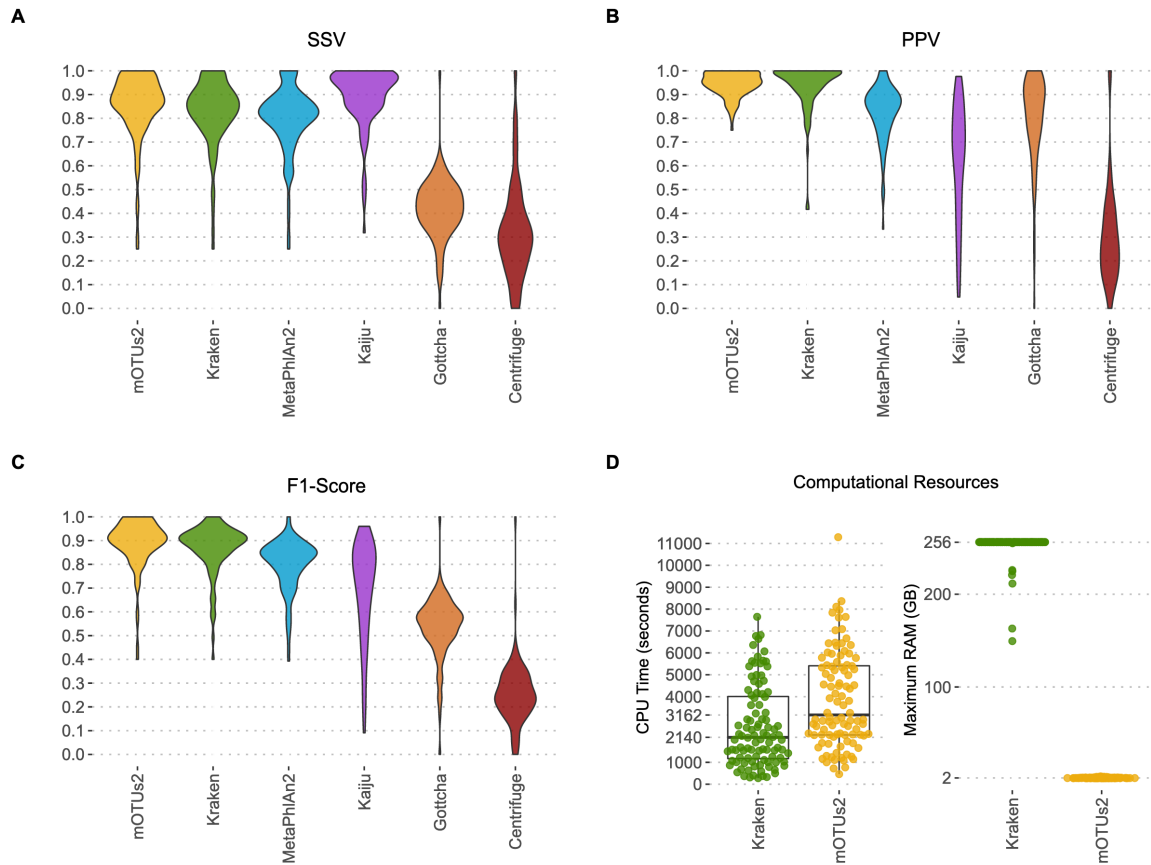


Figure 2.4: Performance estimates for taxonomic classification tools on 100 metagenome simulations following quality filtering and human sequence depletion. Performance is summarized at the genus level in terms of sensitivity (A), positive predictive value (B), and F1 score (C). Computational resources, CPU time and RAM is also shown for the top two performing tools: Kraken and mOTUs2 (D). Kraken utilised 20 threads for most datasets whereas mOTUs2 utilised 17. mOTUs2 output was unfiltered, whereas Kraken had a confidence threshold of 0.2 and a minimum read threshold of 500 in determining positive classifications.

Kraken utilizes > 125 times the RAM of mOTUs2 (figure 2.4D; median 256GB vs 2GB RAM for Kraken and mOTUs2 respectively; $p = 2.2 \times 10^{-16}$ Mann-Whitney U test); Kraken was ran with the database loaded into RAM to improve run-time. Historically, alignment based taxonomic classification tools have been slow, but by using the reduced 40 marker gene database, mOTUs2 has much lower run times.

Kraken Performance at Various Taxonomic Levels

Kraken was run on all bacterial simulated datasets and the performance at the main taxonomic ranks was assessed in terms of F1 score, PPV and sensitivity (figure 2.5). It was noted that there was a worse overall performance on unfiltered data compared with filtering using a minimum read threshold. Additionally, there was a drop off in performance towards species level. On filtered data, genus level performance obtained a median F1 score above 0.9.



Figure 2.5: Violin plots demonstrating performance in terms of F1-score, PPV and SSV for taxonomic ranks between Phylum and Species level on $n=100$ simulated datasets. **A)** Performance of Kraken when run on raw reads with no read threshold. **B)** Performance following a minimum threshold of 500 reads for each classification.

2.3.4 Bacterial Proportion Estimation

Analysing population proportions can provide a deeper understanding of microorganism community structure. Therefore, it is important to assess the performance of tools in predicting proportions. For each true positive result from the top performing pipelines using Kraken and mOTUs2, the output number of reads was compared against the true number of reads in the simulations (figure 2.6). The mOTUs2 pipeline obtained accurate

rankings of read estimates ($R^2 = 0.91$; Spearman's rank-order correlation) whereas our Kraken pipeline predicted the number of reads with a Spearman's rank-order correlation value of $R^2 = 0.69$. We will however note that this value could be improved by using Bracken (discussed later).

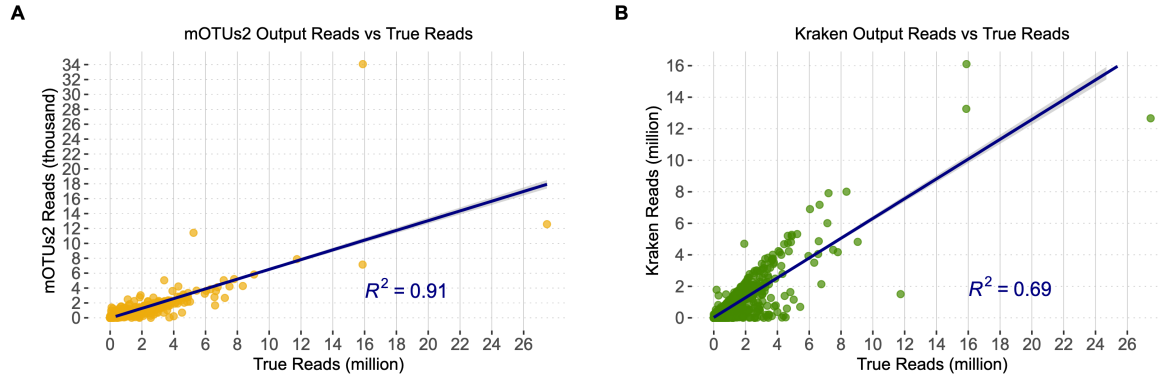


Figure 2.6: Quantitative ability for mOTUs2 and Kraken. mOTUs2 output reads vs true reads (A) and Kraken output reads vs true reads (B). For all true-positive genera classifications (Spearman's rank correlation coefficients 0.91 and 0.69, for $n = 2,084$ and 2,021 true-positive classifications for mOTUs2 and Kraken, respectively). mOTUs2 classifications were left unfiltered whereas Kraken had a confidence threshold of 0.2 and a minimum read threshold of 500.

2.3.5 Bacterial Classification Following Metagenomic Assembly

When run on raw reads, Kraken typically requires post-classification filtering strategies in order to obtain high performance¹⁰⁴ (figure 2.5). Post-classification filtering involves applying criteria to remove low quality classifications from taxonomic results. Applying a metagenomic assembly algorithm to quality-trimmed non-host reads might provide a rapid filtering approach that reduces the need for read-based thresholds.

MetaSPAdes²⁰⁷ was employed on high quality non-human reads from 100 simulated datasets and then Kraken was applied. An F1-score of 0.83 was obtained without any read threshold, which was an improvement over Kraken on raw reads without any filtering strategies (F1=0.54), but lower than Kraken with filtering (F1=0.9). The F1-score was increased to 0.89 when a requirement for a minimum of five classified contigs for classification was applied (figure 2.7A). Filtering out contigs with lower coverage made little difference on performance with the parameters tested. (figure B.3 - B.4).

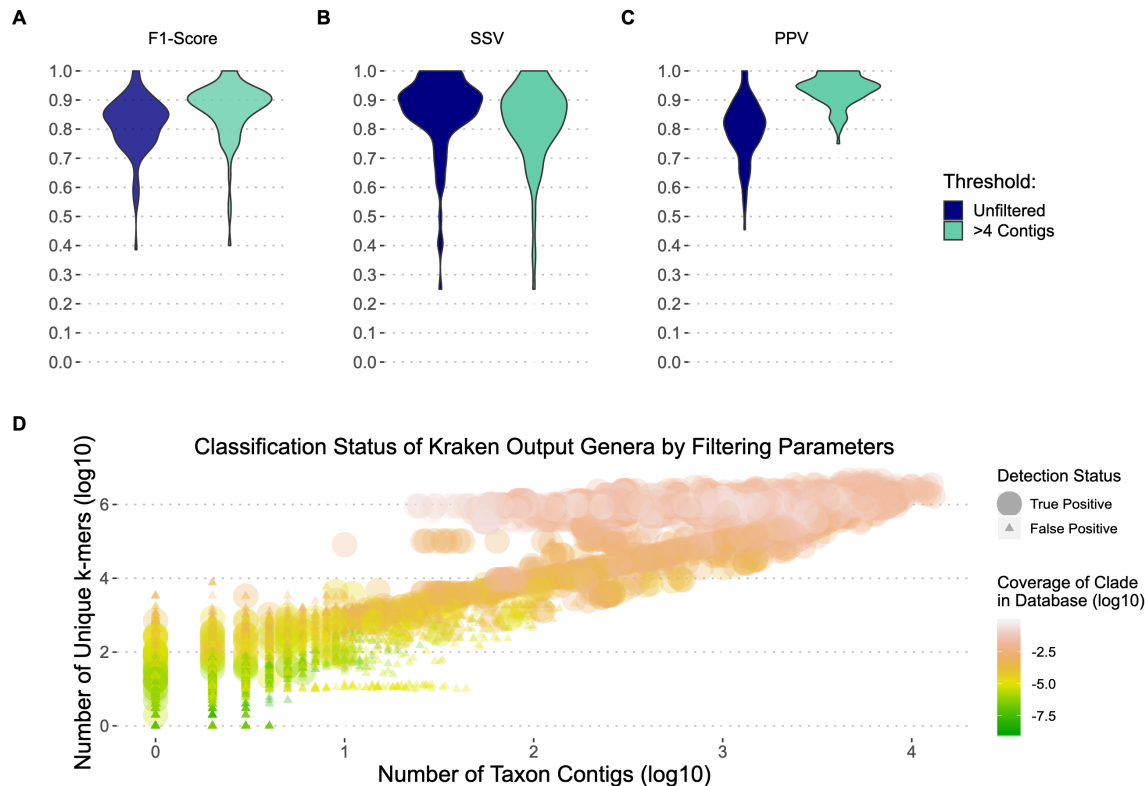


Figure 2.7: Genus-level performance of Kraken on contigs following metagenomic assembly with MetaSPAdes. Performance is summarized by genus-level F1 score (**A**), sensitivity (**B**), and PPV (**C**). A single dataset failed metagenomic assembly, and so, data shown is for 99 of 100 simulated datasets. Performance is shown on raw Kraken classifications with no threshold applied (unfiltered) in dark blue. Light blue denotes the performance when a minimum of 5 contigs assigning to a genus was used. Median values for unfiltered performance were 0.83, 0.88, and 0.81, and for filtered performance were 0.89, 0.85, and 0.94 for F1 score, sensitivity, and PPV, respectively. **D**) KrakenUniq filtering parameters in relation to detection status. The y -axis indicates the number of unique k -mers assigned to a particular taxon, the x -axis represents the number of contigs assigned to a particular taxon ($\log_{10}$), and the colour gradient shows the coverage of the clade in the database ($\log_{10}$). True-positive results are larger circles, whereas false-positive results are smaller triangles. The scatter plot shows 10,450 contigs classified at genus level as data points. The k -mer size was 31.

Filtering these datasets by number of contigs is non-ideal, as it would remove classifications from taxa that assembled well into few contigs. An evolution of Kraken, KrakenUniq²⁰⁸ was run on these contigs to further illuminate the relationship between taxa detection and advanced metrics including the coverage of the clade in the reference database and the number of unique k -mers (figure 2.7D, figure B.5). This analysis reveals that on our datasets, no set of filtering parameters could obtain perfect performance. Upon investigation

of a single dataset, it was observed that 13 out of 17,693 contigs assigning to different genera were responsible for false positive classifications resulting in a drop of PPV to 0.83 (figure B.6). These contigs were extracted and used as input for NCBI's MegaBLAST with standard parameters. Of the thirteen false positive contigs, three were correctly reclassified, three were incorrectly classified, and the remaining seven obtained no significant hits. This highlights that these contigs may suffer from mis-assembly or non-uniqueness that is not improved by use of a tool with a different approach.

2.3.6 Taxonomic Classification - Viral Datasets

The performance of viral classification was established in the presence of bacterial noise by spiking a selection of host-bacterial datasets with 10,000 viral reads for each of ten species. As mOTUs2 does not make viral classifications, Kraken was run on either quality trimmed reads or contigs following metaSPAdes²⁰⁷ assembly (see methods section 2.6). Kraken correctly identified 8/10 virus species from reads as input with post-classification filtering. When using contigs and no filtering strategies, 7/10 species were detected with no viral false positive results (figure 2.8B). Filtering by minimum number of contigs removed most viral classifications. The effect of filtering on viral species classification was not reflected in classification of bacterial genera (figure 2.8A).

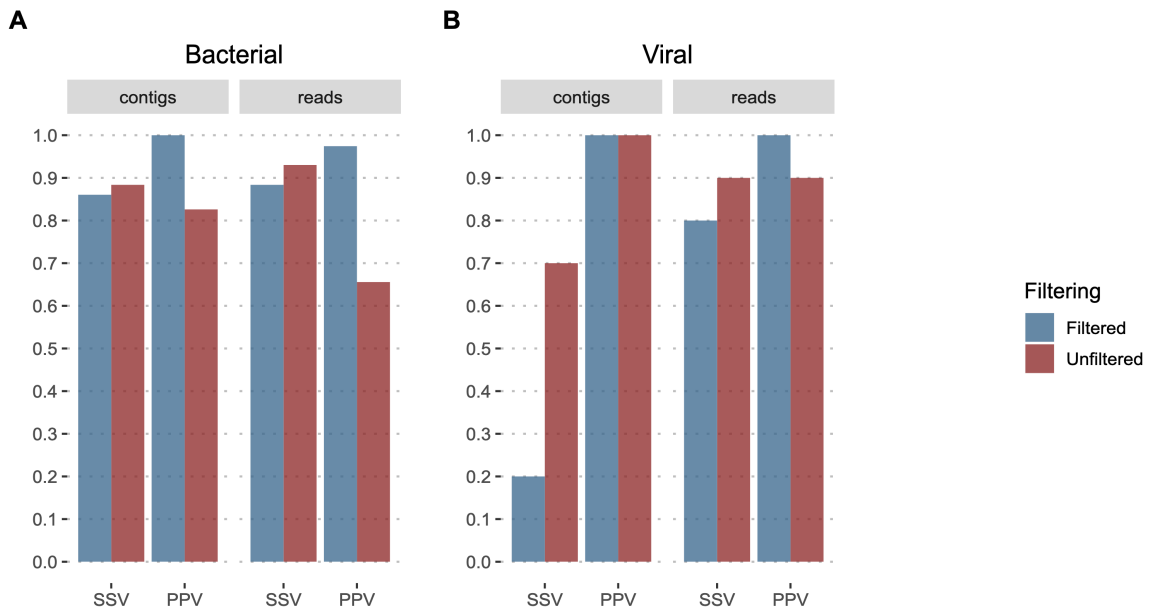


Figure 2.8: Kraken performance on a single dataset containing both bacterial (A) and viral (B) reads. Performance from metagenomic assembly approach is shown on both unfiltered contigs and results filtered by a minimum of 5 contigs required for classification. Kraken performance on raw reads is shown both unfiltered and filtered by a minimum of 100 reads for classification. Bacterial performance is classified at genus level whereas viral performance is regarding species level due to peculiarities in taxonomy.

2.3.7 Bacterial Consensus Classification

Using distinct methods of classification and combining the results has been shown to improve metagenomic classification performance¹⁹². The Kraken/mOTUs2 pipelines outlined here were compared with the alignment based PathSeq²⁰⁹ (earlier versions utilising BLAST²¹¹ and the more recent versions using BWA-MEM) on a reduced selection of eleven simulated bacterial datasets (figure 2.9). A smaller selection of datasets was used due to local resource limitations on storage and compute time. It was found that using an intersection of classifications between any two tools obtained a perfect median PPV score but caused a small drop in sensitivity and resulted in similar F1-scores compared with using single tools. Sensitivity increased to 0.9 when using a consensus approach between all three tools (where only classifications made by 2/3 tools is taken as true). This rise in sensitivity for the consensus approach resulted in a median genus level F1-score of 0.95; which was a better score than any other single tool or intersection of two tools.

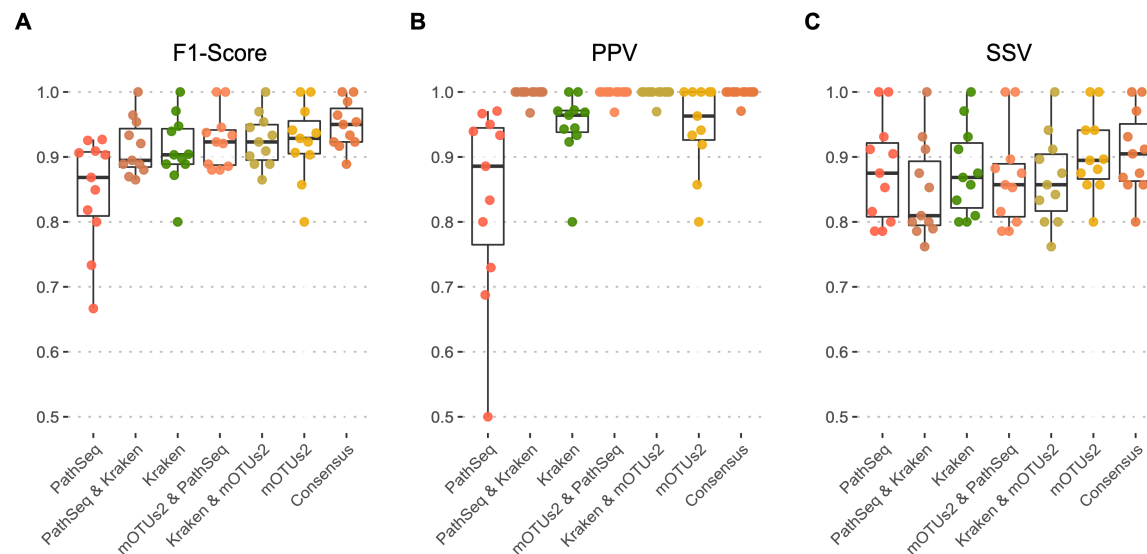


Figure 2.9: mOTUs2, Kraken, and Pathseq form a consensus with near-perfect genus-level classification performance. Box plots with individual data points for $n = 11$ simulated bacterial metagenomes showing genus-level F1 score (A), PPV (B), and SSV (C) for single tools, an intersection of classification between two tools, and a consensus of all three tools. PPV obtained perfect values in the result of an intersection between two tools or a consensus. Sensitivity generally decreases in the event of combining two tools with an intersection but increases to 0.905 (median) in the result of an intersection. This raise in sensitivity resulted in a genus level F1 score in the consensus approach of 0.95. mOTUs2 output files were unfiltered, whereas Kraken had a filter of >4 contigs and PathSeq >1 reads.

2.3.8 Real Cancer Whole Genome Sequence Data

SEPATH pipelines using Kraken and mOTUs2 were run on quality trimmed, human depleted sequencing files (figure 2.10). Kraken identified Alphapapillomavirus in 9/10 cervical squamous cell carcinoma samples, with a high average number of sequencing reads com-

pared to other taxa (figure 2.10A). Alphapapillomavirus was not prominent in stomach cancer or medulloblastoma samples. *Treponema* was identified as present in two samples by both techniques (figure 2.10B) and both tools report high number of reads or relative abundance. This may well represent an interesting diagnostic finding, although further investigation would be required to ascertain the clinical utility. In the stomach cancer samples, both mOTUs2 and Kraken identified *Helicobacter* in 4 and 5 samples respectively as anticipated. *Helicobacter* was not a prominent classification in either of the other tumour types. Kraken reported Lymphocryptovirus in 6/10 samples with a high number of reads in addition to a variety of other genera (figure 2.10C). Despite human read depletion, care should be taken to ensure the true positive nature of Lymphocryptovirus^{212,213}. It is noteworthy that the classification is not prominent in either cervical cancer or medulloblastoma and has previously been associated with gastric oncogenesis^{214,215}.

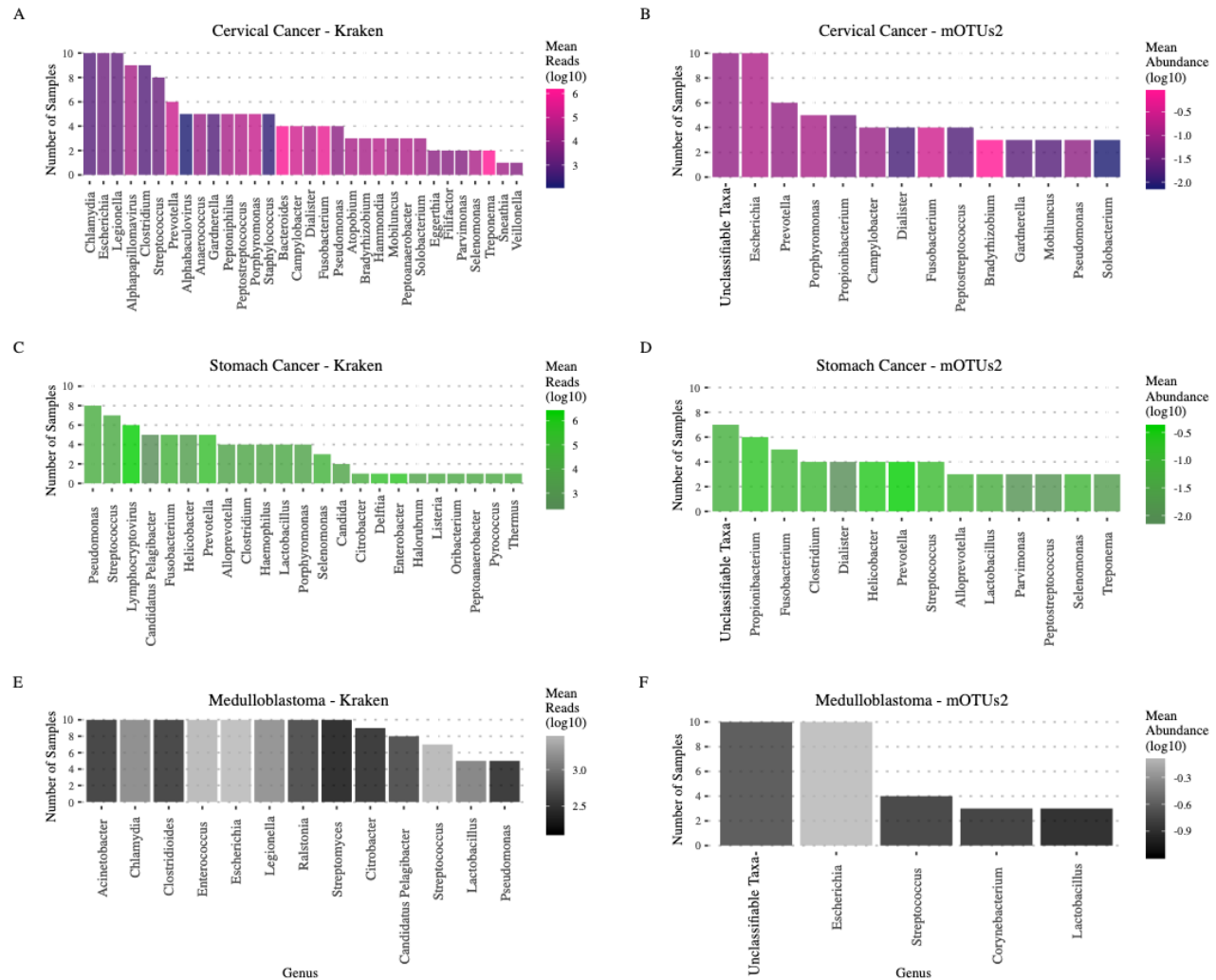


Figure 2.10: The application of SEPATH pipelines on a range of cancer types. Output genera from Kraken (left) and mOTUs2 (right) human-depleted, quality-trimmed reads from whole genome sequencing files. $n = 10$ for each of cervical cancer (A, B), stomach cancer (C, D), and medulloblastoma (E, F). For display purposes, mOTUs2 results were filtered to show taxa that occurred in at least 3 samples. Kraken results were filtered for taxa that were in a minimum of 5 samples, or had a mean read count of over 5,000.

In both cervical and gastric cancer, expansion of these pipelines to larger datasets would help to characterize the role of many other reported genera. Medulloblastoma samples are expected to be mostly sterile and this is well reflected with only very low number of genera at low read counts (number of genera: total reads in all samples 75 : 11,213,997, 102 : 16,269,893, 27 : 138,712 for cervical, gastric and medulloblastoma respectively). Kraken appears to be more sensitive; making a greater number of classifications overall and classifying the same taxa as present in a higher number of samples than mOTUs2.

2.3.9 SEPATH Template Pipelines

The top performing algorithms and parameters for each of the stages discussed above have been combined in a unifying template pipeline implemented in Snakemake²¹⁶ called SEPATH (figure 2.11, https://github.com/UEA-Cancer-Genetics-Lab/sepath_tool_UEA). SEPATH provides three blocks of functionality:

- Conversion of host-aligned BAM files to FASTQ files intentionally preserving microbial reads.
- mOTUs2 bacterial classification run on trimmed and filtered sequencing reads
- Kraken run on quality trimmed reads or metagenomic assembled contigs.

All blocks can be run together or separately and uses either BAM or FASTQ input files. All software dependencies for SEPATH can easily be installed via Conda and a singularity image file containing all required dependencies is provided.

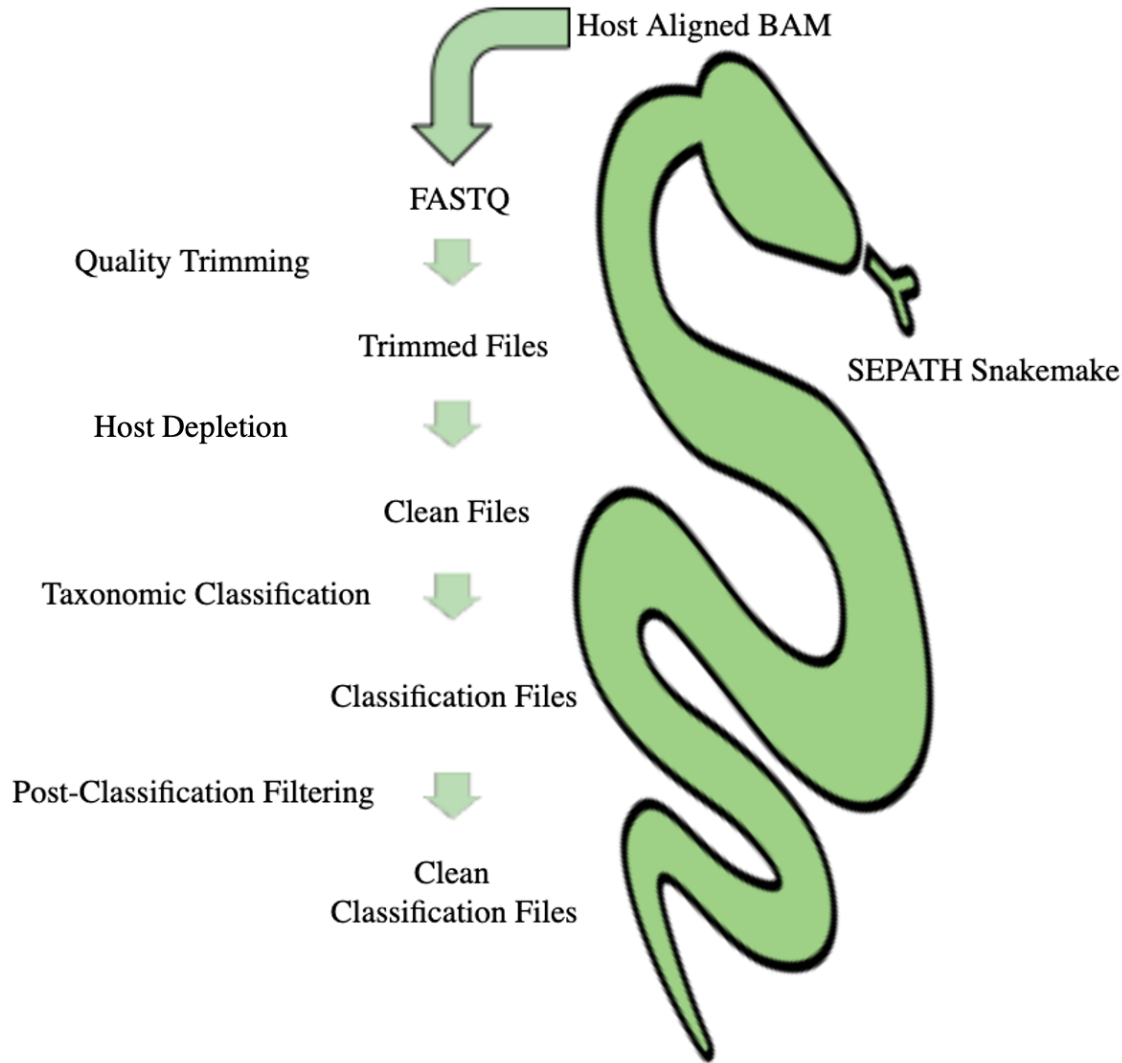


Figure 2.11: SEPATH template computational pipeline. The top-performing pipelines from this benchmark are provided as a template for users to adjust according to their own job scheduling systems and resource availability. SEPATH provides two main pathways: a bacterial pipeline using mOTUs2 classifications on raw sequencing reads and a bacterial and viral pipeline employing Kraken on metagenomic contigs assembled using non-human reads with MetaSPAdes.

2.4 Discussion

Numerous tools were trialled for the purpose of metagenome simulation including GemSIM²¹⁷ and GRINDER²¹⁸. Ultimately, BEAR²¹⁹ was selected because it presented as the most practical tool that had a high throughput, produced paired-end reads, applied an error model to simulated reads based on input sequence data and was the most reasonable to install and get running. It may be of interest to conduct future benchmarks using newer tools for metagenome simulation such as CAMISIM²²⁰. An investigation into the benefits

and drawbacks of each simulation tool would benefit the scientific community.

I have produced pipelines for detecting bacterial genera and viral species in simulated and real whole genome sequence data from cancer samples. These pipelines perform well in terms of sensitivity & PPV and utilize computational resources effectively. We found a drop in performance at the more specific taxonomic levels (figure 2.5).

The two top performing classification tools, Kraken and mOTUs2, have very different underlying mechanics despite achieving similar performance. Kraken begins analysis by breaking down each input read into its constituent k -mers and matching each of these to a user generated database. The sequence is classified probabilistically by the leaf in the highest weighted root-to-leaf path in a taxonomic tree¹⁰³. In comparison to Kraken, mOTUs2 uses a targeted approach by analysing 40 phylogenetic bacterial marker genes for classification. Overall mOTUs2 uses 7,726 marker gene based operational taxonomic units (mOTUs). Classifications are obtained by an alignment to this database using BWA-MEM with default parameters^{104,107}.

mOTUs2 has been developed with abundance in mind. It intuitively estimates the proportion of sequences estimated to originate from unknown taxa (denoted by ‘-1’ in mOTUs2 reports) and adjusts abundance values from detected clades accordingly to account for this. The Kraken read distribution can be improved by using a Bayesian framework to redistribute the assigned reads using Bracken¹⁰⁹. A comparison of relative abundance between mOTUs2 and Bracken was carried out during the production of mOTUs2 as reported in Milanese *et al.* 2019¹⁰⁴; which demonstrated that mOTUs2 appeared to provide more accurate predictions. Kraken is therefore recommended for accurate representations of microbial presence/absence and I suggest that using abundance weighted β -diversity metrics from these pipelines should be interpreted with caution. A caveat of the assembly-Kraken pipeline is that it requires successful metagenomic assembly. Whilst MetaSPAdes worked well on our simulations, idiosyncrasies of different technologies and datasets may hinder successful assembly. In this event it should be recommended to run Kraken classification on quality trimmed and human depleted sequencing reads without assembly.

The data in this chapter supports use of mOTUs2 for quantitative bacterial measurements, which together with the high classification performance on simulated data suggests that both binary and non-binary β -diversity measures would be representative of the true values of the dataset - conferring accuracy in bacterial community profiling. MOTUs2 differs to current methods that rely purely on bacterial reference sequences by incorporating data from MAGs (metagenome assembled genomes), suggesting that mOTUs2 captures a differing scope of classifications to our Kraken database, which was developed using reference genomes. Although both tools are state-of-the-art at the time of writing, they are likely to contain biases in terms of what they can classify. This might pertain to differences in sequencing efforts between sampling sites. The human gut microbiome for example is currently believed to be better characterized than other body sites¹⁰⁴.

Caution must be urged when working at the species level using these tools. At lower taxonomic levels, the retention of BAM files from mOTUs2 could theoretically allow for subsequent investigations at more specific taxonomic nodes (*i.e.* strain level) by investigating single nucleotide variation. Kraken also automatically produces sub-genus level classifications where the input data and database permits. Validating performance at these taxonomic levels would require extensive performance benchmarking which has not been

conducted here. Benchmarking tools and databases as they emerge is an important task as they greatly influence performance. It is hoped that utilities presented here will assist future benchmarking efforts.

Using a laboratory protocol designed for human DNA may be imperfect to sequence microbial DNA, particularly when the biomass of microbes in human tissue is low. Developing a sequencing protocol that is optimal for microbes is a widely researched topic. Differences in protocol at each of the stages of sequencing has been shown to produce different results²²¹.

The DNA of some microorganisms are easier to extract than others²²². Ultra-sonication or agitating samples in the presence of a grinding media (*'bead-beating'*) has been shown to result in higher yields of DNA and statistically significant differences in the recovery of clinically relevant microbes^{222,223}. Also the use of extraction kits from differing manufacturers (Qiagen vs MO BIO) has been shown to differentially deplete particular types of bacteria (*e.g.* gram positive organisms)²²⁴. These facts contribute to a widely reported *'reproducibility crisis'* in microbial bioinformatics²²⁵⁻²³¹ and stresses the importance of validating any results with differing technologies and cohorts.

The use of SEPATH pipelines on real cancer sequence data suggests overall agreement between Kraken and mOTUs2 but reveals important considerations for subsequent analysis. Kraken appears to be more sensitive than mOTUs2 when applied to real data. One reason for this could be due to shorter read lengths (2x100bp in real sample data compared to 2x150bp in simulated data). However, it is later revealed in this thesis (chapter 4) that mOTUs2 is prohibitively insensitive on other real data. Therefore, this difference in sensitivity is likely a peculiarity of mOTUs2 which was not captured by the simulated datasets.

Using sequencing protocols optimised for microbial detection, as compared to human sequencing projects, is likely to result in higher and more even microbial genome coverage and subsequently better performance by mOTUs2. This has been recently demonstrated in the analysis of faecal metagenomes of colorectal cancer patients²³². In this study mOTUs2 assigned interesting *'unknown'* classifications which would not be captured by standard Kraken databases. Therefore, Kraken is recommended as the primary tool of investigation on tissue, but mOTUs2 has great potential in the confirmatory setting and for investigating unknown taxa. A consensus approach of different tools on much larger real datasets would likely help in distinguishing between the peculiarities (particularly false positives) of individual tools and true positive results which would benefit the accurate characterization of human tissue metagenomes.

2.5 Conclusions and Future Directions

A benchmark into metagenomic classification tools has revealed high performing approaches to process host-dominated sequence data with low microbial abundance on a large selection of challenging simulated datasets. These pipelines are provided for the experienced user to adjust according to their own resource availability. Likewise, our simulated metagenomes are available for others to use freely for independent investigations. mOTUs2 provides fast and accurate bacterial classification with good quantitative predictions. MetaSPAdes and Kraken provide bacterial and viral classification with assembled contigs as a useful downstream output. It has been shown that SEPATH forms a consensus alongside PathSeq

to achieve near perfect genus level bacterial classification performance. Using SEPATH pipelines will contribute towards a deeper understanding of the cancer metagenome and generate further hypotheses regarding the complicated interplay between microbes and cancer.

Limitations of this benchmarking exercise are that not all tools and parameters were tested and the simulated datasets do not reflect the full diversity of datasets available. It might be that some algorithms are optimal for certain datasets. Additional benchmarking efforts will be required as new tools and sequencing technologies (*e.g.* long read technology) are produced.

2.6 Methods

2.6.1 Metagenome Simulations

Metagenomes were simulated using a customized version of BEAR (Better Emulation for Artificial Reads)²¹⁹ and using in-house scripts to generate proportions for each reference genome (figure 2.2, <https://github.com/UEA-Cancer-Genetics-Lab/BEAR>). A full list of the 77 species level taxonomy and filenames for each reference genome used is available in appendix section B. These proportions were based on previously analysed cancer data⁹⁶. Firstly, the number of total bacterial reads (in both pairs) was generated by a random selection of positive values from a normal distribution function with a mean of 28,400,000 and a standard deviation of 20,876,020. These parameters were based on empirical observations from microbial classification of PCAWG (pan-cancer analysis of whole genomes) data. The number of human reads in the sample was set to the difference between this number and 600 million (the total number of reads in both pairs). The number of bacterial species were randomly sampled from the reference species available and the number of bacterial reads available were picked from a gamma distribution of semi-random shape (randomising the shape parameter). The number of reads for each bacterial species were distributed among contigs proportionately depending on contig length. This produced a file with contigs and proportions of final reads which was provided to BEAR to generate paired-end FASTA files for each of the 100 metagenomes with approximately 300 million reads per paired-end file. An error model was generated following the BEAR recommendations from a sample provided by Illumina containing paired-end reads that were 150 base pairs in read length (https://basespace.illumina.com/run/35594569/HiSeqX_Nextera_DNA_Flex_Paternal_Trio). This sample was selected to best resemble data originating from within Genomic England's 100,000 genomes project. These simulated metagenomes can be downloaded from the European Nucleotide Archive (<https://www.ebi.ac.uk/ena/data/view/PRJEB31019>).

2.6.2 Tool Performance Benchmarking

Samples were trimmed for quality, read length and adapter content with Trimmomatic¹⁹⁶ prior to running any classification (default parameters were minimum read length = 35 and minimum phred quality of 15 over a sliding window of 4). SEPATH has trimming parameters set as default that prevent any excessive removal of data (including any reads that may be microbial). The default $k = 31$ was used for BBDuk and Kraken. An investigation into the adequacy of this k -mer size and the rationale behind utilising both paired end reads

and single end reads that have lost their pair in quality trimming is provided in appendix section B.0.2.

Performance estimates were obtained by converting all output files into a common file format which were compared against the true composition by string matches and NCBI taxonomic ID. The total number of true positive results, false positive results and false negative results were used to calculate F1-score, sensitivity and PPV were calculated as follows:

$$SSV(recall) = \frac{TP}{TP + FN} \quad (2.1)$$

$$PPV(precision) = \frac{TP}{TP + FP} \quad (2.2)$$

$$F1 - Score = \frac{2}{SSV^{-1} + PPV^{-1}} \quad (2.3)$$

2.6.3 Real Cancer Whole Genome Sequence Analysis

Sequencing data from cancer tissue was obtained from the cancer genome atlas (TCGA-CESC and TCGA-STAD)²³³, International Cancer Genome Consortium (ICGC) PedBrain Tumor Project²³⁴, and ICGC Chinese Gastric Cancer project²³⁵ (not publicly available at time of writing). These sequencing reads were pre-processed through a common pipeline to obtain reads unaligned to the human genome²³⁶ and were additionally quality trimmed and depleted for human reads using SEPATH standard parameters but with a database consisting of human reference genome 38, African ‘*pan-genome*’²⁰⁴ project sequences and COSMIC cancer genes²⁰³. Kraken was run on quality trimmed reads and a confidence threshold of 0.2 was applied to the reports. mOTUs2 was ran for genus level analysis on the same reads using two marker gene minimum and a non-standard minimum alignment length of 50 to account for shorter read length. Kraken files had a minimum read threshold applied of 100 reads for each classification and mOTUs2 results were left unfiltered.

2.6.4 Computational Tools and Settings

All analysis for figures was carried out in R version 3.5.1 (2018-07-02). All scripts and raw data used to make the figures can be found on https://github.com/UEA-Cancer-Genetics-Lab/sepath_paper. This analysis used the following software: picard 2.10.9, samtools v1.5, BEAR (<https://github.com/UEA-Cancer-Genetics-Lab/BEAR> commit: a58df4a01500a54a1e89f42a6c7314779273f9b2), BLAST v2.6.0+, Diamond v0.9.22, MUMmer v3.2.3, Jellyfish v1.1.11, Kaiju v1.6.3, Kontaminant (pre-release, GitHub commit: d43e5e7), KrakenUniq (github commit: 7f9de49a15aac741629982b35955b12503bee27f), MEGAHIT (github commit: ef1bae692ee435b5bcc78407be25f4a051302f74), MetaPhlAn2 v2.6.0, Gottcha v1.0c, Centrifuge v1.0.4, FASTA Splitter v0.2.6, Perl v5.24.1 bzip2 v1.0.5, gzip v1.3.12, and Singularity v3.2.1.

Python v3.5.5 was used with the exception of BEAR, which used Python 2.7.12. Python modules used the following: SeqIO of BioPython v1.68, os, sys, gzip, time, subprocess, and glob. The following are the R packages used and their versions: Cowplot v0.9.3, dplyr v0.7.6, ggExtra v0.8, ggplot2 v3.0.0, ggpubr v0.1.8, ggrepel v0.8.0, purr v0.2.5, ggbeeswarm v0.6.0, see v0.2.0.9, RColorBrewer v1.1-2, readr v1.1.1, reshape2 v1.4.3, tidyr v0.8.1, and tidyverse v1.2.1.

Chapter 3

Microbes of Prostate Cancer Exosomal Urine and Potential Microbe-Host Interactions

Some of the work presented in this chapter is contributing towards a submitted manuscript by Hurst, Meader, Gihawi and Rallapalli *et al.*

3.1 Prostate Cancer Background

Prostate cancer is the most common male cancer in the UK with 48,600 new cases in 2017 (26% of newly diagnosed cancers)²³⁷. Incidence is expected to rise, peaks at ages 75-79 and is elevated in men of African ancestry^{237,238}. Prostate cancer is a comparatively slow cancer to progress, is genetically heterogeneous and often polyclonal in nature²³⁹. Almost 80% of men will survive their disease for ten years or more but despite this, nearly 12,000 men succumb to their cancer every year in the UK²³⁷. Therefore, there exists an unmet clinical need to stratify patients that will develop aggressive tumours from those that will require only limited clinical intervention^{240,241}.

The pathway of clinical investigation begins with a patient exhibiting lower urinary tract symptoms and/or a high PSA (Prostate Specific Antigen) blood test result. Patients are then subject to additional investigations and stratified using the D'Amico risk classification system²⁴² which considers PSA, clinical stage and the Gleason²⁴³ pattern from histopathological analysis of prostate biopsy tissue (figure 3.1, appendix table C.1).

PSA is notoriously non-specific and varies between cohorts with an estimated sensitivity and specificity of 83.4% and 38.9% respectively for detecting prostate cancer (1.1 ng ml^{-1} threshold)²⁴⁴⁻²⁴⁶. Following PSA testing and DRE (Digital Rectal Examination), if there remains the suspicion of prostate cancer further investigations occur: patients are offered a tissue biopsy and MRI (Magnetic Resonance Imaging). Multi-parametric MRI is being introduced to produce high resolution images of the prostate, which is expected to reduce the number of men that require invasive biopsies by 27%²⁴⁷. Standard practice had been to sample via a TRUS (Transrectal Ultrasound) biopsy, but transperineal template biopsies are now more regularly used. This samples more of the prostate, reduces false negative biopsies

as a consequence of undersampling the prostate and reduces the incidence of post-procedure infection²⁴⁸. The Gleason system provides two scores describing the predominant and less-predominant patterns of cellular differentiation in the sample. Scores range from one to five where one describes healthy tissue composed of small, well-formed glands (although rarely observed because most elderly men have some degree of prostatic hyperplasia), and grade five describes tissue with no evidence of glands, with pleomorphic cancerous cells (adenocarcinoma) disseminated throughout the tissue²⁴³. Men with a very high PSA (>20) or a high Gleason are sent for a bone scan to search for osseous metastases.

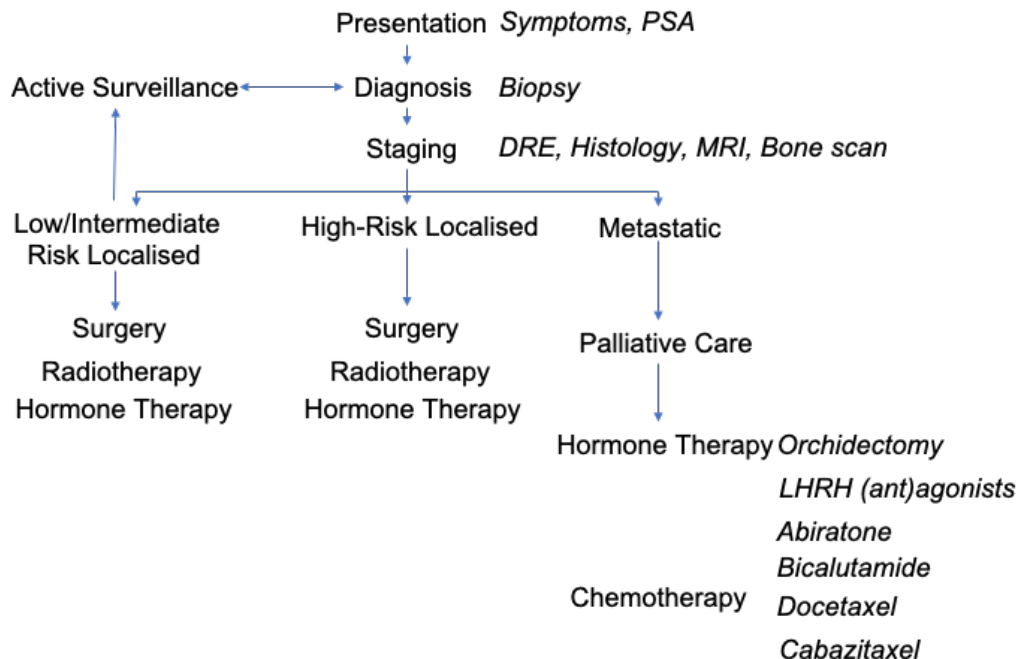


Figure 3.1: The patient pathway for prostate cancer in the UK. Patients are stratified based on the D'Amico classification system²⁴² which uses Digital Rectal Exam (DRE), Magnetic Resonance Imaging (MRI) and histopathology of a template or transrectal ultrasound guided biopsy (TRUS). Patients can also be offered focal therapy, particularly those with low or intermediate risk²⁴⁹.

Men demonstrating signs of metastases are generally subject to palliative care. They are offered hormone therapy (typically alongside radiotherapy) to help limit any androgen driven cancer growth. This can occur in many forms, the most common of which are that of luteinizing hormone releasing hormone (ant)agonists (LHRH), CYP17 inhibitor abiraterone acetate, androgen receptor antagonist bicalutamide and bilateral orchidectomy to prevent gonadal testosterone synthesis. Chemotherapy in the form of systemic microtubule inhibitors (Docetaxel and Cabazitaxel²⁵⁰) can also be offered when prostate cancers develop resistance to hormone therapy.

Arguably the more complicated therapeutic pathway is that of men with intermediate risk localised prostate cancer. Many tumours in this category will remain indolent. However, a subset of these tumours will transition to an aggressive phenotype. Although men can opt

for radical therapy, this is not without notable side effects including sexual and urological dysfunction²⁵¹. Given these life altering side effects, patients may prefer to monitor their condition until intervention is required by frequent MRIs, biopsies and PSA tests (active surveillance)²⁵².

Despite this clinical framework, men are still dying and overtreatment is occurring. There is a need to refine this framework by identifying the tumours that will become aggressive and to define what treatment options are best for these patients. To do this requires two things: better biomarkers of disease progression and a better understanding of the underlying disease biology.

3.1.1 Anatomy of the Prostate

The bladder rests superior to the prostate and protrudes into the urethra which runs medially through the prostate (figure 3.2). The prostate is composed of glandular tissue with an ordinary function to produce and release components of semen into the seminal vesicle. The combination of the secretory nature of the prostate and its proximity to the urethra has sparked interest in the use of urine as a potential pool of prostate related biomarkers^{241,253–255}. Whilst most tumours occur in the peripheral zone and can be palpated with a DRE, 20-30% are estimated to originate from the transitional zone; positioned adjacent to the urethra²⁵⁷. This might hint that urine could be of particular diagnostic importance in transitional zone tumours. Furthermore, urine has the potential to contain information from the entire prostate whereas biopsy (particularly TRUS biopsy) is limited to the sampling site, often meaning that tumours are not sampled and diagnosed.

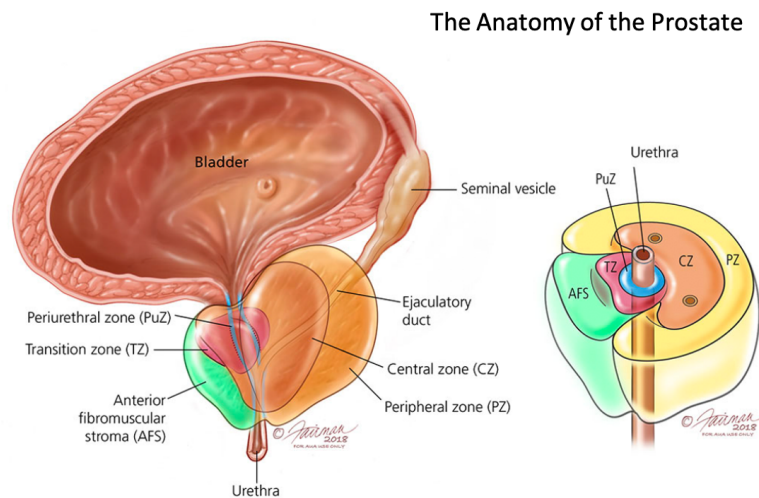


Figure 3.2: The anatomy of the prostate. Adapted from the urology care foundation²⁵⁸.

3.1.2 The History of Microbes and Prostate Cancer

Unlike other cancers, a firm link between microbes and prostate cancer remains elusive^{259–261} but there is a growing body of evidence that suggests there may be an association.

A hallmark of cancer, chronic inflammation has been well-documented in many tissues prior

to its malignant transformation²⁶². Bacterial infection has been known to cause prostatitis for many years^{261,265}; *E. coli* is one such bacteria that is documented to cause clinically pertinent prostate infections and prostatitis¹⁹⁹. Feng *et al.*^{266,267} has found evidence of microbial reads in prostate tissue whole genome/transcriptome sequences, although these are limited in sample size. Some of the observed genera include *Escherichia*, *Propionibacterium*, *Acinetobacter* and *Pseudomonas* (figure 3.3).

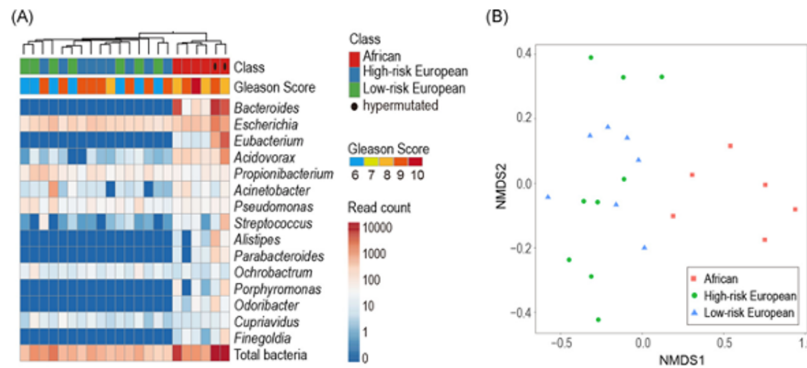


Figure 3.3: Bacteria detected in prostate tissue whole genome sequencing. **A)** A heatmap of the top 15 abundant genera identified by Feng *et al* 2019. **B)** An NMDS (non-metric multidimensional scaling) plot of bacterial communities found in prostate tissue, coloured by geographical location and clinical risk. Image adapted from Feng *et al.* 2019²⁶⁶.

Feng *et al.*, reports differing prostate bacteria communities between African and European populations but no difference between high and low-risk Europeans (figure 3.3). Whilst this study has highlighted the possibility of detecting bacteria in prostate WGS data, it should be noted that there are several limitations. Firstly, all patients received antibiotics prior to sample collection (Ciproxin for African and Ceftriaxone for ethnically European patients). Sample collection also differed between the two cohorts: samples from African participants were obtained by a TRUS biopsy and samples from the Australian cohort were obtained immediately post radical prostatectomy. Although all samples were sequenced together to avoid a batch effect, the differences in antibiotics and sample collection could have a large effect on the estimated microbial community.

Alane *et al.* used 16S ribosomal RNA to investigate the association of the urinary and faecal microbiota with prostate cancer²⁶⁸. This study also had small sample sizes ($N=30$). They found that the urine microbiome clustered separately in those with prostate cancer from those with no diagnosis of prostate cancer. This difference was caused by a decrease in abundance of several bacterial taxa, such as *Lachnospira*, *Clostridium*, *Acetanaerobacterium* and *Faecalibacterium* in patients with prostate cancer. They detected no notable differences in bacterial abundance between benign and malignant tumours. Banerjee *et al.* 2019²⁶⁹ did report different microbial signatures between benign and malignant prostate cancer samples using an array-based technique to identify microbes. These emerging studies suggest that there could indeed be an association between microbes and prostate cancer.

The Bacillus of Calmette and Guerin (BCG) vaccine which provides some immunity to tuberculosis has been used as an immunotherapy in renal cancer since the mid 1900's^{270,271}.

This suggests a possible role for microbes to influence urological cancers. Historically it has been believed that urine was sterile in healthy individuals. Although bacteria are known to be able to colonise the bladder^{272–274} and kidneys (pyelonephritis)^{275–277}; there is emerging evidence of low levels of bacteria in the urine²⁷⁸, which could reflux into prostatic ducts²⁷⁹. Moreover, urinary dysbiosis has been associated with bladder cancer, with cases showing more evidence for *Fusobacterium* compared to healthy participants²⁸⁰. There is therefore a reasonable body of evidence emerging that microbes in the urine/prostate could be associated with prostate cancer.

To summarise, the investigation of the urinary microbiome in prostate cancer is warranted by several strands of evidence. A more concise summary of these examples include:

- Bacteria can cause prostatitis^{261,265}.
- Urine refluxes into prostatic ducts, a potential route of infection²⁷⁹.
- The urethra runs through the prostate which is a shorter distance to the urethral orifice than the bladder and kidneys (also sites of bacterial infection^{272–277}).
- The preliminary identification of microbial DNA in prostate cancer tissue^{266,267}.
- Emerging evidence associating urinary dysbiosis with bladder cancer²⁸⁰ and prostate cancer²⁶⁸.

3.1.3 Urine as a Source of Biomarkers

Urine offers an incredible opportunity for biomarkers. Collection is non-invasive and studies investigating non-bacterial urinary biomarkers for prostate cancer have produced better diagnostic and prognostic tests than standard clinical practice^{241,253,281–283}. There is the potential for microbial identification in similar samples to be a successful biomarker. The availability of a routine prognostic assay with sample collection at home could have profound implications for men living with prostate cancer, particularly for patients on active surveillance.

One such area of research is that of RNA found in urinary extracellular vesicles including exosomes²⁵³. Exosomes are small vesicles (50-100nm) that are ordinarily secreted by all tissues (figure 3.4). Exosomes have been described by researchers as biomarker ‘treasure chests’²⁸⁴. They often contain tissue-type specific markers and have been suggested as an alternative pathway for lysosomal degradation, assisting the cell to remove superfluous content. Exosomes seemingly have an immunological role with proteins/RNA being transferred between cells and donor exosomes can confer expression of MHC (Major Histocompatibility Complex) proteins on recipient cell surfaces²⁸⁵. Exosome production is increased in states of cellular stress and tumour promoting exosomes have been suggested to promote cell survival and modulate the immune response in the tumour microenvironment^{286,287}. The underlying biology behind this and the potential for therapeutic exploitation however is unclear. Antigen presenting cells (*i.e.*, dendritic cells) have been shown to markedly increase the production of extracellular vesicles following a phagocytic event²⁸⁸. Additionally, and perhaps counterintuitively, the phagocytosed bacteria can also be released back into the environment²⁸⁸.

The Biogenesis and Secretion of Small Vesicles in Human Cells

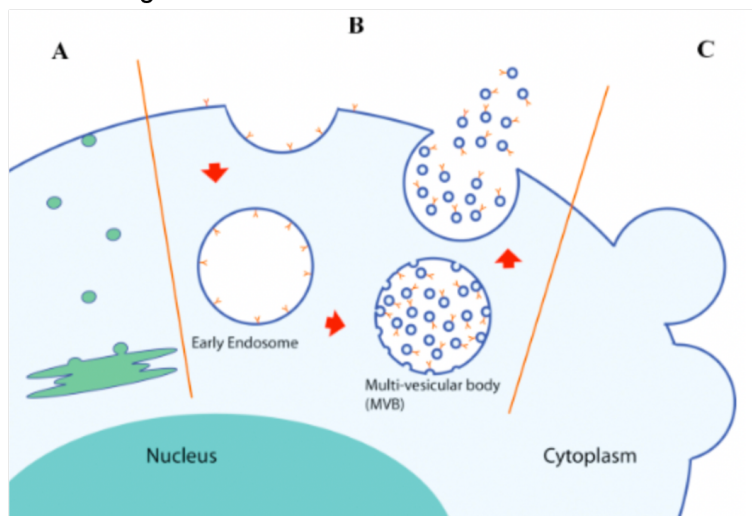


Figure 3.4: The biogenesis and secretion of small vesicles. **A)** The release of signal proteins from the rough endoplasmic reticulum and golgi apparatus. **B)** Merocrine secretion pathway. Cell membranes fold inwards to form early endosomes. The cytoplasmic content is processed into smaller vesicles called exosomes. The endosome then fuses with the plasma cell membrane to release exosomes from the cell. **C)** larger vesicles (50-1000nm) can also be expelled by the cell by the budding outwards as a form of apocrine signalling. Image adapted from Duijvesz *et al.* 2011²⁸⁴.

Studying exosomes in urine could indicate characteristics of cells (in particular stressed/cancerous cells²⁸⁹), and there is the opportunity for microbe specific sequences to be present in exosomes - either as whole bacteria or fragmented nucleic acids²⁹⁰.

This chapter will explore the link between microbes found in urine and tissue and to investigate their associations with prostate cancer (figure 3.5). This will be carried out by investigating two sources of data:

- Urine exosome fractions from participants post-DRE ($N=40$) at Norfolk and Norwich University Hospital (NNUH). These participants were under investigations for prostate cancer at a haematuria clinic.
- Bacterial isolates that have been cultured and sequenced from donor urine. Additional reference genomes are also analysed.

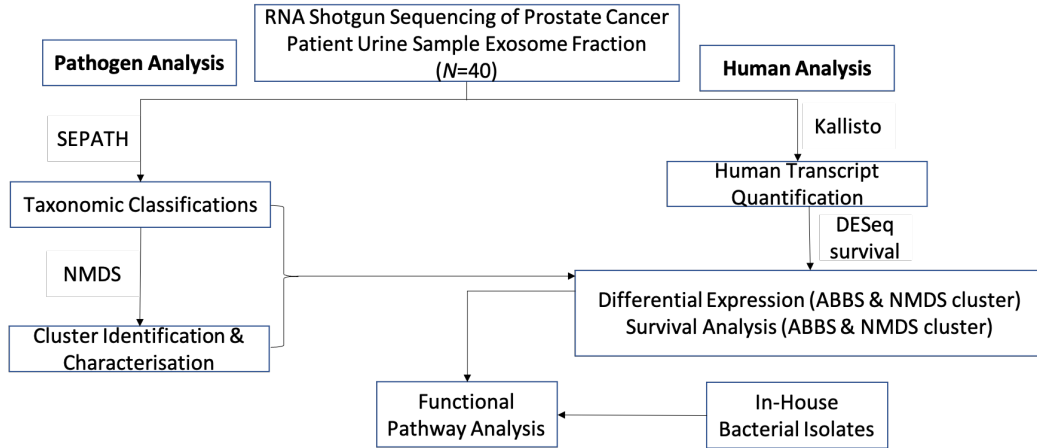


Figure 3.5: An overview of this chapter. Most of this chapter focuses on RNA sequencing from the urine of 40 prostate cancer patients. This was subject to microbial analysis investigating two partitions of the samples, NMDS (Non-Metric Multidimensional Scaling cluster) and ABBS (Anaerobic Bacterial Biomarker Set). These partitions were used to investigate differential expression of human genes and survival analysis. Differentially expressed genes and enriched pathways were determined for these partitions. Additionally, the functional pathways present in in-house isolates were examined and compared. These isolates were obtained from culturing urine samples performed by Dr Rachel Hurst and Dr Emma Meader. Assembly and taxonomic classification was carried out by Dr Ghanasyam Rallapalli.

3.2 Results

3.2.1 Evidence for a Range of Genera in Post-DRE Urine Exosome Samples

SEPATH using Kraken was run with standard parameters on shotgun RNA sequencing data from the exosomal fraction of urine samples from patients with suspected and confirmed prostate cancer ($N=40$). Contaminant genera according to Salter *et al.* 2014¹⁷² were removed. Taxa with < 100 reads were set to zero to limit false positive classifications as described in chapter 2. Sparse taxa were removed for display purposes (figure 3.6). After removing these taxa, there was evidence of 101 genera. After excluding non-zero values, the number of classified sequencing reads (across all classifications) ranged from a minimum of 101 to a maximum of 76,946 (median: 246, interquartile range: 350.2). The top five genera with the highest median sequencing reads (in positive samples) include: *Porphyromonas*, *Mobiluncus*, *Dialister*, *Campylobacter* and *Varibaculum* (figure 3.6).

The top genera is the gram-negative anaerobe, *Porphyromonas*, is a genus containing human pathogens²⁹¹ and does not appear to be a frequent sequencing contaminant. Genera that are frequently reported in urine or the urogenital tract were also detected such as *Mobiluncus*, *Dialister*, *Lactobacillus*, *Veillonella*, *Prevotella*, *Anaerococcus* and *Finegoldia*^{292,293}. *Cutibacterium* (detected in 19 samples) has been shown to cause prostatic inflammation²⁹⁴. Other genera such as *Fusobacterium*, *Porphyromonas*, *Prevotella* have

been associated with cancers, mostly of the digestive system^{295,296}. Porter *et al.* reported genera such as *Varibaculum* and *Anaerococcus* were also found as enriched in the urine of prostate cancer patients²⁹⁷.

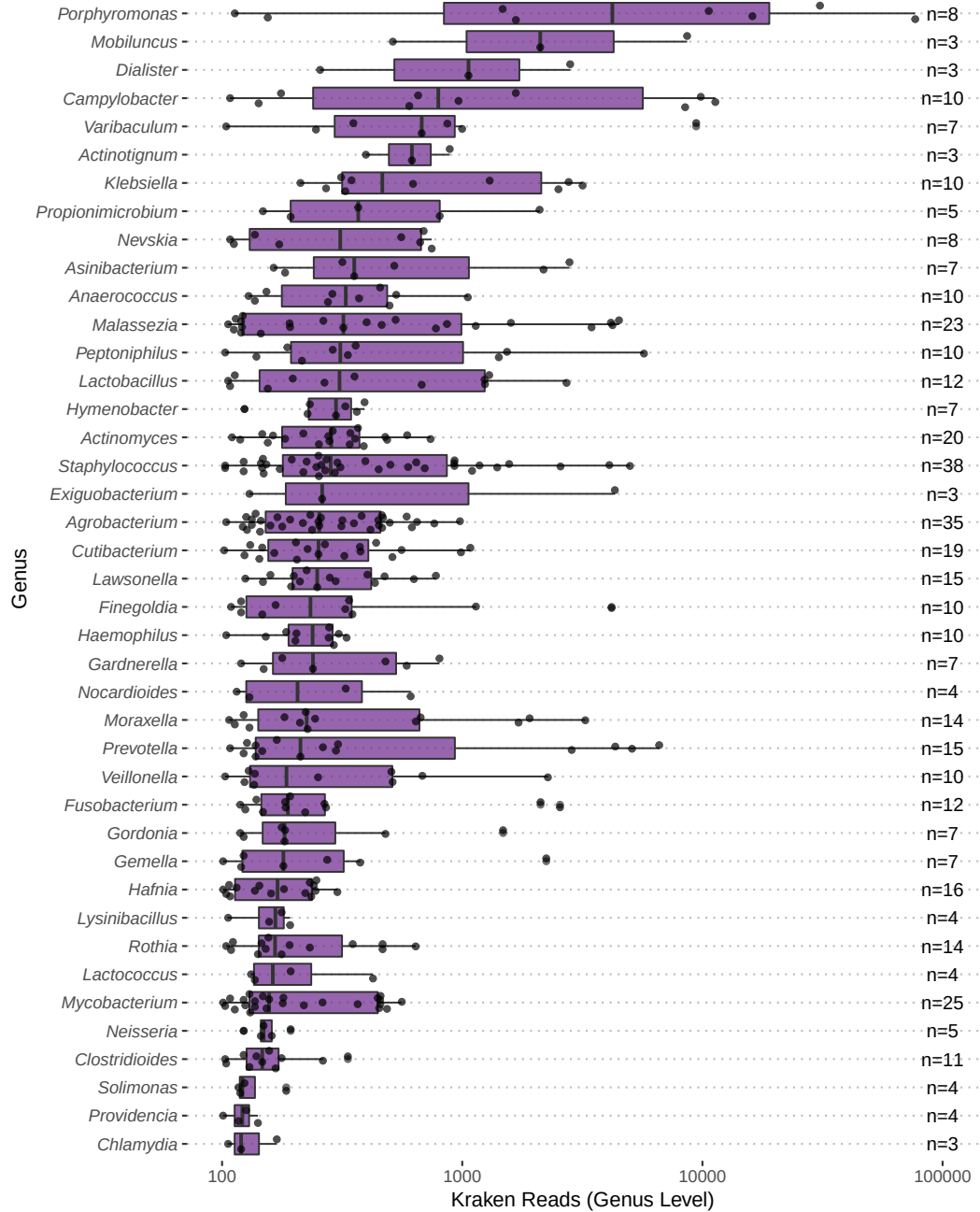


Figure 3.6: The top genera detected in a prostate cancer urine RNA sequencing dataset. Suspected contaminants were removed and taxa present in fewer than three samples are not displayed. Full results are available on the thesis repository. Genera are ordered by median number of reads assigned using Kraken. The number of samples with any evidence for the taxa is indicated on the right.

Although a bulk list of common contaminants were removed, I noted that a few genera are common contaminants and do not seem to typically reside within humans. *Neusikia* is more commonly known as an environmental bacteria²⁹⁸. *Agrobacterium* is likely to be a contaminant as it is frequently identified in soil²⁹⁹, but some evidence might suggest otherwise^{300–303}.

3.2.2 Interrogating a Prognostic Panel of Bacteria

Our laboratory group found that 16S family level classifications of urine exosome samples formed three distinct clusters, including one which demonstrated markedly elevated rates of skeletal metastases (originally analysed by Dr Ghanasyam Rallapalli and repeated by myself; see appendix section C.4 for my version). Based on this, a team effort identified a list of genus level taxa that might contribute towards a poor prognosis of prostate cancer which was termed the Anaerobic Bacterial Biomarker Set (ABBS). ABBS consists of *Ezakiella*, *Peptoniphilus*, *Porphyromonas*, *Anaerococcus*, *Fenollaria* and *Fusobacterium*. Additional information on how this list was determined can be found in appendix section C.4.

I partitioned the patients into those with urine samples that had evidence of at least one ABBS genera present ($n = 19$) and those without ($n = 21$). Patients in the ABBS group had a significantly reduced time to progression ($p = 0.0013$ for all patients, $p = 0.00079$ for patients diagnosed with prostate cancer; figure 3.7). The following were used as markers of progression: metastasis, an increase in histological staging, prostate cancer associated deaths, biochemical recurrence or an increase in PSA sufficient to alter the course of clinical treatment. These multiple endpoints were used due to their impact on patient lives and the fact that prostate cancer is a slow progressing disease, which limits the use of single endpoints in a dataset with limited follow-up time.

All of the participants with at least one of these genera present in their urine either showed signs of worsening disease or they were censored in the RNA sequencing cohort (figure 3.7). Externally, of the data presented in this thesis, I also applied this method back to the original 16S data and to a targeted alignment approach on whole genome sequence data conducted by Dr Ghanasyam Rallapalli, both of which demonstrated similar trends in survival (appendix section C.5).

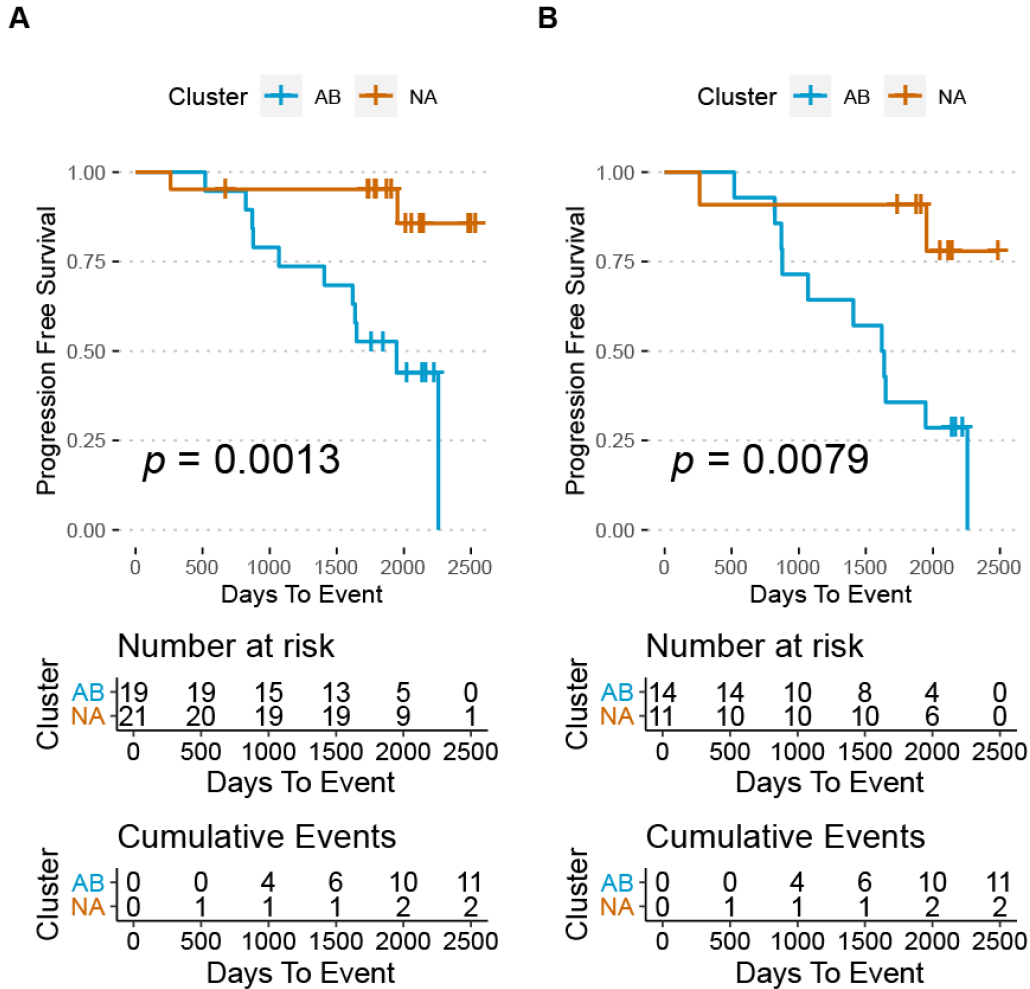


Figure 3.7: Kaplan-Meier plots of time to progression based on whether patients had evidence for ABBS genera in their urine. Samples in AB group contained ABBS genera (blue) and the NA group did not (orange). Vertical lines indicate limits of patient follow-up. **(A)** All samples and **(B)** only cancer samples. Log-rank test was used to compare survival distributions. The ABBS group constitutes urine samples where at least one of the genera (*Fenollaria*, *Ezakiella*, *Peptoniphilus*, *Porphyromonas*, *Anaerococcus*, *Fusobacterium*) were detected in RNAseq data. The following were used as markers of progression: metastasis, an increase in histological staging, prostate cancer associated deaths, biochemical recurrence or an increase in PSA sufficient to alter the course of clinical treatment.

3.2.3 ABBS Isolates from Patients Appear Functionally Related

Bacteria was isolated and cultured in anaerobic conditions from the urine of patients with prostate cancer by Dr Rachel Hurst. These were sequenced with Illumina and Oxford Nanopore Technology and assembled by Dr Ghanasyam Rallapalli using Unicycler³⁰⁴. The assemblies include novel species within the ABBS group of bacteria. To further investigate the functions that these bacteria may exhibit, InterProScan¹⁶⁴ was run on these locally

developed assemblies ($N=50$). To fully represent the ABBS genera, one reference genome of *Fusobacterium nucleatum* *supsp. animalis* was included due to a lack of a pure culture. Pathways were interrogated under numerous ontologies: Gene Ontology (GO) Biological Process, GO Cellular Component, GO Molecular Function, KEGG, MetaCyc and Reactome. The output was the number of ‘hits’ for each metabolic pathway in each strain for each ontology. The resulting matrix was subject to PCA and coloured by ABBS vs non-ABBS genera (figure 3.8).

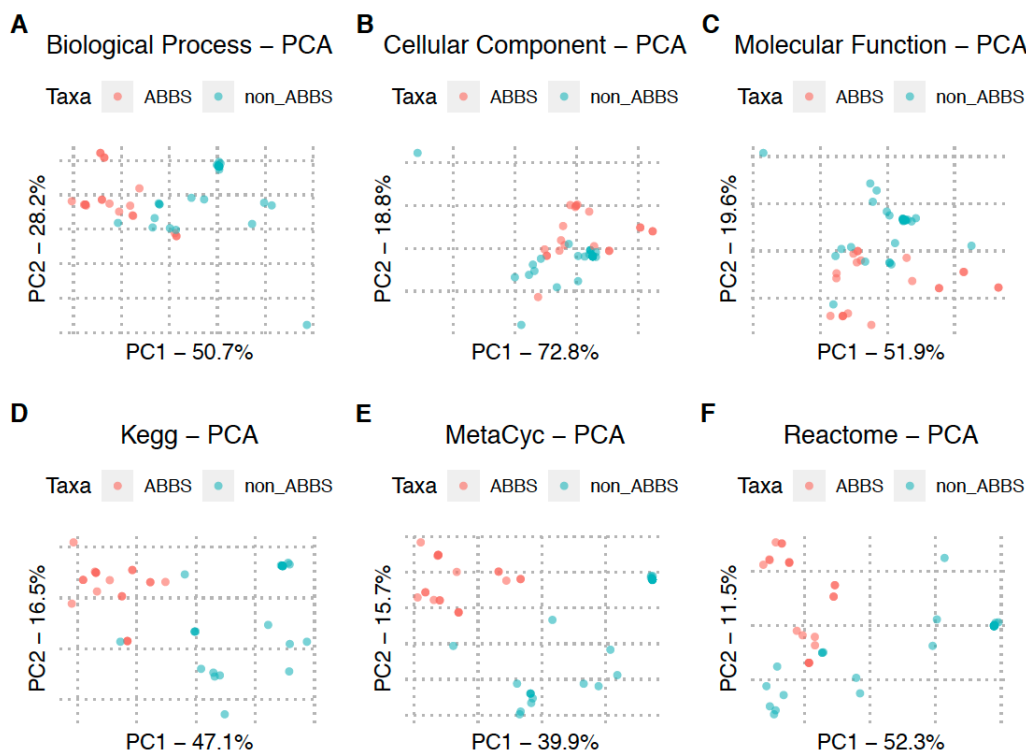


Figure 3.8: Principal components analysis of pathways associated with strains isolated from prostate cancer participants ($N=51$). Isolates belonging to ABBS genera ($n=21$) are highlighted in red and non-ABBS isolates ($n=30$) are blue.

I determined whether each pathway/term were significantly enriched or depleted in genomes belonging to the ABBS genera by determining presence/absence in each genome and performing a Fisher’s Exact test.

Isolates were filtered to only contain one member of each of the following species; ABBS: *F. nucleatum*, *A. urogenitalis*, *Fenollaria sp. nov.*, *Peptoniphilus sp. nov.*, *Porphyromonas sp. nov.*; Non-ABBS: *A. schaali*, *C. acnes*, *E. faecalis*, *F. hominis*, *P. lymphophilum*, *S. epidermidis*, *S. mitis*, *Varibaculum sp. nov.*. A total of 3,591 unique pathways were identified with 1,995,769 total ‘hits’ from all isolates which includes 795 Gene Ontology (GO) Biological Process terms, 99 for GO Cellular Component terms, 1,210 GO Molecular Function terms, 139 KEGG pathways, 800 Metacyc pathways and 548 for Reactome pathways. Total lists of all pathways can therefore be found in the repository associated with this thesis (described in 1.3.2).

197 pathways/terms were significantly different between ABBS and non-ABBS genera ($p < 0.05$), 31 GO Biological Process terms, 3 GO Cellular Component terms, 59 GO Molecular Function terms, 6 KEGG pathways, 36 MetaCyc Pathways and 62 Reactome pathways (figure 3.9). No terms are significant at the 0.05 level after a Benjamini-Hochberg multiple testing correction. Therefore these results should be considered purely for hypothesis generation. Of the 197 differential pathways, 35 were enriched in ABBS genera and 197 enriched in non-ABBS genera (full table of results in thesis repository).



Figure 3.9: Pathways enriched in isolates from ABBS (Anaerobic Bacterial Biomarker Set) genera (right) and from non-ABBS genera (left). Pathways were identified in multiple isolates using InterProScan. The x -axis demonstrates the difference in the percentage of positive samples. A value of 1 indicates that all ABBS bacteria had the pathway and none of the non-ABBS group did. The y -axis indicates the q -value. None of these pathways are significantly differential at the 0.05 level after false discovery adjustment. Gene Ontology (GO) Biological Process (BP), GO Molecular Function (MF), GO Cellular Component (CC).

3.2.4 Ordination Suggests Community Differences

The processed genera read counts from Kraken were converted into a binary presence-absence community matrix which was subject to non-metric multidimensional scaling (NMDS) using the Jaccard distance metric (figure 3.10). The optimal number of clusters was two, as determined by analysing the total within cluster sum of squares for $k = [1, 10]$ (appendix figure C.1). The stress value obtained from NMDS was 0.19 which can be considered fair (discussed in section 1.2.4). PCA was carried out on a log2 normalised community matrix (taxa in < 2 samples removed). We saw a similar pattern to NMDS in that points on the first axis coordinate with NMDS cluster status (figure 3.10). The first two PCA components explained 16.9% and 13.3% of the variation in the dataset (30.4% total).

16 out of 19 samples with ABBS bacteria originate from NMDS cluster B and are significantly overrepresented ($p = 0.0041$; Fisher's exact test). Of the non-ABBS group, 13 samples originated from cluster A and 8 from cluster B. A heatmap of taxa within these samples

highlights that cluster B demonstrates more evidence for various bacterial taxa including the ABBS genera (figure 3.11).

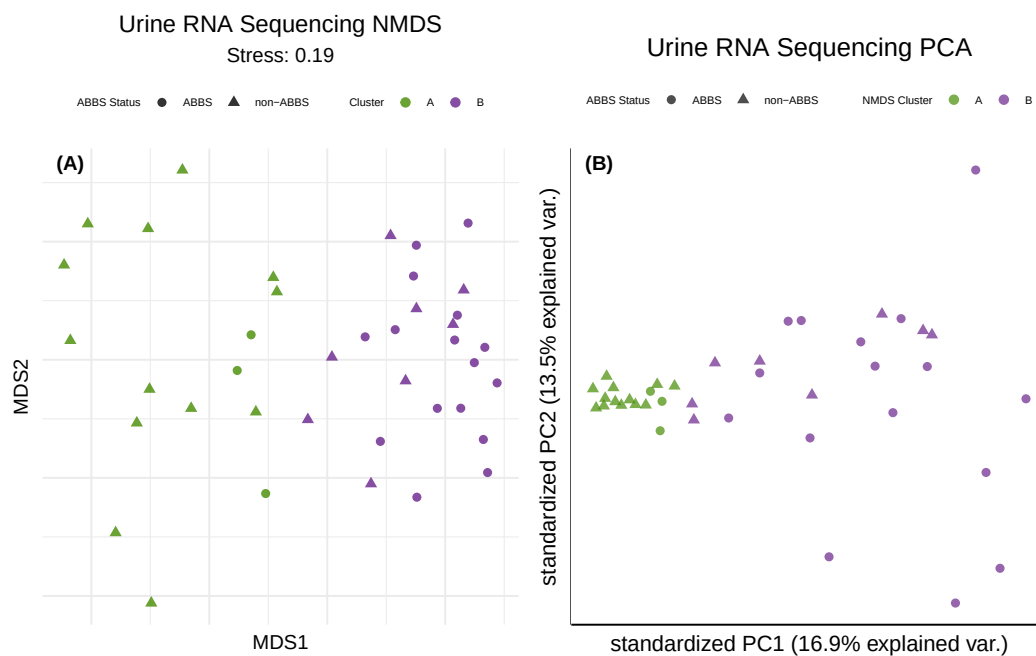


Figure 3.10: Dimensionality reduction of microbial community of urine RNA sequencing Kraken classifications. **A)** Non-metric multidimensional scaling (NMDS) plot obtained using the Jaccard distance. The data points can be optimally partitioned using k -means clustering into two clusters, A (16 samples) and B (24 samples). **B)** Principal Components Analysis PCA plot of the filtered community matrix to remove taxa present in only one sample. Counts were log2 normalised using the decostand function and the first two principal components are displayed. The first three principal components capture 16.9, 13.5 and 11 percent of the total variation in the dataset. Symbols indicate whether the sample has at least one ABBS genera present (circles) or not (triangles).

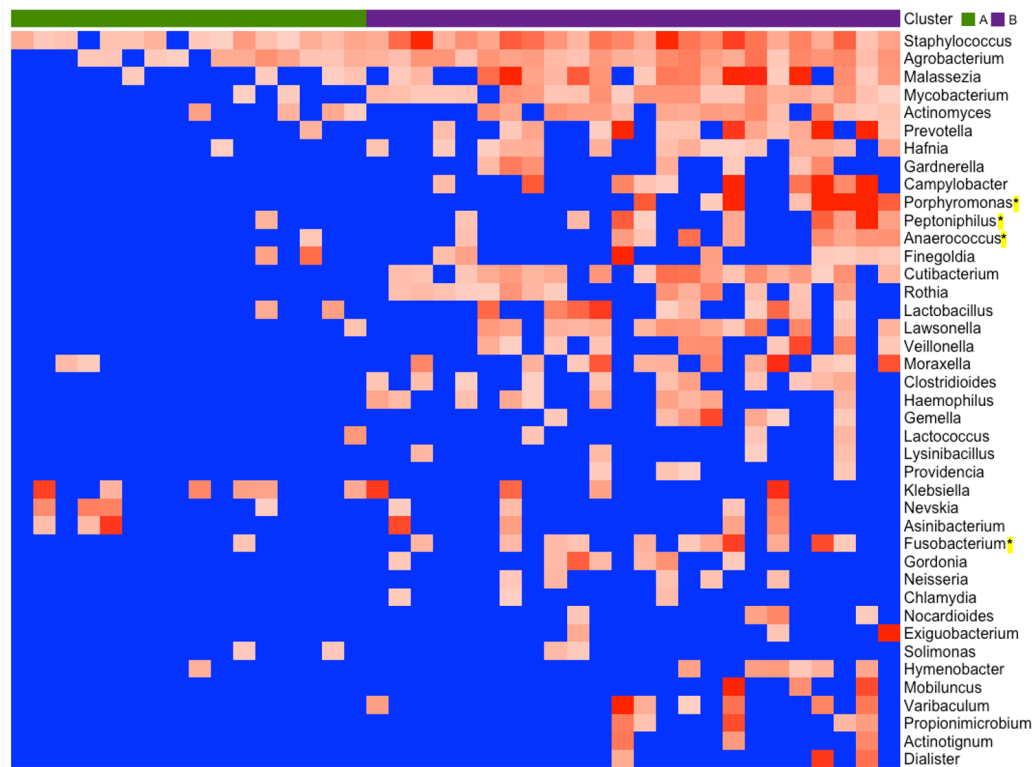


Figure 3.11: Heatmap of log transformed Kraken community data. Blue indicates no evidence for a taxon in a sample, whereas red indicates stronger evidence (legend on log2 of Kraken counts). Samples on the x -axis are ordered by NMDS axis 1. Taxa on the y -axis are ordered by complete-linkage hierarchical clustering and the dendrogram hidden for display purposes. Samples in NMDS cluster A are shown in green and cluster B samples in purple. ABBS genera are highlighted with a yellow asterisk. Sparse taxa in less than two samples were removed for display purposes.

3.2.5 Differential Expression of Human RNA in Exosome Urine

120 human genes were differentially expressed between ABBS groups ($p < 0.05$; DESeq2 test; no multiple testing). Only one gene was significantly differentially expressed ($q < 0.1$): *PCSK5* (appendix section C.6). There is a suspected role for *PCSK5* in anteroposterior development, including urological and anorectal development^{305–310}. *PCSK5* was not found to be significantly differentially expressed between NMDS cluster groups (figure 3.12; $p=0.23$, $q=1$).

235 human genes were significantly differentially expressed between NMDS clusters A and B ($q < 0.1$) (appendix table C.2). Only seven were upregulated in cluster A. Of the 228 genes increased in expression in the cluster B (containing more bacteria), the top hit by fold change is *IGHA1* that encodes for IgA constant heavy chain. The next highest fold changes belong to a long non-coding RNA of unknown function called *AC1000827.3* and *IGKC* (Immunoglobulin Kappa Light Chain Constant Region). Other genes with higher expression in cluster B include *CCL5* and *CD74*.

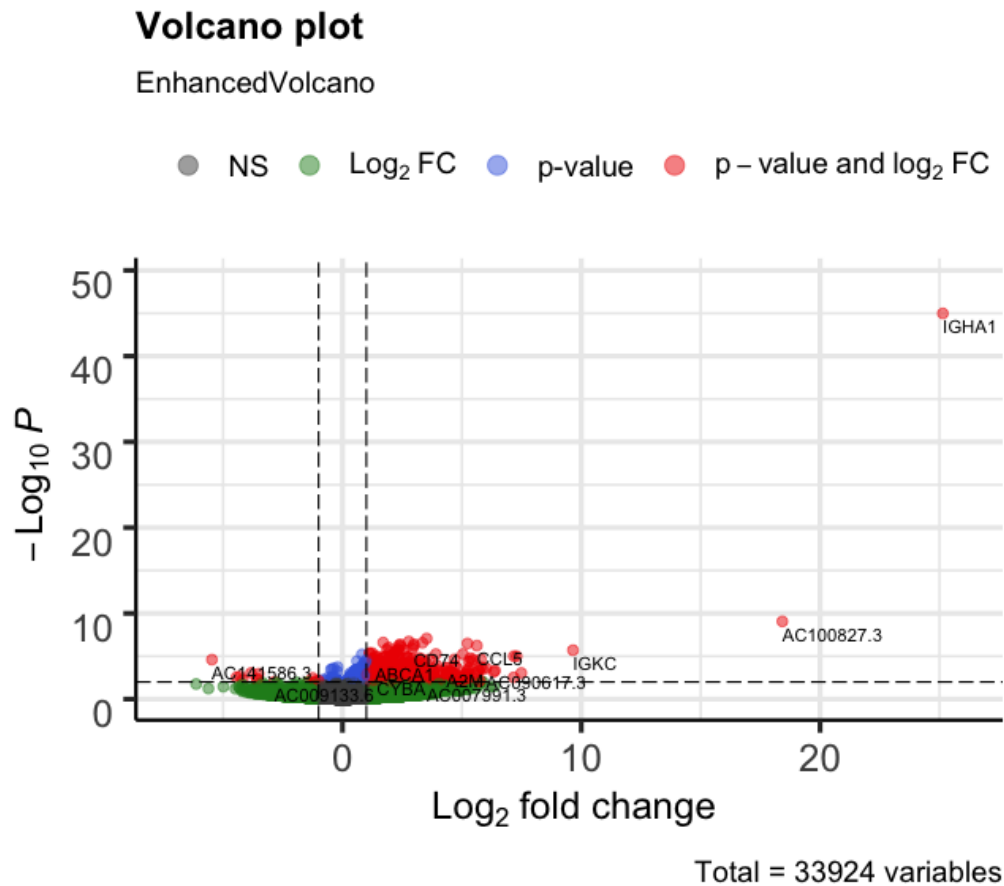


Figure 3.12: Human genes differentially expressed in urine between NMDS cluster B vs cluster A. Samples with positive log2fold change values indicate an increased expression in cluster B samples compared to cluster A samples. The horizontal line indicates $p=0.01$, the vertical lines indicate an absolute log2fold change greater than 1.

3.2.6 Pathway Enrichment Analysis of Differentially Expressed Genes Between Clusters Suggests Immunological Differences

We have identified that many differentially expressed genes that are upregulated in NMDS cluster B have an immunological function. To obtain a more comprehensive picture of the pathways involved in human cells, significantly enriched terms in GO and KEGG pathways were identified.

GO Biological Processes

523 biological process terms were significantly enriched with up-regulated genes in cluster B ($q < 0.05$, clusterProfiler R package; table 3.1). Many enriched pathways are related to immunological functions. See appendix section C.9 for a more complete description of the significant pathways.

Table 3.1: Gene Ontology - Biological Process - Pathway analysis of differentially expressed human genes in urine RNA sequencing. Samples split by NMDS cluster. Gene IDs and q -values have been removed for display purposes (0.05 maximum threshold as determined by the clusterProfiler package).

Description	GeneRatio	BgRatio
leukocyte proliferation	29/305	298/18670
positive regulation of leukocyte activation	30/305	380/18670
positive regulation of cell activation	30/305	394/18670
lymphocyte proliferation	25/305	272/18670
mononuclear cell proliferation	25/305	274/18670
positive regulation of cell-cell adhesion	24/305	255/18670
positive regulation of leukocyte cell-cell adhesion	22/305	218/18670
leukocyte cell-cell adhesion	27/305	337/18670
immune response-activating cell surface receptor signaling pathway	32/305	473/18670
regulation of leukocyte proliferation	22/305	222/18670

GO Cellular Component Terms

53 cellular component terms were associated with up-regulated genes in cluster B ($q < 0.05$, appendix section C.9.2). The top five most significant terms were:

- MHC Class II protein complex
- MHC Class I protein complex
- Collagen-containing extracellular matrix
- Membrane raft
- Membrane microdomain

GO Molecular Function Terms

17 molecular function terms were associated with up-regulated genes in cluster B ($q < 0.05$, appendix section C.9.3). The top five most significant terms were:

- MHC Class II receptor activity
- Structural constituent of a ribosome
- MHC protein complex binding
- Antigen binding
- Rac GTPase binding

These pathways suggest differential immunological responses between clusters. The increase of Rac GTPase binding is supported by the altered expression of *CD74*, which regulates Rac signalling with MIF (Macrophage Inhibitory Factor). Additionally, the top KEGG pathways (appendix section C.9.4) are related to infection response. The top five enriched pathways (228 total) were:

- *Staphylococcus aureus* infection
- Leishmaniasis
- Ribosome biogenesis in eukaryotes

- Antigen processing and presentation
- Hematopoietic cell lineage

3.2.7 Survival Analysis Between Clusters

69% of participants with cancer in cluster B (9/13) demonstrated a progression event whereas 33% of participants with cancer in cluster A (4/12) demonstrated a progression event. However, there was no evidence for a significant difference in time to progression between the NMDS groups ($p = 0.33$ and 0.21 between clusters for all samples and only cancer samples respectively; log-rank test; figure 3.13).

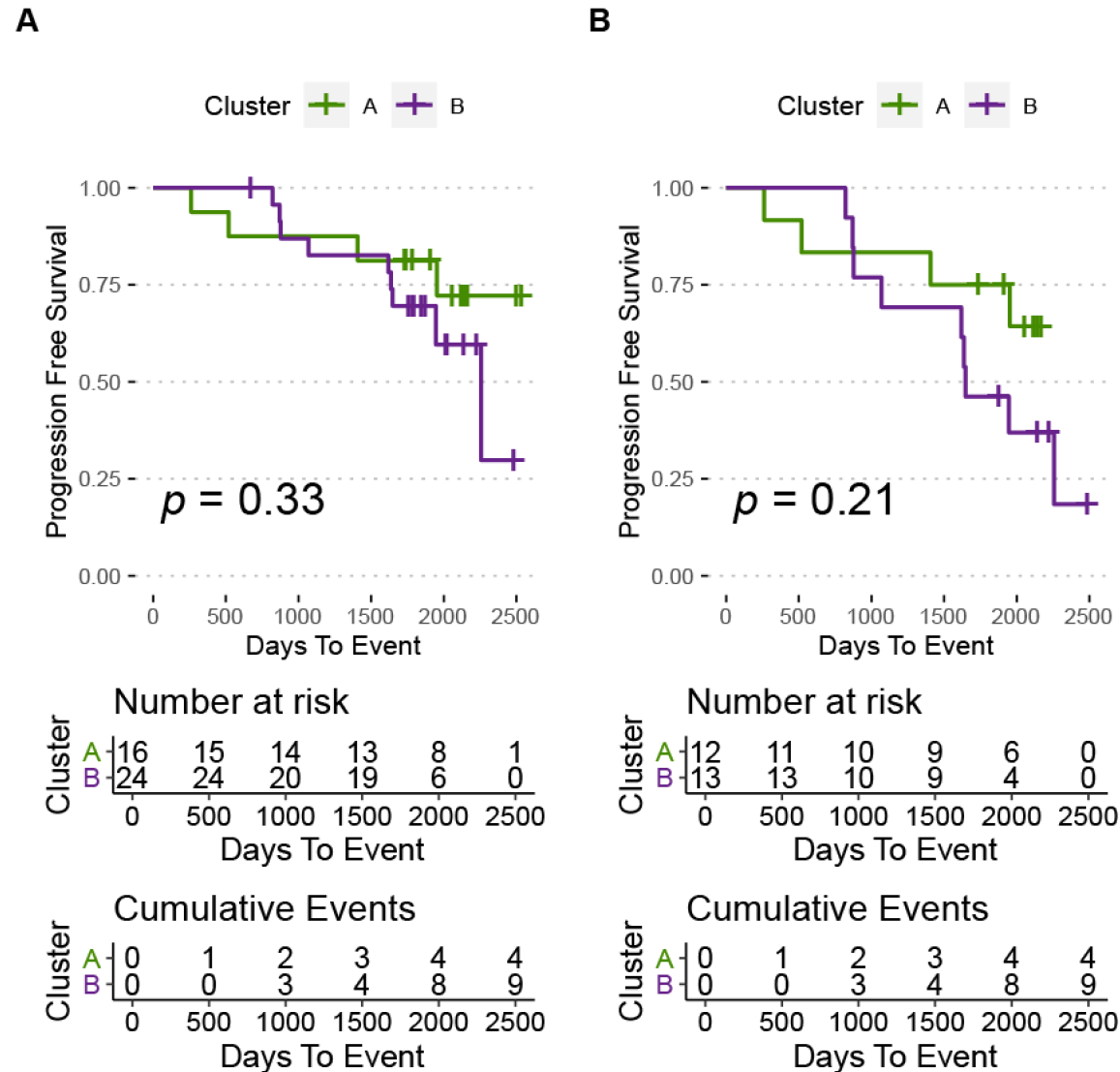


Figure 3.13: Kaplan-Meier plots of time to progression divided by NMDS cluster assignment. For (A) all samples (B). Time to progression was defined as: metastasis, an increase in histological staging, prostate cancer associated deaths, biochemical recurrence or an increase in PSA sufficient to alter the course of clinical treatment. Log-rank test was used to compare survival distributions between groups. NMDS cluster determined from taxa count identified by Kraken on Urine RNAseq data.

3.3 Discussion

3.3.1 Support for Findings in Published Literature

We have shown that there is evidence for microbial DNA in the exosome fraction of urine and that this has an association with progression risk. The successful culture of some of these anaerobic bacteria (Dr Hurst), would indicate that these detected fragments may originate from living bacterial cells.

Previous studies have identified bacteria in the urine of prostate cancer patients^{268,311}. Shrestha *et al.* clustered urine samples based on their bacterial composition and was able to identify a cluster with a higher proportion of cases of cancer. They identified *Streptococcus*, *Actinobaculum*, *Propionimicrobium*, *Varibaculum* and *Anaerococcus* to be differentially abundant in cancer samples compared with samples from patients without cancer. They found no association with PSA and do not report on any associations with metastases or disease progression. I also found that cluster B contained a significantly different presence of *Varibaculum* and *Anaerococcus* in agreement with Shrestha *et al.* (figures 3.11 and C.2).

Feng *et al.* examined metagenomes and metatranscriptomes of 65 tumour and adjacent benign tissue from radical prostatectomy specimens in a Chinese population²⁶⁷. They found some bacteria (*e.g. Escherichia*) as reported here but there was no evidence for the presence of ABBS group bacteria (*Fenollaria*, *Ezakiella*, *Peptoniphilus*, *Porphyromonas*, *Anaerococcus*, *Fusobacterium*). They did not find any significantly differentially abundant bacteria between tumours and matched benign specimens in DNA/RNA, between low and high Gleason scores or an association with metastasis. These differences could be due to the approach used by Feng *et al.*, which may have been insufficiently sensitive to detect ABBS group bacteria or due to natural differences in cohort ethnical composition.

Typically, 24-51% of patients with muscle invasive bladder cancer that undergo radical cystoprostatectomy have some evidence of prostate cancer³¹². Due to this link, the bacteria presented in this chapter may have a role in other urological cancers.

3.3.2 Microbial Presence in the Urine Impacts Immune Host Gene Expression

The expression of a remarkable 417 human genes were found to be significantly differentially expressed between microbial NMDS clusters (cluster A vs B, figure 3.12). Many of these genes are involved in immune responses. IgA is known to be the most predominant immunoglobulin expressed in prostatic fluid and has also been identified to rise in early disease and fall in late stage disease^{313,314}. *IGKC* has been identified as highly expressed in prostate carcinoma but not in “normal” prostate, particularly in African-American men³¹⁵. *CCL5* is expressed by T-lymphocytes but not B-Lymphocytes and is a proinflammatory chemokine³¹⁶. *CCL5* is expressed in prostate cancer cell-lines and activates ERK and Rac proliferative signalling^{317,318}. *CD74* is also known as HLA class II histocompatibility antigen gamma chain and it plays a critical role in antigen presentation and produces a cell surface membrane receptor for the cytokine MIF (Macrophage Migration Inhibitory Factor)). *CD74* is targeted by Milatuzumab, a drug approved for use in multiple myeloma and chronic lymphocytic leukaemia³¹⁹. MIF is released by white blood cells following exposure to bacterial antigens and has been previously shown to be expressed in prostate cancer^{320,321}. Inhibition of MIF attenuates the invasion of prostate cancer cell lines³²². The pathway enrichment analysis also suggests differences in host immunological gene expression.

Given that the difference in clusters is driven by bacterial presence, this suggests the hypothesis that microbial presence impacts host gene expression. The genes involved and pathway enrichment analysis all provide evidence for a host response to infection.

Anti-Tumour Versus Anti-Microbial Antagonism

In evolutionary terms, our ancestors faced a much greater risk with infection than with cancer. Therefore, in the event of an infection, anti-tumour immune response may be dampened in favour of diverting resources to fight infection. This is speculative, but there is some evidence for this associated with myeloid-derived suppressor cells (MDSC).

MDSC are a heterogeneous population of cells that represent a pathologic state of activation of monocytes and neutrophils³²³. The hallmark feature of MDSC cells is to inhibit the ordinary function of T-lymphocytes. Anti-tumour immunosurveillance is believed to be driven by cytotoxic CD8⁺ T-lymphocytes that identify MHC class I and II proteins on nucleated cells and antigen presenting cells respectively³²⁴. Suppression of CD8⁺ T-lymphocytes by MDSC populations represents a common mechanism by which cancerous cells evade the immune response and thereby achieving a hallmark of cancer^{10,325}. In a recent prostate cancer mouse model (characterised by loss of *PTEN* and *SMAD4*), MDSC were shown to be the primary invading immune cell³²⁶. Pertinently, depleting the population of MDSCs limited tumour progression³²⁶.

MDSC populations increase in response to infectious agents at differing body sites including *Mycobacterium tuberculosis*, HIV^{327–329} and the ABBS genera: *Fusobacterium nucleatum*³³⁰ and *Porphyromonas gingivalis*³³¹. There is the possibility that the recruitment of MDSC cells to a tumour site acts as both a cellular response to infection *and* a tumour promoting event.

CCL5 and *CCL2* have been acclaimed as the ‘main chemokines implicated in MDSC migration to tumours’^{332–334}. We found that *CCL5* was the 9th most significantly upregulated gene in NMDS cluster B, which has the greater presence of bacteria (figure 3.12). The next most significantly upregulated genes between clusters is *CD74*, the cell surface receptor for MIF which in addition to its other functions, has been shown to stimulate MDSC production^{335,336}.

3.3.3 Isolate Functional Analysis

The potential functionality of bacteria isolated from urine in the ABBS genera is different from isolated bacteria in the non-ABBS genera (figure 3.8). As all strains were isolated from the urine or prostate secretions, we would expect that they would share common functionality that allows them to adapt to such an environment. It is therefore remarkable that the distinction is observed. 197 pathways/ontology terms were found to be significantly enriched or depleted in the genomes of bacteria in the ABBS genera (figure 3.9). Of particular interest are the pathways enriched in the ABBS group including metalloaminopeptidase activity, biotin biosynthesis and variations in thymidylate synthase activity which will be discussed below.

The ubiquitous pathway of metalloaminopeptidase activity³³⁷ appears exclusive to the ABBS genera. These enzymes are involved in a myriad of biological processes including angiogenesis, hormone processing (including testosterone), DNA synthesis, inflammation and metastasis^{337–344}. Many of these enzymes utilise zinc which is ordinarily found in high concentrations in the prostate^{345,346}. So, a hypothesis could be that the dysfunctional prostate provides a niche for zinc favouring bacteria. Another zinc-related pathway, ‘chromate transport’, is enriched in the ABBS genera. Chromate is a known carcinogen: zinc

chromate induces chromosome instability and DNA double strand breaks³⁴⁷.

We also found “biotin biosynthesis” to be exclusive to ABBS genera. (figure 3.9). Demand and uptake for biotin is high in cancerous cells; there are a wide a range of anti-cancer biotin-drug conjugates³⁴⁸. Biotin mainly acts as a component of enzymes and is involved in protein, lipid and carbohydrate metabolism, thereby potentially influencing cell growth³⁴⁹.

All apart from one strain examined demonstrated evidence of *de novo* biosynthesis of nucleotide bases, however there appears to be a difference in how ABBS versus non-ABBS genera produce nucleotides (figure 3.9). Non-ABBS genera show a preference towards the evolutionarily conserved³⁵⁰ TS (thymidylate synthase) - an orthologue of human TS. ABBS genera on the other hand show a preference for FAD-TS (Flavin Adenine Dinucleotide Dependent Thymidylate Synthase). Many bacteria that use this method are human pathogens³⁵¹.

The increased requirement of tumour cells to use nucleotides as they proliferate renders TS an attractive drug target. 5-fluorouracil (5-FU), a pyrimidine analogue has been used in a range of cancers with varying efficacy since its synthesis in 1957^{352,353}. 5-fluorouracil is metabolised into 5-fluoro-2-deoxyuridine-5-monophosphate (fdUMP) which binds irreversibly to TS³⁵⁴. Upregulation of TS is a well documented resistance mechanism to 5-FU³⁵⁵, so it would seem plausible that FAD-TS, which is structurally distinct to human TS³⁵⁶ could circumvent 5-FU action by failing to entirely deplete the local thymine pool. TS has been suggested as an independent prognostic marker for aggressive prostate cancer³⁵⁷ and gut dysbiosis has been suggested to alter 5-FU efficacy in colorectal cancer³⁵⁸.

Human cells uptake nucleotides and biochemically altered nucleotides as chemotherapeutics: nucleoside analogues³⁵⁹. Therefore, if a bacterial cell were to produce nucleotides and release them to the extracellular matrix, either through efflux or following destruction, then it would be a reasonable assumption that human cells could take these up. Two aspects of FAD-TS that could be investigated further are the production of nucleotide analogues/isomers or the disruption of the human cytosolic nucleotide pool.

3.4 Conclusions and Future Directions

The evidence presented in this chapter supports the existence of microbial DNA in exosome fractions of human urine that are associated with prostate cancer. Successful culture from urine samples suggests that this bacterial DNA originates from living bacterial cells. The relationship between urinary bacteria and aggressive prostate cancer suggests that bacterial constituents in the urogenital tract are associated with or can influence prostate tumours. The fact that these samples are collected at diagnosis supports the latter.

In this chapter, the mechanisms by which microbes might contribute towards the development of aggressive prostate cancer have been speculated upon. This was done by examining the differential expressions of human genes and by analysing functional pathways estimated to be active from the genomes of bacterial isolates.

The interaction between host and bacteria is extraordinarily complex and requires a comprehensive understanding of the molecular biology of both humans and microbes. Here we have hypothesised that the chronic accumulation of bacteria in the vicinity of prostate tissue leads to an inflammatory recruitment of MDSC and a local transition away from

anti-tumour immunity.

By examining the pathways present in (anaerobic) ABBS genera, we have suggested that bacteria might contribute to host cell dysfunction by several mechanisms. We hypothesise that bacterial FAD-TS can contribute towards genome instability either by producing nucleotide analogues or by disrupting the proportions of cytoplasmic nucleotides. As the bacterial load of these genera increases, the amount of genome instability is expected to increase. Therefore, the speed of developing aggressive prostate cancer would be correlated with the bacterial load of ABBS genera. Validating this theory will be complex due to dynamic intra-tissue microbes.

3.4.1 Potential Future Experiments

Potential *in silico* Experiments

Given the small sample sizes of this study, expansion and validation on a larger dataset would be vital for the confirmation of these findings. Although we have detected evidence for bacteria in urine RNA sequencing, additional experiments to identify these genera in direct proximity to prostate cancer cells (either RNA or WGS) would be beneficial. Here we have examined RNA in exosome fractions of urine samples. Although the research group has strong indications that the majority of human RNA comes from the prostate, the same link between microbial RNA has not been made. Therefore, these investigations should be repeated on prostate cancer tissue sequences.

Expansion to a larger dataset would also allow for the identification of a more complete set of microbial genera. Longitudinal studies would help to characterise how dynamic the urine microbiome in prostate cancer is. This would also help validate these bacterial targets as reliable prognostic markers.

Potential *in vitro* Experiments

in vitro experiments are needed to confirm the hypotheses proposing mechanisms of how bacteria impact host cells. We have proposed that the development of aggressive prostate cancer is associated with bacterial constituents of urine. Cell-lines are characteristic of more aggressive tumours and therefore will not be an ideal model of pre-aggressive prostate cancer^{360,361}. It would be interesting to investigate the differential expression produced by exposing cell-lines to lysates of ABBS genera. Obtaining a model of pre-aggressive prostate cancer that encompasses the immune system and other stromal constituents in addition to the complex dynamic of the tissue microbiome is currently impossible. Therefore, FFPE or frozen tissue is the most promising way to illuminate the relationship between prostate cancer and selected bacteria.

MDSC recruitment in prostate cancer tissue should be demonstrated. Wu *et al.* showed that inhibition of interleukin-6 abrogated MDSC recruitment in a murine prostate cancer cell line model which was associated with slower tumour growth and less angiogenesis³⁶². Defining MDSC populations is difficult as they resemble a heterogeneous and phenotypically different set of cells that are myeloid in origin. Wu *et al.* defined MDSC population by Gr1 and CD11b positivity and were isolated with fluorescent activated cell sorting. It would be interesting to investigate MDSC quantity with bacterial presence in prostate cancer FFPE tissue. Furthermore, investigating paired RNA sequencing of urine of these

participants could link intra-tissue bacteria, MDSC populations and urinary bacteria. This could provide support for urinary exosome RNA as a prognostic test or as a method of treatment stratification.

Here we have hypothesised two mechanisms by which FAD-TS might contribute towards genome damage, either by nucleotide analogue production or disrupting local nucleotide pools. More needs to be understood about the basic biochemistry underlying this enzyme. Design of future experiments regarding this topic would be best conducted alongside chemists and structural biologists.

3.4.2 Potential for Clinical Exploitation

The work presented in this chapter suggests a potential clinical use for ABBS genera at the time of diagnosis to predict the aggressiveness of prostate cancer. Detection of these genera could improve current prognostic models. This would be particularly useful as urine samples are used which involve non-invasive sample collection and could limit the use of invasive biopsies. Up to 75% of men with a raised PSA are negative for prostate cancer on biopsy^{363–365}. Therefore, a pre-biopsy prognostic test could have a significant impact on the lives under investigation for prostate cancer. Future adaptations in sample collection may need to be considered to reduce the chance of microbial contamination. It is important to note that the RIDE criteria¹⁸² have not been fully met in this chapter: no negative controls have been sequenced. So the results in this thesis should be considered only in the context of hypothesis generation and not as definitive evidence.

It would be interesting to explore whether there is a casual link between the ABBS genera and cancer. One possible hypothesis is that dysfunctional prostates create a niche for ABBS genera as opposed to the ABBS directly contributing towards disease progression.

Although it seems counter-intuitive, targeting MDSC populations in cancer may prove therapeutically beneficial. It could act to limit inflammation and restore anti-tumour T-cell immunity. Fultang *et al.* 2019 targeted MDSC populations in cancer patients (13 tumour types, including $n=10$ prostate cancers) with Gemtuzumab ozogamicin by binding to CD33³⁶⁶. This drug is approved for acute myeloid leukaemia demonstrating 81% complete response rate (75% in control group)³⁶⁷, but bioavailability or efficacy in the prostate is unknown. Another potential approach is the use of antibiotics. Antibiotics have been used to treat chronic bacterial prostatitis for years³⁶⁸. NICE guidelines recommend antibiotic susceptibility testing. As the ABBS genera are difficult to culture, antibiotic susceptibility testing may not be informative for these genera. Therefore, a thorough investigation needs to be undertaken to determine how these bacteria can be eradicated and if doing so would provide any clinical benefit. A recent meta-analysis suggests that microbiological eradication rates ranged between 40-80% for ciprofloxacin, levofloxacin, azithromycin, doxycycline and clarithromycin but >90% for combination or sequential therapy of both azithromycin and levofloxacin³⁶⁸. Stopiglia *et al.* 2010 investigated 500mg ciprofloxacin twice daily for four weeks upon an elevated PSA result ($N=98$ men)³⁶⁹. They noted no statistically significant reduction in PSA or in the identification of prostate cancer between treatment and control groups³⁶⁹. This could indicate that the treatment regimen is insufficient for microbiological eradication, or that removing bacteria from the prostate has no effect on the tumour. Furthermore, there was no analysis regarding how this impacted the development of aggressive disease specifically. Further examination of ABBS bacteria in the urine could

help to answer these questions and provide a clearer picture of the relationship between ABBS genera and aggressive prostate cancer.

An interesting antibiotic target could be FAD-TS. Although this would eradicate only strains with this enzyme and some bacteria may have both FAD-TS and mammalian TS. This may result in an increase in the proportion of bacteria with mammalian TS. Due to this, demonstrating FAD-TS contributing towards prostate cancer aggression would ideally be required before investigating this as a feasible drug target.

If these hypotheses are proven to be true, the biology of microbial associated prostate cancer may be distinct from non-microbial prostate cancer. There exists the possibility of therapeutic stratification based on the detection of bacteria. Additional work will be required to determine how exactly we can exploit this method of stratification and what treatment options prevail.

3.5 Methods

3.5.1 Urine RNA Exosome Sequencing

RNA was extracted by Martyn Webb from urine of participants according to the protocol and ethics outlined in Webb *et al.* 2020²⁵³. Participants originated from a haematuria clinic at Norfolk and Norwich University Hospitals NHS Foundation Trust (NNUH). The associated metadata was obtained by Dr Kate Manley and the clinical trials data team at NNUH. First-pass urine was collected in 30ml polypropylene tubes and placed on ice. Cell pellets were sedimented (2000g 5 minutes) within 4 hours of sample collection. The supernatant was decanted into a 50ml polypropylene tube (Fisher cat. 35-2070) and placed on ice. Microvesicles were obtained from the urine supernatant as described by Miranda *et al.*, (2010)³⁷⁰. Cellular debris was removed via a 0.8 μ m filter (16592k, Sartorius). The filtrate was then spun in an Amicon Ultra-15 100,000 KDa molecular weight cut-off unit (3400g room temperature 15 minutes) or until the volume was reduced to <500 μ l. 15ml of phosphate buffered saline was added and the sample was respun until the volume was 200 μ l. 700 μ l of RLT buffer (Qiagen Micro RNA RNeasy kit) was added and the RNA extracted as manufacturer's instructions. Average exosomal RNA yields were 340ng. RNA shotgun sequencing was carried out by isolating 100ng of total exosomal RNA which was sequenced by Oxford Gene Technologies using the Illumina TruSeq RNA Sample Prep Kit v2 and sequenced with the Illumina HiSeq 2000 platform and TruSeqv3 chemistry. A mean of 35.7 million 2x100bp paired end reads were sequenced per sample (median 33.8 million, minimum 27.1 million, maximum 94.9 million, IQR 5.5 million). Cohort statistics with and without ABBS detection status are demonstrated in table 3.2.

Table 3.2: Clinical demographics of the Urine RNA sequencing cohort and split by ABBS detection status using Kraken reads. IQR Interquartile range, NEC No evidence of cancer.

Characteristic	ABBS_Pos	ABBS_Neg	Total
N	19	21	40
Age, mean (median; IQR)	69.6 (69; 12.5)	67 (66; 13)	68.3 (67.5; 14.3)
PSA, mean (median; IQR)	11.2 (7.7; 6)	10.3 (4.8, 9)	10.7 (6.9; 9.1)

Clinical Category, n			
Advanced	0	0	0
High Risk (D'amico)	6	6	12
Intermediate Risk (D'amico)	8	4	12
Low Risk (D'amico)	0	0	0
Clinically Benign, NEC	5	11	16
Gleason Score, n			
3+3	0	0	0
3+4	9	4	13
4+3	3	3	6
4+4	2	0	2
4+5	2	4	6
5+4	1	0	1
5+5	0	0	0
Progression/metastasis, n	11	2	13
Leukocytes, + (-; unknown)	8 (6)	4 (10)	12 (16; 12)
Nitrates, + (-; unknown)	0 (14)	0 (14)	0 (28; 12)

3.5.2 RNA Sequencing Bioinformatics

Sequencing data was pre-processed into aligned BAM files by Professor Daniel Brewer. Low quality bases ($q < 30$) were removed from read ends and the TruSeq indexed adapter and TruSeq universal adapter were removed using Cutadapt v1.3 and reads < 32 bp were removed. Reads with more than 5% of Ns or those with low complexity (dust method with maximum score of 10) were filtered using Prinseq 0.20.3. Ribosomal RNA was filtered out using sortmerna V1.89 using the following databases: 5s and 5.8s RFAM database, SILVA SSU Ref NR V111 23s [BAC, ARC] and 23s, 26s, 28s [EUK], SILVA LSU Ref NR V111 16s, 26s, 28s [BAC, ARC] and 18s [EUK]. Filtered reads were aligned to GRCh37 75 human genome (Ensembl) using Tophat (version 2.0.9) and Bowtie (version 2.10) using an expected inner distance between mate pairs of 50 and the Ensembl GRCh37 75 GTF file.

SEPATH¹⁹⁰ was used with standard parameters to obtain Kraken reports for evidence of microbial constituents of urine RNA (as described in chapter 2). Community normalisation including presence-absence and log normalisation was carried out with the `decostand` function from the `Vegan` package³⁷¹. NMDS was carried out with the `Vegan` `metaMDS` function with parameters: `k=2`, `distance='jaccard'`, `try=1000` and `trymax=10000`. PCA was carried out with `prcomp` function with parameters `center=TRUE` and `scale.=TRUE`. NMDS clustering was determined by producing an 'elbow' plot using the `kmeans` function with the number of centers in the range 1 to 10. This suggested two clusters as the most optimal separation of the data. Heatmaps were created with the `Heatmap` function from the `ComplexHeatmap` package³⁷². Survival plots were created with the `survfit` and the `ggsurvplot` functions^{373–375}. Differences between distributions of survival curves were calculated using the log-rank test. The following were used as markers of progression: metastasis, an increase in histological staging, prostate cancer associated deaths, biochemical recur-

rence or an increase in PSA sufficient to alter the course of clinical treatment. Hypothesis testing was carried out with the `wilcox.test` and the `fisher.test` functions³⁷⁶.

Kallisto³⁷⁷ was applied on paired-end quality trimmed sequencing files using an index built by concatenating cDNA and ncRNA sequences for human reference genome 38. Information on generating the Kallisto index can be found in appendix section C.12. Transcript counts were summed by gene and differential expression were performed using DESeq2³⁷⁸ which implements Wald hypothesis testing between groups. Subsequent pathway analysis was carried out using the `enrichGO` and `enrichKEGG` function of the `clusterProfiler` package³⁷⁹.

3.5.3 Isolate Sequencing and Assembly

Isolates were cultured by Dr Rachel Hurst of the University of East Anglia and Dr Emma Meader of NNUH which included anaerobic culturing. Isolates were sequenced by Dr Rachel Hurst. DNA was extracted using Qiagen QIAamp columns and prepared for sequencing with Nextera XT library preparation kit. Sequencing was carried out using Illumina MiSEQ with V3 reagents. The reads produced were 2x300bp. MinION nanopore sequencing was additionally carried out for a select few isolates to achieve higher quality hybrid assemblies. Quality control of this data and assembly was carried out by Dr Ghanasyam Rallapalli using SPAdes (v3.9.034)²⁰⁷ and Unicycler³⁰⁴. Summary statistics for each isolate are available in appendix section C.11.

3.5.4 Prediction of Functional Pathways

Functional analysis of isolates was carried out with InterProScan¹⁶⁴ after converting assembled contigs to protein sequences with Prokka³⁸⁰. The pipeline developed for this purpose is available in appendix section C.13. Prokka was run within a singularity image to obtain protein annotations of each isolate with: `SINGULARITYENV LC_ALL=C singularity exec --cleanenv bin/prokka.sif prokka {input.genome_in} --outdir prokka/{wildcards.sample} --prefix {wildcards.sample} --quiet --force gzip {input.genome_in}`. InterProScan was launched on Prokka output files with: `singularity exec bin/sepathway.sif interproscan.sh -f json --goterms --iprlookup --pathways -T temp -i {input.protein_in} -o {output.ipscn_out}`. Output files were cleaned with `grep` searches and a custom python script. Results within each ontology/framework were collated for all interrogated strains with a custom R script available in C.13.

R packages and versions used at the time of rendering along with system information can be found in appendix section A.1. All other analysis scripts are available on the repository produced for this thesis. All parameters should be assumed to be default settings unless otherwise stated. All package versions unless otherwise stated are available in appendix section A.1.

Chapter 4

Applying SEPATH to Cancer Whole Genome Sequences of Genomics England's 100,000 Genomes Project

4.1 100,000 Genomes Project Overview

The 100,000 Genomes Project was announced in December 2012 with the aim to kick start the transformation of the National Health Service (NHS) into a personalised service of genomic medicine. The department of health established Genomics England as a limited company in order to facilitate prompt delivery of the project. The main objective was to carry out whole genome sequencing (WGS) on 100,000 genomes from patients with cancer (somatic and germline, figure 4.1) and rare diseases (germline only). This target was hit by the end of 2018³⁸¹.

The specific aims of the 100,000 Genomes Project³⁸² are to:

- Identify novel driver mutations in cancer
- Understand evolution of cancer through primary and secondary malignant disease
- Develop cancer therapy treatment stratification based on WGS data and deliver to NHS patients
- Utilise multi-omic approaches to offer additional insight including epigenetics, transcriptomics and proteomics
- Identify novel pathways for cancer therapeutics using WGS

One of the main strengths of the 100,000 Genomes Project over other large-scale sequencing projects lies in its size (figure 4.1). Cancer is a highly complex disease that collectively constitutes over 100 distinct diseases, with seemingly countless numbers of gene mutations (discussed in 1.1.2), that each have varying degrees of impact^{9,383}. Identifying clinically significant mutations is a difficult challenge. Large sample sizes maximise the possibility of finding a clinically significant difference within a population. Lawrence *et al.* 2014 estimated that power saturation in detecting cancer driver mutations is achieved from 600 to 5,000 samples, depending on the type of tumour and the background mutational frequency³⁸⁴.

The 100,000 genomes project has a high target genomic coverage of samples (75X for tumour and 30X for germline samples) that should provide 99% sensitivity for detecting single nucleotide mutations. We hope that these parameters confer sensitivity to detecting and accurately classifying any microbes present. Illumina sequencing was selected as the platform of choice for the 100,000 Genomes Project, because at the start of the project Illumina sequencing technologies were the most widely used, had the highest throughput with a reasonable error profile and presented with the lowest cost of sequencing per base³⁸⁵.

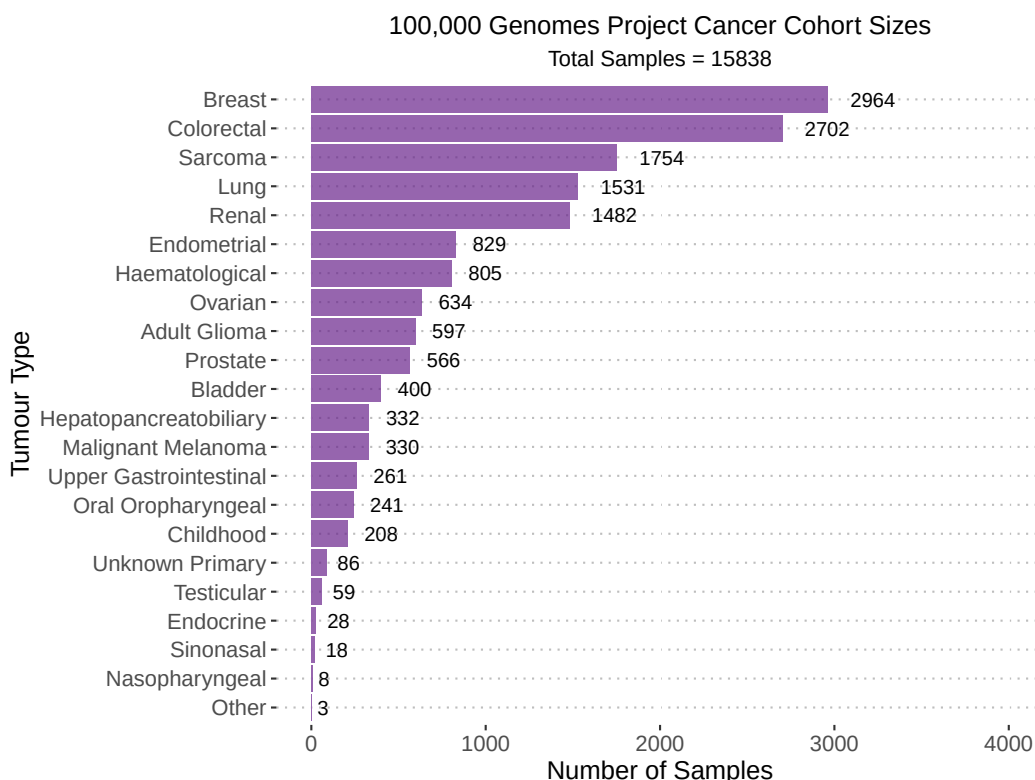


Figure 4.1: Summary of cancer types in the 100,000 Genomes Project. These numbers are based on data release version 9 and resemble only tumour samples and not germline.

This chapter will investigate the microbial constituents in tumour sequences with SEPATH¹⁹⁰ (figure 4.2).

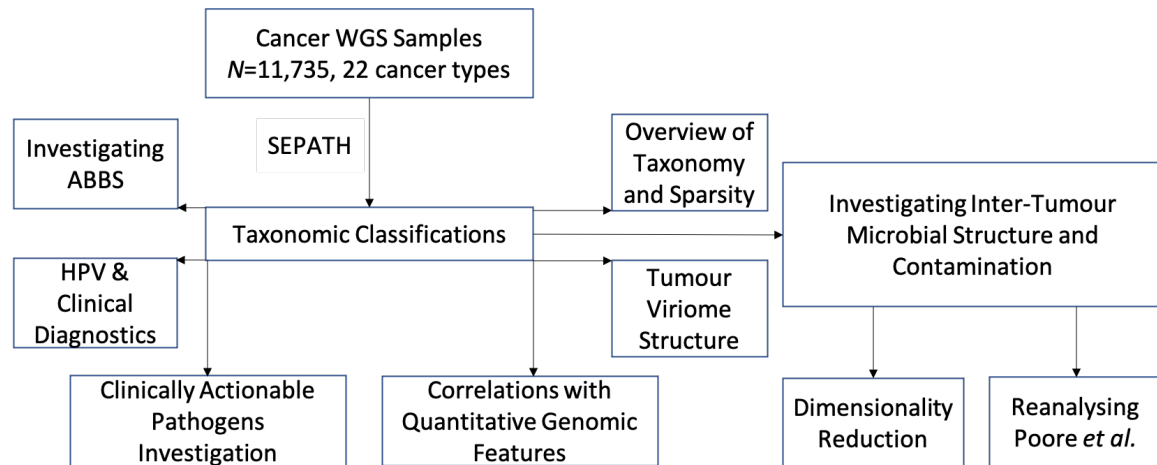


Figure 4.2: An overview of the analyses performed in this chapter. SEPATH was applied on Genomics England 100,000 Genomes Project cancer whole genome sequencing (WGS) data. The following are investigated: ABBS genera (Anaerobic Bacterial Biomarker Set), HPV diagnostics, clinically actionable microbes and pancancer microbial structure.

4.1.1 100,000 Genomes Project Ethical Considerations

The entirety of this work was carried out within the Genomics England research environment due to ethical concerns. All the data presented in this thesis from work on Genomics England data (including graphs, code and datasets) has been submitted to a data ‘airlock’. Data submitted to the airlock is subject to review by the Airlock Review Manager and additionally by an Airlock Review Committee before release.

Diagnosis of Infectious Disease from Whole Genome Sequence Data

Infectious disease is diagnosed based on clinical investigations that involve the evaluation of a symptomatic patient and laboratory testing including microbiological culture. The data presented in the previous chapter demonstrates that it is possible to identify microbial agents (*e.g.*, *H. pylori*) from whole genome sequence data. This gives rise to ethical concerns regarding incidental pathogen detection. As reported in Magiorkinis *et al.* 2019³⁸⁶, the considerations that are high priority to examine include: Evaluation of computational criteria for diagnosis, the actionability of pathogens detected and consequently, the microbiological findings that warrant reporting and those that do not. Four pathogens were suggested by Magiorkinis *et al.* for further clinical evaluation: HIV, HBV, HCV and Human T-Lymphotropic Virus (HTLV). The presence of these pathogens will be evaluated later in this chapter.

Anonymity in Metadata

The 100,000 Genomes Project contains a very large amount of sequencing data and associated metadata. It is theoretically possible to identify patients from their genome code³⁸⁷. It is my opinion that while open data is admirable and key to reproducibility, we must ensure the participants privacy. It is important to consider the anonymity of participants

following the release or publication of data.

Consider the prior knowledge of some potentially unconcealable metadata regarding a participant such as: the anatomical site of the tumour, gender, decade of birth, and general geographical location of a participant (*i.e.*, the NHS trust that a sample was collected at such as 'Northern Ireland'). The more of these metadata variables that are released, the more participants could be uniquely matched to their results. Therefore, much of the raw sample level metadata will remain inside Genomics England research environment.

4.2 Comparable Published Literature

A study by Poore and Kopylova *et al.* 2020¹⁵⁴ entitled: "*Microbiome analyses of blood and tissues suggest cancer diagnostic approach*" was recently published that used a similar approach to SEPATH on whole-genome and whole-transcriptome data from cancer samples within the TCGA (The Cancer Genome Atlas). This approach took unmapped reads from human-aligned reads (BAM files) and classified them using a Kraken database consisting of 59,974 microbial genomes. They report that each tumour type is characterised by specific genera and they were able to predict tumour type from microbial read assignments with an astonishingly high performance. The validity of these claims will be examined in this chapter.

4.3 Chapter Overview

This chapter will detail the search for microbial DNA in cancer by applying SEPATH to cancer whole genome sequence samples within Genomics England's 100,000 Genomes Project. I will evaluate microbial DNA over a variety of tumour types ($N=19$) and compare this to the available literature. I compare Alphapapillomavirus classifications to clinical diagnostic test results and investigate *TP53* within these samples. The potential of SEPATH to identify clinically actionable infectious diseases will be investigated and commented upon. Viral constituents of tumour samples are investigated and compared to the available literature. The impact of contamination within this dataset overall is discussed.

4.4 Results

SEPATH was run on ($N=11,752$) whole genome sequencing tumour samples from the 100,000 genomes project using mOTUs2 and Kraken to classify human depleted sequencing reads. 11,718 shared samples were reported by both Kraken and mOTUS2 including 539 FFPE (Formalin-Fixed Paraffin-Embedded) and 771 PCR (Polymerase Chain Reaction) amplified samples. Metagenomic assembly was trialled on human depleted reads from individual samples but often failed due to insufficient data and so was stopped.

4.4.1 mOTUS2 is Insensitive on Genomics England Data

Kraken detected 1,617 distinct genera across the entire dataset (no read threshold) whereas mOTUs detected 359. 291 genera were shared between both approaches. mOTUs2 reported 68 genera not in Kraken and Kraken reported 1,326 genera that mOTUS2 did not. There are minor differences in taxonomy that limits a perfect comparison between the two tools

appendix D.1. mOTUs often reports ‘unknown’ genera within a family (*i.e.*, ‘unknown *Veillonellaceae*’) which cannot be compared with Kraken genus level results. The taxa reported by mOTUS2 and not by Kraken is available in appendix D.1.

After filtering both matrices to this shared list of genera, 81.3% of values in the Kraken read count matrix (genus vs sample) were empty, whereas 98.39% of values in mOTUs2 matrix were empty. PCoA on mOTUs2 results using a variety of methods can be found in section D.2.

There are only a few taxa that explain most of the variation in the mOTUs2 results (*Akkermansia*, *Alistipes*, *Bacteroides*, *Barnesiella*, *Campylobacter*, *Enterobacteriaceae*, *Eubacterium*, *Faecalibacterium*, *Fusobacterium*, *Gemella*, *Parvimonas*, *Prevotella*, *Roseburia*, *Ruminococcus*, *Streptococcus*, *Propionibacterium*, *Afpia*, *Caulobacter* and *Sphingomonas*, discussed in appendix D.1). The sparsity of the mOTUS2 results rendered it untenable to work with and so mOTUS2 results will not be reported upon further.

4.4.2 An Overview of Kraken Results

The mean number of human depleted reads in all samples was 2,172,490 (median: 1,477,788 inter-quartile range IQR: 1,296,431). The mean unclassified reads in each sample was 1,621,879 (median: 1,259,565, IQR: 1,081,231). The overall number of non-human unclassified reads in each file is shown in appendix figure D.3. The mean number of root classified reads in all samples (reads classified anywhere on the tree of life) was 550,611 (median: 177,561, IQR: 166,375).

The median number of root classified reads for all tumour types are as follows: colorectal 505,490, prostate 245,346, oral 245,328, testicular 211,717, glioma 194,675, bladder 181,421, upper gastrointestinal 181,224, haematological 177,157, melanoma 172,515, lung 170,684, unknown 170,487, renal 168,272, hepatopancreatobiliary 159,199, sarcoma 149,018, childhood 134,380, breast 127,410, sinonasal 118,854, endometrial 118,245 and ovarian 108,120. The median root classified reads is significantly higher in colorectal compared to the rest of the dataset ($p < 2.2 \times 10^{-16}$ unpaired Wilcoxon rank sum test, median of 505,490 in colorectal and 154,524 for all other tumour samples, figure 4.3). Although colorectal cancer demonstrates the best evidence for microbial detection in this cohort, another researcher has registered a research project working on this cancer type and so it was not investigated in any additional depth compared to the other cancer types. Generally, sites with known microbiome constituents demonstrated high levels of evidence for classified sequencing reads (*i.e.*, colorectal, oral). Melanoma is one site that contains a known microbiome but has demonstrated little evidence for classified sequencing reads relative to other tumour types.

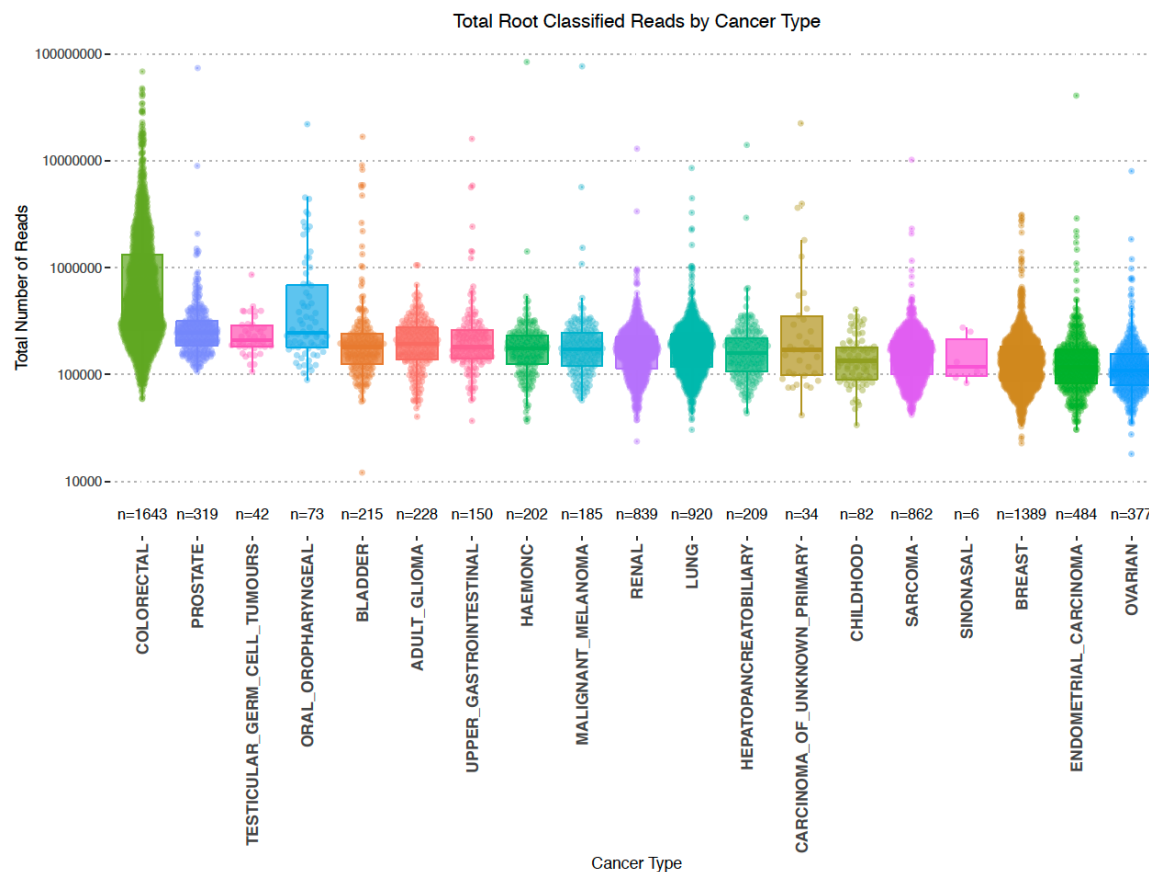


Figure 4.3: Total number of reads classified at any taxonomic level ('root') across whole genome sequences of cancer types within Genomics England's 100,000 Genomes Project. Samples are ordered by the median number of non-human unclassified reads in order from most (colorectal) to least (ovarian) as shown in appendix figure D.3. This analysis includes 361 FFPE samples and 603 PCR samples.

4.4.3 Most Prevalent Kraken Genera

Only sequence data that were generated using a PCR-free library preparation and from fresh frozen tissue were retained for further analysis. Taxa with fewer than 20 assigned reads were set to zero. The genus '*Homo*' is the most prevalent genus being detected in 9,347 of 9,351 samples. The next three most common taxa were *Mycobacterium*, *Toxoplasma* and *Candidatus Pelagibacter* with 9,346, 9,345 and 8,254 positive samples respectively. These taxa are much more prevalent than the next most prevalent taxa (*Achromobacter* in 2,667 samples and *Bacteroides* in 2,619 samples) and may resemble contamination³⁸⁸. These three genera being such a high prevalence would offer very little in determining differences between samples (in terms of presence absence). Because of these reasons, *Mycobacterium*, *Toxoplasma* and *Candidatus Pelagibacter* were removed from further analysis.

The number of positive samples was counted and ranked. There are an interesting variety of genera amongst the top 20 most prevalent (figure 4.4). Whilst *Achromobacter* may resemble a contaminant³⁸⁹, *Bacteroides*, *Enterococcus*, *Prevotella* and others are known to

reside within humans.

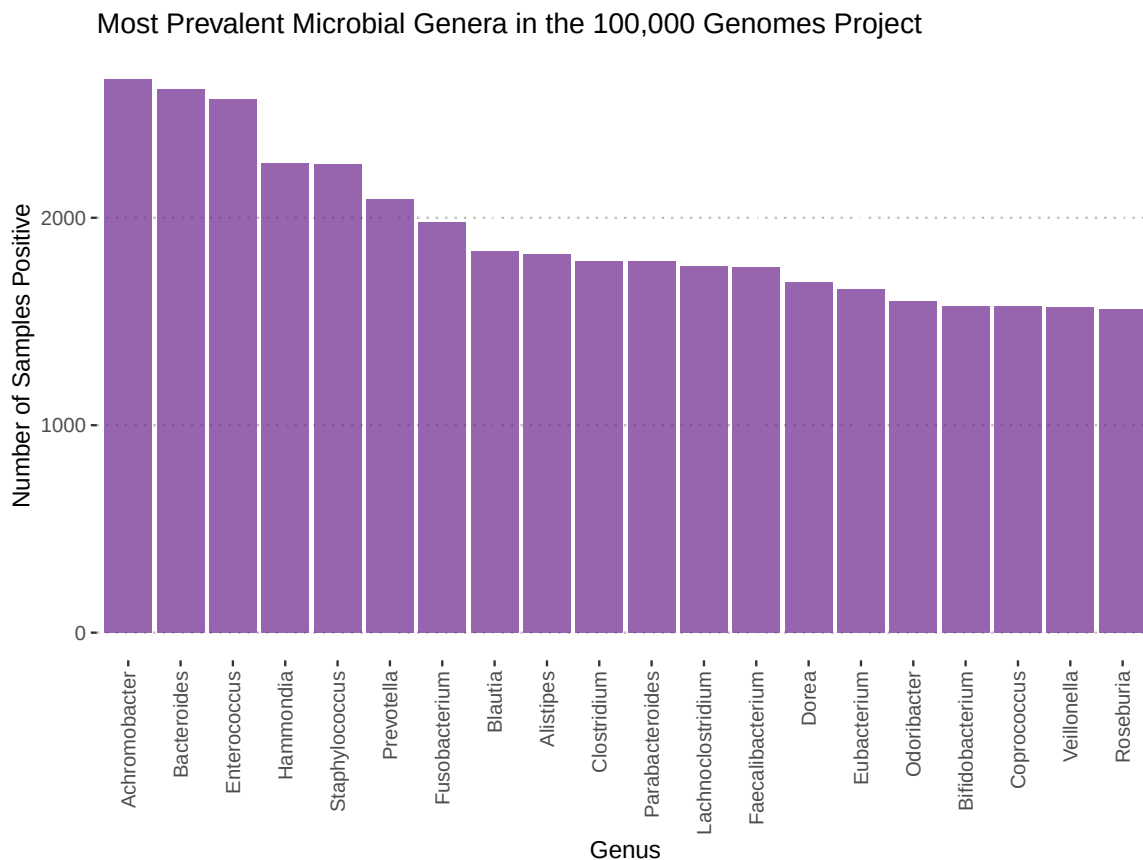


Figure 4.4: The 20 most frequently reported microbial genera across all cancer types in the 100,000 Genomes Project whole genome sequence data after applying SEPATH with Kraken. A minimum of 20 reads was required to determine a positive result. Taxa are ordered with the most prevalent genus (*Achromobacter*) first.

All of the 20 most prevalent taxa have median read values of zero, meaning that for each taxon there are more negative samples than there are positive samples. When considering only positive samples (≥ 20 reads), *Bacteroides* contains the highest median number of classified sequencing reads at 25,544. This is followed by *Fusobacterium* at 4,817, *Alistipes* at 2,553, *Faecalibacterium* at 1,773.5 and *Parabacteroides* at 1,609.

Topmost Prevalent Genera In Each Cancer Type

In this section we will examine the top 20 most prevalent genera detected in tumour tissue in each of the cancer types in the 100,000 genomes project data (figures 4.5 - 4.14).

Sites with known microbiomes generally reflect what we expect. Colorectal cancer is dominated by *Bacteroides*, *Fusobacterium*, *Prevotella*, *Clostridium* and others³⁹⁰. Oral cancer demonstrates *Fusobacterium*, *Prevotella*, *Porphyromonas* as well as *Veillonella*, *Actinomyces* and *Capnocytophaga* in line with other literature³⁹¹. Several parallels between the research in this chapter and the human microbiome project arise such as the dominance

of *Bacteroides* in colorectal samples and the identification of *Prevotella*, *Veillonella* and *Haemophilus* in the oral cavity⁷² (figure 4.13).

The results are sparser in anatomical sites with less well established microbiomes. This may well be because those sites do not host microbes or that the reference genomes used to build the Kraken database are skewed towards well studied microbiome sites. Breast cancer for example, despite having 1,609 samples has much more sparse top genera than colorectal with 1,829 samples. The bacteria present in over 100 breast cancer samples include: *Moraxella*, *Bacteroides*, *Enterococcus*, *Staphylococcus*, *Hammondia* and *Achromobacter*. The brain is considered the most sterile site in the human body. Protected by the blood brain barrier, neurological infections are relatively rare but still possible^{392,393} meaning that finding bacteria in brain tumours cannot be completely ruled out as contamination. The adult glioma results reflect this expected sterility by demonstrating incredibly sparse results with only four taxa present in > 10% of samples: *Staphylococcus*, *Enterococcus*, *Hammondia* and *Achromobacter*. These bacteria are more likely to originate from the skin or kitome.

No cervical cancer exists in the dataset and it is not possible to report on Alphapapillomavirus in this context. However, the diagnostic performance of Alphapapillomavirus detection in oral cancer is investigated later (section 4.4.5). *Helicobacter* has not arisen as one of the most prevalent bacterial genera in upper gastrointestinal cancer being identified in only 19/166 samples (≥ 20 reads). In these 19 positive samples, *Helicobacter* demonstrates a minimum number of reads of 48, 1st quartile 469, median 1,748, mean 8,438, 3rd quartile 8,238 and a maximum of 55,065 which provides relatively strong evidence of detection. This lower than expected prevalence may be because upper gastrointestinal cancers include a range of cancers such as esophageal whereas *Helicobacter* is causative of gastric *non-cardia* carcinoma specifically³⁹⁴. 115 samples with < 20 *Helicobacter* reads were set to zero so it may be that the read threshold is too strict for detection. Of these 115 samples, there is a median of 2 reads, a mean of 4.13 reads and a maximum of 18 reads assigned.

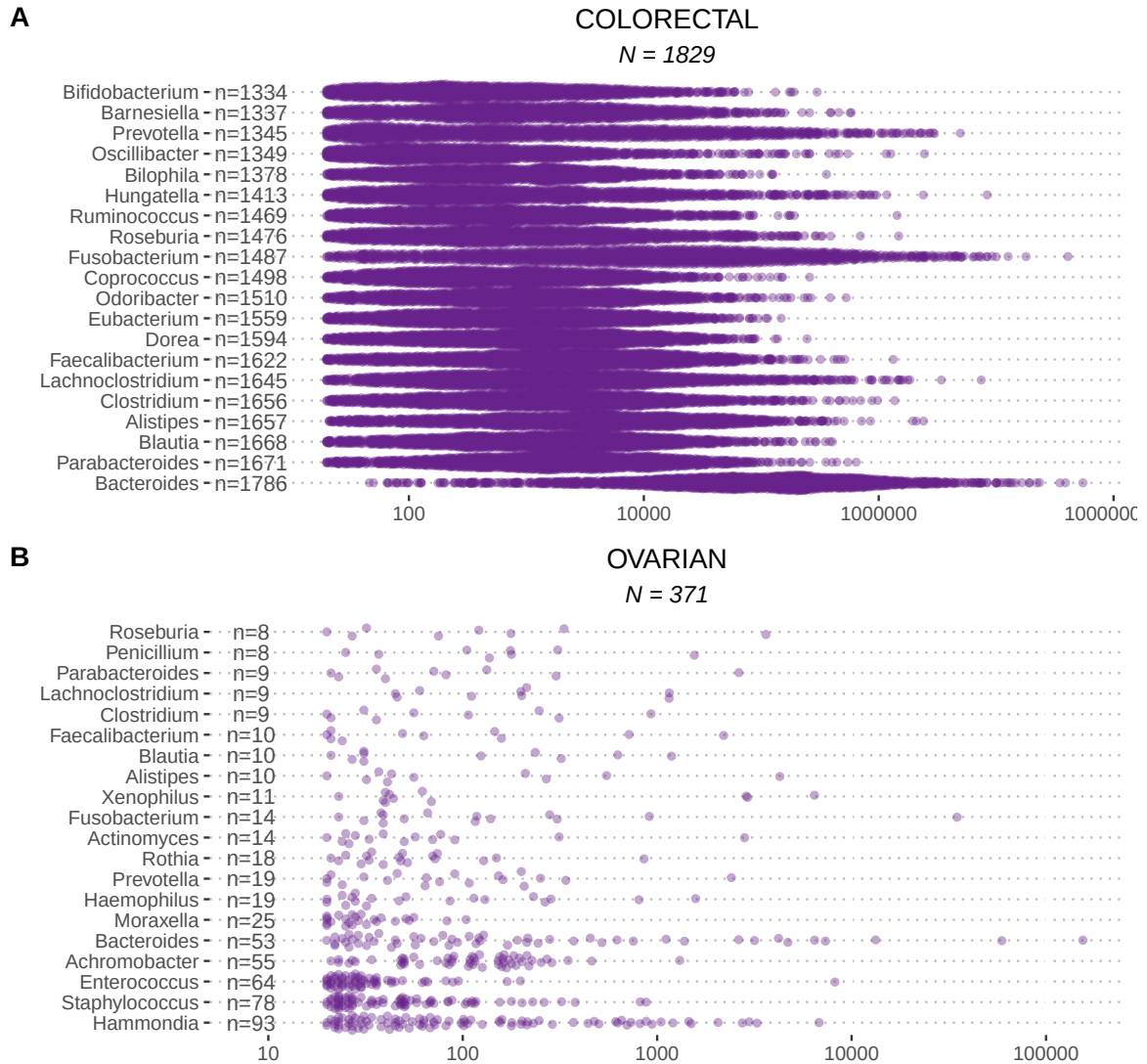


Figure 4.5: Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for **A**) colorectal cancer and **B**) ovarian cancer. The y -axis denotes the number of positive samples.

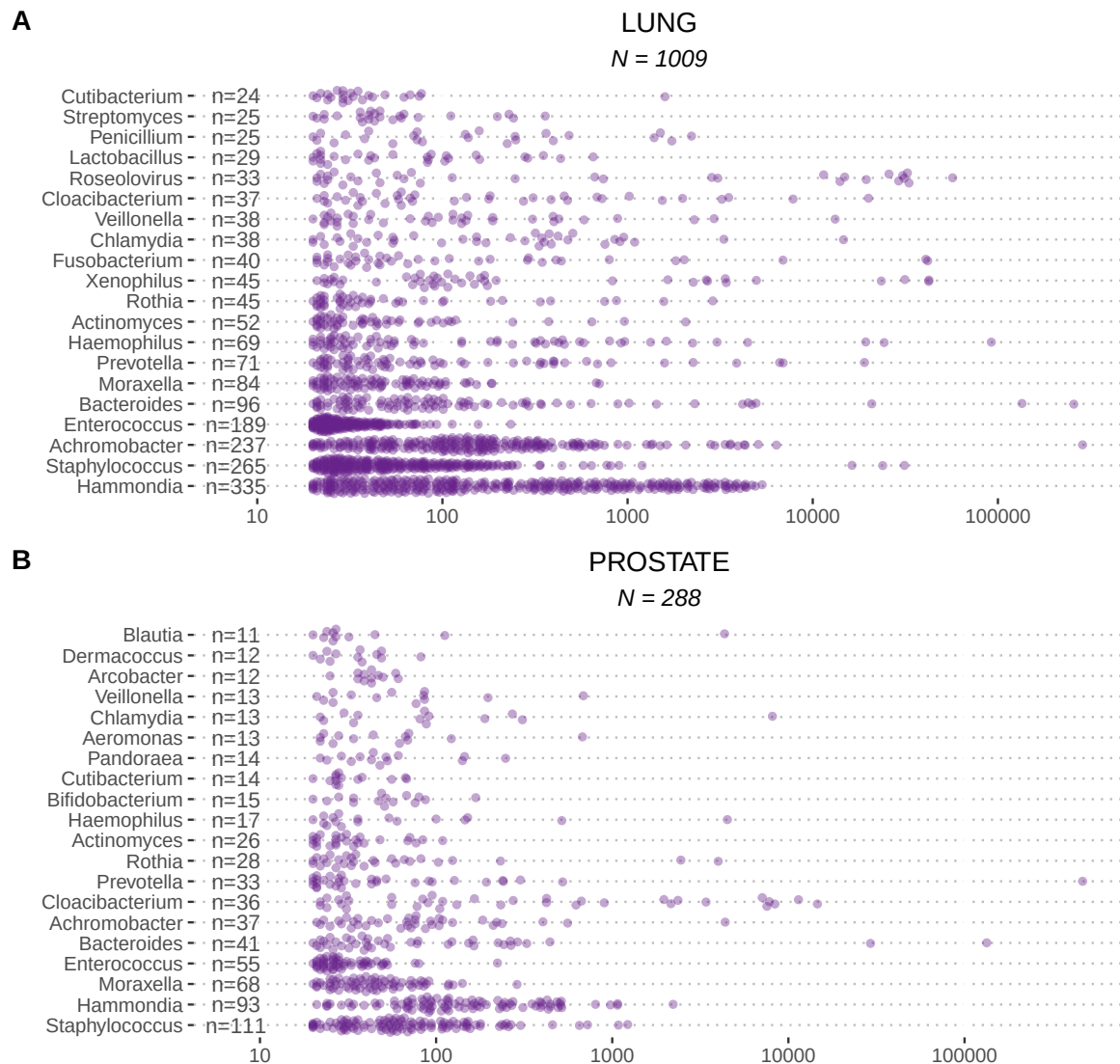


Figure 4.6: Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for **A**) lung cancer and **B**) prostate cancer. The *y*-axis denotes the number of positive samples.

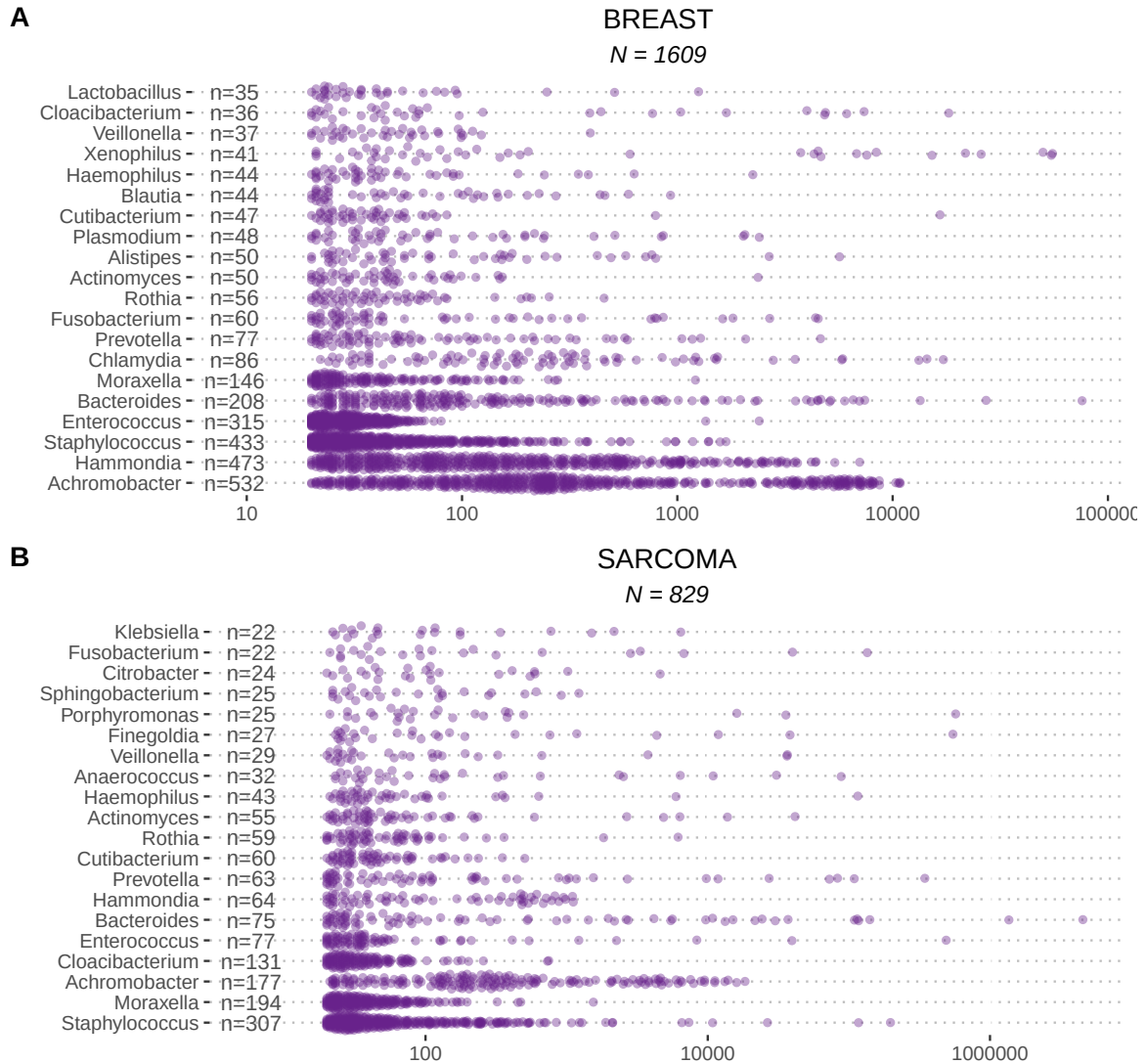


Figure 4.7: Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for **A**) breast cancer and **B**) sarcoma. The y -axis denotes the number of positive samples.

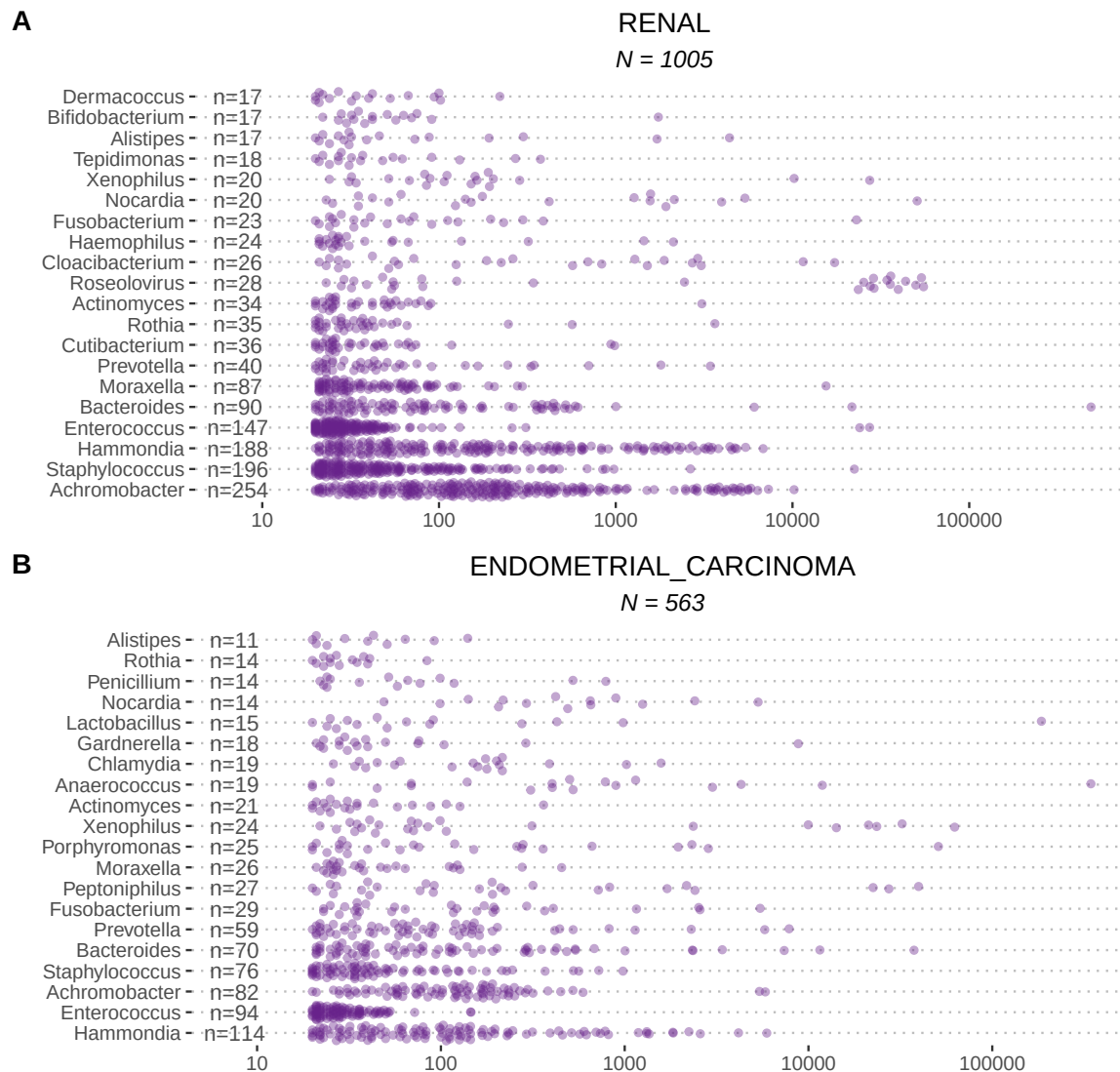


Figure 4.8: Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for **A**) renal cancer and **B**) endometrial cancer. The *y*-axis denotes the number of positive samples.

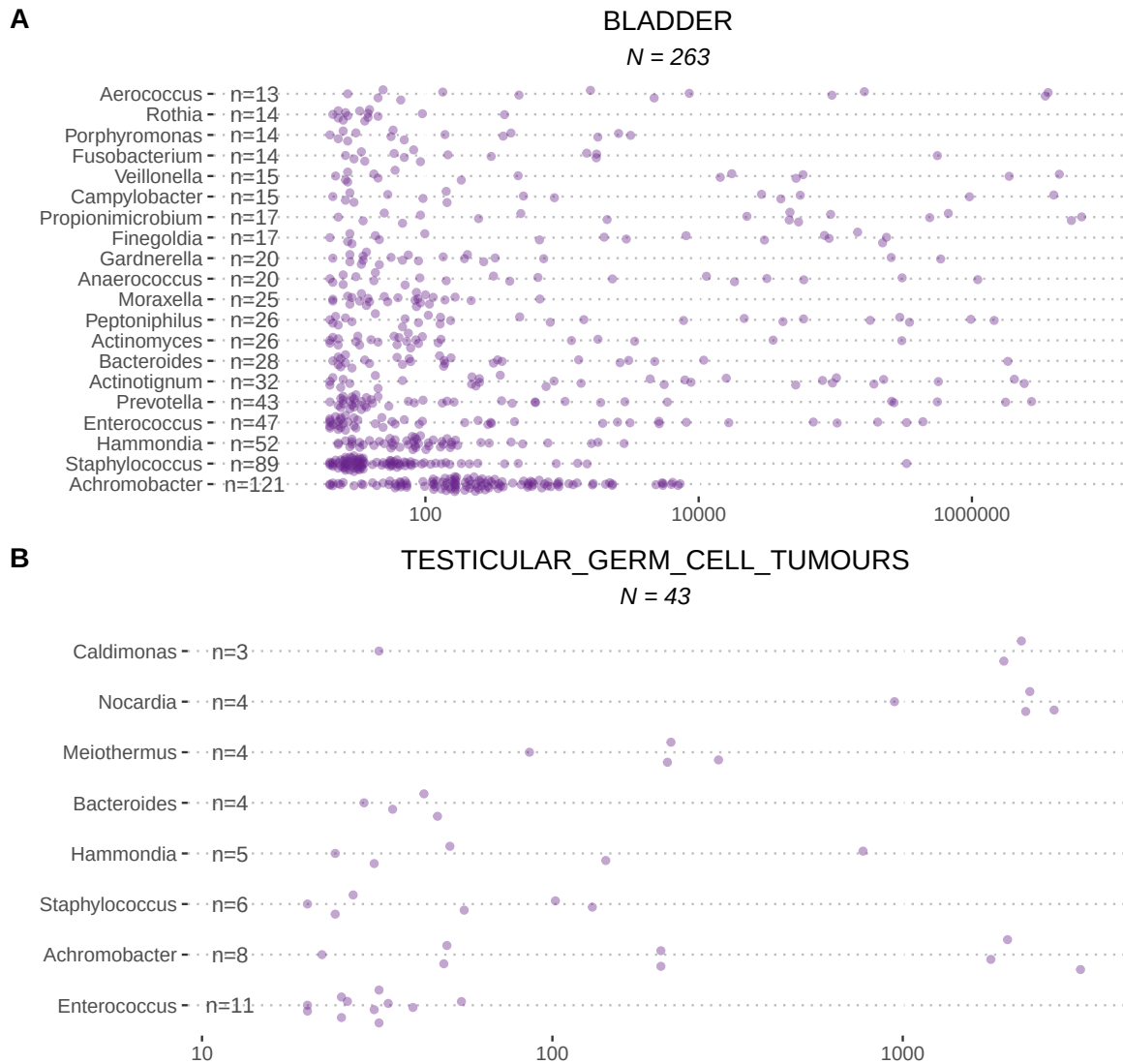


Figure 4.9: Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for **A**) bladder cancer and **B**) testicular cancer. The y -axis denotes the number of positive samples.

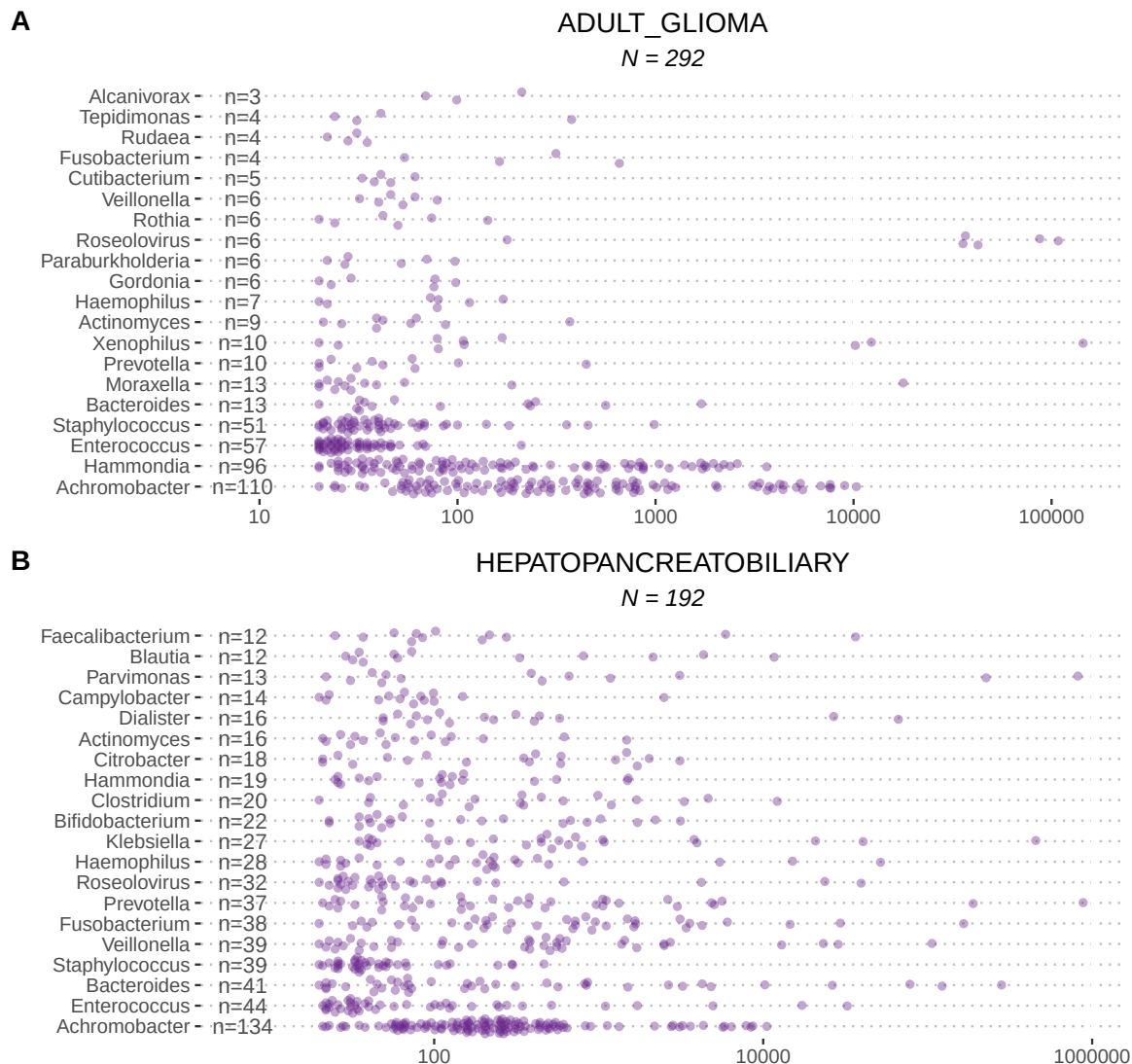


Figure 4.10: Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for **A**) adult glioma and **B**) hepatopancreatobiliary cancer. The y -axis denotes the number of positive samples.

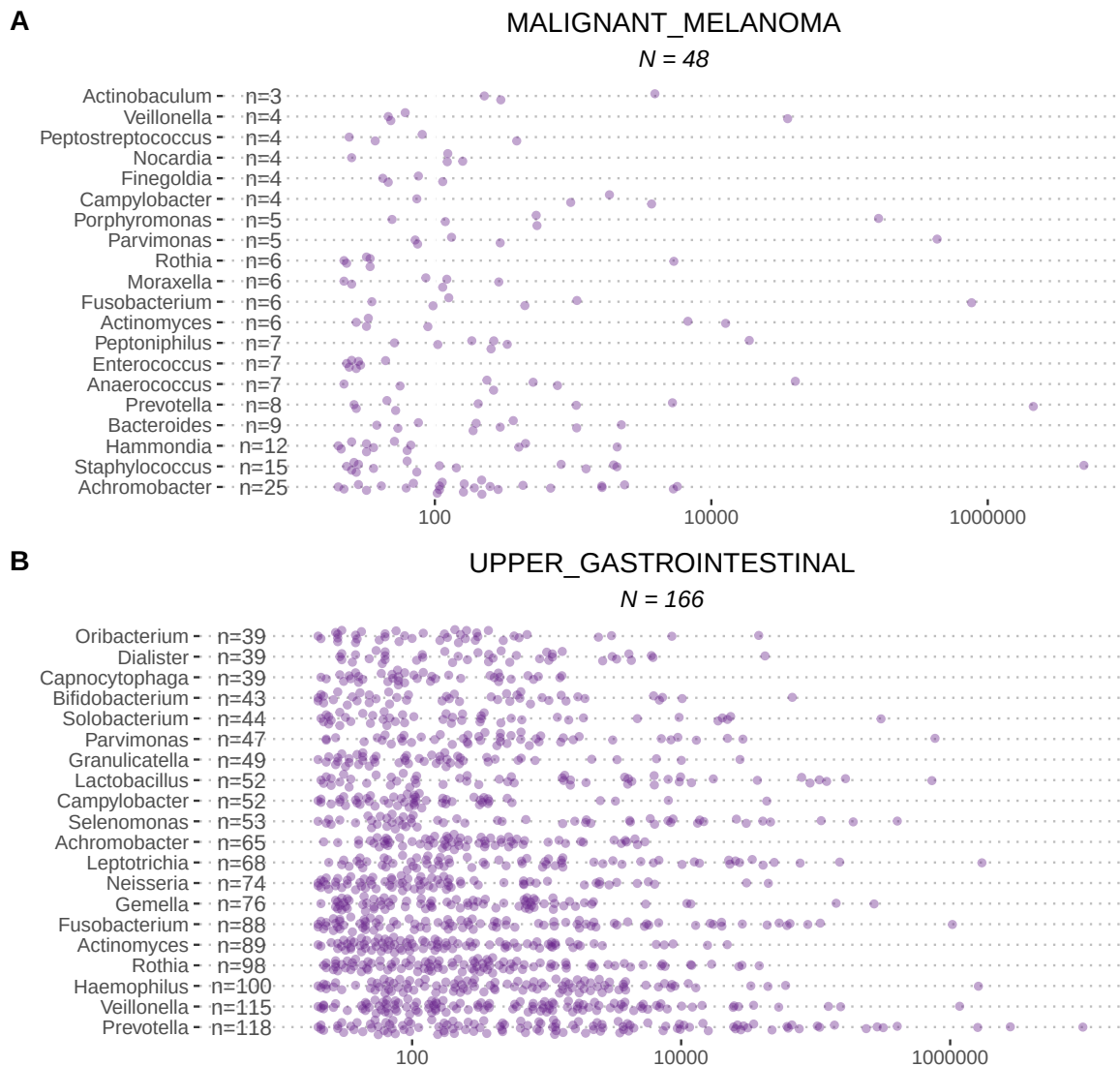


Figure 4.11: Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for **A**) malignant melanoma and **B**) upper gastrointestinal. The y -axis denotes the number of positive samples.

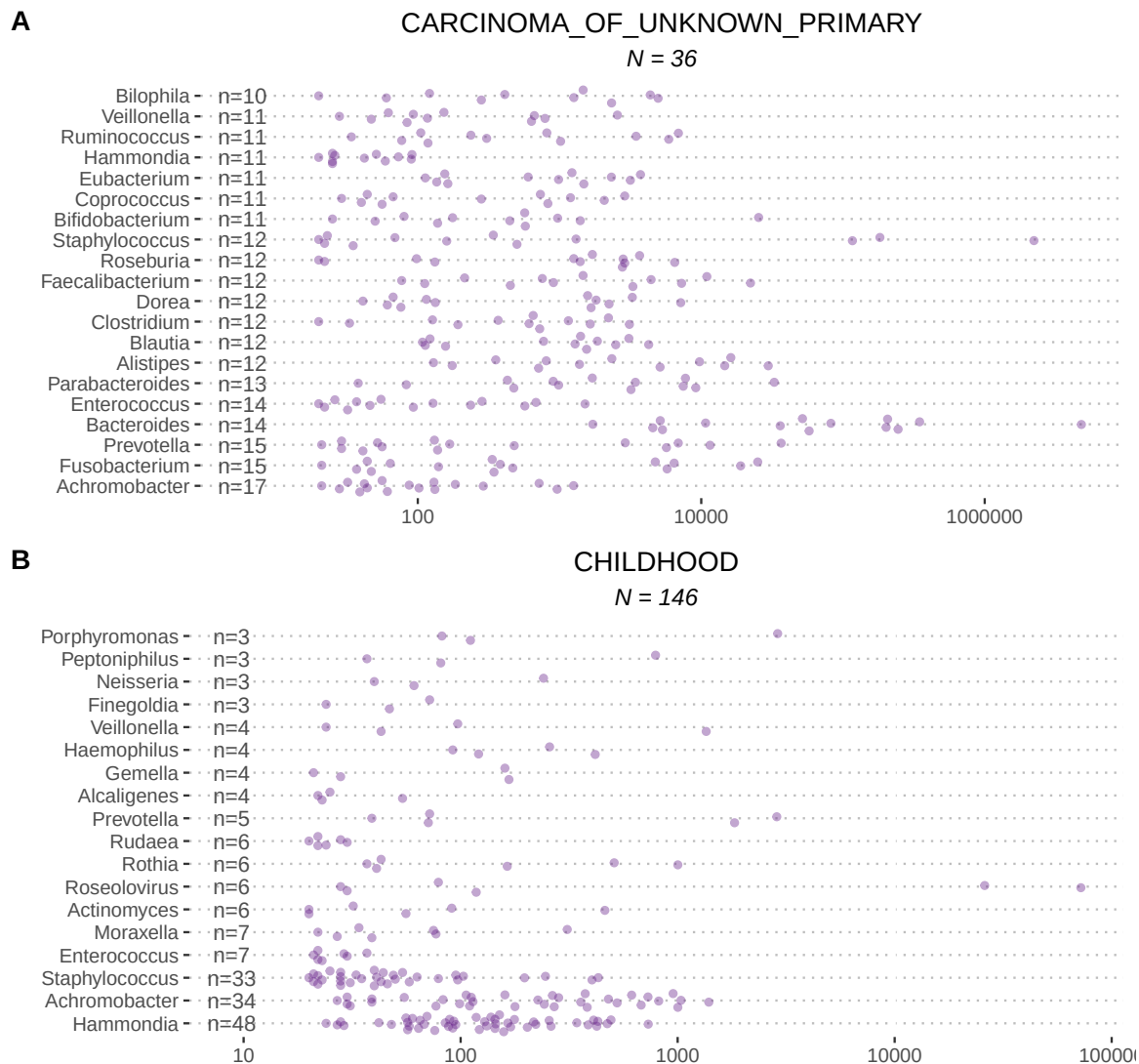


Figure 4.12: Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for **A**) unknown primary and **B**) childhood cancer. The y -axis denotes the number of positive samples.

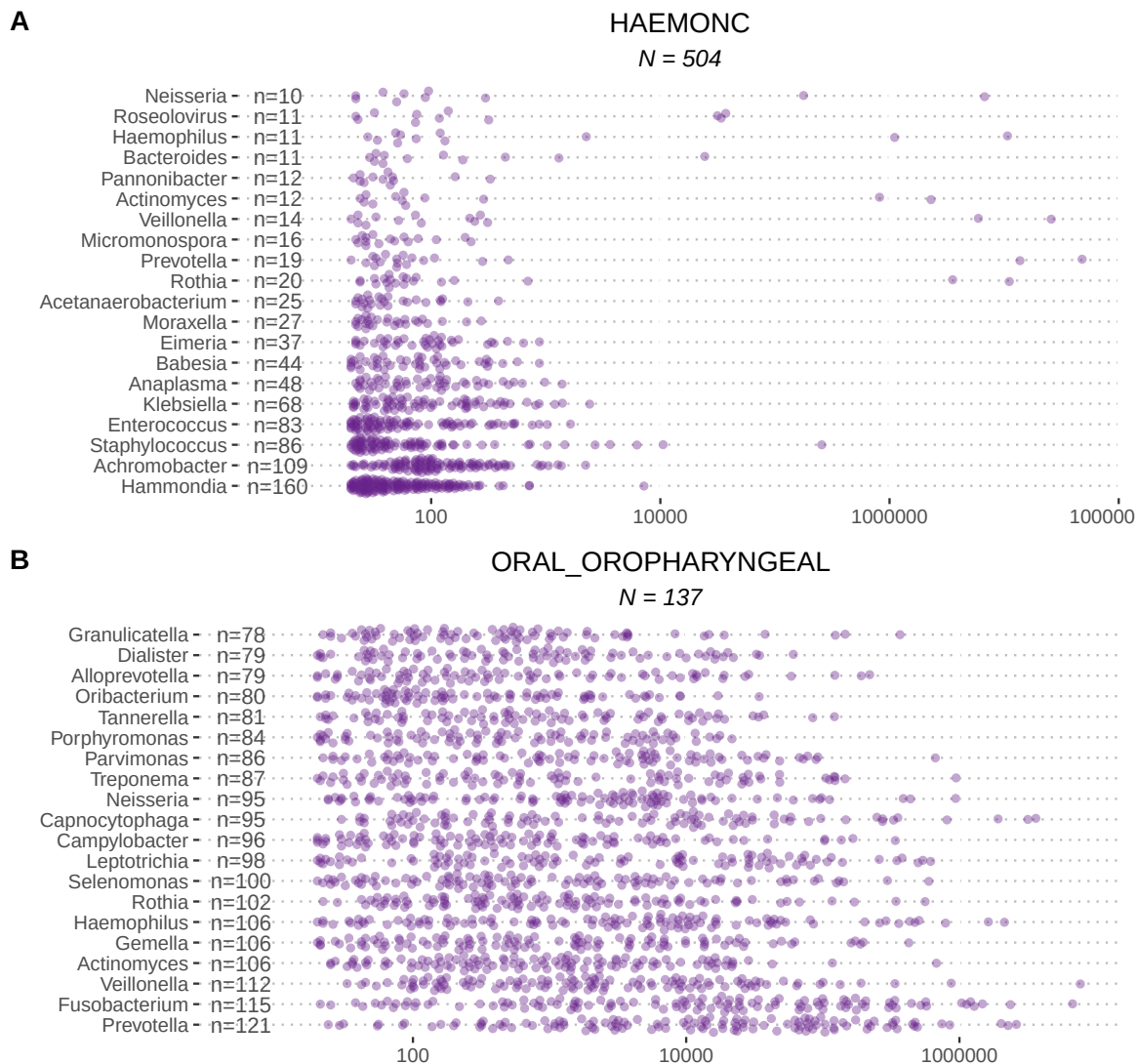


Figure 4.13: Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for **A**) haematological malignancies cancer and **B**) oral and oropharyngeal cancer. The y -axis denotes the number of positive samples.

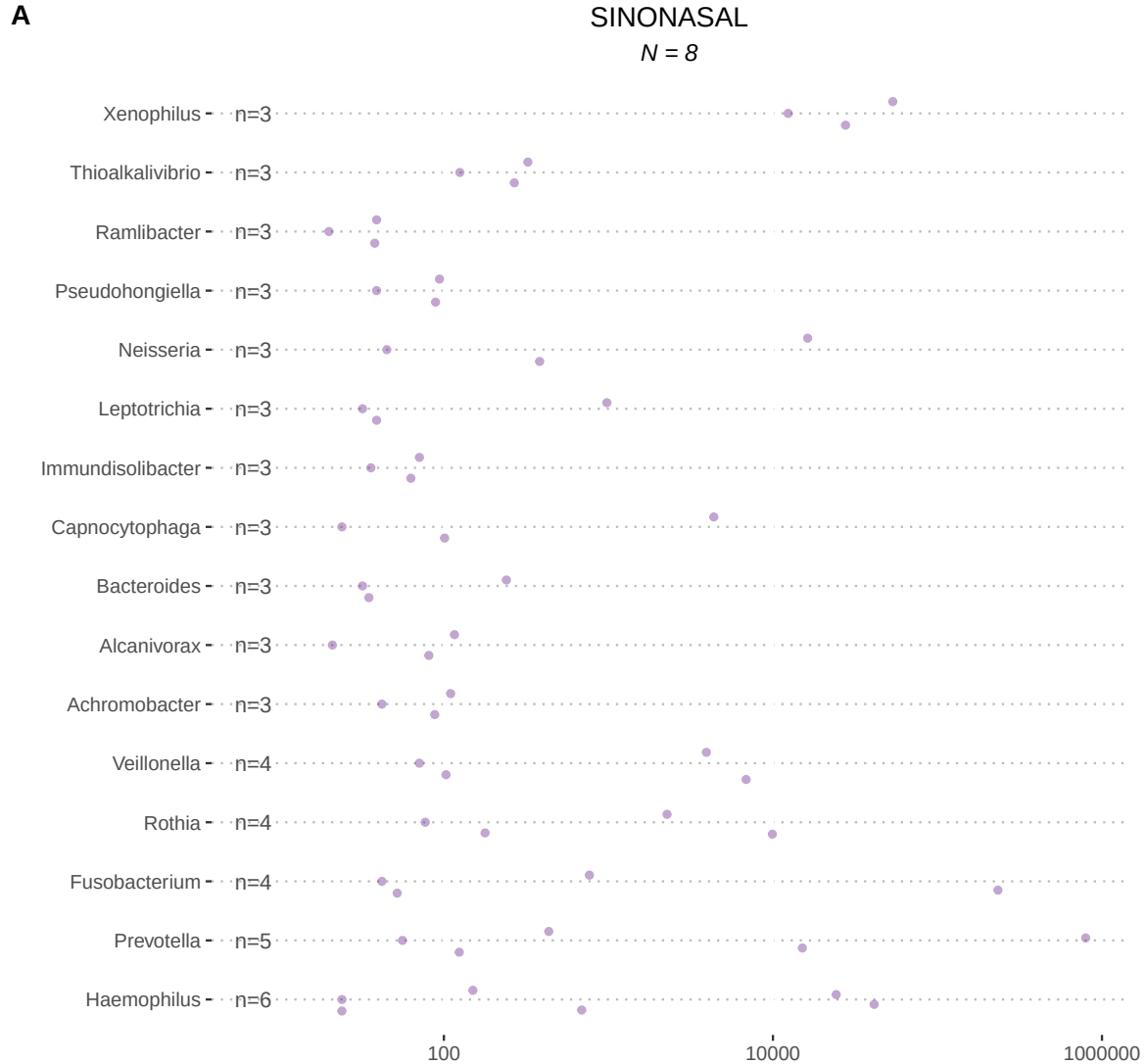


Figure 4.14: Read distributions for the top 20 most prevalent genera detected obtained by running Kraken on whole genome sequence data from Genomics England's 100,000 Genomes Project for sinonasal cancer. The y -axis denotes the number of positive samples.

4.4.4 Sparsity

The sparsity of each cancer type was investigated after setting genera with < 20 reads to zero. Contaminants from Salter *et al.*¹⁷² were removed and additionally *Homo*, *Mycobacterium*, *Toxoplasma* and *Candidatus Pelagibacter* leaving 1,532 genera (figure 4.15). Colorectal was the least sparse cancer type by far (96% of cells in the sample-taxa community matrix were empty). This was followed by Oral at 98%, unknown primary at 98.6%, sinonasal at 98.8%, and upper gastrointestinal at 99%. Glioma demonstrated a sparsity of 99.9%. Three cancer types demonstrated a higher sparsity than Glioma: childhood, endocrine and other. This might indicate that there is a large amount of noise from contamination in these tumour types, or that they have a particularly low diversity.

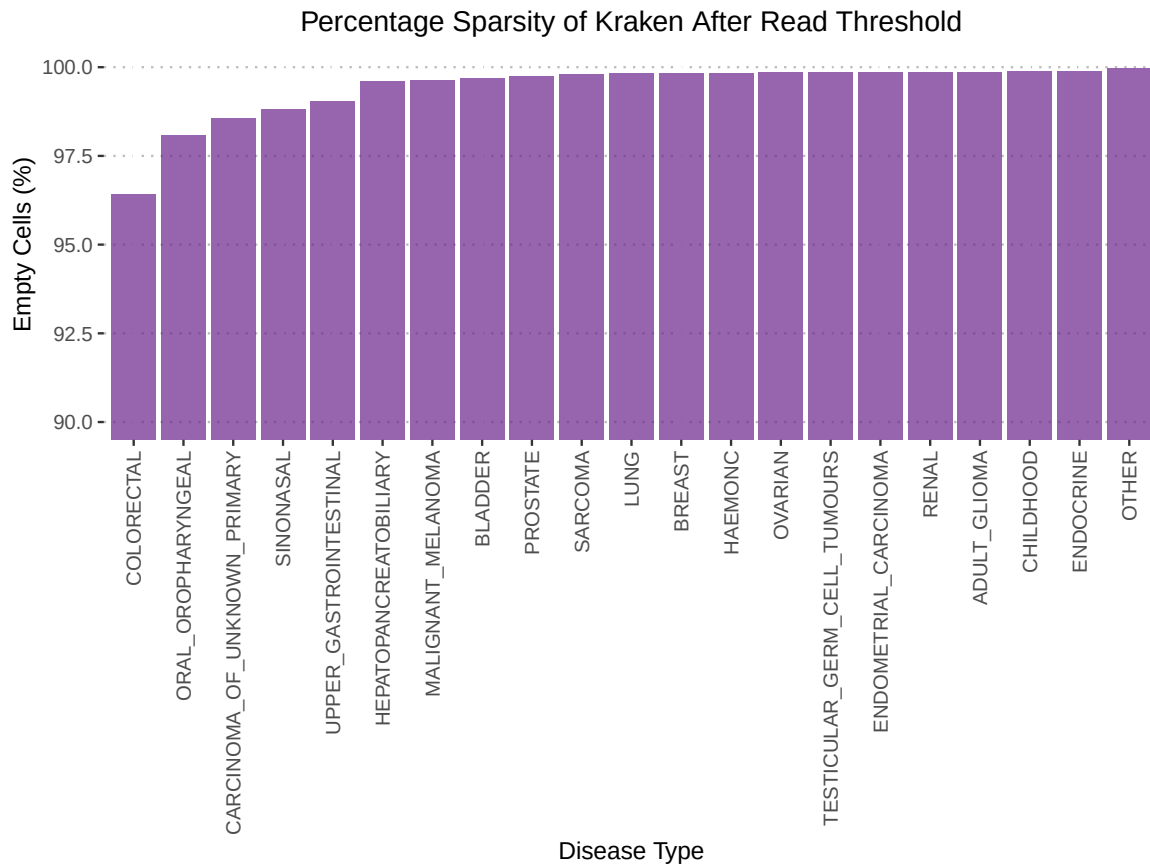


Figure 4.15: Sparsity by cancer type from applying SEPATH with Kraken to Genomics England's 100,000 Genomes project whole genome sequence data. The y axis notes the percentage of cells in the taxa-sample community matrix that are occupied by zero values in each cancer type. This is considering all 1,532 genera detected in the entire dataset after a read minimum read threshold of 20 was applied.

The percentage of positive samples within each taxa in each tumour type are displayed in figure 4.16. Colorectal, oral and upper gastrointestinal cancer are the least sparse. Colorectal boasts the highest number of taxa with positive samples (448 out of 1,532 total genera). This is far greater than the next highest cancer, breast with 285 taxa with any positive samples.

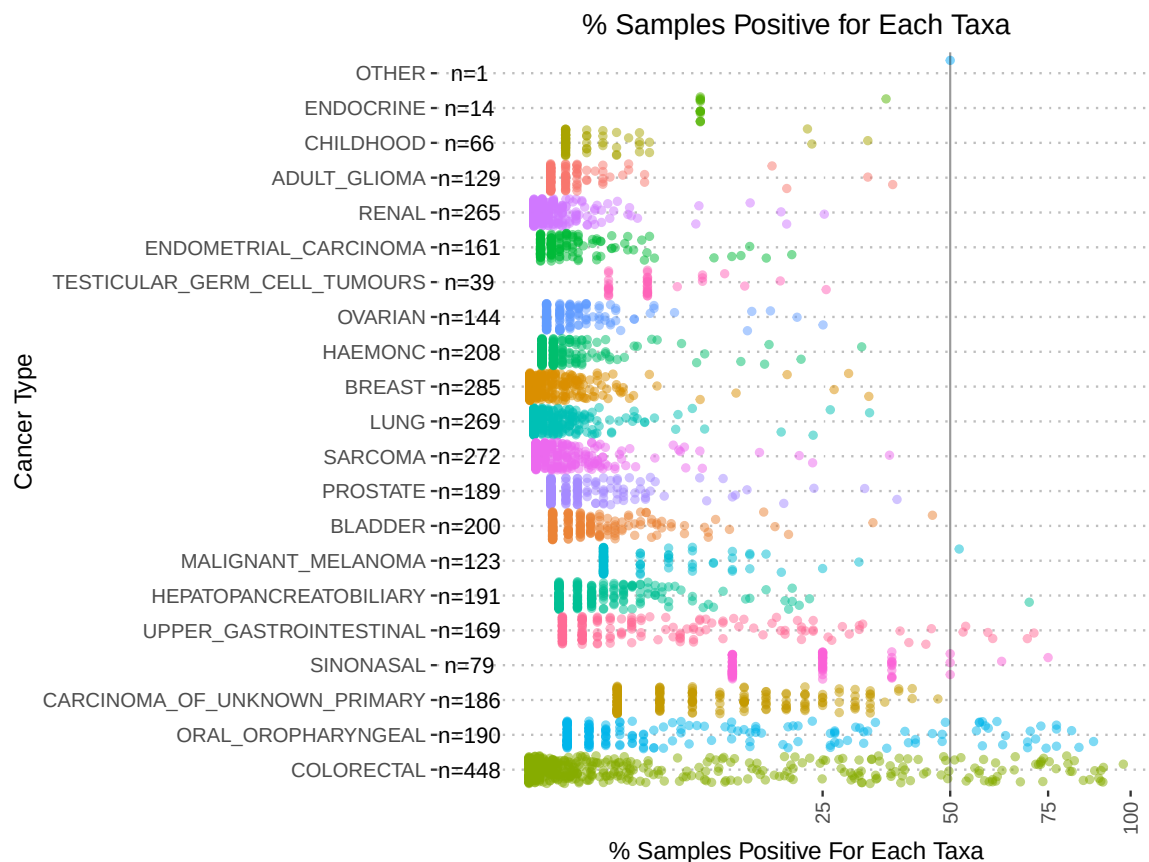


Figure 4.16: Sparsity in Kraken data. Each point represents a genus found in at least one sample. The percentage of positive samples for each taxa in each cancer type, considering 1,532 genera in the entire dataset (minimum read threshold of 20). The number of taxa considered in each cancer type is demonstrated on the *y*-axis. The line indicates taxa with 50 percent of samples positive. It is important to note that taxa with no samples positive were removed from this analysis. Tumour types are in no particular order.

4.4.5 HPV Diagnosis - Comparison of Kraken with Clinical Detection Status

Kraken results from oral/oropharyngeal tumour samples were compared with clinical HPV detection status as listed in the disease subtype of the `cancer_analysis` spreadsheet provided by Genomics England with each data release. Samples were filtered to retain only PCR free, and non-FFPE tissue. There were 48 samples with HPV status and SEPATH results, 38 were HPV⁻, 10 HPV⁺, including 2 samples from metastatic tissue (figure 4.17). SEPATH correctly identified Alphapapillomavirus reads in all 10 HPV⁺ samples. One sample recorded as HPV⁻ demonstrated > 100,000 reads to the Alphapapillomavirus genus which is strong evidence of presence. 101 samples with no clinical HPV result, eight of which were positive with SEPATH (> 100 reads).

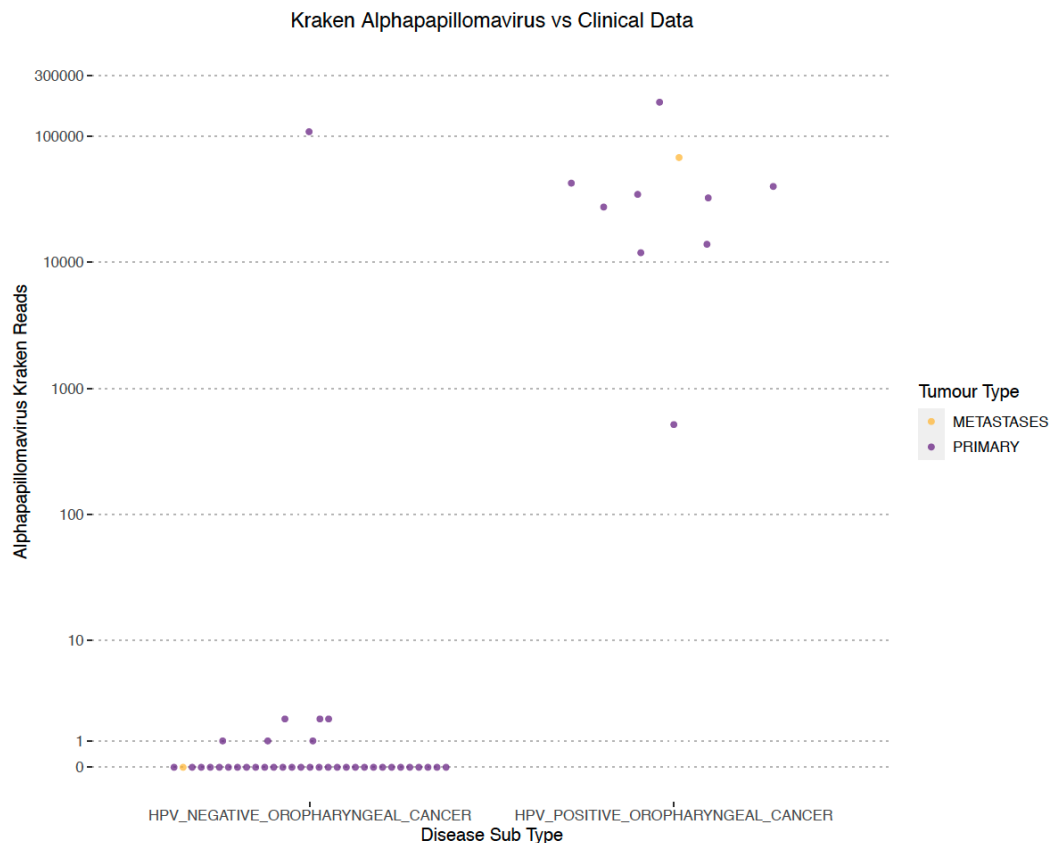


Figure 4.17: SEPATH Alphapapillomavirus reads using Kraken and clinical HPV test status. Samples from metastatic tumours are shown in yellow and primary tumours are purple. 38 samples were negative for HPV, 10 were positive. 101 oral tumour samples had no mention of HPV status for which SEPATH identified eight positive samples (minimum number of reads = 115).

It is known that HPV⁺ oropharyngeal tumours are depleted in the number of somatic *TP53* mutations compared to HPV⁻ tumours^{45,46}. Nonsynonymous mutations with an estimated moderate to high impact detected by Genomics England variant calling pipelines (methods section 4.8.2) were compared with HPV status in 38 oral/oropharyngeal tumour samples. 29 *TP53* variants were identified in 24 HPV⁻ samples from 24 participants. The predicted consequences of these variants are: 18 missense, 4 frameshift, 4 stop gained, 2 splice donor variant and 1 inframe deletion. The sample with high Alphapapillomavirus reads detected with Kraken but a HPV⁻ clinical test reported no variants in the *TP53* gene. This provides a level of evidence that the metadata is incorrect or that the clinical test is a false negative. These results have been reported to the oral / oropharyngeal GeCIP (Genomics England Clinical Interpretation Partnership) to assist their HPV stratification studies.

4.4.6 Reporting Clinically Actionable Results

This section reports on clinically actionable communicable diseases identified in Magiorkinis *et al.*³⁸⁶. This analysis describes the Kraken results of $N=11,735$ tumour samples irrespec-

tive of FFPE or PCR status.

Actionable Viruses with Insufficient Evidence

15 samples demonstrated any evidence for the *Lentivirus* genus (HIV). 11 samples had < 10 reads assigned to the genus. Three samples had 11 reads (two of which were PCR amplified) and one sample had 15 reads (also PCR amplified). This small number of reads is insufficient evidence for the presence of the *Lentivirus* genus.

Two samples showed any evidence for the *Orthohepadnavirus* genus with one and three sequencing reads. This is insufficient evidence for the presence of the Hepatitis B virus. *Hepacivirus* is the genus for Hepatitis C virus. To date, there is no evidence for this genus in the cohort ($N=11,735$). This may be due to the removal of viral sequences by human depletion steps (discussed in chapter 2).

Human T-Cell Lymphotropic Virus

Three participants demonstrated any signs for the presence of the *Deltaretrovirus* genus. One participant with prostate cancer had 5 reads and one female participant with 8 sequencing reads for the *Deltaretrovirus* genus. This is too few reads to explore any further.

One participant with invasive breast ductal carcinoma had 31 reads classified to the *Deltaretrovirus* genus in their tumour sample. This represents an ambiguous result which warrants additional investigation. The tumour sample was PCR-free and was prepared from fresh frozen tissue obtained from surgical resection. SEPATH was re-run on the tumour sample and the same result obtained. The paired blood germline sample also contained *Deltaretrovirus* using SEPATH with 32 sequencing reads uniquely assigned to *Deltaretrovirus*.

Reads with taxonomic classifications (Kraken) to the *Deltaretrovirus* genus were extracted for additional analysis. In total 172 reads were obtained from the tumour and the germline sample. These sequencing reads were subject to a BLAST search (nr/nt database) which obtained small E -values (often 1×10^{-70}) and high percentage identities in the range of 99-100%.

The participant is from the ethnic category 'Black or Black British'. As the patient is from an ethnic minority, extra care should be taken to ensure that reads are truly mapping to pathogens and are not variation in the ethnically African human genome (discussed in section 2.3.2). These reads were aligned to a HTLV-1 reference genome (methods) and viewed with IGV³⁹⁵ (figure 4.18, methods section 4.8.3). The viral reads from both the tumour and germline sample align throughout the HTLV-1 genome. This is additional evidence for the presence of HTLV-1. As a result of this, I have requested an independent validation of these results and will be discussed with the infectious disease group within Genomics England.

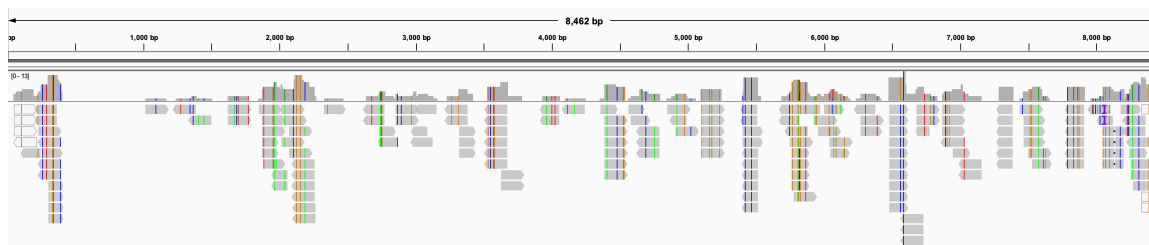


Figure 4.18: Alignment of HTLV-1 classified reads (Kraken) from tumour and germline samples from one participant. The image shows the alignment viewed with IGV. The top track denotes coverage for particular regions (maximum coverage = 13). Coloured regions indicate single nucleotide differences present in the reads and not the reference genomes (orange=G, blue=C, red=T, green=A). In total 172 quality-trimmed, human-depleted reads were subject to alignment (66 and 106 reads from the tumour and germline sample respectively).

4.4.7 Investigating ABBS Genera in Genomics England

The percentage of samples positive for at least one ABBS genera (*Ezakiella*, *Peptoniphilus*, *Porphyromonas*, *Anaerococcus*, *Fenollaria* and *Fusobacterium*) within each clinical stage and tumour type was determined (figure 4.19). Most samples in oral and colorectal cancer were positive for at least one ABBS genera at all clinical stages. Bladder cancer shows a trend of increasing percentage of positive samples as clinical grade increases (apart from stage four) with 60% of stage three cancers being positive for at least one ABBS genus. Endometrial and hepatopancreatobiliary also demonstrate this trend. Some clinical stages have small sample sizes, such as stage four disease in hepatopancreatobiliary ($n=3$) and so care should be taken. Full distributions of reads and clinical stage sample sizes are shown in appendix section D.14.

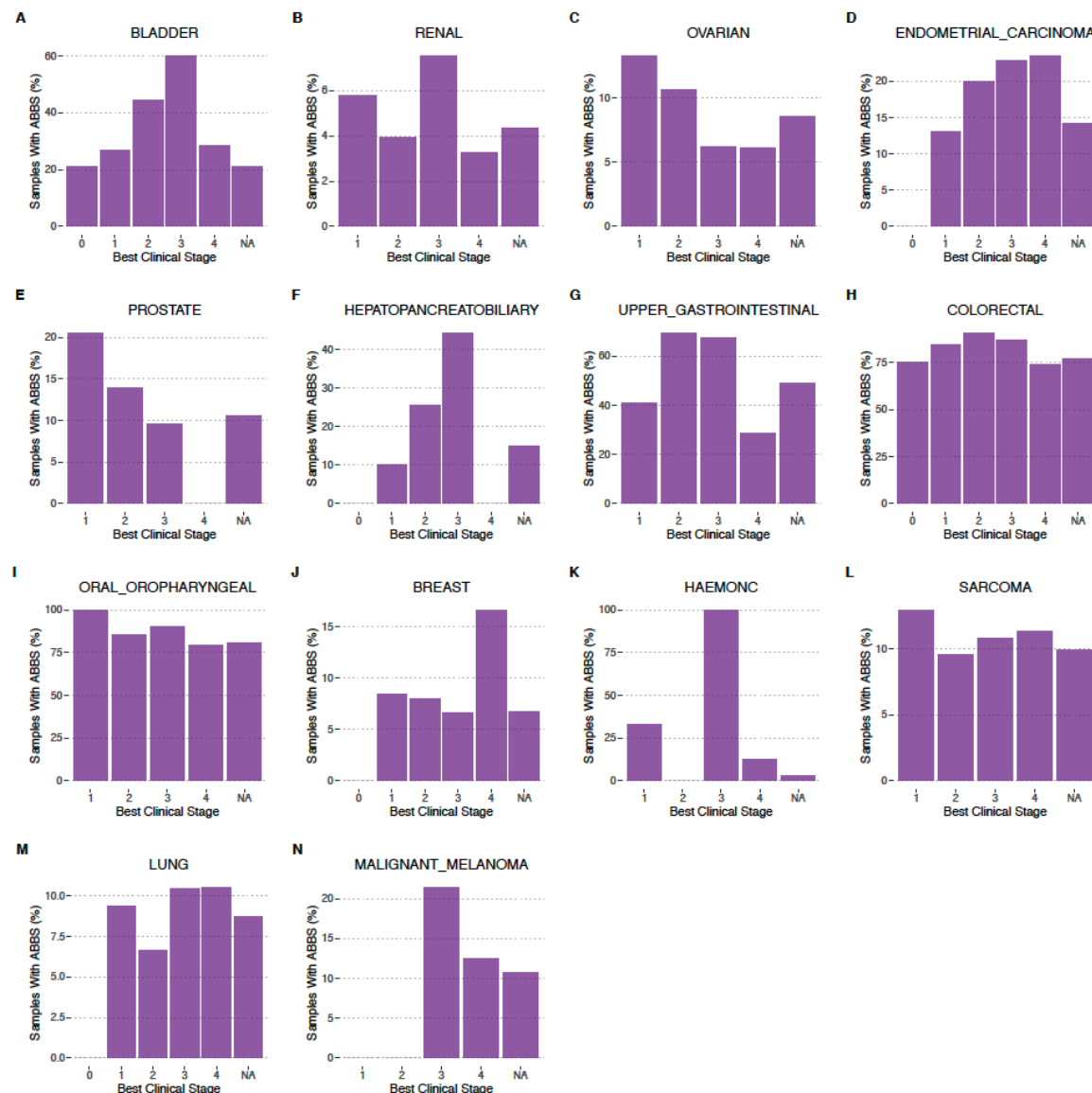


Figure 4.19: The percentage of participants with a positive match for at least one of the ABBS genera at each clinical stage and within each cancer type. Kraken results had a minimum threshold of 10 reads. Please note the y -axis ranges differs between graphs.

4.4.8 Kraken Reported Tumour Viriome Structure

Viral classifications from SEPATH were analysed on a subset of tumour types with the largest cohort sizes ($n=8,541$, figure 4.20). This was due to the lack of tumours with any reported viral constituents. To limit contamination only the following viral genera were considered: *Lymphocryptovirus*, *Orthohepadnavirus*, *Roseolovirus*, *Alphapapillomavirus*, *Cytomegalovirus*, *Gammaretrovirus*, *Alphatorquevirus*, *Lentivirus* and *Alphapolyomavirus*. This 'white-list' of viral taxa was curated based on prevalence estimated by Zapatka *et al.*⁹⁶. A white-list was required to eliminate the prevalence of bacteriophages (which are not truly representative of the virome) upon a string match for 'virus' *i.e.*,

T4virus, *Felixo1virus* and *Kp15virus* (101054, 92048 and 405292 total classified reads in the cohort respectively).

After filtering out samples with no evidence of these viral taxa, 857 (10%) of samples remained. *t*-SNE was applied using a distance matrix of Spearman's correlation values. A density-based clustering algorithm (dbscan, methods section 4.8.4) estimated an optimal of seven clusters.

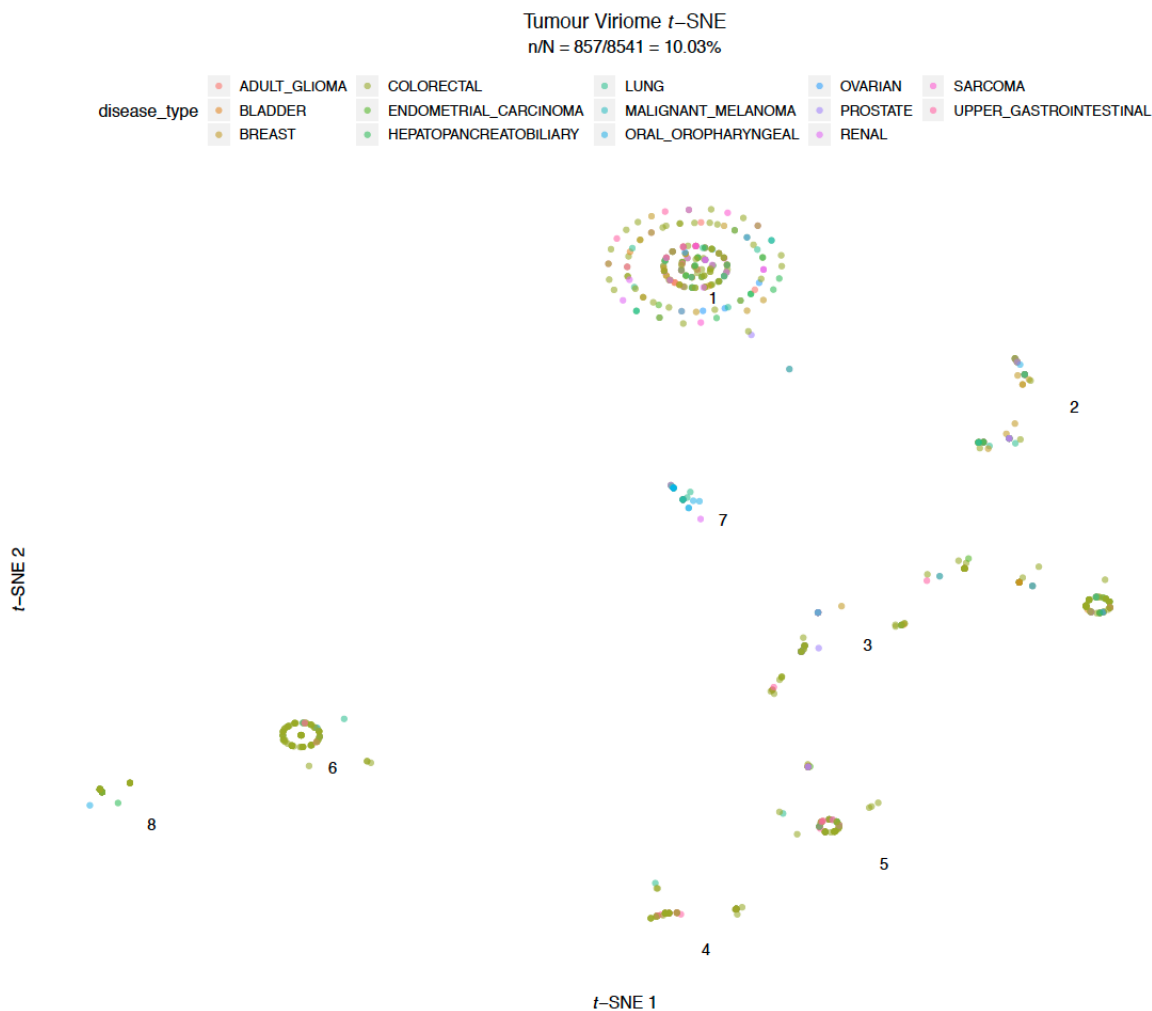


Figure 4.20: A *t*-SNE plot of viral genera detected by Kraken in the 100,000 genomes project. Detection was limited to nine viral genera (*Lymphocryptovirus*, *Orthohepadnavirus*, *Roseolovirus*, *Alphapapillomavirus*, *Cytomegalovirus*, *Gammaretrovirus*, *Alphatorquevirus*, *Lentivirus* and *Alphapolyomavirus*) and 14 cancer types (glioma, colorectal, lung, ovarian, sarcoma, bladder, endometrial, melanoma, prostate, upper gastrointestinal, breast, hepatopancreatobiliary, oral and renal). Samples with no evidence of these genera were removed from analysis. The data for the plot was a distance matrix of Spearman's correlation values with parameters: random seed=8, dimensions=2, theta=0.2, perplexity=50, maximum iterations=5000. Clusters were determined by a density base clustering algorithm from the dbscan package (parameters: eps=3, minimum points=5).

Characterising *t*-SNE Clusters of Viral Community

Based on the clusters determined above, the distribution of clusters amongst disease types (figure 4.21), and the distribution of clusters amongst viral genera (figure 4.22) was investigated. This analysis found *Roseolovirus* to be very high in terms of the number of reads and distributed among multiple cancer types (figure 4.22). *Alphapapillomavirus* was practically exclusive to oral cancer 4.22). *Lymphocryptovirus* was divided between clusters 3 and 4 which contain mixed types of tumours, notably bladder and prostate. *Gammaretrovirus* reads were distributed between clusters 2 and 6 which contained a mixed variety of tumour types but included breast and colorectal. *Cytomegalovirus* reads were distributed among multiple clusters.

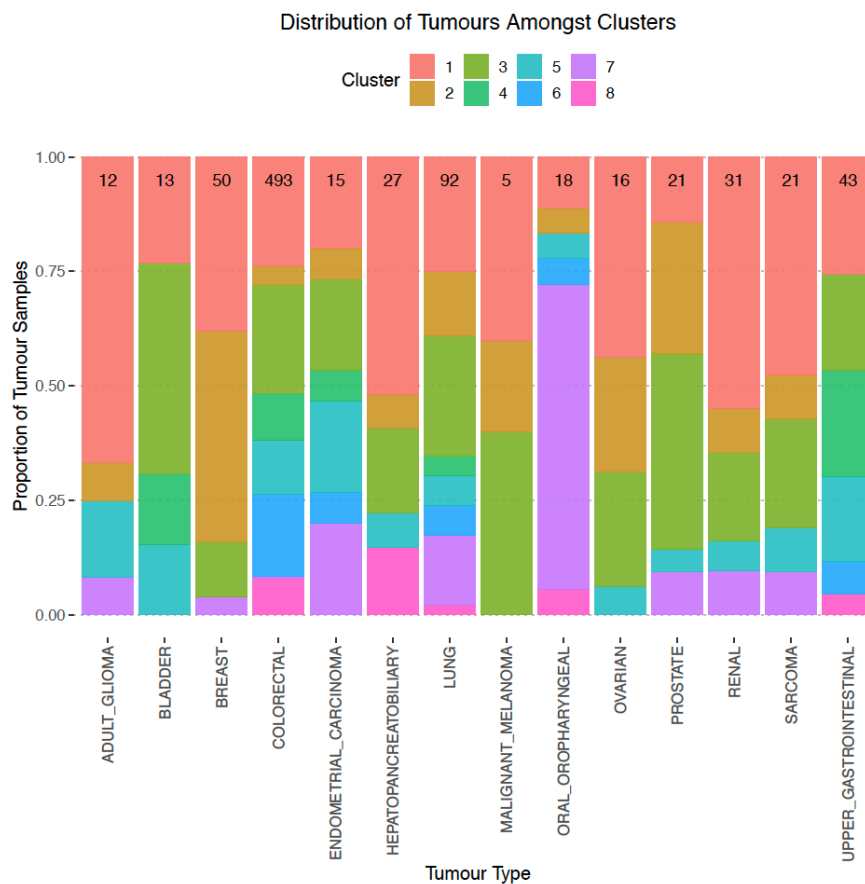


Figure 4.21: The distribution of tumour types amongst *t*-SNE virome clusters (as in figure 4.20). Numbers at the top of each bar indicate the number of samples within individual tumour types

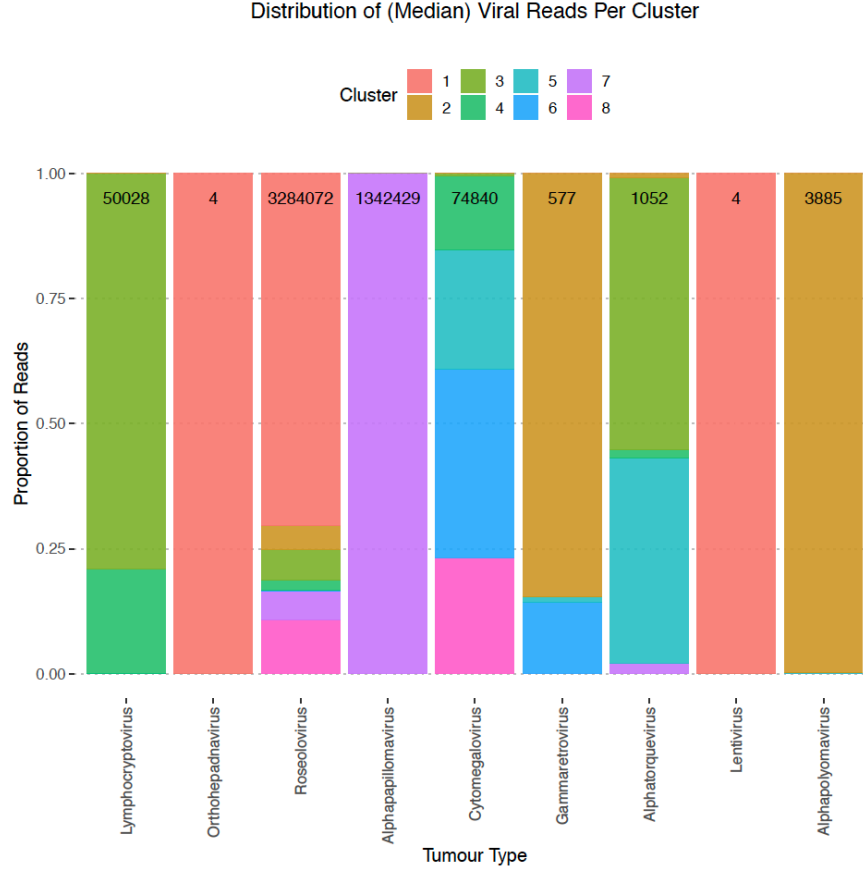


Figure 4.22: The distribution of viral genera amongst t -SNE viriome clusters (as in figure 4.20). Numbers at the top of each bar indicates the number of total number of reads assigning to each viral genus.

4.4.9 The Search for Microbial Community Structure

Recent publications including Poore *et al.* 2020¹⁵⁴ and Nejman *et al.* 2020¹⁷³, suggest that the microbial constituents of tumours is site specific. To reproduce these findings, the community structure of all tumour samples in Genomics England 100,000 Genomes Project (version 7, $N=11,735$) was investigated. No notable microbial community structure was observed that supports this hypothesis. PVCA^{396,397} (Principal Variance Component Analysis, as used in Poore *et al.*) was used to estimate what variables explain variance in the dataset, both on raw Kraken results and on Voom³⁹⁸ normalised Kraken results (used to limit heteroskedasticity and subsequently improve linear modelling). This analysis suggests that most variance in the dataset is unexplained (61% in raw and 57% in Voom) by the variables selected (figure 4.23). It additionally shows that Voom normalising data increases the amount of variation explained by biological and technical variables (*i.e.*, preparation method increases from 12% to 15% in Voom normalised data). There is no variance (values round to 0) attributed to the laboratory the sample was processed at, the laboratory method used or the site the sample was collected from. This suggests that batch correction is not required for this dataset. In Poore *et al.* 12 – 15% of variation in the dataset is explained by preparation method (fresh-frozen, FFPE, EDTA, aspirate *etc.*) followed by by disease type (5 – 7%), tissue source (surgical resection, biopsy *etc.*), library type (PCR amplified vs

PCR-free) and sex. This illustrates the importance of filtering out PCR amplified samples and FFPE samples which can confound microbial analysis.

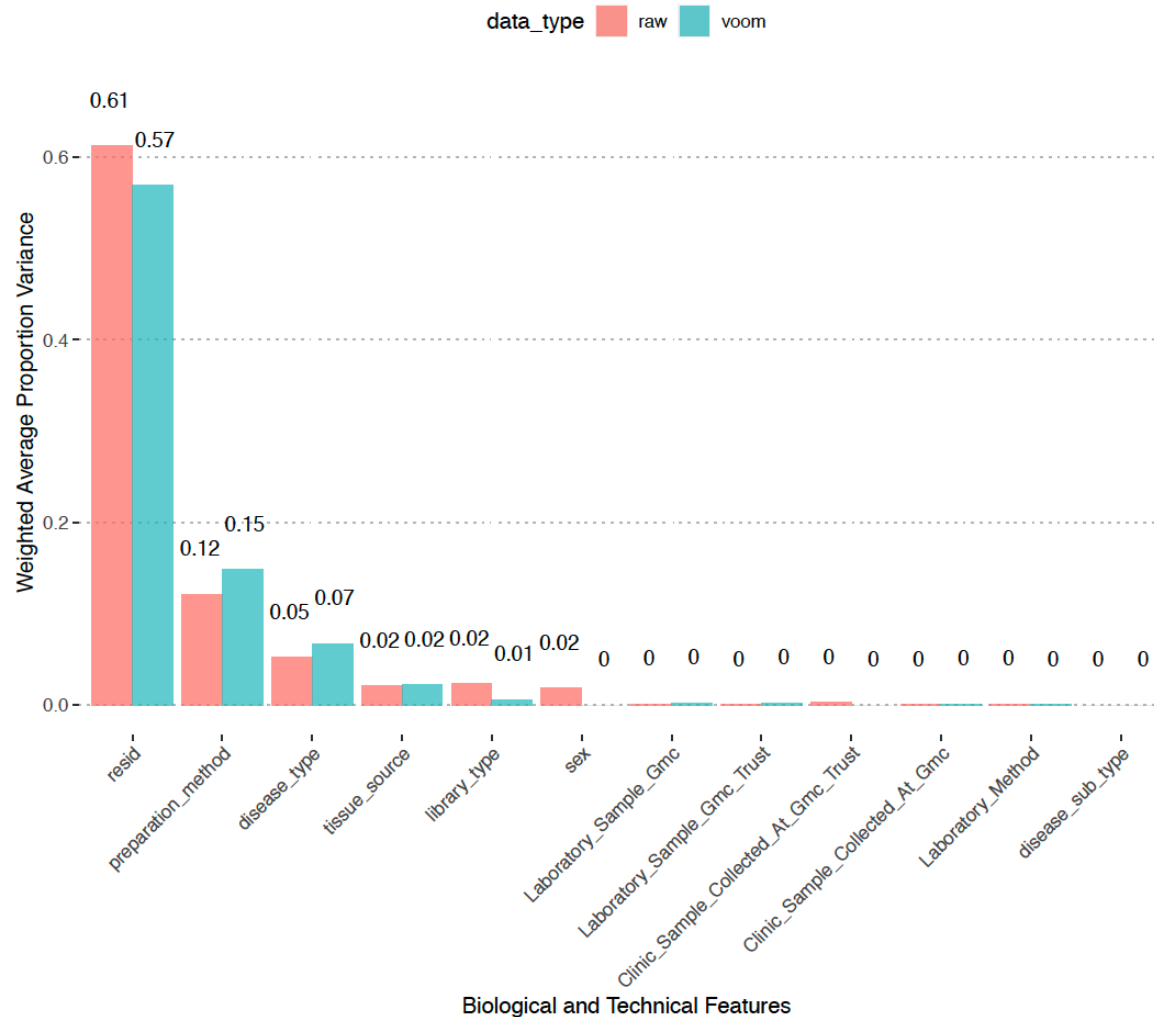


Figure 4.23: Principal Variance Component Analysis (PVCA) to estimate the variance explained by biological and technical variables in the dataset of Kraken classifications of tumour samples. Residual variance that is not explained by the variables investigated is denoted 'resid'. GMC = Genomic Medicine Centre

Modelling Tumour Type

Given the variance explained by tumour type, supervised models might detect differences between tumour types that could not be seen with ordination. Classification models (gradient boosted machines - GBM) were built on a training-set of Voom normalised Kraken data to detect one tumour type vs all others (figure 4.24, methods section 4.8.7). The scripts utilised represent the same procedures as in Poore *et al.*¹⁵⁴. These results show reasonable AUC (Area Under a receiver operating characteristic Curve) results underpinned by good sensitivity and specificity but generally poor PPV (apart from colorectal, AUC=0.97,

F1-score=0.91).

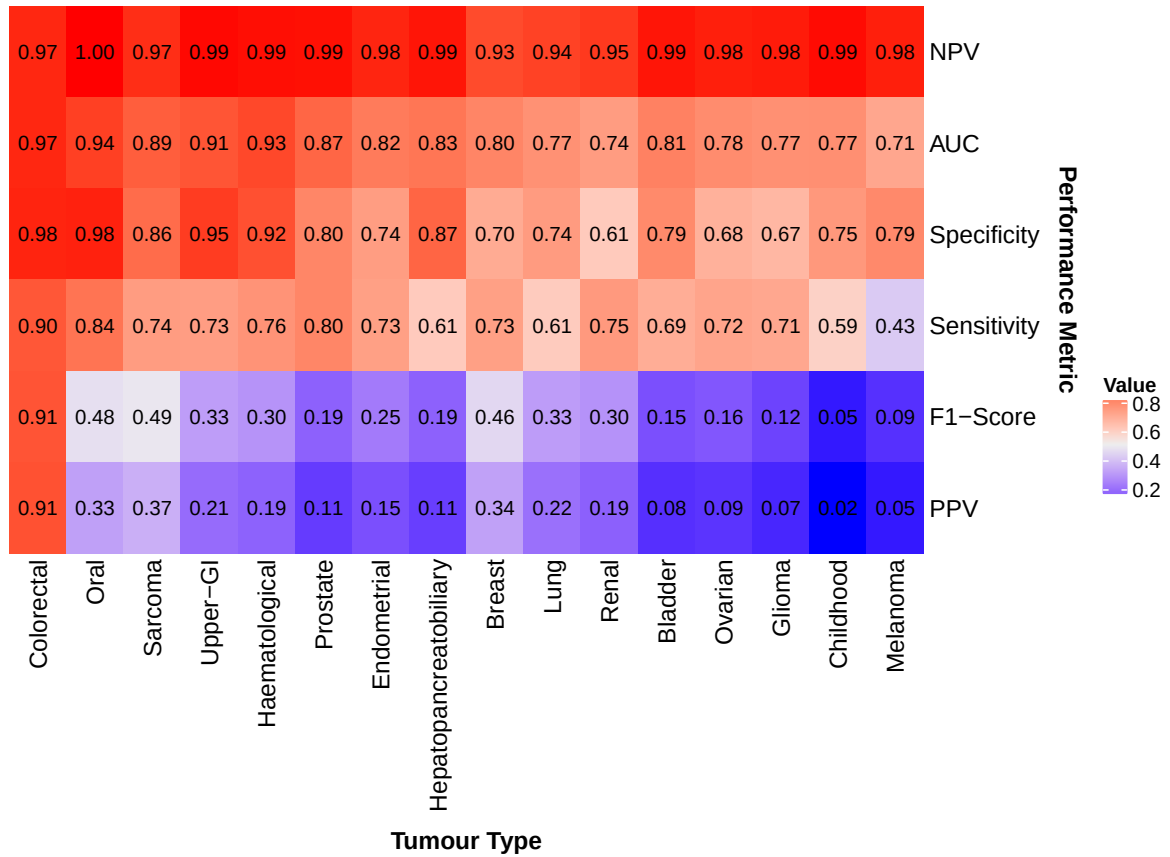


Figure 4.24: Performance of one-vs-all GBM tumour models on Genomics England Voom normalised Kraken data. Contaminants were removed according to Salter *et al.* (NPV = negative predictive value, AUC = area under the ROC curve, PPV = positive predictive value).

Taxa that had a high importance in predicting the tumour type in one-tumour-type-vs-all models were extracted and their stability was tested using Boruta. Of the top ten important features in predicting colorectal cancer for GBM and Boruta, only two taxa overlapped: *Bacteroides* and *Roseolovirus*. Only *Dorea* overlapped between Boruta/GBM top ten features and the top ten features in by Poore *et al.*, which was the sixth most important feature of the Boruta approach. Further details of Boruta/GBM importance is found in appendix section D.9.

In a final effort to search for structure by disease type, a list of the most *informative* taxa from the GBM models was manually curated and a *t*-SNE plot was constructed based only on these taxa (figure 4.25). A much smaller number of the most informative taxa was also investigated (appendix figure D.24).

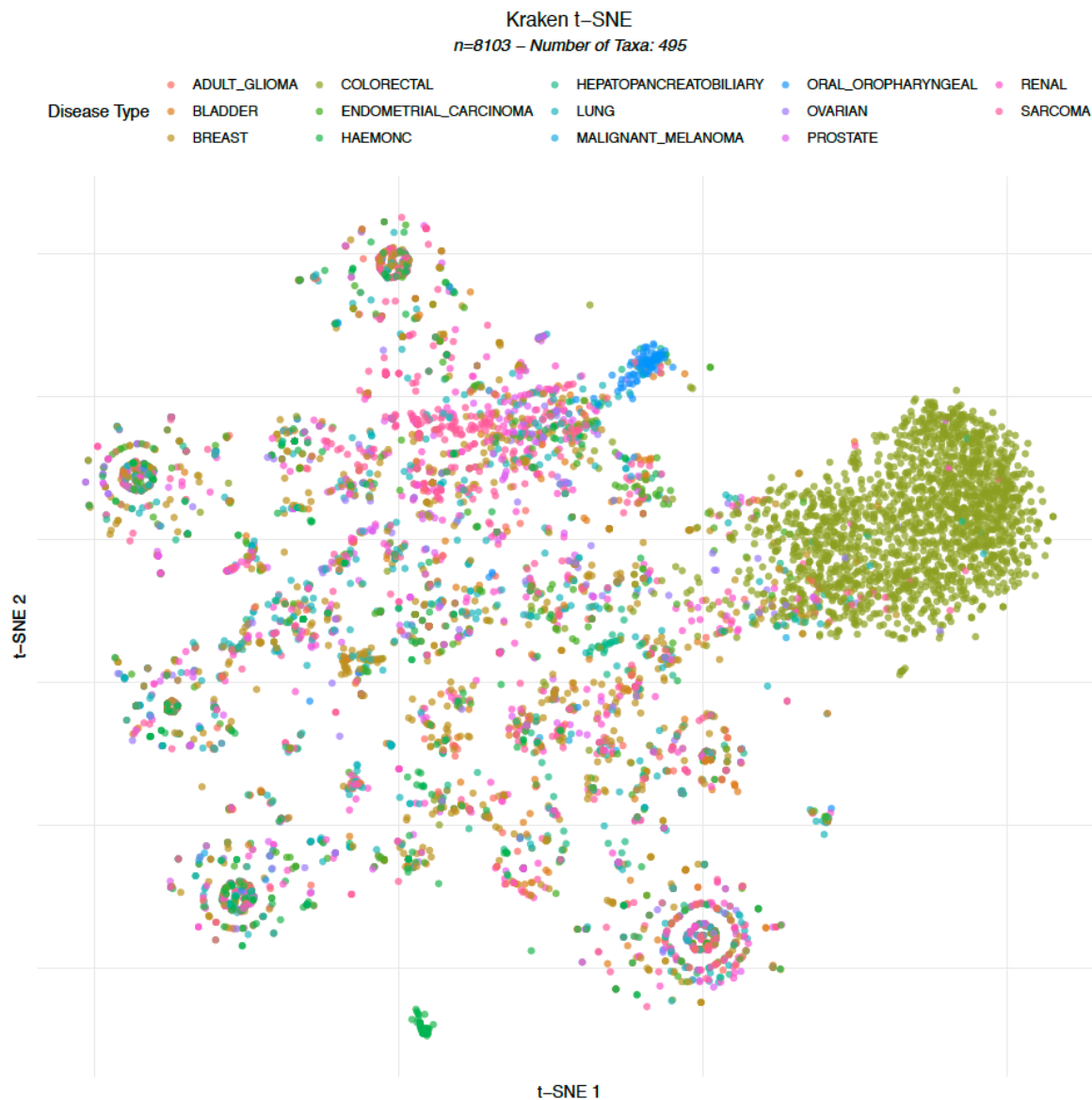


Figure 4.25: *t*-SNE plot of Kraken results coloured by tumour type for the microbial genera that most informative of tumour type. *t*-SNE was carried out on a matrix of Spearman's correlation values between samples. Informative taxa are defined as those taxa that were in the top 1500 most important taxa in one-vs-all GBM models. 495 unique taxa met this criteria in 8103 non-FFPE, PCR free, primary tumour samples further details available in 4.8.5.

4.4.10 Investigating the Poore *et al.* Manuscript

The results presented in this chapter suggest that there are limited microbial specificities between tumour types. This finding is in opposition to Poore *et al.* and so it was decided to investigate their manuscript further.

Human Reads Have Likely Been Misclassified

Poore *et al.* ran Kraken directly on unaligned sequencing reads without filtering. Many of the non-aligned reads are human in origin (section 2.3.2). For this reason, SEPATH runs additional human read depletion using BB DuK. Additionally, human sequences have been included in the reference database for Kraken and is the most frequent classification in the dataset. Additionally, the classifications are filtered such that a minimum of 20% of the k -mers in a sequencing read must be unique to a specific taxon.

The Poore manuscript did not implement any of these additional controls, utilising only 59,974 purely microbial genomes. This has the effect that any human read that has got through will most likely most closely match a microbial taxon. This suggests that human reads may have been misclassified as microbial in origin.

Voom Normalisation Inappropriately Distorts Values for Absent Taxa

Poore *et al.* applied Voom³⁹⁸ to limit heteroskedasticity in Kraken count data and to allow for linear modelling. Voom is ordinarily used for RNA sequencing data whereby there is a base level of expression expected in most genes. Applying this to Kraken count data is inappropriate. Consider an example dataframe with three taxa and five samples.

Table 4.1: A sample community matrix with taxonomic count data.

taxa1	taxa2	taxa3
2	2000	0
5	400	2
10	20	0
20	90	0
10000	62	0

Now we apply Voom normalisation to this community matrix and the zero counts have been transformed to non-zero values (taxa 3; Table 4.2).

Table 4.2: Voom normalisation of a sample community matrix with taxonomic count data.

taxa1	taxa2	taxa3
7.960312	19.56848	17.34661
9.097816	17.24799	19.66853
10.030702	12.95988	17.34661
10.995936	15.10218	17.34661
19.926169	14.56812	17.34661

This means that taxa can appear as important features when they only contain evidence of presence in one sample. Take Hepandensovirus for example. This taxon was present in 1/17,625 samples in a head and neck cancer sample and had two sequencing reads assigned. Poore *et al.* found Hepandensovirus to be by far the most important feature (relative

importance score over 9,000 compared with 900 for the next most important feature) in predicting an adrenocortical carcinoma sample, even when they used the most stringent contamination controls. This virus has been associated with causing mortality in prawns and is therefore highly likely to be a false positive and not truly associated with cancer³⁹⁹.

Supervised Normalisation Introduces Inappropriate Variation

Poore *et al.* demonstrated a prominent batch effect by sequencing site. Therefore, to resolve this they utilised a supervised normalisation technique (SNM) developed for microarrays⁴⁰⁰. They show that variance is preserved by disease type but reduced by sequencing centre and other technical variables. Upon investigating the data, it seems that disease type and sequencing centre are highly linked with eight single cancer types originating from only one sequencing centre (adrenocortical carcinoma, cholangiocarcinoma, mesothelioma, pancreatic adenocarcinoma, pheochromocytoma and paraganglioma, testicular, thymoma and uterine carcinosarcoma). There are nine cancer types with 75% of their samples from one sequencing centre including: kidney renal clear cell carcinoma, glioblastoma, lung squamous cell, breast, kidney, ovarian serous cystadenocarcinoma, kidney renal papillary cell carcinoma, hepatocellular, lymphoid neoplasm, diffuse large B-cell lymphoma and sarcoma. This raises concerns that reducing the batch effect between sequencing centres also removes real differences between cancer types as these variables are linked.

Reviewing Hepandensovirus samples with no Kraken reads, we see that Voom has adjusted these to non-zero counts, and interestingly SNM has introduced variance where variance should not exist (figure 4.26). There is an observable split by sequencing centre that demonstrates that these values have been incorrectly adjusted to account for differences in sequencing centre. It is quite possible that a machine learning algorithm could incorrectly model this variance and interpret it as an important feature in distinguishing between disease types.

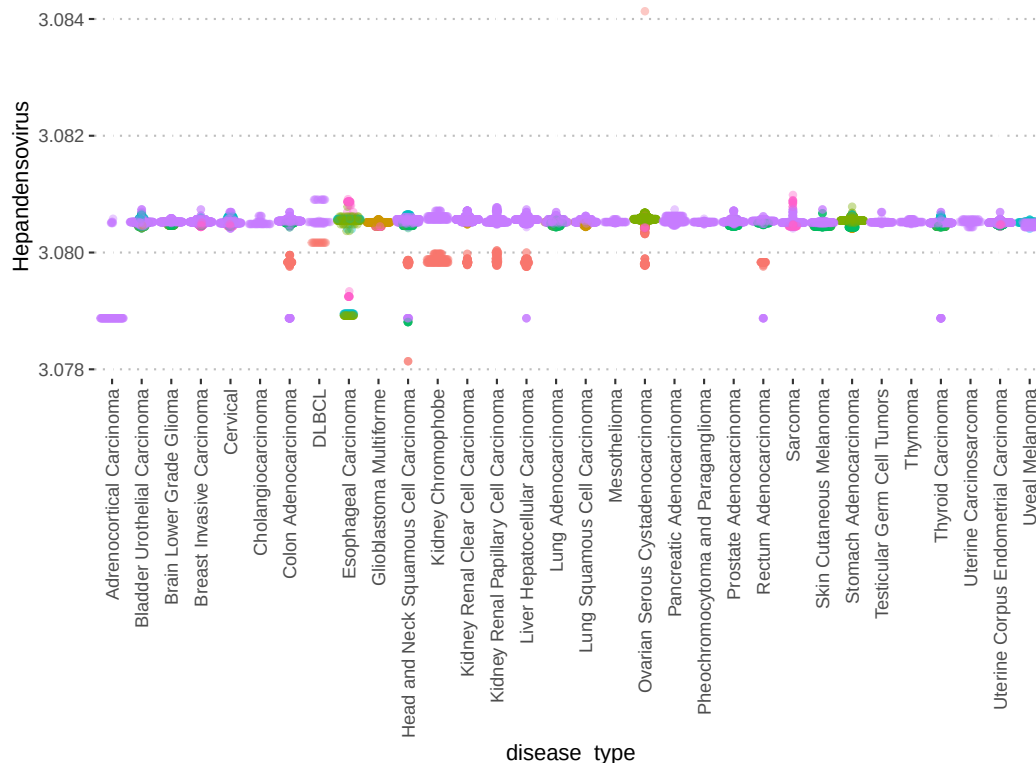


Figure 4.26: Supervised normalisation introduces variance into taxa with no reads where variance should not exist. These samples originally had zero reads for Hepandensovirus. Disease type is on the x -axis, the y -axis represents Voom-SNM normalised values and points are coloured by sequencing centre. The data used was Voom-SNM normalised Kraken classification data as published by Poore *et al.* 2020¹⁵⁴.

TCGA Model Performance is Sensitive to the Sampling Method Used to Balance Datasets

Poore *et al.* balanced the class sizes for each model built by upsampling the data set of the cancer type being predicted to equal the size of the rest of the samples. The one-vs-all tumour classification models from Poore *et al.* (TCGA data) were repeated using downsampling, standard upsampling, SMOTE¹⁴⁹ and ROSE⁴⁰¹ (figure 4.27). Upsampling as presented by Poore *et al.* manuscript is the most performant approach, closely followed by SMOTE. This analysis repeats the most stringent contamination removal on Voom and SNM normalised data.

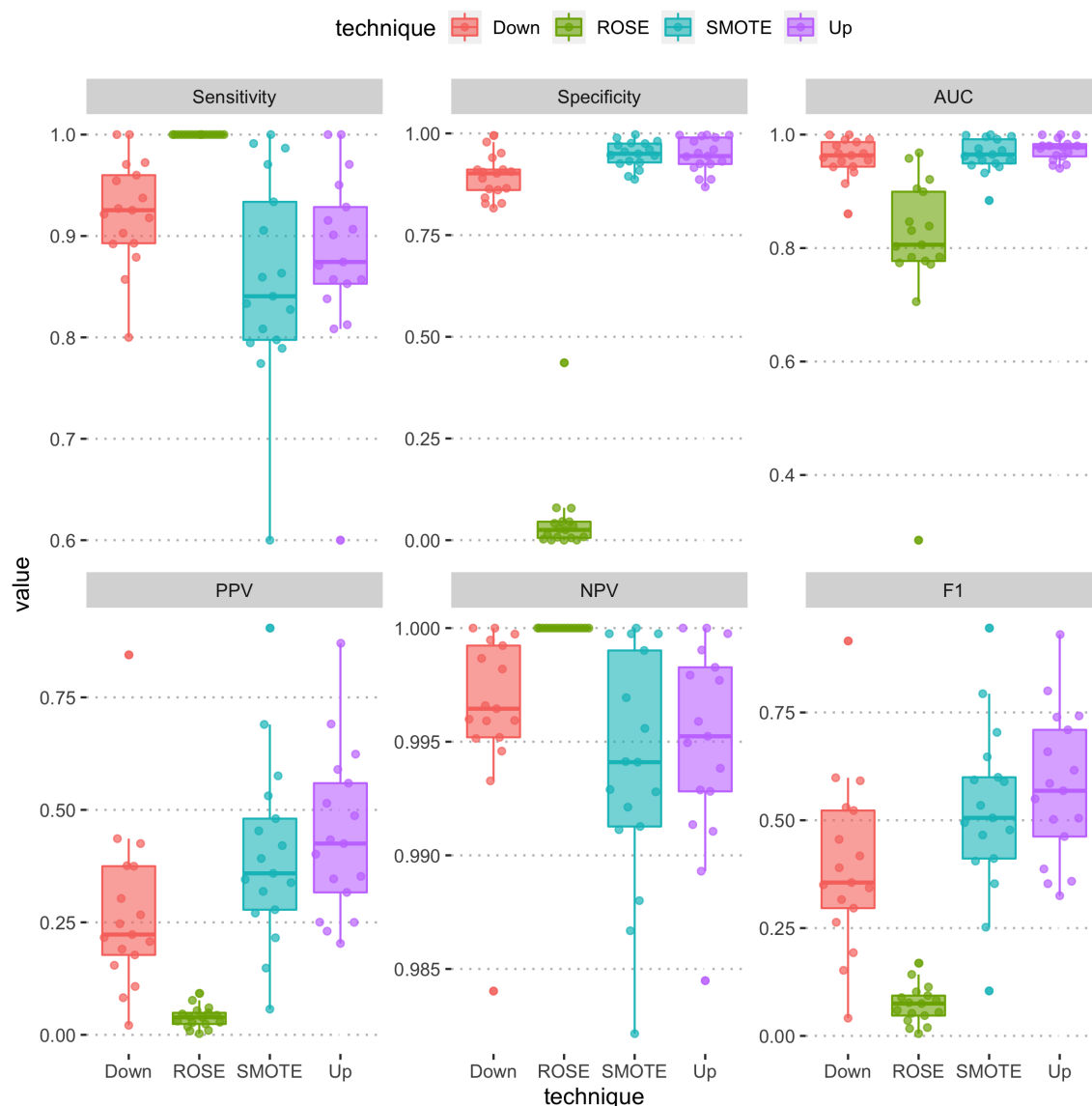


Figure 4.27: Models to predict one tumour type vs all other samples from Poore *et al.* 2020 were reproduced using different sampling techniques. Performance is summarised here in terms of sensitivity, specificity, AUC, positive predictive value, negative predictive value and F1-score. Sampling techniques are for downsampling (red), ROSE (green), SMOTE (blue) and standard upsampling (purple).

Nonsensical Genera are Found to be the Most Important Taxa to Predict Cancer Type in TCGA

Considering only the most important feature for each one-vs-all tumour model in Poore *et al.* (most stringent decontamination), genera are revealed that are nonsensical as they are not known to exist in humans. This data is available from http://cancermicrobiome.ucsd.edu/CancerMicrobiome_DataBrowser.

- Leucothrix - Bladder urothelial carcinoma. Exist naturally in marine habitats⁴⁰².
- Velarivirus - Cervical squamous cell carcinoma and endocervical adenocarcinoma. Grapevines act as the natural host⁴⁰³.
- Tritimovirus - Colon adenocarcinoma. Plants like wheat are the natural hosts⁴⁰⁴.
- Dinovernavirus - Kidney renal cell carcinoma. Known to replicate in mosquitos⁴⁰⁵.
- Bacillarnavirus - Breast invasive carcinoma. Algae is natural host⁴⁰⁶.
- Rymovirus - Ovarian serous cystadenocarcinoma. Infects species of grass⁴⁰⁷.
- *Ignicoccus* - Prostate adenocarcinoma. Archaea identified in marine hydrothermal vents⁴⁰⁸.
- *Salinimicrobium* - Testicular germ cell tumours. A halophilic bacteria identified from marine environments⁴⁰⁹.
- *Thalassomonas* - Uveal melanoma. Causes disease in coral⁴¹⁰.

The top feature for Lymphoid Neoplasm Diffuse Large B-cell Lymphoma was ‘contaminant5RandomSpikes1000’.

Repeating Models with Nonsensical Genera Maintains Performance

Using only the nine nonsensical genera discussed above, the one-vs-all tumour models presented in Poore *et al.* were rebuilt (on Voom-SNM TCGA Kraken data). Near perfect results in terms of sensitivity, specificity, NPV and AUC were obtained.

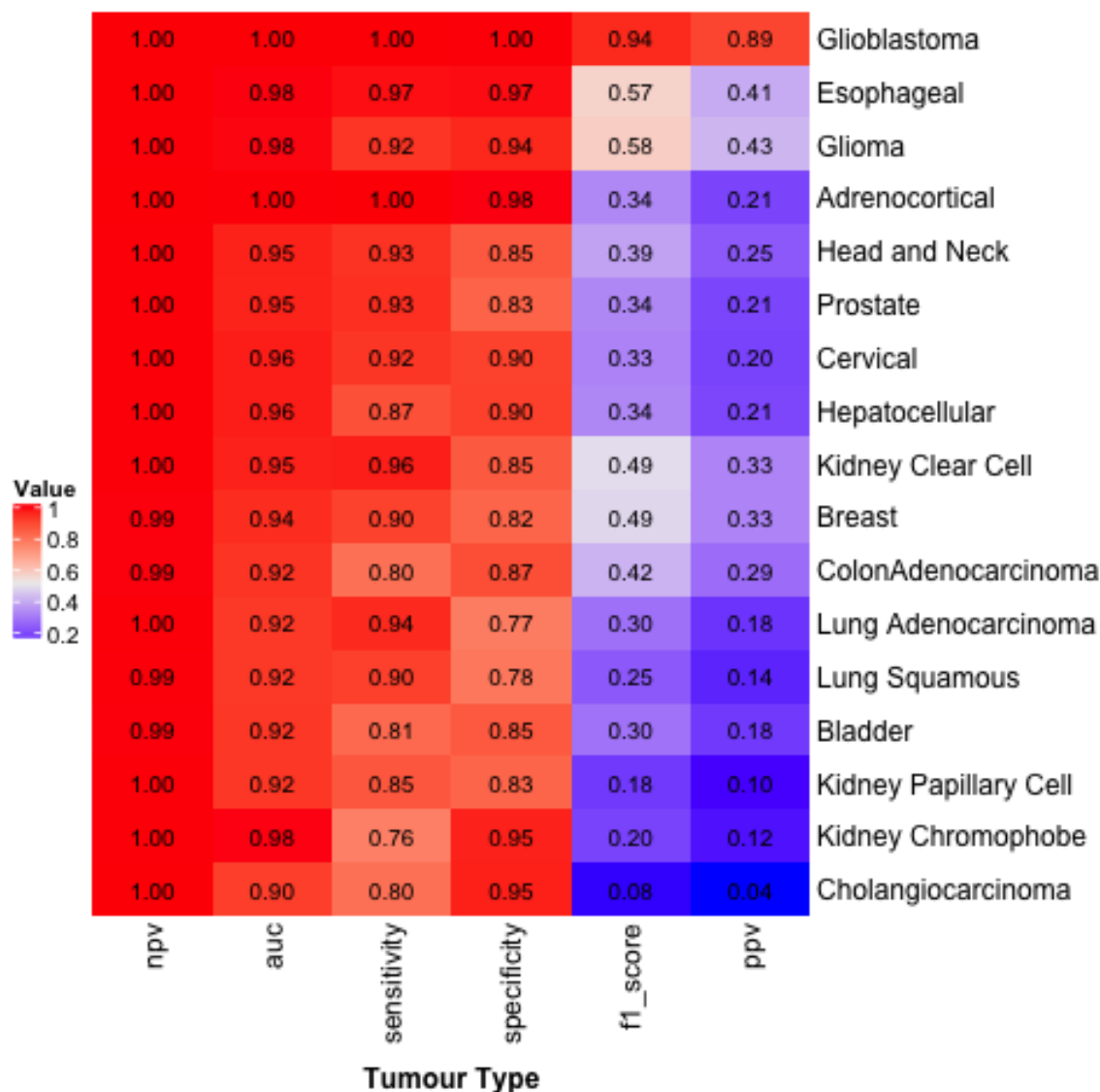


Figure 4.28: TCGA one vs all tumour models were repeated using only nine nonsensical taxa (*Leucothrix*, *Velarivirus*, *Tritimovirus*, *Dinovernavirus*, *Bacillarnavirus*, *Rymovirus*, *Ignicoccus*, *Salinimicrobium* and *Thalassomonas*) that were deemed the most informative in predicting between tumour types. This data was reanalysed from Voom-SNM normalised data with contaminants removed from Poore *et al.* 2020.

4.5 Discussion

4.5.1 Kraken vs mOTUS2 and Sparsity

Running SEPATH with Kraken on Genomics England data identified a range of microbial genera. Many of these genera were missing from mOTUS2 results and it proved insensitive. Although the benchmarks in chapter 2 suggested that mOTUS2 was sensitive, this does not seem to be the case for this dataset. A much smaller number of classified reads is present

in the Genomics England dataset (mean 550,611 ‘root’ classified reads) than was in the simulated dataset (mean 28,400,000 bacterial reads per simulated metagenome figure 2.2 vs 4.3). The number of reads in the simulated datasets were based on the number of unmapped reads observed on the PCAWG project and so the difference observed here likely resembles differences between datasets. Poore *et al.* identified an average of 25 million classified reads per sample in the TCGA dataset which is closer to the simulated datasets¹⁵⁴. Because there is less genomic information available to classify, there may be insufficient representation from the limited number of marker genes that mOTUs2 requires to make a classification. Kraken is much more sensitive than mOTUs2 because it can classify based on any unique 31-mer as opposed to a small subset of genes.

4.5.2 HPV Detection in Genomics England

SEPATH using Kraken identifies Alphapapillomavirus with a performance similar to clinical tests (figure 4.17). As this approach directly interrogates the tumour sample as opposed to the host overall, it may offer an improved approach to stratifying samples. SEPATH also detected HPV in all metastatic samples from patients with a HPV⁺ clinical test result. At the time of writing, these findings are being provided to the researchers involved with utilising HPV data. The somatic *TP53* analysis might suggest that the clinical metadata is inaccurate in that the sample with a high number of Alphapapillomavirus reads (Kraken) is negative for *TP53* variants.

4.5.3 Implications of HTLV Detection

The methods in this thesis show potential to accurately diagnose infection from sequenced human samples but care must be taken to ensure that these decisions have no underlying ethical bias. This is particularly pertinent in the case of HTLV-1 which is endemic to Sub-Saharan Africa. Unseen human variation from ethnically African genomes may have been incorporated into HTLV-1 reference assemblies. Additionally, this variation may be more likely to make it through to classification steps and subsequently be misclassified. Care has been taken to avoid this by including additional African genome variation²⁰⁴ alongside human reference genome GrCh38 in the SEPATH contaminant database.

There was evidence for the presence of HTLV-1 in one participant. This finding was subject to repeat testing, expanding SEPATH to the germline sample and aligning HTLV-1 classified reads to a HTLV-1 reference genome. This is also being independently validated to ensure that these results are not biased, particularly as a relatively small number of sequencing reads ($n=172$) was identified.

HTLV-1 is typically asymptomatic initially, emerging later in life⁴¹¹. If the participant is HTLV-1⁺, the infection may be silently spread through sexual encounters, via breast feeding⁴¹¹ (sample was breast tissue) or by blood transfusion⁴¹¹. Therefore, this diagnosis has larger implications than the original participant and contact tracing may prevent future cancers and other ailments⁴¹¹. There are efforts to reduce the global HTLV-1 burden, currently estimated at 10-20 million people⁴¹². There are limited treatment options for HTLV infections and therefore the main benefits from this diagnosis would be to allow for disease monitoring and the prevention of spread.

4.5.4 Cancer Viriome

The findings in this chapter regarding the tumour virome in-part reflect findings by Zapatka *et al.* 2020⁹⁶ which used a predecessor of SEPATH (without Kraken) as one of three approaches investigated ($N=2,658$ WGS and whole-transcriptome samples from 38 tumor types). In this chapter, evidence of viruses was found in approximately 10% of tumours while Zapatka *et al.* reports 13%. Alphapapillomavirus was associated with oral cancer in both datasets (excluding cervical cancer). *TP53* variants were associated with HPV⁻ samples in both studies. Zapatka *et al.* found that *CDK2NA*, *TERT* and *FAT1* driver mutations were associated with HPV⁻ cancer but this was not examined in the dataset. There are a few minor differences between the two studies: Here Roseolovirus and Lymphocryptovirus was distributed amongst multiple cancer types whereas in Zapatka *et al.* they were principally found in stomach/esophagus respectively; Gammaretrovirus was present in a variety of tumour types including breast and colorectal as opposed to mostly pancreatic in Zapatka *et al.* 2020; Orthohepadnavirus is mostly removed by SEPATH human depletion steps as mentioned previously. Additionally, this analysis used a 'white-list' of taxa due to the presence of bacteriophages and suspected contamination. Examining a wider range of viruses may reveal more findings.

4.5.5 Contaminants and Microbial Structure

This study is not a gold standard microbiome study. The RIDE¹⁸² criteria are not met: there are no negative controls, the experimental design is not focused on microbiome research and therefore the only investigation of contamination has been computational. It is not possible to compare biological samples to controls and so it is not possible to report on the full scope of the contamination and its impact on results.

The search for contaminant genera is a challenging task, particularly in an immunocompromised cohort where genera deemed contaminants may well be pathogenic or commensals not ordinarily able to survive in humans⁴¹³. Lists of common contaminants were covered by taxa informative of technical variables^{172,182} (appendix section D.12). The lists of putative contaminants produced here and available in the literature (Salter *et al.* and Eisenhofer *et al.*) include biologically relevant taxa (*i.e.*, *Bacteroides*). This reflects the impossibility to identify/remove all contaminant genera. Removing varieties of these putative contaminant genera however did not provide a more clearly defined community structure of cancer types. There are four possible reasons for this: there is no community structure by cancer type; it is just a peculiarity of this dataset; the optimal 'black-list' of putative contaminants has not been defined; removing putative contaminants removes biological signal. Figure 4.25 shows the best community structure I achieved, with colorectal and oral cancer being distinguishable from the rest. This could be expected based on prior work by the human microbiome project^{59,72}. It seems likely that the other biological sites contain too much noise in the form of contamination (or no microbial constituents) to discern.

The Genomics England dataset has less microbial reads than TCGA¹⁵⁴, PCAWG or in the simulated datasets (chapter 2). This could be hampering the ability to detect community structure. Technical differences in the sequencing technologies and bioinformatic analyses used could account for this difference, for example different library preparations or alignment to different human reference genomes. Another possibility is that the results in this chapter provide a more realistic view of the cancer microbiome than other studies.

SEPATH performed very well in the detection of HPV and was more discriminatory than the abundance analysis presented in Poore *et al.* 2020¹⁵⁴. The 100,000 Genomes Project is also a more recent project than the others, with a high minimum sequencing coverage (75X for tumour samples), standardised protocols and predominantly PCR-free and non-FFPE samples.

4.5.6 Modelling Cancer Type - Comparing with Poore *et al.*

Several flaws with Poore *et al.* were suggested including inappropriate normalisation, utilising sparse taxa, a lack of reasonable feature importance and likely mis-classified human reads. The claim made by Poore *et al.*¹⁵⁴ that the tumour microbiome is anatomically specific was based on model performance, without visual demonstration (with ordination) and without commenting on tumour type specific microbes. Model performance is impacted by sampling method, with other techniques less performant (figure 4.27). Class imbalance is a prominent issue in machine learning. Upsampling minority classes by such huge factors is likely to cause errors and overfitting. Such imbalanced class sizes also means that the model performance is often no better than the no information rate (null model, see model statistics made available by Poore *et al.*¹⁵⁴).

4.5.7 Searching for Clinical Associations

The proportion of positive cases for at least one ABBS genus was investigated for every clinical stage in every cancer type (figure 4.19). Some cancer types suggest a trend with an increasing clinical stage (bladder, endometrial and hepatopancreatobiliary cancer) but most do not. The bladder cancer link might be supported by research noting an enrichment of *Fusobacterium nucleatum* in bladder cancer cases compared to controls²⁸⁰. *Peptoniphilus* and *Porphyromonas* were found to be significantly enriched in endometrial cancer compared to control cases⁴¹⁴. Whilst this data does support these trends it needs to be backed up with more complete clinical data such as survival and cancer specific features (*e.g.*, PSA). Prostate cancer demonstrates the opposite trend to what might be expected (figure 4.19). The difference is that clinical stage is being investigated as opposed to progression in chapter 3. More data is needed to investigate this further and to be able to investigate the ABBS genera equally in differing cohorts. Associations between individual genera reported by Kraken and clinical metadata were also investigated (appendix section D.7) but ultimately hindered by the sparsity of the data, contamination and limited complete clinical data. It is likely that polymicrobial lists of taxa (perhaps capable of exerting shared functionality) will reveal better clinical associations in future.

4.6 Conclusions

Running SEPATH using Kraken on tumour samples from the 100,000 Genomes Project has revealed interesting genera but was hampered by contamination and the small number of unmapped reads relative to other cancer sequence datasets. Detecting Alphapapillomavirus demonstrates a high overlap with clinical tests suggesting that SEPATH has a utility on this cohort. One participant demonstrated evidence for HTLV which will be subject to independent validation and should be interpreted cautiously. Viriome investigations reflect the literature⁹⁶ with a few differences ascribed to human depletion and variation between cohorts.

The Kraken data was sparse and the clinical data was limited hindering some of the most interesting analyses. Additional analysis may become possible in future as more data is released *i.e.*, survival analyses using metastasis or cancer-specific survival as endpoints. The scope for this is large and could include producing prognostic models and finding associations with mutational signatures or expanding the mutational status of specific genes.

The models produced by Poore *et al.* along with previous work into the human microbiome project suggested the hypothesis that resolving contamination in this cohort might reveal tumour specific microbial community structure. Several flaws were identified in Poore *et al.* and the analysis here failed to reveal any major microbial structure by tumour type, other than colorectal and oral tumours. This might be expected as a disproportionate amount of reference genomes have been isolated from gastrointestinal studies.

4.7 Future Directions

In this chapter, I have investigated the taxonomic classifications of unmapped human-depleted sequencing reads with Kraken. The ability of Kraken to obtain classifications for reads is governed by its reference database. Therefore, there is the potential that the database used is skewed towards more highly researched organisms such as those that are pathogenic or those from more highly studied anatomical sites like the gut.

Assembly of unmapped reads from individual samples (as in chapter 2) often failed on this dataset. Perhaps this was assisted by the lower number of human-depleted reads observed in the Genomics England dataset. An alternative that could be investigated is by pooling unmapped human-depleted reads and performing metagenomic co-assemblies. To investigate this, I suggest the following workflow:

- Create metagenomic co-assemblies with MEGAHIT⁴¹⁵
- Merge multiple co-assemblies and remove redundancy by clustering contigs with CD-HIT⁴¹⁶
- Align non-human unmapped reads to pancancer assembly (*i.e.*, Kallisto³⁷⁷)
- Classify contigs where possible (*i.e.*, DIAMOND⁴¹⁷)
- Investigate sample level taxonomy assigned to assembled contigs

There are some caveats to this approach that should be noted. Co-assembling numerous samples from non-similar anatomical locations may create highly complex mixtures which can cause chimeric contigs⁴¹⁸. Recovery of high-quality MAGs (Metagenome Assembled Genomes) would therefore be challenging if possible at all. Furthermore, diversity between contigs would be impacted by the parameters used for CD-HIT and potentially masking closely related contigs. Comparing these investigations to the Kraken analysis might help to identify a potential bias by sampling site. Also, the rare taxa such as HTLV-1 would be unlikely to completely assemble in this analysis due to low coverage and would therefore be biased based on the prevalence of taxa in all samples.

4.8 Methods

4.8.1 100,000 Genomes Sample Collection and Processing

Samples for this project have been collected from multiple sites across England according to detailed protocols for sample collection and processing^{382,419}. Below is a brief summary of collection and processing methods. Germline samples are either peripheral blood, saliva or cultured fibroblasts (with a minimum of 3 passages). Tumour samples are predominantly fresh-frozen tissue and are obtained surgically and cut to 5x5x2mm. They can also be biopsies or peripheral blood in the case of haematological malignancies⁴¹⁹. Samples are assessed and optimised for tumour content and $> 40\%$ of the nucleated cells must be tumour and $< 20\%$ of cells must be necrotic. Samples must have $> 2\mu\text{g}$ of DNA for the preferred option of PCR-free sequencing. PCR amplification is permitted with a minimum DNA quantity of 500ng when it is not possible to obtain sufficient DNA (nano protocol⁴²⁰). Formalin-fixed paraffin-embedded (FFPE) tissue can be used in exceptional circumstances, but it is not preferred due to poorer quality results⁴²¹.

Fresh frozen samples are cut into adequately sized pieces, placed on foil, covered with OCT (optimal cutting temperature compound) and are immediately lowered into liquid nitrogen, dry ice, wet ice or sprayed with cryospray. Sample dissection took place in a clean environment with sterile equipment cleaned with 70% ethanol solution. Samples are then stored at -80°C until shipment to one of 13 NHS Genomic Medicine Centre (GMC)⁴¹⁹.

Samples are homogenised chemically (Qiagen tissue lyser) or “minced” with one or two scalpels prior to DNA extraction is carried out according to manufacturer’s instructions and the GMC laboratory SOP. DNA quantification is carried out using a validated approach such as Qubit, Picogreen and Glomax Quantifluor. The sample is then made up to a concentration of 20-60ng/ μl and a volume of 65-600 μl in Tris/EDTA DNA buffer and assessed to ensure 60% of fragments are above 23kb⁴¹⁹.

Samples are then processed by Illumina’s internal workflows to produce 150bp paired end sequencing reads on Illumina Hiseq X analysers. Illumina workflow V2 (HiSeq Analysis Software 2) and V4 (NorthStar4) were applied to each sample. Sequencing workflows consist of Illumina’s internally developed tools: Illumina iSAAC Aligner⁴²², iSAAC Variant Caller, iSAAC Copy Number Variant Caller (Canvas) and iSAAC Structural Variant Caller (Manta).

4.8.2 Identification of Mutational Status

Somatic small variants annotated VCF files were generated by Genomics England using the iSAAC Variant Caller. Non-synonymous variants of moderate/high impact were investigated in oral/oropharyngeal cancer participants using functions produced by Genomics England (01.functions.R) available within the research environment.

4.8.3 HTLV Classification

172 sequencing reads from tumour and germline samples with any classification of HTLV-1 (prior to confidence thresholding) were aligned to a HTLV-1 reference genome (NC_001436.1) using BWA mem¹⁰⁷. This alignment was viewed with IGV³⁹⁵ (version 2.9.4). The code for this alignment is available in section D.4. Briefly, BWA index of

HTLV-1 genome was made with the `bwa index` command, then the reads aligned with `bwa mem` (all default parameters, version 0.7.17-r1188). This was converted to a BAM file and sorted with `samtools` (version 1.9).

4.8.4 Viral Classifications

To limit contamination in the form of bacteriophages, only the following viral genera were considered: *Lymphocryptovirus*, *Orthohepadnavirus*, *Roseolovirus*, *Alphapapillomavirus*, *Cytomegalovirus*, *Gammaretrovirus*, *Alphatorquevirus*, *Lentivirus*, *Alphapolyomavirus*. This selection of viral genera was subject to *t*-SNE based on a distance matrix of Spearman's correlation values with parameters: random seed=8, dimensions=2, theta=0.2, perplexity=50, maximum iterations=5000. Clusters were determined by a density base clustering algorithm from the `dbscan` package⁴²³ (version 1.1.5, parameters: `eps`=3, minimum points=5).

4.8.5 Dimensionality Reduction

The sample-taxa kraken community matrix was read into R and a minimum number of 10 reads was required for classification. Samples were filtered to represent non-FFPE, PCR-free, primary tumour samples from cancer types: adult glioma, colorectal, lung, prostate, bladder, endometrial, malignant melanoma, renal, breast, haematological, oral, sarcoma, hepatopancreatobiliary, ovarian. Taxa with total counts across the entire dataset below 100 were removed from further analysis. The following suspected contaminant genera were removed from the matrix according to Salter *et al.* and as discussed throughout this chapter: *Homo*, *Acinetobacter*, *Alcaligenes*, *Bacillus*, *Bradyrhizobium*, *Herbaspirillum*, *Legionella*, *Leifsonia*, *Mesorhizobium*, *Methylobacterium*, *Microbacterium*, *Novosphingobium*, *Pseudomonas*, *Ralstonia*, *Sphingomonas*, *Stenotrophomonas*, *Xanthomonas*, *Sphingobium*, *V5virus*, *K1gvirus*, *Cc31virus*, *Kp15virus*, *T4virus*, *Felixo1virus*, *Afipia*, *Aquabacterium*, *Asticcacaulis*, *Aurantimonas*, *Beijerinckia*, *Bosea*, *Brevundimonas*, *Caulobacter*, *Craurococcus*, *Devosia*, *Hoeflea*, *Ochrobactrum*, *Paracoccus*, *Pedomicrobium*, *Phyllobacterium*, *Rhizobium*, *Roseomonas*, *Sphingopyxis*, *Acidovorax*, *Azoarcus*, *Azospira*, *Burkholderia*, *Comamonas*, *Cupriavidus*, *Curvibacter*, *Delftia*, *Duganella*, *Janthinobacterium*, *Kingella*, *Leptothrix*, *Limnobacter*, *Massilia*, *Methylophilus*, *Methyloversatilis*, *Oxalobacter*, *Pelomonas*, *Polaromonas*, *Schlegelella*, *Sulfuritalea*, *Undibacterium*, *Variovorax*, *Enhydrobacter*, *Enterobacter*, *Escherichia*, *Nevskiae*, *Pseudoxanthomonas*, *Psychrobacter*, *Aeromicrobium*, *Arthrobacter*, *Beutenbergia*, *Brevibacterium*, *Corynebacterium*, *Curtobacterium*, *Dietzia*, *Geodermatophilus*, *Janibacter*, *Kocuria*, *Micrococcus*, *Microthrix*, *Patulibacter*, *Propionibacterium*, *Rhodococcus*, *Tsukamurella*, *Abiotrophia*, *Brevibacillus*, *Brochothrix*, *Facklamia*, *Paenibacillus*, *Streptococcus*, *Chryseobacterium*, *Dyadobacter*, *Flavobacterium*, *Hydrothalea*, *Niastella*, *Olivibacter*, *Pedobacter*, *Wautersiella*, *Deinococcus*, *Acidobacteria*, *Toxoplasma*, *Mycobacterium*, *Candidatus Pelagibacter*¹⁷². The top taxa (1500) that were reported by 1-vs-all GBM models were combined and the intersect between this list and the taxa remaining in the community matrix after filtering was used further. This list of taxa was:

Acanthamoeba, *Acetanaerobacterium*, *Acetitumaculum*, *Acetivibrio*, *Acetobacter*, *Achromobacter*, *Acidaminococcus*, *Acidocella*, *Actinobacillus*, *Actinobaculum*, *Actinomyces*, *Actinotignum*, *Adlercreutzia*, *Aeribacillus*, *Aerococcus*, *Aeromonas*, *Agaricus*, *Aggregatibacter*, *Agrobacterium*, *Akkermansia*, *Alcanivorax*, *Alicyclophilus*, *Alicyclobacillus*, *Alistipes*,

Allisonella, *Allobaculum*, *Allofustis*, *Alloiococcus*, *Alloprevotella*, *Alloscardovia*, *Alphapapillomavirus*, *Alphapolyomavirus*, *Alphatorquevirus*, *Altererythrobacter*, *Alternaria*, *Aminobacter*, *Amycolatopsis*, *Anaerobiospirillum*, *Anaerobium*, *Anaerococcus*, *Anaerocolumna*, *Anaerofustis*, *Anaeroglobus*, *Anaeromassilibacillus*, *Anaerosalibacter*, *Anaerosphaera*, *Anaerosporebacter*, *Anaerostipes*, *Anaerotignum*, *Anaerotruncus*, *Anaplasma*, *Angelakisella*, *Anoxybacillus*, *Aquamicrobium*, *Arcanobacterium*, *Arcobacter*, *Asaccharospora*, *Asinibacterium*, *Aspergillus*, *Atopobium*, *Babesia*, *Bacteroides*, *Bariatricus*, *Barnesiella*, *Beauveria*, *Betapolyomavirus*, *Bifidobacterium*, *Bilophila*, *Bittarella*, *Blastococcus*, *Blastocystis*, *Blas-tomonas*, *Blautia*, *Bordetella*, *Botrytis*, *Brachybacterium*, *Brachymonas*, *Brachyspira*, *Butyricicoccus*, *Butyricimonas*, *Butyrivibrio*, *C2virus*, *Caecibacter*, *Caldicellulosiruptor*, *Caldimonas*, *Campylobacter*, *Candida*, *Candidatus Babela*, *Candidatus Kryptobacter*, *Candidatus Methanomethylophilus*, *Candidatus Microthrix*, *Candidatus Planktophila*, *Candidatus Stoquefichus*, *Capnocytophaga*, *Cardiobacterium*, *Carnobacterium*, *Castel-laniella*, *Catonella*, *Cellulomonas*, *Cellulosilyticum*, *Cellulosimicrobium*, *Centipeda*, *Cetobacterium*, *Chaetomium*, *Chitinophaga*, *Chlamydia*, *Christensenella*, *Chroococcidiop-sis*, *Citrobacter*, *Clavibacter*, *Clavispora*, *Cloacibacillus*, *Cloacibacterium*, *Clostridioides*, *Clostridium*, *Cohnella*, *Collimonas*, *Collinsella*, *Confluentimicrobium*, *Coprobacillus*, *Coprobacter*, *Coprococcus*, *Croceicoccus*, *Cronobacter*, *Cryptobacterium*, *Cryptosporidium*, *Culturomica*, *Cutibacterium*, *Cyberlindnera*, *Cyclospora*, *Cytomegalovirus*, *Debaryomyces*, *Dechloromonas*, *Dehalobacterium*, *Dehalococcoides*, *Deltapolyomavirus*, *Dermabacter*, *Dermacoccus*, *Desulfitobacterium*, *Desulfobulbus*, *Desulfococcus*, *Desulfomicrobium*, *Desulfosporosinus*, *Desulfotomaculum*, *Desulfovibrio*, *Dialister*, *Diaphorobacter*, *Dielma*, *Dokdonella*, *Dolosigranulum*, *Dorea*, *Duodenibacillus*, *Dysgonomonas*, *Edwardsiella*, *Eg-gerthella*, *Eggerthia*, *Eikenella*, *Eimeria*, *Eisenbergiella*, *Elizabethkingia*, *Enorma*, *Ensifer*, *Entamoeba*, *Enterococcus*, *Enterocytozoon*, *Enterorhabdus*, *Erwinia*, *Erysipelatoclostrid-ium*, *Erysipelothrix*, *Erythrobacter*, *Erythroparvovirus*, *Ethanoligenens*, *Eubacterium*, *Exiguobacterium*, *Exophiala*, *Ezakiella*, *Faecalibacterium*, *Faecalibaculum*, *Faecalitalea*, *Fenollaria*, *Filifactor*, *Finegoldia*, *Flavonifractor*, *Fomitiporia*, *Fontimonas*, *Fournierella*, *Frigoribacterium*, *Fron-dihabitans*, *Fusarium*, *Fusicatenibacter*, *Fusobacterium*, *Gabo-nia*, *Gabonibacter*, *Gammaretrovirus*, *Gardnerella*, *Gemella*, *Gemmiger*, *Gemmobacter*, *Geobacillus*, *Giardia*, *Giesbergeria*, *Globicatella*, *Glutamicibacter*, *Gordonia*, *Gordonibacter*, *Gracilibacillus*, *Granulibacter*, *Granulicatella*, *Gregarina*, *Haemophilus*, *Hafnia*, *Halomonas*, *Hammondia*, *Hapalosiphon*, *Helcococcus*, *Helicobacter*, *Herbiconiux*, *Hespellia*, *Histophilus*, *Hk578virus*, *Holdemanella*, *Holdemania*, *Holospira*, *Hungatella*, *Hydrogenophaga*, *Hy-menobacter*, *Hyphomicrobium*, *Immundisolibacter*, *Intestinibacillus*, *Intestinibacter*, *Intestinimonas*, *Jeotgalicoccus*, *Johnsonella*, *Jonquetella*, *Kaistia*, *Kallipyga*, *Kandleria*, *Klebsiella*, *Kluyvera*, *Kosakonia*, *Kp34virus*, *Kytococcus*, *Lachnoanaerobaculum*, *Lach-nobacterium*, *Lachnoclostridium*, *Lachnospira*, *Lactobacillus*, *Lactococcus*, *Lactonifactor*, *Lambdavirus*, *Lampropedia*, *Lautropia*, *Lawsonella*, *Leadbetterella*, *Leclercia*, *Leishmania*, *Lelliottia*, *Leptotrichia*, *Leuconostoc*, *Levyella*, *Libanicoccus*, *Limnohabitans*, *Listeria*, *Luteimonas*, *Lymphocryptovirus*, *Lysinibacillus*, *Lysobacter*, *Mageeibacillus*, *Mailhella*, *Malassezia*, *Mannheimia*, *Marinobacter*, *Marmoricola*, *Marvinbryantia*, *Mastadenovirus*, *Mediterranea*, *Megamonas*, *Megasphaera*, *Meiothermus*, *Merdibacter*, *Methanobacterium*, *Methanobrevibacter*, *Methanomassiliicoccus*, *Methanosphaera*, *Methylibium*, *Methylobacil-lus*, *Methylocaldum*, *Methylococcus*, *Meyerozyma*, *Micromonospora*, *Microvirga*, *Mitsuaria*, *Mitsuokella*, *Mobiluncus*, *Modestobacter*, *Moesziomyces*, *Mogibacterium*, *Mooglevirus*, *Moraxella*, *Morganella*, *Morococcus*, *Mucilaginibacter*, *Mucispirillum*, *Murdochiella*, *My-coplasma*, *Myxococcus*, *N4virus*, *Nakaseomyces*, *Ndongobacter*, *Negativicoccus*, *Neglecta*,

Neisseria, Niameybacter, Nitratedreductor, Nitrobacter, Nitrosomonas, Nitrospira, Nocardia, Nocardioidea, Nonagvirus, Nosocomiicoccus, Nostoc, Obesumbacterium, Oceanobacillus, Odoribacter, Oenococcus, Oerskovia, Oligella, Olsenella, Oribacterium, Ornithobacterium, Oscillibacter, Ottowia, Paenarthrobacter, Paeniclostridium, Paenirhodobacter, Pandoraea, Pannonibacter, Pantoea, Parabacteroides, Paraburkholderia, Parageobacillus, Paraprevotella, Parascardovia, Parasutterella, Parvibaculum, Parvimonas, Pasteurella, Pectobacterium, Pediococcus, Pelagibacterium, Penicillium, Peptoanaerobacter, Peptococcus, Peptoniphilus, Peptostreptococcus, Petrimonas, Phaeobacter, Phascolarctobacterium, Phenyllobacterium, Phocaeicola, Phoea, Phoenicibacter, Plantibacter, Plasmidium, Plesiomonas, Pluralibacter, Porphyrobacter, Porphyromonas, Prevotella, Prevotellamasilia, Propionimicrobium, Proteus, Providencia, Pseudoalteromonas, Pseudobutyrvibrio, Pseudoflavonifractor, Pseudogymnoascus, Pseudohongiella, Pseudonocardia, Pseudopropionibacterium, Pseudoramibacter, Pseudorhodofera, Psychrobacillus, Purpureocillium, Pyrenophora, Rahnella, Ramlibacter, Raoultella, Reyranelia, Rhizobacter, Rhizorhabdus, Rhodobacter, Rhodopseudomonas, Rhodotorula, Rickettsia, Rikenella, Rodentibacter, Romboutsia, Roseburia, Rosenbergiella, Roseolovirus, Rothia, Rtpvirus, Rubellimicrobium, Rudaea, Ruminiclostridium, Ruminococcus, Saccharomyces, Salmonella, Sanguibacter, Sarcina, Scardovia, Schwartzia, Sediminibacterium, Selenomonas, Senegalimassilia, Sep1virus, Septima3virus, Serratia, Seuratvirus, Sextaecvirus, Sfi21dt1virus, Sharpea, Shewanella, Shigella, Shinella, Shuttleworthia, Silanimonas, Simonsiella, Simplexvirus, Sinorhizobium, Sk1virus, Slackia, Sneathia, Solobacterium, Sordaria, Sp18virus, Sp6virus, Sphaerotilus, Sphingobacterium, Spirosoma, Spn3virus, Sporanaerobacter, Sporosarcina, Staphylococcus, Starkeya, Stomatobaculum, Streptobacillus, Streptomyces, Subdoligranulum, Succinatimonas, Succiniclasticum, Succinivibrio, Sulfitobacter, Sulfurospirillum, Suspivirus, Sutterella, Synechocystis, Synergistes, Syntrophomonas, T5virus, T7virus, Tannerella, Tepidimonas, Tepidiphilus, Terrisporobacter, Tessaracoccus, Tetracenococcus, Tetrasphaera, Thauera, Thermoanaerobacter, Thermoanaerobacterium, Thermobacillus, Thermomonas, Thermus, Thioalkalivibrio, Thiomonas, Tolypothrix, Treponema, Trichosporon, Tropheryma, Tropicimonas, Trueperella, Turicella, Tyzzerella, Ureaplasma, Ureibacillus, Vaginella, Veillonella, Vibrio, Weissella, Wickerhamomyces, Williamsia, Xenophilus

A distance matrix (Spearman's) of these 495 taxa was created and *t*-SNE ran on this with parameters: `perplexity=80, verbose=TRUE, max_iter=2000, check_duplicates=TRUE`.

t-SNE plot of Kraken results coloured by tumour type for the microbial genera that most informative of tumour type. *t*-SNE was carried out on a matrix of Spearman's correlation values between samples. Informative taxa are defined as those taxa that were in the top 1500 most important taxa in one-vs-all GBM models. 495 unique taxa met this criteria in 8103 non-FFPE, PCR free, primary tumour samples further details available in 4.8.5. The script used to conduct this analysis can be found in appendix section D.10.

4.8.6 Re-analysing the Poore *et al.* Manuscript

The raw Kraken TCGA results presented by Poore *et al.* 2020¹⁵⁴ were reanalysed (available from ftp://ftp.microbio.me/pub/cancer_microbiome_analysis/). Taxa that existed in TCGA but not in SEPATH on Genomics England could not be compared and may originate from misclassified reads, so the TCGA data was only interrogated for the taxa reported by SEPATH. TCGA samples were filtered to only primary tumours that were sequenced with Illumina HiSeq technology and aligned to HG19 (human reference genome

37). All TCGA samples utilised RNA sequencing technology and no samples originated from FFPE tissue. All samples with no positive taxa were removed and all taxa with < 20 positive samples were filtered out of the dataset. The resulting dataframe had 8,669 samples with 451 different genera.

4.8.7 Supervised learners

The `voom` function was used with default settings as implemented in the `limma` package (version 3.36.5)⁴²⁴ on sample-taxa community matrix of Kraken results.

Scripts were adapted from Poore *et al.* to produce gradient boosted machine learners¹⁵⁴. Test-train splits were produced with the `createDataPartition` function of the `caret` package (version 6.0.86)⁴²⁵. 70% of the data was retained for training sets taking into consideration the disease type. To limit issues caused by unequal sample sizes (for analysis on the Genomics England dataset) tumours with small sample sizes were removed including: carcinoma of unknown primary, endocrine, nasopharyngeal, other, and testicular. Models were trained using the `train` function of the `caret` package with parameters: `method='gbm'`, `preProcess = c('scale', 'center')`, `verbose=TRUE`, `metric='ROC'`, `tuneGrid = caretTuneGrid` where arguments for `caretTuneGrid` were `interaction.depth=seq(1,3)`, `n.trees=floor((1:3)*50)`, `shrinkage=0.1`, `n.minobsinnode=5`. The `trainControl` was set with parameters: `method='repeatedcv'`, `number=3`, `repeats=2`, `summaryFunction=twoClassSummary`, `classProbs=TRUE`, `verboseIter=TRUE`, `savePredictions=TRUE` and `allowParallel=TRUE`. The upsampling strategy was set to 'up', 'ROSE'⁴⁰¹ or 'SMOTE'¹⁴⁹.

Chapter 5

Discussions, Future Work and Concluding Remarks

These final remarks summarise each chapter and expand upon cross-chapter discussion before concluding and outlining future work.

5.1 Chapter Summaries

Developing SEPATH

Chapter 2 outlines the development of SEPATH. Various computational approaches were performed on simulated and real low-biomass host-dominated metagenomic data. Kraken and mOTUs2 delivered the highest performance in terms of classifying metagenomic sequencing reads with median F1 scores at genus level of 0.90 (IQR=0.083) and 0.91 (IQR=0.10) respectively. With median PPV scores of 0.97 (IQR=0.084), 0.95 (IQR=0.080) and median sensitivity scores of 0.86 (IQR=0.123), 0.88 (IQR=0.126) for Kraken and mOTUs2 respectively. This analysis also suggested assembling sequencing reads could provide a performant approach with reduced requirements for read thresholds (genus level F1 score of 0.83). The final pipeline adopted in SEPATH begins with unaligned read extraction from an alignment to a human reference genome, followed by additional host depletion, quality trimming and taxonomic classification with Kraken and mOTUs2. Kraken was good at identifying viruses (8/10 detected correctly), but human depletion removed some viruses. Testing the classification of reads with Kraken and mOTUs2 on a pilot dataset of real cancer tissue revealed genera in line with published literature ($N=30$, with 10 samples for each of medulloblastoma, cervical cancer and gastric cancer). This included detection of *H. pylori* in gastric cancer and Alphapapillomavirus in cervical cancer.

Microbes of Prostate Cancer Exosomal Urine and Potential Microbe-Host Interactions

Chapter 3 investigated the metatranscriptomic composition of urine from men under investigation for prostate cancer. SEPATH with Kraken identified a wide range of genera. The five genera with the highest median number of classified reads were *Porphyromonas*,

Mobiluncus, *Dialister*, *Campylobacter* and *Varibaculum*. Many of the bacterial genera reported by this analysis are found in urogenital flora^{292,293}. Patients with at least one ABBS genus present in their urine (*Ezakiella*, *Peptoniphilus*, *Porphyromonas*, *Anaerococcus*, *Fenollaria* and *Fusobacterium*) had a statistically significant quicker time to disease progression ($p=0.0079$). Dimensionality reduction of this community data suggested the samples could be divided into two clusters. Differential expression and pathway analysis of human genes between the two clusters showed prominent differences in immunological gene expression including *IGHA1*, *IGKC*, *CCL5* and *CD74*. Investigating time to progression between these groups suggested that samples from the cluster containing more bacteria had a worse disease trajectory, but this was not statistically significant ($p=0.21$). Patients with at least one ABBS genus were almost entirely a subset of the cluster which contained more bacterial genera. To formulate hypotheses regarding a possible link between ABBS and poor prognosis prostate cancer, InterProScan was applied to isolates from participants under investigation for prostate cancer. Isolates from the ABBS genera appeared functionally related with a preference for preference for FAD-TS. This also lead to the hypothesis that ABBS bacteria leads to an inflammatory recruitment of MDSC and dampening anti-tumour immunity.

The Application of SEPATH to Cancer Whole Genome Sequence Data from Genomics England's 100,000 Genomes Project

Chapter 4 described the application of SEPATH to cancer sequences from Genomics England's 100,000 Genomes Project. The resulting community matrices were sparse, particularly for results obtained from mOTUs2. Kraken returned fewer classified reads per sample on the Genomics England datasets than from other similar datasets. Machine learning approaches were employed in an attempt to identify putative contaminant genera. Many of the output genera were also reported as common contaminants in literature but removing these taxa did little to resolve overall pan-cancer microbial structure.

The tumour virome was largely in agreement with previously published studies⁹⁶. There was a near perfect performance in SEPATH identifying clinically diagnosed HPV cases in oral cancer. Clinically actionable viruses including HIV, HBV, HCV, and HTLV-1 were investigated. Only one participant had any evidence of these viral pathogens (HTLV-1). This pathogen was identified in the participants germline and tumour sample and the classified sequencing reads were shown to align uniquely to HTLV-1 and across the entirety of a reference genome.

Overall pan-cancer microbial structure did not show the separation by tissue type that I expected. Colorectal and oral cancer samples were distinguishable from other tumour types, but the remaining tumour types were indistinguishable. One-vs-all tumour models were carried out as in Poore *et al*¹⁵⁴. This demonstrated a high performance in determining colorectal cancer samples from all others as suggested (F1-score=0.91). The next best performance was shown in predicting sarcoma and oral cancer (F1 = 0.49, 0.48 respectively). Examination of the Poore *et al.* manuscript revealed limitations in their approach such as inadequately accounting for host contamination, the normalisation methods used and handling of unequal class sizes.

5.1.1 Thesis Aims

The aims of the project were to:

1. Develop computational pipelines to classify microbes in cancer whole genome sequence data (achieved in chapter 2).
2. Apply these pipelines to a simulated benchmarking dataset (achieved in chapter 2).
3. Carry out pilot studies in a range of cancer types (achieved in chapter 2).
4. Apply these pipelines to various datasets (achieved in chapters 3 and 4).
5. Determine the landscape of microbes detectable in tumour sequence data (investigated in chapter 4).
6. Search for associations between microbes and metadata (chapters 3 and 4)
7. Generate hypotheses regarding associations between microbes and cancer (chapter 3).
8. Liaise with cancer-type specific scientists to design experiments to verify findings.
9. Investigate the metagenomic assembly of cancer sequence data and examine recurrent contigs that do not map to known microbial sequences.

Most of my time in this studentship was spent on aim five and this was partially met (further discussions in section 5.2.2). Regarding aim eight, data has been shared with researchers in Genomics England (breast and oral domain) and within my laboratory group. Prostate cancer specific research will continue within our own laboratory group with intentions to approach scientists with knowledge in immunology to help confirm some of the immune specific theories.

In the development of SEPATH, a high performance of taxonomic classification was obtained by assembling unmapped reads from individual samples. Running SEPATH on Genomics England data resulted in fewer than expected non-human reads which hindered metagenomic assembly of individual samples. There is scope for future research to investigate the metagenomic assembly of pooled sequencing reads which may provide sufficient sequencing data for assembly.

5.2 Discussions

5.2.1 The Applications of SEPATH

I produced an optimal pipeline for the detection of microbes in host-dominated low-biomass whole genome sequencing data by benchmarking multiple tools on simulated metagenomic and host datasets. This approach was validated by application on real data and the successful identification of Alphapapillomavirus in cervical cancer and *H. pylori* in stomach cancer (figure 2.10). High performance was obtained using Kraken and mOTUS2. Other manuscripts that have analysed cancer WGS metagenomics have also used Kraken^{154,426}. This is the first time that a comprehensive benchmark of taxonomic classification tools has been performed in host-dominated whole genome sequencing data such as cancer studies. This analysis was non-exhaustive compared to the number of possible approaches. A community driven focus such as CAMI⁹⁸ with a focus on cancer metagenomics would be an ideal approach to benchmarking future efforts.

The high performance of SEPATH (using Kraken) was also observed when it was applied to RNA sequencing data from urine samples in chapter 3. A range of taxa were detected that are known to be found in urine or the urogenital tract. The application of SEPATH

to Genomics England WGS data returned some expected results. Colorectal tumours had the largest median number of classified reads (figure 4.3). The most prevalent genera were *Bacteroides*, *Parabacteroides*, *Blautia*, *Alistipes* and *Clostridium* and the genera with the most reads were *Bacteroides* and *Fusobacterium*. This is in-line with what is expected from the colorectal cancer microbiome⁴²⁷. The oral microbiome also returned expected results with *Prevotella*, *Fusobacterium*, *Veillonella*, *Actinomyces* and *Gemella* as the most prevalent genera⁴²⁸. There was also a high performance in detecting Alphapapillomavirus (figure 4.17). This suggests that SEPATH has general utility in host-dominated WGS data within the limitations of the datasets. SEPATH detected a low number of microbes in the Genomics England dataset compared to our simulated datasets and other genomic studies. The results matrix is exceptionally sparse, particularly in non-colorectal cancers. In this situation I expect the performance of SEPATH to be lower and the amount of noise to increase. It is not clear why the Genomics England cohort would return such a low number of classified sequencing reads as the amount of sequencing data collected is large (see chapter 4).

5.2.2 The Search for Pan-Cancer Microbial Structure

Comprehensively determining the landscape of microbes in the Genomics England dataset was hindered by the lack of microbes detected. This made nearly all investigations challenging. Uncertainty between contaminant and pathogenic/commensal genera (a single genera which may contain pathogenic, commensal and contaminant species) makes reporting on genera speculative.

I focused on developing a heuristic approach to identifying contaminant genera which suggested many taxa that are known contaminants (section D.12). However, distinctions between tumour types did not significantly improve with the removal of these taxa. Without a study designed for microbial detection, developing a heuristic approach to dealing with contamination is very difficult if not impossible. Biologically informative genera are likely to have been suggested as important in model performance and therefore would be marked as a contaminant. There is therefore a delicate balance in considering what taxa to permit in a community matrix. Ubiquitous microbes such as *Mycobacterium*, *Toxoplasma* and *Candidatus Pelagibacter* detected in nearly every sample in the Genomics England dataset may represent contamination (chapter 4). *Mycobacterium* and *Toxoplasma* also have pathogenic potential^{429,430} which complicates the complete removal of these taxa from further consideration.

The human microbiome project⁷² and earlier work⁵⁹ found that the microbial community is specific to the sampling site (gastrointestinal, urogenital, skin, nasal and oral sites). The first two principal components published in 2012⁷² explain 17.4% of the total variation in the dataset which provides evidence for distinguishing microbial features. This suggests that tumours arising from different sampling sites should also boast unique microbial compositions as suggested in more recent publications. The manuscript by Nejman *et al.* is more descriptive than Poore *et al.* regarding the taxa that distinguish between tumour types. Nejman *et al.* investigated 1,526 samples from 7 cancer types (breast, bone, glioblastoma, lung, ovarian, pancreas, melanoma as well as paraffin and negative controls). I view their results with skepticism. They claim that breast cancer boasts the greatest number of species per sample. They have attempted to deal with the issue of contamination by sequencing negative controls, but I still believe they have identified some contaminant genera includ-

ing *Sphingomonas*, *Pseudomonas*, *Enterobacter*, *Corynebacterium* and *Novosphingobium*. *Streptophyta* was significantly associated with triple negative breast cancer, but this is a clade of plants. Despite this, Nejman *et al* has identified some logical microbial pathways, for example identifying a link between lung samples from smokers and phenol or nicotinate degradation. Poore *et al.* produced some high performing models with some significant flaws and therefore require further investigation for any potentially important features. Both Poore *et al.* and Nejman *et al.* have hinted towards tumour-type specific bacteria but failed to fully demonstrate pan-cancer microbial structure that is split by tissue type.

A company called Micronoma was established by the authors of Poore *et al.* that has recently secured \$3 million in seed funding⁴³¹. They have “exclusively licensed the original intellectual property on microbial-based cancer diagnostics and therapeutics, created by Poore and Knight, from the University of California San Diego”. In a recent quote on their website, they claim⁴³¹:

*“Micronoma’s technology has shown that distinct cancer types can be diagnosed sensitively and specifically solely using microbial nucleic acids in human blood and tissues. This approach was first demonstrated using a cancer sequencing database of >10,000 patients from 32 cancer types, then validated on >150 real-world patient samples from Moore’s Cancer Center at UC San Diego Health, across three cancer types and comparing results to those from healthy, non-cancer controls. The results of the study were published in Nature in March 2020”*⁴³¹

Whilst I remain optimistic regarding the future of this field, I do believe that there are several fundamental flaws with the approach presented by Poore *et al.*¹⁵⁴ (see section 4.4.10). This is somewhat of a juxtaposition, but perhaps more discourse would help initiate a move from the ‘black box’ machine learning approaches into hypotheses that can be empirically investigated.

5.2.3 The Search for Clinical Associations

Clinical Associations in Prostate Cancer

The largest potential impact of this work on future cancer patients is the identification of translational clinical associations which could eventually be adopted into clinical practice. The most promising associations presented in this thesis resulted from analysing urine samples from patients under investigation for prostate cancer (chapter 3). We developed an anaerobic bacterial biomarker set using a combination of high-quality biological data and clinical data. This clinical data included: biochemical recurrences, metastases, cancer associated deaths and a long follow up period. This allowed us, as a team, to discover an important clinical association between microbes and progression and formulate some fascinating hypotheses about how bacteria interact with tumours in aggressive prostate cancer. A benefit of using RNA sequencing is that SEPATH could be used to investigate metatranscriptomics, concurrently with Kallisto³⁷⁷ to investigate differential expression of human genes. This helped to formulate various hypotheses that might link bacteria, aggressive prostate cancer and the mechanisms by which that might occur. This is a first step and extensive mechanistic evidence will be required to prove a robust link between microbes and cancer. The work presented in this thesis has led to grant funding to interrogate a larger dataset of prostate cancer specific sequences which includes whole genome and RNA sequencing samples (PPCG, Pan Prostate Cancer Group). The PPCG will generate a wealth

of genomic information (such as variants, mutational spectra) and clinical data that can be analysed alongside microbial data.

The evidence for these associations originates from investigating time-to-event data and partitioning samples by ABBS genera detection on multiple datasets: urine RNA sequencing, the original 16S cohort and a separate targeted alignment approach (appendix C.5). The equivalent time to event data is currently not available in Genomics England meaning that an equivalent investigation on this cohort is not possible. A recent version of the clinical data release (Version 10) has zero recorded pre-treatment PSA results for prostate cancer patients. 43 participants have any measurement of PSA whatsoever (7.5% of the total 574 prostate cancer participants listed in the most recent data). Applying SEPATH data to Genomics England prostate cancer participants did not reveal any association between ABBS positive participants and clinical grade (figure 4.19).

Feng *et al.* 2019 investigated microbial composition of WGS data from 22 men and did not find an association with clinical risk²⁶⁶. Ma *et al.* 2019 investigated the microbial content of prostatic fluid (16S sequencing) of 59 participants ($n=32$ vs 27 with and without prostate cancer respectively). They found that diversity was different in participants with and without cancer but state that there were no ‘idiosyncrasies in the microbial species’ suggesting that they did not find taxa responsible for this difference in diversity⁴³². Shrestha *et al.* 2018³¹¹ investigated the urinary microbiome in men with positive versus negative biopsies for prostate cancer and found differences. They did not however report on any links with the development of aggressive prostate cancer. Several other sources have collated evidence of bacteria potentially driving the acute and chronic inflammation that likely contributes towards prostate cancer tumorigenesis^{433–435}. Porter *et al.* 2018²⁹⁷ carried out a literature review assessing some of this evidence and suggested that microbial constituents at multiple anatomical sites (including the urinary tract, gastrointestinal tract, oral cavity and the prostate) might play an important role in prostate health.

The work in the prostate specific chapter in this thesis (chapter 3) is descriptive of microbes associated with aggressive cancer. Validation of the association between ABBS genera and progression in prostate cancer patients is needed in independent cohorts beyond the East of England and the UK.

Clinical Associations within Genomics England Datasets

Clinical investigations in Genomics England have not been as fruitful. This is due to the limited number of microbial reads detected and the Kraken sparsity along with limited/sparse clinical data and genomic metrics. For example, 84% of mutational signature results in the “cancer analysis” (dated 23rd September 2020) contain ‘NA’ values. If looking at a subset of samples positive for a particular taxon, it becomes highly likely that there is insufficient non-NA values to make a statistically informative comparison. Only 7.7% of participants with a microbial result had demographic information such as weight, height, and smoking status. Of these 51% have no value for height and similar proportions are seen for weight and smoking status. In data release version 11 (released 17 Dec 2020, after the analysis in this thesis), 179 participants have developed metastases (186 total samples, 2% of all that are PCR free, non-FFPE and have a SEPATH result). Six cancer types have > 10 metastatic samples (colorectal 65, ovarian 23, hepatopancreatobiliary 14, renal 14, breast 13, prostate 13). These number of events are too small to have a reasonable statistical

power.

Even when considering prevalence across the entire dataset the combination of limited meta-data and a sparse taxonomic matrix means there is a lack of statistical power. To illustrate this point, I investigated whether there was any significant associations between the microbes detected in primary tumours of each cancer type and metastases status with 9,703 Fisher's exact tests for all taxa (≥ 10 reads minimum, data in Genomics England research environment). Of the 9,703 tests, 162 tests were significant ($p < 0.05$), reducing to only eight after FDR adjustment ($Q < 0.05$). Seven of these tests were for colorectal cancer and one was for prostate cancer. Of these eight significant tests, only one taxon demonstrated a higher proportion in primary tumour samples from participants that developed metastases, *Enterococcus*, which was present in 92% of primary samples from patients that developed metastases compared with 35% of samples that did not develop metastases. *Enterococcus* was one of the most prevalent microbial genera across the entirety of SEPATH results in Genomics England and so it is difficult to ascertain if this result is indicative of contamination. Additionally, the results in the prostate chapter (chapter 3) did not support this taxon as being associated with high risk prostate cancer. It is likely that the state of metadata in Genomics England improves, more associations will be found.

The analysis in chapter 3 suggests that it is the presence of particular polymicrobial groups that are associated with aggressive disease rather than a single genus. Machine learning approaches like random forests that iteratively select different features might help to identify polymicrobial lists in future. Machine learning methods for high dimensional data are scarce in metagenomics, particularly approaches capable of handling sparsity¹²¹. As discussed in chapter 4, inappropriate machine learning practices do little to further the field of investigating microbes associated with cancer. The ability to find taxa associated with aggressive cancer is a product of a few key factors: Rationale to reduce the number of taxa, whether based on differential abundance, prevalence, community structure, a review of relevant literature, or a combination of all of these; the metadata should be impactful and complete. Machine learning techniques can be useful if used appropriately and the sensible use of domain knowledge is critical.

5.2.4 The Search for Functional Pathways

Detecting the metabolic pathways that bacteria have the capacity to use from sequence data is becoming increasingly possible^{164,166}. Chapter 3 interrogated functional potential in two ways: the identification of pathways associated with differentially expressed human genes and investigating pathways in bacterial assemblies. This has highlighted several pathways, both human and bacterial that might be important in the development of aggressive prostate cancer and would benefit from additional experimentation.

Most differentially expressed human genes were immunological, which led to the hypothesis that infection marks a transition from anti-tumour immunity to anti-microbial immunity via the recruitment of MDSCs, most likely via an upregulation of host *CCL5*, *CCL2* and *CD74* (figure 3.12)⁴³⁶. Investigating different pathways between ABBS and non-ABBS bacteria isolated from the urine of prostate cancer patients highlighted the bacterial pathway FAD-TS (figure 3.9). It is not immediately clear how these pathways might influence the development of aggressive prostate cancer.

5.2.5 Limitations

This section discusses the main limitations of the work presented in this thesis.

SEPATH Limitations

The benchmarking of tools (chapter 2) was undertaken on simulated metagenomic data. Finding an appropriate software for metagenomic simulation was a challenge. Other simulation tools were trialled, including GemSIM⁴³⁷ and Grinder²¹⁸, which were too slow to achieve the expected quantity. Installing BEAR²¹⁹ was also a challenge as the software was not functional ‘out-of-box’. There have been several developments in metagenomic simulations since the analysis for chapter 2 was conducted. New tools have been released that might offer better solutions to BEAR including CAMISIM²²⁰ and InSilicoSeq⁴³⁸. The difficulty of installing computational tools and packages was the reason why I developed a singularity for SEPATH. Containerised software aids reproducibility but this work is not completely reproducible by the public due to the large size of the Kraken database and not being permitted to share a file used as part of human depletion obtained from COSMIC²⁰³.

The development of SEPATH would have benefited from a human/microbial sequencing dataset generated in the laboratory with known concentrations of mammalian and microbial biomass. This would have helped to overcome some of the limitations of testing simulated data. A further limitation of the SEPATH pipelines is that they need to be adapted to each type of computing cluster and so require expert knowledge of Snakemake pipelines²¹⁶.

Limitations of the Genomics England and other datasets.

The datasets utilised in this thesis are not designed for microbial studies. There is a large issue with contamination which has not been resolved in this thesis. Contamination is a complex issue in microbial bioinformatics which requires further study. The RIDE criteria¹⁸² (section 1.2.8) cannot be followed by analysing these datasets. The experimental design has not been designed for microbial detection and there are no negative controls. At most this thesis has had to speculate on what taxa are contaminants based on publications that have sequenced and classified nucleic acids in negative controls^{172,182}. These lists of taxa include potentially biologically relevant taxa like *Escherichia* that would be beneficial to investigate. The blanket removal of these taxa in a ‘black-list’ likely removes biological signal. Equally these taxa should not be retained in a community matrix because they likely impact the overall profile observed in a sample. There is no optimal solution to investigating or resolving the issue of contamination used in these datasets but I feel the removal of (known) common sequencing contaminants was the best approach available under the circumstances. Analysis of communities in this thesis was impacted by contamination and false positive classifications. Fortunately, in the case of chapter 3, there were differences in community structure that did not seem to be impacted by contamination.

Analysis of single taxa in the Genomics England dataset was not possible due to sparsity and the overall number of microbial reads was low compared to other datasets. There is a potential to use this data to distinguish between gastrointestinal and non-gastrointestinal associated tumours. Perhaps this could have a clinical utility in sarcoma where retroperitoneal and truncal associated sarcomas demonstrate a worse prognosis⁴³⁹. This would need

to be investigated and would require highly detailed and descriptive anatomical data of the tumours from Genomics England.

Ultimately, without non-cancerous control tissue samples and longitudinal studies it is impossible to distinguish the taxa that are associated with cancer. Future studies should be designed specifically for microbial detection with adequate controls and potentially chemical depletion of host DNA before sequencing⁴⁴⁰. This is very unlikely to happen with Genomics England. It was hoped that identifying associations between taxa and clinical data would provide evidence of those taxa as being associated with cancer. Whilst this has been true in part for the prostate chapter (chapter 3), there have been a lack of clinical associations in Genomics England data.

5.2.6 Potential Impact

Despite the limitations discussed above, there is scope for impact from the findings in this thesis. SEPATH provides a benchmarked containerised approach that could be used by other researchers to investigate microbial nucleic acids. The ABBS genera may have an impact by offering a non-invasive prognostic test for prostate cancer. Releasing the results of SEPATH on Genomics England may assist researchers investigating contamination by providing classifications from samples that are mostly expected to be sterile. Running SEPATH on Genomics England may hold some translational benefit by identifying HTLV-1 in one participant and detecting Alphapapillomavirus with a high performance.

The association between ABBS genera and aggressive prostate cancer holds the clearest potential route of impact. Chapter 3 identified evidence for ABBS genera in 14 pre-diagnosis urine RNA sequencing samples of a total of 25 urine samples from participants that were confirmed to have prostate cancer ($N=40$ total with or without cancer, figure 3.7). Of those 14 ABBS positive samples, 11 demonstrated signs of progression (85% of all 13 progression events in the cohort). This association therefore could have an impact on the nearly 12,000 men that die every year in the United Kingdom as a result of prostate cancer²³⁷. These findings may help reduce the economic impact of prostate cancer, estimated at £136,278,237 annually^{441,442}. Although the samples from this study were pre-diagnosis, the work in this thesis does not suggest a role for diagnostic utility of the ABBS genera but does suggest a use in prognostics when a cancer has already been diagnosed. Therefore, any subsequent investigations into the application of these genera would need to be in addition to the current standard of clinical investigations for prostate cancer which can vary between countries.

5.3 Future Work

5.3.1 Future Work on Genomics England

Associations between cancers and microbes will be found in the future from the Genomics England dataset as the depth and completeness of metadata improves. I did not see the separation of tissue type in the pan-cancer microbial structure which has been observed in recently published manuscripts. Therefore, there needs to be a more in depth investigation into this (see section 4.7). One avenue is to investigate metagenomic assembly of pooled non-human reads from cancer samples within the 100,000 Genomes Project. Further investigating assembled contigs may lead to the discovery of novel species^{443,444}.

The Genomics England dataset contains whole genome sequence data from 40,000 blood samples from patients with rare diseases. There is the possibility to explore whether microbes detected in these participants are associated with the aetiology of idiopathic conditions. A project has since been established within Genomics England to extend the search for microbes into their rare disease cohort. One limitation of this is that each rare disease type will not have a large sample size.

5.3.2 Pan-Prostate Cancer Group Investigations

The work presented in this thesis has led to a successful grant application for an 18-month post-doctoral post. The objectives for this position are as follows.

Objective 1: Gathering data and quantifying the microbial population present in tumour (0-9 months)

WGS and RNA sequencing data will be collected from the first data release of The Pan-Prostate Cancer Group (PPCG) that contains 2,012 cancer samples (1,703 primary, 303 metastatic). Clinical and histopathology information will also be collected. Other secondary outputs such as somatic & germline variants, mutational signatures, and predicted fusion genes will also be obtained. SEPATH will be applied to these samples. The application of SEPATH to Genomics England data was somewhat sparse and it is not yet clear if this was a peculiarity of the Genomics England dataset or if this was a more global feature of prostate tissue. Therefore, contingent approaches will also be investigated such as the targeted alignment approach (section 3.2.2) or by advancing the metagenomic assembly approach from within Genomics England.

Objective 2: Validate and determine novel clinical and pathological associations (9-15 months)

In this objective, the post will validate findings presented in chapter 3 regarding anaerobic bacteria and aggressive prostate cancer. The increased size of the dataset holds higher statistical power that may be sufficient to find additional associations.

Objective 3: Determine how widespread the bacteria associated with tumours travel (10-16 months)

The PPCG dataset contains metastatic tissue from 303 men, 60 of which have a paired primary sample. With this comes the possibility to investigate relationships between primary and metastatic tissue. The post will examine whether ABBS genera are associated with metastatic tissue or if this is purely a feature of primary tissue. As many of these bacteria are intracellular in nature there is the possibility for bacteria to spread with metastatic cells¹⁷³.

Objective 4: Characterising the molecular interaction between microbes and tumour tissue (12-18 months)

Chapter 3 suggested some molecular interactions between microbes and prostate cancer. However, this was partly based on differentially expressed human genes in exosomal RNA. It may be that exosomal RNA has shown to be a better resource for biomarkers, but these findings might not be reflected in prostate tissue. The potential functionality of microbes within PPCG tissue data will therefore be examined using the same approaches outlined in chapter 3. There is rationale for the ABBS genera to be investigated alongside other features

(genomic, transcriptomic and clinical variables) in the development of a prognostic test for prostate cancer. There is preliminary evidence in this thesis to suggest that the presence of bacteria might correlate with a specific immunological expression signature which will be investigated on the PPCG dataset.

5.3.3 Additional Future Work

There are several results from chapter 3 that would benefit from *in vitro* investigation. It was hypothesised that chronic accumulation of bacteria in the vicinity of prostate tissue leads to an inflammatory recruitment of MDSCs. This could be investigated in the PPCG cohort to a certain extent, but would also require results from laboratory experiments. One approach could be to use FISH (Fluorescence *In Situ* Hybridization) to demonstrate the co-localisation of MDSC populations with bacteria and regions of aggressive prostate cancer (by means of high Gleason scores or high proliferation). One relatively short experiment that could be interesting is to culture prostate cancer cell-lines in the presence of lysates of ABBS bacteria. These cell-lines could then be subject to RNA sequencing and differential expression analysis to identify how exactly ABBS genera influence cellular expression. This experiment may be hindered by bacteria/cell-lines adapting to culture methods and losing the influence of the complex stromal microenvironment.

5.3.4 Potential for Clinical Impact

Stronger evidence of the associations in this thesis should be obtained before considering how to achieve the ultimate goal of translational benefit. The most promising results for further work in this thesis is to focus on the use of ABBS genera as a biomarker for aggressive prostate cancer. Doing so would require an additional development cohort and a cross-site prospective study, ideally with repeated sampling to establish the robustness of bacterial detection in the urine.

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) criteria⁴⁴⁵ should be considered in taking this work further for the development of a clinically usable model. TRIPOD is a 22-point checklist aiming to improve the reporting of studies developing, validating, or updating a prediction model. These criteria provide recommendations for all aspects of a scientific study ranging from the study title through to the methods, results and reporting the sources of funding.

Although the association between ABBS genera and aggressive prostate cancer has been found using 16S and tissue WGS data (see section C.4), I would still categorise this association as preliminary. In the TRIPOD criteria, these findings would be classified as analysis type 1a: “Development only”. The PPCG will offer a separate dataset and cohort to investigate these findings further (TRIPOD analysis type 3: “Development and validation using separate data”). Only after these findings are more established can we begin to move into a cohort designed for validation (TRIPOD analysis type 4).

Establishing a validation study should be done with careful consideration using the TRIPOD criteria. Chapter 3 analysed samples from patients undergoing a range of different *initial* treatment options (*e.g.*, radiotherapy, surgery, hormone therapy). Investigating the utility of bacterial biomarkers could be complicated by variety within future patient cohorts. Additionally, future research would benefit from pre-defined clinical endpoints such as metastases and investigating common confounders in microbiome studies which have

not been investigated on this dataset. It may be that RNA sequencing is not the most cost-effective approach to developing a clinical test for bacteria in these samples and so alternative approaches (*e.g.*, quantitative PCR) could be investigated. The main benefit of using RNA sequencing technology is that it could prove additive to other prognostic models investigating human components²⁴¹ with or without additional clinical features.

Appendix A

System Information and Background

A.1 System Information

To enable reproducibility, the system information and all R packages used at the time of rendering this thesis are as follows.

```
#sys info goes off screen if simply called, so produce as a table
# System Information is as follows:
sysinfo <- as.data.frame(Sys.info())
sysinfo %<>% rownames_to_column('variable')
colnames(sysinfo) <- c('Variable', 'Data')

# knitr::kable(caption = "System Information at Time of Rendering",
# sysinfo,
# booktabs = TRUE,
# longtable= TRUE,
# format = "latex") %>%
# kable_styling(bootstrap_options = "striped",
#full_width = F, latex_options = "scale_down")

knitr::include_graphics(here::here("figures", "sysinfo.png"))
```


System Information at Time of Rendering

Variable	Data
sysname	Darwin
release	18.7.0
version	Darwin Kernel Version 18.7.0: Tue Nov 10 00:07:31 PST 2020; root:xnu-4903.278.51~1/RELEASE_X86_64
nodename	UEAC02V73CJHTD6
machine	x86_64
login	agihawi
user	agihawi
effective_user	agihawi

```
# All loaded R packages at the time of rendering:
sessionInfo()
```

R version 3.6.1 (2019-07-05)

Platform: x86_64-apple-darwin15.6.0 (64-bit)

Running under: macOS Mojave 10.14.6

Matrix products: default

BLAS: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib

LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

locale:

```
[1] en_GB.UTF-8/en_GB.UTF-8/en_GB.UTF-8/C/en_GB.UTF-8/en_GB.UTF-8
```

attached base packages:

```
[1] grid      parallel stats4    stats      graphics  grDevices utils
[8] datasets  methods   base
```

other attached packages:

```
[1] flextable_0.6.8          limma_3.42.2
[3] egg_0.4.5                circlize_0.4.13
[5] ggbiplot_0.55            scales_1.1.1
[7] plyr_1.8.6               ClassDiscovery_3.4.0
[9] oompaBase_3.2.9          cluster_2.1.2
[11] VennDiagram_1.6.20       futile.logger_1.4.3
[13] EnhancedVolcano_1.4.0    ComplexHeatmap_2.2.0
[15] pheatmap_1.0.12          DESeq2_1.26.0
[17] SummarizedExperiment_1.16.1 DelayedArray_0.12.3
```

[19] BiocParallel_1.20.1	matrixStats_0.60.1
[21] Biobase_2.46.0	GenomicRanges_1.38.0
[23] GenomeInfoDb_1.22.1	IRanges_2.20.2
[25] S4Vectors_0.24.4	BiocGenerics_0.32.0
[27] ggforce_0.3.3	survival_3.2-13
[29] survminer_0.4.9	gridExtra_2.3
[31] datapasta_3.1.0	lubridate_1.7.10
[33] ggrepel_0.9.1	EnvStats_2.4.0
[35] ggbeeswarm_0.6.0	magrittr_2.0.1
[37] reshape2_1.4.4	vegan_2.5-7
[39] lattice_0.20-44	permute_0.9-5
[41] huxtable_5.4.0	cowplot_1.1.1
[43] ggpubr_0.4.0	kableExtra_1.3.4
[45] forcats_0.5.1	stringr_1.4.0
[47] dplyr_1.0.7	purrr_0.3.4
[49] readr_2.0.1	tidyr_1.1.3
[51] tibble_3.1.4	ggplot2_3.3.5
[53] tidyverse_1.3.1	thesisdown_0.2.0.9000
[55] remotes_2.4.0	

loaded via a namespace (and not attached):

[1] utf8_1.2.2	tidyselect_1.1.1	RSQLite_2.2.8
[4] AnnotationDbi_1.48.0	htmlwidgets_1.5.4	munsell_0.5.0
[7] withr_2.4.2	colorspace_2.0-2	uuid_0.1-4
[10] knitr_1.34	oompaData_3.1.1	rstudioapi_0.13
[13] officer_0.4.0	ggsignif_0.6.3	labeling_0.4.2
[16] GenomeInfoDbData_1.2.2	KMSurv_0.1-5	polyclip_1.10-0
[19] bit64_4.0.5	farver_2.1.0	rprojroot_2.0.2
[22] vctrs_0.3.8	generics_0.1.0	lambda.r_1.2.4
[25] xfun_0.25	markdown_1.1	R6_2.5.1
[28] clue_0.3-59	locfit_1.5-9.4	bitops_1.0-7
[31] cachem_1.0.6	assertthat_0.2.1	vroom_1.5.4
[34] nnet_7.3-16	beeswarm_0.4.0	gtable_0.3.0
[37] rlang_0.4.11	genefilter_1.68.0	systemfonts_1.0.2
[40] GlobalOptions_0.1.2	splines_3.6.1	rstatix_0.7.0
[43] broom_0.7.9	checkmate_2.0.0	yaml_2.2.1
[46] abind_1.4-5	modelr_0.1.8	backports_1.2.1
[49] Hmisc_4.5-0	gridtext_0.1.4	tools_3.6.1
[52] bookdown_0.24	ellipsis_0.3.2	RColorBrewer_1.1-2
[55] Rcpp_1.0.7	base64enc_0.1-3	zlibbioc_1.32.0
[58] RCurl_1.98-1.4	rpart_4.1-15	GetoptLong_1.0.5
[61] zoo_1.8-9	haven_2.4.3	fs_1.5.0
[64] here_1.0.1	futile.options_1.0.1	data.table_1.14.0
[67] openxlsx_4.2.4	reprex_2.0.1	hms_1.1.0
[70] evaluate_0.14	xtable_1.8-4	XML_3.99-0.3
[73] rio_0.5.27	jpeg_0.1-9	mclust_5.4.7
[76] readxl_1.3.1	shape_1.4.6	compiler_3.6.1
[79] crayon_1.4.1	htmltools_0.5.2	mgcv_1.8-36

```
[82] tzdb_0.1.2          Formula_1.2-4          ggtext_0.1.1
[85] geneplotter_1.64.0   DBI_1.1.1              formatR_1.11
[88] tweenr_1.0.2          dbplyr_2.1.1           MASS_7.3-54
[91] Matrix_1.3-4          car_3.0-11             cli_3.0.1
[94] pkgconfig_2.0.3       km.ci_0.5-2            foreign_0.8-76
[97] xml2_1.3.2            svglite_2.0.0          annotate_1.64.0
[100] vipor_0.4.5           webshot_0.5.2          XVector_0.26.0
[103] rvest_1.0.1           digest_0.6.27          rmarkdown_2.10
[106] cellranger_1.1.0      survMisc_0.5.5         htmlTable_2.2.1
[109] gdttools_0.2.3        curl_4.3.2             rjson_0.2.20
[112] lifecycle_1.0.0       nlme_3.1-153           jsonlite_1.7.2
[115] carData_3.0-4         viridisLite_0.4.0      fansi_0.5.0
[118] pillar_1.6.2          fastmap_1.1.0          httr_1.4.2
[121] glue_1.4.2            zip_2.2.0              png_0.1-7
[124] bit_4.0.4             stringi_1.7.4          blob_1.2.2
[127] latticeExtra_0.6-29   memoise_2.0.0
```

```
# The date and time of last rendering:
Sys.time()
```

```
[1] "2021-09-19 07:44:01 BST"
```

A.2 Background

A.2.1 Beta Diversity Measurements

The following provides an overview of community similarities and distance metrics, S . Dissimilarities can be calculated with $D = 1 - S$. N_A , total number of taxa in sample A ; n_A total number of individuals in sample A ; n_{Ai} , individuals of taxon i in sample A ; $\hat{U}$, estimator for fraction of individuals in shared taxa for sample A ; φ_i , phylogenetic branch length of taxon i to root; Φ_{ij} , phylogenetic association between taxa i and j ; C_{ij} interaction similarity between taxa i and j . These formulae are non-exhaustive and are as shown in Schmidt *et al.* 2016⁴⁴⁶ and Barwell *et al.* 2015⁴⁴⁷. Some other examples include: Chord, Mahalanobis, χ^2 , Hellinger, Canberra, Sorensen, Pearson, Spearman's, Kendall's, Dice and Ochai.

Jaccard Classical

Ratio of shared taxa among all observed taxa.

$$Jaccard\ Classical = \frac{|A \cap B|}{|A \cup B|} \quad (A.1)$$

Jaccard Weighted

Average fraction of individuals in shared taxa

$$Jaccard\ Weighted = 0.5 \left(\sum_{i \in A \cap B} \frac{n_{Ai}}{n_A} + \sum_{j \in A \cap B} \frac{n_{Bj}}{n_B} \right) \quad (A.2)$$

Jaccard, Chao

Abundance-based Jaccard index, corrected for unseen shared taxa based on shared rare taxa

$$Jaccard\ Chao = \frac{\hat{U}\hat{V}}{\hat{U} + \hat{V} - \hat{U}\hat{V}} \quad (A.3)$$

Bray–Curtis

Fraction of minimum per-sample abundance of shared taxa

$$Bray\ Curtis = \frac{2 \sum_{i \in AB} \min(n_{Ai}, n_{Bi})}{n_A + n_B} \quad (A.4)$$

Morisita-Horn

Pairwise overlap of taxa abundances, adjusted by per-sample concentration indices

$$Morista\ Horn = \frac{2 \sum_{i \in A \cup B} n_{Ai} n_{Bi}}{\left(\frac{\sum_{i \in A} n_{Ai}^2}{n_A^2} + \frac{\sum_{i \in B} n_{Bi}^2}{n_B^2} \right) n_A n_B} \quad (A.5)$$

UniFrac, unweighted

Fraction of shared branch length on a phylogenetic tree

$$Unweighted\ Unifrac = \frac{\sum_{i \in A \cap B} \Phi_i}{\sum_{j \in A \cup B} \Phi_j} \quad (A.6)$$

UniFrac, weighted

Fraction of shared branch length on a phylogenetic tree, weighted by taxa abundances

$$Weighted\ Unifrac = 1 - \frac{\sum_{i \in A \cup B} \Phi_i \left| \frac{n_{Ai}}{n_A} - \frac{n_{Bi}}{n_B} \right|}{\bar{\Phi}} \quad (A.7)$$

Euclidean distance

Also known as L_2 distance or Pythagorean distance. Euclidean distance equates to the length of the straight line between two points mapped in euclidean space.

$$Euclidean\ Distance = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2} \quad (A.8)$$

Manhattan distance

Also known as L_1 distance or ‘*taxi-cab*’ distance. This is similar to Euclidean distance above, but rather than use the pythagorean distance, Manhattan simply adds the absolute differences of the two dimensions for two objects mapped in euclidean space.

$$\text{Manhattan Distance} = (|y_B - y_A|) + (|x_B - x_A|) \quad (\text{A.9})$$

A.2.2 Computational Methods for taxonomic classification

This section provides a non-exhaustive list of computational tools that can be used for metagenomic classification or the analysis of metagenomic data and provides a brief overview of their functionality and/or performance.

AbundanceBin

k -mer based approach. Can bin raw sequencing reads or contigs, provides species abundance estimates. Performs unsupervised clustering on k -mers into bins that are expected to be the same or similar species. Wu and Ye combined AbundanceBin with MetaCluster for improved performance.⁴⁴⁸

BiMeta

An unsupervised k -mer based approach that analyses overlapping k -mers within non-overlapping reads. Bimeta was shown to have the highest performance when compared with other tests in their analysis on short and long reads. Uses k -mer frequency distribution on sets of non-overlapping reads obtained from microbial genomes.⁴⁴⁹

BLAST

The most widely used local alignment tool and the earliest computational tool utilised for metagenomics analysis, typically very slow. MegaBLAST is often a speed benchmark for various metagenomics tools. A modified version of BLAST is used in a variety of metagenomic tools.^{211,450}

BRAKEN

Bayesian re-estimation of Kraken results. It statistically reassigns k -mer taxonomic classifications to the species level in order to boost performance of within 1% abundance estimates of true abundance values.¹⁰⁹

CARMA

Developed for 454 pyrosequencing. Classifies all fragments as short as 80bp. Identifies domain/protein family fragments in unassembled reads by using protein family profile hidden Markov models. 61-84% order level sensitivity.⁴⁵¹

CLARK

Full/partial k -mer based approach. Boasts precision of up to ~99% and sensitivity of up to ~89% at the Genus level, whilst up to 99% precision and a lower sensitivity that averages about 60% at species level on simulated datasets.⁴⁵²

CompostBin

k -mer based approach. Assigns taxon-specific bins to raw reads. Requires number of taxonomic bins/ number of species in the sample as input to help determine number of bins.⁴⁵³

CONCOCT

Clusters contigs into genomes based on sequence composition and coverage over multiple samples. Boasts 99.8% precision/sensitivity at species level.⁴⁵⁴

Centrifuge

Searches both forward and reverse complement. 16bp match found first, then extended as far as possible. Species is assigned to the read/contig by summing score of exact matches weighted for match-length. Classification Index built by obtaining only unique sequences between species in the database. Uses an indexing scheme based on Burrows-Wheeler transform (BWT) and Ferragina-Manzini (FM) compression. The resulting database uses ~4.2GB RAM, small enough to be used in field metagenomics. Reports ~77% species sensitivity and 98.4% species precision.²⁰⁶

COCACOLA

Bins contigs into OTU's based on sequence composition and coverage across multiple samples. Also takes into consideration the co-alignment to reference genomes and linkage between paired-end reads to improve performance. In their analysis they found that COCACOLA outperformed all other contig binner comparators.⁴⁵⁵

CoMeta

k -mer based approach. Assigns reads to user-defined groups using k -mers. Boasts 98-99% overall sensitivity and 94-99% overall precision and typically uses <100GB RAM, slower than Kraken.⁴⁵⁶

ConStrains

An adaptation of MetaPhlAn to investigate strain level. Following identification, ConStrains uses single nucleotide polymorphism (SNP) patterns to determine intra-species variation to establish strains. Also recovers information about the dominance/rarity of strains. Requires adequate coverage for SNP calling. Therefore, caveat that species must be sufficiently abundant. Uses MetaPhlAn for profiling, then analyses those with sufficient sequencing depth further.⁴⁵⁷

DIAMOND

Used for alignment of short sequence reads, faster than BLAST. DIAMOND is used in tools such as MEGAN6. DIAMOND is used to align input sequencing reads against the NCBI-nr protein reference database.⁴¹⁷

DiScRIBinATE

An adaptation of Sort-ITEMS that reduces binning time by 50% and partially reduces misclassification. Requires BLASTx output for reads against reference sequences as input.⁴⁵⁸

DUDes

Taxonomic binning process. Uses a ‘*top down*’ approach to classify sequences that are matched to taxonomic groups. DUDes does this by comparing number of alignments (bin score) between input sequences and reference sequences associated with taxonomic classifications. The top 50% bins are selected, and P-values are estimated for each node based on alignment strength. Reported species level sensitivity is 95.2%, species level specificity is 99.4% on HMP data.⁴⁵⁹

EBI Metagenomics

Used for rRNA and protein coding. An online metagenomics resource developed by the European bioinformatics institute. It is a fully automated pipeline that has a Python framework and takes advantage of an IBM Platform Load Sharing Facility computing cluster. Applies Qiime for taxonomic prediction.⁴⁶⁰

FACS

k -mer based approach. Reference database: Creates a Bloom filter for each reference genome: a bitmap/hash-based data structure for efficient look-up. Query sequences are broken down into non-overlapping Kmers. Each k -mer interrogated with reference bloom filters. Scores produced based on length of k -mer match. Claims 98.1-99.8% sensitivity and 100% specificity.⁴⁶¹

Focus

k -mer based approach. k -mer counts from reference datasets and metagenomics datasets are obtained using Jellyfish. Nonnegative least squares is used to calculate the optimum abundance of k -mers in the query dataset.⁴⁶²

GAAS

Estimates relative abundance and genome size from metagenomics sequences. Written in Perl and uses a modified BLAST based approach.⁴⁶³

GASiC

A method of improving observed abundances by analysing similarieis in reference sequence data written in Python. Metagenomic reads are aligned to reference genomes and the SAM

alignment is input for GASiC.⁴⁶⁴

Genometa

Uses alignment BAM files to reference sequences, can use multiple alignment tools such as BLAST, Bowtie and BWA. Genometa has a graphical user interface and is a modification of the integrated genome browser.⁴⁶⁵

GOTTCHA

k-mer based approach. A semi-automated approach aimed at specifically reducing the false positive rate associated with classification tools. Eliminates genomic regions responsible for false positive results by analysing distribution and depth of coverage. Machine-learning approach is employed to set a limit of detection. Query sequences are divided into 30-mers after quality trimming, apart from the end fragment which is 30- 59bp. *k*-mers. *k*-mers are mapped to a reference database using BWA-MEM.²⁰⁵

GRAMMy

GRAMMy Analyses sequence composition and accepts reads as input. Reads are mapped to reference sequences using an alignment tool such as BLAT or BLAST to produce a read probability matrix or by measuring *k*-mer frequency. These two measurements are combined to form a read probability matrix file from which relative abundance and taxonomic information is estimated using the expectation maximisation algorithm.⁴⁶⁶

GroopM

A contig binner. Contigs and BAM files mapping to reference sequences are input for GroopM. From these files, a set of high confidence bins is created which can be manually refined, then the unbinned contigs are added to the bins previously generated and the output is FASTA files for each bin.⁴⁶⁷

Illumina BaseSpace Apps:

Typically, applications of previously developed tools. Some examples include: 16S Metagenomics, Bacterial Analysis Pipeline, Genius Metagenomics: Know Now. These tools are often adaptations of other currently available tools and are listed for use on the Illumina BaseSpace browser-based sequence hub available at: <https://basespace.illumina.com/>

INDUS

Alignment free approach. Genome fragments are compared to input sequences by measuring tetranucleotide frequency. Taxon names of the reference sequences that match the input sequences are retrieved. A taxonomic level cut-off is determined based on ‘distance’ between input sequences and genome fragments. The sequence is categorised as unassigned if the taxonomic level is above superkingdom level.⁴⁶⁸

Kaiju

Uses BWT to find maximum exact matches at the protein level in the 6 possible reading frames and matches to a reference database of microbial proteins reporting the taxon or the lowest common ancestor in the case of equally good matches.¹⁰¹

Kallisto

Designed for RNA-seq quantification. Pseudo-aligns reads to references. Measured at almost 100% sensitivity but with much more specificity than Bowtie2.³⁷⁷

KNNLog

Whilst not strictly a classification tool, KNNLog Combines instance-based learning (using k-nearest neighbour) and model-based learning (using logistic regression) to be able to predict classes. OTUs were determined by Mothur sorted by KNNLog. KNNLog was included for completeness due to reference by other metagenomics tool publications.⁴⁶⁹

Kraken

k-mer based approach. One of the most widely used classifiers, Kraken breaks down reads into all possible *k*-mers of a chosen length and assigns each *k*-mer to a taxonomic classification. The lowest common ancestor from all classified *k*-mers within a read is reported for that read. Reported Genus level sensitivity of 80-99%¹⁰³ evolutions: Kraken2⁴⁷⁰ and Krakenuniq²⁰⁸.

***k*-SLAM**

Sequences are split into overlapping *k*-mers, their reverse complement produced and these are sorted by similarity. Overlap of a set number of bases is measured and a full Smith-Waterman alignment is performed. Taxonomy inferred from the LCA. Designed to work with paired-end reads. Reads that map to adjacent locations on the same reference genome are grouped together to form a pseudo-assembly and an alignment score calculated based on this.⁴⁷¹

LikelyBin

Uses unsupervised statistical clustering of environmental shotgun sequences. *k*=17-20 classifies fragments with over 90% accuracy in low-complexity simulated metagenomes.⁴⁷²

LMAT

Uses a similar approach to Kraken in that the taxonomic classification of a read is found by determining the LCA.⁴⁷³

MASH

MASH produces MinHash sketches from large metagenomes and compresses sequence data and then measures the difference between two sketches to return an estimate (Jaccard Index) of the shared number of *k*-mers between 2 sequences.⁴⁴⁴

MaxBin 2.0

Classifies sequences following metagenome assembly. Measures tetranucleotide frequency of assembled contigs.⁴⁷⁴

MEGAN6

Metagenomics analysis pipeline. DNA reads compared with protein reference database using BLAST/DIAMOND and a taxonomy ID is assigned. DIAMOND Can utilise multiple taxonomic systems, such as NCBI/SILVA.^{475,476}

MetaBAT

Input is aligned BAM (binary alignment map) file of reads aligned to previously assembled metagenome. MetaBAT measures the distance between each read and contig by measuring tetranucleotide frequency and abundance. These two measurements are combined, and a matrix is formed of all combined measurements. Contigs are then binned iteratively. Identified the second most bins in their analysis, second to GroopM on synthetic metagenomes but the most on real metagenomes.⁴⁷⁷

MetaBinG

Utilises graphics processing units (GPU's) for accelerated metagenomic classification. 140x faster than Phymm. Uses a parallel programming module called CUDA.⁴⁷⁸

MetaCluster

Unsupervised read/contig binner. Designed specifically for retrieving low abundance species in complex metagenomes. Firstly MetaCluster groups reads from highly abundant species and the remaining reads are subject to noise from low abundance-species. MetaCluster attempts to correct for low coverage of species with low abundance.^{479,480}

MetaDomain

Classifies based on protein domains. Works with DNA fragments as short as 80bp. Protein analysis allows for a more functional analysis of a metagenome and has the potential to nullify silent single nucleotide variation. Showed improvements over HMMER. Not particularly a taxonomic classifier but holds potential for classification Other tools such as HMMer and HMM-GRASPx not included.⁴⁸¹

MetaMeta

A tool produced to combined the results of several metagenomic tools in order to obtain a consensus metagenomics profile that shows improvements compared to use of single tests.¹⁹²

MetaMLST

Requires species isolation by using culture-based techniques. used for strain-level identification and tracking using multilocus sequence typing (MLST).⁴⁸²

MetaPalette

k -mer based approach for all reads. Uses k -mers to match to the k -mer composition (palette) of a given organism. Was shown to have the lowest false positive rate at Phyla level of tools including: MetaPhlAn, Kraken, LMAT and OneCodex.⁹⁹

MetaPhyler

Uses 31 universal phylogenetic marker genes as reference. Classifies based on BLAST scores and has a BLAST score threshold of positivity. ~35% sensitivity at Genus level which increases to ~75% specificity at Phylum level.⁴⁸³

MetaPhlAn

Reference database prepared by identifying clade specific marker genes that contain ~4% of sequenced microbial genes. Provides estimates of relative abundance but does not attempt map every read in an input sample, only clade-specific ones. Provides estimates down to species/strain level. MetaPhlAn2 Uses Bowtie to compare input sequences to reference sequences, previous versions use BLAST^{99,110}.

MetaProb

Assembly based approach. First reads are pseudo-assembled into groups, and the k -mer frequency within these groups are measured and compared to the k -mer frequency of reference genomes.⁴⁸⁴

MetaWATT

Contig binner. A Graphical application written in Java that bins based on multivariate statistics of tetranucleotide frequencies. In their analysis they assembled prior to binning using MetaVelvet.⁴⁸⁵

Metaxa2

Designed for use on 16S rRNA. Uses HMMER to find sequences containing rRNA and then subjects these to analysis using BLAST against a specialised reference database. Worked well for contig length > 400.⁴⁸⁶

MG-RAST

An open source and online metagenomics service containing multiple tools for analysis of microbial community structure and function.⁴⁸⁷

MLTreeMap

A pipeline/framework for metagenomic sequence analysis. Marker genes in input sequences are searched for by using BLASTX. All hits are extracted from the input sequences, alignments undergo some mild gap removal and the subsequent sequences are provided to RAxML to obtain phylogeny. Also performs some functional analysis, including identifying important enzyme families.⁴⁸⁷

mOTU

Uses single copy phylogenetic marker genes to form OUT's from shotgun metagenomic data.⁴⁸⁸

MTR

Groups similar reads prior to classification inspired by LCA. Uses BLAST for alignments. Written in Matlab and C++.^{489,490}

MyCC

Contig binner. Does not require reference sequences as part of function. Based on the recognition of 40 universal marker genes, metagenomics assembly and recognition/classification of contig clusters.⁴⁹¹

NBC

Machine learning approach to metagenomics classification. Every read is classified and presents an attractive alternative to BLAST based approaches at the time of release. A webserver has been released that utilises naïve bayes classifier (NBC). Originally released in 2008 prior to web server implementation.⁴⁹²

One Codex

k -mer based approach with default $k=31$. Demonstrated better performance than other tools in their analysis such as Kraken. MiniKraken, Seed-Kraken and Clark.⁴⁹³

PanPhlAn

Strain level metagenomics profiling by identifying gene composition and transcriptional activity. Does not include abundance estimation.⁴⁹⁴

PaPaRa

Novel method for short read alignment against a reference sequence written in C/C++. Uses phylogenetic placement methods such as EPA or pplacer.⁴⁹⁵

Parallel-META3

Written in C/C++ and/or R. Outputs useful statistics such as alpha/beta diversity. Provides data mining for taxonomic information and metabolic function for metagenomics datasets. Has useful features for 16S rRNA analysis. More efficient CPU usage than PICRUSt and Qiime.⁴⁹⁶

PathSeq

Designed specifically for human sequences. Employs trimming and a primary subtractive phase to remove human sequences from metagenomics sequences using MAQ and 5 human reference genomes. Also aligns sequences to human reference sequences using BLAST.

Remaining sequences are subject to metagenomics assembly using Velvet for novel species identification.^{209,497}

Pathoscope

Uses a Bayesian statistical framework to take into consideration sequence quality and mapping quality. Reassigns ambiguous classifications using the expectation-maximization algorithm.⁴⁹⁸

PCAHIER

Designed for use on short read metagenomics (<1000bp). Employs hierarchical classification and principal component analysis.⁴⁹⁹

PICRUSt

Designed to function on 16S rRNA data. Predicts relative abundance, normalised based on rRNA copy number estimation.¹⁶⁷

PhyloPythia S+

Contig binner. Quoted as performing well at Genus level. Outperformed Kraken, Taxator-tk and MEGAN6 in their analysis.⁵⁰⁰

PhymmBL

Classifies reads as short as 100bp. Genome binner based on phylogenetic analysis.⁵⁰¹

Prophyle

Low computational requirement. k -mer based approach. Assembles contigs at each taxonomic node.⁵⁰²

ProViDE

Specifically developed to measure viral diversity. 186 Qiime A tool used for phylogenetic analysis through marker gene analysis such as 16S rRNA . Qiime offers a platform to help determining phylogenetic structure and OTU's obtained from samples. Typically assumes 97% similarity between sequences in order for them to be classified as the same species.⁵⁰³

Quikr

A k -mer based approach. Produces a vector of taxonomic assignments. Based on the mathematical theory of compressive sensing.⁵⁰⁴

RAlphy

“uses a novel sequence similarity metric with iterative refinement of taxonomic models”⁵⁰⁵

RAxML

A widely used tool for phylogenetic analysis. Can also offer placing short reads into a given reference phylogeny using an adaptation of PaPaRa to determine the evolutionary origin of the reads⁵⁰⁶

Riems

Classifies every read in a metagenomic dataset. Uses BLAST to quantify sequence similarity.⁵⁰⁷

RINS

Uses non-overlapping k -mers of length 25.⁵⁰⁸

RITA

Classifies reads as small as 50bp. Measures agreement between composition and homology. Uses BLASTN, BLASTX and D-BLASTN.⁵⁰⁹

SCIMM

Performs unsupervised clustering of sequences. Suggested to be used alongside a supervised method Phymm in a tool called PHYSCIMM.⁵¹⁰

SECLAF

SECLAF is a webserver that uses deep neural networks for hierarchical classification of sequences.⁵¹¹

SeMeta

Reads are separated into clusters based on composition and a semi-supervised machine learning approach to classifies short length sequences with sufficient overlap.⁵¹²

S-GSOM

Tested on rRNA data. An unsupervised approach. Presented with comparable results to PhyloPythia.⁵¹³

Sigma

Designed for strain level inference of genomes for biosurveillance. Measures statistical certainty of classifications and performs hypothesis testing.⁴⁸⁹

SLIMM

Uses “coverage information of reference genomes to remove unlikely genomes from the analysis and subsequently gain more uniquely mapped reads to assign at lower ranks of a taxonomic tree.”⁵¹⁴

SMART

4bp match performed first and then performs match searching based on a sliding window of 30. Checks reverse complement for every read also.⁵¹⁵

Sourmash

Creates MinHash sketches of sequence data. Minhash is effectively a method of data compression to make comparison between large sequences much more efficient.⁴⁴³

Sort-ITEMS

Functions based on BLAST results to reference sequences. Classifies at an appropriate level in the taxonomic tree by analysing the quality of hits to reference sequences.⁵¹⁶

SPHINX

Uses both sequence composition as well as alignment to infer decisions about taxonomic binning.⁵¹⁷

SRST2

Specifically developed for classification of Illumina sequence data. Designed specifically to detect the presence of genes, alleles and MLST. Recommended for use in the public health surveillance context.⁵¹⁸

StrainSeeker

k -mer based approach. designed to work from sequence data of bacterial isolates. Ultimately aimed at strain typing.⁵¹⁹

SURPI

Uses SNAP and RAPSearch aligners. Boasts speed of performance of 11mins-5 hours on datasets of 7-500 million reads. Has been piloted in the diagnoses of acutely ill patients.⁵²⁰

TACOA

Declared high accuracy until class level.⁵²¹

TAEC

Utilises genome composition and alignment. Requires alignment as input, can be any aligner, not just BLAST.⁵²²

Taxsom

Web based service for public use. Produces self-organising maps. Boasts 90-100% specificity / sensitivity from phylum to genus level.⁵²³

Taxator-tk

Classification based on sequence similarity, uses BLAST.⁵²⁴

Taxoner

Classification pipeline, based on alignment. Uses an aligner such as Bowtie2⁵²⁵

Taxonomer

A web-based tool for metagenomic classification. Provides built in results visualisation. Provides combined nucleotide and protein-based classification. And mRNA transcript profiling.⁵²⁶

Taxy-Pro

Performs taxonomic classification by identification of protein domains.⁵²⁷

TETRA

Uses z-scores from tetramer frequencies to classify metagenomics sequences.⁵²⁸

TIPP

Marker based identification. Self-declared as performing well in the presence of indel errors and novel genomes. Claims improvement on NBC, mOTU, PhymmBL, MetaPhyler and MetaPhlAn.⁵²⁹

TreePhyler

Written in Perl, taxonomic classification based on protein family.⁵³⁰

TWARIT

Classification based on sequence alignment and composition. Uses BWA for alignment and calculates the LCA.⁵³¹

VizBin

Contig Binner – written in Java, provides a great visualisation of results. Employs pattern recognition for cluster identification and delineation. Runs in Windows, Mac and Linux.⁵³²

What's in mypot (WIMP)

Real-time species detection for the MinION sequencer. An application of Kraken⁵³³

WSVDD

A machine learning approach that uses a training dataset to be able to remove outliers from sample data. Performs supervised clustering as a method of classification.⁵³⁴

A.2.3 Metagenomic Assembly Tools

This section contains different tools that can be used for (meta)genomic assembly.

Canopy

Does not use reference genomes as function, more of a metagenome assembly-based approach. Genes are assembled from individual samples and the abundance of each gene is measured. Pearson correlation coefficient is measured between random ‘bait’ seed genes and a fixed coefficient distance is referred to as a canopy. The claim is that this technique helps differentiate between closely related strains.⁵³⁵

IDBA-UD

Metagenomic Assembly based on multiple k -mer lengths and uses a De Bruijn graph assembly approach.⁵³⁶

LSA

Latent Strain Analysis (LSA) partitions reads from metagenomics sequence data by using covariance information to cluster sequences within a fixed amount of memory. Not strictly a tool for metagenomics classification, but a preassembly partitioning tool to separate reads from several species.⁵³⁷

MegaHit

Assembly based on multiple k -mer sizes and De Bruijn Graphs.⁴¹⁵

Meraga

A combination of Meraculus and MegaHIT. Used in CAMI⁹⁸

Metassembler

Combines multiple metagenomic assembly results⁵³⁸

MetaCortex

Identifies sub graphs within a greater De Bruijn Graph assembly. Currently Unpublished – Available online

MetaSpades

Metagenomic Assembly based on multiple k -mer lengths and uses a De Bruijn graph assembly approach²⁰⁷

MetaVelvet

Uses a single k -mer size and De Bruijn graph assembly approach.⁵³⁹

Minia

Metagenome assembly based on De Bruijn graphs and bloom filters.⁵⁴⁰

Omega

Finds overlaps between reads using a prefix/suffix hash table.⁵⁴¹

Ray Meta

Metagenome assembly based on De Bruijn graphs.⁵⁴²

SOAPdenovo

Uses a single k -mer size and De Bruijn graph assembly approach.⁵⁴³

Unicycler

Allows for hybrid assembly - mostly for assembly of isolates and not metagenomic assembly.^{304,544}

Velour

Metagenome assembly based on De Bruijn graphs Reference not found, used in CAMI⁹⁸

A.2.4 Common Contaminants in Negative Controls

Taxa commonly found in negative control samples as appears in Eisenhofer *et al.* 2019. This publication reports taxa identified in negative conatrols of more than one study.

Contaminant Genera:

Actinomyces, *Corynebacterium*, *Arthrobacter*, *Rothia*, *Propionibacterium*, *Atopobium*, *Sediminibacterium*, *Porphyromonas*, *Prevotella*, *Chryseobacterium*, *Capnocytophaga*, *Chryseobacterium*, *Flavobacterium*, *Pedobacter*, *UnclassifiedTM7*, *Bacillus*, *Geobacillus*, *Brevibacillus*, *Paenibacillus*, *Staphylococcus*, *Abiotrophia*, *Granulicatella*, *Enterococcus*, *Lactobacillus*, *Streptococcus*, *Clostridium*, *Coprococcus*, *Anaerococcus*, *Dialister*, *Megaspheera*, *Veillonella*, *Fusobacterium*, *Leptotrichia*, *Brevundimonas*, *Afipia*, *Bradyrhizobium*, *Devosia*, *Methylobacterium*, *Mesorhizobium*, *Phyllobacterium*, *Rhizobium*, *Methylobacterium*, *Phyllobacterium*, *Roseomonas*, *Novosphingobium*, *Sphingobium*, *Sphingomonas*, *Achromobacter*, *Burkholderia*, *Acidovorax*, *Comamonas*, *Curvibacter*, *Pelomonas*, *Cupriavidus*, *Duganella*, *Herbaspirillum*, *Janthinobacterium*, *Massilia*, *Oxalobacter*, *Ralstonia*, *Leptothrix*, *kingella*, *Neisseria*, *Escherichia*, *Haemophilus*, *Acinetobacter*, *Enhydrobacter*, *Pseudomonas*, *Stenotrophomonas*, *Xanthomonas*

Appendix B

Developing SEPATH

B.0.1 Species used for Metagenome Simulation

Table B.1: Species and Filenames used as part of metagenome simulation

Species	Filename
<i>Abiotrophia defectiva</i>	GCF_000160075.2_ASM16007v2_genomic.fna
<i>Actinobaculum massiliense</i>	AGWL01.1.fsa_nt.fna
<i>Actinobaculum schaalii</i>	CP008802.1.fsa_nt.fna
<i>Actinomyces neuvi</i>	AGWP01.1.fsa_nt.fna
<i>Actinomyces turicensis</i>	AGWQ01.1.fsa_nt.fna
<i>Actinotignum urinale</i>	ATUY01.1.fsa_nt.fna
<i>Anaerococcus hydrogenalis</i>	ABXA01.1.fsa_nt.fna
<i>Anaerococcus lactolyticus</i>	JRMW01.1.fsa_nt.fna
<i>Anaerococcus prevotii</i>	AEXM01.1.fsa_nt.fna
<i>Anaerococcus prevotii</i>	NC_013171.1.fsa_nt.fna
<i>Bacteroides fragilis</i>	GCF_000009925.1_ASM992v1_genomic.fna
<i>Blautia hansenii</i>	GCF_000156675.1_ASM15667v1_genomic.fna
<i>Bulleidia extructa</i>	GCF_000177375.1_ASM17737v1_genomic.fna
<i>Campylobacter rectus</i>	GCF_000174175.1_ASM17417v1_genomic.fna
<i>Campylobacter ureolyticus</i>	ARGD01.1.fsa_nt.fna
<i>Campylobacter ureolyticus</i>	NZ_CP012195.1.fsa_nt.fna
<i>Catonella morbi</i>	GCF_000160035.2_ASM16003v2_genomic.fna
<i>Clostridium spiroforme</i>	GCF_000154805.1_ASM15480v1_genomic.fna
<i>Corynebacterium accolens</i>	GCF_000159115.1_ASM15911v1_genomic.fna
<i>Corynebacterium glucuronolyticum</i>	ABYP01.1.fsa_nt.fna
<i>Corynebacterium pseudogenitalium</i>	ABYQ02.1.fsa_nt.fna
<i>Corynebacterium pyruviciproducens</i>	ATBY01.1.fsa_nt.fna
<i>Corynebacterium tuberculostearicum</i>	GCF_000175635.1_ASM17563v1_genomic.fna
<i>Cutibacterium acnes</i>	GCF_000008345.1_ASM834v1_genomic.fna
<i>Dialister invisus</i>	GCF_000160055.1_ASM16005v1_genomic.fna
<i>Enterococcus faecalis</i>	ATIO01.1.fsa_nt.fna

<i>Enterococcus faecalis</i>	GCF_000007785.1_ASM778v1_genomic.fna
<i>Escherichia coli</i>	GCF_000005845.2_ASM584v2_genomic.fna
<i>Eubacterium ventriosum</i>	GCF_000153885.1_ASM15388v1_genomic.fna
<i>Faecalibacterium prausnitzii</i>	GCF_000162015.1_ASM16201v1_genomic.fna
<i>Finegoldia magna</i>	NZ_CM000955.1.fsa_nt.fna
<i>Fusobacterium gonidiaformans</i>	ACDD01.1.fsa_nt.fna
<i>Fusobacterium nucleatum</i>	GCF_000007325.1_ASM732v1_genomic.fna
<i>Gemella haemolysans</i>	GCF_000173915.1_ASM17391v1_genomic.fna
<i>Haemophilus influenzae</i>	GCF_000027305.1_ASM2730v1_genomic.fna
<i>Helicobacter Pylori</i>	GCF_000008525.1_ASM852v1_genomic.fna
<i>Holdemania filiformis</i>	GCF_000157995.1_ASM15799v1_genomic.fna
<i>Kingella oralis</i>	GCF_000160435.1_ASM16043v1_genomic.fna
<i>Leptotrichia goodfellowii</i>	GCF_000516535.1_ASM51653v1_genomic.fna
<i>Listeria monocytogenes</i>	GCF_000196035.1_ASM19603v1_genomic.fna
<i>Mannheimia haemolytica</i>	GCF_000376645.1_ASM37664v1_genomic.fna
<i>Mobiluncus curtisii</i>	NC_014246.1.fsa_nt.fna
<i>Moraxella catarrhalis</i>	GCF_000092265.1_ASM9226v1_genomic.fna
<i>Mycoplasma genitalium</i>	NC_000908.2.fsa_nt.fna
<i>Neisseria elongata</i>	GCF_000818035.1_ASM81803v1_genomic.fna
<i>Neisseria mucosa</i>	GCF_000186165.1_Neis_muco_C102_V1_genomic.fna
<i>Pasteurella dagmatis</i>	GCF_000163475.1_ASM16347v1_genomic.fna
<i>Peptoniphilus coxii</i>	LSDG01.1.fsa_nt.fna
<i>Peptoniphilus duerdenii</i>	AEEH01.1.fsa_nt.fna
<i>Peptoniphilus harei</i>	AENP01.1.fsa_nt.fna
<i>Peptoniphilus sp. BV3C26</i>	AWXB01.1.fsa_nt.fna
<i>Porphyromonas bennonis</i>	AQWR01.1.fsa_nt.fna
<i>Porphyromonas somerae</i>	AQVC01.1.fsa_nt.fna
<i>Prevotella bivia</i>	AJVZ01.1.fsa_nt.fna
<i>Prevotella buccalis</i>	GCF_000177075.1_ASM17707v1_genomic.fna
<i>Prevotella corporis</i>	AUME01.1.fsa_nt.fna
<i>Prevotella disiens</i>	BAIS01.1.fsa_nt.fna
<i>Prevotella timonensis</i>	CBQQ01.1.fsa_nt.fna
<i>Propionimicrobium lymphophilum</i>	AGZR01.1.fsa_nt.fna
<i>Propionimicrobium lymphophilum</i>	AUIB01.1.fsa_nt.fna
<i>Propionimicrobium sp. BV2F7</i>	AZIS01.1.fsa_nt.fna
<i>Rothia dentocariosa</i>	GCF_000164695.2_ASM16469v2_genomic.fna
<i>Ruminococcus torques</i>	GCF_000153925.1_ASM15392v1_genomic.fna
<i>Salmonella enterica</i>	GCF_000195995.1_ASM19599v1_genomic.fna
<i>Sneathia sanguinegens</i>	LOQF01.1.fsa_nt.fna
<i>Staphylococcus aureus</i>	NC_017333.1.fsa_nt.fna
<i>Staphylococcus epidermidis</i>	GCF_000007645.1_ASM764v1_genomic.fna
<i>Streptobacillus moniliformis</i>	GCF_000024565.1_ASM2456v1_genomic.fna
<i>Streptococcus agalactiae</i>	NC_004368.1.fsa_nt.fna
<i>Streptococcus anginosus</i>	NC_022239.1.fsa_nt.fna

<i>Streptococcus mitis</i>	GCF_000027165.1_ASM2716v1_genomic.fna
<i>Streptococcus parasanguinis</i>	GCF_000164675.2_ASM16467v2_genomic.fna
<i>Streptococcus thermophilus</i>	GCF_000253395.1_ASM25339v1_genomic.fna
<i>Treponema vincentii</i>	GCF_000412995.1_Trep_vinc_F0403_V1_genomic.fna
<i>Ureaplasma urealyticum</i>	GCF_000021265.1_ASM2126v1_genomic.fna
<i>Varibaculum cambriense</i>	ATUF01.1.fsa_nt.fna
<i>Veillonella parvula</i>	GCF_000024945.1_ASM2494v1_genomic.fna

B.0.2 BB DuK Parameter Investigation

In a single dataset, the number of human and bacterial reads remaining after running BB-DuK with default settings but changing the k -mer size was examined (figure B.1). Although $k = 31$ is a common setting in many k -mer tools such as Kraken, an additional investigation was carried out to ensure the adequacy of its use in human depletion. Larger k -mers are generally more specific but smaller k -mers are more sensitive. For this reason, small k -mers can be *too* sensitive and result in the loss of too much information, discarding many microbial reads (figure B.1). Additionally, including single-end reads that have lost their pair due to quality trimming can increase the overall retention of microbial reads, although this resulted in a higher number of human reads being retained. The optimal k -mer size appears to be between 30 and 32 (figure B.1).

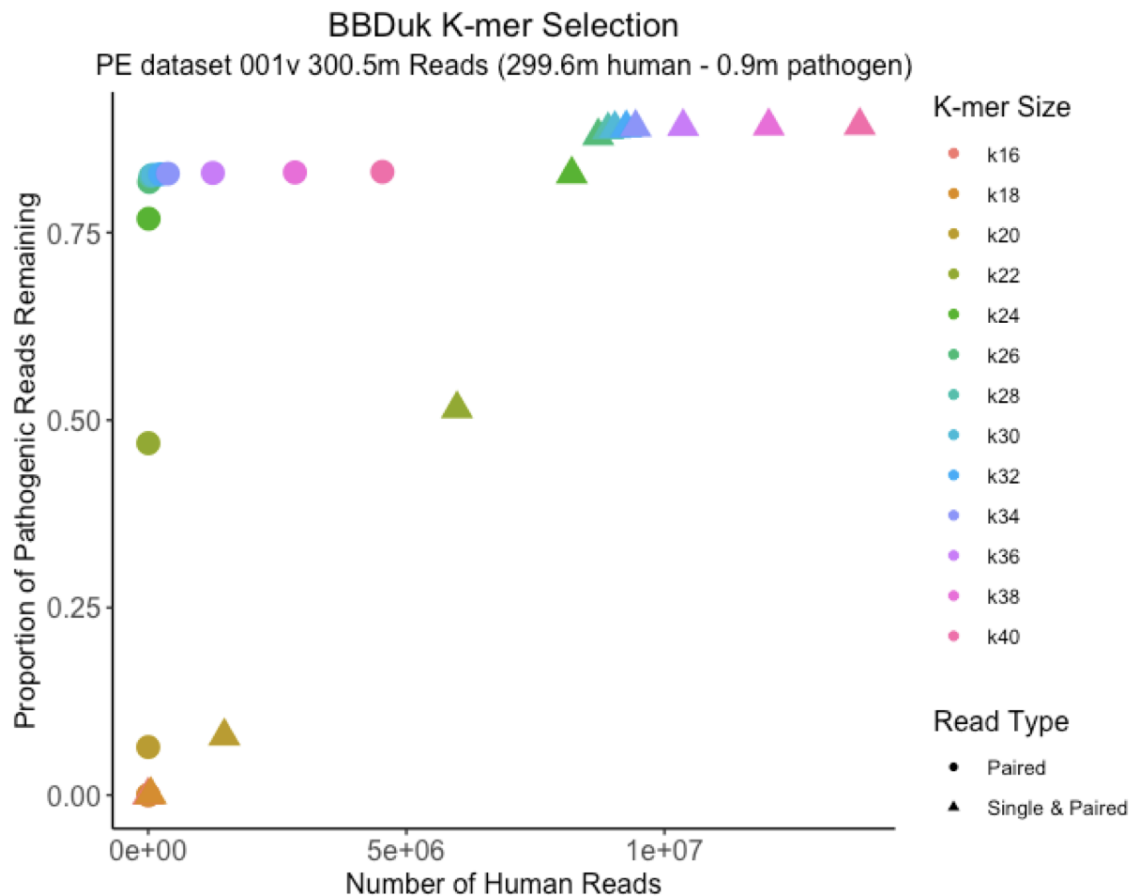


Figure B.1: The choice of k -mer size can influence bacterial and human read depletion. A single simulated dataset was subject to quality trimming and then human depletion using various BBDuk k -mer settings. Additionally, utilising information from single end reads can aid with bacterial read retention. $k=30-32$ appears to be an optimal setting for retaining bacterial reads and removing human reads.

B.0.3 Bacterial Read Retention

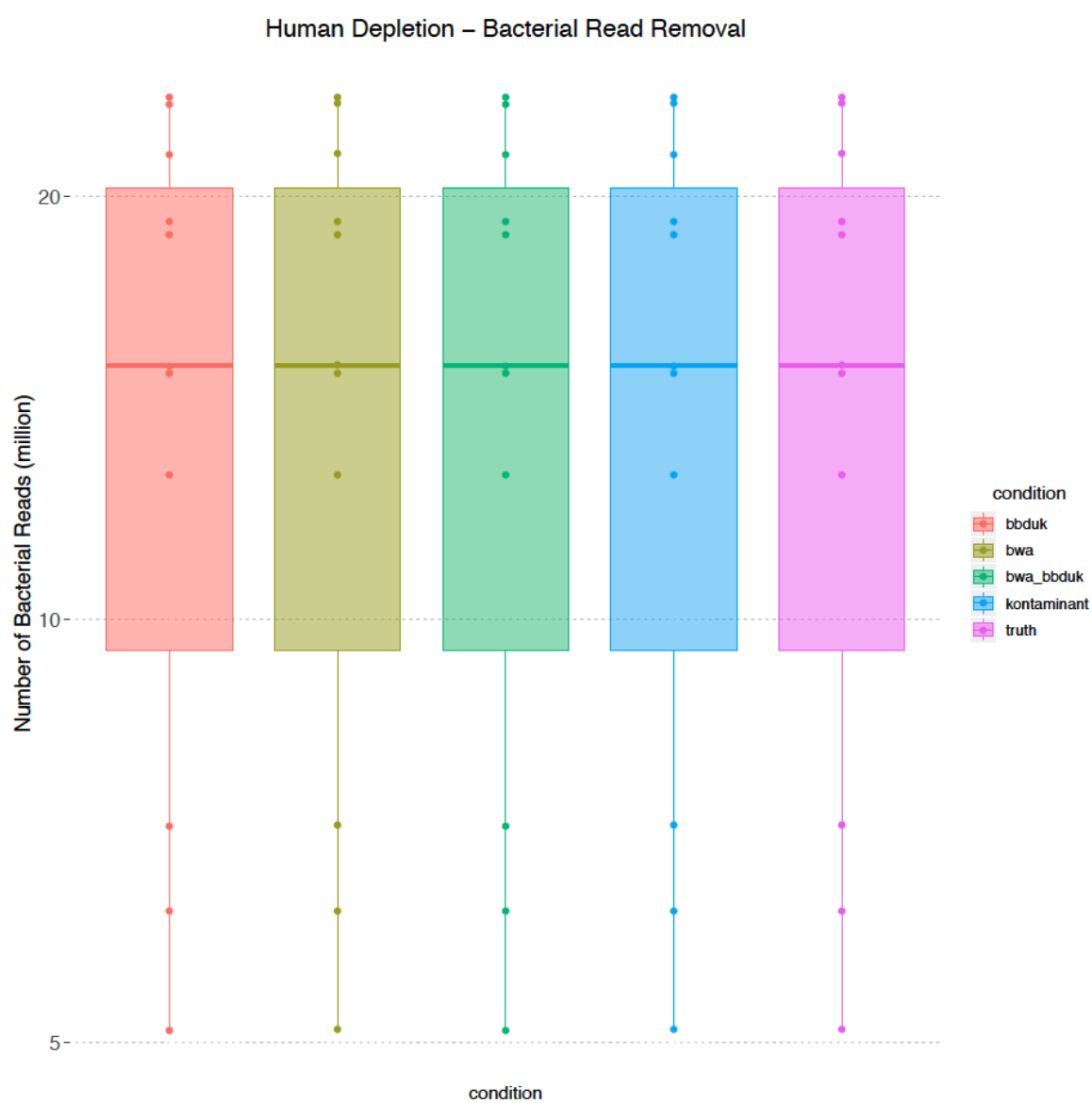


Figure B.2: Retention of bacterial reads using different depletion software.

B.0.4 Contig Coverage in Simulations

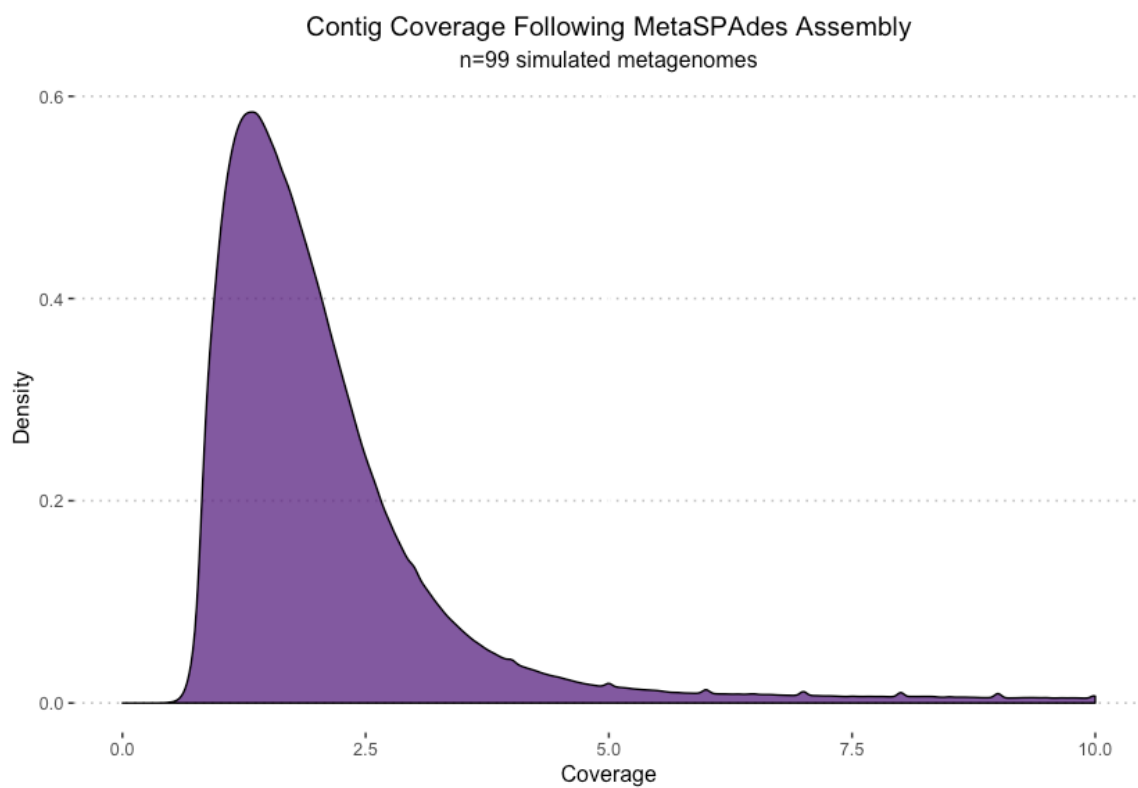


Figure B.3: Coverage of contigs following metagenomic assembly on 99 simulated metagenomes. Higher values not shown in density plot.

B.0.5 Classification Performance by Contig Coverage

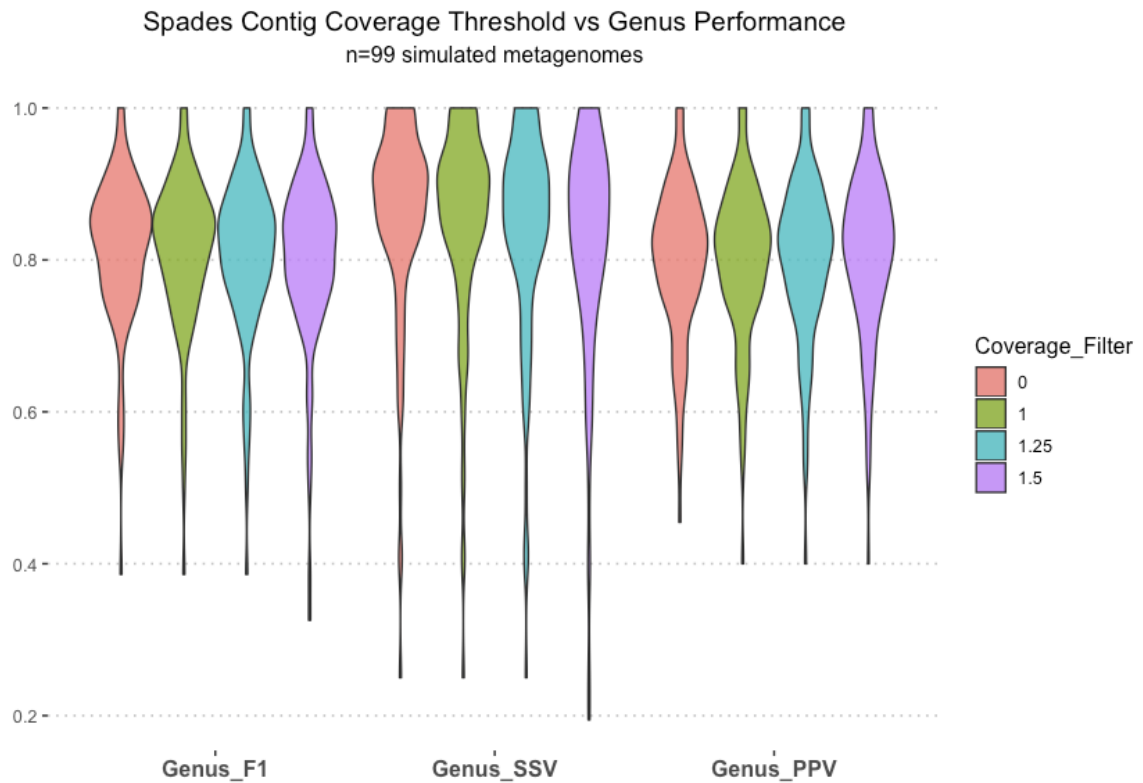


Figure B.4: Violin plot shows genus level performance with increasing minimum contig coverage filter but not to a large degree. Tool used was Kraken on MetaSPAdes contigs.

B.0.6 KrakenUniq Filtering Parameters

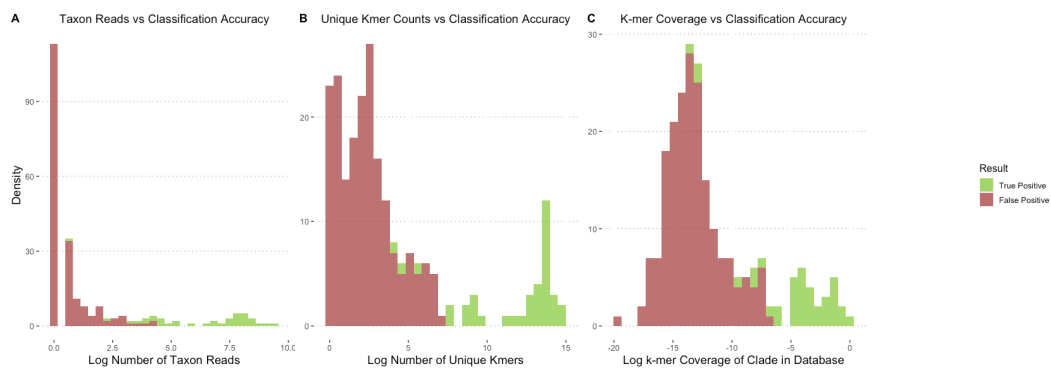


Figure B.5: An in depth look into Krakenuniq filtering parameters vs bacterial classification status for one simulated bacterial dataset.

B.0.7 Viral Dataset Classification



Figure B.6: A more in-depth look into contig parameters vs classification status for one of the viral datasets assembled using MetaSPAdes and classified using Kraken.

Appendix C

Prostate Cancer

C.0.1 D'Amico Classification system

Table C.1: D'amico classification system for prostate cancer

Risk	PSA	Gleason	Clinical_stage	DRE
Low	<10 ng/ml	6 or under	T1-T2a	Localised disease
Intermediate	10-20 ng/ml	7	T2b or T2c	Localised disease
High	>20 ng/ml	8-10	T3-T4	Localised or locally advanced disease

C.1 Urine RNA NMDS Scree and Stress Plot

The optimal number of clusters was identified by finding the ‘*elbow*’ in the scree plot below. This point was selected as $k = 2$ when investigating $k = [1, 10]$. The rationale behind this method is provided in figure 1.6. Two dimensions was selected for ease of visualisation.

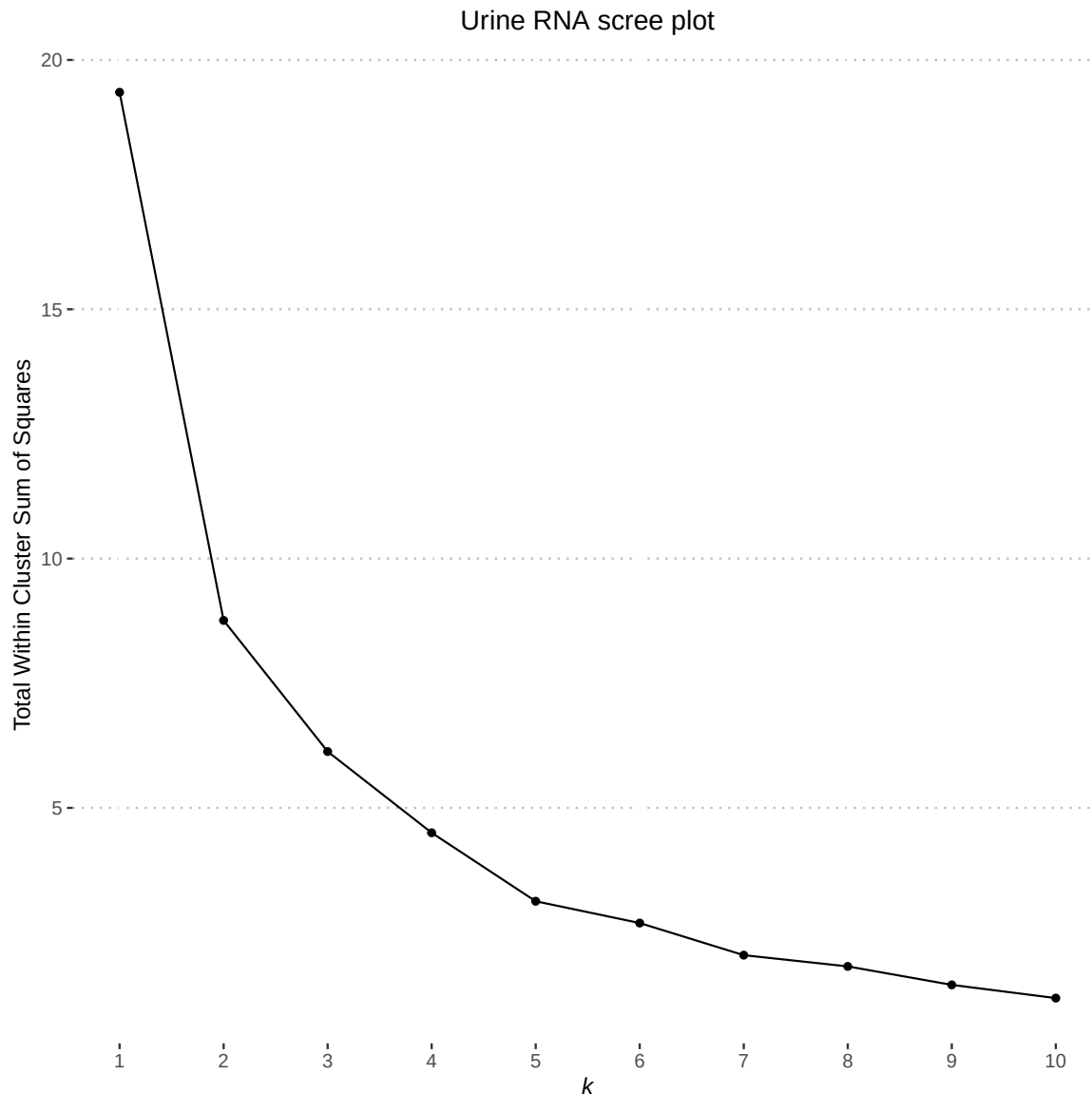


Figure C.1: Urine RNA sequencing scree plot used to determine the optimal number of clusters.

To identify differences between clusters, contingency tables of the binary presence-absence community matrix were formed for each taxa and a Fisher's exact test was carried out in each case (figure C.2). The focus on this analysis is the taxa that are more likely than others to be differentially present between clusters as opposed to a focus on the p -values in particular. Therefore false discovery correction has not been applied at this time. Cluster 'B' appears to contain much more prevalent bacteria. Cluster 'A' appears to be defined by a general absence of genera with only 11 genera displaying a non-significant preference for cluster 'A' (all with p -values without multiple testing correction >0.3). Cluster 'B' is more diverse in terms of the number of genera determined as present in each sample. 90 Genera demonstrated a preference for cluster B, 12 of which seem more significant than the others with a maximum p -value for these genera of 0.007 (without false discovery correction). Additionally, these findings do not appear to be associated with sequencing depth (figure

C.3).

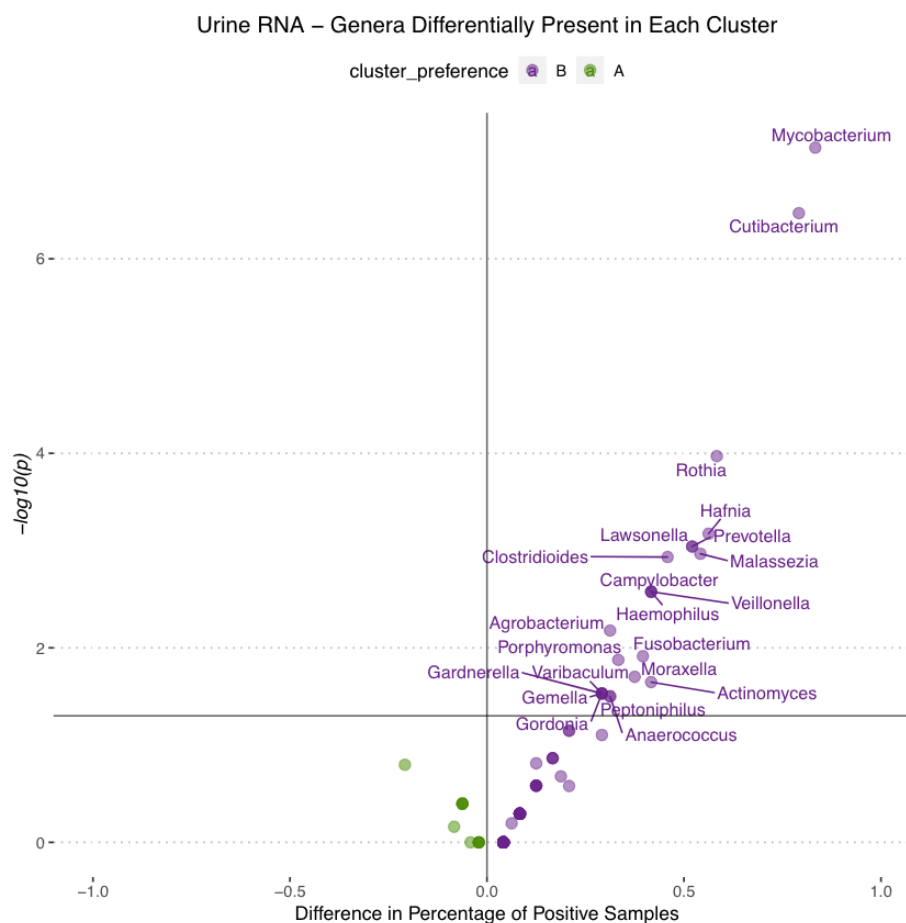


Figure C.2: Analysing differentially present taxa between clusters. Differences in proportions of positive samples between clusters was carried out (Fishers exact test) and is indicated on the y -axis. To investigate directionality, the difference in percentage of positive samples was calculated whereby a value of +1 indicates cluster ‘A’ samples all contain a taxa and none of the samples in cluster ‘B’ contain the taxa. The horizontal line indicates $\alpha=0.05$.

C.2 Urine RNA differences in diversity appear independent of sequencing depth

Comparing the number of sequencing reads in the original file to the number of genera detected in each sample, there appears to be a poor correlation. One sample was removed from analysis because it contained in excess of 80,000,000 reads and was an outlier.

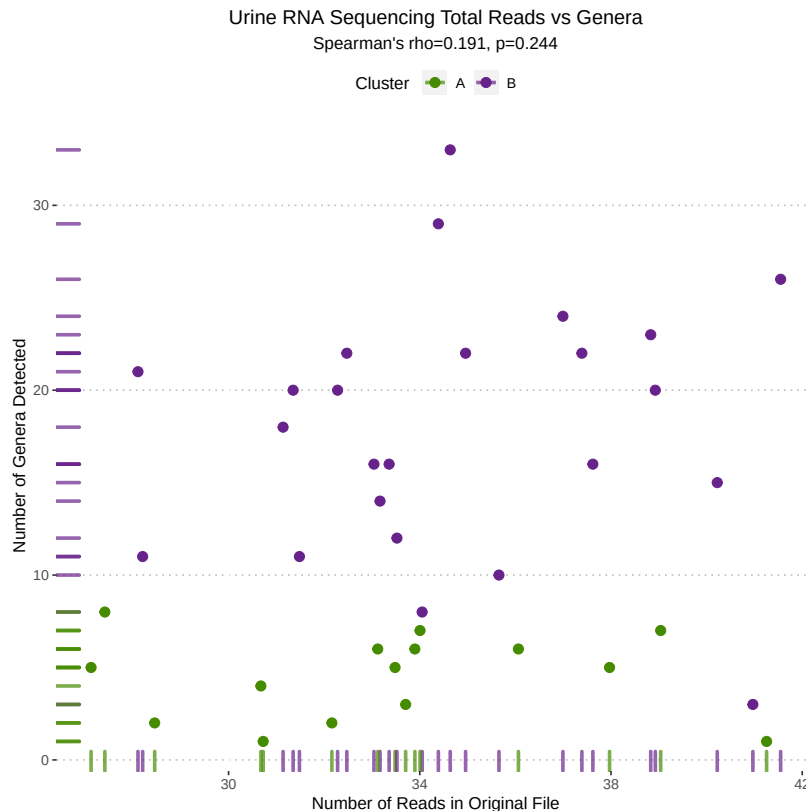


Figure C.3: Urine RNA seq genera diversity appears independent of sequencing depth

C.3 Differentially Expressed Human Genes between NMDS Clusters

Table C.2: Differentially Expressed human genes between NMDS cluster.
Maximum FDR adjusted Q -value for differential expression = 0.1

geneID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
IGHA1	78.651878	25.1418653	1.7717072	14.190756	0.0000000	0.0000000
PCP4L1	40.856240	3.5242331	0.6581527	5.354735	0.0000001	0.0009709
HLA-DQA1	192.985389	1.7076005	0.3312294	5.155342	0.0000003	0.0011519
HLA-DRA	522.273694	2.9877074	0.5889772	5.072705	0.0000004	0.0011519
MT-CO1	23258.889921	2.4116145	0.4760551	5.065830	0.0000004	0.0011519
MT-CO3	8425.791679	2.7787789	0.5334696	5.208880	0.0000002	0.0011519
PIGR	822.145244	3.3469861	0.6509142	5.141977	0.0000003	0.0011519
RPL10P9	91.643216	5.2341902	1.0242169	5.110431	0.0000003	0.0011519
CCL5	29.369828	5.6286349	1.1261053	4.998320	0.0000006	0.0013025
CD74	523.688212	2.9804290	0.5983453	4.981118	0.0000006	0.0013025
MT-CO2	3772.697260	2.3684855	0.4741259	4.995478	0.0000006	0.0013025
MT-CYB	5823.998206	2.4493257	0.4939456	4.958695	0.0000007	0.0013403

MT-TP	73.245424	1.9351121	0.3951642	4.896982	0.0000010	0.0016966
RARRES1	85.770878	2.9288709	0.6045344	4.844838	0.0000013	0.0020512
MT-ND1	2617.160465	2.3226257	0.4854867	4.784119	0.0000017	0.0021623
MT-ND3	1402.577336	2.5735456	0.5378213	4.785131	0.0000017	0.0021623
RAPGEFL1	107.322835	1.8656965	0.3882743	4.805098	0.0000015	0.0021623
SLC14A1	77.331452	2.3473428	0.4895867	4.794540	0.0000016	0.0021623
IGKC	106.735896	9.6526255	2.0291482	4.756984	0.0000020	0.0023439
MT-ATP6	2744.353188	2.2140636	0.4717231	4.693566	0.0000027	0.0030423
EPB41L2	74.500037	1.9919648	0.4275398	4.659134	0.0000032	0.0034269
SPG7	1905.116229	1.1330273	0.2461620	4.602771	0.0000042	0.0042947
ZPR1	5831.419173	1.2515345	0.2725879	4.591307	0.0000044	0.0043403
DCN	11.049349	5.0454071	1.1101539	4.544782	0.0000055	0.0046633
KRT19	2257.594808	1.1504741	0.2520552	4.564374	0.0000050	0.0046633
MT-ND2	2336.479634	2.1901486	0.4813428	4.550081	0.0000054	0.0046633
NCKAP1L	48.479133	3.9102633	0.8607930	4.542629	0.0000056	0.0046633
METTL7A	256.765386	0.8141824	0.1800245	4.522620	0.0000061	0.0049437
MT-ND4	7527.008734	2.1010179	0.4689400	4.480356	0.0000075	0.0056633
MT-ND4L	951.781140	2.0385008	0.4558296	4.472067	0.0000077	0.0056633
SERPINB1	146.418061	2.3329555	0.5213657	4.474701	0.0000077	0.0056633
JCHAIN	88.061155	7.2915511	1.6489675	4.421889	0.0000098	0.0067659
TNNT3	24.227311	7.1733710	1.6227890	4.420397	0.0000099	0.0067659
ACSL5	82.524907	1.4515151	0.3295124	4.405040	0.0000106	0.0069927
ASS1	1052.823740	1.3542587	0.3079146	4.398163	0.0000109	0.0069927
MT-ND5	4361.780591	1.8106837	0.4120434	4.394401	0.0000111	0.0069927
PAQR8	31.524711	1.9915557	0.4540795	4.385919	0.0000115	0.0070223
S100A6	590.813539	1.7630763	0.4023707	4.381721	0.0000118	0.0070223
PCDH7	36.873352	2.7213189	0.6220625	4.374671	0.0000122	0.0070671
MT-ATP8	232.340411	2.3809517	0.5455141	4.364601	0.0000127	0.0072156
ETS1	207.416932	1.7011581	0.3914568	4.345711	0.0000139	0.0076511
FAT2	37.480375	2.7479658	0.6337660	4.335931	0.0000145	0.0076511
GPX2	146.246159	2.6880927	0.6199628	4.335894	0.0000145	0.0076511
DOCK10	33.829194	3.2067508	0.7424626	4.319074	0.0000157	0.0080704
MT-RNR2	523110.214856	2.2254329	0.5164084	4.309444	0.0000164	0.0082426
CORO1A	40.242933	5.3516436	1.2435327	4.303581	0.0000168	0.0082798
TSPAN3	207.117668	1.2627401	0.2951226	4.278697	0.0000188	0.0090647
TRIM29	224.893799	2.9294737	0.6866629	4.266247	0.0000199	0.0093857
ATP2B4	45.430390	3.3476242	0.7911623	4.231274	0.0000232	0.0097337
CTSC	75.781921	1.4165354	0.3342125	4.238428	0.0000225	0.0097337
GAPT	20.687977	5.7972804	1.3686985	4.235615	0.0000228	0.0097337
H19	1200.374417	4.6870531	1.1075058	4.232080	0.0000232	0.0097337
KRT15	230.202468	3.6576103	0.8651794	4.227575	0.0000236	0.0097337
SGK1	202.500285	1.8815855	0.4449408	4.228844	0.0000235	0.0097337
STAB1	36.635695	5.4687603	1.2933005	4.228530	0.0000235	0.0097337
AC141586.3	7.650761	-5.4668738	1.2969772	-4.215088	0.0000250	0.0101045

AL135790.1	1119.771910	1.2468074	0.2969793	4.198296	0.0000269	0.0105425
CARD6	88.934181	2.4066366	0.5733421	4.197558	0.0000270	0.0105425
ZNF486	270.087479	0.9841252	0.2352596	4.183145	0.0000288	0.0110436
ETS2	144.517294	1.5658692	0.3764943	4.159078	0.0000320	0.0120694
ABCA1	106.347210	1.3647309	0.3290808	4.147099	0.0000337	0.0125098
MTRNR2L8	20.647407	3.1536196	0.7629033	4.133708	0.0000357	0.0130479
CLDN4	327.054066	0.6085207	0.1473553	4.129614	0.0000363	0.0130716
DOCK2	24.863125	5.2886800	1.2888742	4.103333	0.0000407	0.0142855
PHYHIP	7.170444	5.3942421	1.3150507	4.101927	0.0000410	0.0142855
LRMP	73.545016	2.7889069	0.6839105	4.077883	0.0000454	0.0156058
PPM1L	187.162104	1.2483839	0.3065115	4.072878	0.0000464	0.0157070
ADGRG6	62.708073	2.1526367	0.5307976	4.055475	0.0000500	0.0166748
TNFSF15	121.097159	0.9596361	0.2368471	4.051712	0.0000508	0.0166997
WNT5A	183.441890	2.0126283	0.4971796	4.048091	0.0000516	0.0167179
KYNU	108.028737	1.2829099	0.3173250	4.042889	0.0000528	0.0168525
AXL	40.399967	2.5063995	0.6232841	4.021279	0.0000579	0.0179699
MIR205HG	63.184430	3.5118380	0.8732840	4.021416	0.0000578	0.0179699
HLA-A	415.322632	0.8968498	0.2239811	4.004131	0.0000622	0.0185615
JAK3	21.200655	3.2995184	0.8239251	4.004634	0.0000621	0.0185615
MT-ND6	702.124018	1.6915896	0.4221629	4.006959	0.0000615	0.0185615
CYBB	78.388892	2.5837573	0.6489788	3.981266	0.0000685	0.0200592
PLAT	62.587291	2.1368641	0.5369578	3.979576	0.0000690	0.0200592
FLNA	131.616923	1.0408249	0.2622465	3.968880	0.0000722	0.0203719
HLA-DQB1	59.619602	2.9395409	0.7408774	3.967648	0.0000726	0.0203719
LAMB3	32.820742	2.5757046	0.6492979	3.966907	0.0000728	0.0203719
TRIM31	44.264117	2.5792471	0.6510822	3.961477	0.0000745	0.0205867
IL7R	27.381899	3.0877571	0.7814874	3.951128	0.0000778	0.0212387
KLRD1	33.297979	1.8841529	0.4772522	3.947919	0.0000788	0.0212691
SCGB3A1	212.707682	2.8305773	0.7187243	3.938335	0.0000820	0.0218762
FMNL3	44.908128	2.6794771	0.6826086	3.925349	0.0000866	0.0228222
CHST2	27.948115	3.1392243	0.8059398	3.895110	0.0000982	0.0255686
RN7SK	6210.226306	1.3366571	0.3466703	3.855701	0.0001154	0.0291758
RNA5S17	21.655528	2.1321594	0.5531300	3.854717	0.0001159	0.0291758
TNFRSF21	324.982586	1.2342051	0.3197904	3.859419	0.0001137	0.0291758
AHR	536.624171	0.8242373	0.2144862	3.842846	0.0001216	0.0299584
CP	48.092302	3.1401580	0.8167026	3.844922	0.0001206	0.0299584
SRD5A2	89.460543	3.9614601	1.0320859	3.838305	0.0001239	0.0301898
C3	344.972077	1.0011546	0.2615084	3.828384	0.0001290	0.0310983
MEF2C	57.231182	1.6808130	0.4400420	3.819665	0.0001336	0.0312219
MT-RNR1	178509.199919	1.9283227	0.5048360	3.819701	0.0001336	0.0312219
MTATP6P1	240.789143	1.9342277	0.5059784	3.822747	0.0001320	0.0312219
ABI3	17.005203	4.7169197	1.2378032	3.810719	0.0001386	0.0320435
TACSTD2	406.180488	0.8266974	0.2179409	3.793219	0.0001487	0.0340419
CAV2	107.109926	1.6456474	0.4348953	3.784008	0.0001543	0.0344941

CTSB	447.966854	0.8698976	0.2299783	3.782520	0.0001552	0.0344941
TNFAIP2	358.314863	1.6826555	0.4447425	3.783438	0.0001547	0.0344941
CD8A	22.569502	2.6591893	0.7046494	3.773776	0.0001608	0.0351413
FP671120.4	507.403450	1.5964334	0.4231146	3.773052	0.0001613	0.0351413
IFI16	106.269999	2.0143701	0.5359510	3.758497	0.0001709	0.0368948
SLCO2B1	25.673353	3.2524591	0.8674640	3.749388	0.0001773	0.0378999
CEACAM7	5.637436	5.7942518	1.5487037	3.741356	0.0001830	0.0385433
GGNBP2	1392.156872	-0.2014896	0.0538675	-3.740469	0.0001837	0.0385433
JAKMIP2	29.258330	3.0377032	0.8138644	3.732444	0.0001896	0.0390692
PHOSPHO1	10.649095	3.4427750	0.9222202	3.733138	0.0001891	0.0390692
ALDH3B1	78.855123	1.5616235	0.4191644	3.725563	0.0001949	0.0397888
ARNTL2	70.422724	1.5935484	0.4294720	3.710482	0.0002069	0.0400612
FGD2	25.184830	2.4201916	0.6524557	3.709358	0.0002078	0.0400612
INPP5D	53.736091	2.5472387	0.6858357	3.714066	0.0002040	0.0400612
LPCAT2	31.329363	3.3711807	0.9073925	3.715240	0.0002030	0.0400612
SOX2	76.987580	1.8186332	0.4899077	3.712195	0.0002055	0.0400612
ST6GALNAC2	20.385429	3.0607857	0.8237514	3.715667	0.0002027	0.0400612
TGFBI	64.432913	2.9134596	0.7856417	3.708382	0.0002086	0.0400612
CLU	63.990187	1.4125394	0.3811731	3.705768	0.0002108	0.0401365
A2M	102.131430	4.3607024	1.1780766	3.701544	0.0002143	0.0404707
S100P	305.456767	1.6645299	0.4508095	3.692313	0.0002222	0.0416219
CXADR	382.379176	0.8546293	0.2320234	3.683376	0.0002302	0.0427560
CLCA4	59.853628	4.3599385	1.1879527	3.670128	0.0002424	0.0446680
MT-TY	38.086352	2.0846347	0.5687708	3.665157	0.0002472	0.0451773
AC090617.3	19.095931	6.0162820	1.6425055	3.662869	0.0002494	0.0452184
ARHGDIB	231.510256	2.1996761	0.6014870	3.657063	0.0002551	0.0456235
PCDH20	6.773383	4.8023345	1.3140867	3.654504	0.0002577	0.0456235
SPRYD4	473.962231	-0.3492058	0.0955118	-3.656152	0.0002560	0.0456235
GRN	143.847154	0.8103962	0.2222590	3.646180	0.0002662	0.0467610
CXCR6	6.550978	5.2721508	1.4520278	3.630888	0.0002824	0.0491191
MS4A6A	44.767668	3.9138563	1.0783336	3.629541	0.0002839	0.0491191
AL365361.1	5.924609	5.8677939	1.6176599	3.627335	0.0002864	0.0491653
DGKH	175.308295	0.7914317	0.2193900	3.607420	0.0003093	0.0523036
LINC01133	74.619292	1.3297574	0.3684457	3.609100	0.0003073	0.0523036
PEAK1	94.568669	1.1179715	0.3101604	3.604495	0.0003128	0.0525043
RNA5S1	19.640037	1.5729058	0.4377738	3.592964	0.0003269	0.0529241
RNA5S16	19.640037	1.5729058	0.4377738	3.592964	0.0003269	0.0529241
RNA5S2	19.640037	1.5729058	0.4377738	3.592964	0.0003269	0.0529241
RNA5S3	19.640037	1.5729058	0.4377738	3.592964	0.0003269	0.0529241
SYTL3	766.377607	-0.4360386	0.1213400	-3.593529	0.0003262	0.0529241
POF1B	548.524102	0.6018246	0.1676181	3.590451	0.0003301	0.0530582
GIMAP6	31.306592	2.6882320	0.7507023	3.580956	0.0003423	0.0535064
GPRC5A	83.861563	1.5996403	0.4460298	3.586398	0.0003353	0.0535064
NPTN	159.937097	0.6140054	0.1714499	3.581253	0.0003420	0.0535064

USP43	1543.287550	-0.4804089	0.1340215	-3.584566	0.0003376	0.0535064
KRT5	194.173123	3.3772364	0.9456891	3.571191	0.0003554	0.0551614
ALOX5	475.625548	0.5725739	0.1604264	3.569075	0.0003582	0.0552304
IGF2BP2	38.190725	1.8097208	0.5081284	3.561542	0.0003687	0.0564558
ASPH	189.654165	1.0903863	0.3067292	3.554882	0.0003781	0.0575166
FGL2	65.603355	3.1803544	0.8952915	3.552312	0.0003819	0.0576943
GPRIN3	42.793184	2.3508943	0.6621388	3.550456	0.0003846	0.0577179
ROBO1	197.437459	1.5476806	0.4363404	3.546957	0.0003897	0.0581050
FCER1A	19.896893	5.7139121	1.6135092	3.541295	0.0003982	0.0589783
CD4	35.366968	1.5947577	0.4508865	3.536938	0.0004048	0.0593954
SLA	25.891778	2.8256866	0.7991181	3.536006	0.0004062	0.0593954
KRTCAP2	191.856870	0.5103239	0.1445849	3.529580	0.0004162	0.0597012
TNFRSF1B	17.704027	3.1487859	0.8919959	3.530045	0.0004155	0.0597012
TOR4A	15.724458	2.3937490	0.6779683	3.530769	0.0004144	0.0597012
CTSD	388.424426	0.8408537	0.2395025	3.510834	0.0004467	0.0613554
GPR174	14.986368	3.8286253	1.0893022	3.514750	0.0004402	0.0613554
KLF5	1237.773955	0.7492385	0.2129464	3.518437	0.0004341	0.0613554
PLAC8	27.597753	1.9412064	0.5525201	3.513368	0.0004425	0.0613554
RBPMS	111.003148	1.1350889	0.3227723	3.516686	0.0004370	0.0613554
RCSD1	20.566581	2.7567392	0.7851834	3.510950	0.0004465	0.0613554
S100A16	427.415903	0.7252863	0.2061679	3.517940	0.0004349	0.0613554
DAPP1	66.540468	1.8330456	0.5227911	3.506268	0.0004544	0.0620417
SLITRK6	44.072344	3.2483092	0.9271016	3.503725	0.0004588	0.0622620
IRX5	139.921433	0.7278160	0.2078880	3.501000	0.0004635	0.0625277
AC138150.1	8.498046	6.3883820	1.8267561	3.497118	0.0004703	0.0630243
LINC01609	18.448592	3.3598861	0.9611388	3.495735	0.0004728	0.0630243
LY86	5.673700	5.2471699	1.5025755	3.492117	0.0004792	0.0635104
COL1A2	11.065043	5.6946578	1.6359479	3.480953	0.0004996	0.0658325
ROCK1P1	228.842320	0.9388568	0.2701591	3.475200	0.0005105	0.0668720
ITGAM	10.715956	3.5196754	1.0148517	3.468167	0.0005240	0.0676012
MSN	545.863069	0.6010864	0.1732282	3.469912	0.0005206	0.0676012
RASAL3	15.844941	2.7045309	0.7799271	3.467672	0.0005250	0.0676012
ANXA8	40.256160	4.3787308	1.2669338	3.456164	0.0005479	0.0685770
ARHGAP23	143.961087	1.6571834	0.4795896	3.455420	0.0005494	0.0685770
CTSS	102.181635	2.3745675	0.6864965	3.458965	0.0005423	0.0685770
ELF3	1073.742060	0.8721463	0.2524455	3.454790	0.0005507	0.0685770
FUT9	23.746312	2.0150688	0.5827478	3.457875	0.0005445	0.0685770
KRT13	1253.533894	3.0596661	0.8855814	3.454980	0.0005503	0.0685770
AKR1C2	63.441418	2.5541443	0.7427928	3.438569	0.0005848	0.0719334
DENND5A	767.373353	-0.2202915	0.0640651	-3.438559	0.0005848	0.0719334
ITGAV	273.658569	0.4842627	0.1408781	3.437460	0.0005872	0.0719334
KLHL6	22.345978	2.2052650	0.6434258	3.427380	0.0006094	0.0732988
LO000005.1	68.126503	1.0943512	0.3193733	3.426559	0.0006113	0.0732988
NAB1	609.273113	0.4983531	0.1453547	3.428530	0.0006069	0.0732988

SDS	6.659906	4.0040366	1.1670865	3.430797	0.0006018	0.0732988
TP63	117.242682	4.5581402	1.3312463	3.423965	0.0006171	0.0736126
ATP2B1	176.522598	0.8332615	0.2439183	3.416149	0.0006351	0.0753614
LPAR5	6.529761	4.3860702	1.2850834	3.413063	0.0006424	0.0758233
INKA1	5.539760	5.2184134	1.5300290	3.410663	0.0006481	0.0760973
CXCL17	34.276727	3.8053550	1.1162174	3.409152	0.0006517	0.0761256
TREM2	8.195561	6.3346854	1.8607172	3.404432	0.0006630	0.0770562
DTX4	378.637004	1.1432294	0.3366116	3.396286	0.0006831	0.0772837
FGFR2	106.880508	1.4506566	0.4267742	3.399120	0.0006760	0.0772837
GAB3	13.864179	2.2546268	0.6646755	3.392071	0.0006937	0.0772837
HLA-DPA1	335.469848	1.5902926	0.4683597	3.395451	0.0006852	0.0772837
MPZL2	226.305637	0.6919839	0.2037855	3.395647	0.0006847	0.0772837
NEAT1	293.916095	0.5397530	0.1596209	3.381468	0.0007210	0.0772837
RNA5S10	9.813440	1.6041577	0.4746826	3.379432	0.0007264	0.0772837
RNA5S11	9.813440	1.6041577	0.4746826	3.379432	0.0007264	0.0772837
RNA5S12	9.813440	1.6041577	0.4746826	3.379432	0.0007264	0.0772837
RNA5S13	9.813440	1.6041577	0.4746826	3.379432	0.0007264	0.0772837
RNA5S14	9.813440	1.6041577	0.4746826	3.379432	0.0007264	0.0772837
RNA5S15	9.813440	1.6041577	0.4746826	3.379432	0.0007264	0.0772837
RNA5S4	9.813440	1.6041577	0.4746826	3.379432	0.0007264	0.0772837
RNA5S5	9.813440	1.6041577	0.4746826	3.379432	0.0007264	0.0772837
RNA5S6	9.813440	1.6041577	0.4746826	3.379432	0.0007264	0.0772837
RNA5S7	9.813440	1.6041577	0.4746826	3.379432	0.0007264	0.0772837
RNA5S8	9.813440	1.6041577	0.4746826	3.379432	0.0007264	0.0772837
SNCG	210.420719	1.7810037	0.5254035	3.389783	0.0006995	0.0772837
ADGRF1	129.973616	1.7889548	0.5298149	3.376566	0.0007340	0.0776483
ITGB6	54.510192	2.4970790	0.7397509	3.375567	0.0007366	0.0776483
NT5E	105.889939	1.2538700	0.3720534	3.370135	0.0007513	0.0788289
MS4A7	26.450813	3.2941079	0.9781891	3.367557	0.0007584	0.0792028
MSR1	25.557924	2.2744209	0.6756503	3.366270	0.0007619	0.0792084
CAPNS2	18.024526	4.6737011	1.3966384	3.346393	0.0008187	0.0839559
FAM30A	4.863323	5.5821229	1.6678821	3.346833	0.0008174	0.0839559
GPR65	6.265981	4.5823578	1.3687241	3.347905	0.0008142	0.0839559
TMPRSS4	52.863965	3.1330193	0.9369284	3.343926	0.0008260	0.0843245
GBP2	251.577005	1.2471207	0.3745241	3.329881	0.0008688	0.0882975
PDLIM1	391.869527	0.9442905	0.2840350	3.324556	0.0008856	0.0895992
MAATS1	99.926871	0.9269900	0.2789723	3.322875	0.0008909	0.0897402
AL391069.2	3.569641	-3.8334296	1.1549774	-3.319052	0.0009032	0.0901765
KLRB1	17.353461	7.4880353	2.2559491	3.319240	0.0009026	0.0901765
SDCBP2	88.809899	0.8716481	0.2639599	3.302199	0.0009593	0.0953537
LYZ	127.697670	2.1129886	0.6401875	3.300578	0.0009649	0.0954874
CD9	668.785130	0.3182638	0.0964837	3.298627	0.0009716	0.0957352
ALCAM	764.389129	0.4368090	0.1325652	3.295051	0.0009840	0.0965422
LGALS9	45.530137	2.0812724	0.6328835	3.288555	0.0010070	0.0983721

LCN2	115.804126	2.1124501	0.6427740	3.286459	0.0010146	0.0986820
WNT10A	6.519044	4.3967793	1.3386285	3.284540	0.0010215	0.0989316
HOPX	74.093108	1.2312790	0.3752886	3.280886	0.0010348	0.0997960

C.4 Discovering the Anaerobic bacterial biomarker set (ABBS)

A team effort was carried out to identify a list of genus level taxa that might contribute towards a poor prognosis of prostate cancer which was termed the Anaerobic Bacterial Biomarker Set (ABBS) and consists of *Peptoniphilus*, *Porphyromonas*, *Anaerococcus*, *Fenollaria*/*Ezakiella* and *Fusobacterium*.

A list of seven bacterial genera was provided by Dr Rachel Hurst based on genera assignment of the 16S dataset, predicted mechanisms of action of the bacteria and preliminary interpretation of the urine RNAseq dataset. The bacteria in the list were identified due to prediction to be associated with aggressive prostate cancer. The list consisting of *Porphyromonas*, *Anaerococcus*, *Fusobacterium*, *Peptoniphilus*, *Ezakiella*(*Fenollaria*), (ABBS above) *Actinotignum*, and *Finegoldia* (the latter two were only present in very few samples or as a marker of anaerobic urine microbiota).

Of all genera detected in the 16S data I found and confirmed that genera *Ezakiella*, *Peptoniphilus*, *Porphyromonas* and *Anaerococcus* were the four most differentially abundant bacterial genera with elevated proportions in the poor prognosis cluster of 16S samples compared to the remainder of the 16S dataset ($Q = 0.03, 0.06, 0.03$ and 0.5 for *Ezakiella*, *Peptoniphilus*, *Porphyromonas* and *Anaerococcus* respectively, Benjamini-Hochberg corrected Wilcoxon signed-rank test. This suggested that those four genera should be subject to further investigation. We additionally decided to include *Fenollaria* in this list as advised by Dr Rachel Hurst because of work carried out by herself and Dr Emma Meader in isolating strains of *Fenollaria* from participants with aggressive prostate cancer. (Dr Ghanasyam Rallapalli determined that our *Fenollaria* strains were highly phylogenetically related to *Ezakiella*). Finally, *Fusobacterium* was included, as detailed in the list above provided by Dr Rachel Hurst, due to predicted mechanisms of action. In addition, we had isolated a relevant strain of bacteria belonging to *Fusobacterium* genera from clinical samples, *Fusobacterium nucleatum* which has also been associated with aggressive cancer at a range of anatomical sites^{55,184,296,545}.

I decided to partition the samples into groups with and without evidence of at least one of these ABBS genera. Applying this method of partitioning samples I demonstrated prominent differences in clinical outcome for those patients in the urine RNA sequencing dataset from NNUH (figure 3.7; clinical updates were kindly provided by Dr Kate Manley and the NNUH team). All of the participants with at least one of these genera present in their urine either showed signs of worsening disease (development of metastasis, an increase in histological staging, an increase in PSA sufficient enough to alter the course of treatment or a prostate cancer associated death) or they were censored in the RNA sequencing cohort ($p=0.0013$ and 0.0079 using log-rank tests, figure 3.7). Externally of the data presented in this thesis I also applied this method back to the original 16S data and to a targeted alignment approach on whole genome sequence data conducted by Dr Ghanasyam Rallapalli both of which demonstrated similar trends in survival ($p=0.045$, $p=0.035$ accordingly;

Log-rank test).

These figures below detail the analysis of differential abundance and constitute supplementary figures for a manuscript in preparation.

16S ribosomal subunit sequencing was carried out for $n=46$ urine exosome samples. Dr Rachel Hurst conducted the sample preparation and sequencing, Dr Ghanasyam Rallapalli wrote pipelines to process the data and obtain taxonomic classifications. The analysis was repeated, figures were updated and differential abundance was investigated by myself.

Samples were clustered by using the Manhattan distance (figure C.4). 48.9% of the variation in the dataset was explained by the first three principal coordinates and k -means clustering grouped the samples into three distinct clusters. Differential abundance investigations into these three clusters (figure C.5) helped to develop the anaerobic bacterial biomarker set (ABBS) which consists of *Ezakiella*, *Fenollaria*, *Peptoniphilus*, *Porphyromonas*, *Anaerococcus* And *Fusobacterium* was included due to mounting evidence of its association with cancer.

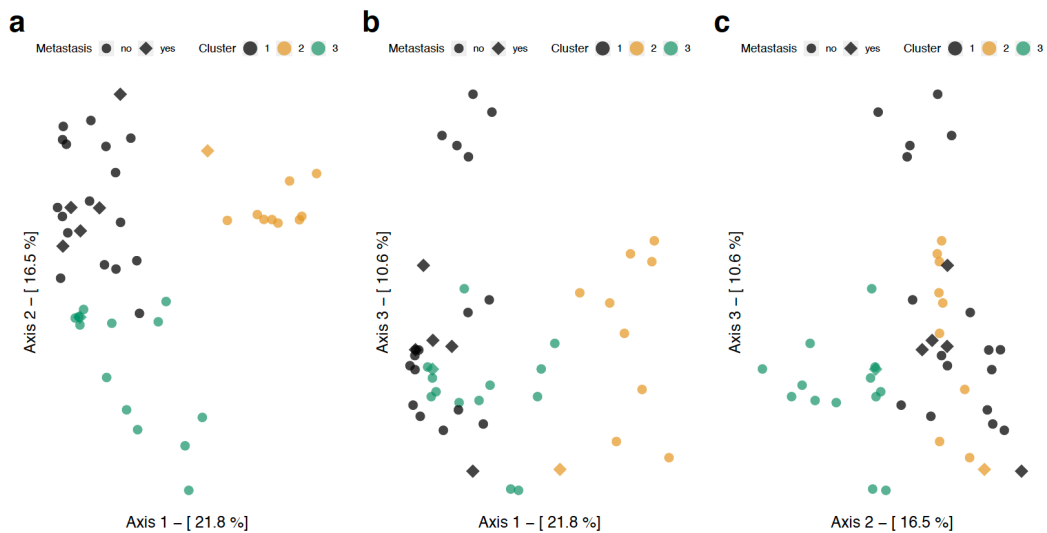


Figure C.4: Urine 16S PCoA (Manhattan distance). k -means clustering separated the data points into three groups. One of these groups (cluster 1, black) contained an extraordinarily large proportion of metastases as indicated by diamonds on the plot. The first three principal coordinates are shown through A-C.

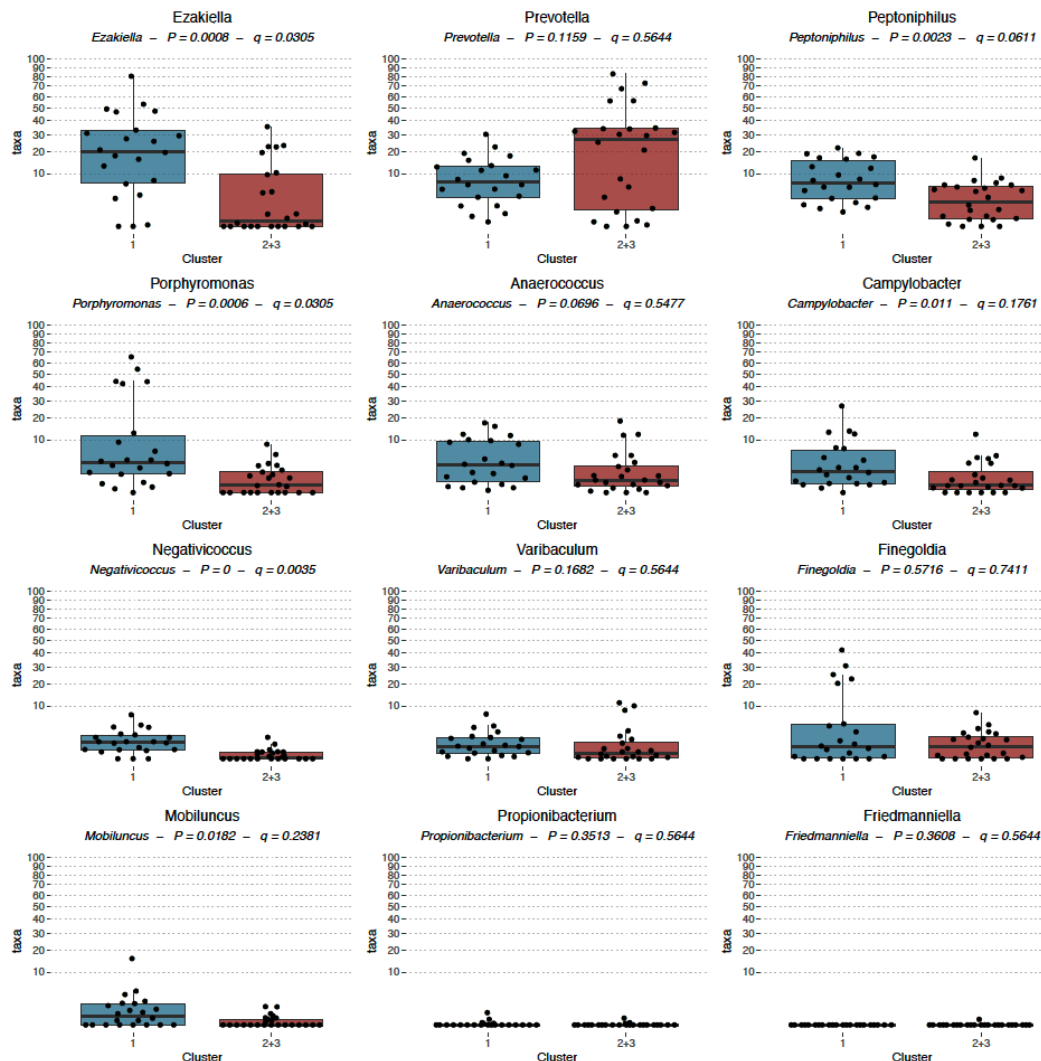


Figure C.5: Box plots to demonstrate differential abundance between the poor prognosis cluster (cluster 1) vs the rest of the dataset (clusters 2 and 3). p -values indicate Wilcoxon signed rank-sum test and Q -values indicate FDR adjusted p -values using Benjamini-Hochberg. Several taxa appear to be higher in abundance in the poor prognosis cluster. These taxa were selected to display the largest differences between clusters.

C.5 Investigating ABBS in a Separate Dataset

Dr Ghanasyam Rallapalli developed a targeted alignment approach to prostate tissue whole genome sequences based on strains that Dr Rachel Hurst and Emma Meader had isolated from the prostate. I investigated survival differences in this cohort after partitioning them by the presence of at least one ABBS genera. Progression events are as used elsewhere in this thesis and include any prostate cancer associated death, metastasis or biochemical recurrence. Overall the group containing ABBS demonstrated a significantly worse prognosis than the participants that had no group with no ABBS detected $p=0.035$.

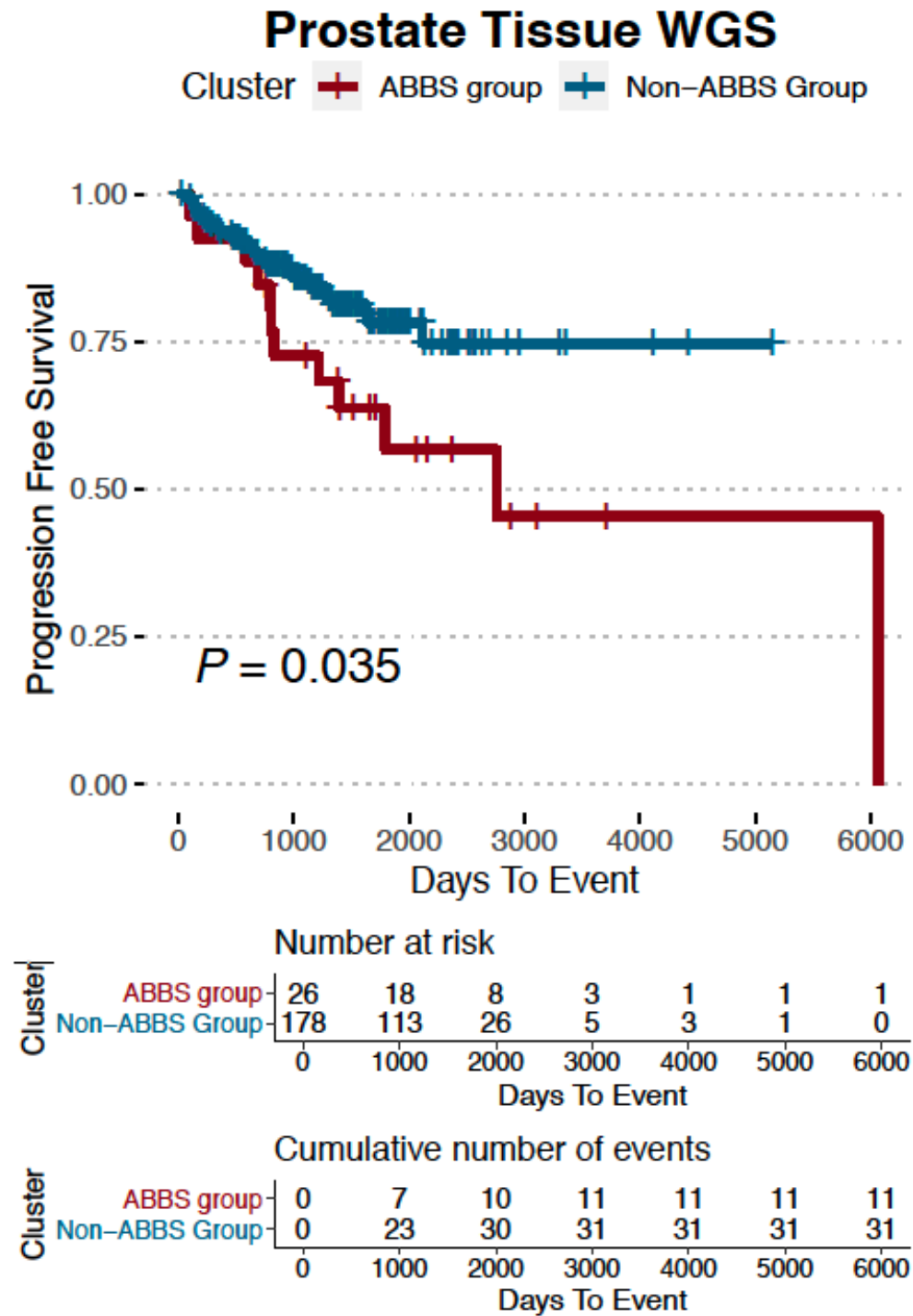


Figure C.6: Targeted alignment survival curve for ABBS bacteria

C.6 Investigating differentially Expressed human genes in urine

The following analysis investigates differential RNA expression in urine of participants under investigation for prostate cancer. The two groups formed based on the detection of at least one of the ABBS genera in their urine (*Esakiella*, *Fenollaria*, *Peptoniphilus*, *Porphyromonas*, *Anaerococcus* And *Fusobacterium*).

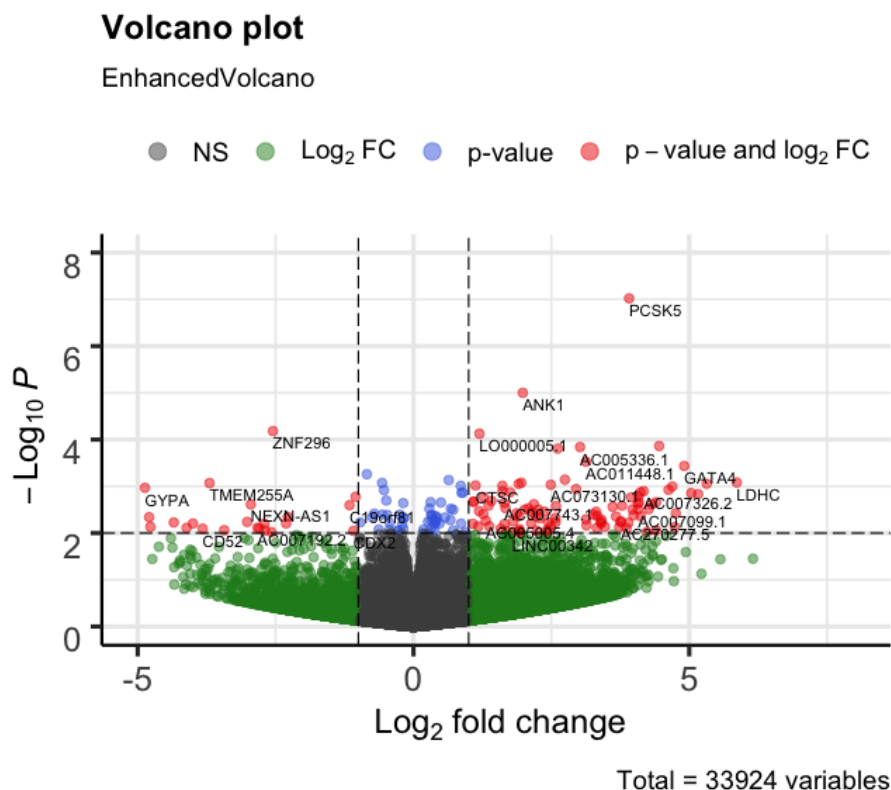


Figure C.7: A volcano plot of human differentially expressed genes in exosome urine RNA sequencing. Samples were partitioned based on the presence of at least one of the ABBS prognostic bacteria. A few genes seem significantly different between samples: *PCSK5* and *ANK1*. The horizontal line on the y -axis denotes $p=0.01$. The directionality is such that a positive log fold change indicates a higher expression in the ABBS group.

C.7 Differentially Expressed Human Genes between ABBS vs non-ABBS Samples

Table C.3: Differentially Expressed human genes between ABBS groups cluster. Maximum FDR adjusted Q -value for differential expression = 0.1

geneID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
--------	----------	----------------	-------	------	--------	------

PCSK5	7895.564	3.908973	0.7326237	5.33558	0.0000001	0.003121
-------	----------	----------	-----------	---------	-----------	----------

C.8 Differential Expression Pathway Enrichment Analysis - ABBS group differential

The following tables attempt to obtain a functional annotations for the most significantly differentially expressed genes.

Genes were filtered for $P < 0.01$ and a $\log_2$ fold change of > 1 or < -1 . They were then ran through the R package *clusterProfiler* for functional pathway enrichment. The resulting tables contain the following:

- **ONTOLOGY** - Ontology type. BP, CC, MF for Biological process, cellular component and molecular function respectively.
- **ID** - Pathway identification
- **Description** - A description of each pathway
- **GeneRatio** - $\frac{k}{n}$, where k is the number of genes in common in the input query and a geneset with a particular function and n denotes the all of the potential genes associated with that particular pathway. Therefore the higher the *GeneRatio* the higher the likelihood that the pathway mentioned is truly enriched.
- **BgRatio** - $\frac{M}{N}$, Where M is the number of genes annotated in that particular geneset and N is the total size of all unique genes in the collection of genesets associated with that pathway (all of the genes with known functions in an organism). Therefore a higher *BgRatio* denotes a higher likelihood that those pathways are truly enriched and a higher degree of that particular pathway in that particular organism.
- **pvalue** - p -value for pathway enrichment
- **p.adjust** - FDR adjusted p value
- **qvalue** - FDR adjusted p value
- **geneID** - Names of the genes highlighted in this dataset with the specified function

Note - There is only very limited documentation on the format of these output tables. And there is limited documentation available on exactly how the package formulates the p -value and what the difference between *p.adjust* and *qvalue* is despite it being possible to set the method of FDR control. Some of these columns have been removed to display all of the data effectively on an A4 page. Results have been ordered by the lowest *q-value* first. To limit the length of tables only the first 5 results are displayed here. The raw files can be found on the GitHub repository associated with this project.

C.8.1 Pathways, ABBS group differential - Gene Ontology Biological Process

Table C.4: Gene Ontology - Biological Process - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by ABBS group.

ID	Description	GeneRatio	BgRatio	pvalue
GO:0071621	granulocyte chemotaxis	5/48	123/18670	0.0000156
GO:0048565	digestive tract development	5/48	134/18670	0.0000236

GO:0097530	granulocyte migration	5/48	141/18670	0.0000302
GO:0055123	digestive system development	5/48	146/18670	0.0000357
GO:0048246	macrophage chemotaxis	3/48	38/18670	0.0001263

C.8.2 Pathways, ABBS group differential - Gene Ontology Cellular Component

Table C.5: Gene Ontology - Cellular Component - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by ABBS group.

ID	Description	GeneRatio	BgRatio	pvalue
GO:0071621	granulocyte chemotaxis	5/48	123/18670	0.0000156
GO:0048565	digestive tract development	5/48	134/18670	0.0000236
GO:0097530	granulocyte migration	5/48	141/18670	0.0000302
GO:0055123	digestive system development	5/48	146/18670	0.0000357
GO:0048246	macrophage chemotaxis	3/48	38/18670	0.0001263

C.8.3 Pathways, ABBS group differential - Gene Ontology Molecular Function

Table C.6: Gene Ontology - Cellular Component - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by ABBS group.

ID	Description	GeneRatio	BgRatio	pvalue
GO:0001085	RNA polymerase II transcription factor binding	4/46	155/17697	0.0006939
GO:0001614	purinergic nucleotide receptor activity	2/46	20/17697	0.0012190
GO:0016502	nucleotide receptor activity	2/46	20/17697	0.0012190
GO:0033613	activating transcription factor binding	3/46	85/17697	0.0013985
GO:0019865	immunoglobulin binding	2/46	24/17697	0.0017591

C.8.4 Pathways, ABBS group differential - Kegg

Table C.7: Kegg - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by ABBS group.

ID	Description	GeneRatio	BgRatio	pvalue
hsa04662	B cell receptor signaling pathway	3/26	82/8048	0.0022385
hsa05150	Staphylococcus aureus infection	3/26	96/8048	0.0035050
hsa04640	Hematopoietic cell lineage	3/26	99/8048	0.0038231
hsa04080	Neuroactive ligand-receptor interaction	5/26	341/8048	0.0041805
hsa00533	Glycosaminoglycan biosynthesis - keratan sulfate	1/26	14/8048	0.0443261

C.9 Differential expression pathway Enrichment Analysis- NMDS Cluster differential

These results are similar to those above, but instead of being partitioned by ABBS group of bacteria, they are partitioned by NMDS cluster.

C.9.1 Pathways, cluster group differential - gene ontology, biological process

Table C.8: Gene Ontology - Biological Process - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by NMDS cluster.

ID	Description	GeneRatio	BgRatio	pvalue
GO:0070661	leukocyte proliferation	29/305	298/18670	0
GO:0002696	positive regulation of leukocyte activation	30/305	380/18670	0
GO:0050867	positive regulation of cell activation	30/305	394/18670	0
GO:0046651	lymphocyte proliferation	25/305	272/18670	0
GO:0032943	mononuclear cell proliferation	25/305	274/18670	0

C.9.2 Pathways, cluster group differential - Gene Ontology Cellular Component

Table C.9: Gene Ontology - Cellular Component - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by NMDS cluster.

ID	Description	GeneRatio	BgRatio	pvalue
GO:0070661	leukocyte proliferation	29/305	298/18670	0
GO:0002696	positive regulation of leukocyte activation	30/305	380/18670	0
GO:0050867	positive regulation of cell activation	30/305	394/18670	0
GO:0046651	lymphocyte proliferation	25/305	272/18670	0
GO:0032943	mononuclear cell proliferation	25/305	274/18670	0

C.9.3 Pathways, cluster group differential - Gene Ontology Molecular Function

Table C.10: Gene Ontology - Cellular Component - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by NMDS Cluster.

ID	Description	GeneRatio	BgRatio	pvalue
GO:0032395	MHC class II receptor activity	6/316	10/17697	0.0000000

GO:0003735	structural constituent of ribosome	16/316	202/17697	0.0000007
GO:0023023	MHC protein complex binding	6/316	25/17697	0.0000041
GO:0003823	antigen binding	13/316	160/17697	0.0000061
GO:0048365	Rac GTPase binding	8/316	69/17697	0.0000307

C.9.4 Pathways, cluster group differential - Kegg

Table C.11: Kegg - Pathway analysis of most differentially expressed human genes in Urine RNA sequencing. Samples split by NMDS cluster.

ID	Description	GeneRatio	BgRatio	pvalue
hsa05150	Staphylococcus aureus infection	17/178	96/8048	0
hsa05140	Leishmaniasis	15/178	77/8048	0
hsa03008	Ribosome biogenesis in eukaryotes	17/178	111/8048	0
hsa04612	Antigen processing and presentation	13/178	78/8048	0
hsa04640	Hematopoietic cell lineage	14/178	99/8048	0

C.10 List of Isolates Interrogated

Table C.12: Isolate Names

isolate_name
a prevotii strain 8B
a prevotii strain 8S
a schaalii strain 41
a schaalii strain 43
a schaalii strain 48
a urogenitalis strain 44b
anaerococcus strain 8S
c acnes strain 5E
e faecalis strain 44a
f hominis strain 5
f nucleatum animalis
fenollaria sp nov strain 24
fenollaria sp nov strain 27A
fenollaria sp nov strain 27B
fenollaria sp nov strain 29
fenollaria sp nov strain 35
fenollaria strain 27A
p asaccharolytica strain 02-09
p asaccharolytica strain 22-07
p harei strain 25
p harei strain 28
p harei strain 33
p harei strain 37

p harei strain 38
p lymphophilum strain 10
p lymphophilum strain 11
p lymphophilum strain 12
p lymphophilum strain 13
p lymphophilum strain 14
p lymphophilum strain 15
p lymphophilum strain 16
p lymphophilum strain 17
p lymphophilum strain 18
p lymphophilum strain 19
p lymphophilum strain 20
p lymphophilum strain 3
p lymphophilum strain 36
p lymphophilum strain 4
p lymphophilum strain 40
p lymphophilum strain 9
peptoniphilus sp nov strain 23
peptoniphilus strain 37
porphyromonas sp nov strain 6C
porphyromonas strain 6C-R
s epidermidis strain 4A
s epidermidis strain 4B
s mitis strain 3A
s mitis strain 3B
s mitis strain 6A
s mitis strain 6B
varibaculum sp nov strain 39

C.11 Isolate Sequencing Statistics

Table C.13: Isolate sequencing statistics in terms of isolate ID, taxonomy, number of sequenced reads and whether additional minION sequencing was undertaken, average coverage, genome size (Mb) and total Number of contigs assembled. nd not determined.

ID	Taxonomy	Num_Reads	Avg_Cov	Size_Mb	Contigs
3	<i>P. lymphophilum</i>	606586	70.39	2.4	5
4	<i>P. lymphophilum</i>	505470	59.25	2.4	13
5	<i>F. hominis</i>	938428	131.95	1.8	25
8B	<i>A. prevotii</i>	611306	85.06	1.9	69
8S	<i>A. prevotii</i>	718672	102.29	1.9	55
9	<i>P. lymphophilum</i>	586150	72.01	2.2	11
10	<i>P. lymphophilum</i>	612724	75.7	2.2	9
11	<i>P. lymphophilum</i>	1731290	115.05	2.2	11

12	<i>P. lymphophilum</i>	533266	66.17	2.2	15
13	<i>P. lymphophilum</i>	155118	20.4	2.2	13
14	<i>P. lymphophilum</i>	530804	65.65	2.2	11
15	<i>P. lymphophilum</i>	1731134	114.47	2.2	9
16	<i>P. lymphophilum</i>	1608910	106.97	2.2	9
17	<i>P. lymphophilum</i>	713506	90.29	2.2	14
18	<i>P. lymphophilum</i>	587656	73.39	2.2	11
19	<i>P. lymphophilum</i>	883298	107.14	2.2	9
20	<i>P. lymphophilum</i>	490554	60.8	2.2	10
23	<i>Peptoniphilus sp. nov.</i>	3225350	241.46	1.9	52
24	<i>Fenollaria sp. nov.</i>	526828 (MinION)	117.49 (MinION)	1.6	1
25	<i>P. harei</i>	2349717 (MinION)	238.39 (MinION)	2	1
27A	<i>Fenollaria sp. nov.</i>	1756304	286.16	1.6	20
27B	<i>Fenollaria sp. nov.</i>	2089770	263.22	1.7	515
28	<i>Peptoniphilus harei</i>	2527068	323.41	2	63
29	<i>Fenollaria sp. nov.</i>	1100996	176.5	1.6	30
33	<i>P. harei</i>	939866	126.06	2	22
35	<i>Fenollaria sp. nov.</i>	1172650	198.01	1.6	19
36	<i>P. lymphophilum</i>	1672504	118.19	2	8
37	<i>P. harei</i>	634482	106.49	1.5	32
38	<i>P. harei</i>	248740	41.76	1.6	61
39	<i>Varibaculum sp. nov.</i>	1783562 (MinION)	164.87 (MinION)	2.2	1
40	<i>P. lymphophilum</i>	1099484	144.48	2	6
41	<i>A. schaalii</i>	1283420	144.87	2.2	59
43	<i>A. schaalii</i>	295046	37.52	2.2	53
44A	<i>E. faecalis</i>	389530	37.77	2.8	91
44B	<i>A. urogenitalis</i>	685980	71.24	2.6	85
48	<i>A. schaalii</i>	234512	30.32	2.2	79
02_09	<i>P. asaccharolytica</i>	625706	69.38	2.3	93
02_15	<i>F. nucleatum (mixed)</i>	463356	nd	nd	nd
22_07	<i>P. asaccharolytica</i>	2049248	127.46	2.3	97
3A	<i>S. mitis</i>	1100752	134.02	1.9	31
3B	<i>S. mitis</i>	860672	98.92	2	71
4A	<i>S. epidermidis</i>	1876892	150.43	2.4	183
4B	<i>S. epidermidis</i>	970352	68.54	2.3	390
5E	<i>C. acnes</i>	3004308	223.14	2	560
6A	<i>S. mitis</i>	619234	64.15	1.8	31
6B	<i>S. mitis</i>	738582	62.18	1.9	44
6C	<i>Porphyromonas sp. nov.</i>	594410	39.61	2.2	142

C.12 Building Kallisto Index

```
#!/bin/bash
```

```
# Download and merge human cDNA and ncDNA files from Ensembl for the index.
```

```
wget ftp://ftp.ensembl.org/pub/current_fasta/\
homo_sapiens/cdna/\
Homo_sapiens.GRCh38.cdna.all.fa.gz
wget ftp://ftp.ensembl.org/pub/current_fasta/\
homo_sapiens/ncrna/\
Homo_sapiens.GRCh38.ncrna.fa.gz
cat Homo_sapiens.GRCh38.cdna.all.fa.gz Homo_sapiens.GRCh38.ncrna.fa.gz >\
Homo_sapiens.GRCh38.rna.fa.gz

#source kallisto dependencies
source ~/programs/kallisto/dependencies

# Build the transcriptome index.
time kallisto index -i hsGRCh38_kallisto Homo_sapiens.GRCh38.rna.fa.gz
```

C.13 SEPATHWAY Pipeline

The pipeline used to analyse the functional potential of isolates (working title SEPATHWAY) is as follows.

```
#!/usr/bin/env python

# Pipeline to extract proteins from assembly with prokka
# and obtain functional annotations with interproscan
import os
import re
import glob
import sys

# Make directories if they don't already exist
os.system('mkdir -pv genomes prokka interproscan/kegg interproscan/metacyc \
interproscan/reactome/ interproscan/go_bp interproscan/go_mf \
interproscan/go_cc interproscan/json_output interproscan/dirty')

# Find list of input gzipped fna files:
genome_files = glob.glob('genomes/*fna.gz')

# Get list of final output files
clean_files = []

for file in genome_files:
    file = os.path.basename(file)
    file = file.split('.fna')[0]
    reactome_file = ("interproscan/reactome/%s_reactome.txt" %(file))
    kegg_file = ("interproscan/kegg/%s_kegg.txt" %(file))
    metacyc_file = ("interproscan/metacyc/%s_metacyc.txt" %(file))
    bp_file = ("interproscan/go_bp/%s_bp.txt" %(file))
```



```
mf_file = ("interproscan/go_mf/%s_mf.txt" %(file))
cc_file= ("interproscan/go_cc/%s_cc.txt" %(file))
clean_files.append(reactome_file)
clean_files.append(kegg_file)
clean_files.append(metacyc_file)
clean_files.append(bp_file)
clean_files.append(mf_file)
clean_files.append(cc_file)
#print(clean_files)

rule all:
    input: [file for file in clean_files]

rule clean_reactome:
    input: "interproscan/dirty/{sample}_reactome.dirty"
    output: "interproscan/reactome/{sample}_reactome.txt"
    priority: 50
    shell: "bin/clean_pathways.py -i {input} -o {output}"

rule clean_kegg:
    input: "interproscan/dirty/{sample}_kegg.dirty"
    output: "interproscan/kegg/{sample}_kegg.txt"
    priority: 50
    shell: "bin/clean_pathways.py -i {input} -o {output}"

rule clean_metacyc:
    input: "interproscan/dirty/{sample}_metacyc.dirty"
    output: "interproscan/metacyc/{sample}_metacyc.txt"
    priority: 50
    shell: "bin/clean_pathways.py -i {input} -o {output}"

rule clean_bp:
    input: "interproscan/dirty/{sample}_bp.dirty"
    output: "interproscan/go_bp/{sample}_bp.txt"
    priority: 50
    shell: "bin/clean_pathways.py -i {input} -o {output}"

rule clean_mf:
    input: "interproscan/dirty/{sample}_mf.dirty"
    output: "interproscan/go_mf/{sample}_mf.txt"
    priority: 50
    shell: "bin/clean_pathways.py -i {input} -o {output}"

rule clean_cc:
    input: "interproscan/dirty/{sample}_cc.dirty"
    output: "interproscan/go_cc/{sample}_cc.txt"
    priority: 50
```

```
shell: "bin/clean_pathways.py -i {input} -o {output}"

rule extract_reactome:
    input: json_in="interproscan/json/{sample}.json"
    output: reactome="interproscan/dirty/{sample}_reactome.dirty"
    priority: 40
    shell: "cat {input.json_in} | grep -B 1 'Reactome' | \
    grep 'name' > {output.reactome}"

rule extract_kegg:
    input: json_in="interproscan/json/{sample}.json"
    output: kegg="interproscan/dirty/{sample}_kegg.dirty"
    priority: 40
    shell: "cat {input.json_in} | grep -B 1 'KEGG' | \
    grep 'name' > {output.kegg}"

rule extract_metacyc:
    input: json_in="interproscan/json/{sample}.json"
    output: metacyc="interproscan/dirty/{sample}_metacyc.dirty"
    priority: 40
    shell: "cat {input.json_in} | grep -B 1 'MetaCyc' | \
    grep 'name' > {output.metacyc}"

rule extract_bp:
    input: json_in="interproscan/json/{sample}.json"
    output: bp="interproscan/dirty/{sample}_bp.dirty"
    priority: 40
    shell: "cat {input.json_in} | grep -B 2 'BIOLOGICAL_PROCESS' | \
    grep 'name' > {output.bp}"

rule extract_mf:
    input: json_in="interproscan/json/{sample}.json"
    output: mf="interproscan/dirty/{sample}_mf.dirty"
    priority: 40
    shell: "cat {input.json_in} | grep -B 2 'MOLECULAR_FUNCTION' | \
    grep 'name' > {output.mf}"

rule extract_cc:
    input: json_in="interproscan/json/{sample}.json"
    output: cc="interproscan/dirty/{sample}_cc.dirty"
    priority: 40
    shell: "cat {input.json_in} | grep -B 2 'CELLULAR_COMPONENT' | \
    grep 'name' > {output.cc}"

rule interproscan:
    input: protein_in="prokka/{sample}/{sample}.faa"
    output: ipscn_out="interproscan/json/{sample}.json"
```

```

priority: 30
shell: "singularity exec bin/sepathway.sif interproscan.sh -f json\
--goterms --iprlookup --pathways -T temp \
-i {input.protein_in} -o {output.ipscn_out}"

rule prokka_annotate:
    input: genome_in="genomes/{sample}.fna"
    output: "prokka/{sample}/{sample}.faa"
    priority: 12
    shell: "SINGULARITYENV_LC_ALL=C singularity exec \
--cleanenv bin/prokka.sif prokka {input.genome_in} \
--outdir prokka/{wildcards.sample} --prefix {wildcards.sample}\
--quiet --force ; gzip {input.genome_in}"

rule gunzip_genomes:
    input: i1="genomes/{sample}.fna.gz"
    output: o1="genomes/{sample}.fna"
    priority: 10
    shell: "gunzip {input.i1}"

```

The resulting pathways from this pipeline were collated with Rscript

```

#!/usr/bin/env Rscript

#load/install packages as required
pkg <- c('dplyr', 'ggplot2', 'purrr', 'readr')
new.pkg <- pkg[!(pkg %in% installed.packages())]
if (length(new.pkg)) {
  install.packages(new.pkg)
}
library(dplyr)
library(ggplot2)
library(purrr)
library(readr)

#define function to collate all results
collate_results <- function(folder) {
  #load and clean filenames
  input_files <- list.files(paste0('../interproscan/', folder),
                           pattern='*.txt',
                           full.names = TRUE)
  filenames <- basename(input_files)
  filenames <- gsub('_bp.txt|_cc.txt|_mf.txt|_kegg.txt|
    _metacyc.txt|_reactome.txt|_genomic',
    '', filenames)

  #load files
  dflist <- lapply(input_files, read_tsv, col_names=c('pathway'))
  names(dflist) <- filenames

```

```

# obtain counts of each pathway in each file
count_list <- dflist
for(df_name in names(dflist)) {
  count_list[[df_name]] <- as.data.frame(table(dflist[[df_name]]))
  colnames(count_list[[df_name]]) <- c('pathway', df_name)
}
collated <- reduce(count_list, full_join, by='pathway') %>%
  replace(., is.na(.), 0)
}

# collate all results
bp <- collate_results('go_bp')
cc <- collate_results('go_cc')
mf <- collate_results('go_mf')
kegg <- collate_results('kegg')
metacyc <- collate_results('metacyc')
reactome <- collate_results('reactome')

# write all results as a table
write.table(bp, file='../interproscan/bp.tsv',
            col.names = TRUE, row.names = FALSE, sep='\t')
write.table(cc, file='../interproscan/cc.tsv',
            col.names = TRUE, row.names = FALSE, sep='\t')
write.table(mf, file='../interproscan/mf.tsv',
            col.names = TRUE, row.names = FALSE, sep='\t')
write.table(kegg, file='../interproscan/kegg.tsv',
            col.names = TRUE, row.names = FALSE, sep='\t')
write.table(metacyc, file='../interproscan/metacyc.tsv',
            col.names = TRUE, row.names = FALSE, sep='\t')
write.table(reactome, file='../interproscan/reactome.tsv',
            col.names = TRUE, row.names = FALSE, sep='\t')

print('Script Complete...')

```


Appendix D

Genomics England and SEPATH

D.1 Taxa Reported by mOTUs2 and not Kraken

Actinobacteria, Alishewanella, Alphaproteobacteria, Aquabacterium, Atopobiaceae, Bacillales, Bacilli, “Bacteria”, Bacteroidaceae, Bacteroidales, Bacteroidetes, Bacteroidia, Bulleidia, Burkholderiales, Caedibacter, Candidatus Gracilibacteria, Candidatus Sacharibacteria, “cellular”, Clostridia, Clostridiaceae, Clostridiales, Colibacter, Comamonadaceae, Coralimargarita, Coriobacteriaceae, Coriobacteriales, Dehalococcoidales, Eggerthellaceae, Eggerthellales, Enterobacterales, Enterobacteriaceae, Erysipelotrichaceae, Faecalibacillus, Firmicutes, Flavobacteriaceae, Flavobacteriia, Fusobacteria, Fusobacteriaceae, Geodermatophilaceae, Haematobacter, Lachnospiraceae, Lactobacillales, Lentisphaerae, Massiliomicrobiota, Methanomicrobia, Microbacteriaceae, Nisseriaceae, Neisseriales, Odoribacteraceae, Pasteurellaceae, Peptostreptococcaceae, Porphyromonadaceae, Prevotellaceae, Pseudaminobacter, Pyramidobacter, Peptostreptococcaceae, Porphyromonadaceae, Prevotellaceae, Pseudoaminobacter, Pyramidobacter, Rhodospirillaceae, Robinsoniella, Rumonicoccaceae, Selenomonadaceae, Selenomonadales, Succinivibrionaceae, Sutterellaceae, Synergistaceae, Tissierallia, “Unknown”, Veillonellaceae, Virmiphilus, Verrucomicrobia.

D.2 Investigating Structure in mOTUs2 data

This section demonstrates a selection of plots produced during an investigation the dimensionality reduction of mOTUs2 data. Data was investigated using bray-curtis (figure D.2) and jaccard presence-absence distance (figure D.1). Abundance values were renormalised following the removal of the ‘-1’ classification denoting unknown taxa.

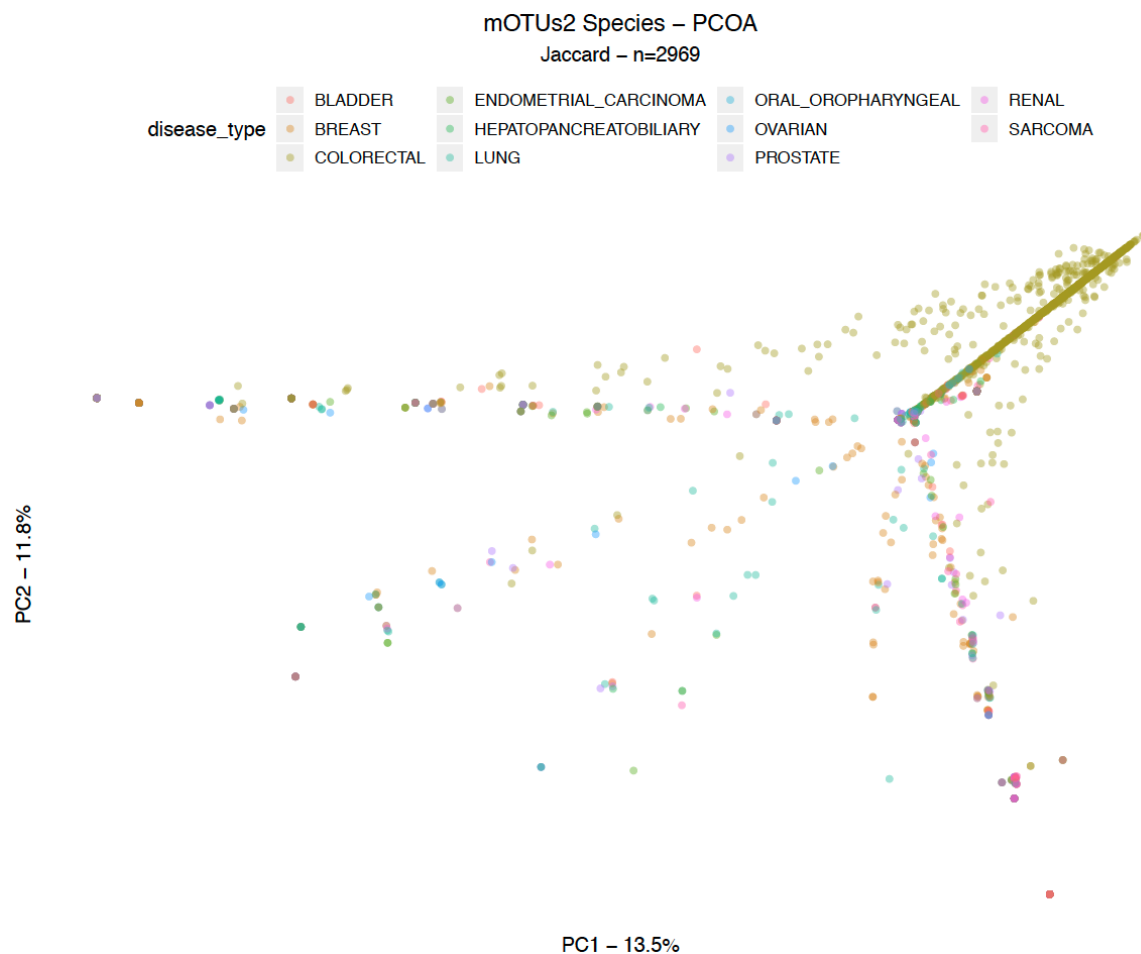


Figure D.1: PCoA of mOTUs2 data using the jaccard distance.

PCoA for mOTUs2 data using renormalised abundance values and the bray-curtis shows three main groups of samples. Kruskal-wallis tests were carried out to identify taxa with differential results between clusters defined manually with a geometric cut-off (bottom $n=456$, right $n=566$ and top $n=1,144$). Plots were constructed for each taxa. As many plots were produced they are not demonstrated here. Instead, the summarised differences between clusters are explained by variations (increased abundance) in the following genera:

Top Cluster:

- *Akkermansia*
- *Alistipes*
- *Bacteroides*
- *Barnesiella*
- *Campylobacter*
- *Enterobacteriaceae*
- *Eubacterium*
- *Faecalibacterium*
- *Fusobacterium*

- *Gemella*
- *Parvimonas*
- *Prevotella*
- *Roseburia*
- *Ruminococcus*
- *Streptococcus*

Bottom Cluster:

- *Propionibacterium*

Right Cluster:

- *Afipia*
- *Caulobacter*
- *Sphingomonas*



Figure D.2: PCoA of mOTUs2 data using the bray-curtis distance.

D.3 Read Complexity

The plot below demonstrate the total number of unmapped reads by cancer type (figure D.3). There is a bimodal distribution and a split by gender in the total non-human reads. There was no obvious bimodal distribution in the ‘root’ classified reads suggesting that the difference is caused by complexity. In searching for a cause for this it was noted that non-human unmapped reads are of higher complexity in females compared to males (figure D.4). Therefore this complexity may be explained by the Y-chromosome.



Figure D.3: Total number of reads in each file following quality trimming and human read depletion prior to classification in sequencing files. The bi-modal distribution in total reads appears to be associated with gender.

A notably higher number of quality-trimmed non-human reads originated from male samples compared to females. Sarcoma was investigated further as it demonstrated the strongest bimodal distribution (figure D.3). It was suspected that these reads may be originating from low complexity of the y-chromosome. BB DuK was used to filter out low complexity reads (low entropy) reads to investigate this. Although the median was brought down following entropy filtering, the majority of samples were entirely removed. This indicates that of the high quality non-human reads, those in males are mostly entirely low complexity whereas in females there are complex reads that are not altered by complexity filtering (only 2 samples lost).

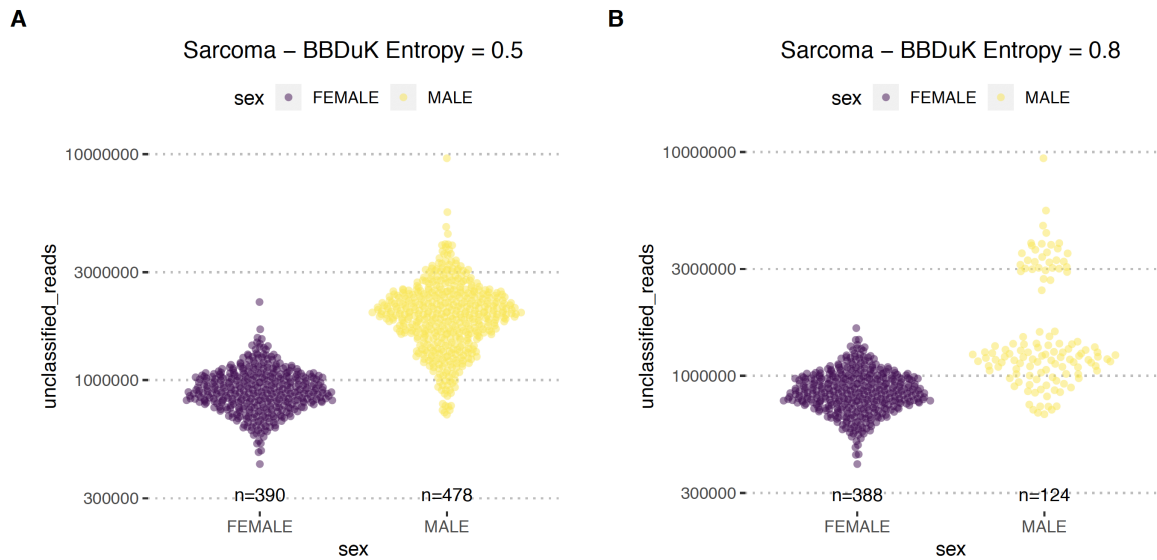


Figure D.4: Low entropy filtering of sarcoma total reads (A) vs high entropy filtering of sarcoma total reads (B). Reads were quality trimmed and human depleted from SEPATH.

D.4 Aligning Reads to HTLV genome

This code was used to download a HTLV-1 reference genome and align HTLV-1 classified reads from a participants tumour and germline sample.

```
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all\
/GCF/000/863/585/GCF_000863585.1_ViralProj15434\
/GCF_000863585.1_ViralProj15434_genomic.fna.gz
mv GCF_000863585.1_ViralProj15434_genomic.fna.gz \
htlv1_ref.fasta.gz

gunzip htlv1_ref.fasta.gz

module add bwa/0.7.17

bwa index htlv1_ref.fasta

bwa mem htlv1_ref.fasta combined_reads.fasta \
> combined_htlv.sam

samtools view -S -b combined_htlv.sam \
> combined_htlv.bam

samtools sort combined_htlv.bam \
> sorted_combined_htlv.bam

samtools index sorted_combined_htlv.bam
```

D.5 Searching for Contaminant Genera

Boruta was employed on data to investigate the taxa that were deemed ‘important’ in predicting metadata variables. The graphical representations of the ranks feature importance are too large for adequate representation in this thesis. Therefore, these files are available in the repository for the thesis but we provide a list of the taxa deemed important here.

Each of the features are provided in reverse order (least important to most important) The metadata variables include both biological and technical variables: age, AT drop, fragment size, chimeric percentage, coverage homogeneity, disease type, GC drop, Genome Build, DNA amount, QC status, library type, preparation method, gender, tissue source, tumour purity and cross contamination percentage,

Three metadata variables had too many important features to extract from the graphs produced within and exported from the Genomics England research environment, Handling NHS trust, Sample concentration and ethnic category.

Boruta Important Features, Age

Lawsonella, Sanguibacter, Collinsella, Trichococcus, Prosthecobacter, Fusicatenibacter, Brochothrix, Candidatus Cloacimonas, Abiotrophia, Bilophila, Anaeromassilibacillus, Lactococcus, Flavonifractor, Ochrobactrum, Fournierella, Ruminiclostridium, Bosea, Angelakisella, Gemmiger, Enterobacter, Romboutsia, Ornithinimicrobium, Nocardia, Bacillus, Elizabethkingia, Xenophilus, Janthinobacterium, Laribacter, Thermus, Anaerostolibacter, Auricoccus, Salmonella, Thermoanaerobacter, Butyrivibrio, Fusobacterium, Dorea, Toxoplasma, Xanthomonas, Odoribacter, Scardovia, Deinococcus, Holdemania, Arthrobacter, Gordonibacter, Mesorhizobium, Lymphocryptovirus, Eubacterium, Oscillibacter, Paraprevotella, Thermoanaerobacterium, Anaerotignum, Senegalimassilia, Helcococcus, Faecalibacterium, Citrobacter, Sphingopyxis, Coprococcus, Sphingobacter, Porphyromonas, Thermomonas, Methanobrevibacter, Methylophilus, Mycobacterium, Aeribacillus, Chryseobacterium, Solimonas, Psychrobacter, Subdoligranulum, Neisseria, Propionibacterium, Gemella, Serratia, Campylobacter, Exiguobacterium, Dolosigranulum, Parabacteroides, Erysipelatoclostridium, Caulobacter, Bifidobacterium, Coprobacillus, Porphyrobacter, Alicyclophilus, Caldimonas, Klebsiella, Cutibacterium, Finegoldia, Clostridium, Ruminococcus, Peptoniphilus, Faecalitalea, Bacteroides, Eggerthella, Blautia, Novosphingobium, Eisenbergiella, Candidatus Planktophilia, Malassezia, Afipia, Azospira, Anaerococcus, Brachybacterium, Lysobacter, Limnochabitans, Sphingobacterium, Burkholderia, Arcobacter, Haemophilus, Nocardioideis, Alistipes, Hammondia, Prevotella, Pelomonas, Brevibacterium, Tyzzerella, Veillonella, Janibacter, Lachnoclostridium, Roseburia, Paracoccus, Hungatella, Clostridioides, Bradyrhizobium, Lactobacillus, Escherichia, Gordonia, Erwinia, Actinomyces, Ralstonia, Intestinimonas, Fusarium, Sphingobium, Methylobacterium, Aeromonas, Rothia, Acidovorax, Micrococcus, Microbacterium, Pseudomonas, Variovorax, Comamonas, Delftia, Rhodococcus, Leuconostoc, Sphingomonas, Saccharomyces, Kocuria, Streptococcus, Staphylococcus, Corynebacterium, Lautropia, Stenotrophomonas, Acinetobacter, Flavobacterium, Moraxella, Curvibacter, Brevundimonas, Cupriavidus, Cloacibacterium

Boruta Important Features, AT Drop

Selenomonas, Roseateles, Carboxydocella, Parvimonas, Anaerotruncus, Shigella, Akkermansia, Tannerella, Mitsuokella, Intestinimonas, Anaerostipes, Ruminiclostridium, Streptobacillus, Collinsella, Erythrobacter, Methylobacter, Aspergillus, Eggerthella, Angelakisella, Thiomonas, Faecalitalea, Methyloversatilis, Agrococcus, Campylobacter, Oribacterium, Coprobacillus, Bacillus, Butyrivibrio, Erysipelatoclostridium, Fusicatenibacter, Fournierella, Janibacter, Lachnoanaerobaculum, Plasmodium, Holdemania, Clostridioides, Massilia, Oscillibacter, Slackia, Kytococcus, Sordaria, Granulicatella, Gemmiger, Bilophila, Paraburkholderia, Methylothena, Johnsonella, Thermomonas, Flavonifractor, Paraprevotella, Roseburia, Gluconacetobacter, Coprococcus, Agrobacterium, Barnesiella, Brevundimonas, Methylophilus, Ruminococcus, Herbiconiux, Butyricimonas, Candidatus Planktophilia, Stenotrophomonas, Tyzzerella, Gemella, Dialister, Devosia, Limnochabitans, Dorea, Flavobacterium, Hungatella, Eubacterium, Bifidobacterium, Comamonas, Lactobacillus, Parabacteroides, Xanthomonas, Novosphingobium, Candidatus Pelagibacter, Alistipes, Sutterella, Blautia, Faecalibacterium, Odoribacter, Solobacterium, Cloacibacterium, Clostridium, Mesorhizobium, Herbaspirillum, Lachnoclostridium, Veillonella, Fusobacterium, Microbacterium, Collimonas, Enhydrobacter, Glutamicibacter, Sulfitobacter, Listeria, Haemophilus, Asticcacaulis, Halomonas, Moraxella, Prevotella, Propionibacterium, Deinococcus, Cryptosporidium, Paracoccus, Neisseria, Bacteroides, Sphingobium, Rhodococcus, Acinetobacter, Acidovorax, Corynebacterium, Rothia, Achromobacter, Rhodopseudomonas, Xenophilus, Marinomonas, Cutibacterium, Pseudomonas, Kocuria, Brevibacterium, Rudaea, Asinibacterium, Pelomonas, Micrococcus, Janthinobacterium, Confluentimicrobium, Delftia, Actinomyces, Nesterenkonia, Streptococcus, Caulobacter, Afipia, Burkholderia, Variovorax, Methylobacterium, Ralstonia, Sphingomonas, Curvibacter, Hammondia, Bradyrhizobium, Cupriavidus, Mycobacterium, Escherichia, Toxoplasma, Enterococcus

Boruta Important Features, Fragment Size

Roseateles, Tidjanibacter, Pleomorphomonas, Flavobacterium, Parvimonas, Gulbenkiania, Thermomonas, Holdemania, Erysipelatoclostridium, Bifidobacterium, Photobacterium, Butyrivibrio, Corynebacterium, Anaeromassilibacillus, Brachybacterium, Coprobacillus, Zymoseptoria, Pelomonas, Sutterella, Fontimonas, Oscillibacter, Fusicatenibacter, Slackia, Meiothermus, Micrococcus, Syntrophus, Flavonifractor, Gabonia, Faecalitalea, Asticcacaulis, Dialister, Veillonella, Paraprevotella, Lactobacillus, Anaerotruncus, Escherichia, Anaerosalibacter, Pseudoflavonifractor, Haemophilus, Intestinimonas, Kocuria, Ralstonia, Anaerostipes, Holdemanella, Cutibacterium, Lachnoanaerobaculum, Porphyromonas, Caulobacter, Angelakisella, Fournierella, Eggerthella, Peptostreptococcus, Methyloversatilis, Staphylococcus, Subdoligranulum, Rhodopseudomonas, Azospirillum, Laribacter, Cyclospora, Methylobacterium, Selenomonas, Coprococcus, Tyzzerella, Ruminiclostridium, Dorea, Roseburia, Fusobacterium, Comamonas, Prevotella, Bilophila, Pseudoalteromonas, Akkermansia, Gemmiger, Geobacter, Nevskia, Hungatella, Pandoraea, Planktothrix, Clostridioides, Eubacterium, Sphingomonas, Nitrospira, Acidovorax, Solobacterium, Methylibium, Odoribacter, Campylobacter, Methylobacter, Methylophilus, Janthinobacterium, Brevibacterium, Ruminococcus, Marinomonas, Nocardia, Lachnoclostridium, Blautia, Collinsella, Burkholderia, Microbacterium, Faecalibacterium, Clostridium, Actinomyces, Parabacteroides, Pseudomonas, Tepidiphilus, Barnesiella,

Alistipes, Cloacibacterium, Sulfitobacter, Afipia, Bacteroides, Xanthomonas, Curvibacter, Cupriavidus, Delftia, Novosphingobium, Sporosarcina, Hammondia, Bradyrhizobium, Enterococcus, Xenophilus, Streptococcus, Azospira, Deinococcus, Acinetobacter, Achromobacter, Sphingobium, Rudaea, Variovorax, Herbaspirillum

Boruta Important Features, Chimeric Percentage

Nitrospira, Prevotellamassilia, Hungatella, Butyricimonas, Clostridium, Parabacteroides, Novosphingobium, Gabonibacter, Collimonas, Cryptosporidium, Roseburia, Blautia, Lachnoclostridium, Bacteroides, Confluentimicrobium, Anaerotignum, Sutterella, Fournierella, Ruminiclostridium, Lachnoanaerobaculum, Trueperella, Candida, Pseudoflavonifractor, Bilophila, Anaeromassilibacillus, Sandarakinorhabdus, Shinella, Barnesiella, Nocardia, Urinacoccus, Hammondia, Roseateles, Eggerthella, Olsenella, Parascardovia, Cytomegalovirus, Megamonas, Acetanaerobacterium, Oscillibacter, Thioalkalivibrio, Variovorax, Alloscardovia, Enterococcus, Johnsonella, Odoribacter, Sanguibacter, Methanobrevibacter, Pararhodospirillum, Methyloversatilis, Romboutsia, Laribacter, Erythrobacter, Mobiluncus, Duodenibacillus, Methylophilus, Anaerotruncus, Sulfurospirillum, Enterorhabdus, Helicobacter, Enterocytozoon, Faecalitalea, Thermomonas, Holdemanella, Methylothena, Methylobacter, Rudaea, Xenophilus

Boruta Important Features, Coverage Homogeneity

Slackia, Glutamicibacter, Methanosarcina, Eggerthella, Rothia, Xanthomonas, Neglecta, Caldimonas, Erysipelatoclostridium, Gulbenkiania, Anaeromassilibacillus, Faecalibacterium, Dyella, Hungatella, Clostridioides, Eisenbergiella, Syntrophus, Anaerotignum, Dorea, Christensenella, Ruminiclostridium, Alistipes, Holdemania, Marvinbryantia, Oxalobacter, Angelakisella, Odoribacter, Sphingobium, Anaerotruncus, Gemmiger, Fontimonas, Cloacibacterium, Gabonia, Tepidiphilus, Lachnoclostridium, Enhydrobacter, Carboxydocella, Oscillibacter, Coprobacillus, Brevibacterium, Subdoligranulum, Merdibacter, Novosphingobium, Holdemanella, Clostridium, Thermomonas, Mycobacterium, Coprococcus, Turicella, Xenophilus, Acetitomaculum, Streptococcus, Fournierella, Butyricimonas, Butyrivibrio, Tyzzerella, Pseudomonas, Senegalimassilia, Intestinimonas, Bilophila, Herbaspirillum, Ruminococcus, Geobacter, Paracoccus, Roseburia, Paraprevotella, Felixolvirus, Cryobacterium, Eubacterium, Achromobacter, Anaerostipes, Parabacteroides, Kocuria, Afipia, Agrococcus, Faecalitalea, Flavonifractor, Peptostreptococcus, Gluconacetobacter, Bacteroides, Actinomyces, Halorubrum, Candidatus Kryptobacter, Blautia, Curvibacter, Methylobacter, Rhodopseudomonas, Cutibacterium, Micrococcus, Pelomonas, Methylobacterium, Burkholderia, Cryptosporidium, Hammondia, Nocardia, Confluentimicrobium, Sulfitobacter, Sphingomonas, Ralstonia, Janthinobacterium, Rudaea, Bradyrhizobium, Asinibacterium, Marinomonas, Caulobacter, Deinococcus, Nesterenkonia, Cupriavidus, Enterococcus

Boruta Important Features, Disease Type

Butyricococcus, Slackia, Anaeromassilibacillus, Pseudoflavonifractor, Fournierella, Methanobrevibacter, Ezakiella, Methanobacterium, Streptobacillus, Agrobacterium, Enterobacter, Anaerotignum, Nitrospira, Lysobacter, Butyrivibrio, Mogibacterium, Tepidimonas, Romboutsia, Eisenbergiella, Trametes, Xenophilus, Butyricimonas, Brachybacterium, Faecalitalea, Geobacillus, Gammaretrovirus, Olsenella, Diaphorobacter, Arcobacter,

Finegoldia, Candida, Fusarium, Anoxybacillus, Eggerthia, Paraburkholderia, Paraprevotella, Penicillium, Alcaligenes, Johnsonella, Trichosporon, Pseudopropionibacterium, Erysipelatoclostridium, Toxoplasma, Confluentimicrobium, Tyzzerella, Propionimicrobium, Peptostreptococcus, Anaerotruncus, Brevibacterium, Cellulomonas, Rudaea, Aggregatibacter, Elizabethkingia, Kingella, Alloprevotella, Klebsiella, Thermus, Ruminiclostridium, Peptoanaerobacter, Janthinobacterium, Saccharomyces, Holdemania, Catonella, Azospira, Janibacter, Leifsonia, Collinsella, Akkermansia, Propionibacterium, Eikenella, Limnhabitans, Anaerostipes, Mesorhizobium, Lachnoanaerobaculum, Clostridioides, Comamonas, Atopobium, Gemmiger, Burkholderia, Dialister, Treponema, Sutterella, Flavonifractor, Oribacterium, Thermomonas, Gordonia, Solobacterium, Megasphaera, Paracoccus, Meiothermus, Ruminococcus, Eggerthella, Tannerella, Lymphocryptovirus, Plasmodium, Roseburia, Lactobacillus, Oscillibacter, Coprobacillus, Intestinimonas, Granulicatella, Microbacterium, Stomatobaculum, Deinococcus, Citrobacter, Angelakisella, Bilophila, Hungatella, Sphingobacterium, Coprococcus, Xanthomonas, Anaerococcus, Selenomonas, Porphyromonas, Bifidobacterium, Cutibacterium, Cronobacter, Dermacoccus, Candidatus Microthrix, Chlamydia, Subdoligranulum, Leptotrichia, Rhodopseudomonas, Eubacterium, Neisseria, Parvimonas, Barnesiella, Leclercia, Faecalibacterium, Odoribacter, Aeromonas, Dorea, Sphingopyxis, Rothia, Haemophilus, Alistipes, Caulobacter, Capnocytophaga, Actinomyces, Blautia, Gemella, Chryseobacterium, Peptoniphilus, Micrococcus, Campylobacter, Afipia, Parabacteroides, Lachnoclostridium, Staphylococcus, Pseudomonas, Fusobacterium, Clostridium, Curvibacter, Sphingomonas, Novosphingobium, Cupriavidus, Brevundimonas, Acidovorax, Roseolovirus, Delftia, Pandoraea, Prevotella, Kocuria, Variovorax, Veillonella, Flavobacterium, Bacteroides, Corynebacterium, Mycobacterium, Acinetobacter, Sphingobium, Streptococcus, Enterococcus, Pelomonas, Stenotrophomonas, Moraxella, Methylobacterium, Rhodococcus, Hammondia, Ralstonia, Achromobacter, Bradyrhizobium, Escherichia, Cloacibacterium

Boruta Important Features, GC drop

Microbacterium, Coprobacillus, Odoribacter, Roseateles, Bifidobacterium, Parabacteroides, Paraprevotella, Flavobacterium, Stenotrophomonas, Coprococcus, Dorea, Comamonas, Flavonifractor, Holdemania, Asticcacaulis, Lachnoclostridium, Butyrivibrio, Brevundimonas, Halorubrum, Intestinimonas, Glutamicibacter, Dialister, Methylothermus, Sutterella, Novosphingobium, Butyricimonas, Rhodopseudomonas, Herbaspirillum, Agrococcus, Intestinibacter, Cloacibacterium, Propionibacterium, Nitrosomonas, Xylanimonas, Methylophilus, Thermomonas, Herbiconiux, Blautia, Eubacterium, Tyzzerella, Solobacterium, Mesorhizobium, Enhydrobacter, Fusobacterium, Mitsuraria, Neisseria, Johnsonella, Veillonella, Leptonema, Gluconacetobacter, Sphingobium, Rhodococcus, Gemella, Clostridium, Haemophilus, Corynebacterium, Acinetobacter, Acidovorax, Methylobacter, Prevotella, Bacteroides, Paracoccus, Cutibacterium, Deinococcus, Kocuria, Janthinobacterium, Rothia, Streptococcus, Xenophilus, Pelomonas, Cryptosporidium, Pseudomonas, Toxoplasma, Brevibacterium, Afipia, Rudaea, Marinomonas, Curvibacter, Confluentimicrobium, Delftia, Micrococcus, Variovorax, Sulfitobacter, Hammondia, Caulobacter, Ralstonia, Bradyrhizobium, Methylobacterium, Actinomyces, Sphingomonas, Mycobacterium, Asinibacterium, Achromobacter, Burkholderia, Escherichia, Nesterenkonia, Enterococcus, Cupriavidus

Boruta Important Features, Genome Build

Delftia, Brevundimonas, Capnocytophaga, Eggerthella, Fournierella, Angelakisella, Anaerostipes, Variovorax, Pseudopropionibacterium, Faecalitalea, Anaerotruncus, Serratia, Lachnoanaerobaculum, Anaerobium, Aquitalea, Aeromonas, Aureimonas, Flavonifractor, Intestinimonas, Alistipes, Oscillibacter, Pelomonas, Lachnoclostridium, Alkalibacterium, Coprobacillus, Xanthomonas, Staphylococcus, Veillonella, Pandoraea, Burkholderia, Weissella, Rothia, Lactobacillus, Cutibacterium, Curvibacter, Chroococcidiopsis, Stenotrophomonas, Sphingobium, Enterococcus, Afipia, Moraxella, Haemophilus, Caulobacter, Paracoccus, Microbacterium, Bifidobacterium, Kocuria, Halomonas, Confluentimicrobium, Cryptosporidium, Actinomyces, Corynebacterium, Prevotella, Sphingomonas, Streptococcus, Rhodococcus, Acinetobacter, Bradyrhizobium, Pseudomonas, Micrococcus, Acidovorax, Methylobacterium, Hammondia, Ralstonia, Mucilaginibacter

Boruta Important Features, DNA amount

Coprobacillus, Sutterella, Mycobacterium, Intestinimonas, Subdoligranulum, Thiomonas, Anaerostipes, Arcobacter, Eisenbergiella, Gemmiger, Tannerella, Butyrivibrio, Plasmodium, Cyberlindnera, Odoribacter, Parabacteroides, Ruminococcus, Capnocytophaga, Nocardioideae, Abiotrophia, Roseburia, Saccharomyces, Peptoniphilus, Eikenella, Meiothermus, Erysipelatoclostridium, Kingella, Dorea, Coprococcus, Desulfotomaculum, Bilophila, Confluentimicrobium, Aerococcus, Rheinheimera, Vaginella, Ruminiclostridium, Dialister, Dolosigranulum, Rhizobium, Hungatella, Elizabethkingia, Anaerococcus, Akkermansia, Fontimonas, Legionella, Enhydrobacter, Blautia, Purpureocillium, Planifilum, Streptomyces, Filifactor, Alicyclophilus, Faecalibacterium, Selenomonas, Lactococcus, Nitrospira, Clostridioides, Wickerhamomyces, Limnohabitans, Aeromonas, Parvimonas, Fusobacterium, Eremococcus, Phascolarctobacterium, Methanobrevibacter, Tyzzerella, Bifidobacterium, Peptoanaerobacter, Anoxybacillus, Barnesiella, Salinicoccus, Gemella, Lawsonella, Pantoea, Chlamydia, Sphingobacterium, Acidithiobacillus, Campylobacter, Sphingopyxis, Lachnoclostridium, Alistipes, Cellulosimicrobium, Yersinia, Enterobacter, Clostridium, Shewanella, Frigoribacterium, Streptobacillus, Chryseobacterium, Janthinobacterium, Serratia, Treponema, Kurthia, Psychrobacter, Porphyromonas, Bacteroides, Massilia, Herbaspirillum, Methylophilus, Nocardia, Lactobacillus, Dermacoccus, Veillonella, Bosea, Haemophilus, Methyloversatilis, Tepidimonas, Novosphingobium, Paracoccus, Roseomonas, Pseudoxanthomonas, Jeotgaliococcus, Gordonia, Neisseria, Paraburkholderia, Candidatus Microthrix, Alkalibacterium, Caulobacter, Arthrobacter, Xanthomonas, Brachybacterium, Rothia, Sulfotobacter, Brevibacterium, Prevotella, Deinococcus, Pseudomonas, Spirosoma, Flavobacterium, Propionibacterium, Actinomyces, Thermomonas, Agrobacterium, Comamonas, Enterococcus, Janibacter, Mesorhizobium, Sphingobium, Brevundimonas, Bacillus, Sphingomonas, Streptococcus, Rudaea, Rhodopseudomonas, Ralstonia, Giesbergeria, Acidovorax, Kocuria, Pelomonas, Cloacibacterium, Rhodococcus, Microbacterium, Burkholderia, Cupriavidus, Stenotrophomonas, Xenophilus, Methylobacterium, Bradyrhizobium, Moraxella, Cutibacterium, Staphylococcus, Curvibacter, Micrococcus, Hammondia, Corynebacterium, Acinetobacter, Afipia, Delftia, Variovorax, Achromobacter

Boruta Important Features, QC Status

Variovorax, Candidatus Planktophilia, Massilia, Cloacibacterium, Elizabethkingia, Curvibacter, Mesorhizobium, Methylobacterium, Pseudomonas, Corynebacterium, Sphingomonas

Feature, Library Type

Caldimonas, Pseudonocardia, Actinobaculum, Nocardiosis, Bosea, Chelatococcus, Dokdonella, Rheinheimera, Pedobacter, Diaphorobacter, Nesterenkonia, Ornithinimicrobium, Novibacillus, Turicella, Rathayibacter, Frigoribacterium, Paenibacillus, Plasmodium, Friedmanniella, Thauera, Limnhabitans, Caldicellulosiruptor, Planococcus, Jeotgaliococcus, Lysinibacillus, Gammapapillomavirus, Nitrobacter, Duganella, Solirubrobacter, Veillonella, Sporosarcina, Streptomyces, Hydrogenophaga, Thermocrispum, Meyerozyma, Chlamydia, Dolosigranulum, Thermobispora, Heterobasidion, Nocardia, Paraburkholderia, Ochrobactrum, Methylophilus, Erythrobacter, Aspergillus, Flaviflexus, Xenophilus, Lactobacillus, Malassezia, Cloacibacterium, Laribacter, Cyberlindnera, Sphaerotilus, Bacillus, Psychrobacter, Marinospirillum, Flavobacterium, Prevotella, Giesbergeria, Massilia, Leucobacter, Roseomonas, Confluentimicrobium, Lactococcus, Arthrobacter, Agrococcus, Haemophilus, Brochothrix, Deinococcus, Neisseria, Auricoccus, Pleomorphomonas, Geobacillus, Nitrosomonas, Azospirillum, Rickettsia, Nocardioidea, Nitrospira, Collimonas, Silanimonas, Sphingopyxis, Curtobacterium, Dietzia, Novosphingobium, Meiothermus, Enterococcus, Mesorhizobium, Enhydrobacter, Glutamicibacter, Plantibacter, Achromobacter, Purpureocillium, Brevundimonas, Streptococcus, Halomonas, Gordonia, Alternaria, Acidocella, Thermus, Pelomonas, Kytococcus, Moraxella, Rudaea, Symbiobacterium, Variovorax, Rothia, Sep1virus, Dermacoccus, Agrobacterium, Rhodopseudomonas, Rhodococcus, Paracoccus, Brachybacterium, Tepidiphilus, Pantoea, Sphingobium, Hammondia, Actinomyces, Pseudomonas, Curvibacter, Parageobacillus, Propionibacterium, Acidovorax, Xanthomonas, Aliicoccus, Stenotrophomonas, Staphylococcus, Aquamicrobium, Sphingomonas, Caulobacter, Afipia, Micrococcus, Janibacter, Burkholderia, Brevibacterium, Comamonas, Kocuria, Asticcacaulis, Corynebacterium, Herbaspirillum, Cutibacterium, Acinetobacter, Ralstonia, Methylobacterium, Bradyrhizobium, Salinicoccus, Cupriavidus, Janthinobacterium, Microbacterium, Delftia

Boruta Important Features, Preparation Method

Carboxydocella, Bifidobacterium, Leishmania, Nitrospira, Azoarcus, Macrococcus, Comamonas, Tsukamurella, Phycococcus, Candidatus Planktophilia, Thermoanaerobacterium, Debaryomyces, Gluconacetobacter, Plantibacter, Plasmodium, Arcobacter, Psychrobacter, Pyrenophora, Rhizobium, Gulbenkiania, Fusobacterium, Pseudomonas, Pseudohongiella, Microcella, Halorubrum, Paraburkholderia, Mycobacterium, Geobacillus, Devosia, Escherichia, Rubrobacter, Methanobacterium, Fontimonas, Pantoea, Pseudogymnoascus, Veillonella, Frigoribacterium, Lactobacillus, Babesia, Reyranella, Lautropia, Dietzia, Enterobacter, Dolosigranulum, Ferribacterium, Jeotgaliococcus, Roseomonas, Pelomonas, Micromonospora, Curtobacterium, Agrococcus, Bergeyella, Sporosarcina, Candidatus Microthrix, Cryptosporidium, Curvibacter, Legionella, Bacteroides, Agaricus, Thermus, Planomicrobium, Sinorhizobium, Mesorhizobium, Ornithinimicrobium, Haemophilus, Ochrobactrum, Cloacibacterium, Prevotella, Leucobacter, Bordetella, Rhodopseudomonas, Brevundimonas, Confluentimicrobium, Xenophilus, Pannonibacter, Geobacter, Moraxella,

Agrobacterium, Nitrosospira, Dermacoccus, Abiotrophia, Sphingopyxis, Rhodococcus, Sulfitobacter, Staphylococcus, Nitrosomonas, Flavobacterium, Meiothermus, Enterococcus, Pelagibacterium, Neisseria, Delftia, Synechocystis, Gulosibacter, Achromobacter, Streptococcus, Stenotrophomonas, Propionibacterium, Novosphingobium, Meyerozyma, Asticcacaulis, Afipia, Caulobacter, Rothia, Acidovorax, Herbaspirillum, Acinetobacter, Janibacter, Acetanaerobacterium, Gordonia, Microbacterium, Bradyrhizobium, Nocardia, Tepidiphilus, Asinibacterium, Turicella, Hammondia, Actinomyces, Brachybacterium, Corynebacterium, Massilia, Paracoccus, Marmoricola, Xanthomonas, Brevibacterium, Variovorax, Anaplasma, Methylobacterium, Kocuria, Micrococcus, Glutamicibacter, Methylophilus, Rudaea, Marinomonas, Sphingobium, Cutibacterium, Nesterenkonia, Burkholderia, Janthinobacterium, Nocardioideis, Sphingomonas, Ralstonia, Eimeria, Deinococcus, Cupriavidus

Boruta Important Features, Gender

Anaeromassilibacillus, Pseudoflavonifractor, Fournierella, Methanobrevibacter, Ezakiella, Methanobacterium, Streptobacillus, Agrobacterium, Enterobacter, Anaerotignum, Nitrosospira, Lysobacter, Butyrivibrio, Mogibacterium, Tepidimonas, Romboutsia, Eisenbergiella, Trametes, Xenophilus, Butyricimonas, Brachybacterium, Faecalitalea, Geobacillus, Gammaretrovirus, Olsenella, Diaphorobacter, Arcobacter, Finegoldia, Candida, Fusarium, Anoxybacillus, Eggerthia, Paraburkholderia, Paraprevotella, Penicillium, Alcaligenes, Johnsonella, Trichosporon, Pseudopropionibacterium, Erysipelatoclostridium, Toxoplasma, Confluentimicrobium, Tyzzerella, Propionimicrobium, Peptostreptococcus, Anaerotruncus, Brevibacterium, Cellulomonas, Rudaea, Aggregatibacter, Elizabethkingia, Kingella, Alloprevotella, Klebsiella, Thermus, Ruminiclostridium, Peptoanaerobacter, Janthinobacterium, Saccharomyces, Holdemania, Catonella, Azospira, Janibacter, Leifsonia, Collinsella, Akkermansia, Propionibacterium, Eikenella, Limnohabitans, Anaerostipes, Mesorhizobium, Lachnoanaerobaculum, Clostridioides, Comamonas, Atopobium, Gemmiger, Burkholderia, Dialister, Treponema, Sutterella, Flavonifractor, Oribacterium, Thermomonas, Gordonia, Solobacterium, Megasphaera, Paracoccus, Meiothermus, Ruminococcus, Eggerthella, Tannerella, Lymphocryptovirus, Plasmodium, Roseburia, Lactobacillus, Oscillibacter, Coprobacillus, Intestinimonas, Granulicatella, Microbacterium, Stomatobaculum, Deinococcus, Citrobacter, Angelakisella, Bilophila, Hungatella, Sphingobacterium, Coprococcus, Xanthomonas, Anaerococcus, Selenomonas, Porphyromonas, Bifidobacterium, Cutibacterium, Cronobacter, Dermacoccus, Candidatus Microthrix, Chlamydia, Subdoligranulum, Leptotrichia, Rhodopseudomonas, Eubacterium, Neisseria, Parvimonas, Barnesiella, Leclercia, Faecalibacterium, Odoribacter, Aeromonas, Dorea, Sphingopyxis, Rothia, Haemophilus, Alistipes, Caulobacter, Capnocytophaga, Actinomyces, Blautia, Gemella, Chryseobacterium, Peptoniphilus, Micrococcus, Campylobacter, Afipia, Parabacteroides, Lachnoclostridium, Staphylococcus, Pseudomonas, Fusobacterium, Clostridium, Curvibacter, Sphingomonas, Novosphingobium, Cupriavidus, Brevundimonas, Acidovorax, Roseolovirus, Delftia, Pandoraea, Prevotella, Kocuria, Variovorax, Veillonella, Flavobacterium, Bacteroides, Corynebacterium, Mycobacterium, Acinetobacter, Sphingobium, Streptococcus, Enterococcus, Pelomonas, Stenotrophomonas, Moraxella, Methylobacterium, Rhodococcus, Hammondia, Ralstonia, Achromobacter, Bradyrhizobium, Escherichia, Cloacibacterium

Boruta Important Features, Tissue Source

Chryseobacterium, Desulfitobacterium, Porphyrobacter, Alphatorquevirus, Thioalkalivibrio, Proteus, Leishmania, Pseudogymnoascus, Aeromonas, Cryptosporidium, Anaerosalibacter, Duganella, Oligella, Gemmata, Azoarcus, Peptococcus, Cyclospora, Bordetella, Actinotignum, Rudaea, Pseudohongiella, Meyerozyma, Barnesiella, Propionimicrobium, Devosia, Enterobacter, Tepidiphilus, Ornithinimicrobium, Veillonella, Babesia, Nitrospira, Methylibium, Photobacterium, Actinoplanes, Haemophilus, Brochothrix, Bacillus, Leuconostoc, Fervidobacterium, Dermacoccus, Dorea, Methylophilus, Agaricus, Clostridium, Massilia, Finegoldia, Fusarium, Pyrenophora, Hydrogenophaga, Serratia, Enhydrobacter, Xanthomonas, Lachnoclostridium, Levyella, Meiothermus, Belliella, Lactococcus, Enterocytozoon, Alistipes, Bosea, Microbacterium, Achromobacter, Propionibacterium, Confluentimicrobium, Gulosibacter, Aurantimicrobium, Xenophilus, Sporosarcina, Thermus, Nitrosomonas, Coprococcus, Brevundimonas, Malassezia, Peptoniphilus, Betapolyomavirus, Cloacibacterium, Lactobacillus, Ruminococcus, Pannonibacter, Marmoricola, Blautia, Mycobacterium, Brevibacterium, Sphaerotilus, Faecalibacterium, Rothia, Actinomyces, Plasmodium, Cutibacterium, Caulobacter, Stenotrophomonas, Geodermatophilus, Pseudomonas, Ochrobactrum, Prevotella, Janthinobacterium, Salmonella, Pelagibacterium, Sphingopyxis, Janibacter, Marinospirillum, Agrobacterium, Burkholderia, Chlamydia, Thiomonas, Novosphingobium, Acidovorax, Alternaria, Escherichia, Gordonia, Paracoccus, Enterococcus, Moraxella, Hammondia, Acinetobacter, Anaplasma, Acetanaerobacterium, Micrococcus, Delftia, Streptococcus, Methylobacterium, Corynebacterium, Staphylococcus, Kocuria, Bacteroides, Nocardioideis, Afipia, Pelomonas, Rhodopseudomonas, Eimeria, Variovorax, Sphingobium, Sphingomonas, Ralstonia, Cupriavidus, Bradyrhizobium, Curvibacter

Boruta Important Features, Tumour Purity,

Flavonifractor, Janibacter, Burkholderia, Leuconostoc, Deinococcus, Granulicatella, Pectobacterium, Solobacterium, Treponema, Anaerostipes, Tyzzerella, Microbacterium, Capnocytophaga, Tannerella, Holdemanella, Brachybacterium, Stenotrophomonas, Oscillibacter, Rhodococcus, Caldimonas, Coprococcus, Staphylococcus, Sphingomonas, Xenophilus, Coprobacillus, Anaerococcus, Cytomegalovirus, Propionibacterium, Peptoniphilus, Hammondia, Leptotrichia, Xanthomonas, Paracoccus, Neisseria, Odoribacter, Cupriavidus, Dorea, Ruminiclostridium, Olsenella, Moraxella, Eubacterium, Ruminococcus, Faecalibacterium, Alistipes, Gemella, Blautia, Micrococcus, Haemophilus, Parabacteroides, Afipia, Gemmiger, Curvibacter, Streptococcus, Barnesiella, Bacteroides, Corynebacterium, Oribacterium, Rothia, Achromobacter, Fusobacterium, Novosphingobium, Pelomonas, Actinomyces, Lachnoanaerobaculum, Roseburia, Bradyrhizobium, Delftia, Hungatella, Kocuria, Ralstonia, Sphingopyxis, Erythroparvovirus, Methylobacterium, Lachnoclostridium, Clostridium, Alphapolyomavirus, Lactobacillus, Cloacibacterium, Prevotella, Sphingobium, Dermacoccus, Acinetobacter, Veillonella, Variovorax, Bifidobacterium, Porphyromonas, Lymphocryptovirus, Roseolovirus

Boruta Important Features, Cross Contamination Percentage

Variovorax, Nosocomiicoccus, Candidatus Planktophila, Roseburia, Curtobacterium, Propionibacterium, Cloacibacterium, Sphingobium, Fusobacterium, Microbacterium, Paracoccus, Arthrosphaera, Sphingopyxis, Bacteroides, Haemophilus, Williamsia, Enterococcus,

Moraxella, Actinomyces, Prevotella, Kocuria, Trichomonas, Psychrobacter, Hymenobacter, Burkholderia, Terrabacter, Rothia, Cutibacterium, Curvibacter, Streptococcus, Sphingomonas, Micrococcus, Methylobacterium, Bradyrhizobium, Ralstonia, Staphylococcus, Thermoanaerobaculum, Chlamydia, Plasmodium, Gammaretrovirus

D.6 Comparing TCGA Kraken genera with SEPATH Kraken Genera

In the raw datasets, SEPATH detected 1,617 genera and Poore *et al.* detected 1,993 on TCGA data¹⁵⁴. TCGA and SEPATH combined report 2,496 and 1,114 for union and intersect respectively. SEPATH detected 503 genera not in TCGA and TCGA detected 879 genera not in SEPATH.

Genera in Both TCGA and SEPATH

Abiotrophia Acetivibrio Acetobacter Acetobacterium Acetonema Acholeplasma Achromobacter Acidaminococcus Acidihalobacter Acidimicrobium Acidiphilium Acidithiobacillus Acidocella Acidothermus Acidovorax Acinetobacter Actinoalloteichus Actinobacillus Actinobaculum Actinomadura Actinomyces Actinomycetospira Actinoplanes Actinopolymorpha Actinopolyspora Actinotignum Adhaeribacter Adlercreutzia Advenella Aequorivita Aeribacillus Aerococcus Aeromicrobium Aeromonas Aestuariibacter Afifella Afipia Aggregatibacter Agrobacterium Agrococcus Agromyces Akkermansia Alcaligenes Alcanivorax Algoriphagus Aliagarivorans Alicyclophilus Alicyclobacillus Aliivibrio Alistipes Alkalibacterium Alkaliphilus Alkanindiges Alloactinosynnema Allobaculum Allochromatium Allofustis Alloiococcus Allokutzneria Allopevotella Alloscardovia Alphabaculovirus Alpha-papillomavirus Alpharetrovirus Alphatorquevirus Altererythrobacter Alteromonas Alysella Aminobacter Aminobacterium Aminomonas Amorphus Amphibacillus Amycolatopsis Anabaena Anaeroarcus Anaerobacillus Anaerococcus Anaerofustis Anaerolinea Anaeromusa Anaeromyxobacter Anaerosalibacter Anaerostipes Anaerotruncus Anaerovibrio Anaerovorax Anaplasma Andrevotia Aneurinibacillus Angustibacter Anoxybacillus Aphanizomenon Apibacter Aquamicrobium Aquimarina Aquincola Aquitalea Arcanobacterium Archangium Arcobacter Arenibacter Arenimonas Arhodomonas Aromatoleum Arsenicococcus Arsenophonus Arthrobacter Arthrospira Asaia Asanoa Ascovirus Asticcacaulis Atopobium Aurantimonas Aureimonas Aviadenovirus Azoarcus Azohydromonas Azonexus Azorhizobium Azospira Azospirillum Azotobacter Azovibrio Bacillus Bacteroides Barnesiella Bartonella Bdellovibrio Beggiatoa Beijerinckia Belliella Belnapia Bergeyella Betabaculovirus Betapapillomavirus Betaretrovirus Betatorquevirus Beutenbergia Bhargavaea Bifidobacterium Bilophila Bizionia Blastochloris Blastococcus Blastomonas Blastopirellula Blattabacterium Blautia Bordetella Borrelia Borreliella Bosea Brachyacterium Brachymonas Brachyspira Bradyrhizobium Brenneria Brevibacillus Brevibacterium Brevundimonas Brochothrix Brucella Buchnera Burkholderia Butyricicoccus Butyrimonas Butyrivibrio Caenimonas Caldalkalibacillus Caldanaerobacter Caldanaerobius Caldibacillus Caldicellulosiruptor Caldicoprobacter Caldilinea Caldimonas Calditerrivibrio Caloramator Calothrix Campylobacter Capnocytophaga Carboxydotherrhus Cardiobacterium Carnobacterium Casaltella Castellaniella Catelliglobospora Catonella Caulobacter Cedecea Celeribacter Cellulomonas Cellulosilyticum Cellulosimicrobium Cellvibrio Centipeda Cetobacterium Chamaesiphon Chania Chelativorans Chelatococcus Chitinibacter

Chitinilyticum Chitinimonas Chitiniphilus Chitinophaga Chlamydia Chloracidobacterium Chlorobaculum Chlorobium Chloroflexus Chlorovirus Chondromyces Christensenella Chromobacterium Chromohalobacter Chroococcidiopsis Chryseobacterium Chrysiogenes Chthoniobacter Circovirus Citreicella Citrobacter Citromicrobium Clavibacter Cloacibacillus Clostridioides Clostridium Cobetia Cohnella Coleofasciculus Collimonas Collinsella Colwellia Comamonas Conchiformibius Conexibacter Confluentimicrobium Congregibacter Coprobacillus Coprobacter Coprococcus Coprothermobacter Corallococcus Coriobacterium Corynebacterium Crinalium Croceicoccus Crocinitomix Cronobacter Cryobacterium Cryocola Cryptobacterium Cryptosporangium Cucumibacter Cupriavidus Curtobacterium Curvibacter Cyanobium Cyanothece Cyclobacterium Cycloclasticus Cyldrospermopsis Cyldrospermum Cystobacter Cytomegalovirus Cytophaga Dechloromonas Deefgea Defluviimonas Defluviitoga Dehalobacter Dehalococcoides Dehalogenimonas Deinococcus Delftia Deltaretrovirus Demequina Denitrobacterium Denitrovibrio Dependoparvovirus Dermabacter Dermacoccus Derxia Desmospora Desulfarculus Desulfatiglans Desulfatibacter Desulfatobacterium Desulfobacter Desulfobacterium Desulfobulbus Desulfocarbo Desulfococcus Desulfocurvus Desulfomicrobium Desulforegula Desulfosporosinus Desulfotalea Desulfotomaculum Desulfovermiculus Desulfovibrio Desulfurispirillum Devosia Dialister Diaphorobacter Dichelobacter Dickeya Dielma Dietzia Dinoroseobacter Dokdonella Dokdonia Dolosigranulum Dongia Dorea Draconibacterium Duganella Dyadobacter Dyella Dysgonomonas Echinicola Ectothiorhodospira Edaphobacter Edwardsiella Effusibacillus Eggerthella Eggerthia Eikenella Elijora Elizabethkingia Elusimicrobium Empedobacter Emticicia Enhydrobacter Enorma Ensifer Enterobacter Enterococcus Enterorhabdus Enterovibrio Eremococcus Erwinia Erysipelatoclostridium Erysipelothrix Erythrobacter Erythroparvovirus Escherichia Ethanoligenens Eubacterium Exiguobacterium Facklamia Faecalibacterium Faecalibaculum Fermentimonas Ferriphasesus Fervidobacterium Fibrella Fibrobacter Fictibacillus Filifactor Fimbriimonas Finegoldia Fischerella Flaviflexus Flaviramulus Flavisolibacter Flavobacterium Flavonifractor Flectobacillus Flexibacter Flexilinea Fluoribacter Fluviicola Formosa Francisella Franconibacter Frankia Frateuria Frigoribacterium Frischella Frondihabitans Fructobacillus Fulvimarina Fusobacterium Gallionella Gammagapillomavirus Gammaretrovirus Gammasphaerolipovirus Gammatorquevirus Gardnerella Gelidibacter Gemella Geminicoccus Geminocystis Gemmata Gemmatimonas Gemmobacter Geoalkalibacter Geobacillus Geobacter Geodermatophilus Geofilum Geothrix Gilliamella Gillisia Glaciecola Glaciibacter Gloeocapsa Gluconacetobacter Gluconobacter Glutamicibacter Glycomyces Gordonia Gordonibacter Gottschalkia Gracilibacillus Gramella Granulibacter Granulicatella Granulicella Granulicoccus Grylotalpicola Gulbenkiania Gulosibacter Gynuella Haematospirillum Haemophilus Hafnia Hahella Halalkalibacillus Halalkalicoccus Haliangium Haliea Halioglobus Haliscomenobacter Haloarcula Halobacteriovorax Halobacterium Haloferax Halolactibacillus Halomicrobium Halomonas Haloplanus Haloquadratum Halorhabdus Halorubrum Halostagnicola Halotalea Haloterrigena Halothermothrix Halothiobacillus Hamadaea Hapalosiphon Helcococcus Helicobacter Heliobacterium Herbaspirillum Herbiconiux Herbidospira Hermiinimonas Histophilus Hoeflea Holdemanella Holdemania Holophaga Holospora Hoyosella Hungatella Hydrocarboniphaga Hydrogenibacillus Hydrogenobacter Hydrogenophaga Hylemonella Hymenobacter Hyphomicrobium Hyphomonas Ideonella Idiomarina Ignatzschineria Ignavibacterium Ilumatobacter Inovirus Inquilinus Intestinibacter Intestinimonas Intraspangium Iridovirus Isophtericola Isosphaera Janibacter Jannaschia Janthinobacterium Jeotgalibaca Jeotgalibacillus Jeotgalicoccus Jiangella Johnsonella Jonesia Jonquetella Joostella Kaistia Kallipyga Kamptonema Kandleria Kangiella Kayvirus Kerstersia Ke-

togulonicigenium Kineococcus Kineosporia Kingella Kitasatospora Klebsiella Kluyvera Kocuria Komagataeibacter Kordia Kosakonia Kozakia Kribbella Ktedonobacter Kurthia Kushneria Kutzneria Kyrpidia Kytococcus Labrenzia Lachnoanaerobaculum Lachnobacterium Lachnoclostridium Lactobacillus Lactococcus Lampropedia Laribacter Lautropia Lawsonella Lawsonia Leadbetterella Lechevalieria Leclercia Leeuwenhoekii Legionella Leifsonia Leisingera Lelliottia Leminorella Lentimicrobium Lentivirus Lentzea Leptolynghya Leptonema Leptospira Leptothrix Leptotrichia Leucobacter Leuconostoc Levyella Limnohabitans Listeria Loktanella Longilinea Luteibacter Luteimonas Luteipulveratus Lutibacter Lymphocryptovirus Lysinibacillus Lysinimicrobium Lysobacter Macrococcus Mageibacillus Magnetospirillum Mahella Mangrovibacter Mannheimia Mardivirus Maribacter Marinithermus Marinobacter Marinobacterium Marinococcus Marinomonas Marinospirillum Marinovum Maritimibacter Marmoricola Marseillevirus Martellella Marvinbryantia Massilia Mastadenovirus Mastigocladopsis Megamonas Megasphaera Meiothermus Mesoflavibacter Mesonia Mesorhizobium Methanobacterium Methanobrevibacter Methanocella Methanocorpusculum Methanoculleus Methanolinea Methanomassiliicoccus Methanomethylovorans Methanosacta Methanosarcina Methanosphaera Methanosphaerula Methanospirillum Methanothermobacter Methylibium Methylobacillus Methylobacter Methylobacterium Methylocaldum Methylocapsa Methyloceanibacter Methylocella Methylococcus Methylocystis Methyломicrobium Methyломonas Methylophaga Methylophilus Methylosarcina Methylosinus Methyлотenera Methyloversatilis Methylovorus Methylovulum Micavibrio Microbacterium Microbispora Micrococcus Microcoleus Microcystis Microlunatus Micromonospora Microterricola Microtetraspora Microvirga Microvirgula Mimivirus Mitsuraria Mitsukella Mobilicoccus Mobiluncus Modestobacter Mogibacterium Molluscipoxvirus Moorea Moorella Moraxella Morganella Moritella Morococcus Mucilaginibacter Mucinivorans Mucispirillum Mupapillomavirus Muricauda Mycetocola Mycobacterium Mycoplasma Myroides Myxococcus Nakamurella Natrialba Natrinema Natronomonas Natronorubrum Neisseria Neorhizobium Neptunomonas Nesterenkonia Nevskia Niabella Niasella Nisaea Nitratireductor Nitriliruptor Nitrobacter Nitrococcus Nitrosococcus Nitrosomonas Nitrososphaera Nitrospira Nitrospira Niveispirillum Nocardia Nocardioidea Nocardioptosis Nodosilinea Nonlabens Nonomuraea Nostoc Noviherspirillum Novispirillum Novosphingobium Nupapillomavirus Obesumbacterium Oblitimonas Oceanicaulis Oceanicola Oceanobacillus Oceanobacter Oceanospirillum Ochrobactrum Octadecabacter Odoribacter Oenococcus Oerskovia Oleiagramonas Oleispira Oligella Oligotropha Olivibacter Olsenella Opitutus Orenia Oribacterium Orientia Ornithinimicrobium Ornithobacterium Orthohepadnavirus Oscillatoria Oscillibacter Ottowia Oxalobacter Paenarthrobacter Paenibacillus Paeniclostridium Paenirhodobacter Paenisporosarcina Paludibacter Paludibacterium Pandoraea Pannonibacter Pantoea Parabacteroides Paraburkholderia Parachlamydia Paracoccus Paraglaciacola Paraprevotella Pararhodospirillum Parascardovia Parvibaculum Parvimonas Pasteurella Patulibacter Paucibacter Paucisalibacillus Pectobacterium Pediococcus Pedobacter Pelagibaca Pelagibacterium Pelistega Pelobacter Pelomonas Pelosinus Peptoclostridium Peptoniphilus Peptostreptococcus Percavirus Phaeobacter Phaeospirillum Phaeovirus Phascolarctobacterium Phenylbacterium Phocaeicola Photobacterium Photorhabdus Phycococcus Phycisphaera Pimelobacter Pirellula Pithovirus Planctomyces Planctopirius Planktomarina Planktothricoides Planktothrix Planococcus Planomicrobium Plantibacter Pleomorphomonas Plesiomonas Pleurocapsa Pluralibacter Polaromonas Polymorphum Polynucleobacter Pontibacter Porphyrobacter Porphyromonas Pragia Prauserella Prevotella Promicromonospora Propionibacterium Propioniceella Propionimicrobium Propionispira Prosthecochloris Prosthecomicrobium Proteiniclasticum Proteiniphilum Proteus

Providencia Pseudanabaena Pseudarthrobacter Pseudoalteromonas Pseudobutyrvibrio Pseudoclavibacter Pseudodonghicola Pseudoduganella Pseudoflavonifractor Pseudogulbenkiania Pseudohongiella Pseudolabrys Pseudomonas Pseudonocardia Pseudooceanicola Pseudopedobacter Pseudophaeobacter Pseudoramibacter Pseudorhodoferax Pseudoxanthomonas Psychrobacillus Psychrobacter Psychroflexus Psychromonas Pusillimonas Rahnella Ralstonia Ramlibacter Raoultella Rathayibacter Renibacterium Reyranella Rhadinovirus Rheinheimera Rhizobacter Rhizobium Rhodanobacter Rhodobacter Rhodococcus Rhodoferax Rhodoluna Rhodomicrobium Rhodonellum Rhodoplanes Rhodopseudomonas Rhodospirillum Rhodothermus Rhodovulum Rickettsia Riemerella Rikenella Risunbinella Roseateles Roseburia Roseiflexus Roseivivax Roseobacter Roseolovirus Roseomonas Roseovarius Rothia Rouxiella Rubellimicrobium Rubritalea Rubritepida Rubrivivax Rubrobacter Rudaea Rudanella Ruegeria Rufibacter Ruminiclostridium Ruminococcus Rummeliibacillus Runella Saccharibacillus Saccharomonospora Saccharopolyspora Saccharothrix Salegentibacter Salimicrobium Salinarimonas Salinibacter Salinicoccus Salinimicrobium Salinimonas Salinispora Salinivibrio Salipiger Salmonella Sandaracinus Sandarakinorhabdus Sanguibacter Scardovia Schlesneria Sciscionella Scytonema Sebaldella Sedimenticola Sedimentitalea Sediminibacillus Sediminibacterium Segetibacter Selenomonas Senegalimasilia Serinicoccus Serratia Sharpea Shewanella Shigella Shimwellia Shinella Shuttleworthia Siccibacter Silanimonas Silvibacterium Simkania Simonsiella Simplexvirus Singulisphaera Sinomonas Sinorhizobium Slackia Sneathia Snodgrassella Sodalis Solibacillus Solimonas Solirubrobacter Solobacterium Sorangium Sphaerobacter Sphaerochaeta Sphaerotilus Sphingobacterium Sphingobium Sphingomonas Sphingopyxis Sphingorhabdus Spirochaeta Spiroplasma Spirosoma Spongiibacter Sporocytophaga Sporolactobacillus Sporomusa Sporosarcina Stackebrandtia Stanieria Staphylococcus Stappia Starkeya Stenotrophomonas Steroidobacter Stigmatella Stomatobaculum Streptacidiphilus Streptobacillus Streptococcus Streptomyces Streptosporangium Subdoligranulum Succinimonas Succinispira Succinivibrio Sulfitobacter Sulfuricella Sulfuricurvum Sulfurihydrogenibium Sulfurimonas Sulfuritalea Sulfurospirillum Sutterella Symbiobacterium Synechococcus Synechocystis Synergistes Syntrophobacter Syntrophobotulus Syntrophomonas Syntrophorhabdus Syntrophus Tannerella Tardiphaga Tatlockia Tatumella Taylorella Tenacibaculum Tepidanaerobacter Tepidimonas Tepidiphilus Teredinibacter Terrabacter Terracidiphilus Terracoccus Terribacillus Terriglobus Terrimonas Terrisporobacter Tessaracoccus Tetragenococcus Tetraparvovirus Tetrasphaera Thalassobacter Thalassobius Thalassolituus Thalassomonas Thalassospira Thauera Thermacetogenium Thermaerobacter Thermanaerovibrio Thermicanus Thermicola Thermithiobacillus Thermoactinomyces Thermoanaerobacter Thermoanaerobacterium Thermoanaerobaculum Thermobacillus Thermobaculum Thermobifida Thermobispora Thermobrachium Thermococcus Thermocrinis Thermocrisum Thermodesulfobacterium Thermodesulfovibrio Thermomonas Thermomonospora Thermonema Thermus Thioalkalivibrio Thiobacillus Thioclava Thioflavicoccus Thiomicrospira Thiomonas Thiorhodovibrio Thiothrix Timonella Tistrella Tolumonas Tolypothrix Trabulsiella Treponema Trichormus Tropheryma Tropicibacter Truepera Trueperella Tsukamurella Tuberibacillus Turicella Turicibacter Turneriella Tyzzerella Uliginosibacterium Ureaplasma Vagococcus Varibaculum Varicellovirus Variovorax Veillonella Verminephrobacter Verrucomicrobium Verrucosipora Vesiculovirus Vibrio Virgibacillus Vitreoscilla Vogesella Waddlia Weeksella Weissella Wenxinia Whispovirus Williamsia Winogradskyella Wohlfahrtiimonas Wolbachia Wolinella Xanthobacter Xanthomonas Xenophilus Xenorhabdus Xylanimonas Xylella Xylophilus Yersinia Yokenella Yonghaparkia Youngiibacter Zavarzinella Zhongshania Zunongwangia Zymbobacter Zymomonas

Genera in TCGA only

3alikevirus *Acaricomus* *Acaryochloris* *Acetohalobium* *Acidianus* *Acidibacillus* *Acidilobus*
Acidiplasma *Acidithrix* *Acidobacterium* *Aciduliprofundum* *Actibacterium* *Actinocatenispora*
Actinokineospora *Actinospica* *Actinosynnema* *Actinotalea* *Aeropyrum* *Aestuariimicro-*
bium *Aestuariivita* *Agarivorans* *Agreia* *Ahjdlikevirus* *Ahrensia* *Alfamovirus* *Algibacter*
Algicola *Algiphilus* *Aliiglaciecola* *Aliihoefta* *Aliiroseovarius* *Alishewanella* *Aliterella*
Alkaliflexus *Alkalilimnicola* *Allexivirus* *Allolevivirus* *Allosalinactinospora* *Alphacoro-*
navirus *Alphaentomopoxvirus* *Alphafusellovirus* *Alphamesonivirus* *Alphapartitivirus*
Alphasphaerolipovirus *Alphaturrivirus* *Alphavirus* *Altibacter* *Amantichitinum* *Ambidenso-*
virus *Aminiphilus* *Ammonifex* *Ampelovirus* *Amphritea* *Ampullavirus* *Amycolicococcus*
Anaerobaculum *Anaerophaga* *Anaerosporemusa* *Ancylobacter* *Anditalea* *Andromedali-*
kevirus *Aparavirus* *Aphanocapsa* *Aphthovirus* *Apscaviroid* *Aquabacterium* *Aquamavirus*
Aquaparamyxovirus *Aquareovirus* *Aquifex* *Archaeoglobus* *Arcticibacter* *Ardenticatena*
Arenavirus *Arenitalea* *Arsukibacterium* *Arterivirus* *Asfivirus* *Atadenovirus* *Aureispira*
Austwickia *Avastrovirus* *Avibacterium* *Avihepadnavirus* *Avipoxvirus* *Avulavirus* *Babuvirus*
Bacilladnavirus *Bacillarnavirus* *Bacteriovorax* *Badnavirus* *Bafinivirus* *Balneatrix* *Balneola*
Barnavirus *Barnyardlikevirus* *Basfia* *Basilea* *Batrachovirus* *Bcep22likevirus* *Bcep78likevirus*
Bdellomicrovirus *Begomovirus* *Bellilinea* *Benyvirus* *Bermanella* *Betacoronavirus* *Betaento-*
mopoxvirus *Betafusellovirus* *Betalipothrixvirus* *Betapartitivirus* *Bibersteinia* *Bicaudavirus*
Bignuzlikevirus *Bocaparovirus* *Bornavirus* *Botrexvirus* *Bppunalikevirus* *Brackiella* *Br-*
acovirus *Brevidensovirus* *Bromovirus* *Bronlikevirus* *Bryobacter* *Bulleidia* *Buttiauxella*
Bymovirus *C2likevirus* *C5likevirus* *Caballeronia* *Caedibacter* *Caenispirillum* *Cafeteriavirus*
Caldimicrobium *Caldisericum* *Caldisphaera* *Caldithrix* *Caldivirga* *Caminibacter* *Candidatus*
Accumulibacter *Candidatus* *Amoebophilus* *Candidatus* *Aquiluna* *Candidatus* *Arthromitus*
Candidatus *Atelocyanobacterium* *Candidatus* *Azobacteroides* *Candidatus* *Babela* *Candidatus*
Blochmannia *Candidatus* *Brocadia* *Candidatus* *Cardinium* *Candidatus* *Cloacimonas* *Can-*
didatus *Competibacter* *Candidatus* *Contendobacter* *Candidatus* *Desulforudis* *Candidatus*
Endolissoclinum *Candidatus* *Evansia* *Candidatus* *Glomeribacter* *Candidatus* *Halobonum*
Candidatus *Hamiltonella* *Candidatus* *Hepatobacter* *Candidatus* *Ishikawaella* *Candidatus*
Jettenia *Candidatus* *Korarchaeum* *Candidatus* *Koribacter* *Candidatus* *Liberibacter* *Candi-*
datus *Magnetobacterium* *Candidatus* *Methanomethylophilus* *Candidatus* *Methanoperedens*
Candidatus *Methanoplasma* *Candidatus* *Methylopumilus* *Candidatus* *Microthrix* *Candi-*
datus *Midichloria* *Candidatus* *Moranella* *Candidatus* *Nasuia* *Candidatus* *Neoehrlichia*
Candidatus *Nitrosoarchaeum* *Candidatus* *Nitrosocosmicus* *Candidatus* *Nitrosopelagicus*
Candidatus *Nitrosotenuis* *Candidatus* *Odyssella* *Candidatus* *Paracaedibacter* *Candidatus*
Pelagibacter *Candidatus* *Phaeomarinobacter* *Candidatus* *Photodesmus* *Candidatus* *Phyto-*
plasma *Candidatus* *Portiera* *Candidatus* *Profftella* *Candidatus* *Protochlamydia* *Candidatus*
Puniceispirillum *Candidatus* *Regiella* *Candidatus* *Riesia* *Candidatus* *Ruthia* *Candidatus*
Schmidhempelia *Candidatus* *Soleaferrea* *Candidatus* *Solibacter* *Candidatus* *Stoquefichus*
Candidatus *Sulcia* *Candidatus* *Symbiobacter* *Candidatus* *Tachikawaea* *Candidatus* *Xiphine-*
matobacter *Capillovirus* *Capripoxvirus* *Carlavirus* *Carmovirus* *Carnimonas* *Catabacter*
Catellibacillus *Catenovulum* *Catenulispora* *Catenuloplanes* *Caulimovirus* *Cavemovirus*
Cecembia *Cellulophaga* *Cephalotococcus* *Cervidpoxvirus* *Cesiribacter* *Charlielikevirus*
Che8likevirus *Che9clikevirus* *Chelonobacter* *Cheravirus* *Chilikevirus* *Chitinispirillum*
Chitinivibrio *Chloriridovirus* *Chlorogloeopsis* *Chloroherpeton* *Chrysovirus* *Chthonomonas*
Cilevirus *Citricoccus* *Cjwunalikevirus* *Closterovirus* *Cocadviroid* *Coccolithovirus* *Coleviroid*
Coltivirus *Commensalibacter* *Comovirus* *Copiparvovirus* *Coralimargarita* *Corndog-*

likevirus Corticovirus Cosavirus Cosenzaea Coxiella Cp220likevirus Cp8unalikevirus
Crenobacter Criblamydia Crinivirus Cripavirus Croceibacter Croceitalea Crocodylid-
porvirus Crocosphaera Cucumovirus Curtovirus Cyanobacterium Cypovirus Cyprinivirus
D3likevirus Dactylococcopsis Dactylosporangium Dasania Deferribacter Defferisoma
Deltabaculovirus Deltacoronavirus Deltalipothrixvirus Deltapapillomavirus Demetria Der-
matophilus Desulfatibacillum Desulfatirhabdium Desulfatitalea Desulfobacca Desulfobacula
Desulfocapsa Desulfohalobium Desulfomonile Desulfonatronospira Desulfonatronovibrio
Desulfonatronum Desulfosarcina Desulfospira Desulfotignum Desulfurella Desulfurispora
Desulfurivibrio Desulfurobacterium Desulfurococcus Desulfuromonas Dethiobacter Dethio-
sulfatarculus Dethiosulfovibrio Dicipivirus Dictyoglomus Dinovernavirus Diplorickettsia
Diplosphaera Dolichospermum Domibacillus Dyoetapapillomavirus Dyolambdapapillo-
mavirus Dyoomikronpapillomavirus Dyopipapillomavirus Ebolavirus Ehrlichia Emaravirus
Enamovirus Endomicrobium Endornavirus Endozoicomonas Enteractinococcus Enterovirus
Entomoplasma Ephemerovirus Epilithonimonas Epsilonretrovirus Epsilontorquevirus
Epulopiscium Eudoraea Euryhalocaulis Ewingella Fabavirus Faecalicoccus Falsirhodobacter
Fangia Felixounalikevirus Ferrimicrobium Ferrimonas Ferroglobus Ferroplasma Ferrovum
Fervidicella Fibrisoma Fijivirus Filomicrobium Flagellimonas Flammeovirga Flavihu-
mibacter Flavivirus Flexistipes Flexithrix Fodinicurvata Foveavirus Fulvivirga Furovirus
Gaetbulibacter Galbibacter Gallaecimonas Gallibacterium Gallivirus Gammabaculovirus
Gammacoronavirus Gammalipothrixvirus Gayadomonas Geitlerinema Gemmatirosa
Geoglobus Geopsychrobacter Geovibrio Gilvimarinus Gimesia Globulovirus Gloebac-
ter Glossinavirus Gorillibacterium Gracilimonas Grimontia Gyrovirus Haematobacter
Haladaptatus Halanaeroarchaeum Halanaerobium Halapricum Halarchaeum Hallella
Halobacillus Halobacteroides Halobellus Halobiforma Halococcus Halocynthiibacter Halofer-
ula Halogeometricum Haloglycomyces Halogranum Halolamina Halonatronum Halopiger
Haloplasma Halorhodospira Halosimplex Halothece Halovivax Hantavirus Hapunalikevirus
Hassallia Helcobacillus Hellea Henipavirus Henriciella Hepacivirus Hepandensovirus Hepa-
tovirus Higrevirus Hippea Hirschia Hk578likevirus Hordeivirus Hpunalikevirus Humibacter
Hyalangium Hydrogenivirga Hydrogenobaculum Hydrogenovibrio Hyperthermus Hypovirus
I3likevirus Ichnovirus Ictalurivirus Idaeovirus Iebhlikevirus Iflavirus Ignicoccus Ilarvirus
Iltovirus Ilyobacter Imtechella Indibacter Influenzavirus A Influenzavirus C Iotatorquevirus
Ipomovirus Isavirus Iteradensovirus Jejuia Jerseylikevirus Kappapapillomavirus Kappa-
torquevirus Kibdelosporangium Kiloniella Kineosphaera Kinetoplastibacterium Knoellia
Kobuvirus Kordiimonas Kosmotoga L5likevirus Labilithrix Lacimicrobium Lacinutrix
Lambdallikevirus Lambdapapillomavirus Lamprocystis Lebetimonas Leeia Leishmaniovirus
Lentibacillus Lentisphaera Leporiporvirus Leptolinea Leptospirillum Leucothrix Lev-
ilinea Levivirus Lewinella Limnobacter Limnochorda Litoreibacter Lolavirus Longispora
Luminiphilus Lunatimonas Luteovirus Lutibaculum Luz24likevirus Lymphocystivirus
Lyngbya Lyssavirus Macavirus Maculavirus Magnetococcus Magnetospira Mamastrovirus
Mammarenavirus Mandarivirus Mangrovimonas Marafivirus Marburgvirus Maribius
Maricaulis Marichromatium Marinagarivorans Marinifilum Marinilabilia Marinimicrobium
Mariniradius Marinitoga Mariprofundus Maritalea Marivirga Marnavirus Mastigo-
coleus Mastrevirus Megabirnavirus Megalocytivirus Meganema Megrivirus Melioribacter
Melissococcus Mesoaciditoga Mesoplasma Mesotoga Metallosphaera Metapneumovirus
Methanocaldococcus Methanococcoides Methanococcus Methanofollis Methanolalobium
Methanolalophilus Methanolacinia Methanolobus Methanomicrobium Methanoplanus
Methanoregula Methanosalum Methanothermococcus Methanothermus Methanotorris
Methermicoccus Methylacidiphilum Methyloferula Methylohalobius Methylomarinum

Methylopila Microbulbifer Microchaete Microscilla Microvirus Mimoreovirus Mitovirus Mizugakiibacter Moellerella Morbillivirus Mulikevirus Muribacter Muromegalovirus Muscavirus Mycoflexivirus Mycoreovirus Myxosarcina N15likevirus N4likevirus Nafulsella Nairovirus Nanovirus Narnavirus Natranaerobius Natronobacterium Natronococcus Natronolimnobiuss Nautella Nautilia Necropsobacter Negevirus Neochlamydia Neorickettsia Neosynechococcus Nepovirus Nereida Nesiotobacter Nitratifractor Nitratiruptor Nitrincola Nitritalea Nitrosopumilus Nitrospina Nodularia Norovirus Novirhabdovirus Nucleorhabdovirus Nyavirus Oceanibaculum Oceanibulbus Oceanimonas Oceaniovalibus Oceanithermus Okavirus Oleavirus Oleiphilus Olleya Omegalikevirus Omegapapillomavirus Omikronpapillomavirus Orbivirus Ornatilinea Ornithinibacillus Orthobunyavirus Orthopneumovirus Orthopoxvirus Oscillochloris Ostreavirus Owenweeksia Oxobacter P100virus P23likevirus P2likevirus Paeniglutamicibacter Palaeococcus Panicovirus Paraclostridium Paramesorbizobium Paraoerskovia Parapoxvirus Parechovirus Parvularcula Pasivirus Passerivirus Pbiunalikevirus Pbunalikevirus Pecluvirus Pedosphaera Pegivirus Pelamoviroid Pelodictyon Perhabdovirus Perlucidibaca Persephonella Persicobacter Pestivirus Petrotoga Pgonelikevirus Phaeodactylibacter Phaseolibacter Phi29likevirus Phic3unalikevirus Phicbklikevirus PhiCD119likevirus Phieco32likevirus Phietalikevirus PhiJlunalikevirus Phikmlikevirus Phikzlikevirus Phipapillomavirus Phlebovirus Phyllobacterium Phytoreovirus Picophilus Pilimelia Pipapillomavirus Piscihepevirus Piscirickettsia Planomonospora Plasmavirus Plectrovirus Plesiocystis Pneumovirus Poacevirus Polaribacter Polemovirus Polerovirus Polycyclovorans Polyomavirus Pomovirus Pontibacillus Ponticaulis Ponticoccus Porticoccus Potexvirus Potyvirus Prasinovirus Proboscivirus Prochlorococcus Prochlorothrix Prolixibacter Protoparvovirus Prymnesiovirus Pseudacidovorax Pseudaminobacter Pseudobacteroides Pseudoglutamicibacter Pseudohalaea Pseudorhodobacter Pseudothermotoga Pseudovibrio Psychrilyobacter Psychroserpens Puniceibacterium Pyramidobacter Pyrinomonas Pyrococcus Pyrodictium Pyrolobus Ranavirus Raphidiopsis Reinekea Respirovirus Rhodopirellula Rhodovibrio Rhopapillomavirus Richelia Rickettsiella Rivularia Robiginitalea Robiginitomaculum Robinsoniella Rosadnavirus Rosavirus Roseibacterium Roseibium Roseivirga Rotavirus Ruania Rubidibacter Rubinisphaera Rubivirus Rubulavirus Rudivirus Rymovirus Saccharibacter Saccharicrinis Saccharophagus Saccharosporillum Sadwavirus Sagittula Sakobuvirus Salinarchaeum Salinibacillus Salinibacterium Salinisphaera Salinispira Salisaeta Salivirus Salsuginibacillus Sanguibacteroides Sap6likevirus Sapelovirus Saprospira Schizot4likevirus Schleiferia Sclerodarnavirus Scutavirus Sediminibacter Sediminicola Sediminimonas Segniliparus Sellimonas Senecavirus Sfi1unalikevirus Sfi21dtunalikevirus Shimazuella Shimia Siadenovirus Siansivirga Sicinivirus Sideroxydans Silviavirus Simidua Simplicispira Skermanella Skunalikevirus Smaragdicoccus Sneathiella Sobemovirus Solendovirus Solitalea Soymovirus Sp6likevirus Spiribacter Spirillospora Spiromicrovirus Spirulina Spo1virus Sporichthya Spounalikevirus Spumavirus Staphylothermus Stenoxybacter Streptomonospora Suipoxvirus Sulfobacillus Sulfolobus Sulfurovum Sunxiuqinia Syntrophaceticus Syntrophothermus T4likevirus T5likevirus T7likevirus Tamiana Tanticharoenia Tateyamaria Taupapillomavirus Tectivirus Tenuivirus Tepidicaulis Terasakiella Terrimicrobium Thalassobacillus Thalassobaculum Thalassobium Thalassotalea Thermanaerothermox Thermodesulfatator Thermodesulfobium Thermofilum Thermogemmatisspora Thermogladius Thermomicrobium Thermopetrobacter Thermoplasma Thermoproteus Thermorudis Thermosediminibacter Thermosinus Thermosipho Thermosphaera Thermosulfidibacter Thermosulfurimonas Thermosynechococcus Thermotoga Thermovibrio Thermovirga Thetatorquevirus Thioalkalimicrobium Thiocapsa Thiocystis Thiohalocapsa Thiolapillus Thioploca Thiorhodococcus Thiorhodospira Tibrovirus

Tm4likevirus Tobamovirus Tombusvirus Tomitella Torradovirus Tospovirus Totivirus Tremovirus Trichodesmium Trichomonasvirus Trichovirus Tritimovirus Tunalikevirus Tungrovirus Twortlikevirus Tymovirus Ulvibacter Varicosavirus Velarivirus Ventosimonas Vesivirus Victorivirus Viridibacillus Vitellibacter Vitivirus Viunalikevirus Vulcanisaeta Vulgatibacter Waikavirus Wenzhouxiangella Wigglesworthia Woodsholea Xanthomarina Xenococcus Xp10likevirus Yaniella Yatapoxvirus Yualikevirus Zetatorquevirus Zhouia Zobellia Zooshikella

Genera in SEPATH only

Ab18virus Acanthamoeba Acetanaerobacterium Acetitomaculum Acetoanaerobium Acidipropionibacterium Acidisphaera Acidomonas Acytostelium Agaricus Aidingimonas Albidovulum Albimonas Algoriella Aliicoccus Alkalimonas Alkalispirochaeta Alkalitalea Allisonella Alpha3microvirus Alphapolyomavirus Alternaria Anabaenopsis Anaerobiospirillum Anaerobium Anaerobranca Anaerocolumna Anaeroglobus Anaeromassilibacillus Anaerosphaera Anaerosporobacter Anaerotignum Anaerovirgula Angelakisella Antartobacter Aphanomyces Aquaspirillum Aquiflexum Aquimonas Arabia Arachidicoccus Arsenicibacter Arthrobotrys Asaccharospora Ascoidea Asinibacterium Aspergillus Atopococcus Atopostipes Aurantimicrobium Auraticoccus Aureobasidium Aureococcus Auricoccus Auricularia Babesia Bariatricus Batrachochytrium Beauveria Bernardetia Betapolyomavirus Bigelowiella Bipolaris Bittarella Blastocystis Blastomyces Blumeria Botrytis Bpp1virus Brevefilum Budvicia C2virus C5virus Caecibacter Candida Candidatus Accumulibacter Candidatus Amoebophilus Candidatus Aquiluna Candidatus Arthromitus Candidatus Azobacteroides Candidatus Babela Candidatus Blochmannia Candidatus Brocadia Candidatus Cloacimonas Candidatus Competibacter Candidatus Dactylopiibacterium Candidatus Desulforudis Candidatus Frackibacter Candidatus Hamiltonella Candidatus Hoaglandella Candidatus Kryptobacter Candidatus Kryptonium Candidatus Methanomethylophilus Candidatus Methanoplasma Candidatus Methylopusillus Candidatus Microthrix Candidatus Nanopelagicus Candidatus Nitrosotenuis Candidatus Nucleicultrix Candidatus Paracaedibacter Candidatus Pelagibacter Candidatus Phaeomarinobacter Candidatus Phytoplasma Candidatus Planktophila Candidatus Protochlamydia Candidatus Regiella Candidatus Solibacter Candidatus Stoquefichus Candidatus Symbiobacter Candidatus Thiodiazotropha Candidatus Thioglobus Candidimonas Capronia Carboxydocella Cc31virus Chaetomium Chipapillomavirus Chishuiella Chivirus Chondrocystis Chryseolinea Citreimonas Cladophialophora Clavispora Cloacibacterium Clostridiisalibacter Cnuella Cnuibacter Coccidioides Colletotrichum Coniophora Coniosporium Cordyceps Cp1virus Cr3virus Crenotalea Cronusvirus Cryptococcus Cryptosporidium Culturomica Cutaneotrichosporon Cutibacterium Cvm10virus Cyberlindnera Cyclospora Cyphellophora D3virus Debaryomyces Defluviitalea Dehalobacterium Deltapolyomavirus Dendrosporobacter Desnuesiella Desulfofustis Desulfoluna Desulfoviregula Dethiosulfatibacter Devriesea Dichomitus Dictyostelium Diplodia Duodenibacillus Ea92virus Edaphobacillus Eimeria Eisenbergiella Elstera Elvirus Emiliana Endocarpon Entamoeba Enterocytozoon Eutypa Exophiala Ezakiella Faecalitalea Felixo1virus Fenollaria Fibroporia Filimonas Fomitiporia Fonsecaea Fontimonas Fortiea Fournierella Fri1virus Friedmanniella Fuerstia Fusarium Fusicatenibacter G7civirus Gabonia Gabonibacter Gaeumannomyces Garciella Gemmiger Gemycircularvirus Gemykrogvirus Geosporobacter Geotoga Giardia Gibbsiella Giesbergeria Glarea Globicatella Gloeophyllum Granulosicoccus Gregarina Grosmannia Guillardia Haloparvum Halopenitus Halorientalis Hammondia Hapnavirus

Hathewayia Hemiselmis Herbinix Hespellia Heterobasidion Hk578virus Homo Hp1virus Hydrococcus Hydrogenoanaerobacterium Hyphopichia Ichthyophthirius Ignavigranum Immundisolibacter Inediibacterium Insolitispirillum Intestinibacillus Isaria Izhakiella Jatrophihabitans Jd18virus Jeongeupia Jerseyvirus Jwxvirus K1gvirus Kazachstania Kiritimatiella Kluyveromyces Komagataella Kp15virus Kp34virus Kp36virus Kpp10virus Kpp25virus Krasilnikoviella Kuraishia Kwoniella Laccaria Lachnospira Lactonifactor Lacunisphaera Lambdavirus Leishmania Leptomonas Leptosphaeria Libanicoccus Liberibacter Limimonas Limnoraphis Lineavirus Litorimicrobium Lodderomyces Lonsdalea Lutispora Magnaporthe Mailhella Malassezia Mameiella Marinactinospora Marinilactibacillus Mariniphaga Marssonina Massilibacterium Mastigocladus Mediterranea Melampsora Merdibacter Metallibacterium Metarhizium Meyerozyma Microcella Moesziomyces Mooglevirus Murdochhiella Muvirus N4virus Naegleria Nakaseomyces Naumovozyma Ndongobacter Negativicoccus Neglecta Neoasaea Neofusicoccum Neomegalonema Neomicrococcus Neurospora Niameybacter Nit1virus Nitrosovibrio Nitrospirillum Nonagvirus Nonanavirus Nosocomiicoccus Novibacillus Np1virus Oceanicella Oceanisphaera Oceanivirga Ogataea Ohtaekwangia Oleibacter Oryza P22virus P2virus Pa6virus Pacificimonas Paenalcaldigenes Pajaroellobacter Paludisphaera Pamx74virus Papillibacter Paracoccidioides Parageobacillus Paramecium Parapedobacter Paraphaeosphaeria Parastagonospora Parasutterella Pbnavirus Pelagirhabdus Penicilliopsis Penicillium Peptoanaerobacter Peptococcus Perkinsus Pestalotiopsis Petrimonas Phaeoacremonium Phaeodactylum Phanerochaete Phialocephala Phico32virus Phietavirus Phifelvirus Phikmvvirus Phikzvirus Phoea Phoenicibacter Phormidesmis Phormidium Phycomyces Phytophthora Pichia Pisciglobus Planifilum Plantactinospora Plasmodium Pleurotus Podospora Postia Pradovirus Prevotellamassilia Pricia Propionispora Propionivibrio Prosthecobacter Proteiniborus Proteocatella Pseudarcicella Pseudocercospora Pseudodesulfovibrio Pseudogymnoascus Pseudopropionibacterium Pseudorhodoplanes Pseudosphingobacterium Pseudoanthobacter Pseudozyma Puccinia Purpureocillium Pyrenophora Raineyella Rb69virus Rhinocladia Rhizorhabdus Rhodotorula Rodentibacter Romboutsia Roseinatronobacter Rosenbergiella Rtpvirus Rubrivirga Rugamonas Ruminobacter Saccharomyces Salinihabitans Salipaludibacillus Saprolegnia Sarcina Scedosporium Scheffersomyces Schizophyllum Schizosaccharomyces Schwartzia Sclerotinia Se1virus Secunda5virus Seonamhaeicola Sep1virus Septima3virus Serpula Setosphaeria Seuratvirus Sextaecvirus Sfi11virus Sfi21dt1virus Siccationidurans Sinomicrobium Siphonobacter Sk1virus Soonwooa Sordaria Sp18virus Sp6virus Spathaspora Sphaerospermopsis Sphaerulina Spizellomyces Spn3virus Sporanaerobacter Sporobacter Sporolituus Stereum Succinatimonas Succiniclasticum Sulfuriferula Sulfurifustis Suspivirus T1virus T4virus T5virus T7virus Talaromyces Tepidibacillus Tepidibacter Tepidimicrobium Tetrahymena Thalassiosira Thermoflexibacter Thermogutta Thermoleophilum Thermophaeus Thermotheleomyces Thielavia Thiobacimonas Tidjanibacter Tin2virus Tissierella Tistlia Tlsvirus Torulaspora Toxoplasma Trametes Tranquillimonas Traorella Trichococcus Trichoderma Trichomonas Trichophyton Trichosporon Tropicimonas Trypanosoma Tuber Tumebacillus Twortvirus Ureibacillus Urinacoccus Ustilago V5virus Vaginella Variibacter Verruconis Verticillium Vhmlvirus Vittaforma Wallemia Wickerhamomyces Williamwhitmania Xp10virus Yangia Yarrowia Ydn12virus Yuavirus Yuhushiella Zea Zobelella Zygosaccharomyces Zymoseptoria

D.7 Searching for Correlations

Using the database filtered of only PCR-free and fresh frozen tissue, 1,004 genera were compared against 180 different quantitative variables such as surrogates for disease aggression such as mutational signature, single nucleotide variation, copy number variation, the number of insertions and deletions. Of a total of 158,861 Spearman's correlation tests. 719 tests had $\rho > 0.5$ or < -0.5 with an adjusted $p < 0.1$. Due to the sparsity of the dataframe, the summary statistics for these 'significant' correlations regarding the number of taxa in each test is as follows: min 6, median 15, IQR 14, max 166. Only 38 comparisons involved taxa that were present in more than 100 samples (approximately 1% of samples).

After filtering for significant correlations with ($Q < 0.1, |\rho| > 0.5$) the top top ten highest absolute of rho values are for Septima3virus and read coverage (a phage of a common contaminant bacteria⁵⁴⁶). *Syntrophus* was positively correlated with non-synonymous SNVs, SNV Het/Hom ratio, SNVs in coding regions and SNVs in genes ($\rho = 0.987$ in all cases). This may be a fascinating find but this analysis is hindered by the detection of this taxa in six samples in the entire dataset. *Desulfobulbus* is correlated with autosome coverage ($\rho = 0.984$) but again this is based on a total observed 6 taxa. *Oligella* is negatively correlated with signature 29 based on an observation of six taxa ($\rho = -0.98$). Mastadenovirus had a negative correlation with SNVs in mature miRNA ($\rho = -0.97$, number of positive samples = 6). This was suggested as a contaminant in a recent investigation into the cancer virome⁹⁶.

To switch the focus, three bacteria demonstrated a significant positive correlation in above ten samples for SNVs: *Sandarakinorhabdus*, *Oenococcus*, *Microcella* ($\rho = 0.53, 0.65, 0.74$ and $n=24, 21, 12$ respectively). Species of *Sandarakinorhabdus* have been isolated from freshwater lakes⁵⁴⁷⁻⁵⁴⁹. *Oenococcus* is found in wine⁵⁵⁰. Similarly it is well adapted to living in low pH environments with high lactic acid content⁵⁵¹. So at least in part there is a rationale for the identification of *Oenococcus* from tumour samples. There were only two samples in the entire dataset with above 10 classified reads to this genus (353 and 14 reads). This correlation is therefore unlikely to be a significant find. *Microcella* species additionally appear to have been isolated from water sources⁵⁵² and was present in 12 samples in this analysis and therefore is additionally unlikely to be a significant find.

Only considering correlations with above 50 taxa detected there is only one taxa that demonstrates a correlation with a variable which is a surrogate for disease aggression (*i.e.* CNVs, SNVs, indels, structural variation): *Trypanosoma* and structural variation. These correlations however are negative and not particularly strong ($-0.5 < \rho < -0.57$) despite strong significance ($0.0033 < Q < 0.00035$). These variables include:SV Deletions (All), SV Insertions, SV Insertions (All), SV Tandem Duplications, SV Insertions in genes, SV Breakends, SV Tandem Duplications in genes, SV Breakends in genes, SV Breakends (All), SV Tandem Duplications (All), SV Deletions. *Trypanosoma* is a parasitic protozoa with two notable (often asymptomatic) pathogenic species: *Trypanosoma brucei* that causes sleeping sickness and *Trypanosoma cruzi* that causes Chagas disease⁵⁵³. *Trypanosoma* infection has previously been associated with less aggressive cancer by eliciting an anti-tumour immune response^{554,555}. This suggests that this finding is true despite a small correlation value. In the raw dataset however there is a lack of overwhelming read evidence for *Trypanosoma*. Considering the number of reads in positive samples, there is a 1st quartile of 1, a median of 1.5, 3rd quartile of 4 and a maximum of 20. The positive samples (with no read threshold

applied) include a range of cancer types: 1 adult glioma, 3 bladder, 8 breast, 8 colorectal, 1 endocrine, 7 endometrial, 3 haemonec (Haematological oncology), 6 lung, 6 oral, 2 prostate, 5 renal, 2 sarcoma and 2 sinonasal.

D.8 Repeating Models with Different Genera

Below shows a repeat of the TCGA one-vs-all tumour models with stringent contamination removed

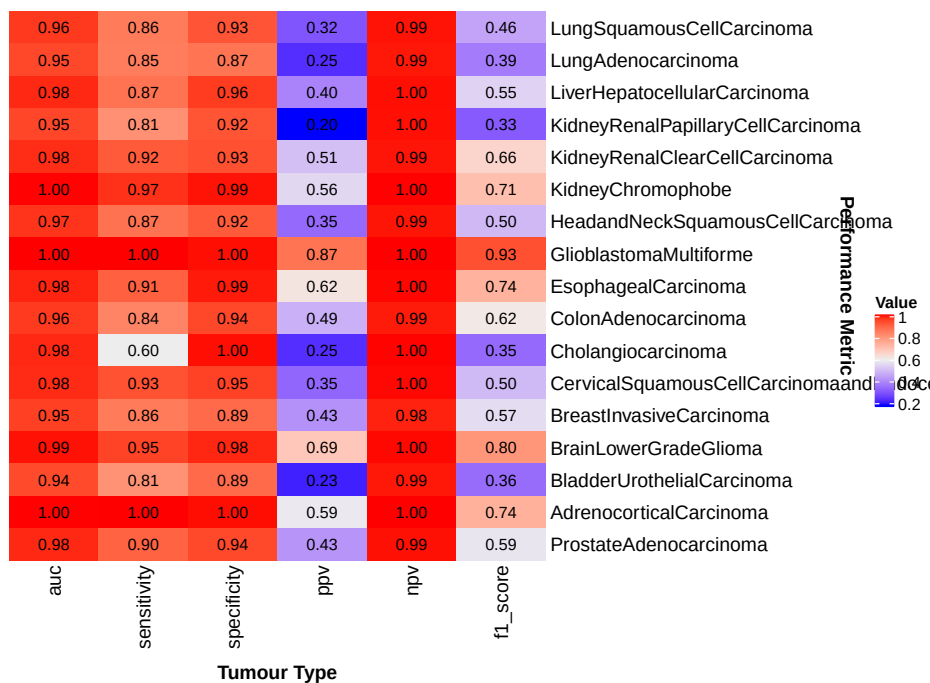


Figure D.5: A repeat of the TCGA one-vs-all tumour classification models with with stringent filters for contamination removed.

D.8.1 Repeating TCGA models with different genera

The TCGA models with stringent contaminants removed were repeated using only genera that were identified in both TCGA and SEPATH on (figure below). In both cases we can see that the model performance rarely changes and in fact often improves performance overall.

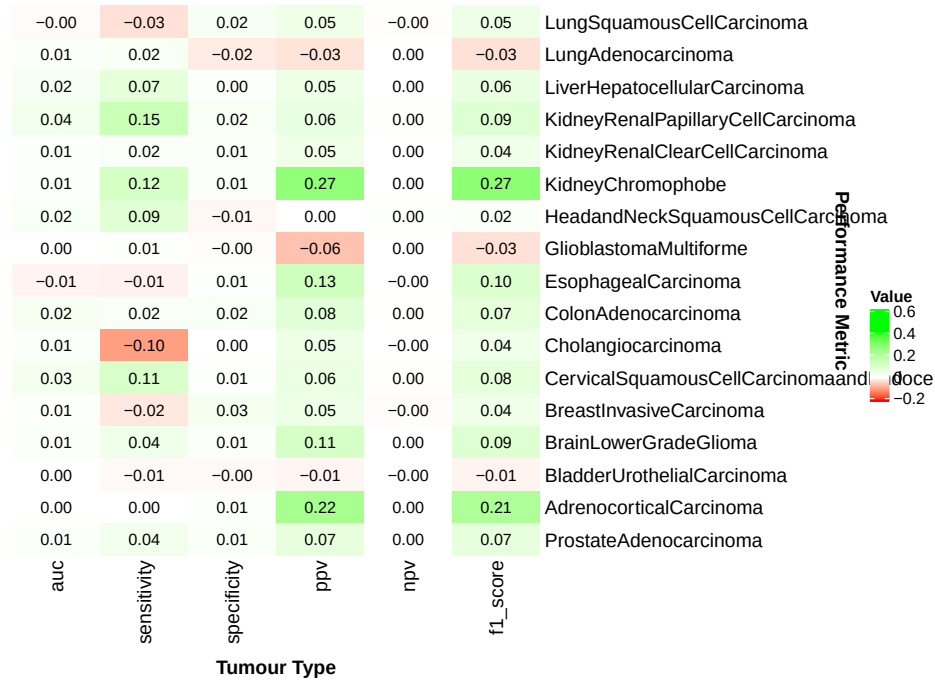


Figure D.6: A repeat of the TCGA one-vs-all tumour classification models with with stringent filters for contamination removed. These only include genera found in both SEPATH on and TCGA. Values in the cells denote the change from the models when ran on all genera in the original TCGA publication

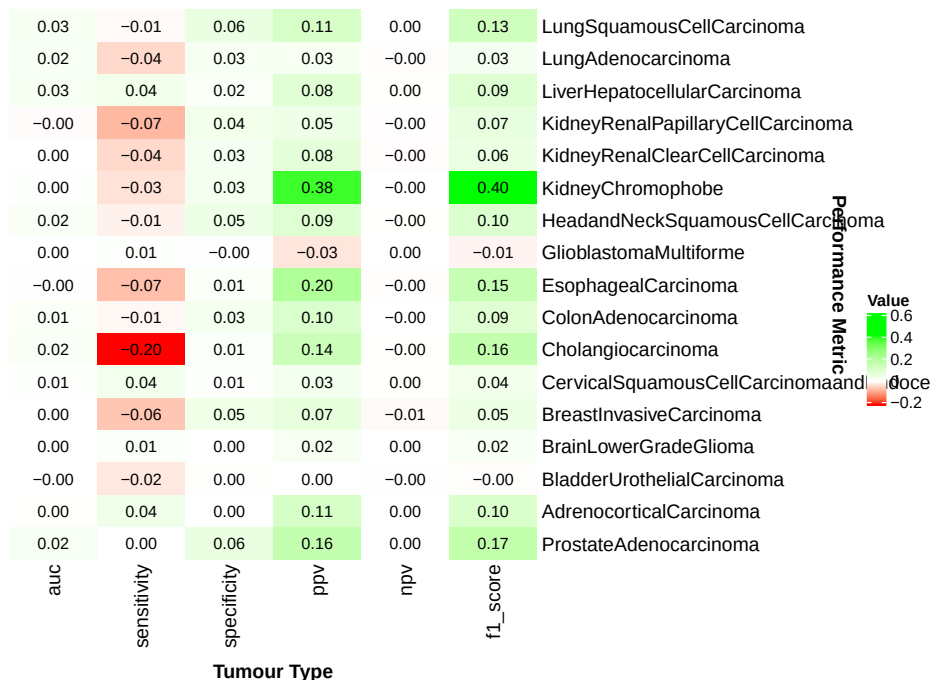


Figure D.7: A repeat of the TCGA one-vs-all tumour classification models with with stringent filters for contamination removed. These only include genera found in both SEPATH on and TCGA. Values in the cells denote the change from the models when ran on all genera in the original TCGA publication

The models were additionally repeated using only genera found in TCGA and not with SEPATH on Genomics England Data (figure D.8)

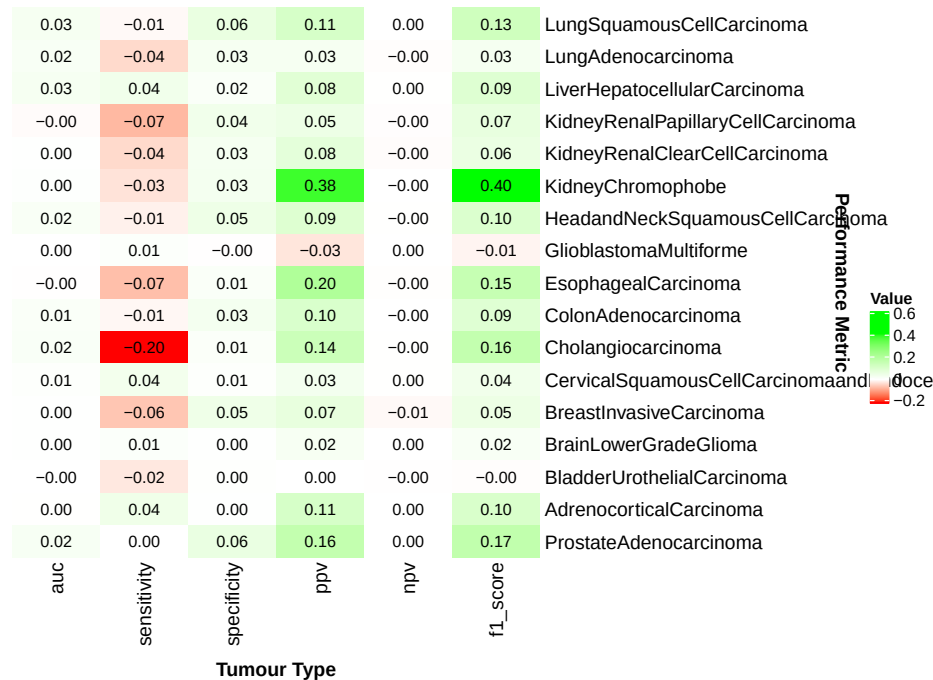


Figure D.8: A repeat of the TCGA one-vs-all tumour classification models with with stringent filters for contamination removed. These only include genera found in TCGA but not in SEPATH. Values in the cells denote the change from the original model

D.9 GBM and Boruta Tumour Type - Feature Importance

This section describes the feature importance of GBM models and using boruta for classification models of one tumour type vs all others. Only the top few features will be shown. The remainder of the feature importance can be found on the GitHub repository for this thesis. Please note that feature importance is not comparable between approaches and is relative to the other taxa in the analysis.

Regarding Boruta, only features with a median importance value above the maximum shadow variable are confirmed as important.

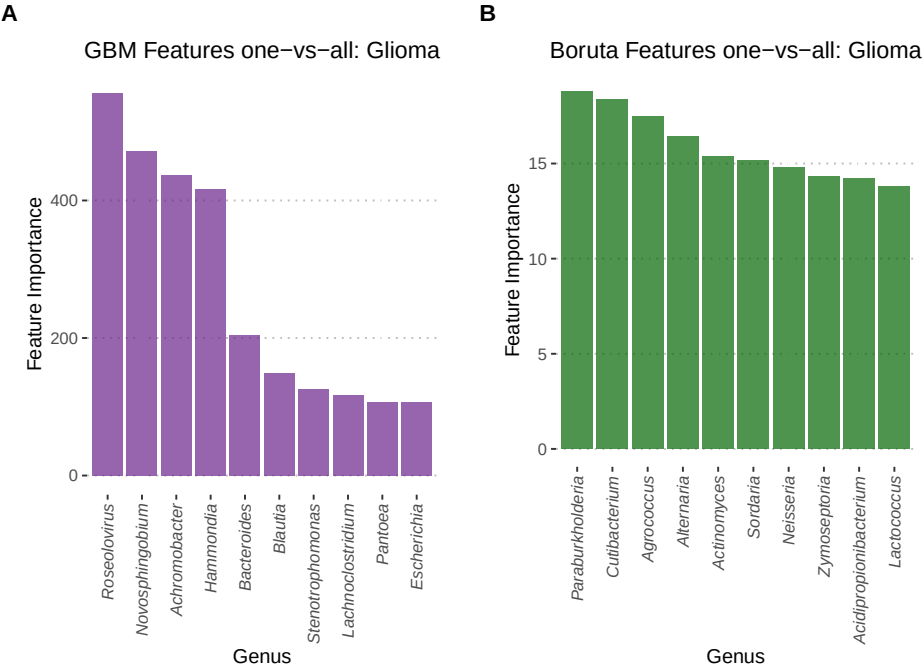


Figure D.9: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.

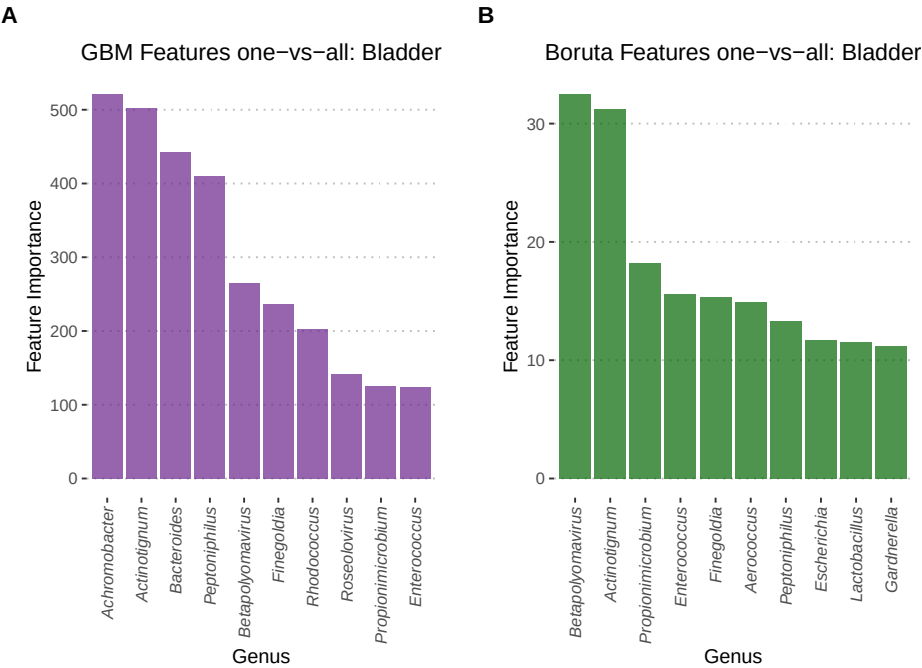


Figure D.10: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.

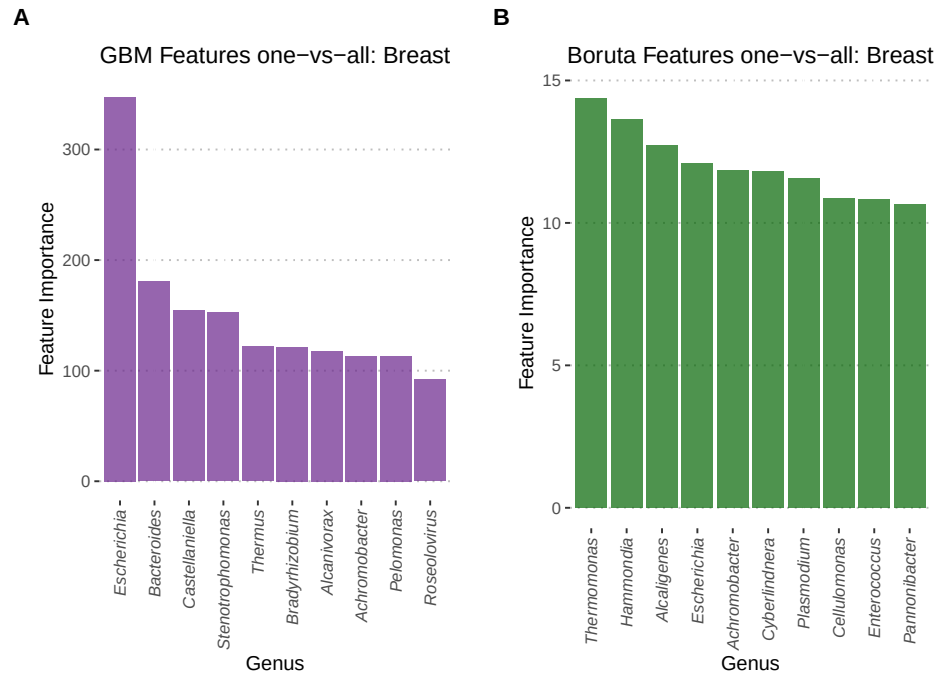


Figure D.11: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.

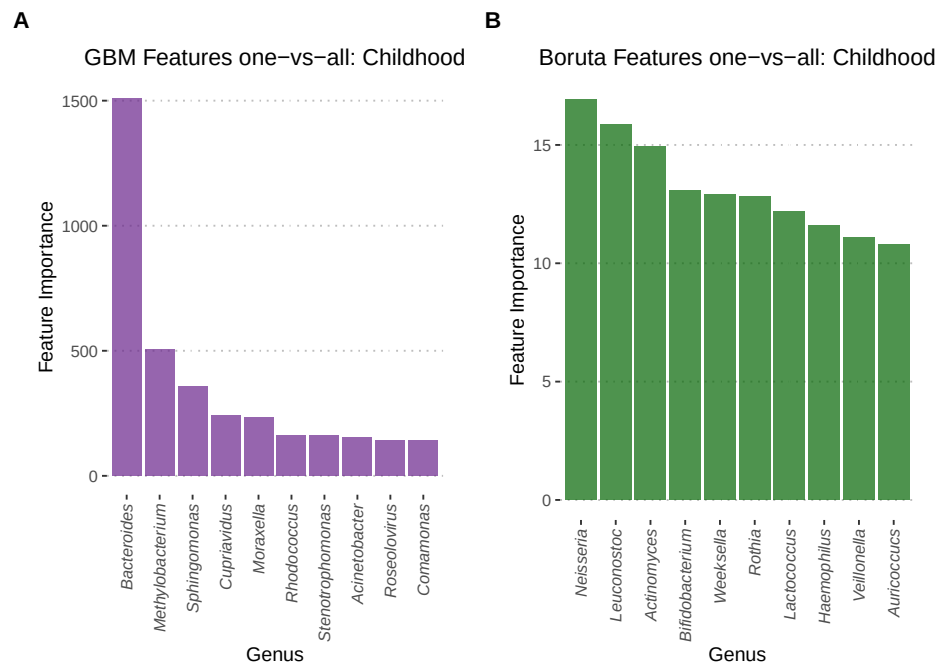


Figure D.12: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.

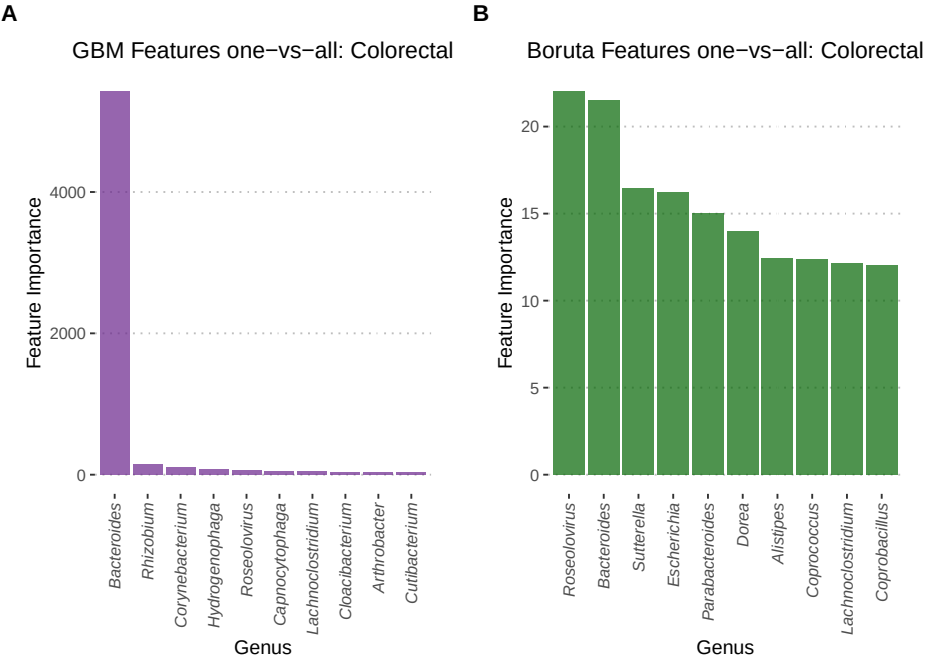


Figure D.13: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.

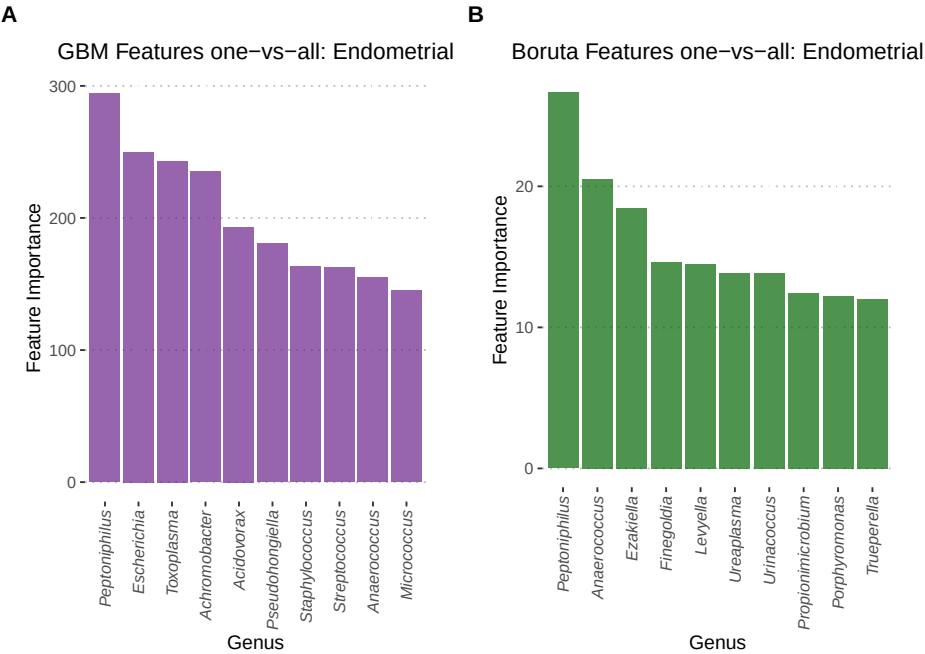


Figure D.14: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.

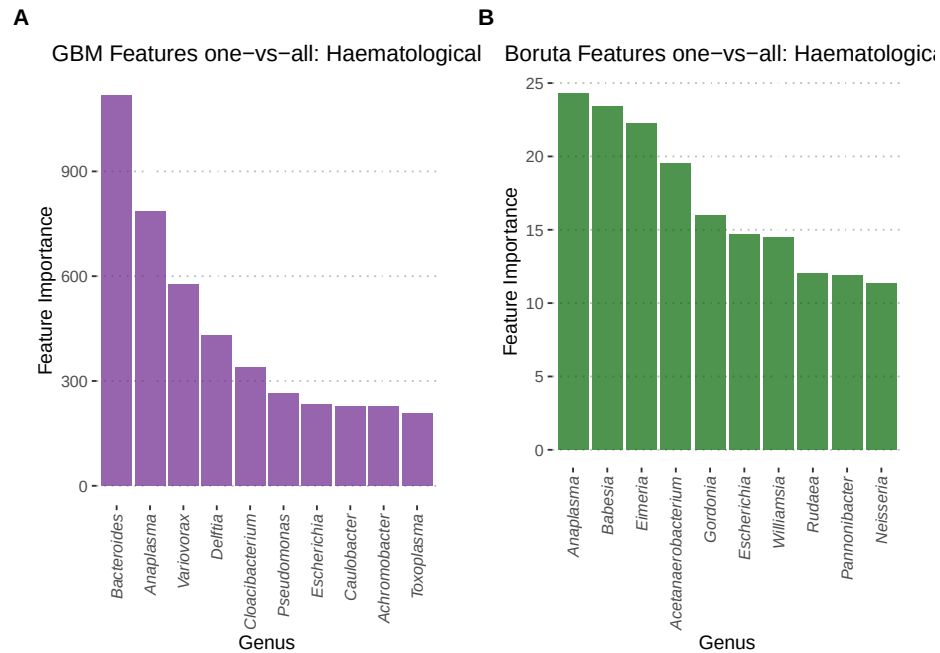


Figure D.15: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.

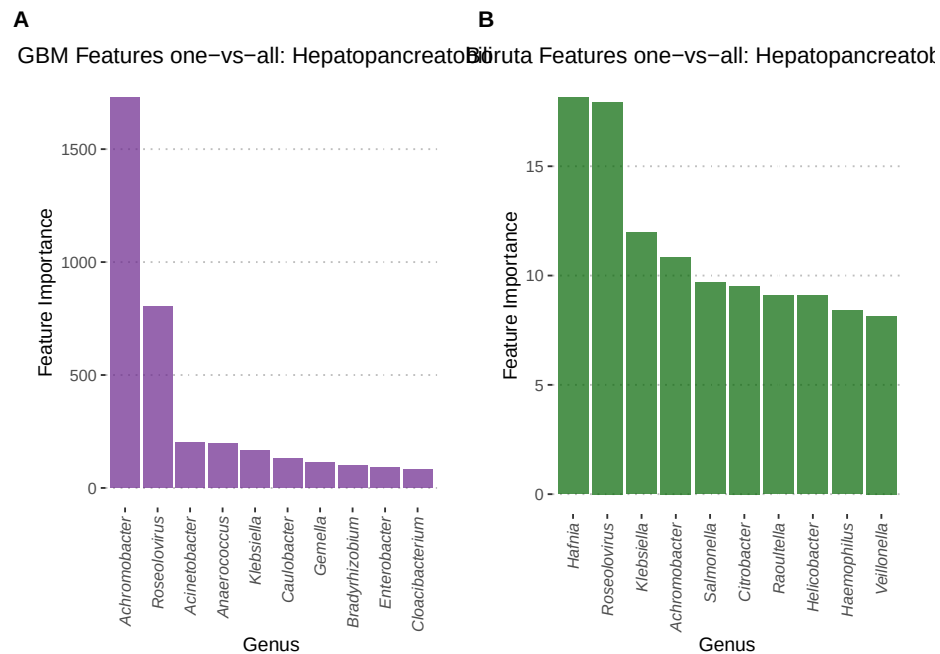


Figure D.16: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.

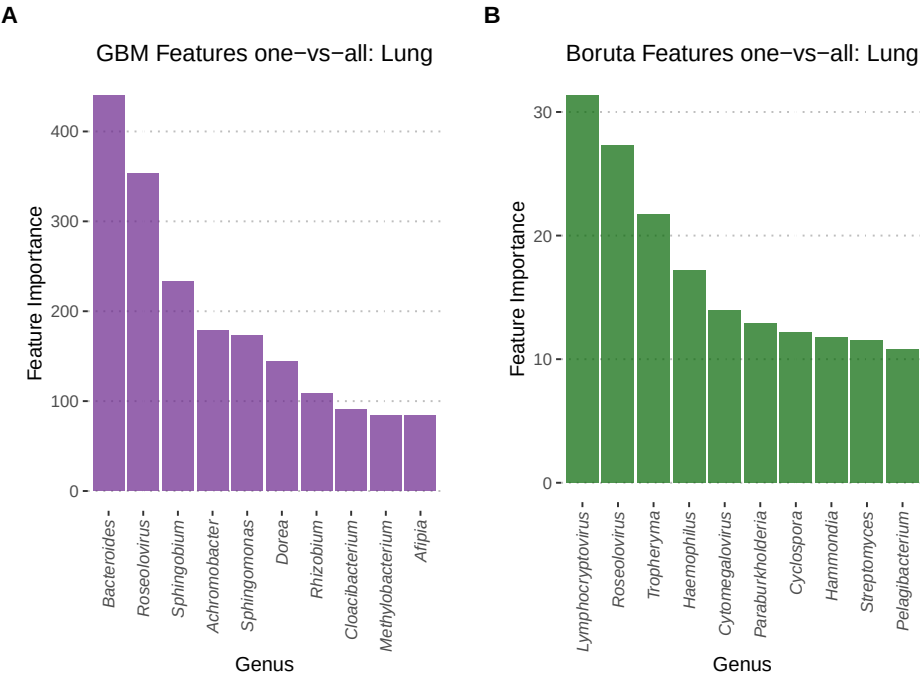


Figure D.17: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.

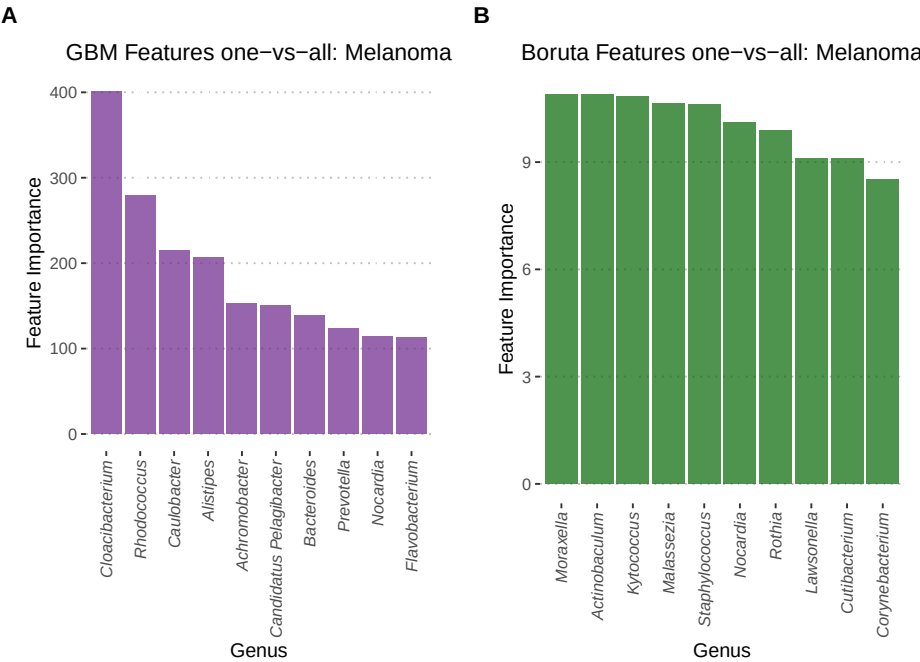


Figure D.18: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.

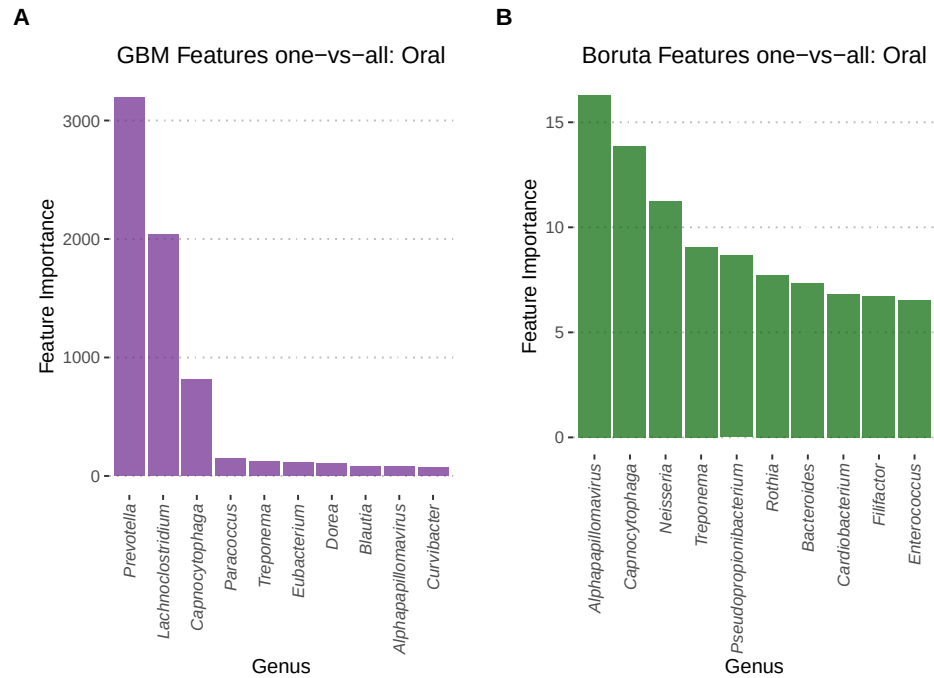


Figure D.19: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.

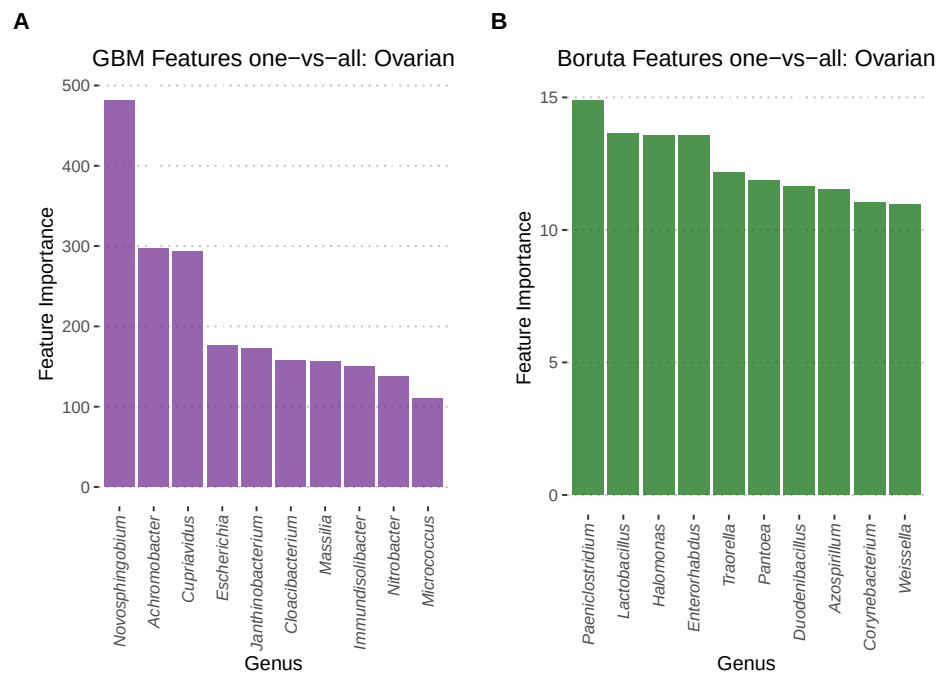


Figure D.20: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.

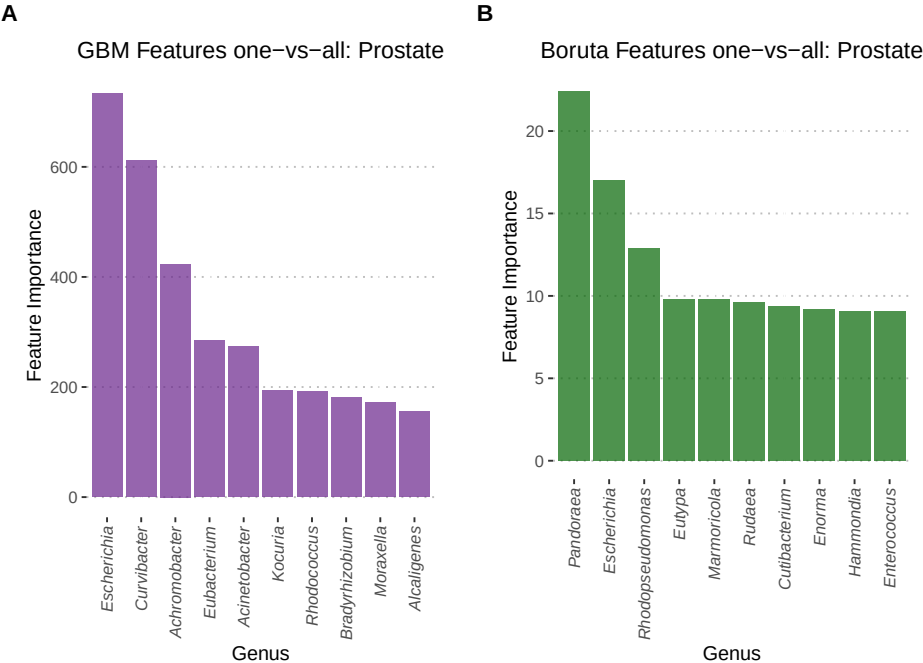


Figure D.21: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.

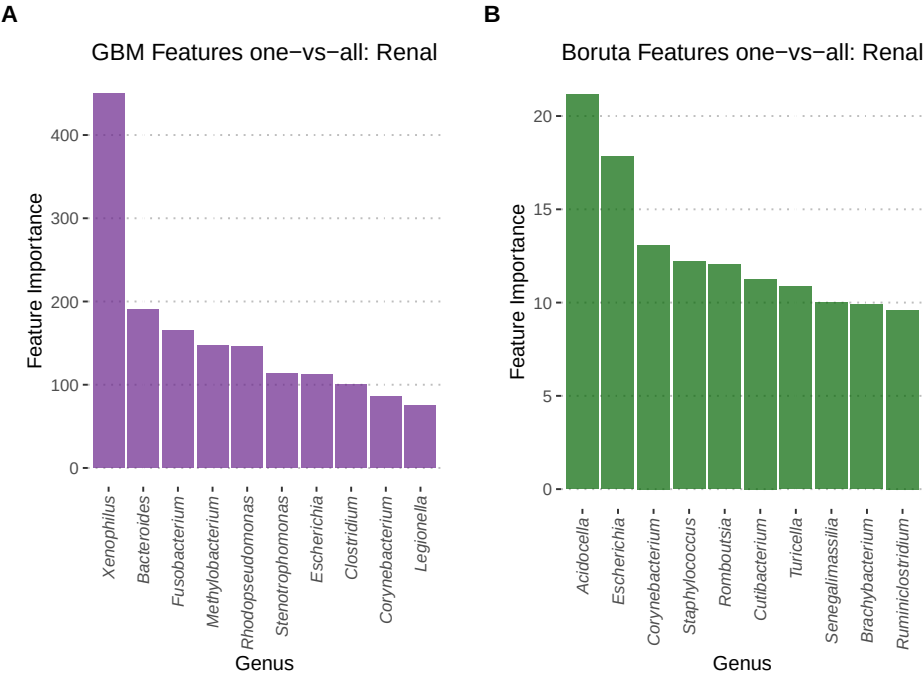


Figure D.22: Feature Importance for GBM and Boruta of one-vs-all tumour classification models.


```

      'ENDOMETRIAL_CARCINOMA', 'MALIGNANT_MELANOMA',
      'RENAL', 'BREAST', 'HAEMONC',
      'ORAL_OROPHARYNGEAL', 'SARCOMA',
      'HEPATOPANCREATOBILIARY', 'OVARIAN'))

kraken %<>% select(c(Genus, intersect(colnames(kraken),
  cancer_analysis$`Tumour Sample Platekey`)))
kraken$taxa_totals <- rowSums(kraken[2:ncol(kraken)])
kraken %<>% filter(taxa_totals > 100)
kraken %<>% select(-taxa_totals)

#remove common contaminants
kraken %<>% filter(!Genus %in% c('Homo', 'Acinetobacter', 'Alcaligenes',
  'Bacillus', 'Bradyrhizobium', 'Herbaspirillum', 'Legionella',
  'Leifsonia', 'Mesorhizobium', 'Methylobacterium', 'Microbacterium',
  'Novosphingobium', 'Pseudomonas', 'Ralstonia',
  'Sphingomonas', 'Stenotrophomonas', 'Xanthomonas', 'Sphingobium',
  'V5virus', 'K1gvirus', 'Cc31virus', 'Kp15virus', 'T4virus',
  'Felixovirus', 'Afipia', 'Aquabacterium', 'Asticcacaulis',
  'Aurantimonas', 'Beijerinckia', 'Bosea', 'Bradyrhizobium',
  'Brevundimonas', 'Caulobacter', 'Craurococcus', 'Devosia',
  'Hoeflea', 'Mesorhizobium', 'Methylobacterium', 'Novosphingobium',
  'Ochrobactrum', 'Paracoccus', 'Pedomicrobium', 'Phyllobacterium',
  'Rhizobium', 'Roseomonas', 'Sphingobium', 'Sphingomonas', 'Sphingopyxis',
  'Acidovorax', 'Azoarcus', 'Azospira', 'Burkholderia', 'Comamonas',
  'Cupriavidus', 'Curvibacter', 'Delftia', 'Duganella',
  'Herbaspirillum', 'Janthinobacterium', 'Kingella', 'Leptothrix',
  'Limnobacter', 'Massilia', 'Methylophilus', 'Methyloversatilis',
  'Oxalobacter', 'Pelomonas', 'Polaromonas', 'Ralstonia', 'Schlegelella',
  'Sulfuritalea', 'Undibacterium', 'Variovorax', 'Acinetobacter',
  'Enhydrobacter', 'Enterobacter', 'Escherichia', 'Nevskiae',
  'Pseudomonas', 'Pseudoxanthomonas', 'Psychrobacter',
  'Stenotrophomonas', 'Xanthomonas', 'Aeromicrobium',
  'Arthrobacter', 'Beutenbergia', 'Brevibacterium',
  'Corynebacterium', 'Curtobacterium', 'Dietzia',
  'Geodermatophilus', 'Janibacter', 'Kocuria',
  'Microbacterium', 'Micrococcus', 'Microlunatus',
  'Patulibacter', 'Propionibacterium',
  'Rhodococcus', 'Tsukamurella', 'Abiotrophia', 'Bacillus',
  'Brevibacillus', 'Brochothrix', 'Facklamia', 'Paenibacillus',
  'Streptococcus',
  'Chryseobacterium', 'Dyadobacter', 'Flavobacterium', 'Hydrotalea',
  'Niastella', 'Olivibacter', 'Pedobacter', 'Wautersiella',
  'Deinococcus',
  'Acidobacteria', 'Toxoplasma', 'Mycobacterium',
  'Candidatus Pelagibacter'))

```

```

#load in GBM importance on non-voom normalised data
top_most_imp <- 1500
imp <- read_tsv('ADULT_GLIOMA_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)
imp2 <- read_tsv('COLORECTAL_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)
imp3 <- read_tsv('LUNG_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)
imp4 <- read_tsv('PROSTATE_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)
imp5 <- read_tsv('BLADDER_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)
imp6 <- read_tsv('ENDOMETRIAL_CARCINOMA_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)
imp7 <- read_tsv('MALIGNANT_MELANOMA_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)
imp8 <- read_tsv('RENAL_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)
imp9 <- read_tsv('BREAST_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)
imp10 <- read_tsv('HAEMONC_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)
imp11 <- read_tsv('ORAL_OROPHARYNGEAL_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)
imp12 <- read_tsv('SARCOMA_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)
imp13 <- read_tsv('HEPATOPANCREATOBILIARY_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)
imp14 <- read_tsv('OVARIAN_raw_imp.tsv') %>%
  arrange(-Overall) %>% head(top_most_imp)

imp_taxa <- unique(c(imp$variable, imp2$variable,
                    imp3$variable, imp4$variable,
                    imp5$variable, imp6$variable,
                    imp7$variable, imp8$variable,
                    imp9$variable, imp10$variable,
                    imp11$variable, imp12$variable,
                    imp13$variable, imp14$variable))

rm(imp, imp2, imp3, imp4, imp5, imp6, imp7,
    imp8, imp9, imp10, imp11, imp12, imp13, imp14)

kraken %<>% filter(Genus %in% imp_taxa)

kraken %<>% column_to_rownames('Genus')

```

```

kraken <- as.data.frame(t(kraken))

# remove taxa with no positive samples
kraken_pa <- decostand(kraken, method='pa')

kraken$num_pos <- rowSums(kraken_pa)
kraken %<>% rownames_to_column('tumour_sample_platekey')
kraken %<>% filter(num_pos > 0) %>% select(-num_pos)
kraken %<>% column_to_rownames('tumour_sample_platekey')

krak_dist <- distanceMatrix(
  as.data.frame(t(kraken)), metric='spearman')
cancer_analysis %<>%
  filter(`Tumour Sample Platekey` %in% rownames(kraken))
cancer_analysis$tumour_sample_platekey <- \
  cancer_analysis$`Tumour Sample Platekey`

tsne <- Rtsne(krak_dist, dims = 2, perplexity=80,
  verbose=TRUE, max_iter = 2000, check_duplicates=TRUE)

t_sne <- as.data.frame(tsne[["Y"]])
t_sne$tumour_sample_platekey <- rownames(kraken)

tsne_data <- merge(t_sne, cancer_analysis,
  by='tumour_sample_platekey', all=FALSE)

tsneplot <- ggplot(tsne_data, aes(x=V1, y=V2, col=`Disease Type`)) +
  geom_point(alpha=0.7) +
  theme_minimal() +
  labs(x='t-SNE 1',
    y='t-SNE 2',
    title= 'Kraken t-SNE',
    subtitle=paste0('n=', nrow(kraken),
      ' - Number of Taxa: ', ncol(kraken) )) +
  theme(plot.title=element_text(hjust=0.5),
    plot.subtitle=element_text(hjust=0.5, face='italic'),
    panel.grid.major = element_blank(),
    axis.ticks = element_blank(),
    axis.text=element_blank(),
    legend.position='top')

```

D.11 Kraken Community Structure

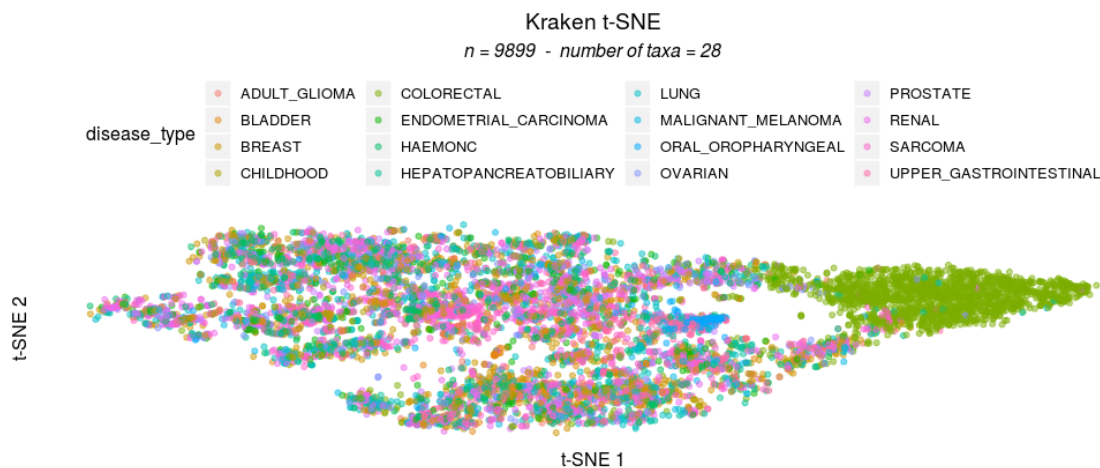


Figure D.24: t -SNE plot of kraken results coloured by tumour type for only the microbial genera that are most informative of tumour type. t -SNE was carried out on a matrix of Spearman's correlation values between samples.

D.12 Searching for Contaminants

This section of the appendix further describes the heuristic approach to identifying common contaminant genera in Genomics England's 100,000 Genomes Project.

Boruta determined 417 genera were informative in predicting technical variables which should have no association with taxa (*i.e.* tissue source, library type, illumina QC status, genome build, preparation method and sample handling GMC trust). The top twenty genera reported were: *Enterococcus*, *Curvibacter*, *Methylobacterium*, *Sphingomonas*, *Variovorax*, *Actinomyces*, *Bradyrhizobium*, *Burkholderia*, *Cutibacterium*, *Hammondia*, *Kocuria*, *Micrococcus*, *Ralstonia*, *Sphingobium*, *Streptococcus*, *Cloacibacterium*, *Pseudomonas*, *Afpia*, *Bacteroides*, *Caulobacter*, *Confluentimicrobium* (full lists available in appendix D.5). Of these 417 genera, 58 out of 66 of contaminant genera evaluated by Eisenhofer *et al.* 2019¹⁸² (which encompasses the Salter list¹⁷²). The eight taxa in Eisenhofer that are not represented by this list are: *Atopobium*, *Sediminibacterium*, *Unclassifiedtm7*, *Brevibacillus*, *Megasphaera*, *Devosiamethylobacterium*, *Phyllobacterium*, *Leptothrix*.

Potential contaminants were ranked according to two lists: the number of citations in a recent meta-analysis of contamination by Eisenhofer *et al.* 2019¹⁸² and the number of times a feature (genus) was confirmed 'important' by Boruta. I hypothesised that incrementally removing contaminant genera would improve the resolution of microbial community structure between tumour types.

Varying levels of both the Boruta contaminants and the Eisenhofer contaminants were removed from Kraken community matrices and PCoA was carried out using the Jaccard distance with different read thresholds applied to the Kraken data to limit false positive results. Clusters were determined with k -means (number of centers set to the number of categories for each variable) and Cramer's V test was used to measure the association

between cluster and biological variables: sex, ethnicity and disease type. Using different Kraken read thresholds and lists of contaminants did not show significant associations with biological variables. Overall disease type showed the best association with cluster, but this was still low (maximum Cramer's V value 0.326, where 1 resembles perfect association). Overall, 245 different combinations of parameters were tested programatically using the Jaccard distance alone (appendix section D.12.1).

Taxa investigated from Eisenhofer *et al.* 2019¹⁸² are as follows:

Actinomyces, Corynebacterium, Arthobacter, Rothia, Propionibacterium, Atopobium, Sediminibacterium, Porphyromonas, Prevotella, Chyseobacterium, Capnocytophaga, Chyseobacterium Flavobacterium, Pedobacter, UnclassifiedTM7, Bacillus, Geobacillus, Brevibacillus, Paenibacillus, Staphylococcus, Abiotrophia, Granulicatella, Enterococcus, Lactobacillus, Streptococcus, Clostridium, Coprococcus, Anaerococcus, Dialister, Megaphaera, Villonella, Fusobacterium, Leptotrichia, Brevundimonas, Afipia, Bradyrhizobium, Devosia, Methylobacterium, Mesorhizobium, Phyllobacterium, Rhizobium, Methylobacterium, Phyllobacterium, Roseomonas, Novosphingobium, Sphingobium, Sphingomonas, Achromobacter, Burkholderia, Acidovorax, Comamonas, Herbaspirillum, Janthinobacterium, Massilia, Oxalobacter, Ralstonia, Leptothrix, Kingella, Neisseria, Escherichia, Haemophilus, Acinetobacter, Enhydrobacter, Pseudomonas, Stenotrophomonas, Xanthomonoas

Using Boruta, 428 genera showed some level of importance in predicting technical variables. The graph below demonstrates the number of unique genera included after filtering the Eisenhofer list and the boruta list by their top quantiles. As the Eisenhofer list is much smaller, changing the Eisenhofer quantile has a much smaller effect on the number of genera than adjusting the boruta quantile.

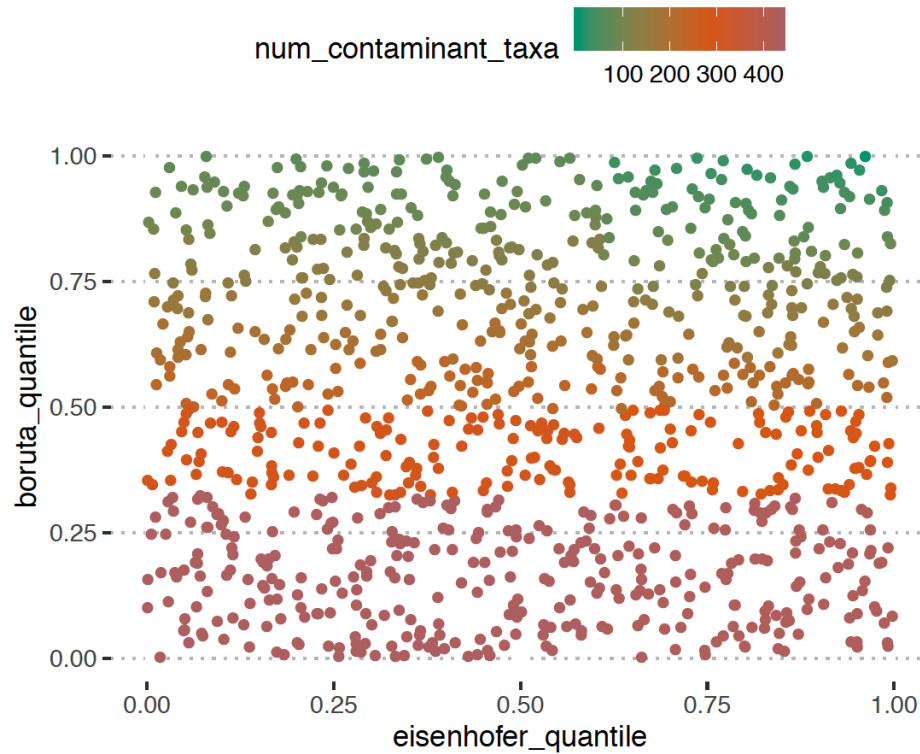


Figure D.25: The number of microbial genera determined as contaminants after filtering ranked eisenhofer and boruta lists for their top quantiles of contaminants. The colour of the points resembles the number of unique genera as indicated in the legend.

D.12.1 CramersV results

The graph below shows crammers V values for association between generated k -means cluster and the variable of choice. Overall disease type shows the strongest association compared with sex and ethnicity (figure D.26). The highest crammers value observed

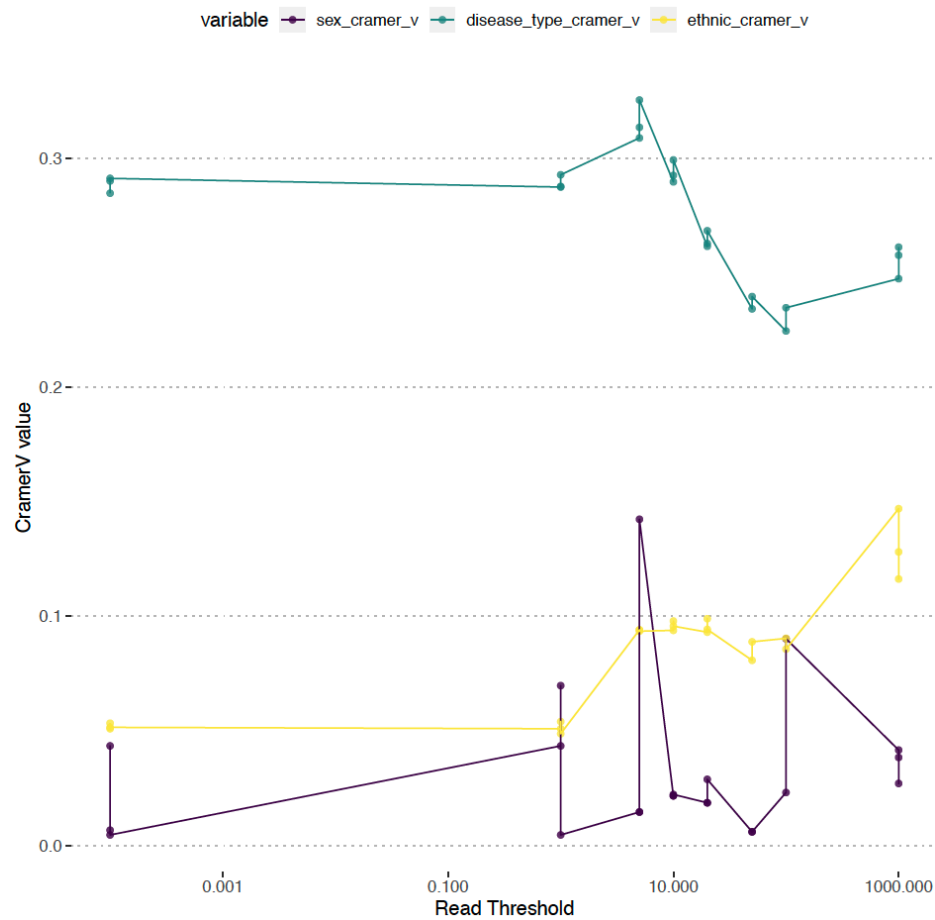


Figure D.26: A sample of the results from searching for structure by disease type, sex and ethnicity. This particular result includes the top 30 percent of estimated contaminants removed and all of contaminants in Eisenhofer et al were removed. Different kraken read thresholds were applied. This shows the parameters with the highest estimated performance for disease type.

[1] "Sex Cramers V Summary:"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0002076	0.0231232	0.1183232	0.0925136	0.1505014	0.1712450

[1] "Ethnicity Cramers V Summary:"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.04756	0.07136	0.08277	0.08533	0.09900	0.14703

[1] "Disease Type Cramers V Summary:"

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.2235	0.2577	0.2817	0.2773	0.2980	0.3255

D.13 100,000 Genomes and SEPATH - Negative Findings

This section describes some of the more interesting negative findings from running SEPATH (with Kraken) on the tumour samples of the 100,000 Genomes project.

Relationships with Mutational Burden

The number of microbial reads classified with Kraken was investigated in relation to the tumour mutational burden. The only tumour type that demonstrated an interesting pattern was colorectal. This split was based on the mutational burden and not associated with the total number of classified reads. It is suspected that this split more likely denotes the difference between left-side and right-side tumours which are known to contain a higher mutational burden. Investigating this further would overlap with researchers with projects registered specifically for investigating pathogens in colorectal samples.



Figure D.27: Results from running Kraken on Tumour Samples from the 100,000 Genomes Project. The y -axis denotes the total number of classified reads and the x -axis denotes the mutational burden in terms of somatic coding variants per megabase (mb).

D.14 ABBS Genera in Genomics England

The full sample sizes and distributions of ABBS genera across different clinical stages and different cancer types is shown below. The y -axis indicates the number of taxa reads + 0.01 on a log scale. The clinical stage is indicated on the x -axis. The total number of participants within each clinical stage is also demonstrated on each x -axis.

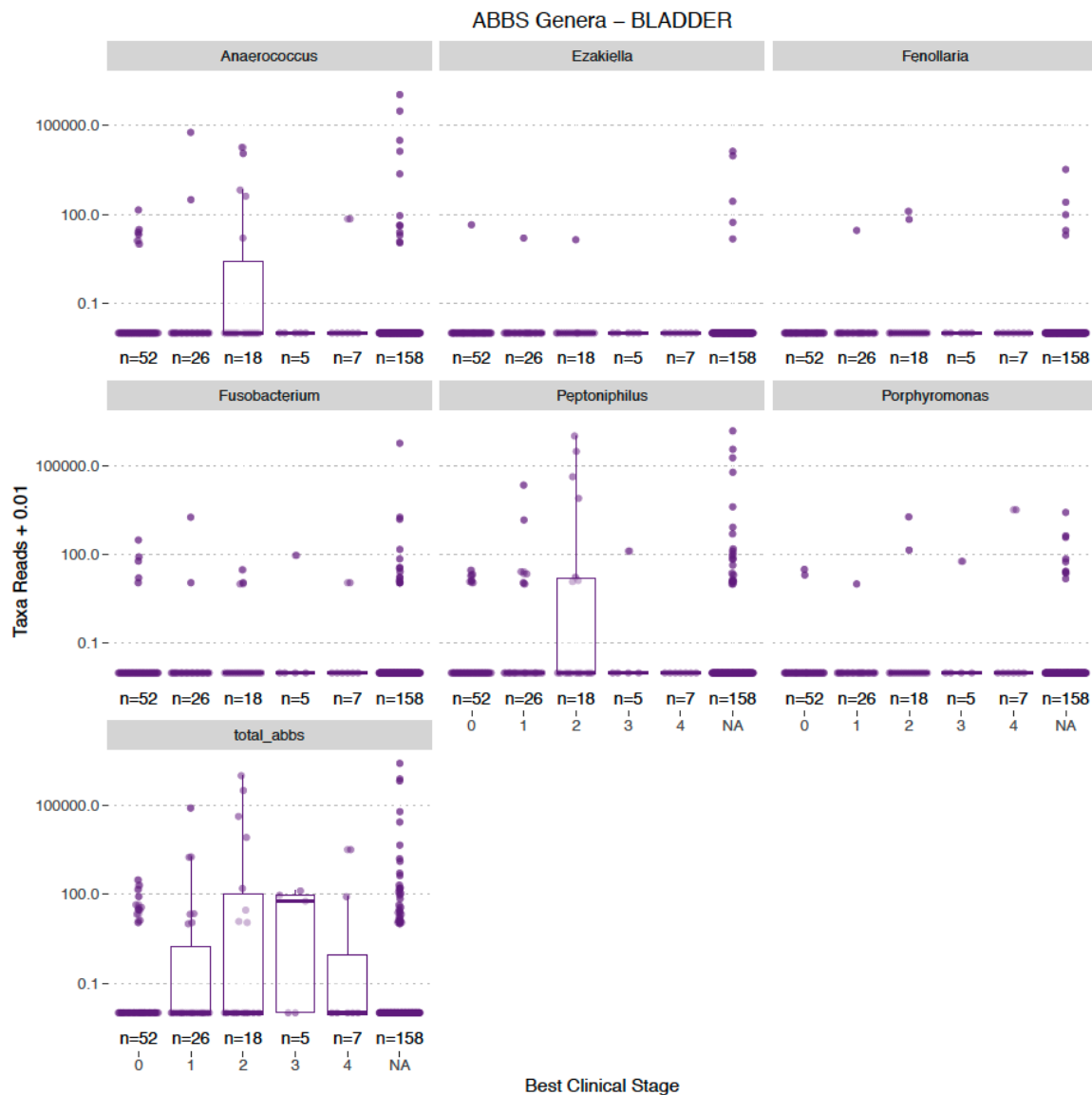


Figure D.28: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlaid on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

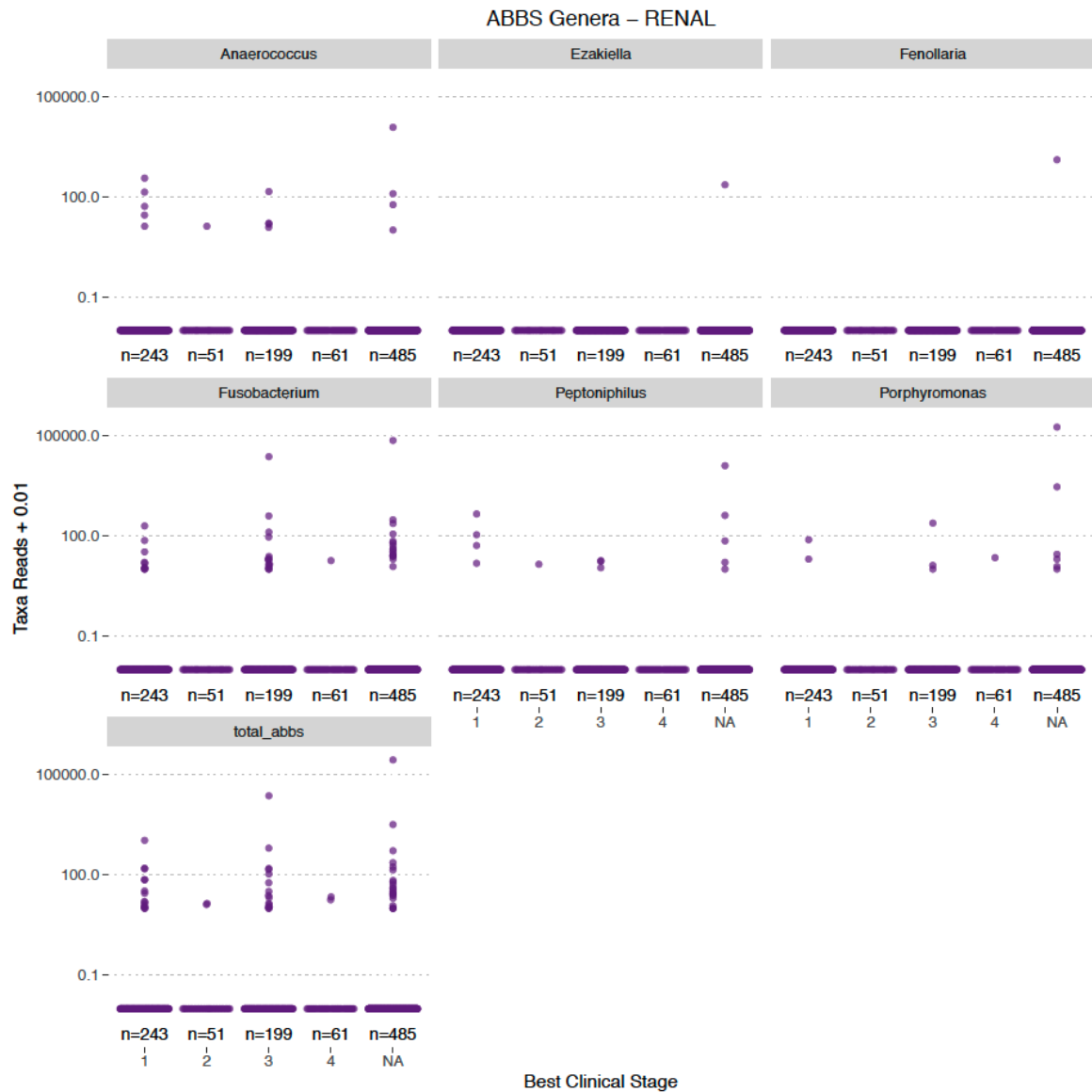


Figure D.29: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlaid on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

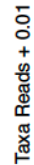


Figure D.30: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

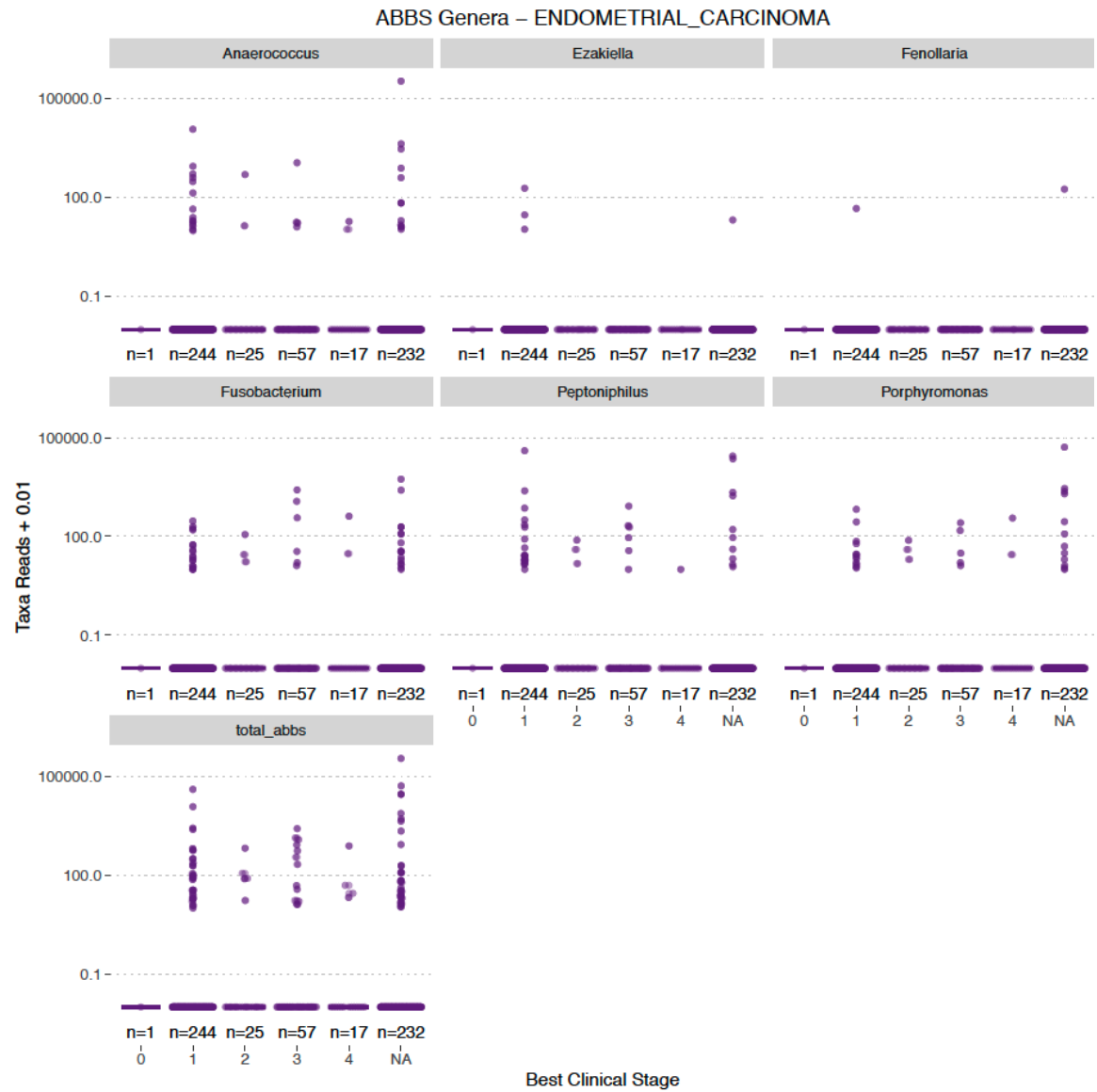


Figure D.31: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlaid on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

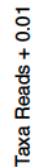


Figure D.32: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

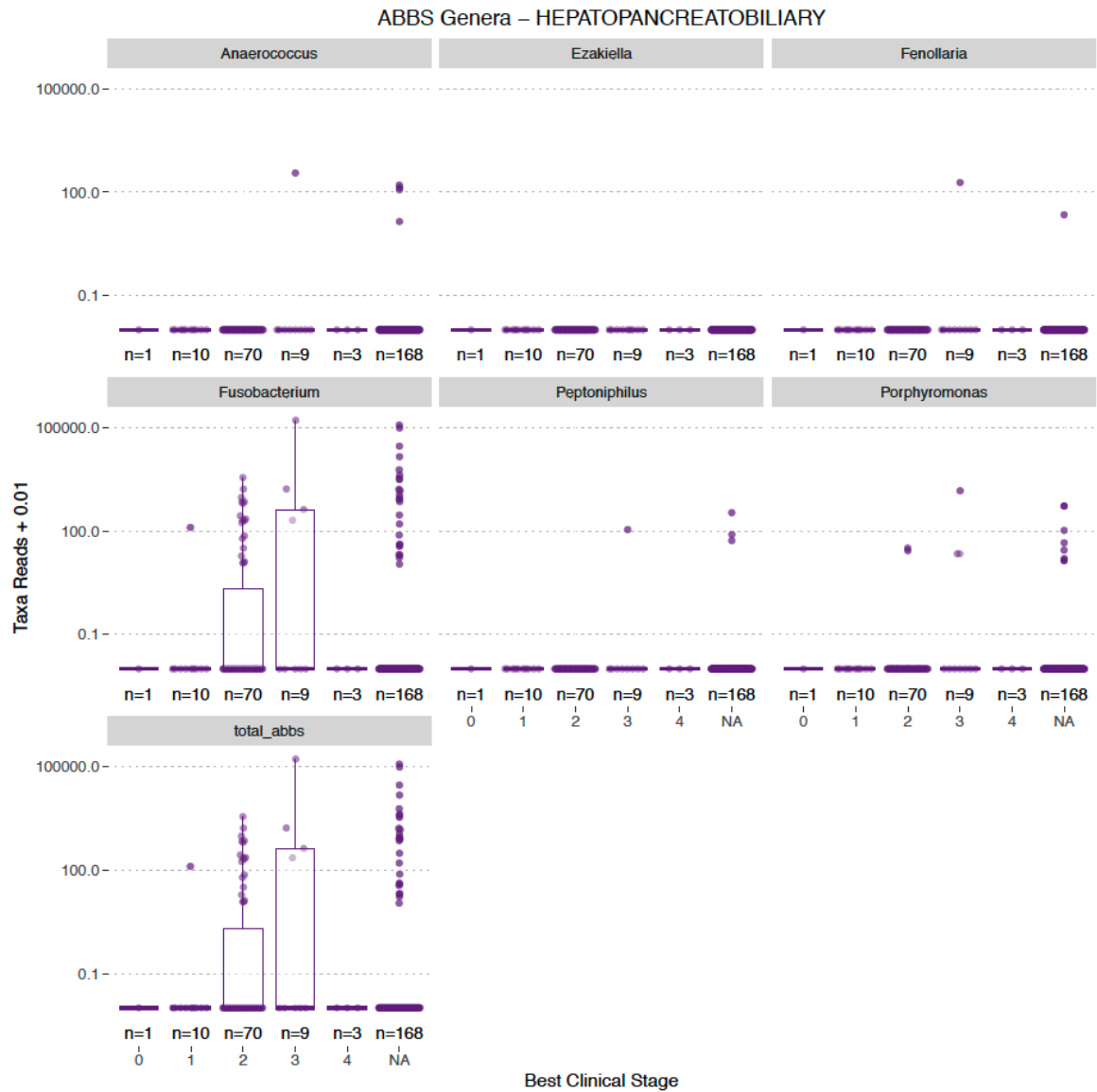


Figure D.33: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlaid on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

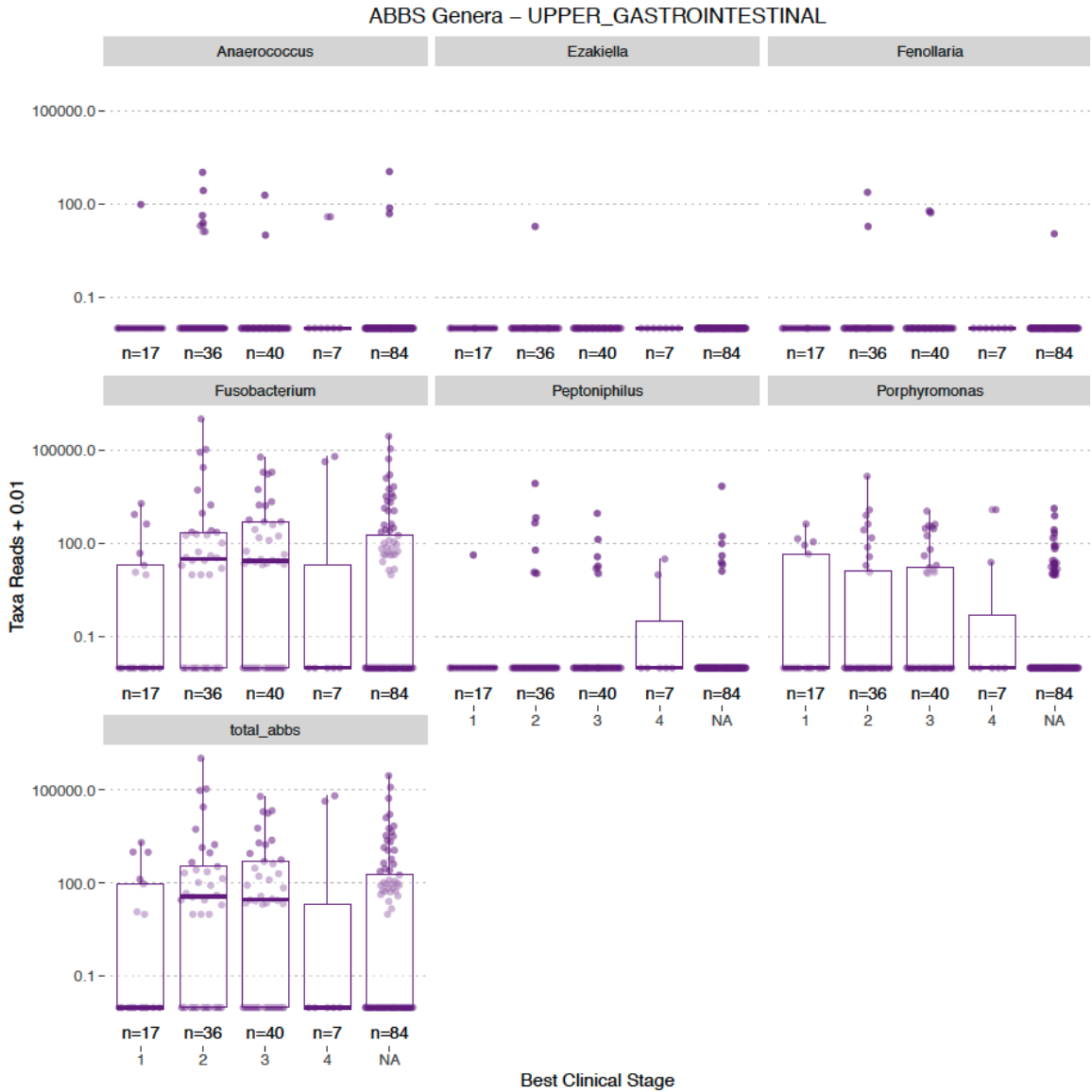


Figure D.34: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

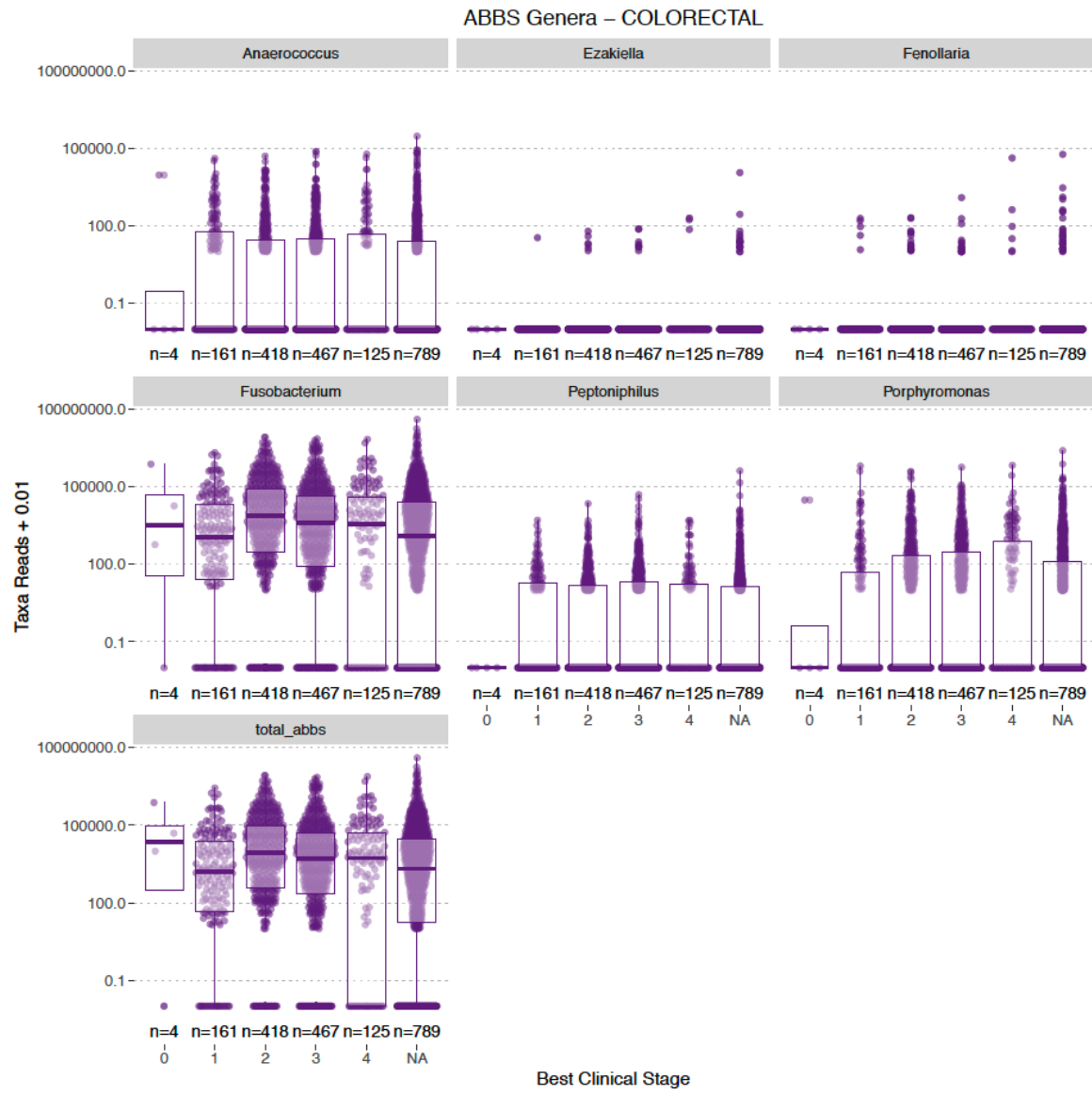


Figure D.35: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlaid on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

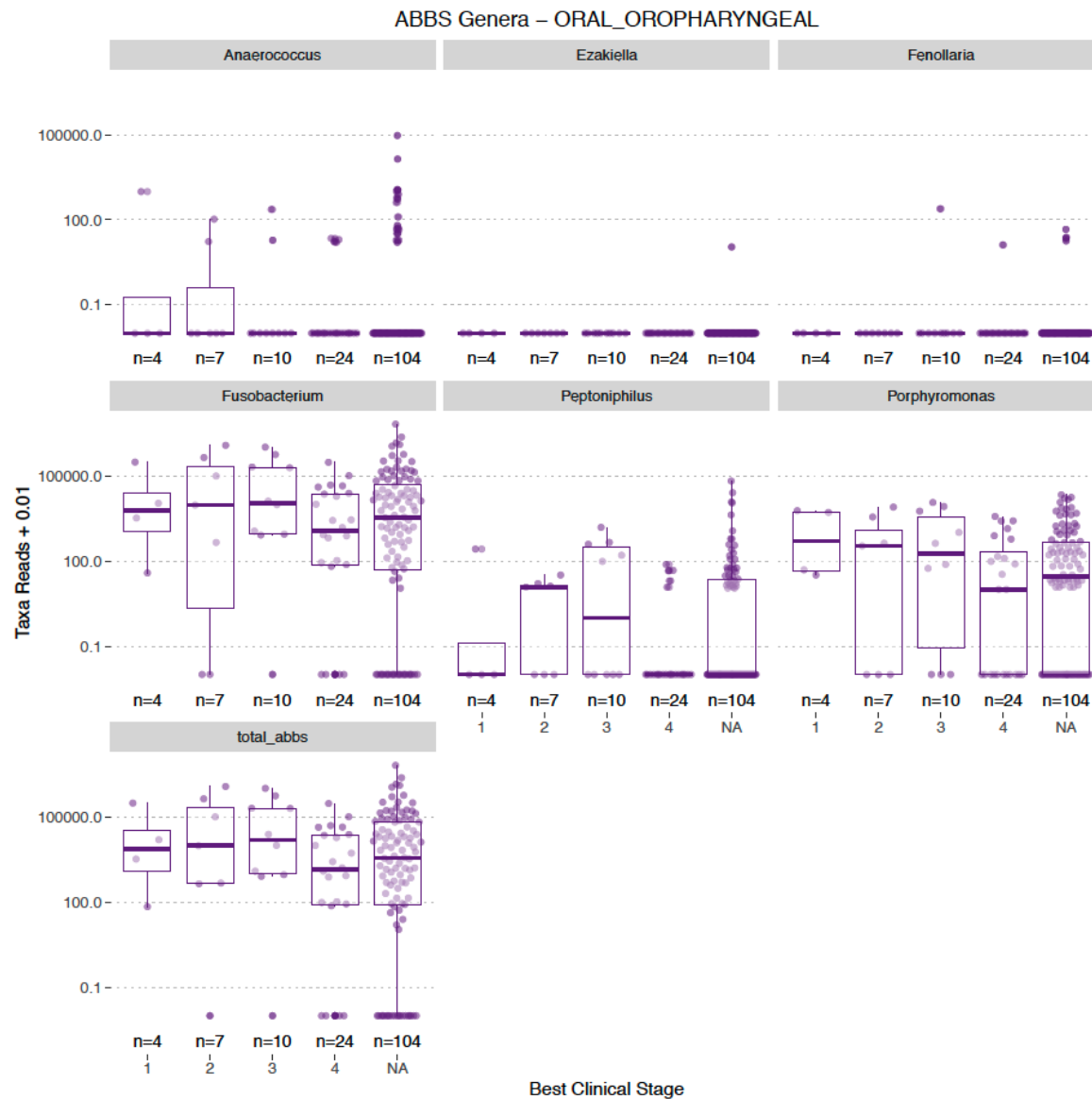


Figure D.36: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlaid on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

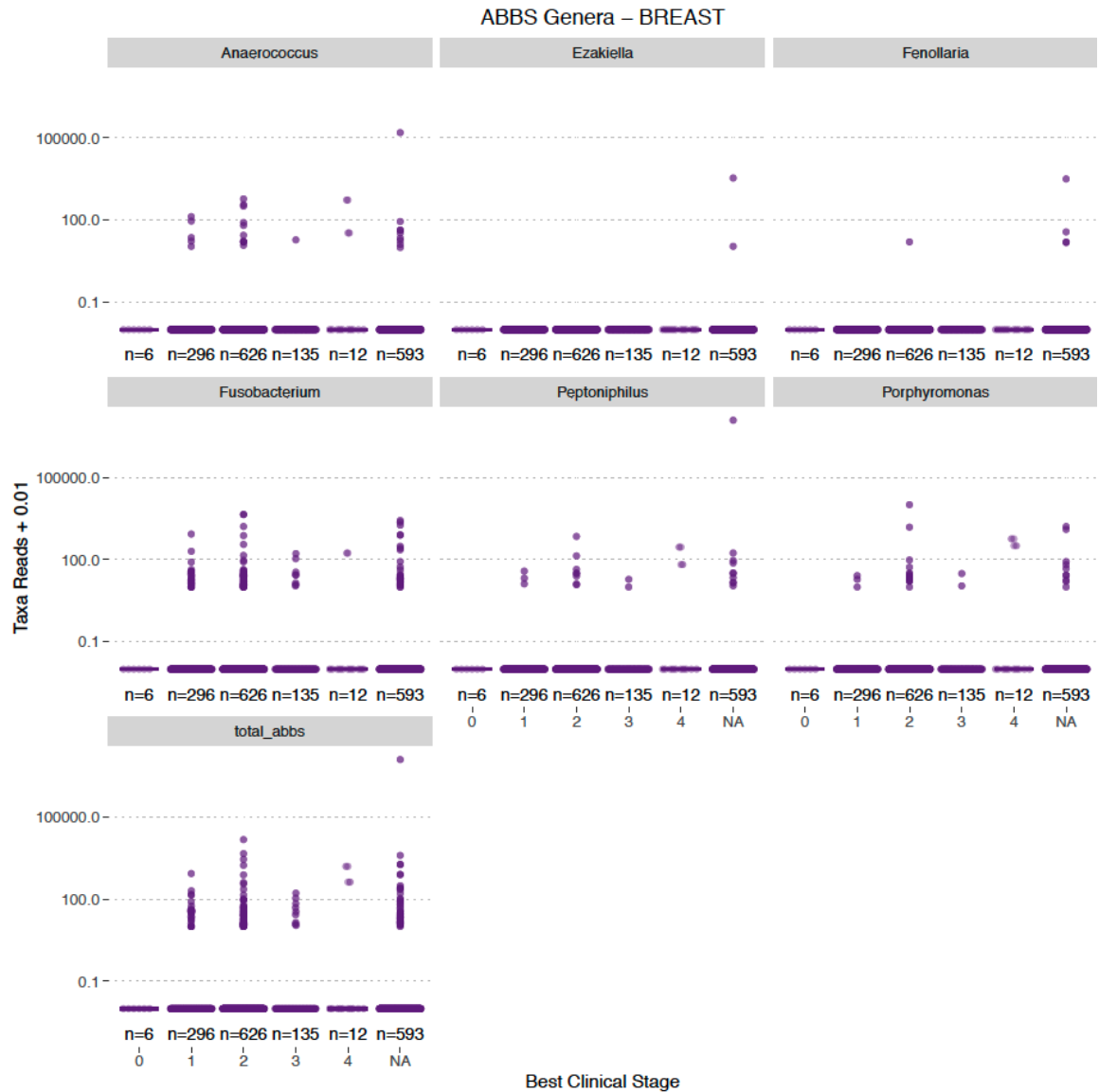


Figure D.37: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlaid on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

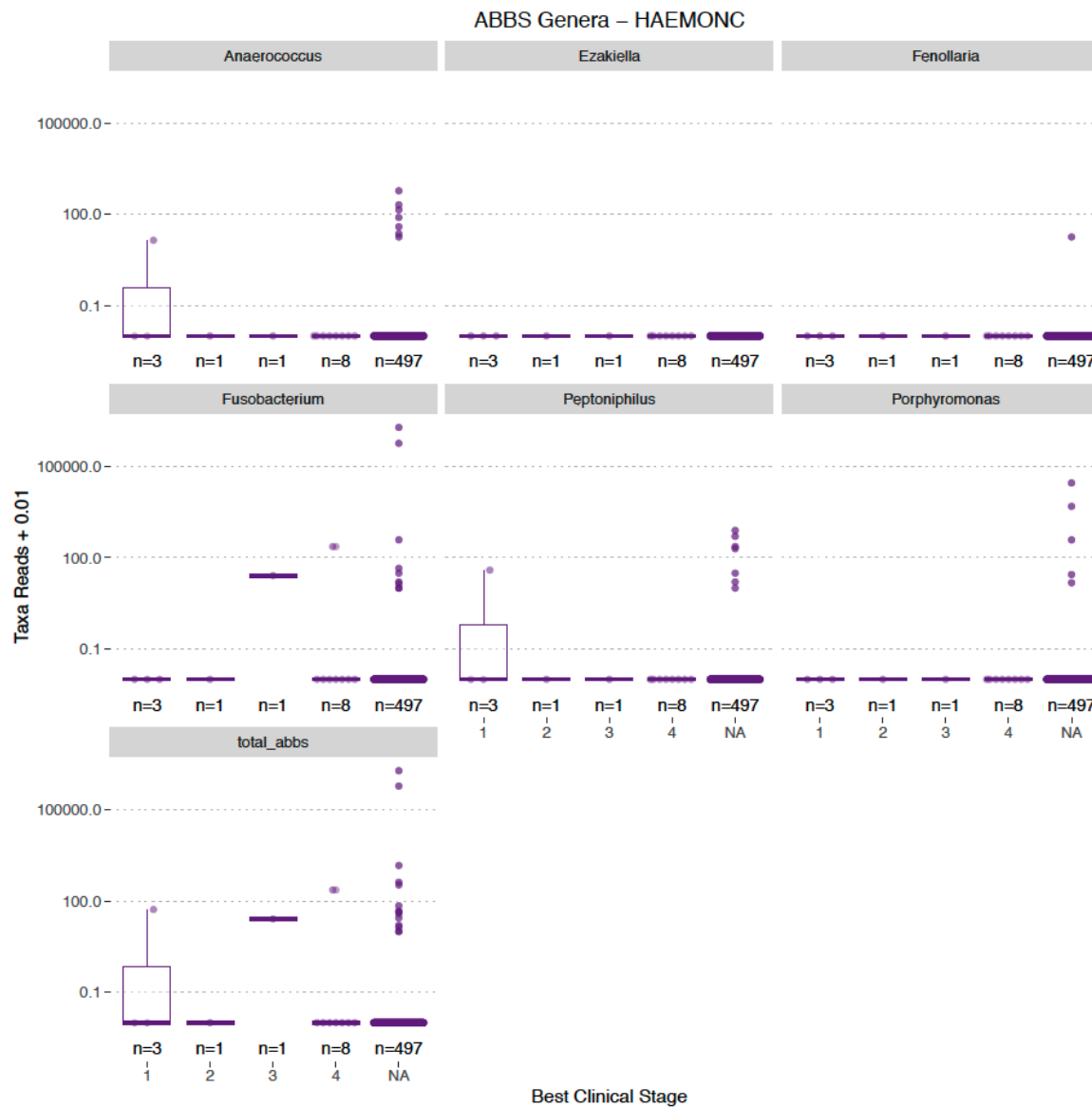


Figure D.38: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlaid on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

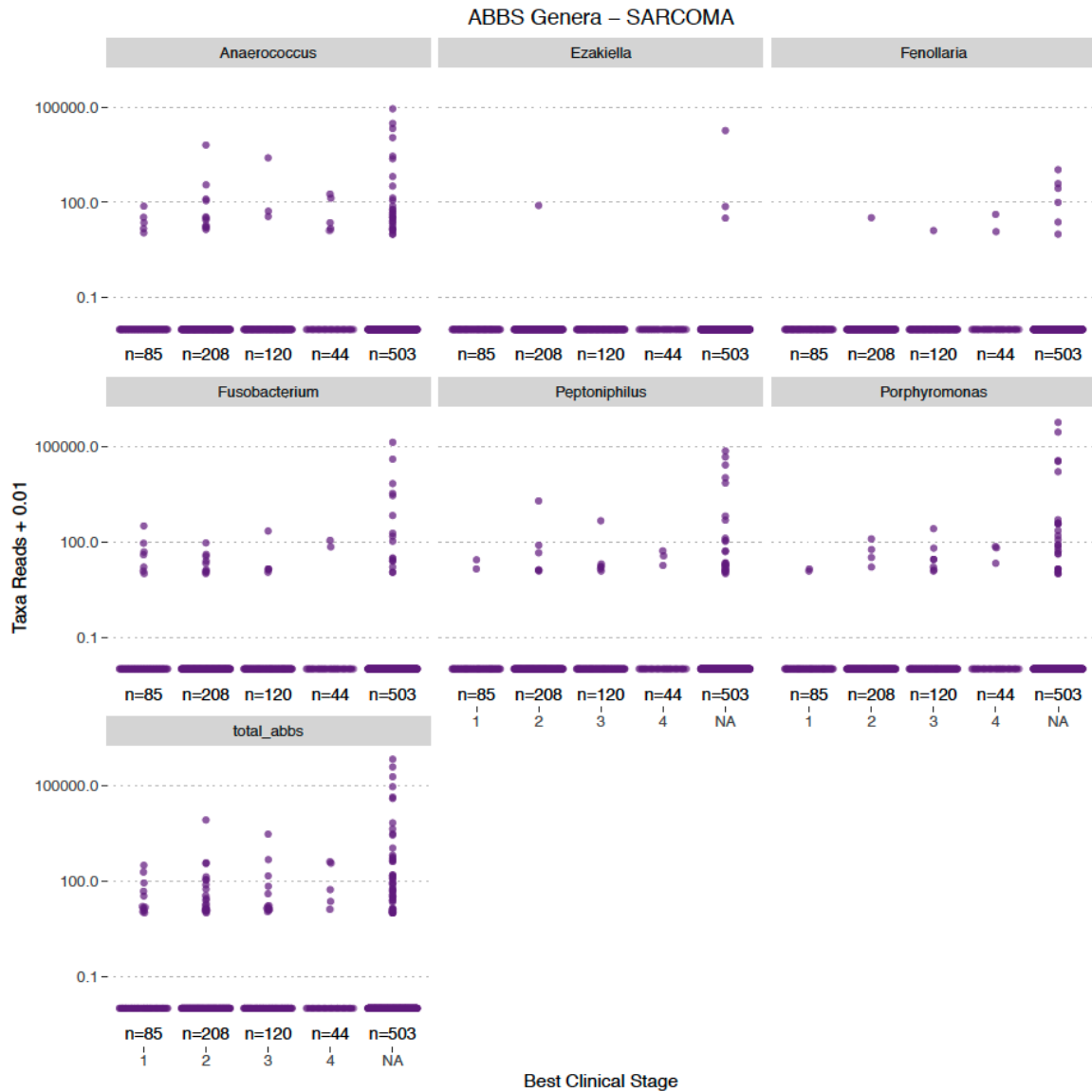


Figure D.39: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

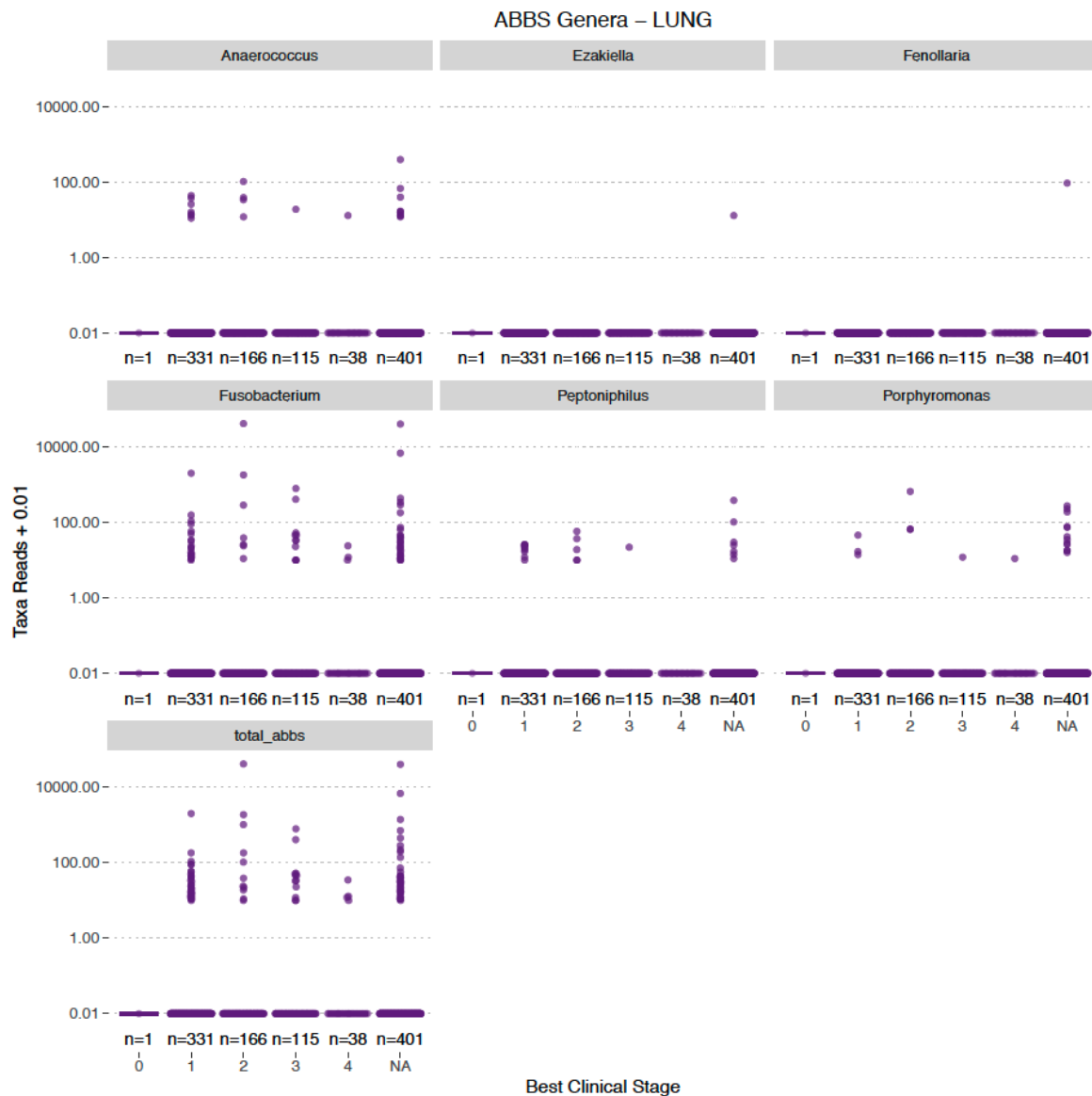


Figure D.40: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlayed on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

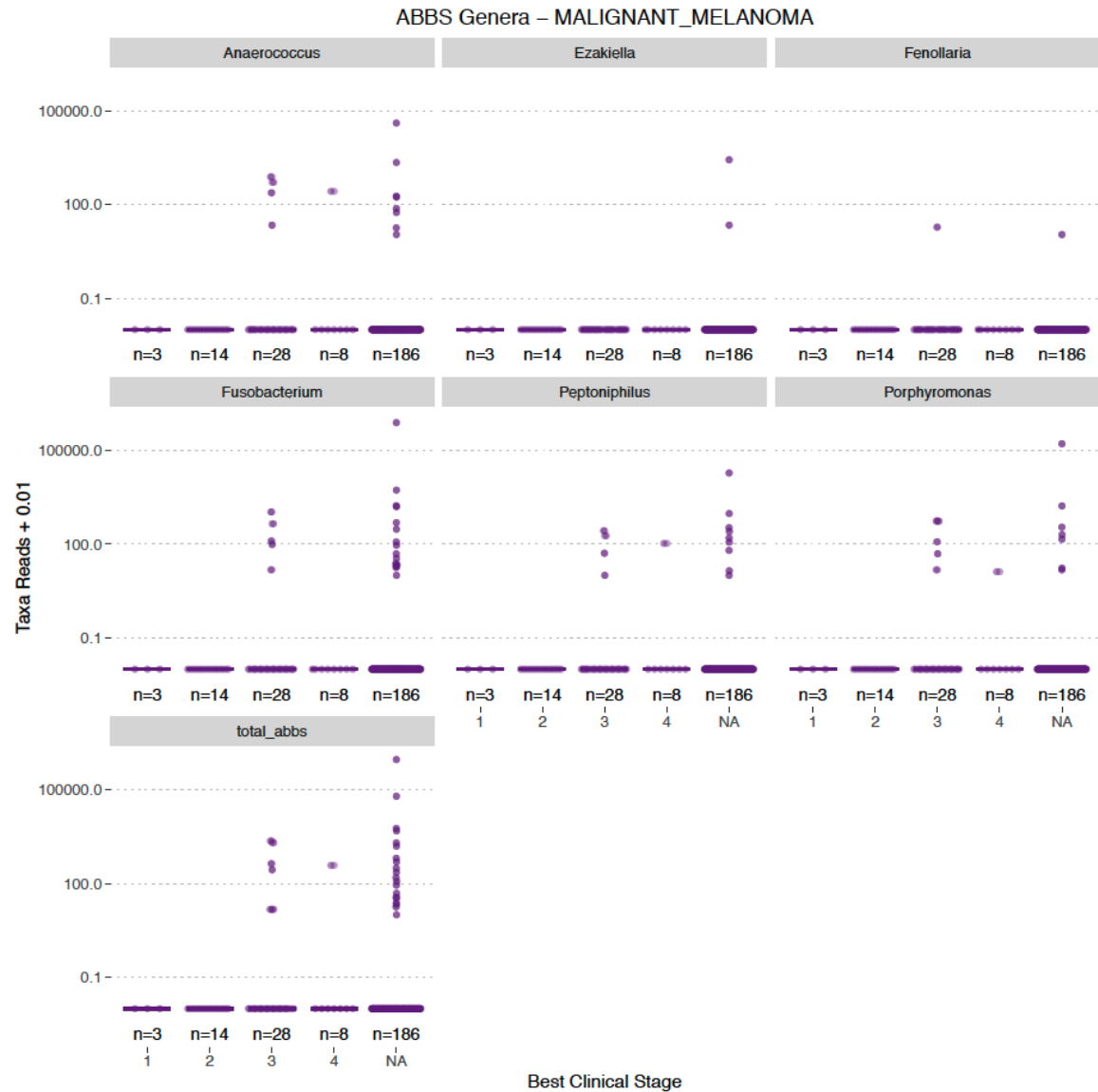


Figure D.41: The distributions of classified reads for ABBS genera from the 100,000 Genomes Project using Kraken with SEPATH. These are box plots overlaid on a beeswarm plot. If no box is visible then it means that the median and the first and third quartiles do not show enough variance to be visible.

References

1. Munoz N, Castellsague X, Gonzalez AB de, Gissmann L. Chapter 1: HPV in the etiology of human cancer. *Vaccine*. 2006;24 Suppl 3:S3/1-10. doi:10.1016/j.vaccine.2006.05.115
2. Song ZQ, Zhou LY. Helicobacter pylori and gastric cancer: Clinical aspects. *Chin Med J (Engl)*. 2015;128(22):3101-3105. doi:10.4103/0366-6999.169107
3. Vale N, Gouveia M, Rinaldi G, et al. The role of estradiol metabolism in urogenital schistosomiasis-induced bladder cancer. *Tumour Biology*. 2017:1-11. doi:10.1177/1010428317692247
4. Ringehan M, McKeating JA, Protzer U. Viral hepatitis and liver cancer. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2017;372(1732):20160274. doi:10.1098/rstb.2016.0274
5. Montagnana G M, Lippi. Cancer diagnostics: Current concepts and future perspectives. *Annals of Translational Medicine*. 2017;5(13):268. doi:10.21037/atm.2017.06.20
6. Weinberg RA. *The Biology of Cancer*. Garland Science, Taylor & Francis Group; 2014.
7. Ruddon R. *Cancer Biology*. Vol 4. Oxford University Press
8. Chandar N, Viselli S. *Cell and Molecular Biology*. Wolters Kluwer; 2010.
9. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature*. 2009;458(7239):719-724. doi:10.1038/nature07943
10. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*. 2000;100:57-70.
11. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. *Cell*. 2011;144(5):646-674. doi:10.1016/j.cell.2011.02.013
12. Whole Genomes Consortium IP-CA of. Pan-cancer analysis of whole genomes. *Nature*. 2020;578(7793):82-93. doi:10.1038/s41586-020-1969-6
13. WANG H, HAN H, MOUSSES S, VONHOFF D. Targeting loss-of-function mutations in tumor-suppressor genes as a strategy for development of cancer therapeutic agents. *Seminars in Oncology*. 2006;33(4):513-520. doi:10.1053/j.seminoncol.2006.04.013
14. Sudhakar A. History of cancer, ancient and modern treatment methods. *Journal of Cancer Science & Therapy*. 2009;01(02):i-iv. doi:10.4172/1948-5956.100000e2
15. Lederman M. The early history of radiotherapy: 1895-1939. *International Journal of Radiation OncologyBiologyPhysics*. 1981;7(5):639-648. doi:10.1016/0360-3016(81)90379-5

16. DeVita VT, Chu E. A history of cancer chemotherapy. *Cancer Research*. 2008;68(21):8643-8653. doi:10.1158/0008-5472.can-07-6611
17. WANG H, HAN H, MOUSSES S, VONHOFF D. Targeting loss-of-function mutations in tumor-suppressor genes as a strategy for development of cancer therapeutic agents. *Seminars in Oncology*. 2006;33(4):513-520. doi:10.1053/j.seminoncol.2006.04.013
18. Helleday T. The underlying mechanism for the PARP and BRCA synthetic lethality: Clearing up the misunderstandings. *Molecular Oncology*. 2011;5(4):387-393. doi:10.1016/j.molonc.2011.07.001
19. Iqbal N, Iqbal N. Imatinib: A breakthrough of targeted therapy in cancer. *Chemotherapy Research and Practice*. 2014;2014:1-9. doi:10.1155/2014/357027
20. Reichardt P. The story of imatinib in GIST - a journey through the development of a targeted therapy. *Oncology Research and Treatment*. 2018;41(7-8):472-477. doi:10.1159/000487511
21. Wagle N, Emery C, Berger MF, et al. Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. *Journal of Clinical Oncology*. 2011;29(22):3085-3096. doi:10.1200/jco.2010.33.2312
22. Chatterjee N, Bivona TG. Polytherapy and targeted cancer drug resistance. *Trends in Cancer*. 2019;5(3):170-182. doi:10.1016/j.trecan.2019.02.003
23. Bellmunt J, Powles T, Vogelzang NJ. A review on the evolution of PD-1/PD-l1 immunotherapy for bladder cancer: The future is now. *Cancer Treatment Reviews*. 2017;54:58-67. doi:10.1016/j.ctrv.2017.01.007
24. Holzinger A, Barden M, Abken H. The growing world of CAR t cell trials: A systematic review. *Cancer Immunology, Immunotherapy*. 2016;65(12):1433-1450. doi:10.1007/s00262-016-1895-5
25. Zaha DC. Significance of immunohistochemistry in breast cancer. *World Journal of Clinical Oncology*. 2014;5(3):382. doi:10.5306/wjco.v5.i3.382
26. Lanari C, Wargon V, Rojas P, Molinolo AA. Antiprogesterins in breast cancer treatment: Are we ready? *Endocrine-Related Cancer*. 2012;19(3):R35-R50. doi:10.1530/erc-11-0378
27. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*. 2003;100(14):8418-8423. doi:10.1073/pnas.0932692100
28. Narod SA, Iqbal J, Miller AB. Why have breast cancer mortality rates declined? *Journal of Cancer Policy*. 2015;5:8-17. doi:10.1016/j.jcpo.2015.03.002
29. Dubsky P, J C Brase, Jakesz R, et al. The endopredict score provides prognostic information on late distant metastases in er+/her2- breast cancer patients. *British Journal of Cancer*. 2013;109(12):2959-2964. doi:10.1038/bjc.2013.671
30. Bogaards JA, Coupe VMH, Xiridou M, Meijer CJLM, Wallinga J, Berkhof J. Long-term impact of human papillomavirus vaccination on infection rates, cervical abnormalities, and cancer incidence. *Epidemiology*. 2011;22(4):505-515. <http://www.jstor.org/stable/23047685>.

31. Crosbie EJ, Einstein MH, Franceschi S, Kitchener HC. Human papillomavirus and cervical cancer. *The Lancet*. 2013;382(9895):889-899. doi:10.1016/s0140-6736(13)60022-7
32. Lichtman M. A bacterial cause of cancer: An historical essay. *Oncologist*. 2017;22(5):542-548. doi:10.1634/theoncologist.2017-0007
33. Temin HM, Rubin H. Characteristics of an assay for rous sarcoma virus and rous sarcoma cells in tissue culture. *Virology*. 1958;6(3):669-688. doi:10.1016/0042-6822(58)90114-4
34. Zheng Z. Viral oncogenes, noncoding rnas and rna splicing in human tumor viruses. *International Journal of Biological Sciences*. 2010;6(7):730-755.
35. Plummer M, Martel C de, Vignat J, Ferlay J, Bray F, Franceschi S. Global burden of cancers attributable to infections in 2012: A synthetic analysis. *The Lancet Global Health*. 2016;4(9):e609-e616. doi:10.1016/s2214-109x(16)30143-7
36. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A*. 2016;113(21):5970-5975. doi:10.1073/pnas.1521291113
37. Pagano JS, Blaser M, Buendia MA, et al. Infectious agents and cancer: Criteria for a causal relation. *Semin Cancer Biol*. 2004;14(6):453-471. doi:10.1016/j.semcancer.2004.06.009
38. Hausen H zur. Viruses in human cancers. *European Journal of Cancer*. 1999;35(8):1174-1181. doi:10.1016/s0959-8049(99)00113-6
39. Tripathi LK, Nailwal TK. Metagenomics: Applications of functional and structural approaches and meta-omics. In: *Recent Advancements in Microbial Diversity*. Elsevier; 2020:471-505. doi:10.1016/b978-0-12-821265-3.00020-7
40. IARC. Cancer and infections. *Global Cancer Observatory*. <https://gco.iarc.fr/causes/infections/data-sources-methods>.
41. Pappas G. Infectious causes of cancer: An evolving educational saga. *Clinical Microbiology and Infection*. 2009;15(11):961-963. doi:10.1111/j.1469-0691.2009.03034.x
42. Zhu C, Wang Y, Cai C, Cai Q. Bacterial infection and associated cancers. In: *Advances in Experimental Medicine and Biology*. Springer Singapore; 2017:181-191. doi:10.1007/978-981-10-5765-6_11
43. Kajitani N, Satsuka A, Kawate A, Sakai H. Productive lifecycle of human papillomaviruses that depends upon squamous epithelial differentiation. *Frontiers in Microbiology*. 2012;3. doi:10.3389/fmicb.2012.00152
44. Narisawa-Saito M, Kiyono T. Basic mechanisms of high-risk human papillomavirus-induced carcinogenesis: Roles of e6 and e7 proteins. *Cancer Sci*. 2007;98(10):1505-1511. doi:10.1111/j.1349-7006.2007.00546.x
45. Dai M, Clifford GM, Calvez F le, et al. Human papillomavirus type 16 and TP53 mutation in oral cancer. *Cancer Research*. 2004;64(2):468-471. doi:10.1158/0008-5472.can-03-3284
46. Wang Z, Xia R-H, Ye D-X, Li J. Human papillomavirus 16 infection and TP53 mutation: Two distinct pathogeneses for oropharyngeal squamous cell carcinoma

- in an eastern chinese population. Liu X, ed. *PLOS ONE*. 2016;11(10):e0164491. doi:10.1371/journal.pone.0164491
47. Kimple AJ, Torres AD, Yang RZ, Kimple RJ. HPV-associated head and neck cancer: Molecular and nano-scale markers for prognosis and therapeutic stratification. *Sensors*. 2012;12(4):5159-5169. doi:10.3390/s120405159
48. Elinav E, Nowarski R, Thaiss CA, Hu B, Jin C, Flavell RA. Inflammation-induced cancer: Crosstalk between tumours, immune cells and microorganisms. *Nat Rev Cancer*. 2013;13(11):759-771. doi:10.1038/nrc3611
49. Kalisperati P, Spanou E, Pateras IS, et al. Inflammation, DNA damage, helicobacter pylori and gastric tumorigenesis. *Frontiers in Genetics*. 2017;8. doi:10.3389/fgene.2017.00020
50. Hatakeyama M. Oncogenic mechanisms of the helicobacter pylori CagA protein. *Nature Reviews Cancer*. 2004;4(9):688-694. doi:10.1038/nrc1433
51. Agazie YM, Hayman MJ. Molecular mechanism for a role of SHP2 in epidermal growth factor receptor signaling. *Molecular and Cellular Biology*. 2003;23(21):7875-7886. doi:10.1128/mcb.23.21.7875-7886.2003
52. Murata-Kamiya N, Kurashima Y, Teishikata Y, et al. Helicobacter pylori CagA interacts with e-cadherin and deregulates the beta-catenin signal that promotes intestinal transdifferentiation in gastric epithelial cells. *Oncogene*. 2007;26(32):4617-4626. doi:10.1038/sj.onc.1210251
53. MacDonald BT, Tamai K, He X. Wnt/beta-catenin signaling: Components, mechanisms, and diseases. *Developmental Cell*. 2009;17(1):9-26. doi:10.1016/j.devcel.2009.06.016
54. Shields MS, Althoff KN, Pfeiffer RM, et al. HIV infection, immunosuppression, and age at diagnosis of non-AIDS-defining cancers. *Clinical Infectious Diseases*. December 2016:ciw764. doi:10.1093/cid/ciw764
55. Brennan CA, Garrett WS. Fusobacterium nucleatum symbiont, opportunist and oncobacterium. *Nature Reviews Microbiology*. 2018;17(3):156-166. doi:10.1038/s41579-018-0129-6
56. Fulbright LE, Ellermann M, Arthur JC. The microbiome and the hallmarks of cancer. Leong JM, ed. *PLOS Pathogens*. 2017;13(9):e1006480. doi:10.1371/journal.ppat.1006480
57. Savage DC. Microbial ecology of the gastrointestinal tract. *Annual Review of Microbiology*. 1977;31(1):107-133. doi:10.1146/annurev.mi.31.100177.000543
58. Wilson KH, Blichington RB. Human colonic biota studied by ribosomal DNA sequence analysis. *Applied and environmental microbiology*. 1996;62(7):2273-2278. doi:10.1128/aem.62.7.2273-2278.1996
59. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. *Science*. 2009;326(5960):1694-1697. doi:10.1126/science.1177486
60. Kamps R, Brandao RD, Bosch BJ, et al. Next-generation sequencing in oncology: Genetic diagnosis, risk prediction and cancer classification. *Int J Mol Sci*. 2017;18(2). doi:10.3390/ijms18020308

61. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. *Front Plant Sci.* 2014;5:209. doi:10.3389/fpls.2014.00209
62. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. *Nature.* 2007;449(7164):804-810. doi:10.1038/nature06244
63. Gevers D, Pop M, Schloss PD, Huttenhower C. Bioinformatics for the human microbiome project. *PLoS Comput Biol.* 2012;8(11):e1002779. doi:10.1371/journal.pcbi.1002779
64. Backhed F, Roswall J, Peng Y, et al. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host & Microbe.* 2015;17(6):852. doi:10.1016/j.chom.2015.05.012
65. Yatsunenko T, Rey FE, Manary MJ, et al. Human gut microbiome viewed across age and geography. *Nature.* 2012;486(7402):222-227. doi:10.1038/nature11053
66. Subramanian S, Huq S, Yatsunenko T, et al. Persistent gut microbiota immaturity in malnourished bangladeshi children. *Nature.* 2014;510(7505):417-421. doi:10.1038/nature13421
67. Schloissnig S, Arumugam M, Sunagawa S, et al. Genomic variation landscape of the human gut microbiome. *Nature.* 2012;493(7430):45-50. doi:10.1038/nature11711
68. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature.* 2006;444(7122):1027-1031. doi:10.1038/nature05414
69. Forslund K, Falk Hildebrand, Nielsen T, et al. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature.* 2015;528(7581):262-266. doi:10.1038/nature15766
70. Foster JA, Neufeld K-AM. Gutbrain axis: How the microbiome influences anxiety and depression. *Trends in Neurosciences.* 2013;36(5):305-312. doi:10.1016/j.tins.2013.01.005
71. Jackson MA, Verdi S, Maxan M-E, et al. Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nature Communications.* 2018;9(1). doi:10.1038/s41467-018-05184-7
72. Structure, function and diversity of the healthy human microbiome. *Nature.* 2012;486(7402):207-214. doi:10.1038/nature11234
73. Einstein MH, Baron M, Levin MJ, et al. Comparison of the immunogenicity and safety of Cervarix and Gardasil human papillomavirus (HPV) cervical cancer vaccines in healthy women aged 18-45 years. *Human Vaccines.* 2009;5(10):705-719. doi:10.4161/hv.5.10.9518
74. Lei J, Ploner A, Elfström KM, et al. HPV vaccination and the risk of invasive cervical cancer. *New England Journal of Medicine.* 2020;383(14):1340-1348. doi:10.1056/nejmoa1917338
75. Jong FA de, Kehler DFS, Mathijssen RHJ, et al. Prophylaxis of irinotecan-induced diarrhea with neomycin and potential role for UGT1A1 genotype screening: A double-blind, randomized, placebo-controlled study. *The Oncologist.* 2006;11(8):944-954. doi:10.1634/theoncologist.11-8-944

76. Ye Q, Shao X, Shen R, Chen D, Shen J. Changes in the human gut microbiota composition caused by helicobacter pylori eradication therapy: A systematic review and meta-analysis. *Helicobacter*. 2020;25(4). doi:10.1111/hel.12713
77. Wu W, Yang Y, Sun G. Recent insights into antibiotic resistance in *Helicobacter pylori* eradication. *Gastroenterology Research and Practice*. 2012;2012:1-8. doi:10.1155/2012/723183
78. Ingman WV. The gut microbiome: A new player in breast cancer metastasis. *Cancer Research*. 2019;79(14):3539-3541. doi:10.1158/0008-5472.can-19-1698
79. Borody TJ, Peattie D, Campbell J. Therapeutic potential of the human gastrointestinal microbiome. *Drug Development Research*. 2013;74(6):385-392. doi:10.1002/ddr.21093
80. Nikhra V. Therapeutic potential of gut microbiome manipulation: Concepts in fecal microbiota transplantation. *Current Research in Diabetes & Obesity Journal*. 2019;11(1). doi:10.19080/crdoj.2019.11.555803
81. Pirofski L-a, Casadevall A. Q&A: What is a pathogen? A question that begs the point. *BMC Biology*. 2012;10(1). doi:10.1186/1741-7007-10-6
82. Ohkusa T, Yoshida T, Sato N, Watanabe S, Tajiri H, Okayasu I. Commensal bacteria can enter colonic epithelial cells and induce proinflammatory cytokine secretion: A possible pathogenic mechanism of ulcerative colitis. *Journal of Medical Microbiology*. 2009;58(5):535-545. doi:10.1099/jmm.0.005801-0
83. Watson J, Crick F. The structure of dna. *Cold Spring Harb Symp Quant Bio*. 1953;(18):123-131. doi:10.1101/SQB.1953.018.01.020
84. Sanger F, Coulson A. A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *J Mol Bio*. 1975;94:441-448.
85. Maxam A, Gilbert W. A new method for sequencing dna. *Proceedings of the National Academy of Sciences*. 1977;74(2):560-564.
86. Heather JM, Chain B. The sequence of sequencers: The history of sequencing dna. *Genomics*. 2016;107(1):1-8. doi:10.1016/j.ygeno.2015.11.003
87. Chen CY. DNA polymerases drive dna sequencing-by-synthesis technologies: Both past and present. *Front Microbiol*. 2014;5:305. doi:10.3389/fmicb.2014.00305
88. 2018. https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.
89. Jain M, Koren S, Miga KH, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. 2018. doi:10.1038/nbt.4060
90. Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*. 2015;13(5):278-289. doi:10.1016/j.gpb.2015.08.002
91. Kchouk M, Elloumi M. Error correction and denovo genome assembly for the minion sequencing reads mixing illumina short reads. *International Conference on Bioinformatics and Biomedicine*. 2015:1785.

92. Johnson SS, Zaikova E, Goerlitz DS, Bai Y, Tighe SW. Real-time DNA sequencing in the antarctic dry valleys using the oxford nanopore sequencer. *Journal of Biomolecular Techniques : JBT*. 2017;28(1):2-7. doi:10.7171/jbt.17-2801-009
93. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. *Mol Cell*. 2015;58(4):586-597. doi:10.1016/j.molcel.2015.05.004
94. Makhzami S, Rambow F, Delmas V, Larue L. Efficient gene expression profiling of laser-microdissected melanoma metastases. *Pigment Cell Melanoma Res*. 2012;25(6):783-791. doi:10.1111/pcmr.12013
95. Nimwegen KJ van, Soest RA van, Veltman JA, et al. *Clin Chem*. 2016;62(11):1458-1464. doi:10.1373/clinchem.2016.258632
96. Zapatka M, Borozan I, Brewer DS, et al. The landscape of viral associations in human cancers. *bioRxiv*. 2018. doi:10.1101/465757
97. Allali I, Boukhatem N, Bouguenouch L, et al. Gut microbiome of moroccan colorectal cancer patients. *Med Microbiol Immunol*. 2018;207(3-4):211-225. doi:10.1007/s00430-018-0542-5
98. Sczyrba A, Hofmann P, Belmann P, et al. Critical assessment of metagenome interpretation-a benchmark of metagenomics software. *Nat Methods*. 2017;14(11):1063-1071. doi:10.1038/nmeth.4458
99. Truong D, Franzosa E, Tickle T, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods*. 2015;12(10):902-903. doi:10.1038/nmeth.3589
100. Papudeshi B, Haggerty JM, Doane M, et al. Optimizing and evaluating the reconstruction of metagenome-assembled microbial genomes. *BMC Genomics*. 2017;18(1):915. doi:10.1186/s12864-017-4294-1
101. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nat Commun*. 2016;7:11257. doi:10.1038/ncomms11257
102. Vollmers J, Wiegand S, Kaster A. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! *PLoS One*. 2017;12(1):e0169662. doi:10.1371/journal.pone.0169662
103. Wood D, Salzberg S. Kraken - ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. 2014;15(3). doi:10.1186/gb-2014-15-3-r46.
104. Milanese A, Mende DR, Paoli L, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun*. 2019;10(1):1014. doi:10.1038/s41467-019-08844-4
105. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;27(6):764-770. doi:10.1093/bioinformatics/btr011
106. Mende DR, Letunic I, Huerta-Cepas J, et al. proGenomes: A resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Research*. 2016;45(D1):D529-D534. doi:10.1093/nar/gkw989
107. Li H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv:1303.3997v1*. 2013;[q-bio.GN].

108. Weiss S, Xu ZZ, Peddada S, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*. 2017;5(1):27. doi:10.1186/s40168-017-0237-y
109. Lu J, Breitwieser F, Thielen P, Salzberg S. Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science*. 2017;3. doi:10.7717/peerj-cs.104
110. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods*. 2012;9(8):811-814. doi:10.1038/nmeth.2066
111. Parks DH, Beiko RG. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*. 2010;26(6):715-721. doi:10.1093/bioinformatics/btq041
112. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337-350. doi:10.1007/s10654-016-0149-3
113. Armstrong RA. When to use the bonferroni correction. *Ophthalmic Physiol Opt*. 2014;34(5):502-508. doi:10.1111/opo.12131
114. Chen SY, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. *J Thorac Dis*. 2017;9(6):1725-1729. doi:10.21037/jtd.2017.05.34
115. Benjamini Y, Hochberg Y. Controlling the false discovery rate- a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995;57(1):289-300.
116. Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation. *Genome Biol*. 2010;12(R60).
117. Morgan XC, Tickle TL, Sokol H, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol*. 2012;13(9):R79. doi:10.1186/gb-2012-13-9-r79
118. Huttenhower C. 2019. <http://huttenhower.sph.harvard.edu/maaslin>.
119. Friedman J. Stochastic gradient boosting. *Computational Statistics & Data Analysis*. 2002;38(4):367-378. doi:https://doi.org/10.1016/S0167-9473(01)00065-2
120. Paulson J, Olson N, Wagner J, Talukder H, Pop M, Bravo H. MetagenomeSeq - statistical analysis for sparse high-throughput sequencing. 2019. doi:10.18129/B9.bioc.metagenomeSeq
121. 2018. doi:10.18129/B9.bioc.SIAMCAT
122. Shade A. Diversity is the question, not the answer. *ISME J*. 2017;11(1):1-6. doi:10.1038/ismej.2016.118
123. Mosca A, Leclerc M, Hugot JP. Gut microbiota diversity and human diseases: Should we reintroduce key predators in our ecosystem? *Front Microbiol*. 2016;7:455. doi:10.3389/fmicb.2016.00455
124. Faith D. Conservation evaluation and phylogenetic diversity. *Biological Conservation*. 1992;61(1):1-10. doi:https://doi.org/10.1016/0006-3207(92)91201-3

125. McCoy CO, Matsen FA th. Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. *PeerJ*. 2013;1:e157. doi:10.7717/peerj.157
126. Morris EK, Caruso T, Buscot F, et al. Choosing and using diversity indices: Insights for ecological applications from the german biodiversity exploratories. *Ecology and Evolution*. 2014;4(18):3514-3524. doi:10.1002/ece3.1155
127. Ahn J, Sinha R, Pei Z, et al. Human gut microbiome and risk for colorectal cancer. *J Natl Cancer Inst*. 2013;105(24):1907-1911. doi:10.1093/jnci/djt300
128. Wong RG, Wu JR, Gloor GB. Expanding the unifrac toolbox. *PLoS One*. 2016;11(9):e0161196. doi:10.1371/journal.pone.0161196
129. Lozupone C, Knight R. UniFrac: A new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*. 2005;71(12):8228-8235. doi:10.1128/AEM.71.12.8228-8235.2005
130. Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol*. 2007;73(5):1576-1585. doi:10.1128/AEM.01996-06
131. Lozupone CA, Knight R. The unifrac significance test is sensitive to tree topology. *BMC Bioinformatics*. 2015;16:211. doi:10.1186/s12859-015-0640-y
132. Jaccard P. The distribution of the flora in the alpine zone. *New Phytologist*. 1912;11(2). doi:https://doi.org/10.1111/j.1469-8137.1912.tb05611.x
133. Faith DP, Minchin PR, Belbin L. Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*. 1987;69(1-3):57-68. doi:10.1007/bf00038687
134. Tripathi A, Marotz C, Gonzalez A, et al. Are microbiome studies ready for hypothesis-driven research? *Curr Opin Microbiol*. 2018;44:61-69. doi:10.1016/j.mib.2018.07.002
135. Kuczynski J, Liu Z, Lozupone C, McDonald D, Fierer N, Knight R. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Methods*. 2010;7(10):813-819. doi:10.1038/nmeth.1499
136. Torgerson W. Multidimensional scaling: I. Theory and method. *Psychometrika*. 1952;17(1):401-219.
137. Wickelmaier F. An introduction to mds. *Sound Quality Research Unit - Aalborg University, Denmark*. 2003. <https://homepage.uni-tuebingen.de/florian.wickelmaier/pubs/Wickelmaier2003SQRU.pdf>.
138. Buttigieg PL, Ramette A. A guide to statistical analysis in microbial ecology: A community-focused, living review of multivariate data analyses. *FEMS Microbiol Ecol*. 2014;90(3):543-550. doi:10.1111/1574-6941.12437
139. 2014;2019. <https://sites.google.com/site/mb3gustame/dissimilarity-based-methods/>.
140. Goodrich JK, Di Rienzi SC, Poole AC, et al. Conducting a microbiome study. *Cell*. 2014;158(2):250-262. doi:10.1016/j.cell.2014.06.037

141. Shepard R. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*. 1962;27(2):125-140. doi:<https://doi.org/10.1007/BF02289630>
142. Kruskal J. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*. 1964;29(2):115-129. doi:<https://doi.org/10.1007/BF02289694>
143. Kruskal J. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*. 1964;29(1). doi:<https://doi.org/10.1007/BF02289565>
144. 2018;2019. http://biol09.biol.umontreal.ca/PLcourses/Comparison_nMDS_PCoA.pdf.
145. Arndt D, Xia J, Liu Y, et al. METAGENassist: A comprehensive web server for comparative metagenomics. *Nucleic Acids Research*. 2012;40(W1):W88-W95. doi:10.1093/nar/gks497
146. Lev-Sagie A, Goldman-Wohl D, Cohen Y, et al. Vaginal microbiome transplantation in women with intractable bacterial vaginosis. *Nature Medicine*. 2019;25(10):1500-1504. doi:10.1038/s41591-019-0600-6
147. Bonaccorso G. *Machine Learning Algorithms: Popular Algorithms for Data Science and Machine Learning*. Packt; 2018.
148. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 2002;16:321-357. doi:10.1613/jair.953
149. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*. 2013;14(1). doi:10.1186/1471-2105-14-106
150. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019;1(5):206-215. doi:10.1038/s42256-019-0048-x
151. Kursa MB, Jankowski A, Rudnicki WR. Boruta – a system for feature selection. *Fundamenta Informaticae*. 2010;101(4):271-285. doi:10.3233/FI-2010-288
152. Dave M, Higgins PD, Middha S, Rioux KP. The human gut microbiome: Current knowledge, challenges, and future directions. *Translational Research*. 2012;160(4):246-257. doi:10.1016/j.trsl.2012.05.003
153. Sun Y, Todorovic S, Goodison S. Local-learning-based feature selection for high-dimensional data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2010;32(9):1610-1626. doi:10.1109/tpami.2009.190
154. Poore GD, Kopylova E, Zhu Q, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature*. 2020;579(7800):567-574. doi:10.1038/s41586-020-2095-1
155. Huerta-Cepas J, Szklarczyk D, Heller D, et al. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*. 2018;47(D1):D309-D314. doi:10.1093/nar/gky1085
156. Tatusov RL. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*. 2000;28(1):33-36. doi:10.1093/nar/28.1.33

157. Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*. 2000;28(1):27-30. doi:10.1093/nar/28.1.27
158. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. *Protein Science*. 2019;28(11):1947-1951. doi:10.1002/pro.3715
159. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Research*. 2018;47(D1):D590-D595. doi:10.1093/nar/gky962
160. El-Gebali S, Mistry J, Bateman A, et al. The pfam protein families database in 2019. *Nucleic Acids Research*. 2018;47(D1):D427-D432. doi:10.1093/nar/gky995
161. Haft DH. The TIGRFAMs database of protein families. *Nucleic Acids Research*. 2003;31(1):371-373. doi:10.1093/nar/gkg128
162. Hunter S, Apweiler R, Attwood TK, et al. InterPro: The integrative protein signature database. *Nucleic Acids Research*. 2009;37(Database):D211-D215. doi:10.1093/nar/gkn785
163. Galloway-Peña J, Hanson B. Tools for analysis of the microbiome. *Digestive Diseases and Sciences*. 2020;65(3):674-685. doi:10.1007/s10620-020-06091-y
164. Jones P, Binns D, Chang H-Y, et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics*. 2014;30(9):1236-1240. doi:10.1093/bioinformatics/btu031
165. Yong-Xia Z, Ge Z. MD5 research. In: *2010 Second International Conference on Multimedia and Information Technology*. IEEE; 2010. doi:10.1109/mmit.2010.186
166. Douglas GM, Maffei VJ, Zaneveld JR, et al. PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology*. 2020;38(6):685-688. doi:10.1038/s41587-020-0548-6
167. Langille MG, Zaneveld J, Caporaso JG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol*. 2013;31(9):814-821. doi:10.1038/nbt.2676
168. Franzosa EA, McIver LJ, Rahnavard G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*. 2018;15(11):962-968. doi:10.1038/s41592-018-0176-y
169. Caspi R, Billington R, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Research*. 2017;46(D1):D633-D639. doi:10.1093/nar/gkx935
170. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. *Nature Genetics*. 2000;25(1):25-29. doi:10.1038/75556
171. Consortium GO. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Research*. 2018;47(D1):D330-D338. doi:10.1093/nar/gky1055
172. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*. 2014;12(1). doi:10.1186/s12915-014-0087-z
173. Nejman D, Livyatan I, Fuks G, et al. The human tumor microbiome is composed of tumor typespecific intracellular bacteria. *Science*. 2020;368(6494):973-980. doi:10.1126/science.aay9189

174. Stinson L, Keelan J, Payne M. Identification and removal of contaminating microbial DNA from PCR reagents: Impact on low-biomass microbiome analyses. *Letters in Applied Microbiology*. 2018;68(1):2-8. doi:10.1111/lam.13091
175. Olomu IN, Pena-Cortes LC, Long RA, et al. Elimination of “kitome” and “splashome” contamination results in lack of detection of a unique placental microbiome. *BMC Microbiology*. 2020;20(1). doi:10.1186/s12866-020-01839-y
176. Nielsen ML, Raahave D, Stage JG, Justesen T. Anaerobic and aerobic skin bacteria before and after skin-disinfection with chlorhexidine: An experimental study in volunteers. *Journal of Clinical Pathology*. 1975;28(10):793-797. doi:10.1136/jcp.28.10.793
177. Ross AA, Müller KM, Weese JS, Neufeld JD. Comprehensive skin microbiome analysis reveals the uniqueness of human skin and evidence for phylosymbiosis within the class mammalia. *Proceedings of the National Academy of Sciences*. 2018;115(25):E5786-E5795. doi:10.1073/pnas.1801302115
178. Venkateswaran K, Vaishampayan P, Cisneros J, Pierson DL, Rogers SO, Perry J. International space station environmental microbiome microbial inventories of ISS filter debris. *Applied Microbiology and Biotechnology*. 2014;98(14):6453-6466. doi:10.1007/s00253-014-5650-6
179. Gu W, Miller S, Chiu CY. Clinical metagenomic next-generation sequencing for pathogen detection. *Annual Review of Pathology: Mechanisms of Disease*. 2019;14(1):319-338. doi:10.1146/annurev-pathmechdis-012418-012751
180. Minich JJ, Sanders JG, Amir A, Humphrey G, Gilbert JA, Knight R. Quantifying and understanding well-to-well contamination in microbiome research. Segata N, ed. *mSystems*. 2019;4(4). doi:10.1128/msystems.00186-19
181. Steinegger M, Salzberg SL. Terminating contamination: Large-scale search identifies more than 2, 000, 000 contaminated entries in GenBank. *Genome Biology*. 2020;21(1). doi:10.1186/s13059-020-02023-1
182. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in low microbial biomass microbiome studies: Issues and recommendations. *Trends in Microbiology*. 2019;27(2):105-117. doi:10.1016/j.tim.2018.11.003
183. Strong MJ, Xu G, Morici L, et al. Microbial contamination in next generation sequencing: Implications for sequence-based analysis of clinical samples. Rall GF, ed. *PLoS Pathogens*. 2014;10(11):e1004437. doi:10.1371/journal.ppat.1004437
184. Han YW. *Fusobacterium nucleatum*: A commensal-turned pathogen. *Current Opinion in Microbiology*. 2015;23:141-147. doi:10.1016/j.mib.2014.11.013
185. Rasko DA, Rosovitz MJ, Myers GSA, et al. The pangenome structure of *Escherichia coli*: Comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *Journal of Bacteriology*. 2008;190(20):6881-6893. doi:10.1128/jb.00619-08
186. Gonzalez A, Vázquez-Baeza Y, Pettengill JB, Ottosen A, McDonald D, Knight R. Avoiding pandemic fears in the subway and conquering the platypus. *mSystems*. 2016;1(3). doi:10.1128/msystems.00050-16

187. Afshinnikoo E, Meydan C, Chowdhury S, et al. Modern methods for delineating metagenomic complexity. *Cell Systems*. 2015;1(1):6-7. doi:10.1016/j.cels.2015.07.007
188. Hsu T, Joice R, Vallarino J, et al. Urban transit system microbial communities differ by surface type and interaction with humans and the environment. Knight R, ed. *mSystems*. 2016;1(3). doi:10.1128/msystems.00018-16
189. Ackelsberg J, Rakeman J, Hughes S, et al. Lack of evidence for plague or anthrax on the new york city subway. *Cell Systems*. 2015;1(1):4-5. doi:10.1016/j.cels.2015.07.008
190. Gihawi A, Rallapalli G, Hurst R, Cooper CS, Leggett RM, Brewer DS. SEPATH: Benchmarking the search for pathogens in human tissue whole genome sequence data leads to template pipelines. *Genome Biology*. 2019;20(1). doi:10.1186/s13059-019-1819-8
191. Gihawi A, Rallapalli G, Hurst R, Cooper C, Leggett R, Brewer D. SEPATH: Benchmarking the search for pathogens in human tissue whole genome sequence data leads to template pipelines - analysis repository. September 2019. doi:10.5281/zenodo.3387205
192. Piro VC, Matschkowski M, Renard BY. MetaMeta: Integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome*. 2017;5(1):101. doi:10.1186/s40168-017-0318-y
193. CAMI. CAMI2. *CAMI2 - Microbiome COSI*. <https://www.microbiome-cosi.org/cami/cami/cami2>.
194. Whiteside SA, Razvi H, Dave S, Reid G, Burton JP. The microbiome of the urinary tract—a role beyond infection. *Nat Rev Urol*. 2015;12(2):81-90. doi:10.1038/nrurol.2014.361
195. Bedarf JR, Beraza N, Khazneh H, et al. Much ado about nothing? Off-target amplification can lead to false-positive bacterial brain microbiome detection in healthy and parkinson's disease individuals. *Microbiome*. 2021;9(1). doi:10.1186/s40168-021-01012-1
196. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170
197. 2018. <https://www.ncbi.nlm.nih.gov/genome>.
198. Kraal L, Abubucker S, Kota K, Fischbach MA, Mitreva M. The prevalence of species and strains in the human microbiome: A resource for experimental efforts. *PLoS One*. 2014;9(5):e97279. doi:10.1371/journal.pone.0097279
199. Lax A. Bacterial toxins and cancer - a case to answer? *Nature Reviews*. 2005;3:343-349.
200. Leggett RM, Ramirez-Gonzalez RH, Clavijo BJ, Waite D, Davey RP. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Front Genet*. 2013;4:288. doi:10.3389/fgene.2013.00288
201. Daly GM, Leggett RM, Rowe W, et al. Host subtraction, filtering and assembly validations for novel viral discovery using next generation sequencing data. *PLoS One*. 2015;10(6):e0129059. doi:10.1371/journal.pone.0129059
202. 2018. <https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/>.

203. Catalogue of Somatic Mutations in Cancer - COSMIC. Data downloads. 2018;2018. <https://cancer.sanger.ac.uk/cosmic/download>.
204. Sherman RM, Forman J, Antonescu V, et al. Assembly of a pan-genome from deep sequencing of 910 humans of african descent. *Nat Genet.* 2019;51(1):30-35. doi:10.1038/s41588-018-0273-y
205. Freitas TA, Li PE, Scholz MB, Chain PS. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res.* 2015;43(10):e69. doi:10.1093/nar/gkv180
206. Kim D, Song L, Breitwieser F, Salzberg S. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Res.* 2017;26(12):1721-1729. doi:10.1101/gr.210641.116
207. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. MetaSPAdes: A new versatile metagenomic assembler. *Genome Res.* 2017;27(5):824-834. doi:10.1101/gr.213959.116
208. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: Confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.* 2018;19(1):198. doi:10.1186/s13059-018-1568-0
209. Kestic A, Ojesina A, Pedamallu C, et al. PathSeq: Software to identify or discover microbes by deep sequencing of human tissue. *Nature Biotechnology.* 2011;29(5):393-396. doi:10.1038/nbt0511-393
210. Walker MA, Pedamallu CS, Ojesina AI, et al. GATK PathSeq: A customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. Berger B, ed. *Bioinformatics.* July 2018. doi:10.1093/bioinformatics/bty501
211. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. *J Mol Biol.* 1990;215(3).
212. Salzberg SL, Breitwieser FP, Kumar A, et al. Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system. *Neurol Neuroimmunol Neuroinflamm.* 2016;3(4):e251. doi:10.1212/NXI.0000000000000251
213. Laurence M, Hatzis C, Brash DE. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One.* 2014;9(5):e97876. doi:10.1371/journal.pone.0097876
214. Mesri EA, Feitelson MA, Munger K. Human viral oncogenesis: A cancer hallmarks analysis. *Cell Host Microbe.* 2014;15(3):266-282. doi:10.1016/j.chom.2014.02.011
215. Iizasa H, Nanbo A, Nishikawa J, Jinushi M, Yoshiyama H. Epstein-barr virus (ebv)-associated gastric carcinoma. *Viruses.* 2012;4(12):3420-3439. doi:10.3390/v4123420
216. Koster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics.* 2012;28(19):2520-2522. doi:10.1093/bioinformatics/bts480
217. McElroy KE, Luciani F, Thomas T. GemSIM: General, error-model based simulator of next-generation sequencing data. *BMC Genomics.* 2012;13:74. doi:10.1186/1471-2164-13-74

218. Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW. Grinder: A versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.* 2012;40(12):e94. doi:10.1093/nar/gks251
219. Johnson S, Trost B, Long JR, Pittet V, Kusalik A. A better sequence-read simulator program for metagenomics. *BMC Bioinformatics.* 2014;15 Suppl 9:S14. doi:10.1186/1471-2105-15-S9-S14
220. Fritz A, Hofmann P, Majda S, et al. CAMISIM: Simulating metagenomes and microbial communities. *Microbiome.* 2019;7(1). doi:10.1186/s40168-019-0633-6
221. Wylezich C, Papa A, Beer M, Höper D. A versatile sample processing workflow for metagenomic pathogen detection. *Scientific Reports.* 2018;8(1). doi:10.1038/s41598-018-31496-1
222. Starke R, Jehmlich N, Alfaro T, et al. Incomplete cell disruption of resistant microbes. *Scientific Reports.* 2019;9(1). doi:10.1038/s41598-019-42188-9
223. Zhang B, Brock M, Arana C, Dende C, Hooper L, Raj P. Impact of bead-beating intensity on microbiome recovery in mouse and human stool: Optimization of DNA extraction. June 2020. doi:10.1101/2020.06.15.151753
224. Panek M, Paljetak H Čipčić, Barešić A, et al. Methodology challenges in studying human gut microbiota effects of collection, storage, DNA extraction and next generation sequencing technologies. *Scientific Reports.* 2018;8(1). doi:10.1038/s41598-018-23296-4
225. De Goffau M, Charnock-Jones S, Smith G, Parkhill J. Batch effects account for the main findings of an in utero human intestinal bacterial colonization study. 2020. doi:10.17863/CAM.54041
226. Kim Y-M, Poline J-B, Dumas G. Experimenting with reproducibility: A case study of robustness in bioinformatics. *GigaScience.* 2018;7(7). doi:10.1093/gigascience/giy077
227. Gruening B, Sallou O, Moreno P, et al. Recommendations for the packaging and containerizing of bioinformatics software. *F1000Research.* 2019;7:742. doi:10.12688/f1000research.15140.2
228. Kanwal S, Khan FZ, Lonie A, Sinnott RO. Investigating reproducibility and tracking provenance a genomic workflow case study. *BMC Bioinformatics.* 2017;18(1). doi:10.1186/s12859-017-1747-0
229. Ravel J, Wommack KE. All hail reproducibility in microbiome research. *Microbiome.* 2014;2(1). doi:10.1186/2049-2618-2-8
230. Brooks JP. Challenges for case-control studies with microbiome data. *Annals of Epidemiology.* 2016;26(5):336-341.e1. doi:10.1016/j.annepidem.2016.03.009
231. Knight R, Vrbanac A, Taylor BC, et al. Best practices for analysing microbiomes. *Nature Reviews Microbiology.* 2018;16(7):410-422. doi:10.1038/s41579-018-0029-9
232. Wirbel J, Pyl PT, Kartal E, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med.* 2019;25(4):679-689. doi:10.1038/s41591-019-0406-6

233. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet.* 2013;45(10):1113-1120. doi:10.1038/ng.2764
234. Northcott PA, Buchhalter I, Morrissy AS, et al. The whole-genome landscape of medulloblastoma subtypes. *Nature.* 2017;547(7663):311-317.
235. Xing R, Zhou Y, Yu J, et al. Whole-genome sequencing reveals novel tandem-duplication hotspots and a prognostic mutational signature in gastric cancer. *Nat Commun.* 2019;10(1):2037.
236. Campbell PJ, Getz G, Stuart JM, Korbel JO, Stein LD. Pan-cancer analysis of whole genomes., ed. *bioRxiv.* 2017. doi:10.1101/162784
237. Prostate cancer statistics. *Cancer Research UK.* April 2020. <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer#heading-Five>.
238. Kimura T, Egawa S. Epidemiology of prostate cancer in asian countries. *International Journal of Urology.* 2018;25(6):524-531. doi:10.1111/iju.13593
239. Gundem G, Loo PV, Kremeyer B, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature.* 2015;520(7547):353-357. doi:10.1038/nature14347
240. Luca B, Brewer D, Edwards D, et al. DESNT: A poor prognosis category of human prostate cancer. *Eur Urol Focus.* 2017. doi:10.1016/j.euf.2017.01.016
241. Connell SP, Yazbek-Hanna M, McCarthy F, et al. A four-group urine risk classifier for predicting outcomes in patients with prostate cancer. *BJU International.* 2019;124(4):609-620. doi:10.1111/bju.14811
242. Hernandez DJ, Nielsen ME, Han M, Partin AW. Contemporary evaluation of the d'Amico risk classification of prostate cancer. *Urology.* 2007;70(5):931-935. doi:10.1016/j.urology.2007.08.055
243. Humphrey PA. Gleason grading and prognostic factors in carcinoma of the prostate. *Modern Pathology.* 2004;17(3):292-306. doi:10.1038/modpathol.3800054
244. Crawford ED, Rove KO, Trabulsi EJ, et al. Diagnostic performance of PCA3 to detect prostate cancer in men with increased prostate specific antigen: A prospective study of 1,962 cases. *Journal of Urology.* 2012;188(5):1726-1731. doi:10.1016/j.juro.2012.07.023
245. Thompson IM. Operating characteristics of prostate-specific antigen in men with an initial PSA level of 3.0 ng/mL or lower. *JAMA.* 2005;294(1):66. doi:10.1001/jama.294.1.66
246. Hunt TC, Ambrose JP, Haaland B, Hanson HA, O'Neil BB. MP75-061emDECISION FATIGUE IN PROSTATE-SPECIFIC ANTIGEN (PSA) TESTING. *The Journal of Urology.* 2020;203:e1143-e1144. doi:10.1097/ju.0000000000000961.06
247. Ahmed HU, Bosaily AE-S, Brown LC, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): A paired validating confirmatory study. *The Lancet.* 2017;389(10071):815-822. doi:10.1016/s0140-6736(16)32401-1
248. Pepe P, Garufi A, Priolo GD, Pennisi M. Multiparametric MRI/TRUS fusion prostate biopsy: Advantages of a transperineal approach. *Anticancer Research.* 2017;37(6).

doi:10.21873/anticancerres.11695

249. NICE. Prostate cancer overview. *Prostate cancer - NICE Pathways*. <https://pathways.nice.org.uk/pathways/prostate-cancer#path=view:/pathways/prostate-cancer/assessing-suspected-prostate-cancer.xml&content=view-node:nodes-information-and-support-for-decision-making>.
250. Oudard S, Fizazi K, Sengeløv L, et al. Cabazitaxel versus docetaxel as first-line therapy for patients with metastatic castration-resistant prostate cancer: A randomized phase III trial FIRSTANA. *Journal of Clinical Oncology*. 2017;35(28):3189-3197. doi:10.1200/jco.2016.72.1068
251. Frey AU, Sønksen J, Fode M. Neglected side effects after radical prostatectomy: A systematic review. *The Journal of Sexual Medicine*. 2014;11(2):374-385. doi:10.1111/jsm.12403
252. experience A personal. Active surveillance. *Prostate Cancer UK*. <https://prostatecanceruk.org/prostate-information/treatments/active-surveillance>.
253. Webb M, Manley K, Olivan M, et al. Methodology for the at-home collection of urine samples for prostate cancer detection. *BioTechniques*. 2020;68(2):65-71. doi:10.2144/btn-2019-0092
254. Becerra MF, Atluri VS, Bhattu AS, Punnen S. Serum and urine biomarkers for detecting clinically significant prostate cancer. *Urologic Oncology: Seminars and Original Investigations*. March 2020. doi:10.1016/j.urolonc.2020.02.018
255. Eskra JN, Rabizadeh D, Pavlovich CP, Catalona WJ, Luo J. Approaches to urinary detection of prostate cancer. *Prostate Cancer and Prostatic Diseases*. 2019;22(3):362-381. doi:10.1038/s41391-019-0127-4
256. Ploussard G, Taille A de la. Urine biomarkers in prostate cancer. *Nature Reviews Urology*. 2010;7(2):101-109. doi:10.1038/nrurol.2009.261
257. Lee JJ, Thomas I-C, Nolley R, Ferrari M, Brooks JD, Leppert JT. Biologic differences between peripheral and transition zone prostate cancer. *The Prostate*. 2014;75(2):183-190. doi:10.1002/pros.22903
258. UrologyCareFoundation. Prostatitis (infection of the prostate). *Urology Care Foundation The official Foundation of the American Urological Association*. [https://www.urologyhealth.org/urologic-conditions/prostatitis-\(infection-of-the-prostate\)](https://www.urologyhealth.org/urologic-conditions/prostatitis-(infection-of-the-prostate)).
259. Hrbacek J, Urban M, Hamsikova E, Tachezy R, Heracek J. Thirty years of research on infection and prostate cancer: No conclusive evidence for a link. A systematic review. *Urologic Oncology: Seminars and Original Investigations*. 2013;31(7):951-965. doi:10.1016/j.urolonc.2012.01.013
260. Puhr M, Marzo AD, Isaacs W, et al. Inflammation, microbiota, and prostate cancer. *European Urology Focus*. 2016;2(4):374-382. doi:10.1016/j.euf.2016.08.010
261. Shinohara DB, Vaghasia AM, Yu S-H, et al. A mouse model of chronic prostatic inflammation using a human prostate cancer-derived isolate of *Propionibacterium acnes*. *The Prostate*. 2013;73(9):1007-1015. doi:10.1002/pros.22648

262. Terzić J, Grivennikov S, Karin E, Karin M. Inflammation and colon cancer. *Gastroenterology*. 2010;138(6):2101-2114.e5. doi:10.1053/j.gastro.2010.01.058
263. Kim SC, Glynn RJ, Giovannucci E, et al. Risk of high-grade cervical dysplasia and cervical cancer in women with systemic inflammatory diseases: A population-based cohort study. *Annals of the Rheumatic Diseases*. 2014;74(7):1360-1367. doi:10.1136/annrheumdis-2013-204993
264. De Marzo AM, Platz EA, Sutcliffe S, et al. Inflammation in prostate carcinogenesis. *Nat Rev Cancer*. 2007;7(4):256-269. doi:10.1038/nrc2090
265. Meares EM. Prostatitis. *Medical Clinics of North America*. 1991;75(2):405-424. doi:10.1016/s0025-7125(16)30462-x
266. Feng Y, Jaratlerdsiri W, Patrick SM, et al. Metagenomic analysis reveals a rich bacterial content in high-risk prostate tumors from african men. *The Prostate*. 2019;79(15):1731-1738. doi:10.1002/pros.23897
267. Feng Y, Ramnarine VR, Bell R, et al. Metagenomic and metatranscriptomic analysis of human prostate microbiota from patients with prostate cancer. *BMC Genomics*. 2019;20(1). doi:10.1186/s12864-019-5457-z
268. Alanee S, El-Zawahry A, Dynda D, et al. A prospective study to examine the association of the urinary and fecal microbiota with prostate cancer diagnosis after transrectal biopsy of the prostate using 16sRNA gene analysis. *The Prostate*. 2018;79(1):81-87. doi:10.1002/pros.23713
269. Banerjee S, Alwine JC, Wei Z, et al. Microbiome signatures in prostate cancer. *Carcinogenesis*. 2019;40(6):749-764. doi:10.1093/carcin/bgz008
270. Alexandroff AB, Jackson AM, Donnell MAO, James K. BCG immunotherapy of bladder cancer: 20 years on. *The Lancet*. 1999;353(9165):1689-1694. doi:10.1016/s0140-6736(98)07422-4
271. Butel R, Ball R. The distribution of BCG prostatitis: A clue for pathogenetic processes? *The Prostate*. 2018;78(15):1134-1139. doi:10.1002/pros.23688
272. Bajic P, Wolfe AJ, Gupta GN. The urinary microbiome: Implications in bladder cancer pathogenesis and therapeutics. *Urology*. 2019;126:10-15. doi:10.1016/j.urology.2018.12.034
273. Bajic P, Kuiken MEV, Burge BK, et al. Male bladder microbiome relates to lower urinary tract symptoms. *European Urology Focus*. 2020;6(2):376-382. doi:10.1016/j.euf.2018.08.001
274. Homma Y, Gotoh M, Kawauchi A, et al. Clinical guidelines for male lower urinary tract symptoms and benign prostatic hyperplasia. *International Journal of Urology*. 2017;24(10):716-729. doi:10.1111/iju.13401
275. Heyns CF. Urinary tract infection associated with conditions causing urinary tract obstruction and stasis, excluding urolithiasis and neuropathic bladder. *World Journal of Urology*. 2011;30(1):77-83. doi:10.1007/s00345-011-0725-9
276. Ruiz J, Simon K, Horcajada JP, et al. Differences in virulence factors among clinical isolates of escherichia coli causing cystitis and pyelonephritis in women and prostatitis in

- men. *Journal of Clinical Microbiology*. 2002;40(12):4445-4449. doi:10.1128/jcm.40.12.4445-4449.2002
277. Lin K-Y, Chiu N-T, Chen M-J, et al. Acute pyelonephritis and sequelae of renal scar in pediatric first febrile urinary tract infection. *Pediatric Nephrology*. 2003;18(4):362-365. doi:10.1007/s00467-003-1109-1
278. Frimodt-Møller N. The urine microbiome contamination or a novel paradigm? *EBioMedicine*. 2019;44:20-21. doi:10.1016/j.ebiom.2019.05.016
279. Kirby RS, Lowe D, Bultitude MI, Shuttleworth KED. Intraprostatic urinary reflux: An aetiological factor in abacterial prostatitis. *British Journal of Urology*. 1982;54:729-731.
280. Popović VB, Šitum M, Chow C-ET, Chan LS, Roje B, Terzić J. The urinary microbiome associated with bladder cancer. *Scientific Reports*. 2018;8(1). doi:10.1038/s41598-018-29054-w
281. Nilsson J, Skog J, Nordstrand A, et al. Prostate cancer-derived urine exosomes: A novel approach to biomarkers for prostate cancer. *British Journal of Cancer*. 2009;100(10):1603-1607. doi:10.1038/sj.bjc.6605058
282. Laxman B, Morris DS, Yu J, et al. A first-generation multiplex biomarker analysis of urine for the early detection of prostate cancer. *Cancer Research*. 2008;68(3):645-649. doi:10.1158/0008-5472.can-07-3224
283. Varambally S, Laxman B, Mehra R, et al. Golgi protein GOLM1 is a tissue and urine biomarker of prostate cancer. *Neoplasia*. 2008;10(11):1285-IN35. doi:10.1593/neo.08922
284. Duijvesz D, Luijckx T, Bangma CH, Jenster G. Exosomes as biomarker treasure chests for prostate cancer. *European Urology*. 2011;59(5):823-831. doi:10.1016/j.eururo.2010.12.031
285. Liu Q, Rojas-Canales DM, Divito SJ, et al. Donor dendritic cell-derived exosomes promote allograft-targeting immune response. *Journal of Clinical Investigation*. 2016;126(8):2805-2820. doi:10.1172/jci84577
286. Chen W, Jiang J, Xia W, Huang J. Tumor-related exosomes contribute to tumor-promoting microenvironment: An immunological perspective. *Journal of Immunology Research*. 2017;2017:1-10. doi:10.1155/2017/1073947
287. O'Neill C, Gilligan K, Dwyer R. Role of extracellular vesicles (EVs) in cell stress response and resistance to cancer therapy. *Cancers*. 2019;11(2):136. doi:10.3390/cancers11020136
288. Lindenbergh MFS, Wubbolts R, Borg EGF, Veld EM van 'T, Boes M, Stoorvogel W. Dendritic cells release exosomes together with phagocytosed pathogen - potential implications for the role of exosomes in antigen presentation. *Journal of Extracellular Vesicles*. 2020;9(1):1798606. doi:10.1080/20013078.2020.1798606
289. Xu R, Rai A, Chen M, Suwakulsiri W, Greening DW, Simpson RJ. Extracellular vesicles in cancer implications for future improvements in cancer care. *Nature Reviews Clinical Oncology*. 2018;15(10):617-638. doi:10.1038/s41571-018-0036-9
290. Barreiro K, Holthofer H. Urinary extracellular vesicles. A promising shortcut to novel

- biomarker discoveries. *Cell and Tissue Research*. 2017;369(1):217-227. doi:10.1007/s00441-017-2621-0
291. Meka A, Bakthavatchalu V, Sathishkumar S, et al. Porphyromonas gingivalis infection-induced tissue and bone transcriptional profiles. *Molecular Oral Microbiology*. 2010;25(1):61-74. doi:10.1111/j.2041-1014.2009.00555.x
 292. Aragón IM, Herrera-Imbroda B, Queipo-Ortuño MI, et al. The urinary tract microbiome in health and disease. *European Urology Focus*. 2018;4(1):128-138. doi:10.1016/j.euf.2016.11.001
 293. Das P, Baker KK, Dutta A, et al. Menstrual hygiene practices, WASH access and the risk of urogenital infection in women from odisha, india. Wilson BA, ed. *PLOS ONE*. 2015;10(6):e0130777. doi:10.1371/journal.pone.0130777
 294. Ugge H, Carlsson J, Söderquist B, Fall K, Andén O, Davidsson S. The influence of prostatic cutibacterium acnes infection on serum levels of IL6 and CXCL8 in prostate cancer patients. *Infectious Agents and Cancer*. 2018;13(1). doi:10.1186/s13027-018-0204-7
 295. Karpiński T. Role of oral microbiota in cancer development. *Microorganisms*. 2019;7(1):20. doi:10.3390/microorganisms7010020
 296. Shang F-M, Liu H-L. Fusobacterium nucleatum and colorectal cancer: A review. *World Journal of Gastrointestinal Oncology*. 2018;10(3):71-81. doi:10.4251/wjgo.v10.i3.71
 297. Porter CM, Shrestha E, Peiffer LB, Sfanos KS. The microbiome in prostate inflammation and prostate cancer. *Prostate Cancer and Prostatic Diseases*. 2018;21(3):345-354. doi:10.1038/s41391-018-0041-1
 298. Cui Y, Chun S-J, Cho A-R, et al. Nevskia lacus sp. Nov., a gammaproteobacterium isolated from a eutrophic lake. *Antonie van Leeuwenhoek*. 2018;112(5):723-729. doi:10.1007/s10482-018-1206-6
 299. Natural transformation of pseudomonas fluorescens and agrobacterium tumefaciens in soil. *Applied and Environmental Microbiology*. 2001;67(6):2617-2621. doi:10.1128/aem.67.6.2617-2621.2001
 300. Paphitou NI, Rolston KVI. Catheter-related bacteremia caused by agrobacterium radiobacter in a cancer patient: Case report and literature review. *Infection*. 2003;31(6):421-424. doi:10.1007/s15010-003-3175-5
 301. Alnor D, Frimodt-Møller N, Espersen F, Frederiksen W. Infections with the unusual human pathogens agrobacterium species and ochrobactrum anthropi. *Clinical Infectious Diseases*. 1994;18(6):914-920. doi:10.1093/clinids/18.6.914
 302. Hwang H-H, Yu M, Lai E-M. Agrobacterium-mediated plant transformation: Biology and applications. *The Arabidopsis Book*. 2017;15:e0186. doi:10.1199/tab.0186
 303. Kunik T, Tzfira T, Kapulnik Y, Gafni Y, Dingwall C, Citovsky V. Genetic transformation of HeLa cells by agrobacterium. *Proceedings of the National Academy of Sciences*. 2001;98(4):1871-1876. doi:10.1073/pnas.98.4.1871
 304. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. Phillippy AM, ed. *PLOS Computational Biology*. 2017;13(6):e1005595. doi:10.1371/journal.pcbi.1005595

305. Cao H, Mok A, Miskie B, Hegele RA. Single-nucleotide polymorphisms of the proprotein convertase subtilisin/kexin type 5 (PCSK5) gene. *Journal of Human Genetics*. 2001;46(12):730-732. doi:10.1007/s100380170008
306. Seidah NG, Prat A. The biology and therapeutic targeting of the proprotein convertases. *Nature Reviews Drug Discovery*. 2012;11(5):367-383. doi:10.1038/nrd3699
307. Essalmani R, Zaid A, Marcinkiewicz J, et al. In vivo functions of the proprotein convertase PC5/6 during mouse development: Gdf11 is a likely substrate. *Proceedings of the National Academy of Sciences*. 2008;105(15):5750-5755. doi:10.1073/pnas.0709428105
308. Szumska D, Pieleś G, Essalmani R, et al. VACTERL/caudal regression/currarino syndrome-like malformations in mice with mutation in the proprotein convertase pcsk5. *Genes & Development*. 2008;22(11):1465-1477. doi:10.1101/gad.479408
309. Expression of pcsk5 in renal cancer - the human protein atlas. *Expression of PCSK5 in renal cancer - The Human Protein Atlas*. <https://www.proteinatlas.org/ENSG00000099139-PCSK5/pathology/renal.cancer>.
310. Iatan I, Dastani Z, Do R, et al. Genetic variation at the proprotein convertase subtilisin/kexin type 5 gene modulates high-density lipoprotein cholesterol levels. *Circulation: Cardiovascular Genetics*. 2009;2(5):467-475. doi:10.1161/circgenetics.109.877811
311. Shrestha E, White JR, Yu S-H, et al. Profiling the urinary microbiome in men with positive versus negative biopsies for prostate cancer. *Journal of Urology*. 2018;199(1):161-171. doi:10.1016/j.juro.2017.08.001
312. Lopez-Beltran A, Cheng L, Montorsi F, Scarpelli M, Raspollini MR, Montironi R. Concomitant bladder cancer and prostate cancer: Challenges and controversies. *Nature Reviews Urology*. 2017;14(10):620-629. doi:10.1038/nrurol.2017.124
313. Shah N. Diagnostic significance of levels of immunoglobulin a in seminal fluid of patients with prostatic disease. *Urology*. 1976;8(3):270-272. doi:10.1016/0090-4295(76)90384-8
314. Silva JAF, Biancardi MF, Stach-Machado DR, Reis LO, Sant'Anna OA, Carvalho HF. The origin of prostate gland-secreted IgA and IgG. *Scientific Reports*. 2017;7(1). doi:10.1038/s41598-017-16717-3
315. Hao L, Thomas S, Greer T, et al. Quantitative proteomic analysis of a genetically induced prostate inflammation mouse model via custom 4-plex DiLeu isobaric labeling. *American Journal of Physiology-Renal Physiology*. 2019;316(6):F1236-F1243. doi:10.1152/ajprenal.00387.2018
316. Schall TJ, Jongstra J, Dyer BJ, et al. A human t cell-specific molecule is a member of a new gene family. *The Journal of Immunology*. 1988;141(3):1018-1025. <https://www.jimmunol.org/content/141/3/1018>.
317. Kato T, Fujita Y, Nakane K, et al. CCR1/CCL5 interaction promotes invasion of taxane-resistant PC3 prostate cancer cells by increasing secretion of MMPs 2/9 and by activating ERK and rac signaling. *Cytokine*. 2013;64(1):251-257. doi:10.1016/j.cyto.2013.06.313

318. Vaday GG, Peehl DM, Kadam PA, Lawrence DM. Expression of CCL5 (RANTES) and CCR5 in prostate cancer. *The Prostate*. 2006;66(2):124-134. doi:10.1002/pros.20306
319. Frölich D, Blafeld D, Reiter K, et al. The anti-CD74 humanized monoclonal antibody, milatuzumab, which targets the invariant chain of MHC II complexes, alters b-cell proliferation, migration, and adhesion molecule expression. *Arthritis Research & Therapy*. 2012;14(2):R54. doi:10.1186/ar3767
320. Meyer-Siegler KL, Iczkowski KA, Vera PL. Further evidence for increased macrophage migration inhibitory factor expression in prostate cancer. *BMC Cancer*. 2005;5(1). doi:10.1186/1471-2407-5-73
321. Tawadros T, Alonso F, Jichlinski P, et al. Release of macrophage migration inhibitory factor by neuroendocrine-differentiated LNCaP cells sustains the proliferation and survival of prostate cancer cells. *Endocrine-Related Cancer*. 2012;20(1):137-149. doi:10.1530/erc-12-0286
322. Meyer-Siegler KL, Iczkowski KA, Leng L, Bucala R, Vera PL. Inhibition of macrophage migration inhibitory factor or its receptor (CD74) attenuates growth and invasion of DU-145 prostate cancer cells. *The Journal of Immunology*. 2006;177(12):8730-8739. doi:10.4049/jimmunol.177.12.8730
323. Veglia F, Perego M, Gabrilovich D. Myeloid-derived suppressor cells coming of age. *Nature Immunology*. 2018;19(2):108-119. doi:10.1038/s41590-017-0022-x
324. Gray JI, Westerhof LM, MacLeod MKL. The roles of resident, central and effector memory CD4 t-cells in protective immunity following infection or vaccination. *Immunology*. 2018;154(4):574-581. doi:10.1111/imm.12929
325. Ostrand-Rosenberg S, Sinha P, Beury DW, Clements VK. Cross-talk between myeloid-derived suppressor cells (MDSC), macrophages, and dendritic cells enhances tumor-induced immune suppression. *Seminars in Cancer Biology*. 2012;22(4):275-281. doi:10.1016/j.semcancer.2012.01.011
326. Wang G, Lu X, Dey P, et al. Targeting YAP-dependent MDSC infiltration impairs tumor progression. *Cancer Discovery*. 2015;6(1):80-95. doi:10.1158/2159-8290.cd-15-0224
327. Plessis N du, Loebenberg L, Kriel M, et al. Increased frequency of myeloid-derived suppressor cells during active tuberculosis and after RecentMycobacterium tuberculosis-Infection suppresses t-cell function. *American Journal of Respiratory and Critical Care Medicine*. 2013;188(6):724-732. doi:10.1164/rccm.201302-0249oc
328. Gabrilovich DI, Nagaraj S. Myeloid-derived suppressor cells as regulators of the immune system. *Nature Reviews Immunology*. 2009;9(3):162-174. doi:10.1038/nri2506
329. Dorhoi A, Plessis ND. Monocytic myeloid-derived suppressor cells in chronic infections. *Frontiers in Immunology*. 2018;8. doi:10.3389/fimmu.2017.01895
330. Kostic AD, Chun E, Robertson L, et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host & Microbe*. 2013;14(2):207-215. doi:10.1016/j.chom.2013.07.007
331. Su L, Xu Q, Zhang P, Michalek SM, Katz J. Phenotype and function of myeloid-derived suppressor cells induced by porphyromonas gingivalis infection. McCormick B, ed. *Infection*

- and Immunity*. 2017;85(8). doi:10.1128/iai.00213-17
332. Kumar V, Patel S, Tcyganov E, Gabrilovich DI. The nature of myeloid-derived suppressor cells in the tumor microenvironment. *Trends in Immunology*. 2016;37(3):208-220. doi:10.1016/j.it.2016.01.004
333. Murdoch C, Giannoudis A, Lewis CE. Mechanisms regulating the recruitment of macrophages into hypoxic areas of tumors and other ischemic tissues. *Blood*. 2004;104(8):2224-2234. doi:10.1182/blood-2004-03-1109
334. Qian B-Z, Li J, Zhang H, et al. CCL2 recruits inflammatory monocytes to facilitate breast-tumour metastasis. *Nature*. 2011;475(7355):222-225. doi:10.1038/nature10138
335. Simpson KD, Templeton DJ, Cross JV. Macrophage migration inhibitory factor promotes tumor growth and metastasis by inducing myeloid-derived suppressor cells in the tumor microenvironment. *The Journal of Immunology*. 2012;189(12):5533-5540. doi:10.4049/jimmunol.1201161
336. Yaddanapudi K, Putty K, Rendon BE, et al. Control of tumor-associated macrophage alternative activation by macrophage migration inhibitory factor. *The Journal of Immunology*. 2013;190(6):2984-2993. doi:10.4049/jimmunol.1201650
337. Mucha A, Drag M, Dalton JP, Kafarski P. Metallo-aminopeptidase inhibitors. *Biochimie*. 2010;92(11):1509-1529. doi:10.1016/j.biochi.2010.04.026
338. Jarrard DF. Does zinc supplementation increase the risk of prostate cancer? *Archives of Ophthalmology*. 2005;123(1):102. doi:10.1001/archophth.123.1.102
339. Schreiber CL, Smith BD. Molecular imaging of aminopeptidase n in cancer and angiogenesis. *Contrast Media & Molecular Imaging*. 2018;2018:1-15. doi:10.1155/2018/5315172
340. Weglarz-Tomczak E, Talma M, Giurg M, Westerhoff HV, Janowski R, Mucha A. Neutral metalloaminopeptidases APN and MetAP2 as newly discovered anticancer molecular targets of actinomycin d and its simple analogs. *Oncotarget*. 2018;9(50):29365-29378. doi:10.18632/oncotarget.25532
341. Zhang X, Xu W. Aminopeptidase n (APN/CD13) as a target for anti-cancer agent design. *Current Medicinal Chemistry*. 2008;15(27):2850-2865. doi:10.2174/092986708786242840
342. Wickström M, Larsson R, Nygren P, Gullbo J. Aminopeptidase n (CD13) as a target for cancer chemotherapy. *Cancer Science*. 2011;102(3):501-508. doi:10.1111/j.1349-7006.2010.01826.x
343. Prasad AS, Mantzoros CS, Beck FW, Hess JW, Brewer GJ. Zinc status and serum testosterone levels of healthy adults. *Nutrition*. 1996;12(5):344-348. doi:10.1016/s0899-9007(96)80058-x
344. CHU Q, CHI Z-H, ZHANG X, et al. A potential role for zinc transporter 7 in testosterone synthesis in mouse leydig tumor cells. *International Journal of Molecular Medicine*. 2016;37(6):1619-1626. doi:10.3892/ijmm.2016.2576
345. To PK, Do MH, Cho J-H, Jung C. Growth modulatory role of zinc in prostate cancer and application to cancer therapeutics. *International Journal of Molecular Sciences*. 2020;21(8):2991. doi:10.3390/ijms21082991

346. Amadi C, Aleme BM. The prevalence of zinc deficiency among men with and without prostate cancer in port harcourt, nigeria. *Nutrition and Cancer*. 2019;72(6):1018-1025. doi:10.1080/01635581.2019.1664600
347. Xie H, Holmes AL, Young JL, et al. Zinc chromate induces chromosome instability and DNA double strand breaks in human lung cells. *Toxicology and Applied Pharmacology*. 2009;234(3):293-299. doi:10.1016/j.taap.2008.10.010
348. Maiti S, Paira P. Biotin conjugated organic molecules and proteins for cancer therapy: A review. *European Journal of Medicinal Chemistry*. 2018;145:206-223. doi:10.1016/j.ejmech.2018.01.001
349. DAKSHINAMURTI K, CHALIFOUR L, BHULLAR RP. Requirement for biotin and the function of biotin in cells in culture. *Annals of the New York Academy of Sciences*. 1985;447(1 Biotin):38-55. doi:10.1111/j.1749-6632.1985.tb18424.x
350. Islam Z, Gurevic I, Strutzenberg TS, Ghosh AK, Iqbal T, Kohen A. Bacterial versus human thymidylate synthase: Kinetics and functionality. Roujeinikova A, ed. *PLOS ONE*. 2018;13(5):e0196506. doi:10.1371/journal.pone.0196506
351. Koehn EM, Perissinotti LL, Moghram S, et al. Folate binding site of flavin-dependent thymidylate synthase. *Proceedings of the National Academy of Sciences*. 2012;109(39):15722-15727. doi:10.1073/pnas.1206077109
352. Grogan BC, Parker JB, Guminski AF, Stivers JT. Effect of the thymidylate synthase inhibitors on dUTP and TTP pool levels and the activities of DNA repair glycosylases on uracil and 5-fluorouracil in DNA. *Biochemistry*. 2011;50(5):618-627. doi:10.1021/bi102046h
353. Christensen S, Roest BV der, Besselink N, et al. 5-fluorouracil treatment induces characteristic t-g mutations in human cancer. *Nature Communications*. 2019;10(1). doi:10.1038/s41467-019-12594-8
354. Longley DB, Harkin DP, Johnston PG. 5-fluorouracil: Mechanisms of action and clinical strategies. *Nature Reviews Cancer*. 2003;3(5):330-338. doi:10.1038/nrc1074
355. Copur S, Aiba K, Drake JC, Allegra CJ, Chu E. Thymidylate synthase gene amplification in human colon cancer cell lines resistant to 5-fluorouracil. *Biochemical Pharmacology*. 1995;49(10):1419-1426. doi:10.1016/0006-2952(95)00067-a
356. Gaurav K, Adhikary T, Satpati P. DUMP/f-dUMP binding to thymidylate synthase: Human versus mycobacterium tuberculosis. *ACS Omega*. 2020;5(28):17182-17192. doi:10.1021/acsomega.0c01224
357. Burdelski C, Strauss C, Tsourlakis MC, et al. Overexpression of thymidylate synthase (TYMS) is associated with aggressive tumor features and early PSA recurrence in prostate cancer. *Oncotarget*. 2015;6(10):8377-8387. doi:10.18632/oncotarget.3107
358. Yuan L, Zhang S, Li H, et al. The influence of gut microbiota dysbiosis to the efficacy of 5-fluorouracil treatment on colorectal cancer. *Biomedicine and Pharmacotherapy*. 2018;108:184-193. doi:10.1016/j.biopha.2018.08.165
359. Koczor CA, Torres RA, Lewis W. The role of transporters in the toxicity of nucleoside and nucleotide analogs. *Expert Opinion on Drug Metabolism and Toxicology*. 2012;8(6):665-676.

- doi:10.1517/17425255.2012.680885
360. Burdall SE, Hanby AM, Lansdown MR, Speirs V. Breast cancer cell lines: Friend or foe? *Breast Cancer Research*. 2003;5(2). doi:10.1186/bcr577
361. Wilding JL, Bodmer WF. Cancer cell lines for drug discovery and development. *Cancer Research*. 2014;74(9):2377-2384. doi:10.1158/0008-5472.can-13-2971
362. Wu C-T, Hsieh C-C, Lin C-C, Chen W-C, Hong J-H, Chen M-F. Significance of IL-6 in the transition of hormone-resistant prostate cancer and the induction of myeloid-derived suppressor cells. *Journal of Molecular Medicine*. 2012;90(11):1343-1355. doi:10.1007/s00109-012-0916-x
363. Martin RM, Donovan JL, Turner EL, et al. Effect of a low-intensity PSA-based screening intervention on prostate cancer mortality. *JAMA*. 2018;319(9):883. doi:10.1001/jama.2018.0154
364. Lane JA, Donovan JL, Davis M, et al. Active monitoring, radical prostatectomy, or radiotherapy for localised prostate cancer: Study design and diagnostic and baseline results of the ProtecT randomised phase 3 trial. *The Lancet Oncology*. 2014;15(10):1109-1118. doi:10.1016/s1470-2045(14)70361-4
365. Connell SP, Reilly EO, Tuzova A, et al. Development of a multivariable risk model integrating urinary cell DNA methylation and cell-free RNA data for the detection of significant prostate cancer. *The Prostate*. 2020;80(7):547-558. doi:10.1002/pros.23968
366. Fultang L, Panetti S, Ng M, et al. MDSC targeting with gemtuzumab ozogamicin restores t cell immunity and immunotherapy against cancers. *EBioMedicine*. 2019;47:235-246. doi:10.1016/j.ebiom.2019.08.025
367. Castaigne S, Pautas C, Terré C, et al. Effect of gemtuzumab ozogamicin on survival of adult patients with de-novo acute myeloid leukaemia (ALFA-0701): A randomised, open-label, phase 3 study. *The Lancet*. 2012;379(9825):1508-1516. doi:10.1016/s0140-6736(12)60485-1
368. Rees J, Abrahams M, Doble A, and AC. Diagnosis and treatment of chronic bacterial prostatitis and chronic prostatitis/chronic pelvic pain syndrome: A consensus guideline. *BJU International*. 2015;116(4):509-525. doi:10.1111/bju.13101
369. Stopiglia R, Ferreira U, Silva M, Matheus W, Denardi F, Reis L. Prostate specific antigen decrease and prostate cancer diagnosis: Antibiotic versus placebo prospective randomized clinical trial. *Journal of Urology*. 2010;183(3):940-945. doi:10.1016/j.juro.2009.11.044
370. Miranda KC, Bond DT, McKee M, et al. Nucleic acids within urinary exosomes/microvesicles are potential biomarkers for renal disease. *Kidney International*. 2010;78(2):191-199. doi:10.1038/ki.2010.106
371. Oksanen J, Blanchet FG, Friendly M, et al. *Vegan: Community Ecology Package*.; 2019. <https://CRAN.R-project.org/package=vegan>.
372. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multi-dimensional genomic data. *Bioinformatics*. 2016.
373. Kassambara A, Kosinski M, Biecek P. *Survminer: Drawing Survival Curves Using 'Ggplot2'*.; 2020. <https://CRAN.R-project.org/package=survminer>.

374. Therneau TM. *A Package for Survival Analysis in R*; 2020. <https://CRAN.R-project.org/package=survival>.
375. Terry M. Therneau, Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. New York: Springer; 2000.
376. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2019. <https://www.R-project.org/>.
377. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic rna-seq quantification. *Nat Biotechnol*. 2016;34(5):525-527. doi:10.1038/nbt.3519
378. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biology*. 2010;11:R106. doi:10.1186/gb-2010-11-10-r106
379. Yu G, Wang L-G, Han Y, He Q-Y. ClusterProfiler: An r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*. 2012;16(5):284-287. doi:10.1089/omi.2011.0118
380. Seemann T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-2069. doi:10.1093/bioinformatics/btu153
381. *Genomics England*. December 2018. <https://www.genomicsengland.co.uk/the-uk-has-sequenced-100000-whole-genomes-in-the-nhs/>.
382. Mark C, Jim D, Dennys M, et al. The 100,000 genomes project protocol. *genomicsengland*. November 2017. https://www.genomicsengland.co.uk/wp-content/uploads/2017/03/GenomicEnglandProtocol_151117-v4-Wales.pdf.
383. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz J L. A., Kinzler KW. Cancer genome landscapes. *Science*. 2013;339(6127):1546-1558. doi:10.1126/science.1235122
384. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014;505(7484):495-501. doi:10.1038/nature12912
385. Dijk EL van, Auger H, Jaszczyzyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends Genet*. 2014;30(9):418-426. doi:10.1016/j.tig.2014.07.001
386. Magiorkinis G, Matthews PC, Wallace SE, et al. Potential for diagnosis of infectious disease from the 100, 000 genomes project metagenomic dataset: Recommendations for reporting results. *Wellcome Open Research*. 2019;4:155. doi:10.12688/wellcomeopenres.15499.1
387. Karvelas N, Peter A, Katzenbeisser S, Tews E, Hamacher K. Privacy-preserving whole genome sequence processing through proxy-aided ORAM. In: *Proceedings of the 13th Workshop on Privacy in the Electronic Society*. ACM; 2014. doi:10.1145/2665943.2665962
388. Goig GA, Blanco S, Garcia-Basteiro AL, Comas I. Contaminant DNA in bacterial sequencing experiments is a major source of false genetic variability. *BMC Biology*. 2020;18(1). doi:10.1186/s12915-020-0748-z
389. Gray JS, Birmingham JM, Fenton JI. Got black swimming dots in your cell culture? Identification of achromobacter as a novel cell culture contaminant. *Biologicals*. 2010;38(2):273-277. doi:10.1016/j.biologicals.2009.09.006

390. Flemer B, Lynch DB, Brown JMR, et al. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut*. 2016;66(4):633-643. doi:10.1136/gutjnl-2015-309595
391. Chattopadhyay I, Verma M, Panda M. Role of oral microbiome signatures in diagnosis and prognosis of oral cancer. *Technology in Cancer Research & Treatment*. 2019;18:153303381986735. doi:10.1177/1533033819867354
392. Paul R, Lorenzl S, Koedel U, et al. Matrix metalloproteinases contribute to the blood-brain barrier disruption during bacterial meningitis. *Annals of Neurology*. 1998;44(4):592-600. doi:10.1002/ana.410440404
393. Raghavan M, Linden PK. Newer treatment options for skin and soft tissue infections. *Drugs*. 2004;64(15):1621-1642. doi:10.2165/00003495-200464150-00002
394. S. Hansen SA K. K. Melby. Helicobacter pylori infection and risk of cardia cancer and non-cardia gastric cancer: A nested case-control study. *Scandinavian Journal of Gastroenterology*. 1999;34(4):353-360. doi:10.1080/003655299750026353
395. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nature Biotechnology*. 2011;29(1):24-26. doi:10.1038/nbt.1754
396. Boedigheimer MJ, Wolfinger RD, Bass MB, et al. Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. *BMC Genomics*. 2008;9(1):285. doi:10.1186/1471-2164-9-285
397. Scherer A, ed. *Batch Effects and Noise in Microarray Experiments*. John Wiley & Sons, Ltd; 2009. doi:10.1002/9780470685983
398. Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*. 2014;15(2):R29. doi:10.1186/gb-2014-15-2-r29
399. Walsh R, Fauze KL, Crockford M, Jones B, Owens L. Genomic heterogeneity and prevalence of hepadensovirus in penaeus esculentus from western australia, and p. Merguiensis from the gulf of carpentaria, australia. *Aquaculture*. 2017;471:43-48. doi:10.1016/j.aquaculture.2017.01.006
400. Mecham BH, Nelson PS, Storey JD. Supervised normalization of microarrays. *Bioinformatics*. 2010;26(10):1308-1315. doi:10.1093/bioinformatics/btq118
401. Lunardon N, Menardi G, Torelli N. ROSE: A Package for Binary Imbalanced Learning. *R Journal*. 2014;6(1):82-92.
402. Bland JA, Brock TD. The marine bacterium leucothrix mucor as an algal epiphyte. *Marine Biology*. 1973;23(4):283-292. doi:10.1007/bf00389335
403. Yu H, Qi S, Chang Z, Rong Q, Akinyemi IA, Wu Q. Complete genome sequence of a novel velarivirus infecting areca palm in china. *Archives of Virology*. 2015;160(9):2367-2370. doi:10.1007/s00705-015-2489-9
404. Rabenstein F, Schubert J, Seifers DL, Stenger DC, French R. Phylogenetic relationships, strain diversity and biogeography of tritimonoviruses. *Journal of General Virology*. 2002;83(4):895-906. doi:10.1099/0022-1317-83-4-895

405. Zhai Y, Attoui H, Jaafar FM, et al. Isolation and full-length sequence analysis of armigeres subalbatus totivirus, the first totivirus isolate from mosquitoes representing a proposed novel genus (artivirus) of the family totiviridae. *Journal of General Virology*. 2010;91(11):2836-2845. doi:10.1099/vir.0.024794-0
406. Short SM, Staniewski MA, Chaban YV, Long AM, Wang D. Diversity of viruses infecting eukaryotic algae. *Current Issues in Molecular Biology*. 2020;29-62. doi:10.21775/cimb.039.029
407. Webster DE, Beck DL, Rabenstein F, Forster RLS, Guy PL. An improved polyclonal antiserum for detecting ryegrass mosaic rymovirus. *Archives of Virology*. 2005;150(9):1921-1926. doi:10.1007/s00705-005-0531-z
408. Paper W, Jahn U, Hohn MJ, et al. Ignicoccus hospitalis sp. Nov., the host of “nanoarchaeum equitans”. *International Journal of Systematic and Evolutionary Microbiology*. 2007;57(4):803-808. doi:10.1099/ijs.0.64721-0
409. Nedashkovskaya OI, Vancanneyt M, Kim SB, Han J, Zhukova NV, Shevchenko LS. Salinimicrobium marinum sp. Nov., a halophilic bacterium of the family flavobacteriaceae, and emended descriptions of the genus salinimicrobium and salinimicrobium catena. *International Journal of Systematic and Evolutionary Microbiology*. 2010;60(10):2303-2306. doi:10.1099/ijs.0.019166-0
410. Thompson FL, Barash Y, Sawabe T, Sharon G, Swings J, Rosenberg E. Thalassomonas loyana sp. Nov., a causative agent of the white plague-like disease of corals on the eilat coral reef. *International Journal of Systematic and Evolutionary Microbiology*. 2006;56(2):365-368. doi:10.1099/ijs.0.63800-0
411. Epidemiology, treatment, and prevention of human t-cell leukemia virus type 1-associated diseases. *Clinical Microbiology Reviews*. 2010;23(3):577-589. doi:10.1128/cmr.00063-09
412. Willems L, Hasegawa H, Accolla R, et al. Reducing the global burden of HTLV-1 infection: An agenda for research and action. *Antiviral Research*. 2017;137:41-48. doi:10.1016/j.antiviral.2016.10.015
413. Klastersky J, Aoun M. Opportunistic infections in patients with cancer. *Annals of Oncology*. 2004;15:iv329-iv335. doi:10.1093/annonc/mdh947
414. Walther-António MRS, Chen J, Multinu F, et al. Potential contribution of the uterine microbiome in the development of endometrial cancer. *Genome Medicine*. 2016;8(1). doi:10.1186/s13073-016-0368-y
415. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*. 2015;31(10):1674-1676. doi:10.1093/bioinformatics/btv033
416. Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22(13):1658-1659. doi:10.1093/bioinformatics/btl158
417. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using diamond. *Nature Methods*. 2015;12(1):59-60.

418. Ayling M, Clark MD, Leggett RM. New approaches for metagenome assembly with short reads. *Briefings in Bioinformatics*. 2019;21(2):584-594. doi:10.1093/bib/bbz020
419. 'Sample handling guidance version 4'. *Genomicsengland.co.uk*. 2020. <https://www.genomicsengland.co.uk/wp-content/uploads/2018/02/Sample-Handling-Guidance-v4.0.pdf>.
420. 'TruSeq nano dna library prep'. *Illumina*. 2015. https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqnanodna/truseq-nano-dna-library-prep-guide-15041110-d.pdf.
421. Robbe P, Niko Popitsch, Knight SJL, et al. Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: Pilot study for the 100, 000 genomes project. *Genetics in Medicine*. 2018;20(10):1196-1205. doi:10.1038/gim.2017.241
422. Raczy C, Petrovski R, Saunders CT, et al. Isaac: Ultra-fast whole-genome secondary analysis on illumina sequencing platforms. *Bioinformatics*. 2013;29(16):2041-2043. doi:10.1093/bioinformatics/btt314
423. Hahsler M, Piekenbrock M, Doran D. dbscan: Fast density-based clustering with R. *Journal of Statistical Software*. 2019;91(1):1-30. doi:10.18637/jss.v091.i01
424. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 2015;43(7):e47. doi:10.1093/nar/gkv007
425. Kuhn M. *Caret: Classification and Regression Training*; 2020. <https://CRAN.R-project.org/package=caret>.
426. Borchmann S. An atlas of the tissue and blood metagenome in cancer reveals novel links between bacteria, viruses and cancer. September 2019. doi:10.1101/773200
427. Feng Q, Liang S, Jia H, et al. Gut microbiome development along the colorectal adenoma-macarcinoma sequence. *Nature Communications*. 2015;6(1). doi:10.1038/ncomms7528
428. Jenkinson HF. Beyond the oral microbiome. *Environmental Microbiology*. 2011;13(12):3077-3087. doi:10.1111/j.1462-2920.2011.02573.x
429. Russell DG. Mycobacterium tuberculosis: Here today, and here tomorrow. *Nature Reviews Molecular Cell Biology*. 2001;2(8):569-578. doi:10.1038/35085034
430. Mukhopadhyay D, Arranz-Solís D, Saeij JPJ. Influence of the host and parasite strain on the immune response during toxoplasma infection. *Frontiers in Cellular and Infection Microbiology*. 2020;10. doi:10.3389/fcimb.2020.580425
431. *Micronoma*. <https://micronoma.com/331-2/>.
432. Ma X, Chi C, Fan L, et al. The microbiome of prostate fluid is associated with prostate cancer. *Frontiers in Microbiology*. 2019;10. doi:10.3389/fmicb.2019.01664
433. Sfanos KS, Yegnasubramanian S, Nelson WG, Marzo AMD. The inflammatory microenvironment and microbiome in prostate cancer development. *Nature Reviews Urology*. 2017;15(1):11-24. doi:10.1038/nrurol.2017.167

434. Sfanos KS, Marzo AMD. Prostate cancer and inflammation: The evidence. *Histopathology*. 2011;60(1):199-215. doi:10.1111/j.1365-2559.2011.04033.x
435. Sfanos KS, Sauvageot J, Fedor HL, Dick JD, Marzo AMD, Isaacs WB. A molecular analysis of prokaryotic and viral DNA sequences in prostate tissue from patients with prostate cancer indicates the presence of multiple and diverse microorganisms. *The Prostate*. 2008;68(3):306-320. doi:10.1002/pros.20680
436. Mitchell RA, Yaddanapudi K. Stromal-dependent tumor promotion by MIF family members. *Cellular Signalling*. 2014;26(12):2969-2978. doi:10.1016/j.cellsig.2014.09.012
437. <McElroy et al 2012 - gemsim- general, error-model based simulator of next-generation sequencing data.pdf>.
438. Gourel H, Karlsson-Lindsjö O, Hayer J, Bongcam-Rudloff E. Simulating illumina metagenomic data with InSilicoSeq. Hancock J, ed. *Bioinformatics*. 2018;35(3):521-522. doi:10.1093/bioinformatics/bty630
439. Gutierrez JC, Perez EA, Franceschi D, Moffat FL, Livingstone AS, Koniaris LG. Outcomes for soft-tissue sarcoma in 8249 cases from a large state cancer registry. *Journal of Surgical Research*. 2007;141(1):105-114. doi:10.1016/j.jss.2007.02.026
440. Charalampous T, Kay GL, Richardson H, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nature Biotechnology*. 2019;37(7):783-792. doi:10.1038/s41587-019-0156-5
441. Sangar VK, Ragavan N, Matanhelia SS, Watson MW, Blades RA. The economic consequences of prostate and bladder cancer in the UK. *BJU International*. 2005;95(1):59-63. doi:10.1111/j.1464-410x.2005.05249.x
442. Roehrborn CG, Black LK. The economic burden of prostate cancer. *BJU International*. 2011;108(6):806-813. doi:10.1111/j.1464-410x.2011.10365.x
443. Brown C, Irber L. Sourmash: A library for minhash sketching of dna. *J Open Source Softw*. 2016;1(1). doi:10.1101/029827
444. Ondov BD, Treangen TJ, Melsted P, et al. Mash: Fast genome and metagenome distance estimation using minhash. *Genome Biol*. 2016;17(1):132. doi:10.1186/s13059-016-0997-x
445. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine*. 2015;162(1):W1. doi:10.7326/m14-0698
446. Schmidt TSB, Rodrigues JFM, Mering C von. A family of interaction-adjusted indices of community similarity. *The ISME Journal*. 2016;11(3):791-807. doi:10.1038/ismej.2016.139
447. Barwell LJ, Isaac NJB, Kunin WE. Measuring beta-diversity with species abundance data. Coulson T, ed. *Journal of Animal Ecology*. 2015;84(4):1112-1122. doi:10.1111/1365-2656.12362
448. Wu YW, Ye Y. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J Comput Biol*. 2011;18(3):523-534. doi:10.1089/cmb.2010.0245

449. Vinh le V, Lang TV, Binh le T, Hoai TV. A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads. *Algorithms Mol Biol.* 2015;10(1):2. doi:10.1186/s13015-014-0030-4
450. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schaffer AA. Database indexing for production megablast searches. *Bioinformatics.* 2008;24(16):1757-1764. doi:10.1093/bioinformatics/btn322
451. Krause L, Diaz NN, Goesmann A, et al. Phylogenetic classification of short environmental dna fragments. *Nucleic Acids Res.* 2008;36(7):2230-2239. doi:10.1093/nar/gkn038
452. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: Fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics.* 2015;16:236. doi:10.1186/s12864-015-1419-2
453. Chatterji S, Yamazaki I, Bai Z, Eisen J. CompostBin: A dna composition-based algorithm for binning environmental shotgun reads. In: Vingron M, Wong L, eds. *Research in Computational Molecular Biology.* Vol 4955. Berlin, Heidelberg: Springer; 2008.
454. Alneberg J, Bjarnason BS, Bruijn I de, et al. Binning metagenomic contigs by coverage and composition. *Nat Methods.* 2014;11(11):1144-1146. doi:10.1038/nmeth.3103
455. Lu YY, Chen T, Fuhrman JA, Sun F. COCACOLA: Binning metagenomic contigs using sequence composition, read coverage, co-alignment and paired-end read linkage. *Bioinformatics.* 2017;33(6):791-798. doi:10.1093/bioinformatics/btw290
456. Kawulok J, Deorowicz S. CoMeta: Classification of metagenomes using k-mers. *PLoS One.* 2015;10(4):e0121453. doi:10.1371/journal.pone.0121453
457. Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. ConStrains identifies microbial strains in metagenomic datasets. *Nat Biotechnol.* 2015;33(10):1045-1052. doi:10.1038/nbt.3319
458. Ghosh T, Haque M, Mande S. DiScRIBinATE:- a rapid method for accurate taxonomic classification of metagenomic sequences. *BMC Bioinformatics.* 2010;11(7).
459. Piro VC, Lindner MS, Renard BY. DUDes: A top-down taxonomic profiler for metagenomics. *Bioinformatics.* 2016;32(15):2272-2280.
460. Hunter S, Corbett M, Denise H, et al. EBI metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.* 2014;42(Database issue):D600-6. doi:10.1093/nar/gkt961
461. Stranneheim H, Kaller M, Allander T, Andersson B, Arvestad L, Lundeborg J. Classification of dna sequences using bloom filters. *Bioinformatics.* 2010;26(13):1595-1600. doi:10.1093/bioinformatics/btq230
462. Silva GG, Cuevas DA, Dutilh BE, Edwards RA. FOCUS: An alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ.* 2014;2:e425. doi:10.7717/peerj.425
463. Angly FE, Willner D, Prieto-Davo A, et al. The gaas metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol.* 2009;5(12):e1000593. doi:10.1371/journal.pcbi.1000593

464. Lindner MS, Renard BY. Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Res.* 2013;41(1):e10. doi:10.1093/nar/gks803
465. Davenport CF, Neugebauer J, Beckmann N, et al. Genometa—a fast and accurate classifier for short metagenomic shotgun reads. *PLoS One.* 2012;7(8):e41224. doi:10.1371/journal.pone.0041224
466. Xia LC, Cram JA, Chen T, Fuhrman JA, Sun F. Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One.* 2011;6(12):e27992. doi:10.1371/journal.pone.0027992
467. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: An automated tool for the recovery of population genomes from related metagenomes. *PeerJ.* 2014;2:e603. doi:10.7717/peerj.603
468. Mohammed MH, Ghosh TS, Reddy RM, Reddy CV, Singh NK, Mande SS. INDUS - a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. *BMC Genomics.* 2011;12 Suppl 3:S4. doi:10.1186/1471-2164-12-S3-S4
469. Liu Z, Bensmail H, Tan M. Efficient feature selection and multiclass classification with integrated instance and model based learning. *Evol Bioinform Online.* 2012;8:197-205. doi:10.4137/EBO.S9407
470. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. *Genome Biology.* 2019;20(1). doi:10.1186/s13059-019-1891-0
471. Ainsworth D, Sternberg MJE, Racz C, Butcher SA. K-slam: Accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. *Nucleic Acids Res.* 2017;45(4):1649-1656. doi:10.1093/nar/gkw1248
472. Kislyuk A, Bhatnagar S, Dushoff J, Weitz JS. Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics.* 2009;10:316. doi:10.1186/1471-2105-10-316
473. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics.* 2013;29(18):2253-2260. doi:10.1093/bioinformatics/btt389
474. Wu Y, Simmons B, Singer S. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics.* 2016;32(4):605-607. doi:https://doi.org/10.1093/bioinformatics/btv638
475. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res.* 2007;17(3):377-386. doi:10.1101/gr.5969107
476. Huson DH, Beier S, Flade I, et al. MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput Biol.* 2016;12(6):e1004957. doi:10.1371/journal.pcbi.1004957
477. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ.* 2015;3:e1165. doi:10.7717/peerj.1165
478. Jia P, Xuan L, Liu L, Wei C. MetaBinG - using gpus to accelerate metagenomic sequence classification. *PLoS ONE.* 2011;6(11). doi:10.1371/journal.pone.0025353.t001

479. Wang Y, Leung HC, Yiu SM, Chin FY. MetaCluster 4.0: A novel binning algorithm for ngs reads and huge number of species. *J Comput Biol.* 2012;19(2):241-249. doi:10.1089/cmb.2011.0276
480. Wang Y, Leung HC, Yiu SM, Chin FY. MetaCluster 5.0: A two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics.* 2012;28(18):i356-i362. doi:10.1093/bioinformatics/bts397
481. Zhang Y, Sun Y. METADOMAIN: A profile hmm-based protein domain classification tool for short sequences. *Biocomputing 2012.* 2011:271-282.
482. Zolfo M, Tett A, Jousson O, Donati C, Segata N. MetaMLST: Multi-locus strain-level bacterial typing from metagenomic samples. *Nucleic Acids Res.* 2017;45(2):e7. doi:10.1093/nar/gkw837
483. Liu B, Gibbons T, Ghodsi M, Pop M. MetaPhyler: Taxonomic profiling for metagenomic sequences. In: *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* IEEE; 2010. doi:10.1109/bibm.2010.5706544
484. Girotto S, Pizzi C, Comin M. MetaProb: Accurate metagenomic reads binning based on probabilistic sequence signatures. *Bioinformatics.* 2016;32(17):i567-i575. doi:10.1093/bioinformatics/btw466
485. Strous M, Kraft B, Bisdorf R, Tegetmeyer HE. The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front Microbiol.* 2012;3:410. doi:10.3389/fmicb.2012.00410
486. Bengtsson-Palme J, Hartmann M, Eriksson KM, et al. METAXA2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol Ecol Resour.* 2015;15(6):1403-1414. doi:10.1111/1755-0998.12399
487. Wilke A, Bischof J, Gerlach W, et al. The mg-rast metagenomics database and portal in 2015. *Nucleic Acids Res.* 2016;44(D1):D590-4. doi:10.1093/nar/gkv1322
488. Sunagawa S, Mende DR, Zeller G, et al. Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods.* 2013;10(12):1196-1199. doi:10.1038/nmeth.2693
489. Ahn TH, Chai J, Pan C. Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics.* 2015;31(2):170-177. doi:10.1093/bioinformatics/btu641
490. Gori F, Folino G, Jetten MS, Marchiori E. MTR: Taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks. *Bioinformatics.* 2011;27(2):196-203. doi:10.1093/bioinformatics/btq649
491. Lin HH, Liao YC. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci Rep.* 2016;6:24175. doi:10.1038/srep24175
492. Rosen GL, Reichenberger ER, Rosenfeld AM. NBC: The naive bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics.* 2011;27(1):127-129. doi:10.1093/bioinformatics/btq619
493. Minot SS, Krumm N, Greenfield NB. One codex: A sensitive and accurate data platform for genomic microbial identification. *BioRxIV.* 2015. doi:10.1101/027607

494. Scholz M, Ward DV, Pasolli E, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods*. 2016;13(5):435-438. doi:10.1038/nmeth.3802
495. Berger SA, Stamatakis A. Aligning short reads to reference alignments and trees. *Bioinformatics*. 2011;27(15):2068-2075. doi:10.1093/bioinformatics/btr320
496. Jing G, Sun Z, Wang H, et al. Parallel-meta 3: Comprehensive taxonomical and functional analysis platform for efficient comparison of microbial communities. *Sci Rep*. 2017;7:40371. doi:10.1038/srep40371
497. 2018. <https://github.com/ChandraPedamallu/PathSeq>.
498. Francis OE, Bendall M, Manimaran S, et al. Pathoscope: Species identification and strain attribution with unassembled sequencing data. *Genome Res*. 2013;23(10):1721-1729. doi:10.1101/gr.150151.112
499. Zheng H, Wu H. Short prokaryotic dna fragment binning using a hierarchical classifier based on linear discriminant analysis and principal component analysis. *J Bioinform Comput Biol*. 2010;8(6):995-1011.
500. Gregor I, Droge J, Schirmer M, Quince C, McHardy AC. PhyloPythiaS+: A self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ*. 2016;4:e1603. doi:10.7717/peerj.1603
501. Brady A, Salzberg SL. Phymm and phymmbl: Metagenomic phylogenetic classification with interpolated markov models. *Nat Methods*. 2009;6(9):673-676. doi:10.1038/nmeth.1358
502. Brinda K, Salikhov K, Pignotti S, Kuchеров G. ProPhyle 0.3.1.0. *Zenodo*. 2017. doi:10.5281/zenodo.1045429
503. Ghosh TS, Monzoorul H, Komanduri D, Mande S. ProViDE: A software tool for accurate estimation of viral diversity in metagenomic samples. *Bioinformatics*. 2011;6(2):91-94.
504. Koslicki D, Foucart S, Rosen G. Quikr: A method for rapid reconstruction of bacterial communities via compressive sensing. *Bioinformatics*. 2013;29(17):2096-2102. doi:10.1093/bioinformatics/btt336
505. Nalbantoglu O, Way S, Hinrichs S, Sayood K. RAIphy: Phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinformatics*. 2011;12(41).
506. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312-1313. doi:10.1093/bioinformatics/btu033
507. Scheuch M, Hoper D, Beer M. RIEMS: A software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. *BMC Bioinformatics*. 2015;16:69. doi:10.1186/s12859-015-0503-6
508. Bhaduri A, Qu K, Lee C, Ungewickell A, Khavari P. Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics*. 2012;28(8):1174-1175. doi:10.1093/bioinformatics/bts100

509. MacDonald NJ, Parks DH, Beiko RG. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Res.* 2012;40(14):e111. doi:10.1093/nar/gks335
510. Kelley DR, Salzberg SL. Clustering metagenomic sequences with interpolated markov models. *BMC Bioinformatics.* 2010;11:544. doi:10.1186/1471-2105-11-544
511. Szalkai B, Grolmusz V. Near perfect protein multi-label classification with deep neural networks. *Methods.* 2018;132:50-56. doi:10.1016/j.ymeth.2017.06.034
512. Le VV, Tran LV, Tran HV. A novel semi-supervised algorithm for the taxonomic assignment of metagenomic reads. *BMC Bioinformatics.* 2016;17:22. doi:10.1186/s12859-015-0872-x
513. Chan CK, Hsu AL, Halgamuge SK, Tang SL. Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics.* 2008;9:215. doi:10.1186/1471-2105-9-215
514. Dadi TH, Renard BY, Wieler LH, Semmler T, Reinert K. SLIMM: Species level identification of microorganisms from metagenomes. *PeerJ.* 2017;5:e3138. doi:10.7717/peerj.3138
515. Lee AY, Lee CS, Van Gelder RN. Scalable metagenomics alignment research tool (smart): A scalable, rapid, and complete search heuristic for the classification of metagenomic sequences from complex sequence populations. *BMC Bioinformatics.* 2016;17:292. doi:10.1186/s12859-016-1159-6
516. Monzoorul Haque M, Ghosh TS, Komanduri D, Mande SS. SOrt-items: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics.* 2009;25(14):1722-1730. doi:10.1093/bioinformatics/btp317
517. Mohammed M, Ghosh TS, Singh N, Mande SS. SPHINX — an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics.* 2011;27(1):22-30.
518. Inouye M, Dashnow H, Raven L, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* 2014;6(11):90. doi:10.1186/s13073-014-0090-6
519. Roosaare M, Vaher M, Kaplinski L, et al. StrainSeeker: Fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. *PeerJ.* 2017;5:e3353. doi:10.7717/peerj.3353
520. Naccache S, Federman S, Veeraraghavan N, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res.* 2014;24(7):1180-1192. doi:10.1101/gr.171934.113
521. Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW. TACOA: Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics.* 2009;10:56. doi:10.1186/1471-2105-10-56
522. Sohn M, An L, Pookhao N, Li Q. Accurate genome relative abundance estimation for closely related species in a metagenomic sample. *BMC Bioinformatics.* 2014;15(1):242.

523. Weber M, Teeling H, Huang S, et al. Practical application of self-organizing maps to inter-relate biodiversity and functional data in ngs-based metagenomics. *ISME J.* 2011;5(5):918-928. doi:10.1038/ismej.2010.180
524. Droge J, Gregor I, McHardy AC. Taxator-tk: Precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics.* 2015;31(6):817-824. doi:10.1093/bioinformatics/btu745
525. Pongor LS, Vera R, Ligeti B. Fast and sensitive alignment of microbial whole genome sequencing reads to large sequence datasets on a desktop pc: Application to metagenomic datasets and pathogen identification. *PLoS One.* 2014;9(7):e103441. doi:10.1371/journal.pone.0103441
526. Flygare S, Simmon K, Miller C, et al. Taxonomer: An interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol.* 2016;17(1):111. doi:10.1186/s13059-016-0969-1
527. Klingenberg H, Asshauer KP, Lingner T, Meinicke P. Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. *Bioinformatics.* 2013;29(8):973-980. doi:10.1093/bioinformatics/btt077
528. Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO. TETRA: A web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics.* 2004;5:163. doi:10.1186/1471-2105-5-163
529. Nguyen NP, Mirarab S, Liu B, Pop M, Warnow T. TIPP: Taxonomic identification and phylogenetic profiling. *Bioinformatics.* 2014;30(24):3548-3555. doi:10.1093/bioinformatics/btu721
530. Schreiber F, Gumrich P, Daniel R, Meinicke P. Treephyler: Fast taxonomic profiling of metagenomes. *Bioinformatics.* 2010;26(7):960-961. doi:10.1093/bioinformatics/btq070
531. Reddy RM, Mohammed MH, Mande SS. TWARIT: An extremely rapid and efficient approach for phylogenetic classification of metagenomic sequences. *Gene.* 2012;505(2):259-265. doi:10.1016/j.gene.2012.06.014
532. Laczny CC, Sternal T, Plugaru V, et al. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome.* 2015;3(1):1. doi:10.1186/s40168-014-0066-1
533. Juul S, Izquierdo F, Hurst A, et al. What's in my pot? Real-time species identification on the minion. *BioRxIV.* 2015. doi:10.1101/030742
534. Hou T, Liu F, Liu Y, Zou QY, Zhang X, Wang K. Classification of metagenomics data at lower taxonomic level using a robust supervised classifier. *Evol Bioinform Online.* 2015;11:3-10. doi:10.4137/EBO.S20523
535. Nielsen HB, Almeida M, Juncker AS, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol.* 2014;32(8):822-828. doi:10.1038/nbt.2939
536. Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-ud: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics.* 2012;28(11):1420-1428. doi:10.1093/bioinformatics/bts174

537. Cleary B, Brito IL, Huang K, et al. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat Biotechnol.* 2015;33(10):1053-1060. doi:10.1038/nbt.3329
538. Wences A, Schatz M. Metassembler: Merging and optimizing de novo genome assemblies. *Genome Biol.* 2015;16:207. doi:10.1186/s13059-015-0764-4
539. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: An extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 2012;40(20):e155. doi:10.1093/nar/gks678
540. Chikhi R, Rizk G. Space-efficient and exact de bruijn graph representation based on a bloom filter. *Algorithms for Molecular Biology.* 2013;8(22).
541. Haider B, Ahn TH, Bushnell B, Chai J, Copeland A, Pan C. Omega: An overlap-graph de novo assembler for metagenomics. *Bioinformatics.* 2014;30(19):2717-2722. doi:10.1093/bioinformatics/btu395
542. Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J. Ray meta: Scalable de novo metagenome assembly and profiling. *Genome Biol.* 2012;13(R122).
543. Luo R, Liu B, Xie Y, et al. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience.* 2012;1(18).
544. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics.* 2017;3(10). doi:10.1099/mgen.0.000132
545. Audirac-Chalifour A, Torres-Poveda K, Bahena-Román M, et al. Cervical microbiome and cytokine profile at various stages of cervical cancer: A pilot study. Tornesello ML, ed. *PLOS ONE.* 2016;11(4):e0153274. doi:10.1371/journal.pone.0153274
546. Flores V, Sepúlveda-Robles O, Cazares A, Kameyama L, Guarneros G. Comparative genomic analysis of pseudomonas aeruginosa phage PaMx25 reveals a novel siphovirus group related to phages infecting hosts of different taxonomic classes. *Archives of Virology.* 2017;162(8):2345-2355. doi:10.1007/s00705-017-3366-5
547. Gich F, Overmann J. Sandarakinorhabdus limnophila gen. Nov., sp. Nov., a novel bacteriochlorophyll a-containing, obligately aerobic bacterium isolated from freshwater lakes. *International Journal of Systematic and Evolutionary Microbiology.* 2006;56(4):847-854. doi:10.1099/ijs.0.63970-0
548. Cai H, Cui H, Zeng Y, An M, Jiang H. Sandarakinorhabdus cyanobacteriorum sp. Nov., a novel bacterium isolated from cyanobacterial aggregates in a eutrophic lake. *International Journal of Systematic and Evolutionary Microbiology.* 2018;68(3):730-735. doi:10.1099/ijsem.0.002571
549. Liu Y, Yao T, Jiao N, et al. Microbial diversity in the snow, a moraine lake and a stream in himalayan glacier. *Extremophiles.* 2011;15(3):411-421. doi:10.1007/s00792-011-0372-5
550. Guerrini S, Mangani S, Granchi L, Vincenzini M. Biogenic amine production by oenococcus oeni. *Current Microbiology.* 2002;44(5):374-378. doi:10.1007/s00284-001-0021-9

551. Hernández T, Estrella I, Pérez-Gordo M, Alegria EG, Tenorio C, Moreno-Arribas MV. Contribution of malolactic fermentation by *Oenococcus Oeni* and *Lactobacillus plantarum* to the changes in the nonanthocyanin polyphenolic composition of red wine. *Journal of Agricultural and Food Chemistry*. 2007;55(13):5260-5266. doi:10.1021/jf063638o
552. Tiago I, Morais PV, Costa MS da, Veri'ssimo A. *Microcella alkaliphila* sp. Nov., a novel member of the family microbacteriaceae isolated from a non-saline alkaline groundwater, and emended description of the genus *microcella*. *International Journal of Systematic and Evolutionary Microbiology*. 2006;56(10):2313-2316. doi:10.1099/ijs.0.64320-0
553. Velásquez AMA, Francisco AI, Kohatsu AAN, et al. Synthesis and tripanocidal activity of ferrocenyl and benzyl diamines against *trypanosoma brucei* and *trypanosoma cruzi*. *Bioorganic & Medicinal Chemistry Letters*. 2014;24(7):1707-1710. doi:10.1016/j.bmcl.2014.02.046
554. Ubillos L, Freire T, Berriel E, et al. *Trypanosoma cruzi* extracts elicit protective immune response against chemically induced colon and mammary cancers. *International Journal of Cancer*. 2015;138(7):1719-1731. doi:10.1002/ijc.29910
555. Ramírez-Tolosa G, Abello P, Ferreira A. Is the antitumor property of *trypanosoma cruzi* infection mediated by its calreticulin? *Frontiers in Immunology*. 2016;7. doi:10.3389/fimmu.2016.00268