Metagenomics and transcriptomics data from human colorectal cancer

Tina Visnovska1,2, Patrick J. Biggs3, Sebastian Schmeier1, Frank A. Frizelle4 & Rachel V. Purcell4

Colorectal cancer is a heterogenous and mostly sporadic disease, the development of which is associated with microbial dysbiosis. Recent advances in subtype classification have successfully stratified the disease using molecular profiling. To understand potential relationships between molecular mechanisms differentiating the subtypes of colorectal cancer and composition of gut microbial community, we classified a set of 34 tumour samples into molecular subtypes using RNA-sequencing gene expression profiles and determined relative abundances of bacterial taxonomic groups. To identify bacterial community composition, 16S rRNA amplicon metabarcoding was used as well as whole genome metagenomics of the non-human part of RNA-sequencing data. The generated data expands the collection of the data sources related to the disease and connects molecular aspects of the cancer with environmental impact of microbial community.

Background & Summary
Colorectal cancer (CRC) is one of the most common types of cancer worldwide, in terms of both incidence and mortality1. Most cases of CRC are sporadic with no known genetic link. Environmental factors are therefore likely to play a critical role in the development of the disease, and a key characteristic of the colon is that it houses the largest proportion of the human microbiome, suggesting that this might play a role in causing CRC. Recent data points to the importance of the microbial communities in the gut, the microbiome, and possible links to the development of CRC2–5. If this is the case, understanding the role of the microbiome in CRC will have profound effects on cancer rates, since it is potentially relatively easily to manipulate, using diet, pre- and probiotics and faecal transplants6–9. However, despite the intense interest in the field and increasing evidence pointing to a role for the microbiome in CRC, convincing connections with clinical parameters and outcome are rarely seen.

CRC is a highly heterogeneous disease, with varying clinical outcomes, response to therapy, and morphological features, and molecular subtyping systems based on CpG-island methylation, microsatellite instability and gene mutations have shown strong associations with outcome and response to therapy in CRC10–13. Contrary to other microbiome studies, where CRC is treated as a single disease entity, we focused on the association between Consensus Molecular Subtypes (CMS) of colorectal cancer and gut microbiome patterns in the accompanying primary publication14. We stratified a set of CRC tumour samples into CMS according to their gene expression profiles15 and assessed differences in bacterial communities among CMS. The gene expression profiles were generated using RNA sequencing, and 16S rRNA metabarcoding as well as metagenomic analysis of non-human portion of the RNA sequencing data were employed for bacterial taxa quantification. We analysed the enrichment/depletion of bacterial species in one subtype compared to the other subtypes and showed enrichment of certain oral bacteria associated with CMS, which was validated using targeted quantitative PCR.

The data generated in this study combine various views of each sample as multiple different methods were used to obtain information about the samples. This allows us to study associations between the results of the particular methods. Making the raw sequencing data available together with the scripts used for data processing and analysis, we enable reuse of the data and extend the collection of the data sources related to CRC, for which the aetiology is not yet well understood.

1School of Natural and Computational Sciences, Massey University, Auckland, New Zealand. 2Present address: Bioinformatics Core Facility, Oslo University Hospital Radium, Oslo, Norway. 3School of Fundamental Sciences, Massey University, Palmerston North, New Zealand. 4Department of Surgery, University of Otago, Christchurch, New Zealand. Correspondence and requests for materials should be addressed to T.V. (email: tina.visnovska@rr-research.no)
**Fig. 1** Workflow of sample and data processing. Samples and data are shown in grey and processes highlighted in blue.

| SampleID | CMS | Age | Gender | Site | Side | Stage |
|----------|-----|-----|--------|------|------|-------|
| CRC_01   | CMS2| 73  | M      | Colon| Left | 1     |
| CRC_02   | U/C | 62  | F      | Colon| Left | 3     |
| CRC_03   | U/C | 76  | F      | Colon| Right| 2     |
| CRC_04   | CMS1| 88  | F      | Colon| Right| 2     |
| CRC_05   | U/C | 68  | F      | Colon| Right| 3     |
| CRC_06   | CMS3| 63  | M      | Colon| Right| 1     |
| CRC_07   | CMS2| 81  | F      | Colon| Left | 2     |
| CRC_08   | CMS2| 74  | M      | Colon| Right| 3     |
| CRC_09   | CMS1| 83  | F      | Colon| Right| 2     |
| CRC_10   | CMS2| 81  | M      | Colon| Left | 1     |
| CRC_11   | CMS3| 79  | F      | Colon| Left | 3     |
| CRC_12   | CMS3| 79  | F      | Colon| Right| 1     |
| CRC_13   | CMS2| 74  | F      | Colon| Right| 2     |
| CRC_14   | U/C | 83  | M      | Colon| Left | 2     |
| CRC_15   | CMS3| 77  | F      | Colon| Right| 3     |
| CRC_16   | CMS3| 84  | F      | Colon| Right| 3     |
| CRC_17   | U/C | 77  | M      | Colon| Left | 3     |
| CRC_18   | CMS2| 58  | M      | Colon| Right| 3     |
| CRC_19   | CMS2| 77  | M      | Colon| Left | 2     |
| CRC_20   | CMS3| 74  | M      | Colon| Right| 2     |
| CRC_21   | CMS1| 75  | F      | Colon| Right| 2     |
| CRC_22   | CMS3| 78  | F      | Rectum| N/A| 3     |
| CRC_23   | CMS2| 78  | F      | Colon| Left | 2     |
| CRC_24   | CMS3| 45  | F      | Colon| Right| 1     |
| CRC_25   | CMS1| 78  | F      | Colon| Right| 2     |
| CRC_26   | U/C | 67  | M      | Colon| Right| 3     |
| CRC_27   | CMS1| 75  | F      | Colon| Right| 3     |
| CRC_28   | CMS3| 78  | M      | Colon| Left | 1     |
| CRC_29   | CMS2| 67  | M      | Colon| Right| 2     |
| CRC_30   | CMS2| 80  | M      | Colon| Left | 2     |
| CRC_31   | CMS2| 74  | F      | Colon| Right| 2     |
| CRC_32   | CMS2| 68  | F      | Colon| Left | 4     |
| CRC_33   | CMS2| 80  | F      | Colon| Right| 3     |
| CRC_34   | CMS1| 81  | M      | Colon| Right| 3     |

**Table 1.** Patient metadata for Predict colorectal cancer cohort. Gender categories M for male and F for female are used; column stage is post-operative Tumour-Node-Metastasis staging; U/C in CMS column stands for unclassified; and N/A in the side column stands for data not available.
Methods

Here, we present a more condensed version of the methods that led to data and analyses in the primary publication14. The workflow is shown in Fig. 1 and the names of the partial processes (depicted in blue in the figure) are used as titles in this section to structure the text. We make the raw sequencing data freely available in NCBI Sequence Read Archive16, and scripts together with more downstream analysis results are accessible as the Zenodo dataset17.

Sample collection & handling. Tumour tissue was collected from 34 patients undergoing surgery for resection of colorectal tumours. None of the patients had received chemotherapy prior to surgery, and all patients provided written, informed consent. This study was carried out with approval from the University of Otago Human Ethics Committee (ethics approval number: H16/037). Table 1 shows patient metadata for the cohort. At the time of surgery, CRC tumour cores were taken and immediately frozen in liquid nitrogen and initially stored at $-80°C$. They were subsequently transferred to RNAlater ICE™ (Qiagen), and equilibrated for at least 48 hours at $-20°C$, prior to nucleic acid extraction. RNA and DNA were extracted from 15–20 mg each of tissue using RNEasy Plus Mini Kit (Qiagen) and DNeasy Blood and Tissue Mini Kit (Qiagen), respectively. Tissue disruption was carried out using a Retsch Mixer Mill. RNA extraction included a DNAse treatment step, and DNA extraction included overnight incubation with proteinase K, and treatment with RNAse A. Purified nucleic acids were quantified using the NanoDrop 2000c spectrophotometer (Thermo Scientific, Asheville, NC, USA), and stored at $-80°C$. Nucleic acids were extracted from all tumour samples in a single batch by one operator, to avoid inter-batch variation.

RNA-seq. Library preparation and ribosomal RNA depletion was carried out using Illumina TruSeq stranded total RNA library prep V1 and Ribo-Zero Gold. The ribosomal RNA depletion step has potentially removed a portion of bacterial ribosomal RNA alongside of the human one, hence losing some information on bacteria. However, the same method of depletion was used on all the samples thus the potential loss would effect all of them in a similar manner. RNA sequencing was carried out using the Illumina HiSeq, 2500 V4 platform, to produce 125 bp paired end reads. Each sample library was split equally to two HiSeq lanes and the sequences from the two lanes were merged for each sample during the data processing phase.

Read mapping, Gene expression quantification, and Profile classification. Adapters and low quality segments were removed from the sequenced reads using fastq-mcf from EAUtils18 and SolexaQA ++19. The cleaned reads were mapped to the GRCh38 reference human genome with STAR20 and the read count for each HAVANA annotated gene in every sample was calculated with htsq-count21. The read counts were transformed to gene expression profiles measured in transcripts-per-million (TPM) with DESeq222. The published CMS classifier15 was used to assign a molecular subtype of the disease to each sample based on the gene expression profiles (for more details see14). We identified six samples as CMS1, 13 samples as CMS2 and nine samples as CMS3. No samples were classified as CMS4, and six samples were unclassified.

Fig. 2. Per base quality of raw sequencing data, sample CRC_16. Output of FASTQC: (a) RNA sequencing, (b) 16S rRNA amplicon sequencing.
Assignment of reads to bacterial taxa. A Kraken\textsuperscript{23} database was built containing all NCBI Refseq complete genomes or chromosome-level genomes (January 2017) and additional genomes of bacteria proposed to play a role in CRC, disregarding their genome status. The used bacterial genomes are listed in the files Supplementary_table_K1.xlsx (all complete and chromosome-level genomes) and Supplementary_table_K2.xlsx (of interest specifically for CRC) in the folder data/kraken accompanied repository. All RNA-seq reads that were not uniquely mapped to the human genome reference sequence were used as input to Kraken using this custom database for taxonomic classification per sample. Altogether, 2231 different bacterial species were detected in at least one sample and only 1.4% of the analysed sequence were not assigned to any bacterial species. We visualised bacterial abundances per CRC subtype using Krona\textsuperscript{24} and the interactive plots are available at http://crc.sschmeier.com.

Differential analysis of bacterial species in CMS. We analysed the enrichment/depletion of bacterial species in one subtype compared to the other subtypes employing a strategy similar to differential expression analysis. Using edgeR\textsuperscript{25}, we identified bacterial taxa with considerable abundance differences among the subtypes.

Table 2. RNAseq, read counts and their ratios in various data processing stages for each sample. N/A in the CRC_14 sample stands for data not available.
16S rRNA metabarcoding. Libraries containing 16S rRNA were prepared with 20 ng of DNA for each sample using primer pairs flanking the V3 and V4 hypervariable regions of the 16S rRNA gene and Illumina sequencing adaptors and barcodes were added using limited cycle PCR. Amplicon sequencing was carried out using the Illumina MiSeq platform, and paired end reads of length 250 bp were generated.

Metabarcoding data analysis. Short overlapping forward and reverse reads coming from the same fragment were joined together with FLASH\(^{26}\) to form sequences of the V3-V4 hypervariable 16S rRNA region. Afterwards, low quality regions were removed from the resulting fragments with SolexaQA\(^{++}\)\(^{19}\). Microbiome analysis was carried out with the QIIME bioinformatics pipeline\(^{27}\) using the Greengenes database\(^{28}\) for taxonomy assignment. No further normalisation of the data was performed.

Data Records

Sequenced genomic data from both RNA-seq and 16S rRNA metabarcoding experiments are stored in the Sequence Read Archive as the study SRP117763\(^{16}\). Data resulting from the analyses presented here are located in the folder data of the Zenodo repository\(^{17}\). The data are separated into several subfolders:

- The folder expr contains raw read counts in subfolder raw_counts, tpm-based expression profiles of all samples stored in file tpm.readyForClassifier.tsv and also file CMSclassifiedCRC.tpm.havana.tsv containing the CMS subtype classification. These files are the main outcomes of gene expression profile classifications.

| sampleID | sequenced read pairs (count) | base quality \(\geq 30\) (in %) | cleaned fragments (count) | cleaned in sequenced (in %) |
|----------|-----------------------------|-----------------------------|--------------------------|--------------------------|
| CRC_01   | 333335                      | 89.75                       | 176823                   | 53.05                    |
| CRC_02   | 238221                      | 92.15                       | 126462                   | 53.09                    |
| CRC_03   | 356650                      | 91.98                       | 187474                   | 52.57                    |
| CRC_04   | 307676                      | 92.35                       | 165991                   | 53.95                    |
| CRC_05   | 261798                      | 93.33                       | 148547                   | 56.74                    |
| CRC_06   | 122630                      | 92.36                       | 67416                    | 54.98                    |
| CRC_07   | 175589                      | 94.39                       | 104310                   | 59.41                    |
| CRC_08   | 210849                      | 93.22                       | 119255                   | 56.56                    |
| CRC_09   | 238258                      | 94.06                       | 133700                   | 56.12                    |
| CRC_10   | 233536                      | 92.30                       | 129813                   | 55.59                    |
| CRC_11   | 291890                      | 87.52                       | 148406                   | 50.84                    |
| CRC_12   | 173621                      | 93.14                       | 96744                    | 55.72                    |
| CRC_13   | 204471                      | 92.43                       | 113588                   | 55.55                    |
| CRC_14   | 255851                      | 92.52                       | 141822                   | 55.43                    |
| CRC_15   | 254700                      | 93.37                       | 145899                   | 57.28                    |
| CRC_16   | 210011                      | 94.06                       | 126141                   | 60.06                    |
| CRC_17   | 197765                      | 92.96                       | 110784                   | 56.02                    |
| CRC_18   | 161324                      | 92.88                       | 90441                    | 56.06                    |
| CRC_19   | 147498                      | 93.66                       | 82425                    | 55.88                    |
| CRC_20   | 235318                      | 92.33                       | 127779                   | 54.30                    |
| CRC_21   | 169421                      | 93.64                       | 96627                    | 57.03                    |
| CRC_22   | 249364                      | 93.55                       | 144146                   | 57.81                    |
| CRC_23   | 171152                      | 91.54                       | 91281                    | 53.33                    |
| CRC_24   | 102066                      | 91.90                       | 54880                    | 53.77                    |
| CRC_25   | 334496                      | 93.67                       | 195656                   | 58.49                    |
| CRC_26   | 265504                      | 93.22                       | 150713                   | 56.76                    |
| CRC_27   | 69391                       | 93.02                       | 40037                    | 57.70                    |
| CRC_28   | 137873                      | 91.43                       | 74333                    | 53.91                    |
| CRC_29   | 176936                      | 94.06                       | 107348                   | 60.67                    |
| CRC_30   | 202971                      | 94.46                       | 118078                   | 58.17                    |
| CRC_31   | 220216                      | 93.44                       | 126807                   | 57.58                    |
| CRC_32   | 108880                      | 93.34                       | 61688                    | 56.66                    |
| CRC_33   | 219198                      | 94.42                       | 128793                   | 58.76                    |
| CRC_34   | 305438                      | 93.88                       | 178759                   | 58.53                    |
Results of the metagenomics analysis of the non-human genomic content of RNA-seq are located in folder kraken together with two tables (Supplementary_table_K*.xlsx) containing lists of bacterial species used in this metagenomics analysis.

The folder 16S contains the biom file otu_table.biom resulting from the 16S rRNA metabarcoding analysis with QIIME and two partial abundance tables otu_table_sorted_*.txt.gz. The abundance tables are derived from the biom file and were used further for data visualisation in the primary publication as well as for the metagenomics method comparison.

Technical Validation

RNA-seq raw data quality. The quality of raw sequenced reads from RNA-seq experiments was assessed with FASTQC and was very good. A pair of representative per base quality plots of corresponding forward and reverse read pairs for one sample is shown in Fig. 2a). Regardless of the raw data quality, all the samples underwent routine data cleaning to ensure that no base was called with a Phred quality below 20. In Table 2, we show number of reads passing various data processing stages together with relative proportion of the reads passing two different stages.

16S rRNA sequencing raw data quality. In Fig. 2b), we show quality of the 16S rRNA sequencing raw data for sample CRC_16. The other samples’ 16S rRNA quality plots looked similar. It can be seen that per base quality varied a little bit more along the 16S rRNA reads when compared to the RNA-seq reads, but overall the quality was very good for the 16S rRNA sequencing as well. Please note that the read length for the 16S rRNA sequencing was twice the read length of the RNA-seq, which together with differences between the used sequencing instruments explains differences in the quality plots. All the 16S rRNA samples underwent routine data cleaning to ensure that no base was called with a Phred quality below 20. In Table 3, we show number of reads passing various data processing stages together with relative proportion of the reads passing two different stages.

Code Availability

All the code used to process the genomic data is freely available as a part of the provided Zenodo repository and the code is located in the folder named scripts. The scripts folder also contains dependencies listed in the file used_packages_and_their_versions.tsv. Depending on the scripts’ functionality, they are separated into various folders:

- The folder rnaseq-subtype-classification contains scripts used for reading mapping, gene expression quantification, and profile classification.
- The folder kraken/human-unmapped contains scripts to assign reads to bacterial taxa.
- The folder kraken/diff-expr-taxa contains scripts for differential analysis of bacterial species in CMS.
- The folder 16S-metabarcoding contains scripts for metabarcoding data analysis.

References

1. Ferlay, J. et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. International Journal of Cancer 136, E359–E386 (2015).
2. Ahn, I. et al. Human gut microbiome and risk for colorectal cancer. Journal of the National Cancer Institute 105, 1907–1911 (2013).
3. Marchesi, J. R. et al. Towards the human colorectal cancer microbiome. PLoS One 6, e20447 (2011).
4. Gao, Z., Guo, B., Gao, R., Zha, Q. & Qin, H. Microbiota disbiosis is associated with colorectal cancer. Frontiers in Microbiology 6, 20 (2015).
5. Suhamit, E. et al. Microbial dysbiosis in colorectal cancer (CRC) patients. PLoS One 6, e16393 (2011).
6. Rauch, M. & Lynch, S. The potential for probiotic manipulation of the gastrointestinal microbiome. Current Opinion in Biotechnology 23, 192–201 (2012).
7. Preidis, G. A. & Versalovic, J. Targeting the human microbiome with antibiotics, probiotics, and prebiotics: gastroenterology enters the metagenomics era. Gastroenterology 136, 2015–2031 (2009).
8. Gupta, A. & Khanna, S. Fecal microbiota transplantation. The Journal of the American Medical Association 310, 102–102 (2017).
9. Nakatsu, G. et al. Gut mucosal microbiome across stages of colorectal carcinogenesis. Nature Communications 6, 8727 (2015).
10. Jass, J. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. Histopathology 50, 113–130 (2007).
11. Leggett, B. & Whitehall, V. Role of the serrated pathway in colorectal cancer pathogenesis. Gastroenterology 138, 2088–2100 (2010).
12. Network, C. G. A. et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330 (2012).
13. Domingo, E. et al. Use of multivariate analysis to suggest a new molecular classification of colorectal cancer. The Journal of Pathology 229, 441–448 (2013).
14. Purcell, R. V., Visnovska, M., Biggs, P. J., Schmeier, S. & Fritzsche, F. A. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. Scientific Reports 7, 11590 (2017).
15. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. Nature Medicine 21, 1350 (2015).
16. NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRP117763 (2017).
17. Schmeier, S., Visnovska, M., Biggs, P. J., Purcell, R. V. & Fritzsche, F. A. Scripts and data attached to colorectal cancer study by Purcell, 2017. Zenodo, https://doi.org/10.5281/zenodo.2604777 (2018).
18. Aromesty, E. Comparison of sequencing utility programs. The Open Bioinformatics Journal 7, 1–8 (2013).
19. Cox, M. P., Peterson, D. A. & Biggs, P. J. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11, 485 (2010).
20. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
21. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169 (2015).
22. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq. 2. Genome Biology 15, 550 (2014).
23. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology 15, R46 (2014).
24. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics* **12**, 385 (2011).
25. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
26. Magić, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
27. Caporaso, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335 (2010).
28. DeSantis, T. Z. et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* **72**, 5069–5072 (2006).

**Acknowledgements**

The authors would like to thank Helen Morrin and the staff at the Cancer Society Tissue Bank, Christchurch for their enthusiasm and support for this project, the patients involved, for generously participating in this study, as well as New Zealand Genomics Limited (NZGL) for the sequencing and support in study design. Funding sources (Rachel Purcell): Maurice and Phyllis Paykel Trust, Gut Cancer Foundation (NZ), with support from the Hugh Green Foundation, Colorectal Surgical Society of Australia and New Zealand (CSSANZ).

**Author Contributions**

T.V. carried out bioinformatics analysis and contributed to manuscript writing. P.B. was involved in study design and data analysis. S.S. was involved in study design, bioinformatics analysis and was a contributor to manuscript preparation. F.F. was involved in study design and clinical aspects of the study. R.P. carried out nucleic acid and sequencing preparation of tumour samples and was a contributor to manuscript writing. All authors read and approved the final manuscript.

**Additional Information**

**Competing Interests:** The authors declare no competing interests.

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit [http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/).

The Creative Commons Public Domain Dedication waiver [http://creativecommons.org/publicdomain/zero/1.0/](http://creativecommons.org/publicdomain/zero/1.0/) applies to the metadata files associated with this article.

© The Author(s) 2019