Searching for Pathogens in Cancer Sequence Data

Gihawi, A.[1], Hurst, R.[1], Leggett, R.M.[2], Cooper, C.S.[1], Brewer, D.S.[1,2], Genomics England Research Consortium[3]

[1] Bob Champion Research and Education Building, University of East Anglia, Norwich, UK
[2] Earlham Institute, Norwich, UK
[3] Genomics England, London, UK

Background

The association of *H. pylori* with gastric adenocarcinomacarcinoma[1] and the successful Human papillomavirus (HPV) vaccine, estimated to prevent ≤80% of cervical cancer cases[2, 3] attest to the prominent role that pathogens play in cancer.

When cancer tissue is sequenced, any microorganisms present in the sample can be incidentally sequenced simultaneously[4]. We have been investigating this in over 10,000 samples from cancer cohort of the 100,000 genomes project by Genomics England. It is hoped that this dataset will prove to be a rich resource for detecting microorganisms and linking them to disease.

We previously benchmarked software to devise the optimal approach to detect mircorogranisms in cancer whole genome sequences. The top performing pipelines are provided in a tool called SEPATH[5] which performs the following:

- Extracts unmapped reads from BAM files
- Quality trimming & human read removal
- Metagenomic classification - Kraken/mOTUs2[6, 7]

Methods

SEPATH is currently being applied to all samples from within the 100,00 genomes project to search for evidence of and associations between pathogens and disease. The resulting non-human reads were pooled by cancer type and subject to metagenomic assembly MEGAHIT[8]. The resulting contigs were classified with Diamond[9] using a database built from NCBI genomes available here. From both approaches, all potential contaminants according to Salter *et al.*[10] and any mammalian classifications were removed from further investigation.

So far we have sifted through in excess of 2.25×10^{14} base pairs from >10,000 tumour samples to search for evidence of pathogens. This is set to increase as more sequence data is released (figure 1) and as we search through the rare disease cohort.

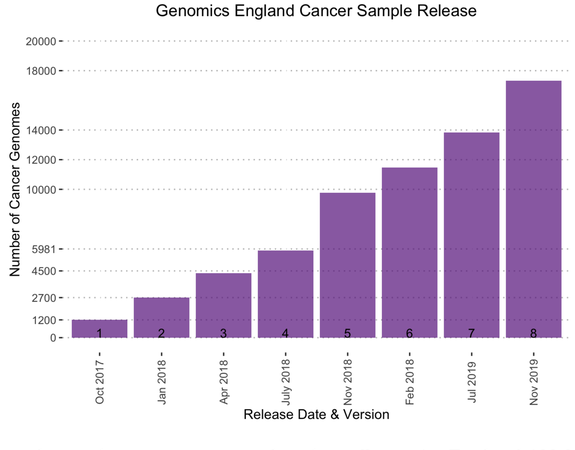

Figure 1: The release of tumour sequence data from Genomics England 100,000 Genomes Project. SEPATH has been applied to data up to version 6.

Results

SEPATH using Kraken

We filtered samples to retain only those that were PCR-free, fresh tissue tumours. Classifications with fewer than 20 reads were disregarded. Here we report on the top 20 most frequently occuring genera within each cancer type. Although some environmental contaminants likely remain, some pathogens have emerged that have previously been suspected to contribute to tumorigenesis such as *Bacteroides*, *Prevotella* and *Fusobacterium*[11].

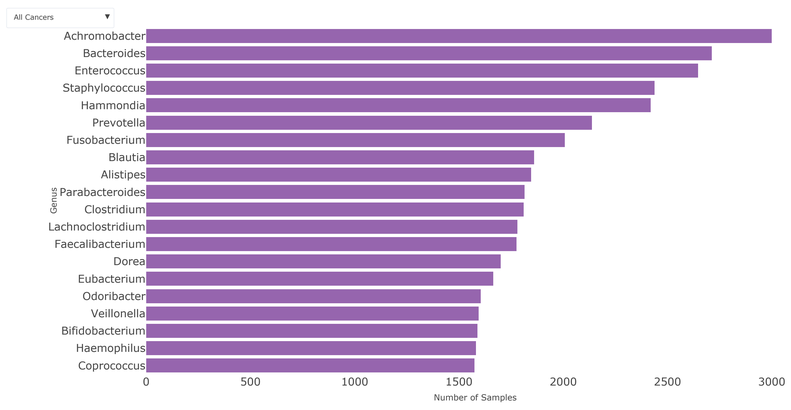

Figure 2: Results from the 100,000 Genomes Project. Genus level classifications were obtained by using SEPATH with Kraken. Results are displayed by cancer type.

Classification of Assembled Contigs

Although reads with any similarity to the human genome have been removed prior to assembly, we see hundreds of thousands of assembled genomic fragments accross the entire dataset for a variety of genera. The most common of which in the whole dataset are for *Prevotella*, *Bacteroides*, *Treponema* and *Fusobacterium*.

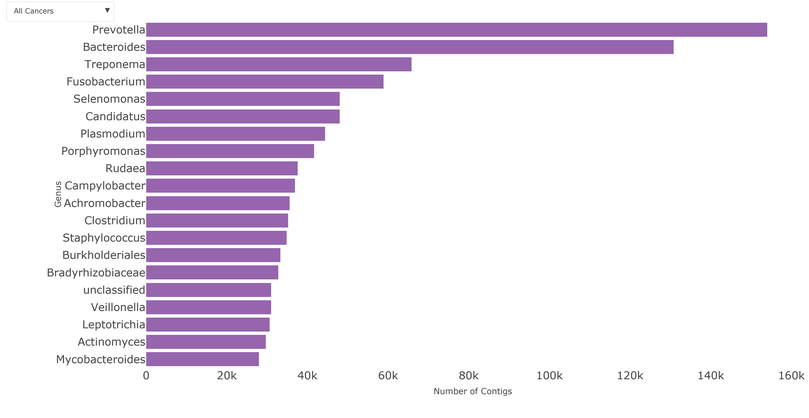

Figure 3: Genus level classifications of contigs from metagenomic assembly of non-human reads within the 100,000 genomes project.

Conclusion

Although the data overall is sparse and rife with potential contaminants, there are a range of interesting genera being reported accross cancer types by both SEPATH and by investigating assembled contigs. Some of these genera have previously been suggested to contribute to tumourigenesis.

Future Directions

Despite using machine learning efforts to refine lists of informative taxa, our efforts so far have not identified any obvious structure by cancer type. This is contradictory to recent publications[12, 13] and will be investigated further.

- This data will be interrogated for associations with metadata.
- The raw quality-trimmed and non-human reads have been mapped back to the contigs and this will be assessed as a second layer of information.
- Contigs will be subject to analysis to investigate the functional potential of pathogens reported.

Final note

This poster was produced using posterdown. The associated GitHub can be found here.

References

1. 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:e609–16.

2. 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;22:505–15. http://www.jstor.org/stable/23047685.

3. Crosbie EJ, Einstein MH, Franceschi S, Kitchener HC. Human papillomavirus and cervical cancer. The Lancet. 2013;382:889–99. doi:10.1016/s0140-6736(13)60022-7.

4. Magiorkinis G, Matthews PC, Wallace SE, Jeffery K, Dunbar K, Tedder R, 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.

5. 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 Biol. 2019;20:208. doi:10.1186/s13059-019-1819-8.

6. Wood D, Salzberg S. Kraken - ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15.

7. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh HJ, Cuenca M, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun. 2019;10:1014. doi:10.1038/s41467-019-08844-4.

8. 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;31:1674–6. doi:10.1093/bioinformatics/btv033.

9. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using diamond. Nature Methods. 2015;12:59–60.

10. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology. 2014;12. doi:10.1186/s12915-014-0087-z.

11. Shang F-M, Liu H-L. Fusobacterium nucleatum and colorectal cancer: A review. World Journal of Gastrointestinal Oncology. 2018;10:71–81. doi:10.4251/wjgo.v10.i3.71.

12. Nejman D, Livyatan I, Fuks G, Gavert N, Zwang Y, Geller LT, et al. The human tumor microbiome is composed of tumor typespecific intracellular bacteria. Science. 2020;368:973–80. doi:10.1126/science.aay9189.

13. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature. 2020;579:567–74. doi:10.1038/s41586-020-2095-1.