Translational Opportunities of Detecting Microbes in Cancer Whole Genome Sequences

Gihawi, A. [1,]

🐦 @AbrahamGihawi
✉️ A.Gihawi@uea.ac.uk

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

Microbial infections cause 15% of cancers globally [1]. The most significant oncomicrobes are *Helicobacter pylori* that causes gastric carcinoma [1] and Human papillomavirus (HPV) that causes several types of cancer. HPV vaccination has been successful in preventing infection and tumourigenesis worldwide[2]. As we transition into an era of genomic medicine, it is becoming increasingly possible to characterise the tumour-microbe landscape by interrogating high-throughput cancer sequencing data. We have developed a benchmarked approach to report evidence for microbial constituents found within human tissue samples (SEPATH [3]). This pipeline has been applied to cancer sequences within Genomics England's 100,000 Genomes Project. A wide range of microbial taxa are detected and here we report on the potential translational opportunities that tumour-microbial data offers.
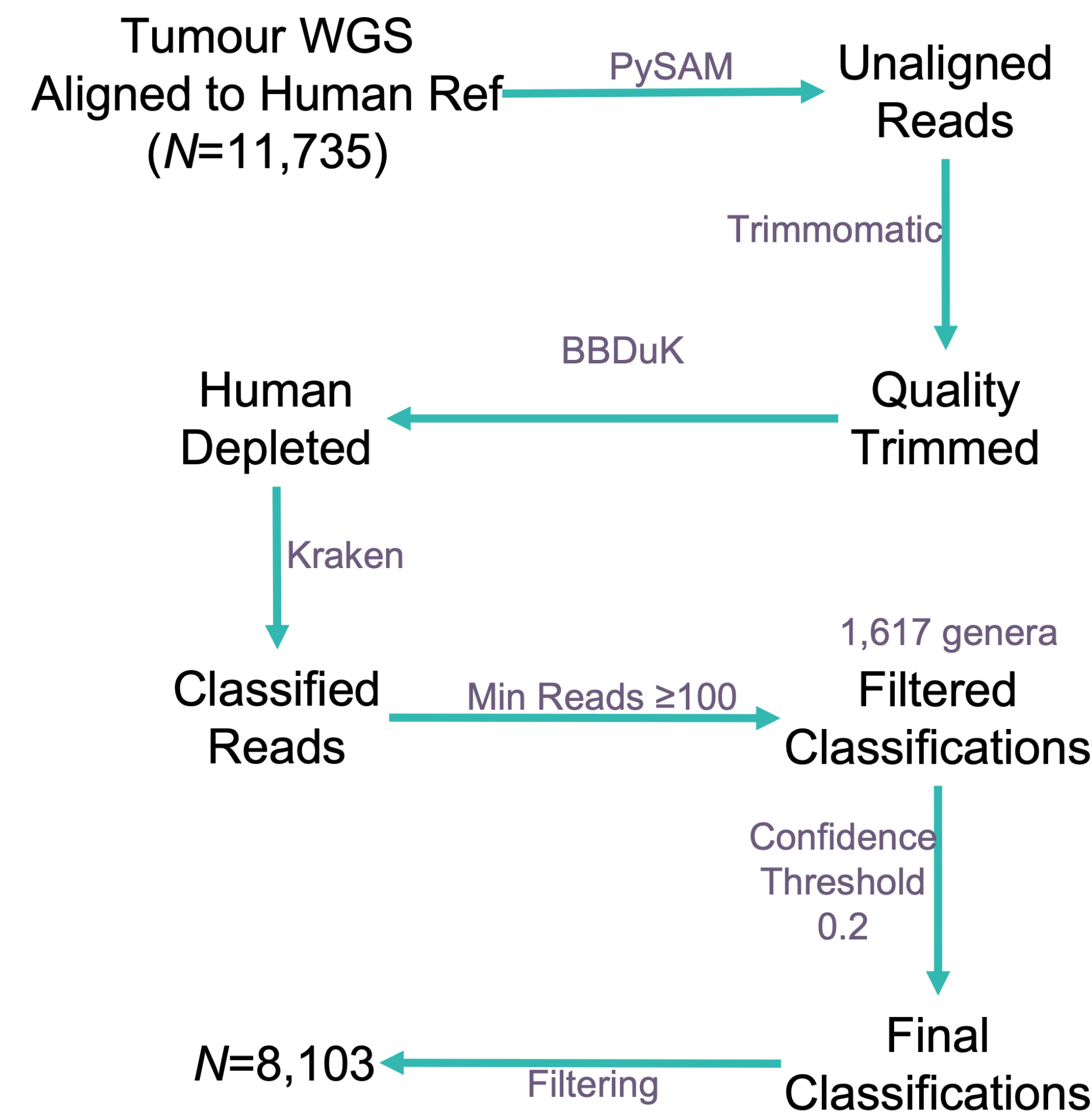
Methods

Figure 1: SEPATH was applied to tumour whole genome sequences (WGS, N=11,735) from Genomics England's 100,000 Genomes Project. This pipeline extracts high quality reads that are unaligned to the human genome, runs additional human read removal and they are subject to taxonomic classification. A total of 1,617 microbial genera were detected in the cohort. The final matrix was filtered to retain dominant tumour types that were PCR Free, from fresh-frozen, primary tissue (n=8,909)

Pan-Cancer Structure

A feature selection algorithm (Boruta) identified microbial genera (495) most indicative of tumour type. A distance matrix was constructed using the Spearman's distance with the `distanceMatrix` function (ClassDiscovery package). The distance matrix was subject to *t*-SNE (Rtsne package) with parameters: `dims=2, perplexity=80, max_iter=2000, check_duplicates=TRUE.`

Mutation Calling

Somatic variant VCF files were generated by Genomics England. Non-synonymous *TP53* variants of moderate-high impact were investigated in the oral/oropharyngeal cohort. Samples with no variants were considered wild-type.

HTLV-1 Identification

Deltaretrovirus sequencing reads from one participant were extracted from both tumour and germline samples and were subject to a Megablast search (nr/nt collection). Reads were aligned to a HTLV-1 reference NC_001436.1 using BWA-MEM with default parameters [4]. The resulting alignment was viewed with IGV (version 2.9.4) [5].

Results

Pan-Cancer Structure

Oral and colorectal tumours contain distinctive microbial elements compared to other cancer types (fig. 2).
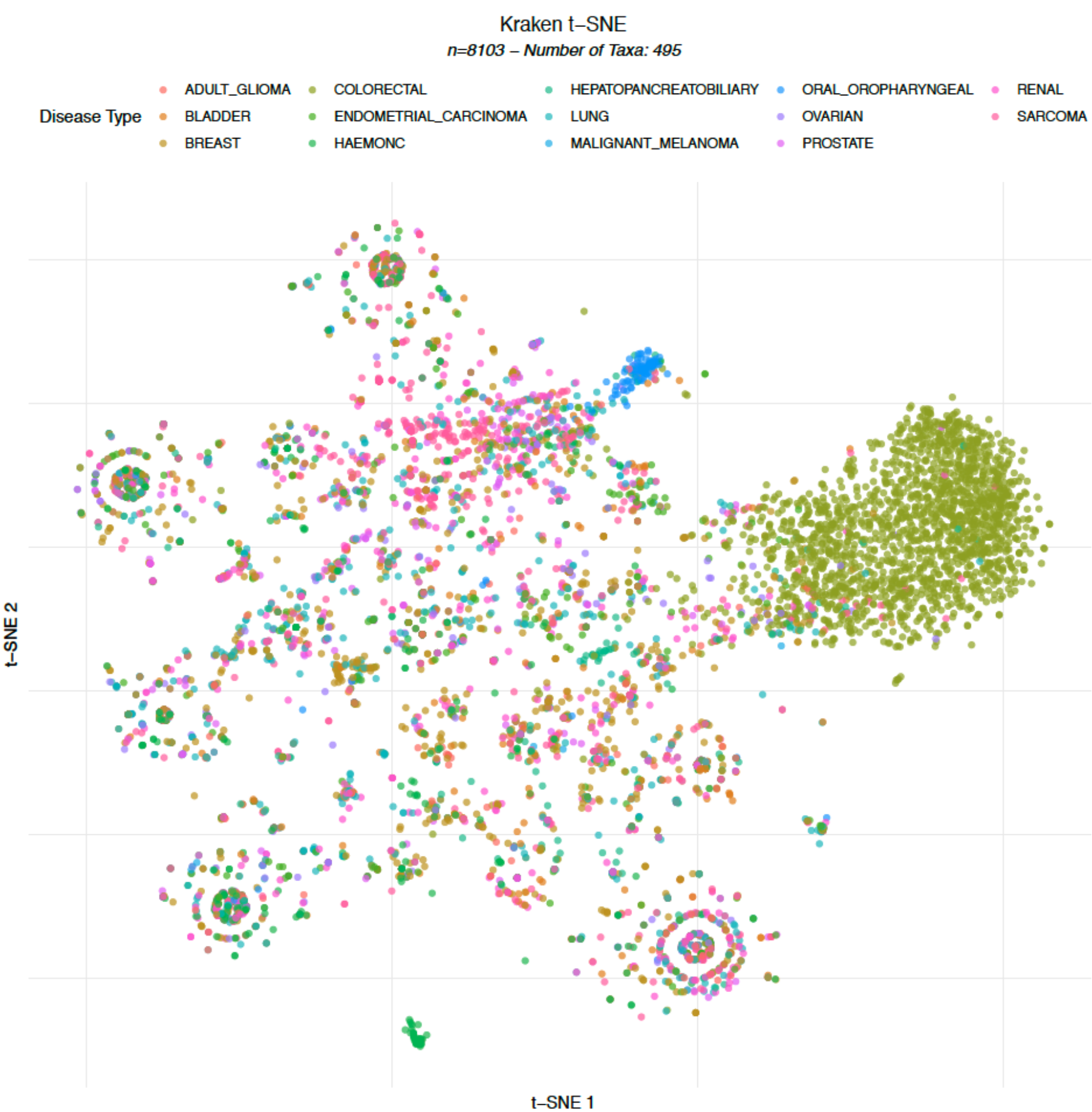

Figure 2: t-SNE plot demonstrating microbial structure for n=8,103 tumour samples that were subject to whole genome sequencing. Oral and Colorectal cancers typically contain microbial elements that distinguish them from other tumour types.

Alphapapillomavirus

Detection of Alphapapillomavirus overlaps with diagnostic tests for HPV and aligns with somatic *TP53* mutational status (fig. 3). 24/38 HPV$^-$ tumours contain non-synonymous *TP53* variants. All 10 HPV$^+$ tumours have wild-type *TP53* as expected [6]. Alphapapillomavirus was detectable in both primary as well as metastatic tumour samples.
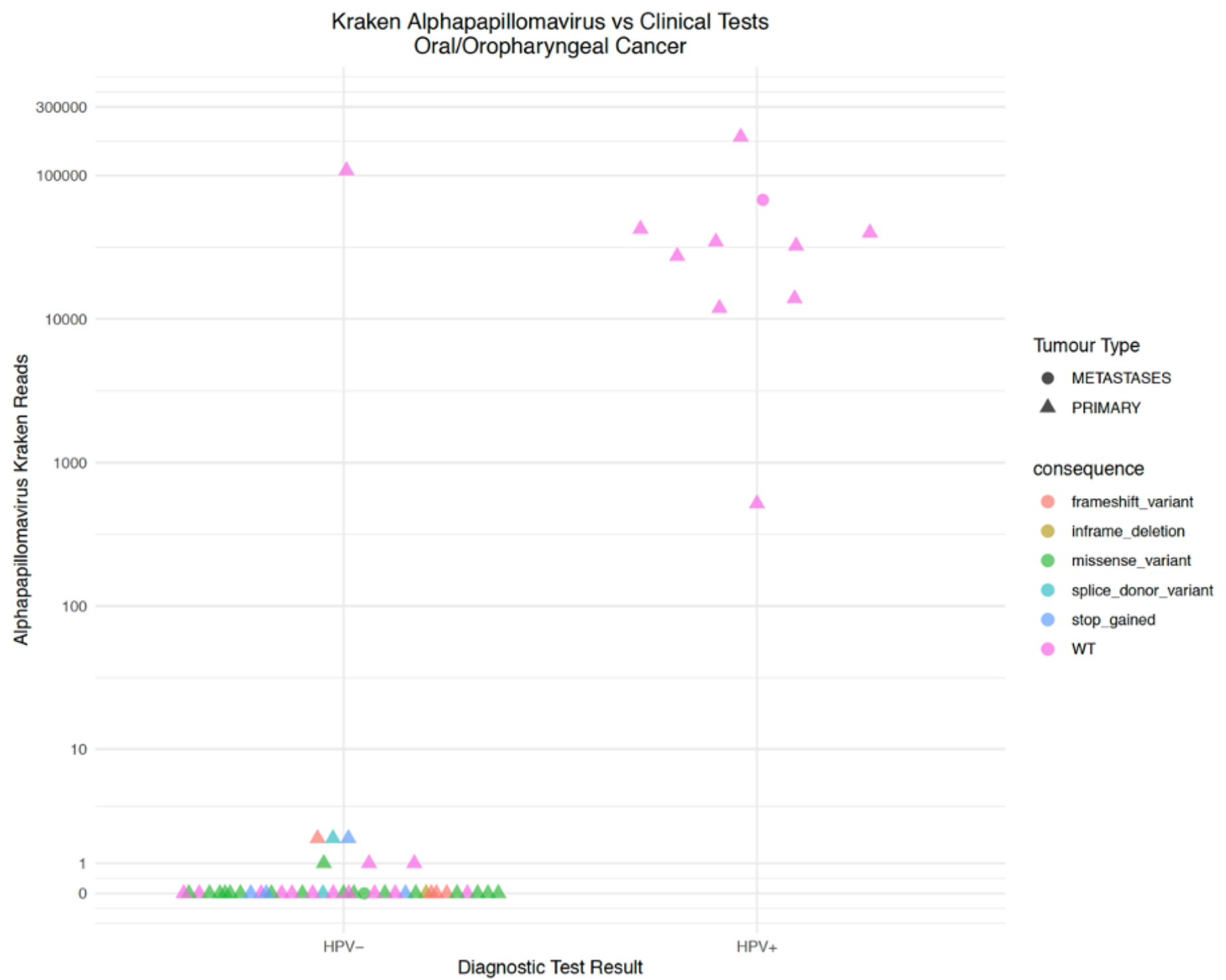

Figure 3: Alphapapillomavirus detection in Oral Cancer with TP53 mutational status.

HTLV-1 Detection

We found evidence for Deltaretrovirus in one participant. A similar result was observed in the participant's germline sample. Classified reads were extracted and uniquely matched HTLV-1 upon a BLAST search. Furthermore, the classified sequencing reads align along the length of a HTLV-1 reference genome providing additional evidence for detection (fig. 4).
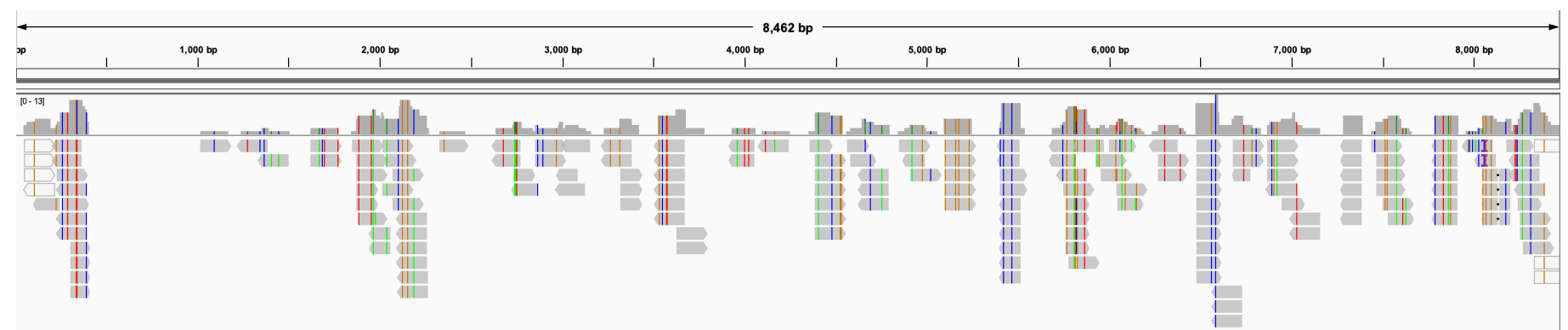

Figure 4: Alignment of 172 sequencing reads from a participants tumour and germline samples across a HTLV-1 reference genome (8,462 base pairs).

Conclusions

- Microbial nucleic acids are detectable within tumour whole genome sequence data.
- Colorectal and oral cancers typically contain distinctive microbial elements.
- Detection of HPV is comparable to the clinical standard. This may provide insights into 101 samples currently with no HPV test result, 8 of which are estimated positive.
- It might be possible to detect infectious diseases like HTLV-1 but this needs to be compared to a diagnostic standard.

Acknowledgements

Note: This poster was produced in R. Details and code to reproduce this poster can be found: github.com/Agihawi/gel_summit_2022

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. Hibbitts S, Tristram A, Beer H, McRea J, Rose B, Hauke A, et al. UK population based study to predict impact of HPV vaccination. J Clin Virol. 2014;59:109–14. doi:10.1016/j.jcv.2013.12.002.
3. 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.
4. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25:1754–60. doi:10.1093/bioinformatics/btp324.
5. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature Biotechnology. 2011;29:24–6. doi:10.1038/nbt.1754.
6. Shi C, Liu S, Tian X, Wang X, Gao P. A TP53 mutation model for the prediction of prognosis and therapeutic responses in head and neck squamous cell carcinoma. BMC Cancer. 2021;21. doi:10.1186/s12885-021-08765-w.

UEA University of East Anglia   BIG C   Genomics england

Reproduce Poster