Extended text of manuscript submitted Wellcome Open Research

Potential for diagnosis of infectious disease from the 100,000 Genomes Project Metagenomic Dataset: Recommendations for reporting results

Gkikas Magiorkinis[1], Philippa C Matthews[2,3,4], Susan E. Wallace[5], Katie Jeffery[2,3], Kevin Dunbar[6], Richard Tedder[7], Jean L. Mbisa[6], Bernadette Hannigan[6,8], Effy Vayena[9], Peter Simmonds[2], Daniel S. Brewer[10,11], Abraham Gihawi[10], Ghanasyam Rallapalli[10], Lea Lahnstein[12], Tom Fowler[12], Christine Patch[12], Fiona Maleady-Crowe[12], Anneke Lucassen[13], Colin S. Cooper[10]

Affiliations

1. Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, Athens, Greece
2. University of Oxford, Oxford, United Kingdom
3. Oxford University Hospitals NHS Foundation Trust, Oxford, UK
4. Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK
5. University of Leicester, Leicester, United Kingdom
6. Public Health England, United Kingdom
7. Imperial College, London, United Kingdom
8. University of Cambridge, Cambridge, United Kingdom
9. Swiss Federal Institute of Technology (ETH), Zurich, Switzerland
10. University of East Anglia, Norwich, United Kingdom
11. Earlham Institute, Norwich, United Kingdom
12. Genomics England, London, United Kingdom
13. Faculty of Medicine, University of Southampton, Southampton, United Kingdom

Correspondence to: Gkikas Magiorkinis, gmagi@med.uoa.gr

Summary of research aims of the Integrated pathogens and Mobile Elements GeCIP domain

The aim of the Functional Integrated pathogens and Mobile Elements GeCIP domain is two-fold, firstly to identify microbial sequences in a metagenomics approach and secondly to identify novel integrations of human mobile elements. This manuscript is specifically focusing on incidental findings of the first aim, more specifically microbial DNA sequences in HTS genomic data that do not align to the human genome. Infectious agents are involved in the development of a variety of human cancers and some rare conditions. Our hypotheses are (a) that infectious agents can be directly linked to cancer development, but as yet remain undefined and (b) that some rare conditions may result from infections. It is expected that DNA sequences from these infectious agents will be identified in both blood and cancer samples from *100,000 Genomes Project* participants.

At present these findings do not influence individual patient management, but through a process of collating and analysing these data we may be able to underpin future research that addresses the significance of specific micro-organisms in a specific disease context.

Preliminary data on the identification performance of the proposed pathogens metagenomics pipeline presented by the Integrated Pathogens and Mobile Elements GeCIP domain

(A detailed analysis on the performance of the metagenomics pipeline is provided on the Github website of SEPATH https://github.com/Agihawi/SEPATH. The performance further is described in Gihawi, Abraham, Rallapalli, Ghanasyam, Hurst, Rachel, Cooper, Colin, Leggett, Richard M. and Brewer, Daniel (2019) SEPATH: benchmarking the search for pathogens in human tissue whole genome sequence data leads to template pipelines. Genome Biology.)

High throughput sequencing technologies have facilitated wide application of metagenomic studies to gain insights into ecological and host-associated microbiomes (1, 2). Various bioinformatic tools for metagenome studies have been developed in the last decade (3), with much needed benchmark studies performed in the last two years, providing critical information about the interpretation of the results (4-6). The most important step in metagenomic analyses is identifying constituting organisms and estimating their abundance from sequence reads. However, accurate profiling and interpretation of a microbiome from whole genome sequence data of an individual is challenging for various reasons. Uneven sampling poses one hurdle and it is further confounded by the assignment difficulty of sequence reads to the origin genome. Typically, metagenomic tools assign a sequence read to an organism and place it in the tree of life however, there are branches of the tree of life that are under/un represented due to difficulty in culturing certain microorganisms outside their host/preferred environment. The estimated bacterial species richness on earth ranges from millions to billions (7, 8). Nonetheless, there are only about 130,000 bacterial genomes available on NCBI and only ~10,000 of bacterial genomes are at either Complete Genome or Chromosomal level (9). Unlike the New York City subway microbiome (10) and cow rumen microbiome (11) studies, where less than 50% of metagenomic reads were classified, human microbiome studies are in a better position as 83,000 of bacterial genomes in NCBI were human host derived (9).

Metagenomic analysis is primarily carried out in two ways; classification is performed either on assembled contigs from sequence reads or directly on sequence reads (3).

Assembly based approaches require additional resources and often need at least 10X coverage or higher to be able to assemble constituent metagenomes (6). Direct read classification methods can be broadly divided in to marker, alignment and composition-based approaches (4). Marker sequence-based profiling methods use a curated list of sequences and provide microbial composition as relative abundance using only taxa that were recognized by the software. Alignment-based approaches tend to be slow due to the time it takes to search whole databases. Alternative composition-based approaches make use of constituting *k*-mers (sequences of length *k*) to classify each read, using a premade database of *k*-mers. *K*-mer based tools have shown higher accuracy up to genus level (4) and provide high recall even at lower sequencing depths and are relatively faster (5). *K*-mer lengths between 21 and 31 are preferable, as they have high sensitivity and precision (12) and can distinguish up to 90% of k-mers between isolates from same genus (13). Longer k-mers could improve the species specificity, however, there is a trade-off with loss of sensitivity, loss of assignments as well as increased computation times possibly due to errors in the sequencing reads (12, 13).

Metagenomic analysis is further confounded by the various contaminating sequences appearing from sequencing library preparation kits or laboratory reagents (14, 15), as well contaminants incorporated in the genome assemblies of microbes (16). Therefore, it is vital to include negative controls for each sequencing preparation to identify such issues. One of the major causes of contamination is low template of target material (15), although it shouldn't be an issue for whole genome/exome sequencing. Moreover, certain trends in microbial abundance patterns could be useful hints for detection of contaminants; such as identification in particular batches of samples sequenced together or in samples prepared in a particular lab. Additionally, software such as SourceTracker (17) that use a Bayesian approach to estimate contaminants could be used as part of the analysis pipeline.

We have made use of 100 blood sample sequences from The International Genome Sample Resource (18) for metagenomic analysis using Kraken (19), a *k*-mer based read classifier. We were perplexed to find *Toxoplasma gondii*, *Leishmania major* and *Plasmodium falciparum* in all 100 samples. In addition, we found *Ralstonia picketti*, *Escheria coli* in more than 2/3rd of samples, as well as common kit contaminants

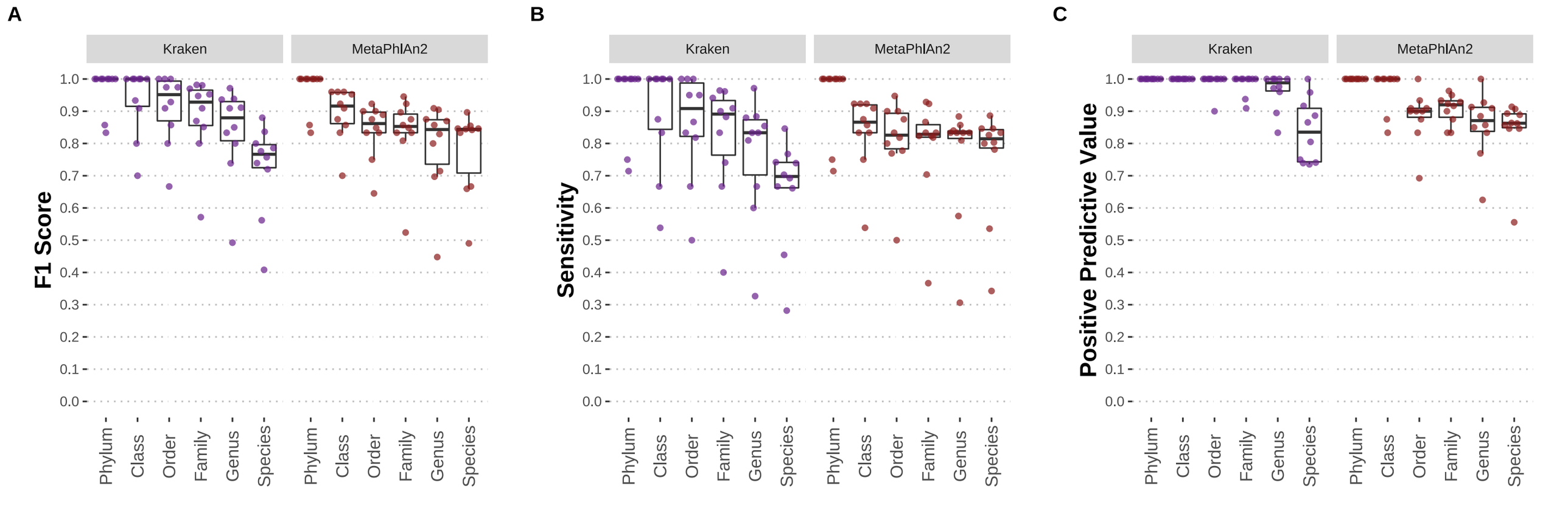
Stenotrophomonas maltophilia, *Pseudomonas chlororaphis* and *Lysobacter enzymogenes* in around 50% of the samples. To investigate further we made use of direct alignment-based approach using BWA (20) against selected reference sequences. We found that *Taxoplasma gondii ME49* genomic scaffolds KE140619, KE140546 have very strong match to human sequence. Similarly, *Parvimonas micra* genomic contig AXUQ01000047 is 100% identical to pJET1.2 derived pCD vectors from Derntl et al (21). Therefore, microbes present in all/most samples would require further investigation, if possible to remove contamination from respective assemblies. Individual differences between the host from which the microbe assembly was generated and the reference human sequence may have missed the filtering against human genome. There were low complex regions in *Leishmania* and *Plasmodium* references that pulled low complex short reads that may not be of high specificity. Low complexity regions in the reference sequence could lead to false positives, therefore references need to be masked for such regions using software like DustMasker (22).

Metagenomic read classifiers achieved highest sensitivity up to family and genus level but sensitivity has seen a drop at species and sub-species level (5, 6, 23) (Extended Figure A-C). Investigation into species and sub-species level needs independent analysis based on genus-level results. False positives tend to be high for taxa that are covered with more genome assemblies (5) as well as due to detection limits of the software (6). Therefore care should be taken while extrapolating identification to clinical assignation and avoid causing fears of pandemic (24). Additionally, recent clinical studies have applied cautious approaches in which they employed metagenomic classification tools, such as Kraken or alignment-based approach, followed by independent alignment of sequence reads using BLAST (25, 26). We intend to follow a similar strategy and are in the process of verifying our pipeline using diverse simulated metagenomic populations. In addition to implementation of combinatorial approaches, integration of output from multiple metagenomic tools have reduced the false positives and provided reliable results (27). Regardless of the limitations of the tools implemented or the approaches followed, it is highly recommended to verify the presence of the identified microbe using independent molecular or clinical techniques before proceeding with treatment.

References

1. Knight R, Callewaert C, Marotz C, Hyde ER, Debelius JW, McDonald D, et al. The Microbiome and Human Biology. Annual Review of Genomics and Human Genetics. 2017;18(1):65-86.
2. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nature Reviews Genetics. 2017;19:9.
3. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Briefings in Bioinformatics. 2017:bbx120-bbx.
4. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Scientific Reports. 2016;6:19233.
5. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biology. 2017;18(1):182.
6. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Nature Methods. 2017;14:1063.
7. Larsen BB, Miller EC, Rhodes MK, Wiens JJ. Inordinate Fondness Multiplied and Redistributed: the Number of Species on Earth and the New Pie of Life. The Quarterly Review of Biology. 2017;92(3):229-65.
8. Schloss PD, Girard RA, Martin T, Edwards J, Thrash JC. Status of the Archaeal and Bacterial Census: an Update. mBio. 2016;7(3).
9. Dheensa S, Shkedi-Rafid, Shiri C, Gillian B, Gabrielle, Schonstein L, et al. Management of Incidental Findings in Clinical Genomic Sequencing Studies. eLS. 2016.
10. Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, et al. Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. Cell Systems. 2015;1(1):72-87.
11. Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, et al. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. Nature Communications. 2018;9(1):870.
12. 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(1):236.
13. Koslicki D, Falush D. MetaPalette: a *k*-mer Painting Approach for Metagenomic Taxonomic Profiling and Quantification of Novel Strain Variation. mSystems. 2016;1(3).
14. 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(1):87.
15. Thoendel M, Jeraldo P, Greenwood-Quaintance KE, Yao J, Chia N, Hanssen AD, et al. Impact of Contaminating DNA in Whole Genome Amplification Kits Used for Metagenomic Shotgun Sequencing for Infection Diagnosis. Journal of Clinical Microbiology. 2017.
16. Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. Standards in Genomic Sciences. 2015;10(1):18.
17. Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, et al. Bayesian community-wide culture-independent microbial source tracking. Nature Methods. 2011;8:761.

18. The Genomes Project C. A global reference for human genetic variation. Nature. 2015;526:68.
19. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology. 2014;15(3):R46.
20. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
21. Derntl C, Kiesenhofer DP, Mach RL, Mach-Aigner AR. Novel Strategies for Genomic Manipulation of Trichoderma reesei with the Purpose of Strain Engineering. Applied and Environmental Microbiology. 2015;81(18):6314-23.
22. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. Journal of Computational Biology. 2006;13(5):1028-40.
23. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Research. 2016;26(12):1721-9.
24. Gonzalez A, Vázquez-Baeza Y, Pettengill JB, Ottesen A, McDonald D, Knight R. Avoiding Pandemic Fears in the Subway and Conquering the Platypus. mSystems. 2016;1(3).
25. Zhang C, Cleveland K, Schnoll-Sussman F, McClure B, Bigg M, Thakkar P, et al. Identification of low abundance microbiome in clinical samples using whole genome sequencing. Genome Biology. 2015;16(1):265.
26. Salzberg SL, Breitwieser FP, Kumar A, Hao H, Burger P, Rodriguez FJ, et al. Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system. Neurology - Neuroimmunology Neuroinflammation. 2016;3(4).
27. Piro VC, Matschkowski M, Renard BY. MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling. Microbiome. 2017;5(1):101.


A
B
C
Kraken
MetaPhlAn2
F1 Score
Sensitivity
Positive Predictive Value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Phylum
Class
Order
Family
Genus
Species