Third-year progress report for thesis advisory committee

This manuscript (permalink) was automatically generated from mbhall88/TAC3_Report@9d05ee8 on September 21, 2020.

Authors

- **Michael B. Hall**
  0000-0003-3683-6208 · mbhall88 · mbhall88
  EMBL-EBI; University of Cambridge · Funded by EMBL International PhD Programme (EIPP)

Part A: Progress Report

Examining bacterial variation with genome graphs

Thesis Advisory Committee

- Zamin Iqbal (Supervisor) - EMBL-EBI
- John Marioni (Chair) - EMBL-EBI
- Georg Zeller - EMBL Heidelberg
- Estée Török - University of Cambridge

Starting Date: 12/10/2017
Qualifying Assessment Date: 06/07/2018
Second TAC Meeting: 15/10/2019
Third TAC Meeting: 13/10/2020

Executive summary

Genomics is now ubiquitous in clinical and public health microbiology, at least in the developed world. However, many significant challenges remain.

- Bacterial genomes harbour huge amounts of diversity, even within a species, and traditional reference-based approaches are problematic.
- Much of the variation in bacteria is fundamentally inaccessible to short reads.
- Long Nanopore reads are noisy, and SNP calling with this data is not properly benchmarked or standardised.
- Since *Mycobacterium tuberculosis* (Mtb) infects so many people, there is potential for considerable impact for clinical applications.
- There is also much to be gained from a high-quality pan-genome of Mtb as well as a detailed map of its enigmatic *pe/ppe* gene repertoire.

These motivations drive the following PhD thesis structure:

1. Develop algorithms and software for variant discovery using bacterial genome graphs, building on work of a previous student in the lab (my first paper, second author).
2. Benchmark Nanopore versus Illumina SNP calling, showing our algorithms meet the needs of clinical and public health users, validate, and publish (second paper).
3. Improve upon current whole-genome sequencing-based drug resistance prediction for Mtb using genome graphs.
4. Curate a high-quality reference pan-genome for Mtb that includes a detailed map of the *pe/ppe* genes.

Motivation and Background

In 2018, 1.4 million people died of tuberculosis (TB) globally, and over 10 million people fell ill to the disease. There were also 484,000 new cases of rifampicin-resistant TB reported, of which 78% were multi-drug resistant (MDR) [1]. Drug-resistant TB is not a new problem; the first clinical trial using Streptomycin to treat TB reported resistance and this outcome influenced the creation of the four-drug first-line regimen used today. Standard of care requires phenotypic testing of the infecting organism against these four drugs to ensure that appropriate treatment is prescribed. However, *Mycobacterium tuberculosis* (Mtb), the causative agent of TB, is a slow-growing organism and the gold-standard phenotypic tests (“mix bug and drug”) take around two months to complete. Thus, the traditional clinical diagnostic and treatment regimen is slow and expensive. Whole-genome sequencing offers a faster solution; recently it was shown that equivalent results are achievable by sequencing Mtb grown in “liquid culture” (also known as MGIT, Mycobacterial Growth Indicator Tube) after two weeks of culture in contrast to the two-month traditional (Lowenstein-Jensen) culture method [2]. A number of genes are implicated in drug resistance and predicting resistance from sequencing data based on a catalogue of resistance single-nucleotide polymorphisms (SNPs) and indels (insertions or deletions) works with high-specificity [3,4]. For the four first-line drugs, a study by the CRyPTIC consortium is the first of its kind to demonstrate that phenotyping is not required if genotype predicts susceptibility [5]. However, as the genetic basis for drug resistance is not entirely understood, there is still a sensitivity gap that differs drug-by-drug.

Public health requirements for TB diagnostics are **resistance prediction, species identification, and clustering** of genomes. The clusters are intended to contain potential transmission events, thus enabling more effective contact tracing. All of the main requirements are currently successfully achieved with Illumina sequencing technology. However, there are reasons to consider abandoning Illumina for Oxford Nanopore Technology’s (ONT) sequencer. First, there is cost - Illumina has raised its reagent prices considerably. Second, the burden of TB lies primarily in places where there is neither capital nor infrastructure for purchasing or maintaining a large machine. The third is speed: Votintseva *et al.* showed that it was feasible to go from patient to result via a Nanopore sequencer in 12.5 hours [6]. Since this publication, the yield and quality of Nanopore has risen. Therefore, a diagnostic that delivers results while the patient is in the clinic is now imaginable.

As an aside, it is worth noting that in high prevalence areas, there is no interest in transmission analysis, and therefore all of the information that they need (species and resistance) could be attained via deep amplicon sequencing. We accept this but ignore it and focus on a whole-genome solution nevertheless; both because of the better phylogenetic resolution, and because we do not yet know the relevant genetic mutations associated with new drugs.

The idea of using a single, linear reference genome to represent a population of individuals is not ideal. Even more so in the context of a bacterium such as *Salmonella enterica*, where two individuals can differ to such a degree that they only share 16% of their genes [7,8]. This fluidity of genetic content leads to the concept of a pan-genome; defined as the set of all genes observed in a species. Some genes are more common than others within this pan-genome, leading to the distinction of the ‘core’ and ‘accessory’ genome with the core being those shared across (most) all individuals and accessory being everything else. The effect of using a single-reference genome to describe variation within samples from the same species is best illustrated in Figure 1. In this toy example, the maximum number of SNPs we can hope to discover is only 56%. Clearly, comparing many samples requires better methods than single-reference based ones.

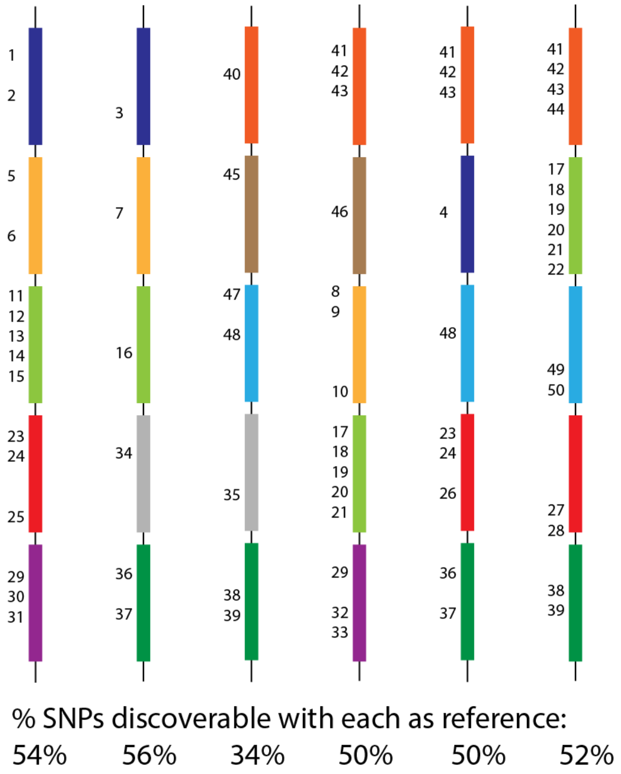

Figure 1: An illustration of how reference bias can impact variant-calling. Each vertical column signifies an individual genome, with the coloured blocks representing genes. Numbers label 50 segregating SNPs. The percentages at the bottom express the proportion of SNPs which can be detected using each of the 6 'genomes' as a reference by mapping perfect reads from the remaining 5 to this reference.

One of the first questions that arise from a computational point-of-view is how to apply current methods, which assume a single reference, to a pan-genome? An approach to answering this question that has been gaining considerable traction in recent years is that of genome graphs. This new paradigm uses a population reference graph (PRG) as its equivalent of a single, linear, reference genome. A PRG is built to represent variation seen within a population, conceding the fact that no single genome can accurately represent an entire species. To construct such a PRG, one takes a multiple sequence alignment (MSA) and collapses shared sequence and creates "bubbles", or branch points, where they do not (see Figure 2 for an illustration of this process). We define a collection of PRGs as a pan-genome reference graph (PanRG), which we will interchangeably refer to as a pan-genome. Thus far, genome graph methods have focused mainly on eukaryotes. Given the rich diversity of genetic content in the prokaryote world, it would seem genome graphs are more suited to utilisation there.

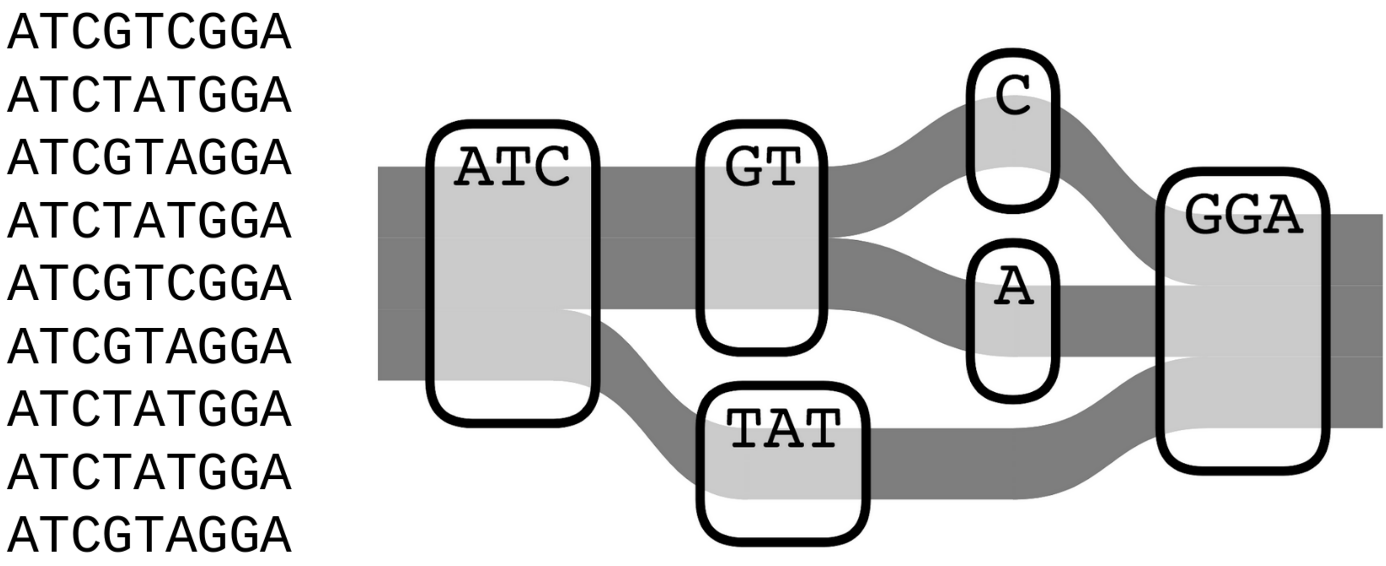

Figure 2: An illustration of how a population reference graph (PRG) is constructed. Regions (columns) of shared sequence are collapsed into a single node. Those that differ are split into "bubbles", or branching nodes.

Nanopore sequencing yields very long reads (up to 2.2Mbp [9]), the mean/mode read identity tends to fall in the range of 87-94% (using ONT's `guppy` basecaller), while on a consensus level, it can achieve 99.94% with assembly polishing [10].
Variant calling with Nanopore sequencing data has seen a somewhat slow development. Currently, the main tools that have had reasonable testing done and are versatile enough to detect both SNPs and (some) indels are `nanopolish` [11], Clair [12] and `medaka` [13]. A recent benchmark showed that Nanopore variant calling provides reliable diagnostic information for *Neisseria gonorrhoeae* [14]. However, to date, there has been no extensive Nanopore variant calling benchmark done for Mtb. Given the potential benefits of using genome graphs and long-read Nanopore sequencing for bacterial genomics, it makes sense to try and blend the two.

Pandora is a method being developed in the group to genotype across the *entire* pan-genome of a bacterial sample - not just the core. It does this by working with a PRG, rather than a linear reference. The method is based on the following intuition (similar to that behind the Li and Stephens model in population genetics): genomes evolve by recombination and mutation, and thus we ought to be able to approximate a $N + 1$ genome as a mosaic of the first N genomes. `pandora` maps Nanopore reads to a graph encoding of a PRG, infers a mosaic, and provides genotypes at all variants in the PanRG. Mapping is done using minimising k-mers [15] in a similar vein to that done by `minimap` [16], and is therefore fast. By using Nanopore sequencing data, it is also possible to infer gene order as, in general, a single read will contain multiple genes, as opposed to Illumina sequencing where multiple reads are required to span a single gene. Note `pandora` does not (prior to this work) include any facility for discovering novel variation.

Summary

This PhD seeks to develop new methods to enable Nanopore-based diagnostics and epidemiology for Mtb and other bacteria. By using PRGs as a strong prior [17,18], we should be able to mitigate the Nanopore error biases and indel issues. In the process, we aim to construct a high-quality reference pan-genome for Mtb that we hope will open previously inaccessible parts of its genome for investigation.

Part B: Training and career development

Publication Strategy

> Briefly outline the publication strategy; set priorities if needed.

List of publications, papers in press, preprints, manuscripts submitted/in preparation to date

Work plan and timeline for thesis submission

> Provide a work plan to be completed prior to thesis writing and a timeline for submission.

List of scientific courses and conferences attended to date and planned for next year

> Please list all events in reverse chronological order.

List of additional training, teaching and other relevant activities to date

> Please list all training and other activities in reverse chronological order.

Career development plan

Please describe your current long-term career aims (i.e. 3-5 years after PhD).

Please comment on the types of position you would like to apply to for after the PhD and your expected application timeline?

> If applying for postdoc positions, which fields/fellowships are you considering? If you are applying for non-academic careers, what type of role? Do you have target companies or organisations in mind?

What do you see as your strengths (2-3 skills)?

What do you see as your areas for improvement1 (2-3 areas)?

What are your career development priorities until the end of your contract?

> Please take into account both the scientific and non-scientific skills you will need to work on a successful and timely completion of the PhD. List the skills you still need to acquire to achieve your longer-term career aims.

What actions will you take to develop these skills?

> This may include training courses, opportunities to practice and develop these skills, or seeking feedback/guidance. N.B. fellows should take 1-2 non-scientific trainings or career development workshops per year, and at least one international conference during their PhD.

Part C: Impact of COVID-19 pandemic

> As outlined in the Procedures for COVID-19 related fellows’ contract extensions, students and their GTL should document the impact of the corona crisis on the PhD project. The TAC should assess the approximate project delay and discuss how this could be mitigated over the remaining PhD period. Please use the space below to outline the impact of corona crisis on your project (e.g. lost productivity / project delays).

References

1. **Global tuberculosis report 2019**
   Organisation mondiale de la santé
   (2019)
   ISBN: 9789241565714

2. **Mycobacterial DNA Extraction for Whole-Genome Sequencing from Early Positive Liquid (MGIT) Cultures**
   Antonina A. Votintseva, Louise J. Pankhurst, Luke W. Anson, Marcus R. Morgan, Deborah Gascoyne-Binzi, Timothy M. Walker, T. Phuong Quan, David H. Wyllie, Carlos Del Ojo Elias, Mark Wilcox, … Derrick W. Crook
   Journal of Clinical Microbiology (2015-04) https://doi.org/f65dtt
   DOI: 10.1128/jcm.03073-14 · PMID: 25631807 · PMCID: PMC4365189

3. **Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis**
   Phelim Bradley, N. Claire Gordon, Timothy M. Walker, Laura Dunn, Simon Heys, Bill Huang, Sarah Earle, Louise J. Pankhurst, Luke Anson, Mariateresa de Cesare, … Zamin Iqbal
   Nature Communications (2015-12-21) https://doi.org/f755tg
   DOI: 10.1038/ncomms10063 · PMID: 26686880 · PMCID: PMC4703848

4. **Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study**
   Timothy M Walker, Thomas A Kohl, Shaheed V Omar, Jessica Hedge, Carlos Del Ojo Elias, Phelim Bradley, Zamin Iqbal, Silke Feuerriegel, Katherine E Niehaus, Daniel J Wilson, … Tim EA Peto
   The Lancet Infectious Diseases (2015-10) https://doi.org/f3jjtq
   DOI: 10.1016/s1473-3099(15)00062-6 · PMID: 26116186 · PMCID: PMC4579482

5. **Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing**
   The CRyPTIC Consortium and the 100,000 Genomes Project
   New England Journal of Medicine (2018-10-11) https://doi.org/d9kj
   DOI: 10.1056/nejmoa1800474 · PMID: 30280646 · PMCID: PMC6121966

6. **Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples**
   Antonina A. Votintseva, Phelim Bradley, Louise Pankhurst, Carlos del Ojo Elias, Matthew Loose, Kayzad Nilgiriwala, Anirvan Chatterjee, E. Grace Smith, Nicolas Sanderson, Timothy M. Walker, … Zamin Iqbal
   Journal of Clinical Microbiology (2017-05) https://doi.org/f94vt4
   DOI: 10.1128/jcm.02483-16 · PMID: 28275074 · PMCID: PMC5405248

7. **Why prokaryotes have pangenomes**
   James O. McInerney, Alan McNally, Mary J. O'Connell
   Nature Microbiology (2017-03-28) https://doi.org/gfw8gq
   DOI: 10.1038/nmicrobiol.2017.40 · PMID: 28350002

8. **panX: pan-genome analysis and exploration**
   Wei Ding, Franz Baumdicker, Richard A Neher
   Nucleic Acids Research (2018-01-09) https://doi.org/gczkbr
   DOI: 10.1093/nar/gkx977 · PMID: 29077859 · PMCID: PMC5758898

9. **BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files**
Alexander Payne, Nadine Holmes, Vardhman Rakyan, Matthew Loose
Bioinformatics (2019-07-01) https://doi.org/gfnkxj
DOI: 10.1093/bioinformatics/bty841 · PMID: 30462145 · PMCID: PMC6596899

10. **Performance of neural network basecalling tools for Oxford Nanopore sequencing**
Ryan R. Wick, Louise M. Judd, Kathryn E. Holt
Genome Biology (2019-06-24) https://doi.org/gf4jwm
DOI: 10.1186/s13059-019-1727-y · PMID: 31234903 · PMCID: PMC6591954

11. **Real-time, portable genome sequencing for Ebola surveillance**
Joshua Quick, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, … Miles W. Carroll
Nature (2016-02-03) https://doi.org/f88652
DOI: 10.1038/nature16996 · PMID: 26840485 · PMCID: PMC4817224

12. **Exploring the limit of using a deep neural network on pileup data for germline variant calling**
Ruibang Luo, Chak-Lim Wong, Yat-Sing Wong, Chi-Ian Tang, Chi-Man Liu, Chi-Ming Leung, Tak-Wah Lam
Nature Machine Intelligence (2020-04-06) https://doi.org/d9kq
DOI: 10.1038/s42256-020-0167-4

13. **nanoporetech/medaka**
GitHub
https://github.com/nanoporetech/medaka

14. **High precision *Neisseria gonorrhoeae* variant and antimicrobial resistance calling from metagenomic Nanopore sequencing**
Nicholas D. Sanderson, Jeremy Swann, Leanne Barker, James Kavanagh, Sarah Hoosdally, Derrick Crook, Teresa L. Street, David W. Eyre, The GonFast Investigators Group
Genome Research (2020-09) https://doi.org/ghcbjr
DOI: 10.1101/gr.262865.120 · PMID: 32873606

15. **Reducing storage requirements for biological sequence comparison**
M. Roberts, W. Hayes, B. R. Hunt, S. M. Mount, J. A. Yorke
Bioinformatics (2004-07-15) https://doi.org/dkhs8w
DOI: 10.1093/bioinformatics/bth408 · PMID: 15256412

16. **Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences**
Heng Li
Bioinformatics (2016-07-15) https://doi.org/f8zxc3
DOI: 10.1093/bioinformatics/btw152 · PMID: 27153593 · PMCID: PMC4937194

17. **A Natural Encoding of Genetic Variation in a Burrows-Wheeler Transform to Enable Mapping and Genome Inference**
Sorina Maciuca, Carlos del Ojo Elias, Gil McVean, Zamin Iqbal
Lecture Notes in Computer Science (2016) https://doi.org/d9ks
DOI: 10.1007/978-3-319-43681-4_18

18. **Improved genome inference in the MHC using a population reference graph**
Alexander Dilthey, Charles Cox, Zamin Iqbal, Matthew R Nelson, Gil McVean

Nature Genetics (2015-04-27) https://doi.org/f7dt83
DOI: 10.1038/ng.3257 · PMID: 25915597 · PMCID: PMC4449272