Christophe Georgescu
19, allée des primevères
78 100 Saint-Germain-en-Laye
+33 (0)6 16 48 46 77

Using colored graphs to build a comprehensive map of genomic variability in *Mycobacterium tuberculosis* strains

Supervisor : Thomas Abeel

Academic year 2014-2015

Period of the internship : 1st of March 2015 - 31th of July 2015

The Delft Bioinformatics Lab
Delft University of Technology
Faculty of Electrical Engineering, Mathematics, and Computer Science
Mekelweg 4
2628 CD Delft, The Netherlands
+31 (0)15 27 86052

TU Delft Delft University of Technology

université PARIS DIDEROT PARIS 7

Contents

Introduction

The bacteria *Mycobacterium tuberculosis* (TB) is an obligate human pathogen from which 9 million people fell ill in 2013, and that killed 1.5 million people, most of them in countries with less health care capacities such as the former USSR and South Africa. For the latter, it is a real threat for people also suffering from HIV as the immune system is compromised[2,3]. Even without co-infections, it is the second highest cause of death due to a single pathogen, and the number of people infected with multidrug-resistant tuberculosis (MDR-TB) and extensively drug-resistant tuberculosis (XDR-TB) is increasing[4]. This is problematic as most TB infections remain latent with as much as an estimated 2 billion people with latent infection worldwide. Only about 10% of infected people develop an illness over their lifetime, yet the costs of treating MDR-TB and XDR-TB infections rises quickly, and the likelihood of the treatment's success gets much smaller.

Differences in virulence have been observed among populations originating from different regions, and this is thought to be due to a co-evolution with H. sapiens as it spread across the world [12, 13]. One example of known difference between the 'Ancient' and more virulent 'Modern' lineages, is the deletion of 2.8kb region of the genome called TbD1 [14].

Another important information about TB is that it lives within macrophages, meaning it is isolated from other bacterium and thus, unlike most other bacteria, does not use horizontal genetic transfers to acquire new genes. It is thought that since the start of the co-evolution with *H. sapiens*, TB has a reducing genome that only varies through deletions, mutations, inversions and active insertion sequences (ISs) [8, 11].

The first goal of this study is to compare a large number of TB genomes from different strains and lineages produced by recent large scale genome sequencing projects, to identify all the variants between them. The second is to make a pan-genome, a "blueprint" for the TB genome that would account for all found locations and combinations of specific DNA elements, and allow to infer a genome close to the ancestral genome of TB under the previous hypothesis. Both those results are to keep track of the origin of each information they carry so as to be able to overlap data about drug resistance and search for yet unidentified [5, 6], maybe lineage-specific, mechanisms that lead to drug resistance.

Material and Methods

The data used during this study are two TB reference strains that are well known, H37Rv [1, 10] and F11 [15], and 670 strains genome scaffolds from the TB-ARC initiative at Broad Institute [16]. Of those 670 genome scaffolds, most were not completely assembled and I developped a tool written in Python to order and orient the scaffolds using a reference genome, and the GSA library from REVEAL (see below) to extract Maximal Unique Matches (MUMs) which is based on the Needleman-Wunsch algorithm.

To build a phylogenetic tree from a distance matrix, the R packages phytools and its dependencies *ape* and *maps* were used. The tree was built following the neighbor joining algorithm that allows for different rates of evolution between branches, and it was midpoint rooted.

The output genomes were used for the multi-genome alignment and variant calling under the form of a colored graph, with a modified version of REVEAL (REcursiVe Exact-matching ALigner) [18] and its precursor ProBubble [17], that allowed a phylogenetic tree in Newick format as input to determine the aligning order. The phylogenetic tree reading was done using the Biopython module.

The primers sequences used for the identification of ISs, Phage sequences and documented repeats were obtained from literature [7, 9] and through annotations of the H37rv genome.

The identification of tandem repeats (TR) in genomes was done using the Tandem Repeats Finder tool [19] with default settings.

For graph visualisation, Cytoscape was used, and for genome viewing, annotation comparison and reading, GenomeView was used.

Results and Discussion: DNA alignment graph constructions

Ordering and Orientating scaffolds

Most genomes present in the 670 strains dataset used are not completely assembled and contain a variable number of scaffolds. To be able to compare those genomes using both existing and new tools, there was first a need to finish assembling them. In order to do so, for each

scaffold of the genome we want to finish, we search for all MUMs that are at least 1 000 bp long and store them, ordered by start position with regards to the reference. Then we only keep those that have another MUM within 1 000 bp of their tips on both the reference and the scaffold so as to discard mobile elements that were copied to another region of the genome. The remaining MUMs are collapsed in groups depending on their orientation if they are close together (<1000 bp), All which are in the same orientation as the reference are collapsed into a single group.

Scaffolds are then ordered for analysis by growing start position of their first MUM with regard to the reference because the start of the circular genome is by convention the origin of replication. The final orientation of the first scaffold is determined by the fact that the first MUM has to be in the same orientation as the reference. If it already is, the current orientation is kept, else it is inverted. We then look at each group of MUMs along the scaffold to know the expected orientation of the first group of the next scaffold. We also look out for gaps with respect to the reference, that usually indicate the start of an inversion. If a gap of at least 5 000 bp is found, the program looks at the average position of the two groups of MUMs around the gap in regards to the scaffold, and at their current orientation. Depending on those 2 values, they are 4 noteworthy possible outcomes illustrated in Fig. 1. If the scaffold contains a part of the sequence before the start of the inversion, it is kept in its current position, and only its future orientation may change. The program then searches for all scaffolds part of the inversion, that is, all scaffolds that have their MUMs in the gap, and inverts the order in which they are to be analysed. However, if the scaffold contains a part after the end of the inversion, the inversion of scaffolds analysis order includes the current scaffold.

Further filtering of MUMs is done when a small group of MUMs that overlaps with already analysed sequence with regard to the reference is encountered, and the scaffold still contains other groups of MUMs. When such a case is found, if the next group of MUMs fits without too large a gap, it is analysed as previously described, else the next scaffold is checked to see if it would fit in the gap, and if it is, both scaffolds are swapped in the analysis order so as to fill the gap first. This swapping can happen as long as there is a gap that can be filled by further scaffolds.

Finally, before writing the output, the program checks whether there are groups of MUMs that have a smaller starting position in regards to the scaffold than the origin of replication. If such groups of MUMs are found, they are removed from the first scaffold and appended to the end of the last scaffold, so as to keep the origin of replication as the start of the genome sequence.

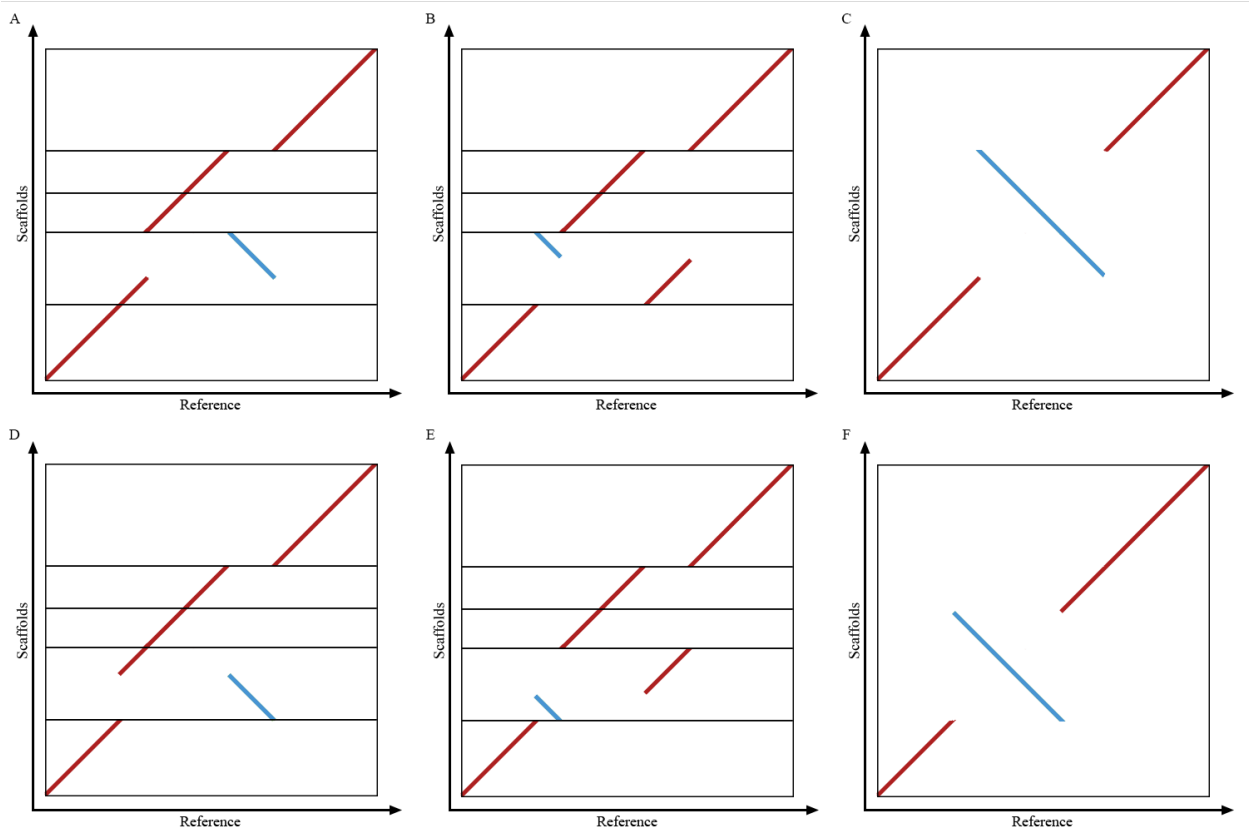

Fig. 1 : Plots of the MUMs between a reference genome (x-axis) and the scaffolds of the genome we want to order and orient (y-axis) sorted by growing start position of the first MUM of each scaffold in regards to the reference genome. Direct MUMs are in red and reverse complement ones in blue. The result from the ordering and orientating of A-B is C and of D-E is F. In A-B the 2nd scaffold contains sequence before the start of the inversion, so only the 3rd and 4th scaffolds' order need to be inverted. In D-E the 2nd scaffold contains sequence after the end of the inversion, so the 2nd, 3rd and 4th scaffolds' order need to be inverted.

Distance Matrix Building

To be able to effectively compare all the completed genomes present in our dataset, we needed to decide on an order in which to perform the comparisons based on genetic distances, which comes down to using a phylogenetic tree. In order to be able to quickly calculate a precise enough approximation of all pairwise genetic distances between two genomes, I wrote a tool in Java that efficiently calculates the Jaccard Similarity between the k-mers of all pairs of genomes given. Since the genome of TB is approximately 4.5 million bp, the size of k has to such as 4^k to be significantly higher than that number, but not too high either as this would penalise too much the distance between genomes due to SNPs, reducing their precision, and uselessly increase the costs in time and memory of the program. With this knowledge, the size of k that was used and verified to produce good results compared to SNPs distances is k = 15, but this setting can be defined by the user depending on the size of the genomes used. Furthermore, to remove the bias induced by inversions, both direct and reverse k-mers were used.

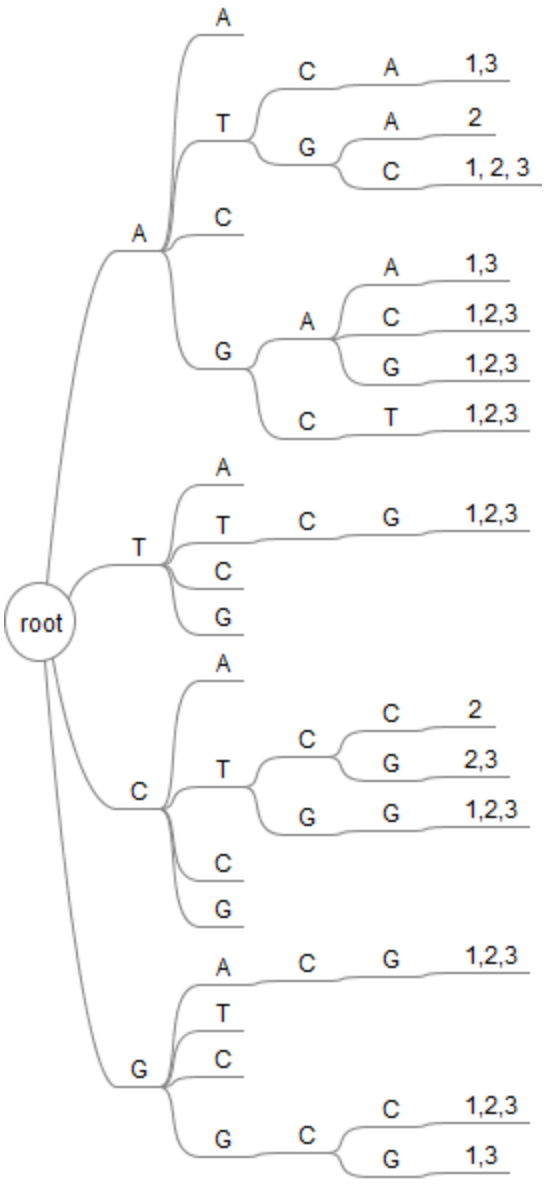

Fig. 2 : Example of the structure of the uneven tree used for calculatng the genetic distance. In this case k = 4 and the first 2 levels of the tree are created at the initialization. The remaining 2 levels are added whenever a k-mer that needs it is encountered. At each leaf's tip is the list of the IDs of genomes that contain that k-mer.

To allow for linear memory complexity, an uneven tree structure was used to store k-mers as illustrated in Fig. 2 (for k = 4) where only existing paths are created. To allow for k-mers to be added to the tree from multiple genomes at once with multiple threads, the 8 first levels of the tree are created at the start of program as they will statistically all exist. The

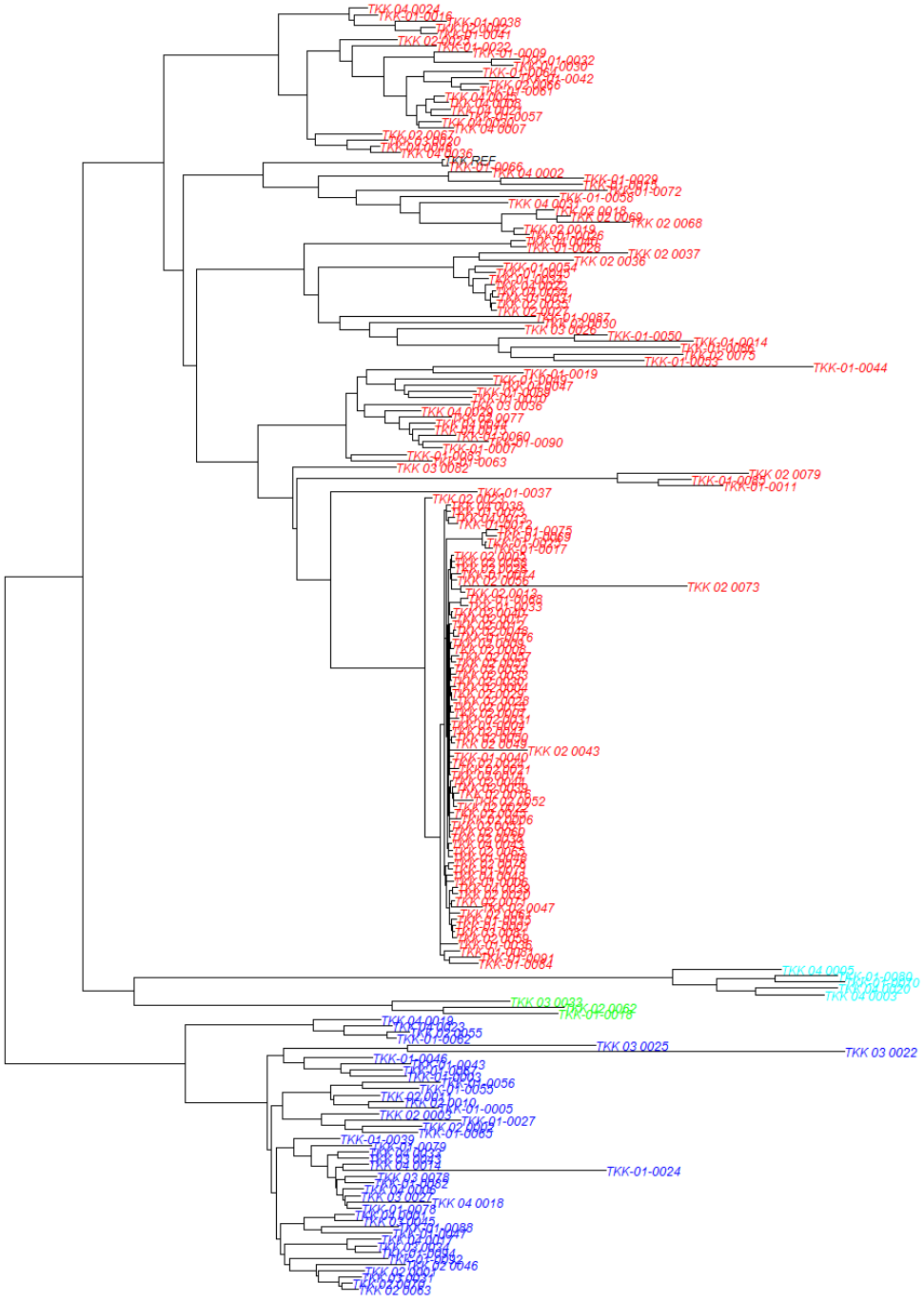

Fig. 3: Midpoint rooted tree obtained using the Neighbor-Joining algorithm on the distance matrix produced by the program, each tip being a different strain. The colors indicate the lineage of which the strains are part. Cyan is for lineage 1, blue for lineage 2, green for lineage 3, and red for lineage 4. The black strain is H37rv, which is part of lineage 4.

remaining 7 levels are only created when needed, with their access being synchronized to be thread-safe, that is, to prevent multiple concurrent nodes creation in the same node, which would result in loss of data.

Each final node, or leaf, contains an array with an identifier for each genome that contains the k-mer it represents. Since frequencies are not used, all identifiers can only appear once per leaf. This means that frequency variations due to a variation in the number of occurrences of the same IS does not affect computing. Once all k-mers have been added to the tree, each leaf is used to fill an union and an intersection matrix of k-mers between genomes. Those matrices are then used to calculate the Jaccard similarity matrix that gives all pairwise genetic distances between genomes by dividing the intersection matrix by the union matrix.

In order to improve execution speed in cases where the same dataset is used with only adding new genomes to it, it is possible to save the tree at the end of an execution and then reuse it when running the program again with the larger dataset.

The phylogenetic tree of 205 TB strains built using the method described earlier and the matrix distance obtained with this program can be seen in Fig. 3. We can easily see that all 4 lineages present in this dataset are correctly grouped within the tree and distinguished between them, which shows that the distance matrix is a good approximation of the genetic distance between strains.

Variant Calling

Using the ordered and oriented genomes as input, I used REVEAL to produce graphs of datasets containing 10, 16, 38, 100, 205 and 671 strains. The resulting graphs were linear only when all the strains from the dataset had the same inversions if no phylogenetic tree was used to determine an efficient merging order. But by using one, all the graphs obtained were linear, which allows to actually use them for analysis. This results are currently being used by 5 groups of Msc students at TU Delft to build an interactive visualization tool for sequence graphs as the current state-of-the-art, Cytoscape, has severe limitations that make it unsuited for large graphs analysis.

Results and Discussion: Structural pan-genome reconstruction

Pan-Genome

To be able to build a pan-genome from the few hundreds strains we assembled, we need to first artificially remove all inversions within them and annotate the inversions that were found. This is because otherwise, aligning a genome that has an inversion to one that does not would result in complex bubble structures that are not both not suited for continuing aligning new genomes to it and for the "blueprint" we want to build, as it would create duplicate paths that are all incomplete because they would miss features present in other paths. In order to do this, we use a variation of the program to order and orient scaffolds that reverts all found inversions and returns their start and end positions.

Then the genome of each strain needs to be represented as a graph where the nodes indicate a specific DNA element, such as an IS or a region of deletion. To search for ISs, Phages, known repeats, deletion regions, the genome is searched for 20 bp primers on both sides of the element with 2 mismatches allowed per primer. It is important to note that the primers taken are not completely on the edge of the element for ISs as that part has a higher variability than the rest of its sequence. For this reason, the primers are taken 40 bp from the edges. Furthermore, for elements such as ISs that have an expected size, if the program finds a match for both primers, it verifies whether the size found is of the same order as the expected one, and if it is the case, it is kept and later added to the graph as a node. One important structural detail of the graph is that no node contains any actual sequence, they only contain information relative to the element they model such as start and end position in that specific strain and an identifier if it is a documented element such as IS6110. Furthermore they contain a reference to the strain it was found in, referred to as the color, which is later used in merged graphs to keep track of the origins of each element.

Before any element is found, the genome is represented as a single regular sequence node. Whenever an element is found, the regular sequence node where the element fits is split into 3 nodes that are the element node, and the previous and the following regular sequence nodes. Edges are added from the previous regular sequence node to the element node, then from

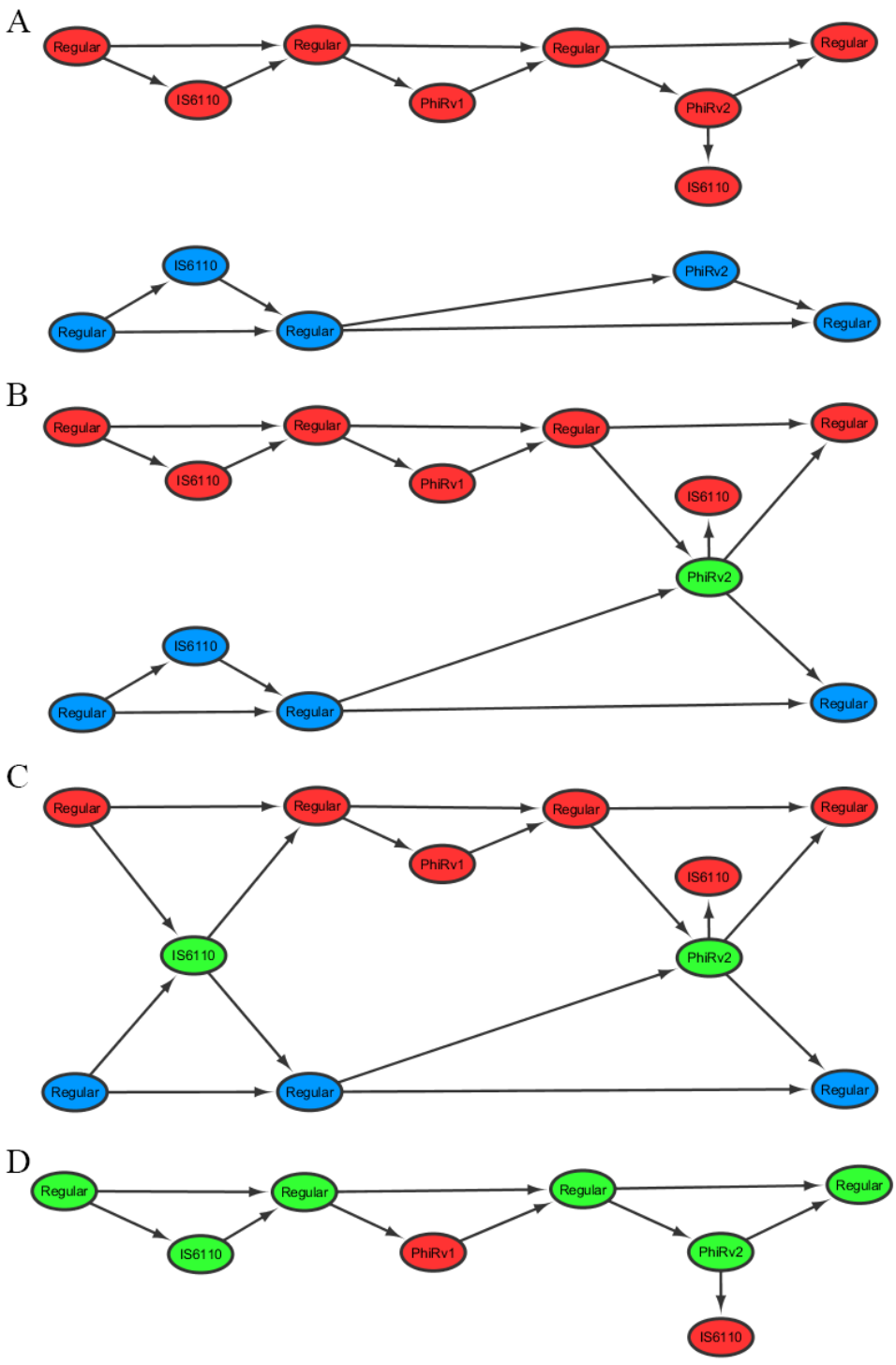

Fig. 4: Example of the merging steps for two single strain graphs, based on H37rv and F11 strains. The color of the node indicates which strains it is associated to, red for H37rv, blue for F11 and green for both. PhiRv1 and 2 are the two known Phage sequences in TB, and IS6110 is an active Insertion Sequence. A : the two graphs before merging. B : Merging of common Phages. C : Merging of common ISs. D : Merging of the remaining regular nodes.

the element node to the next regular sequence node, and an edge between the previous and the following regular sequence node, that models the fact that this element may be absent in other strains when they will be compared. If an element fits within another element, it is put in a node that extends the node of the element it fits in, unless it is a tandem-repeat (TR), in which case the TR node is not added. The order in which the elements are added to the graph is based on their size, genetic stability and significance, as for example there are two known insertion of a phage sequence into the genome of TB, and those have a much higher likelihood to have happened only once and then be conserved than the insertion of an IS that is a highly mobile element, or a TR that has a high length variability. The order used is to first add phages, then deletion regions, ISs and finally TRs. The last step is to add inversions if they were found in the pre-processing step of the genome. Currently, the tandem repeat finder settings are very conservative to only take the 50 largest TR, calculated by multiplying the period size and repeat count.

Merging two graphs is then done as illustrated in Fig. 4 by comparing the most stable elements first, in the same order they were added to the single-strain graphs. To check whether two element nodes from different graphs should be merged, if they are of the same type, their identifier, if they have any, is compare. If those matches, using the strain-specific start and end positions stored in the information sets, 20 flanking bp on each side of the elements, taken with an offset to be sure to be out of the element, are read from the file that contains the sequence directly at the position needed, and are then compared. If the mismatch count is lesser or equal to 2 for each of the two short sequences, the nodes are merged. Merging two nodes actually means adding the properties of the node from the smaller graph to the node from the bigger graph. The information sets stored in the “absorbed” node are appended to the ones of the remaining node, and so are the incoming and outgoing edges connecting it with the rest of the graph, thus making a “bridge” between the two graphs and incrementally aligning them overall. The element nodes that can not be merged between graphs because they are only present in one of them are kept and inserted on the graph that does not have them, in the same way as for single-strain graphs, using the knowledge brought by the adjacent element nodes that were merged to determine the position where they should be inserted.

The implementation is done in Java, and to reduce the time it takes for the program to run, multiple single-strain graphs are built in parallel, the order being the one there are going to be merged in, which is taken from the phylogenetic tree built previously. When all the graphs required by a merging are ready, the merger starts in parallel. Multiple clades, or subgroups within the tree, can be merged at the same time as long as all the graphs needed are ready.

For visualisation, the graphs are currently outputted to both SIF format for visualizing the graph itself in Cytoscape, and GFF3 format to be able to compare the genetic features found on the graph to any sequence using GenomeView.

Conclusion

The tools developed above allow the use of large datasets for comparative genotype analysis by both finishing the ordering and orientating of assembly scaffolds, and defining an order in which the genomes should be incrementally compared. Although currently only 671 strains are used for this study, once all tools are finished, the goal is to increase that number to a few thousands.

By using a combination of the colored graphs obtained from both REVEAL and the pan-genome building, as well as the phylogenetic trees, overlapping the information on drug resistances of all strains should allow to identify both SNPs sources of drug resistance, and larger structural mechanisms that induce drug resistances. The pan-genome will also prove useful for new reference-based genome assembling, as the blueprint it provides has less limitations than a single reference genome. It will also help us improve our understanding of what the ancestral TB genome looked like when it split from other species, and thus more insight on how TB evolution works.

References

1. S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, III, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies,K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M.-A. Rajandream, J. Rogers, S. Rutter, K.Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead & B. G. Barrell. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537-544 (11 June 1998) | doi:10.1038/31159;

2. Edlin, B. R., J. I. Tokars, M. H. Grieco, J. T. Crawford, J. Williams, E. M. Sordillo, K. R. Ong, J. O. Kilburn, S. W. Dooley, and K. G. Castro. 1992. An outbreak of multidrug-resistant tuberculosis among hospitalized patients with the acquired immunodeficiency syndrome. N. Engl. J. Med. 326:1514–1521.
3. Fischl, M. A., R. B. Uttamchandani, G. L. Daikos, R. B. Poblete, J. N. Moreno, R. R. Reyes, A. M. Boota, L. M. Thompson, T. J. Cleary, and S. Lai. 1992. An outbreak of tuberculosis caused by multiple-drug-resistant tubercle bacilli among patients with HIV infection. Ann. Intern. Med. 117:177–183.
4. Breathnach, A. S., A. de Ruiter, G. M. Holdsworth, N. T. Bateman, D. G. O'Sullivan, P. J. Rees, D. Snashall, H. J. Milburn, B. S. Peters, J. Watson, F. A. Drobniewski, and G. L. French. 1998. An outbreak of multi-drug-resistant tuberculosis in a London teaching hospital. J. Hosp. Infect. 39:111–117.
5. Heym, B., P. M. Alzari, N. Honore, and S. T. Cole. 1995. Missense mutations in the catalase-peroxidase gene, *katG*, are associated with isoniazid resistance in *Mycobacterium tuberculosis*. Mol. Microbiol. 15:235–245.
6. Hirano, K., M. Takahashi, Y. Kazumi, Y. Fukasawa, and C. Abe. 1997. Mutation in *pncA* is a major mechanism of pyrazinamide resistance in *Mycobacterium tuberculosis*. Tuber. Lung Dis. 78:117–122.
7. Supply, P., Mazars, E., Lesjean, S., Vincent, V., Gicquel, B. and Locht, C. (2000), Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. Molecular Microbiology, 36: 762–771. doi: 10.1046/j.1365-2958.2000.01905.x
8. Bentley SD, Comas I, Bryant JM, Walker D, Smith NH, Harris SR, Thurston S, Gagneux S, Wood J, Antonio M, Quail MA, Gehre F, Adegbola RA, Parkhill J, de Jong BC. The genome of Mycobacterium africanum West African 2 reveals a lineage-specific locus and genome erosion common to the M. tuberculosis complex. PLoS Negl Trop Dis. 2012;6(2) e1552. doi:10.1371/journal.pntd.0001552.PMID: 22389744; PMCID: PMC3289620.
9. Stephen V. Gordon, Beate Heym, Julian Parkhill, Bart Barrell, and Stewart T. Cole. New insertion sequences and a novel repeated sequence in the genome of *Mycobacterium tuberculosis* H37Rv Microbiology, April 1999 145:881-892; doi:10.1099/13500872-145-4-881
10. Claudio U. Köser, Stefan Niemann, David K. Summers, John A.C. Archer, Overview of errors in the reference sequence and annotation of Mycobacterium tuberculosis H37Rv, and variation amongst its isolates, Infection, Genetics and Evolution, Volume 12, Issue 4, June 2012, Pages 807-810, ISSN 1567-1348, http://dx.doi.org/10.1016/j.meegid.2011.06.011.
11. Anthony G. Tsolaki, Aaron E. Hirsh, Kathryn DeRiemer, Jose Antonio Enciso, Melissa Z. Wong, Margaret Hannan, Yves-Olivier L. Goguet de la Salmoniere, Kumiko Aman, Midori Kato-Maeda, and Peter M. Small. Functional and evolutionary genomics of *Mycobacterium tuberculosis*: Insights from genomic deletions in 100 strains PNAS 2004 101 (14) 4865-4870; published ahead of print March 15, 2004, doi:10.1073/pnas.0305634101
12. Brites D, Gagneux S. Co-evolution of *Mycobacterium tuberculosis* and *Homo sapiens*. *Immunological Reviews*. 2015;264(1):6-24. doi:10.1111/imr.12264.
13. Sebastien Gagneux, Kathryn DeRiemer, Tran Van, Midori Kato-Maeda, Bouke C. de Jong, Sujatha Narayanan, Mark Nicol, Stefan Niemann, Kristin Kremer, M. Cristina Gutierrez, Markus Hilty, Philip C. Hopewell, and Peter M. Small. Variable host–pathogen compatibility in *Mycobacterium tuberculosis* PNAS 2006 103 (8) 2869-2873; published ahead of print February 13, 2006, doi:10.1073/pnas.0511240103
14. Gordon, S. V., Brosch, R., Billault, A., Garnier, T., Eiglmeier, K. and Cole, S. T. (1999), Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. Molecular Microbiology, 32: 643–655. doi:10.1046/j.1365-2958.1999.01383.x
15. *Mycobacterium tuberculosis* Comparative Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/)
16. TB-ARC initiative, Broad Institute (broadinstitute.org)
17. http://www.liacs.nl/assets/Masterscripties/CS-Studiejaar-2013-2014/2013-2014JasperLinthorst.pdf
18. REVEAL : https://github.com/jasperlinthorst/reveal
19. Tandem Repeat Finder https://tandem.bu.edu/trf/trf.definitions.html

Acknowledgments

I would like to express my deep gratitude to Dr Thomas Abeel, my research supervisor, for his guidance and constructive discussions throughout my internship, as well as for providing me with the data needed to my research. I would also like to thank Mr. Thies Gehrmann for his help in working with the compute servers, and all the researchers of the laboratory for our talks about my research project.

Finally, I wish to thank my family for all their support and encouragement throughout this internship.

Abstract

Even after the discovery of antibiotics, *Mycobacterium tuberculosis* (TB) remains one of the most lethal human pathogens. To make things worse, in recent years, drug-resistant strains have become more common worldwide. This is mostly due to bad choice of antibiotics to use depending on the strain, or treatments stopped before their end. Some of the mechanisms that lead to drug-resistance are known, but not all of them, and it thus important to continue studying them and how they are acquired. One way of doing so is by using new comparative algorithms on large genomic datasets that cover a wide range of TB genomic diversity. Such studies can also help improve our understanding of TB evolution and what its ancestral genome looked like when it became an obligate human pathogen. This could in turn help other future investigations focused on TB.

Résumé

Même depuis l'apparition des antibiotiques, *Mycobacterium tuberculosis* (TB) reste l'un des agents pathogènes humains les plus létales. Et pour ne pas améliorer la situation, au cours des dernières années, une augmentation du nombre de souches résistantes aux antibiotiques et de leur étendue a été mondialement observée. Ceci est principalement du à de mauvais choix d'antibiotiques à utiliser en fonction de la souche, ou a des traitements interrompus avant leur fin. Certains des mécanismes menant à une résistance aux antibiotiques sont connus, mais pas tous, et il est donc important de continuer à étudier ces mécanismes et leur acquisition. Une façon de procéder est en utilisant de nouveaux algorithmes de comparaisons sur de larges sets de données génomiques couvrant un panel étendu de diversité de TB. De telles études peuvent aussi aider à améliorer notre compréhension de l'évolution de TB ainsi que ce à quoi ressemblait son génome ancestral lorsqu'elle est devenue un pathogène obligatoire de l'homme. Ceci pourrait à son tour être utile dans le cadre d'autres recherches se focalisant sur TB.