Spolmap: an enriched visualization of CRISPR diversity

Christophe Guyeux, Guislaine Refrégier, and Christophe Sola

April 7, 2022

Abstract

In the study of the evolution of various bacteria, the content of the CRISPR locus has proven to be quite useful. This locus has been made famous because it allows for simple and inexpensive genome editing. And bacteriologists are used to studying this locus, through tools such as spoligotyping, in order to experimentally be able to determine the lineage or even the sub-lineage of a given strain, and to deduce an optimal antibiotic cocktail. The problem is that the study of the content of this locus is very often delicate and difficult. Therefore, we propose in this paper a new way of representing them, which makes sense biologically speaking, and which allows a simplified and enriched study of the CRISPR content. After explaining how to extract this locus from Whole Genome Sequencing data, we propose an embedding of this locus in a high dimensional space, followed by a reduction to dimension 2, which makes sense of the content. This method is applied to the case of the *Mycobacterium tuberculosis* complex, and a discussion is proposed to list the advantages of this approach.

1 Introduction

Tuberculosis remains one of the most deadly diseases in the world today, and its incidence has even increased in recent years following the COVID epidemic. This disease is caused by a bacterium called *Mycobacterium tuberculosis*, which was described more than 100 years ago. But since the discovery of the first antibiotics, little progress has been made in the fight against this bacterium, and the development of resistant to multi-resistant strains is certainly a problem. Therefore, any additional knowledge on this bacterium and its evolution is welcome.

To a lesser extent, this can also be established for other diseases such as salmonella or legionella. And the various bacteria involved in these diseases have the particularity to be studied through the content of the CRISPR locus. In some of them, such as the bacteria of the *Mycobacterium tuberculosis* complex (MTC), this locus is no longer active and now only faces deletions. In this case, the difference between the current content and the ancestral content [10,11] is a

specific characteristic of a strain, which allows it to be classified, for example, in a particular lineage. This barcode allowing the analysis of strains based on their content in CRISPR is called spoligotype in *M.tuberculosis*. In other bacteria, this locus remains active, but it also contains sub-patterns which can be studied to gain knowledge. But in any case, the deciphering and analysis of this content is still a delicate task, and there is currently no tool to help study these complex motifs.

The objective of this paper is to propose a new way to represent these CRISPR motifs, illustrating it in the case of MTC. The idea is mainly to plunge them intelligently into an N-dimensional space, and then to do a quality dimension reduction, to obtain a planar view that makes sense biologically speaking, and that is easier to study. The various steps required to achieve this result are fully detailed, from downloading the genome directly from the sequencing, to extracting the lineage information and the CRISPR locus content. The latter is obtained here by a De Bruijn graph approach, after extraction of the reads of interest, and requires a manual step. Finally, the embedding is also fully detailed.

The remainder of this article is as follows. In the next section, basic recalls regarding the CRISPR locus and the spoligotyping technics is recalled. Section 3 is devoted to the proposed approach, which is fully detailed. It is experimented in Section 4 in the case of the *Mycobacterium tuberculosis* complex. This result section is followed by a discussion that extends this work. This article ends by a conclusion section, in which the contribution is summarized and intended future work is outlined.

2 Basic recalls

The CRISPR locus of Mycobacterium tuberculosis complex (MTC), the agent of tuberculosis (TB), was first described in 1993 as the "Direct Repeat" locus [9, 20]. It consists of 36 nucleotide-repeats interspersed with single spacers averaging 37nt (range: 25-45nt). The repeats were quickly referred to as direct repeats and abbreviated as such (DR), and the sequences of a single spacer + a DR were called direct-variant repeats (DVR). The first two isolates sequenced (*M. tuberculosis* H37Rv and *M. bovis BCG*) yielded 43 different spacer sequences. The detection of their presence/absence led to the development of the innovative method of "spoligotyping" [15]. This method has become very popular because of its ease of implementation and its digital format. It has indeed allowed us to decipher the structure of the global MTC population [4]. More recently, whole genome sequencing (WGS) studies have indeed confirmed that for the 6 major human lineages (L1 to L6) and many sub-lineages, the spoligotypic signature allows an approximate taxonomic assignment [16]. However, some generic signatures remain either meaningless, imprecise, or convergent, which largely justifies the use of SNPs as preferred taxonomic markers, whether at the global or national level [6], or for L4 [22], L1 [19], or L2 [12, 21].

As in other species with functional CRISPRs, this locus is accompanied

by a set of CRISPR-associated genes (cas). Their number and nature make the MTC CRISPR type fall into the Type III-A group in the CRISPR-Cas taxonomy [17]. The CRISPR-Cas locus has recently been shown to be active in the H37Rv system [26]. Yet some or all of the region is deleted in several MTC sublines [8]. Another important question is whether deletion of some of the cas genes in the CRISPR-Cas locus can promote genomic instability in some epidemic strains of MTC [23].

The genomic diversity of the CRISPR locus was studied in detail as early as 2000 in a study by J. van Emdben et al. showing that spacer duplications, spacer variations and IS6110 insertion sites could be found in the different phylogenetic lineages of TCM [25]. However, it involved a very small sample ($n = 34$) and did not include any investigation of cas genes [3, 7]. Understanding the evolutionary dynamics of this locus now requires exploration of the CRISPR-Cas region on an extensive data set.

The classical in vitro approach of spoligotyping lists the presence or absence of a well-known list of spacers in a sample. This robust method has been widely applied in vitro [15]. However, this approach did not explore many features, such as whether the order of spacers is different in one strain or the other. It also did not reveal whether there was duplication of any part of the locus. Finally, it did not provide information on the presence of insertions such as IS6110, nor on the existence of single nucleotide polymorphisms (SNPs) in its direct repeats or spacers. This masks potential functionally important changes in the loci, and makes it impossible to conduct in-depth evolutionary studies. New in silico approaches (SpolPred, SpoTyping) have been developed to produce spoligotypes from genome reads [5]. Although these methods reveal the presence/absence of spacers in a similar manner, they have the same limitations as in vitro spoligotyping techniques. Last, but not least, their exploitation is frequently difficult due to the existence of numerous patterns that are difficult to relate together.

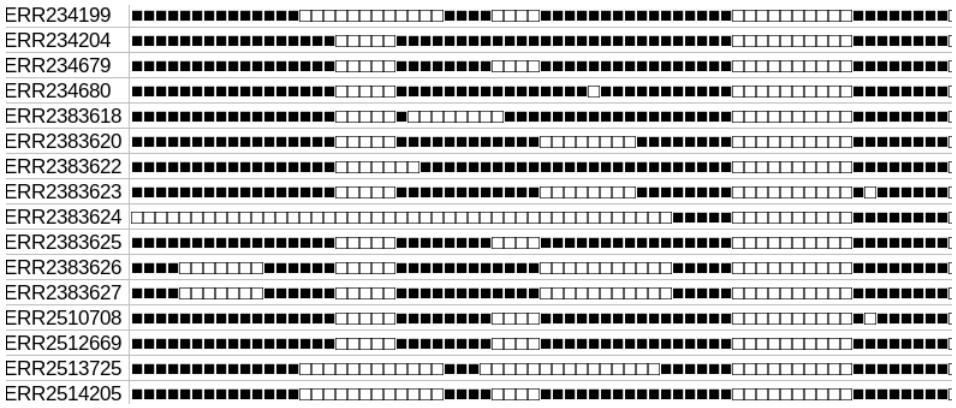

Figure 1: Example of spoligotypes of Lineage 5, defined by their accession numbers.

3 The proposed approach

Firstly, we have to download the genomes of interest, in the form of a Sequence Read Archive (SRA), for example with the fastq-dump command from the NCBI SRA Toolkit [1].

The first key step is then to extract the spoligotype from this SRA. Various tools exist in the literature [5], but none of them are suitable for our approach. Some of them require assembled genomes; however, the CRISPR locus is rich in repeated sequences (DR and IS6110), and its very difficult assembly often leads to gross errors. Others are compatible with SRA-type inputs, but the quality of the spoligotypes produced has proven insufficient for our needs. The problem is that these tools are not specific to MTBC: some just tell if a CRISPR locus is present, while others try to find the content of the locus without a priori knowledge of the spacer sequences to find. This is why we have chosen to follow the new approach proposed in [13].

We first build a blast database from the reads contained in the SRA file, then we blast the sequences of interest (spacers, DR, and CAS genes). To increase the diversity of the retained sequences and to guard against the discarding of reads containing mutations, we transform these reads that match into k-mers, where k is three-fourths the size of the reads. We then construct a De Bruijn graph from these k-mers, in which the nodes are these sequences, and there is an edge from a node i to a node j if and only if a suffix of i is a prefix of j.

We then traverse each of the connected components of this graph G. A first node is drawn at random, and we traverse the related component from vertex to vertex, as long as it is possible. The vertices thus traversed are removed from the G graph, and this traversal produces by concatenation of the sequences a part of the CRISPR locus. We then identify the elements of this part using the list of sequences of interest (spacers, DRs, CAS and IS6110 genes), and we thus obtain a first contig with the details of its content. This process is repeated until the vertices of G are exhausted (the process necessarily has an end). The contigs are then sorted by size, and the final assembly is done by hand.

Note that, with few exceptions, there is always at least one IS6110 in the CRISPR locus. Given the size of this insertion sequence, compared to k, as well as its large number of copies in the genome, contig construction by iterating on G necessarily stops when an IS6110 is encountered. Similarly, we have recently shown the existence of duplicated spacers (singly or in tandem), and these duplications are also a cause of stopping contig reconstruction [20]. These elements explain why a human final step is required.

Once the CRISPR of the strain has been reconstructed and the spoligotype deduced, we still need to determine the lineage of the genome. This is done by taking the list of SNPs per lineage from Coll [6], then extracting from the h37Rv reference a 40 base pair sequence around the SNP position, and blasting the result onto the database defined above. A majority vote is then needed to assign a lineage to this strain.

Let us now assume that the set of our spoligotypes of interest contains a total of N different gaps, e.g. (15,26); (30,34); (51,60) for the first spoligotype

in Fig. 1. The next step is to transform each spoligotype into a point in an N-dimensional space, as follows. The gaps are sorted according to the lexicographic order, and an integer from 1 to N is then assigned to each gap positioned according to this order. The vector corresponding to the considered spoligotype is then constituted as follows: we place a 1 at each associated gap position, and a 0 everywhere else. In this way, we obtain a binary vector of size N, where each distinct spoligotype has a different position in space. In this N-dimensional space, points close in Manhattan distance correspond to similar spoligotypes. It remains then to make a reduction to dimension 2, using the t-SNE algorithm [24].

The implementation has been realized in Python 3.10, and an interface provided in Tk is available upon reasonable request.

4 Obtained results

An example of what Spolmap can lead to is shown in Figure 2 in the MTC case. In this figure, each point corresponds to one strain (in WGS genome form), while the color of these points is made according to the strain lineage. In this picture, we can find:

- the lineage 1, indo-oceanian, in pink at the bottom of the picture;
- the lineage 2, Beijing, in two blue clusters: a spread out cluster at the center of the figure, corresponding to ancient strains, and a concentrated one at its bottom left, for the modern ones;
- the Middle-East lineage 3 in violet, on the right part of the cloud;
- the Americano-European lineage 4 in red, occupying the upper half of the figure;
- the two African lineages 5 and 6, respectively in brown and dark green, a little bit off-center;
- the Ethiopian lineage 7 in yellow, alone at the center of the cloud;
- the animal strains in green, a few circles in the bottom left part of the cloud.

Many conclusions can be drawn from this point cloud obtained from the spoligotypes. First, there are as many clusters as there are lineages, with sub-clusters associated with sub-lineages. Some lineages are very well separated and present a really pure cluster, such as lineages 5, 6 and 7. We also find, in the upper right part, lineages 2 to 4, and in the lower left part, lineages 1, 5, 6 and animal, and we know that these two subgroups are phylogenetically separated. The clusters of lineages 1 to 4 extend to the center of the cloud, arguing for a common origin of the tuberculosis complex, whose ancestor is probably *M.canettii* **??**.

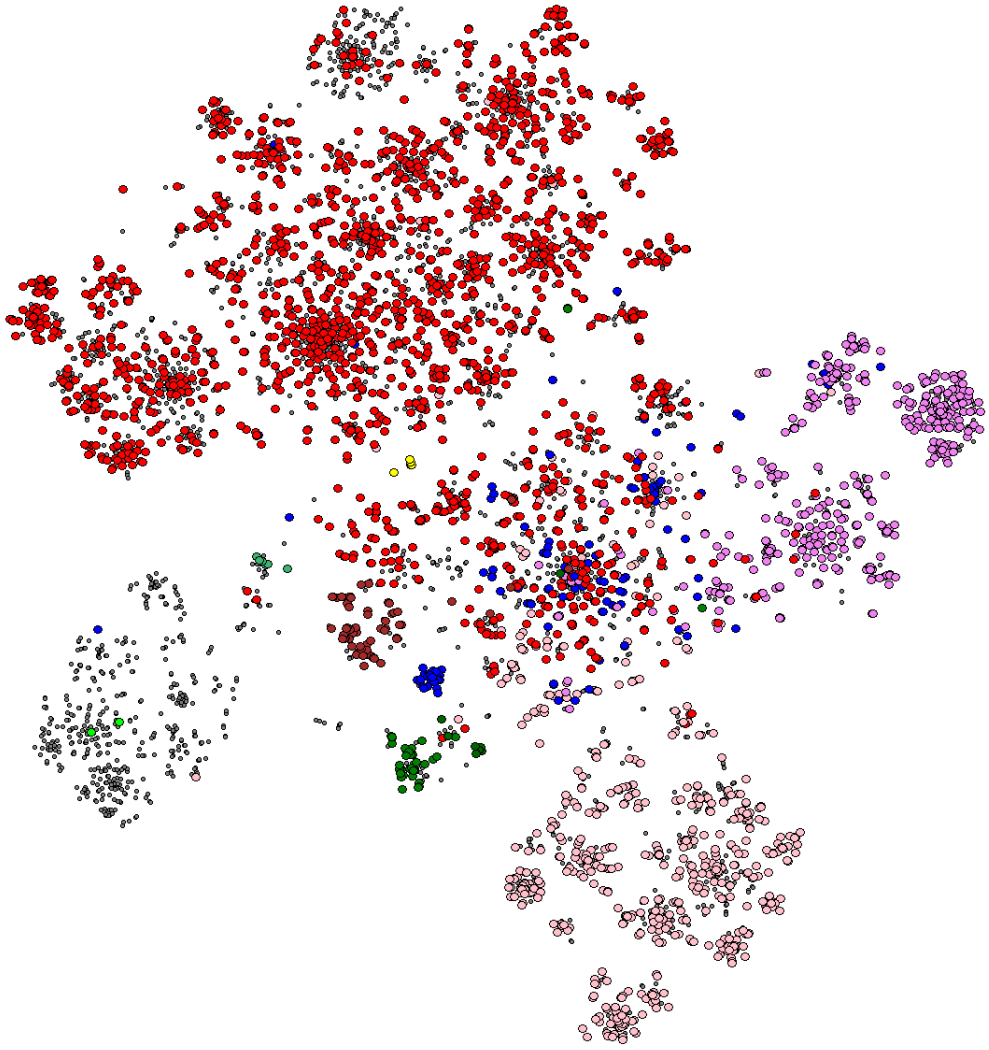

Figure 2: A 2D visualization of the MTC's spoligotypes

In some sublineages, the corresponding subcluster is only partially colored, suggesting a poor definition of said sublineage (an overly restrictive lineage SNP). This is evident in the circular cluster at the top of lineage 4, for example. We also see a whole big gray cluster with a few green dots in it, which would tend to show our very poor knowledge of animal TB.

Another lesson is that, in general, the complex is well described by the existing lineages: apart from the animal lineage, there are no large new gray clusters to investigate. However, there are several small grey clusters, the size of the lineage 8 cluster, which are isolated, such as the small ten points between

the animal lineage and lineage 6 (in dark green). These small clusters probably reflect small exotic lineages, which should be further investigated to have a full understanding of the TB mycobacterial complex.

Finally, it is undeniable that the SNP-based lineage data and the spoligotype hole data are strongly correlated, arguing for a co-occurrence of these two evolutionary mechanisms at the same time.

5 Discussion

The spoligotype has long been considered useful for many lineage identifications, with for example the absence of spacers 18-22 on the one hand, and 51-60 on the other, as a definition of Lineage 5, cf. Figure 1. However, its interpretation is often quite delicate, if one is content to focus on a linear representation. We have shown that a representation of the latter in high dimension followed by a reduction to the 2-dimension reveals something quite coherent, and a vision both summarized and useful.

We also saw that this approach made it possible to find new lineages or sub-lineages, to highlight definitional problems in the latter, as well as poorly explored areas in the diversity of the considered species. Such an approach is not limited to the bacteria that cause tuberculosis. It can potentially be applied to any bacterial species with a CRISPR locus that is no longer functional (and therefore, for which the number of spacers is finite), if exist. It can also be used in bacteria whose locus is active, but for which subgroups of spacers appear, and allow a use for characterization, such as in salmonella or legionella (or, in some plant pathogenic bacteria).

Note that we have only used an elementary definition of distance between two spoligotypes, and other choices are possible. Similarly, t-SNE is not the only recent tool for dimension reduction, and techniques such as the so-called Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP, [18]) could lead to other representations, equally useful and complementary.

Finally, dimension reduction techniques are often coupled with outlier detection methods [2,14], and the latter seem promising either to rule out a strain that does not belong to the complex under consideration, or to highlight new lineages that were previously unknown (recall that lineages 8 and 9 have been discovered in the last three years: there are probably new things to discover).

6 Conclusion

Based on spoligotyping in *M.tuberculosis*, we have proposed a new way of representing the CRISPR locus, which both makes biological sense, and makes the study easier and more thorough. This approach has been fully detailed, from genome upload to locus extraction, through plotting in high-dimensional space and to the final dimension reduction step. This approach allows to detect outliers, to show the diversity of the studied strains and their respective rela-

tionships. It also allows to detect new lineages or sub-lineages, and to highlight possible inconsistencies.

For our next works, we wish to make this tool accessible through a neat interface, and propose versions for tuberculosis, salmonella and legionella. We then wish to integrate all available genomes (more than 100000 genomes in the case of *M. tuberculosis*), and then to search for unknown lineages. Finally, we wish to integrate this representation in a larger and complete tool, including for example the determination of lineages and MIRU-VNTRs in *M.tuberculosis*.

References

[1] SRA toolkit development team. `https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software`. Accessed: 2022-03-16.

[2] Irad Ben-Gal. Outlier detection. In *Data mining and knowledge discovery handbook*, pages 131–146. Springer, 2005.

[3] Charles Bland, Teresa L Ramsey, Fareedah Sabree, Micheal Lowe, Kyndall Brown, Nikos C Kyrpides, and Philip Hugenholtz. Crispr recognition tool (crt): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC bioinformatics*, 8(1):1–8, 2007.

[4] Karine Brudey, Jeffrey R Driscoll, Leen Rigouts, Wolfgang M Prodinger, Andrea Gori, Sahal A Al-Hajoj, Caroline Allix, Liselotte Aristimuño, Jyoti Arora, Viesturs Baumanis, et al. Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (spoldb4) for classification, population genetics and epidemiology. *BMC microbiology*, 6(1):1–17, 2006.

[5] Francesc Coll, Kim Mallard, Mark D Preston, Stephen Bentley, Julian Parkhill, Ruth McNerney, Nigel Martin, and Taane G Clark. Spolpred: rapid and accurate prediction of mycobacterium tuberculosis spoligotypes from short genomic sequences. *Bioinformatics*, 28(22):2991–2993, 2012.

[6] Francesc Coll, Ruth McNerney, José Afonso Guerra-Assunção, Judith R Glynn, João Perdigão, Miguel Viveiros, Isabel Portugal, Arnab Pain, Nigel Martin, and Taane G Clark. A robust snp barcode for typing mycobacterium tuberculosis complex strains. *Nature communications*, 5(1):1–5, 2014.

[7] Kiatichai Faksri, Eryu Xia, Jun Hao Tan, Yik-Ying Teo, and Rick Twee-Hee Ong. In silico region of difference (rd) analysis of mycobacterium tuberculosis complex from sequence reads using rd-analyzer. *BMC genomics*, 17(1):1–10, 2016.

[8] Paul Jeffrey Freidlin, Israel Nissan, Anna Luria, Drora Goldblatt, Lana Schaffer, Hasia Kaidar-Shwartz, Daniel Chemtob, Zeev Dveyrin, Steven Robert Head, and Efrat Rorman. Structure and variation of crispr

and crispr-flanking regions in deleted-direct repeat region mycobacterium tuberculosis complex strains. *BMC genomics*, 18(1):1–14, 2017.

[9] Peter MA Groenen, Annelies E Bunschoten, Dick van Soolingen, and Jan DA van Errtbden. Nature of dna polymorphism in the direct repeat cluster of mycobacterium tuberculosis; application for strain differentiation by a novel typing method. *Molecular microbiology*, 10(5):1057–1065, 1993.

[10] Christophe Guyeux, Bashar Al-Nuaimi, Bassam AlKindy, Jean-François Couchot, and Michel Salomon. On the reconstruction of the ancestral bacterial genomes in genus mycobacterium and brucella. *BMC Systems Biology, IWBBIO 2017 Special Issue*, 12(5):100, 2018.

[11] Christophe Guyeux, Michel Salomon, Bashar Al-Nuaimi, Bassam AlKindy, and Jean-François Couchot. Ancestral reconstruction and investigations of genomic recombination on some pentapetalae chloroplasts. *Journal of Integrative Bioinformatics*, *:20180057, 2019.

[12] Christophe Guyeux, Gaetan Senelle, Guislaine Refrégier, Florence Bretelle-Establet, Emmanuelle Cambau, and Christophe Sola. Connection between two historical tuberculosis outbreak sites in japan, honshu, by a new ancestral mycobacterium tuberculosis l2 sublineage. *Epidemiology and Infection*, 150:e56, 2022.

[13] Christophe Guyeux, Christophe Sola, Camille Noûs, and Guislaine Refrégier. Crisprbuilder-tb: "crispr-builder for tuberculosis". exhaustive reconstruction of the crispr locus in mycobacterium tuberculosis complex using sra. *PLOS Computational Biology*, 17(3):1–21, 03 2021.

[14] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

[15] Judith Kamerbeek, LEO Schouls, Arend Kolk, Miranda Van Agterveld, Dick Van Soolingen, Sjoukje Kuijper, Annelies Bunschoten, Henri Molhuizen, Rory Shaw, Madhu Goyal, et al. Simultaneous detection and strain differentiation of mycobacterium tuberculosis for diagnosis and epidemiology. *Journal of clinical microbiology*, 35(4):907–914, 1997.

[16] Midori Kato-Maeda, S Gagneux, LL Flores, EY Kim, PM Small, EP Desmond, and PC Hopewell. Strain classification of mycobacterium tuberculosis: congruence between large sequence polymorphisms and spoligotypes. *The International journal of tuberculosis and lung disease*, 15(1):131–133, 2011.

[17] Kira S Makarova, Yuri I Wolf, and Eugene V Koonin. Classification and nomenclature of crispr-cas systems: where from here? *The CRISPR Journal*, 1(5):325–336, 2018.

[18] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[19] Prasit Palittapongarnpim, Pravech Ajawatanawong, Wasna Viratyosin, Nat Smittipat, Areeya Disratthakit, Surakameth Mahasirimongkol, Hideki Yanai, Norio Yamada, Supalert Nedsuwan, Worarat Imasanguan, et al. Evidence for host-bacterial co-evolution via genome sequence analysis of 480 thai mycobacterium tuberculosis lineage 1 isolates. *Scientific reports*, 8(1):1–14, 2018.

[20] Guislaine Refrégier, Christophe Sola, and Christophe Guyeux. Unexpected diversity of crispr unveils some evolutionary patterns of repeated sequences in mycobacterium tuberculosis. *BMC genomics*, 21(1):1–12, 2020.

[21] Egor Shitikov, Sergey Kolchenko, Igor Mokrousov, Julia Bespyatykh, Dmitry Ischenko, Elena Ilina, and Vadim Govorun. Evolutionary pathway analysis and unified classification of east asian lineage of mycobacterium tuberculosis. *Scientific reports*, 7(1):1–10, 2017.

[22] David Stucki, Daniela Brites, Leïla Jeljeli, Mireia Coscolla, Qingyun Liu, Andrej Trauner, Lukas Fenner, Liliana Rutaihwa, Sonia Borrell, Tao Luo, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nature genetics*, 48(12):1535–1543, 2016.

[23] 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. *Proceedings of the National Academy of Sciences*, 101(14):4865–4870, 2004.

[24] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[25] JDA Van Embden, T Van Gorkom, K Kremer, R Jansen, BAM Van der Zeijst, and LM Schouls. Genetic variation and evolutionary origin of the direct repeat locus of mycobacterium tuberculosis complex bacteria. *Journal of bacteriology*, 182(9):2393–2401, 2000.

[26] Wenjing Wei, Shuai Zhang, Joy Fleming, Ying Chen, Zihui Li, Shanghua Fan, Yi Liu, Wei Wang, Ting Wang, Ying Liu, et al. Mycobacterium tuberculosis type iii-a crispr/cas system crrna and its maturation have atypical features. *The FASEB Journal*, 33(1):1496–1509, 2019.