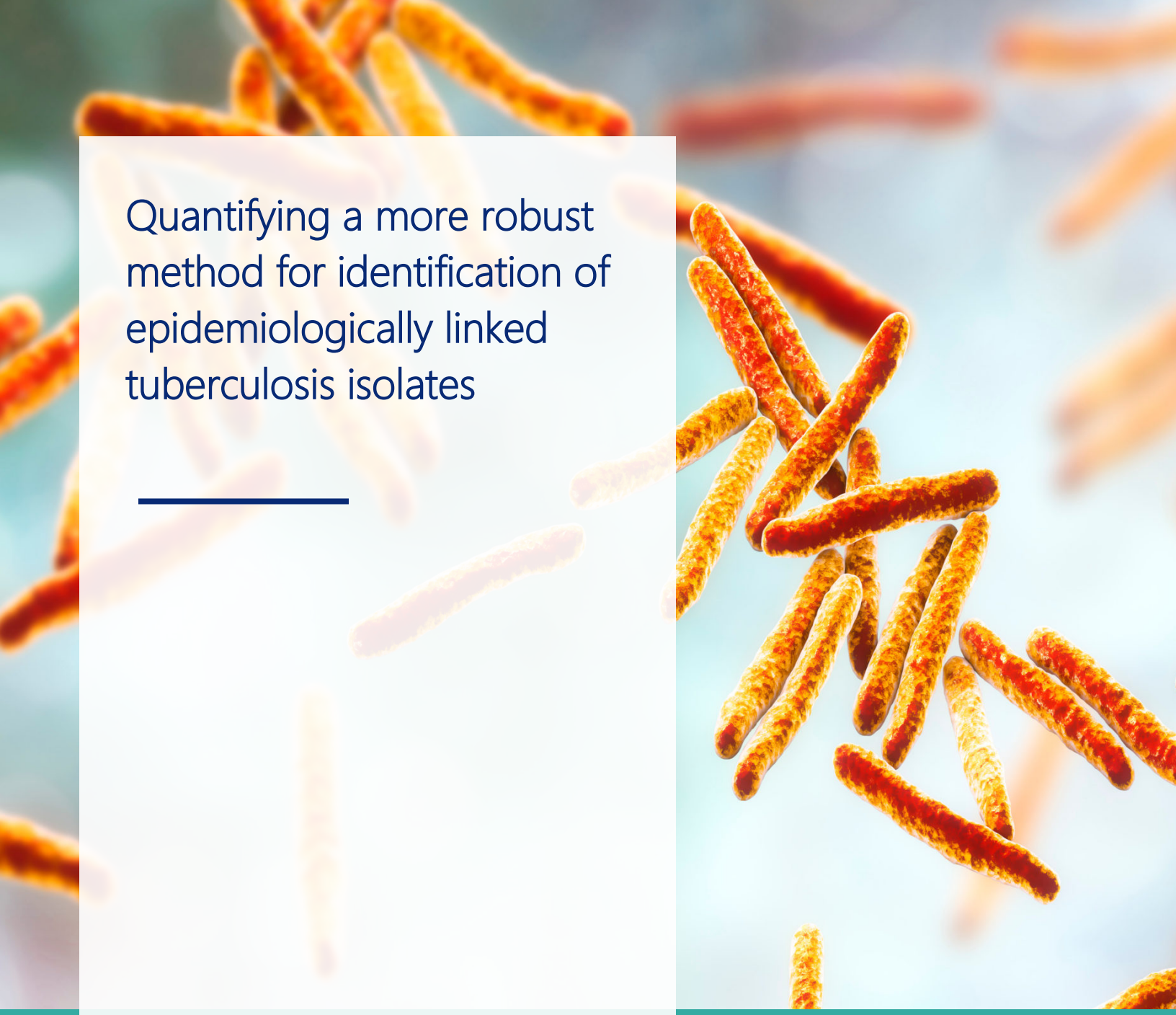


Quantifying a more robust method for identification of epidemiologically linked tuberculosis isolates

AUGUST 27

National Institute for Public Health and the Environment
Authored by: Thierry Haddad
Supervised by: Han de Neeling, Richard Anthony

Summary

Background

Tuberculosis (TB) is a deadly infectious disease caused by members of the *Mycobacterium tuberculosis* complex (MTBC), which can be found in both an infectious active and non-infectious latent stage. TB spread is controlled through contact-tracing, where testing and fingerprint typing of the bacteria can be used for epidemiological linking of patients through clustering. This is strengthened by the slow growth rate of MTBC, resulting in genetically highly similar isolates. Recent advances in high-throughput whole-genome sequencing (WGS) allowed for more accurate TB clustering through calculating SNP distance for cluster identification. However, this method is still limited by technical analytical variations, resulting in isolates sometimes not falling in cluster thresholds; technical replicates end up in different clusters due to slightly different distances calculated. Due to the high clonal growth of MTBC, a less quantitative method is worth investigating, whereby unique cluster-specific SNP's (csSNP's) might be helpful for cluster identification.

Materials & methods

3086 Dutch MTBC isolates were retrieved from the RIVM database. These were spread over 1854 clusters, which were used as reference. Substitutions and deletions were not considered. csSNP's Were calculated for every cluster, so that each csSNP is only found in that particular cluster and is present in every isolate in that cluster. Clusters without csSNP's were merged with single isolates to find additional csSNP's. Afterwards, newly added isolates were assigned a WGS cluster according to csSNP overlap and performance was benchmarked.

Results

1840 Of 1854 WGS clusters were able to produce csSNP's, resulting in a success percentage of 99.24%. Of the 14 clusters without csSNP's, 7 were still able to be assigned csSNP's after merging with single isolates. Minimum spanning network graphs were generated to show that these single isolates were genetically close to the cluster and even identified a WGS merge of identical isolates that were split over clusters due to technical variance. Newly added isolates were generally correctly assigned their WGS clusters if the respective cluster consisted of 2 or more isolates, minus 8 isolates. However, filtering of single isolates csSNP's was necessary due to frequent overlap. Lastly, several intra-*pncA* csSNP's were identified, with reasonably uniform intra-cluster pyrazinamide reactions, but future work is needed.

Conclusion

Overall, the csSNP method shows robustness for existing clusters of 2 or more isolates, even allowing wrongly assigned technical replicates to be properly merged. However, single isolate csSNP's perform less well and can cause cluster overlap in newly added isolates. Further research is needed to address clustering of single isolates and solving the few leftover clusters without csSNP's, but the algorithm could be used in future work for rapid screening of clusters of interest.

Introduction

To this date, tuberculosis (TB) remains one of the major causes of death worldwide. The total number of deaths during 2018 resulting from TB is estimated to be around 1.500.000, while the total number of people that fell ill from TB is estimated to be around 10.000.000 (WHO, 2019). In the Netherlands specifically, the 2019 report has shown a total of 759 TB patients (RIVM, 2020). An infectious disease that can be found in humans and several other animals, TB is caused by infection with the bacterial species of the *Mycobacterium tuberculosis* complex, a rod-shaped bacteria with an inconsistent Gram-staining due to its waxy coat (mycolic acid), and can be spread through coughing. This complex harbors the eponymous *Mycobacterium tuberculosis* (MTB), *Mycobacterium bovis* (including the Bacillus Calmette–Guérin (BCG) strain that is used for vaccinations) and *Mycobacterium africanum* amongst others. Infection with TB can result in either latent or active TB; people infected with latent TB have no symptoms and do not have TB disease, people with active TB have symptoms, have TB disease and are infectious to other people. TB infections can typically be found in the lungs (pulmonary TB), although other locations could also be infected (extrapulmonary TB).

Diagnosis of TB is dependent on whether the infection is active or latent (National Coordination of Infectious Disease Control, National Institute for Public Health and the Environment). Diagnosis of active TB is important for rapid identification of infectious patients. Posterior-anterior chest X-rays are taken for the identification of active pulmonary TB, which can appear as TB consolidations. Samples are auramine stained and investigated through a microscope; mycobacteria will be stained orange, but can only give a reliable positive result when the concentration of TB bacteria in the sample is high. A polymerase chain reaction (PCR) is performed after positive result, to differentiate atypical (non-TB causing) from typical (TB causing) mycobacteria, and the sample is cultured in mycobacteria growth indicator tubes (MGIT's) to identify potential resistances. TB is treated through the application of multiple antibiotics like pyrazinamide, rifampicin and isoniazid, but resistant TB is an increasingly frequent problem in the form of multi or extensively drug-resistant TB. Fast and robust identification of resistances is needed to limit spread throughout the population. Latent TB is diagnosed through a tuberculin skin test (also called Mantoux test), where the skin is injected with tuberculin and the size of the swelling is measured, corresponding to the body's reaction to the tuberculin, or through a blood test (interferon-gamma release

assay), where T-lymphocytes response is measured against *Mycobacterium tuberculosis* antigens; if the antigens are recognized by the T-lymphocytes, interferon-gamma will be secreted and can be measured.

A remarkable feature of the *Mycobacterium tuberculosis* species is the slow growth rate, equating to roughly 16 to 24 hours depending on setting and state (Beste et al., 2009). MTB is assumed to have an even slower or potentially halted growth rate during the latent phase. The combination of slow growth rate and the potential to stay latent for decades results in highly clonal bacteria in the MTBC, sharing 99.99% sequence identity (Dos Vultos et al., 2008). This MTBC feature has proven useful during the typing of isolates, as several methods have been used to detect "fingerprints" of the strains. Such methods include restriction fragment length polymorphism (RFLP) typing (Hermans PW, 1990), spoligotyping (Kremer, Bunschoten, Schouls, van Soolingen, & van Embden) and variable number of tandem repeats (VNTR) typing (Supply, 2005), of which the first two are laboratory-intensive and the latter has been standard until typing through whole-genome sequencing (WGS) was shown to be more efficient in the Netherlands.

The transition from VNTR to WGS typing resulted in less clusters being generated, but a higher percentage of clusters that could be verified on epidemiological linkage through contact tracing by the municipal health services. WGS typing works by mapping reads from a high-throughput next-generation sequencing method to the MTB reference genome H37Rv, after which single nucleotide polymorphisms (SNP's) can be identified through software like Breseq, depending on an allele frequency of 80% and sequencing coverage of at least 5 reads (Jajou et al., 2018). Through the identification of SNP's, the pair-wise SNP distance between TB isolates can be quantified, after which isolates can be added to existing clusters if SNP distance falls within a certain threshold, or can be put in new single-isolate "clusters" if too distant. An SNP distance of less than 6 was considered a direct epidemiological link, while an SNP distance of 6 to 12 was considered indeterminate (Walker et al., 2013).

However, as this method is based on the identification and quantification of SNP's, it will inherently be limited in performance through sequencing shortcomings. SNP distance is dependent on the characteristics of the analytical pipeline and NGS sequencing method that were used, where deviations in analytical methods can lead to technical variance. Variations in SNP distance can result in isolates being placed on the wrong side of a cluster threshold, or clusters growing in case a newly typed isolates falls

between isolates just outside the threshold. Additionally, as SNP distance is a method based on quantification, it does not allow for a rapid and easy identification of clusters in a collection of isolates. To meet this end, the distances between all known isolates needs to be re-calculated.

Therefore, it may be beneficial to explore the potential of a TB typing method which is largely unaffected by the impact of technical variance, while making use of the clonal characteristic of MTB. Previous research has shown the ability to discern TB clade lineages and sub-lineages through the identification of a minimal set of SNP's that are unique to their (sub)lineage (Coll et al., 2014). The persistency of these SNP's throughout the clades brings forth the question of whether this method is applicable to smaller-scale settings, in this case epidemiological linked clusters in the Netherlands. Therefore, this project aims to identify and quantify the potential of cluster-specific and cluster-defining SNP's, for a more robust identification of epidemiologically linked TB isolates, largely independent of the limiting factors SNP distance knows. Hence, this research attempts to answer the following hypothesis: Due to the slow clonal replication of M. tuberculosis, and recent availability of high-throughput and high-coverage next-generation sequencing, single nucleotide polymorphisms could be used for the identification of epidemiologically linked M. tuberculosis isolates.

Materials & methods

Mycobacterium complex isolates

Whole genome sequencing data were accessed through a MySQL database, with two tables of interest. One dataset contained the collection of all *Mycobacterium tuberculosis* complex isolates sequenced in The Netherlands, where every entry represented a distinct mutation for a given isolate. Each such mutation possessed a supplement of annotation, including strain id, run id, gene locus, reference allele, mutation type, etc. The second dataset contained clusters of epidemiologically linked Mycobacterium complex isolates, where a cluster represented a genetic SNP distance of <12 and epidemiological linkage a genetic SNP distance of <6 (Jajou et al., 2018). An inner join was performed to retrieve all the mutations per WGS cluster. The scope of this research was limited to single nucleotide polymorphisms; insertions and deletions were not considered. This dataset was used as a reference set further down the pipeline.

Identification of cluster-specific SNPs

The detection, quantification and validation of cluster-specific SNPs (csSNP's) was done through an R script (version 3.6.0). Mutations were identified and quantified based on a combination of location on the genome and nucleotide, and would be considered cluster-specific only if it adhered to two limitations:

I) it would only be located within intra-cluster Mycobacterium isolates and be absent in other clusters (cluster-specific)

II) it was present in every isolate which is a member of said cluster (cluster-defining)

Thus, an algorithm was created for the identification of cluster-specific SNPs. All WGS clusters were collected in a set, which in turn was iterated over on individual cluster basis. Per cluster, the SNP data for every member isolate was subtracted from the reference dataset and (sub)lineage (Coll et al., 2014) of every isolate was gathered. An anti-join (or left-exclusive join) was performed on the cluster against the reference minus the respective cluster. If one or more cluster-specific SNPs were found, limitation 1 was met. Subsequently, these cluster-specific SNPs were quantified, in such a way that those that were not omni-present (intra-cluster frequency equals cluster size) were trimmed off. If any SNP's remained, limitations 2 was also met and the SNP was

considered cluster-specific and cluster-defining. Figure 2 illustrates the workflow for initial csSNP calculation.

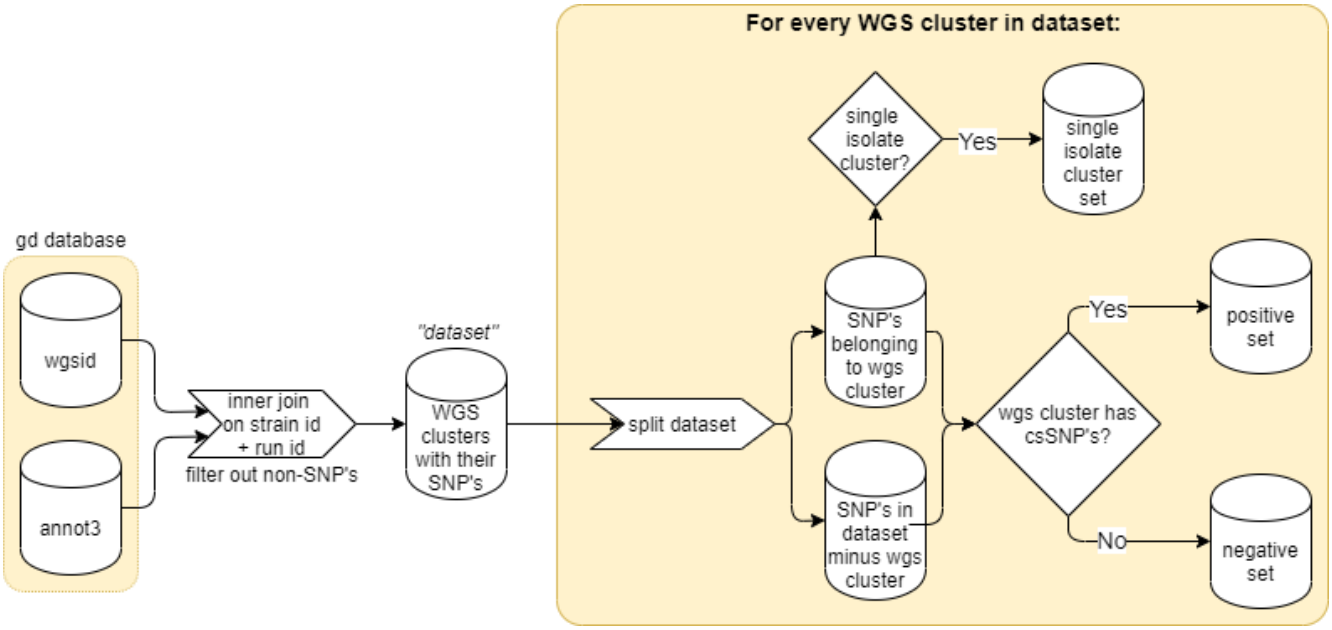


Figure 1: Workflow for the initial calculation of cluster-specific SNP's.

Clusters without cluster-specific SNP's

For cases where no cluster-specific and cluster-defining SNPs were found, a greedy iterative approach was performed where single isolates (clusters with only a single isolate member) were attached to the cluster and an additional search was performed, whereby the concatenation of cluster and isolate acted as a new, distinct cluster. The set of clusters without csSNP's was ordered in descending order, as clusters with more isolates were considered more informative for potential epidemiological linkage. Simultaneously, every iteration the isolate and cluster were subtracted from the original reference for csSNP identification. The single isolate cluster set was ordered according to csSNP presence, so that those without csSNP's had priority during iteration. If this combination of cluster and isolate did not result in limitation 1 and 2 adhering csSNP's, the isolate was removed and the previous steps were repeated for subsequent single isolates. Likewise, when one or more csSNP's were found, the concatenated isolate was removed from the pool of single isolates for subsequent cluster iterations. The workflow can be seen in Figure 2.

If the merging of a cluster and a single-isolate cluster resulted in csSNP's, the average genetic pair-wise distance was calculated using a custom R script and the

"distance3" MySQL database. For every isolate member of the cluster, the genetic distance with the single isolate was calculated and divided by the cluster size. If an isolate was missing from the database, the cluster size was subtracted by 1.

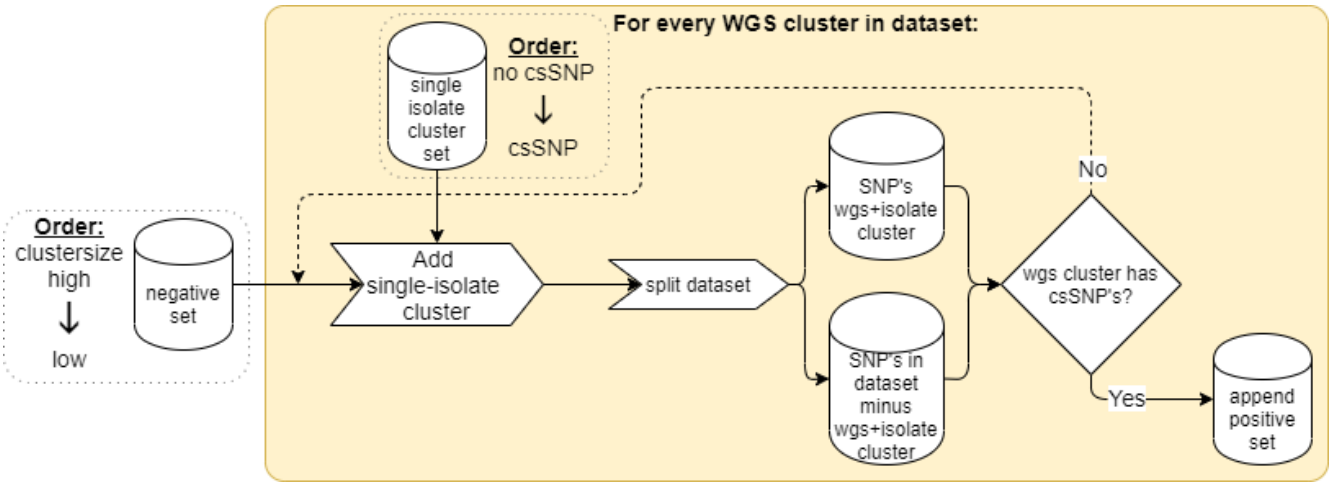


Figure 2: Workflow illustrating re-calculation of csSNP's after single-isolate merging.

Assigning clusters to newly added isolates

Isolates added to the database after the initial calculation of csSNP's were retrieved by filtering on isolates with a "run id" > 200702, which corresponds to the 2nd of July 2020. The assigned WGS clusters id's for these new isolates where collected from the wgsid database and used as reference. If no id was assigned, "None" was noted instead.

The SNP's of every isolate were compared to the library of previously calculated csSNP's, without filtering. The WGS id's of matching csSNP's and the amount of csSNP's overlap were assigned to the isolate and the assigned WGS id and WGS id resulting from csSNP matching were compared. Afterwards, the csSNP matching was repeated, but with a filter so that only csSNP's of clusters with two or more isolates were matched.

Lastly a small analysis was performed on intra-*pncA* csSNP's. For this the library of csSNP's was filtered on csSNP's inside the 2288681-2289240 nucleotide range of the *pncA* gene. This filtered subset was matched against the "annot3" database to retrieve measured pyrazinamide reactions.

Results & discussion

Mycobacterium complex isolates

The isolate and mutation data represented a "snapshot" of the database as of the 2nd of July 2020. The annotated dataset contained a total of 6645903 registered mutations (position in the genome and allele combination), across a total of 3086 distinct isolates. The WGS cluster dataset contained 1854 distinct clusters, harboring a total of 2560 distinct isolates and 2752 isolates when including duplicates (additional in-house sequenced isolates). Of these clusters, 1514 consisted of single-isolate clusters. The largest cluster was A003 with 53 isolates (Figure 3), which is the cluster of the BCG strain used for vaccination. An inner join was performed on the datasets resulting in an annotated dataset of clusters with 3157485 mutation entries for 2752 isolates placed in 1854 clusters. A comparison of the datasets can be seen in Table 1.

	Pre inner-join	Post inner-join
Mutations	6645903	3157485
Of which SNP's	4208200	3157485
Unique isolates	3086	2752

Table 1: Comparison of the mutation dataset pre- and post-merger based on WGS id clusters.

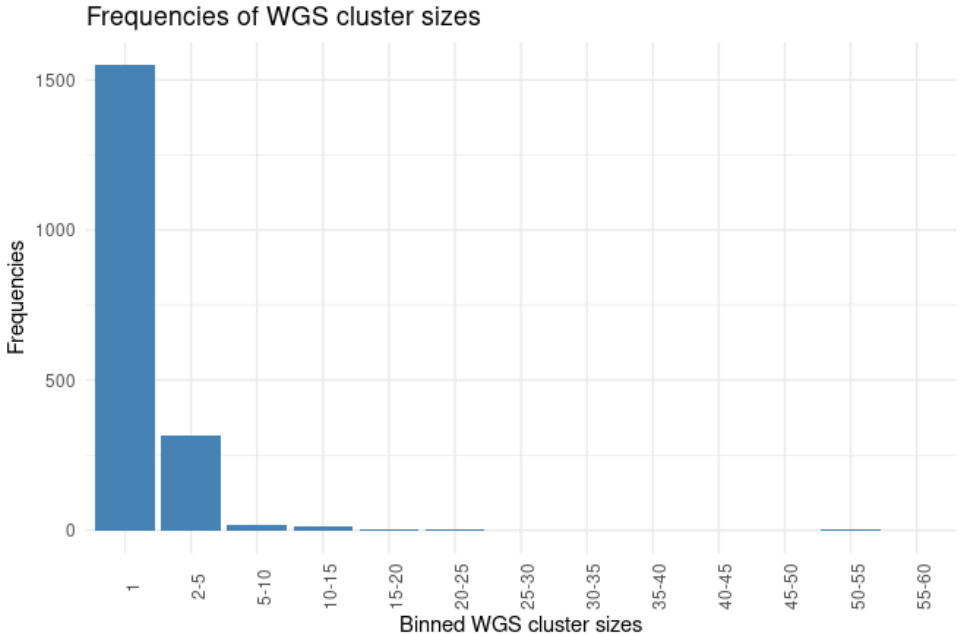


Figure 3: Frequencies of WGS cluster sizes. Noteworthy is that most clusters only have a single isolate, but there's also an outlier with 53 isolates.

Identification of cluster-specific SNP's

The algorithm described in methods was applied on the dataset that was formed through the inner join. This resulted in a dataset of 124907 cluster-specific and cluster-identifying SNP's, across a total of 1840 clusters. 14 Clusters did not provide csSNP's. Overall, 99.24% of all WGS clusters successfully produced csSNP's.

A comparison was made of the binned number of csSNP's found on a per-cluster basis, against the frequency of such number of csSNP's occurring, seen in Figure 4A. The most occurring number of csSNP's are roughly located around 50.

Additionally, a comparison of WGS cluster size against the number of csSNP's found on an individual cluster basis can be seen in Figure 4B. As the number of isolates in a WGS cluster increases, the overall number of csSNP's tends to go down. The highest numbers of csSNP's per cluster can be found in the smallest WGS cluster sizes. This inverse relation between cluster size and number of csSNP's can be expected, as the probability of intra-cluster differences on a per-base level increases for every isolate it has.

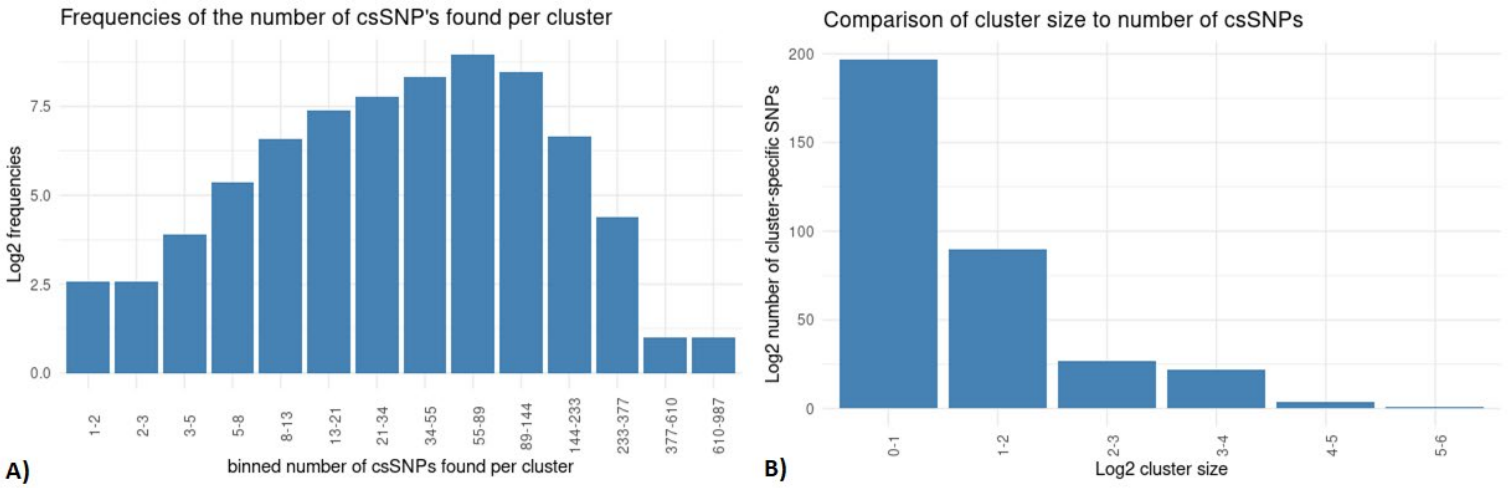


Figure 4: Distributions of csSNP group sizes. A) Frequencies of the number of csSNP's found per cluster. B) Number of csSNP's found per cluster per binned cluster size.

Clusters without cluster-specific SNP's

A set of 14 WGS clusters did not produce useable csSNP's (i.e. not adhering to both limitations), and an iterative, greedy approach of merging a single-isolate cluster to these 14 clusters was performed. This resulted in 7 out of the 14 clusters still producing csSNP's, the total number of csSNP's in the library containing all WGS clusters and their csSNP's going up from 124907 to 124988 and the total of single-isolate clusters dropping from 1514 to 1507. The successful merges can be seen in Table 2.

Additionally, the average pair-wise genetic distance between base cluster and single-isolate cluster was calculated to understand how these single-isolates relate to the cluster. As not all isolates were present in the distance3 database, cluster-pairs B771-B770 and A507-A428, and one isolate from A528-A656 could not be calculated. Overall, the single-isolates appeared to be close to the base cluster, generally just outside the genetic distance of 12 that was used as cluster threshold (Jajou et al., 2018). An outlier in Table 2 was the cluster-pair A037-B601, with a respective average distance of 64.2 between the cluster isolates.

Table 2: Results of identification of csSNP's for clusters that failed the initial check. Each row shows the base cluster, its size, the single-isolate cluster used for merging and the average pair-wise genetic distance between the base cluster isolates and the single isolate. NA's were inserted when the distance database did not have the isolates. *: one isolate was not found in the distance database.

Cluster	Size	Merged with	Avg SNP distance
A037	5	B601	64.2
B491	4	B772	14.5
A022	2	A444	19
A528	2	A656	15*
A799	2	B667	18.5
B771	2	B770	NA
A507	1	A428	NA

To further map the surroundings of these cluster pairs, a web-based interface (kolber03.php, made by J.C. Kolber) was used, which calculates the minimum spanning network graph of an isolate and its neighbours depending on a set threshold. A graph was created for every cluster-pair that was able to generate an average genetic distance; these can be seen in Appendix A. Per graph, the nodes contain the number corresponding to an isolate and the edges the minimum genetic distance between two isolates. Additionally, a list is shown with isolate-pair minimum distances. They were

color-coded according to their characteristic: blue represents the base cluster, green the single-isolate cluster used for merging with the base cluster, black for nearby isolates (within the threshold needed for the single-isolate to appear) which were assigned to a different cluster, and gray for isolates which were not present in the WGS database (and thus outside the scope of this project). Cluster pairs B491-B772, A022-A444 and A799-B667 showed no other nearby isolates, meaning that the single-isolate that was used for merging, was the closest isolate to the cluster and able to yield one or more csSNP's while laying outside the 12 SNP distance threshold. Of particular interest is cluster pair B491-B772, where four identical isolates have slightly different minimum SNP distances. Here, three fall in the base cluster and one was added as a single-isolate. But through the calculation of csSNP's, these identical isolates were merged. This was evident of the before mentioned drawback of using genetic distance as a clustering method; variance in sequencing may result in genetic distances falling just outside cluster thresholds. Lastly, cluster pairs A037-B601 and A528-A656 had shown the presence of nearby isolates with a minimum genetic distance lower than the isolate used for merging. As they were closer to the base cluster, there's a reasonable probability that these isolates could've also been merged. However, for this project the single isolate list was only iterated once. Future work could quantify single-isolate cluster merging for more iterations and more than one single-isolate cluster.

Cluster identification of newly added isolates

Five weeks after the initial calculation of csSNP's, the database was scanned for newly added isolates. 74 New isolates and their assigned WGS cluster id's were collected. Per isolate, the collection of SNP's were matched against the reference set of known csSNP's, allowing for multiple hits if that would occur. The results can be seen in Appendix B. Per row, the identifier, isolate id and run id and assigned WGS cluster id and its size were noted. After that, a vector of all WGS clusters with shared csSNP's, including the respective cluster size and the number of csSNP's overlap. Lastly, if the assigned WGS cluster id was equal to the WGS cluster id calculated from csSNP matching, a Boolean would be set to TRUE, if not FALSE. Cluster identification appeared to do well on the 28 isolates that were assigned cluster sizes >=2, with 20 isolates being correctly matched with csSNP's.

Additionally, at the bottom Appendix B several control isolates can be seen. As these were not used for routine tracing, they have not been assigned an official WGS cluster ID. Yet, despite being technical replicates (due to their different run ID's, which means different sequencing dates), they've all been put in the same WGS cluster when using the csSNP algorithm, giving another indication of the robustness of the csSNP method when facing technical variance.

Performance on isolates with cluster size 1 was very poor with not a single match, but this was expected as this meant they were assigned to their own individual cluster, for which previously no csSNP's were calculated. Therefore, the algorithm could only assign existing WGS clusters to new isolates if the existing clusters contained 2 or more isolates, which is a limiting factor of the algorithm. Isolates that were not assigned to an existing cluster were presumably not closely related to any other isolate in the database and could potentially be defined as a new unique cluster of one isolate, but as there were still 7 WGS clusters without csSNP's it would not be fully certain.

However, several isolates with cluster size >=2 had more than a single csSNP id match, with sometimes a csSNP from a single-isolate cluster matching with just a single csSNP. As was shown in Figure 4B, single-isolate clusters tend to have many csSNP's due to being on their own. A filtering step was introduced, such that only csSNP's of clusters with two or more isolates would be assigned. The result can be seen in Appendix C and the core differences can be seen in Table 3.

Table 3: Effects of no filtering and filtering out csSNP matches from single-isolates. Filtering resulted in reduced frequency of having more than one csSNP match.

	No filtering	**Filter out single-isolates**
Isolates	74	74
Correct csSNP's	20	20
Correct csSNP's, excluding >1 csSNP matches	17	20
Correct csSNP's for cluster sizes >=2	20	20
Correct csSNP's for cluster size 1	0	0
>1 csSNP cluster matches	34	3

While the number of correctly matched csSNP's were not changed on their own, the results became more robust. As some isolates with a matching csSNP also contained a csSNP match of a single-isolate cluster, this would be an ambiguous result. Filtering managed to remove this ambiguity from clusters, resulting in accurate csSNP matching performance for especially larger clusters, as these tend to have a lot of csSNP overlap.

Lastly a short analysis was performed on the presence of csSNP's in the *pncA* gene. The csSNP library was filtered for csSNP's within the range of 2288681-2289240. The results are displayed in Appendix D. As mutations in the *pncA* gene are correlated with pyrazinamide resistance, identification of csSNP's inside this gene could potentially be of interest for rapid resistance identification (T.J. BROWN, 2000). While not every reaction to pyrazinamide was tested, there seems to be reasonable intra-cluster uniformity for pyrazinamide response. However, not every csSNP leads to pyrazinamide resistance. Future work could lead to more robust identification of pyrazinamide resistance defining csSNP's, but limitations due to evolutionary pressure should be taken into consideration, as identical resistance-related mutations might appear in non-epidemiologically linked TB isolates due to homoplasy, which could lead to faulty transmission assumptions.

Overall, the hypothesis that cluster-defining SNP's could be used for the identification of epidemiological linked isolates appeared favorable. More than 99% of the initial WGS clusters were able to produce csSNP's for cluster identification, and half of the remaining clusters were able to produce csSNP's after additions of closely related single isolates – future improvements of this algorithm could potentially solve the rest as well. The algorithm has shown to be robust against technical variance, with identical isolates being clustered through shared csSNP's, where traditional SNP distance clustering would assign them differently. However, the algorithm was limited to assigning WGS clusters to new isolates only if the assigned cluster consisted of 2 or more isolates. New isolates without any csSNP overlap were considered new single-isolate clusters for the csSNP algorithm, but several clusters without csSNP's remained.

Conclusion & further research

Cluster identification is typically performed through a multiple sequence alignment (MSA) of the isolate sequences against reference sequence (of the H37Rv reference strain), after which pair-wise SNP's are counted as a measure of genetic distance. Subsequently, the isolates are clustered according to this genetic distance. However, the sequencing technique and the SNP calling algorithm can affect the number of SNP's found, so that the genetic distance can vary even in-between runs of identical isolates, as was seen in cluster pair B491-B772 in Appendix A. Through the identification and quantification of cluster-specific SNPs, in such way that these could be used as robust cluster markers, a more robust identification of epidemiologically linked isolates could be achieved.

Therefore, csSNP's were calculated for epidemiologically linked isolates, whereby WGS genetic distance clustering was used as reference. Overall, the robustness and reliability of csSNP's as a metric for the identification of epidemiological linkage appeared favorable. More than 99% of all clusters were able to produce csSNP's, with only 14 clusters (which also contained single-isolate clusters) failing the initial check. Merging of a single nearby single-isolate was sufficient for 7 of the 14 clusters to yield csSNP's, and caused the re-location of 4 identical isolates from 2 clusters to a single one – demonstrating the value of this approach for identifying missed clustered isolates due to technical analytical limitations resulting from sequence alignment and SNP calling / filtering. Afterwards, newly added isolates were matched, whereby almost all of the clusters with two or more isolates, given that one or more csSNP's were found for these clusters beforehand, were correctly matched to their respective WGS clusters through csSNP identification.

As csSNP performance on clusters with multiple isolates seems promising, future work could instead focus on more robust merging of single-isolates to expand current clusters and produce csSNP's for the remaining few clusters lacking such. Additionally, the csSNP algorithm could be used in future work to screen collections of isolates for clusters of particular interest, independent of laboratory and without the need to share sequencing data and perform analysis and clustering, or rapidly screen for important clusters like high-risk TB strains.

References

Beste, D. J., Espasa, M., Bonde, B., Kierzek, A. M., Stewart, G. R., & McFadden, J. (2009). The genetic requirements for fast and slow growth in mycobacteria. *PLoS One, 4*(4), e5349. doi:10.1371/journal.pone.0005349

Coll, F., McNerney, R., Guerra-Assuncao, J. A., Glynn, J. R., Perdigao, J., Viveiros, M., . . . Clark, T. G. (2014). A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat Commun, 5*, 4812. doi:10.1038/ncomms5812

Dos Vultos, T., Mestre, O., Rauzier, J., Golec, M., Rastogi, N., Rasolofo, V., . . . Gicquel, B. (2008). Evolution and diversity of clonal bacteria: the paradigm of Mycobacterium tuberculosis. *PLoS One, 3*(2), e1538. doi:10.1371/journal.pone.0001538

Hermans PW, v. S. D., Dale JW, et al. (1990). Insertion element IS986 from Mycobacterium tuberculosis: a useful tool for diagnosis and epidemiology of tuberculosis. *J Clin Microbiol. , 1990;28(9):2051-2058.* Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC268102/

Jajou, R., de Neeling, A., van Hunen, R., de Vries, G., Schimmel, H., Mulder, A., . . . van Soolingen, D. (2018). Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study. *PLoS One, 13*(5), 1-11. doi:10.1371/journal.pone.0197556

Kremer, K., Bunschoten, A., Schouls, L., van Soolingen, D., & van Embden, J. Spoligotyping protocol. 1-19. Retrieved from https://www.rivm.nl/sites/default/files/2018-11/protocol spoligotyping.pdf

Supply, P. (2005). Multilocus Variable Number Tandem Repeat Genotyping of Mycobacterium tuberculosis. *Institut Pasteur de Lille*(May), 73-73. Retrieved from http://www.miru-vntrplus.org/MIRU/miruinfo.faces

T.J. BROWN, Ö. T., G.L. FRENCH. (2000). Simultaneous identification and typing of multi-drug-resistant Mycobacterium tuberculosis isolates by analysis of pncA and rpoB.

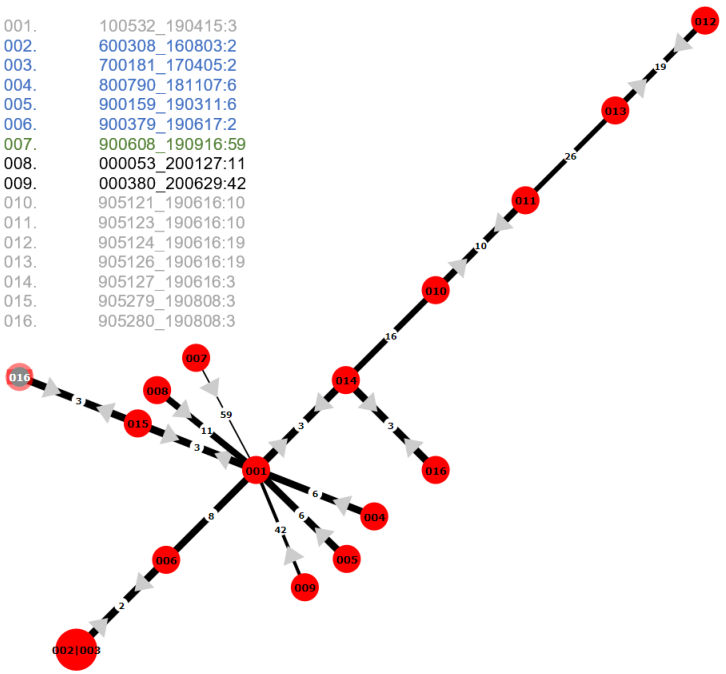
Walker, T. M., Ip, C. L. C., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., . . . Peto, T. E. A. (2013). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious Diseases, 13*(2), 137-146. doi:10.1016/s1473-3099(12)70277-3

WHO. (2019). *Global tuberculosis report 2019*. https://www.who.int/tb/publications/global_report/en/: World Health Organization.

Appendix A

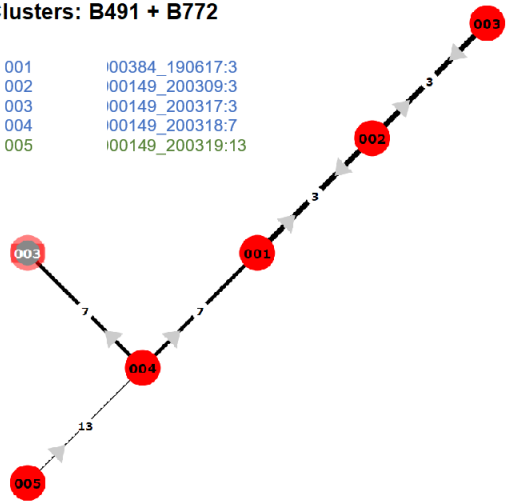
Clusters: A037 + B601

001. 100532_190415:3
002. 600308_160803:2
003. 700181_170405:2
004. 800790_181107:6
005. 900159_190311:6
006. 900379_190617:2
007. 900608_190916:59
008. 000053_200127:11
009. 000380_200629:42
010. 905121_190616:10
011. 905123_190616:10
012. 905124_190616:19
013. 905126_190616:19
014. 905127_190616:3
015. 905279_190808:3
016. 905280_190808:3



Clusters: B491 + B772

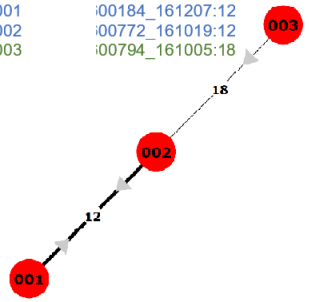
001 00384_190617:3
002 00149_200309:3
003 00149_200317:3
004 00149_200318:7
005 00149_200319:13



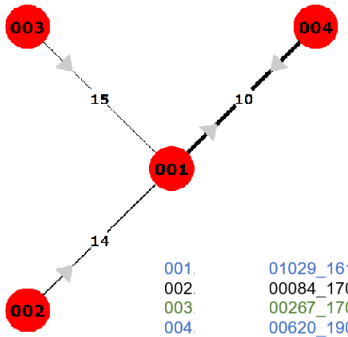
Color legend:
- Base cluster
- Concatenated cluster (single-isolate cluster)
- Nearby isolates in different WGS cluster
- Nearby isolates not found in WGS clusters

Clusters: A022 + A444

001 00184_161207:12
002 00772_161019:12
003 00794_161005:18



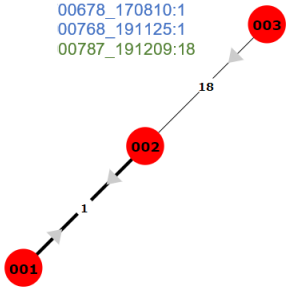
Clusters: A528 + A656



001. 01029_161214:10
002. 00084_170301:14
003. 00267_170517:15
004. 00620_190930:10

Clusters: A799 + B667

001 00678_170810:1
002 00768_191125:1
003 00787_191209:18



Appendix B

ID	isolate	run	assigned cluster	assigned cluster size	assigned WGS has csSNP	found csSNP's + its size + csSNP overlap	csSNP matches wgsid
00113_200724	00113	200724	B746	15	TRUE	B746, 15, 11	TRUE
00113_200725	00113	200725	B746	15	TRUE	B746, 15, 11	TRUE
00123_200724	00123	200724	B764	12	TRUE	B764, 12, 185	TRUE
00123_200725	00123	200725	B764	12	TRUE	B764, 12, 185	TRUE
00130_200724	00130	200724	B757	9	TRUE	B757, 9, 131	TRUE
00130_200725	00130	200725	B757	9	TRUE	B757, 9, 131	TRUE
00377_200713	00377	200713	A169	8	TRUE	A169, 8, 21, B708, 1, 2, B841, 1, 1	TRUE
00100_200724	00100	200724	B752	7	FALSE	None	FALSE
00100_200725	00100	200725	B752	7	FALSE	None	FALSE
00392_200706	00392	200706	B477	5	TRUE	B477, 5, 26	TRUE
00392_200720	00392	200720	B477	5	TRUE	B477, 5, 26	TRUE
00407_200713	00407	200713	A938	4	TRUE	A938, 4, 114	TRUE
00414_200713	00414	200713	A371	4	TRUE	A371, 4, 48, B335, 1, 1	TRUE
00426_200727	00426	200727	A081	4	TRUE	A081, 4, 72	TRUE
00387_200706	00387	200706	B358	3	TRUE	A559, 1, 1, B358, 3, 34	TRUE
00394_200706	00394	200706	B537	3	TRUE	B537, 3, 89	TRUE
00431_200727	00431	200727	B848	3	TRUE	B848, 3, 111	TRUE
00438_200727	00438	200727	A018	3	TRUE	A018, 3, 82	TRUE
00360_200706	00360	200706	B872	2	TRUE	B872, 2, 85	TRUE
00395_200706	00395	200706	A540	2	TRUE	A540, 2, 2	TRUE
00396_200706	00396	200706	B868	2	TRUE	B868, 2, 11	TRUE
00450_200727	00450	200727	B292	2	TRUE	B292, 2, 13	TRUE
00376_200720	00376	200720	B875	2	FALSE	A817, 1, 63	FALSE
00384_200720	00384	200720	B878	2	FALSE	A859, 2, 1, A987, 1, 1, A993, 1, 134, B241, 1, 1, B627, 1, 1, B846, 5, 3	FALSE
00416_200713	00416	200713	B895	2	FALSE	A817, 1, 1, B087, 1, 1, B114, 1, 1, B540, 2, 1, B688, 1, 1	FALSE
00416_200727	00416	200727	B895	2	FALSE	A817, 1, 1, B087, 1, 1, B114, 1, 1, B540, 2, 1, B688, 1, 1	FALSE
00424_200720	00424	200720	B902	2	FALSE	A067, 2, 2	FALSE
00424_200727	00424	200727	B902	2	FALSE	A067, 2, 2	FALSE
00358_200706	00358	200706	B881	1	FALSE	A497, 1, 7, A634, 1, 1, B022, 1, 1	FALSE
00375_200706	00375	200706	B882	1	FALSE	A223, 1, 1, A305, 1, 1, A374, 1, 1, A408, 1, 1, A160, 2, 1, B481, 1, 1, B775, 1, 1	FALSE
00391_200706	00391	200706	B883	1	FALSE	B499, 1, 1	FALSE
00393_200706	00393	200706	B884	1	FALSE	A413, 1, 1, A434, 1, 1, A530, 1, 1	FALSE
00398_200706	00398	200706	B885	1	FALSE	A454, 1, 1, A765, 1, 1, A947, 1, 34	FALSE

	00399_200706	000399	200706 B886	1	FALSE	B075, 1, 21, B855, 1, 1	FALSE
	00400_200706	000400	200706 B887	1	FALSE	None	FALSE
	00401_200713	000401	200713 B891	1	FALSE	A266, 1, 2, A498, 1, 1, B333, 1, 1	FALSE
	00402_200706	000402	200706 B888	1	FALSE	A423, 1, 44	FALSE
	00403_200706	000403	200706 B889	1	FALSE	A020, 16, 1, A021, 2, 1, A533, 1, 1, A714, 1, 1, A150, 2, 1, B273, 2, 1	FALSE
	00404_200706	000404	200706 B890	1	FALSE	A049, 2, 1, A733, 1, 1	FALSE
	00406_200727	000406	200727 B906	1	FALSE	None	FALSE
	00409_200713	000409	200713 B892	1	FALSE	A183, 1, 1, A188, 1, 1, A189, 1, 1, B302, 1, 2, A197, 1, 1, A199, 1, 1, A221, 1, 1, A222, 1, 1, A231, 1, 1, A239, 1, 1, A240, 1, 1, A247, 1, 1, A249, 1, 1, A251, 1, 1, A254, 1, 1, A255, 1, 2, A259, 1, 2, A267, 1, 1, A283, 1, 1, A285, 2, 1, A296, 1, 1, A323, 1, 1, A329, 1, 1, A334, 1, 2, A335, 1, 1, A342, 1, 1, A347, 1, 1, A357, 1, 1, A362, 1, 1, A060, 6, 1, A373, 1, 2, A380, 1, 1, A393, 1, 2, A395, 1, 1, A396, 1, 1, A405, 1, 1, A416, 1, 1, A419, 1, 1, A425, 1, 1, A429, 1, 2, A430, 1, 1, A431, 1, 1, A067, 2, 1, A433, 1, 1, A438, 1, 2, A448, 1, 1, A458, 1, 1, A466, 1, 1, A073, 5, 1, A479, 1, 1, A483, 1, 1, A487, 1, 1, A489, 1, 3, A490, 1, 1,	FALSE
	00410_200713	000410	200713 B893	1	FALSE	A231, 1, 15, B045, 1, 1	FALSE
	00412_200713	000412	200713 B894	1	FALSE	B810, 1, 1	FALSE
	00418_200720	000418	200720 B899	1	FALSE	B648, 1, 70, B705, 1, 1	FALSE
	00419_200720	000419	200720 B900	1	FALSE	B177, 1, 17	FALSE
	00420_200720	000420	200720 B901	1	FALSE	B173, 3, 1, B487, 1, 1, B600, 1, 5	FALSE
	00422_200713	000422	200713 B896	1	FALSE	A059, 3, 13, B294, 1, 1, B696, 1, 1	FALSE
	00423_200713	000423	200713 B897	1	FALSE	None	FALSE
	00425_200713	000425	200713 B898	1	FALSE	A301, 1, 1, B193, 1, 1, B739, 1, 1	FALSE
	00429_200727	000429	200727 B907	1	FALSE	B785, 1, 1	FALSE

	00430_200727	00430	200727	B908	1	FALSE	B105, 1, 2	FALSE
	00434_200720	00434	200720	B903	1	FALSE	A342, 1, 1, A391, 1, 1, A872, 1, 1, B540, 2, 1, B608, 1, 1	FALSE
	00436_200727	00436	200727	B909	1	FALSE	B236, 1, 80	FALSE
	00439_200727	00439	200727	B910	1	FALSE	A381, 1, 90	FALSE
	00441_200727	00441	200727	B911	1	FALSE	A296, 1, 1, A341, 1, 3, A462, 1, 1, A556, 1, 1, B090, 1, 1	FALSE
	00442_200727	00442	200727	B912	1	FALSE	A896, 1, 1, B510, 2, 1	FALSE
	00443_200727	00443	200727	B913	1	FALSE	A187, 1, 1, B696, 1, 5, B809, 1, 1	FALSE
	00444_200727	00444	200727	B914	1	FALSE	A364, 1, 1, B279, 1, 1	FALSE
	00447_200720	00447	200720	B904	1	FALSE	A677, 1, 1, B163, 1, 1, B547, 1, 1, B859, 1, 1	FALSE
	00448_200720	00448	200720	B905	1	FALSE	B075, 1, 26, B855, 1, 1	FALSE
	00449_200727	00449	200727	B915	1	FALSE	A559, 1, 1, B289, 1, 72	FALSE
	00454_200713	00454	200713	None	0	FALSE	A926, 1, 1	FALSE
	00440_200720	00440	200720	None	0	FALSE	A193, 1, 2, B522, 1, 16, A263, 1, 1, A778, 1, 9, A868, 2, 12, A908, 1, 1, B022, 1, 1, B218, 1, 2	FALSE
	5071__200720	5071_	200720	None	0	FALSE	A792, 1, 1, A887, 1, 1, B456, 1, 1	FALSE
	5072__200720	5072_	200720	None	0	FALSE	A792, 1, 1, A887, 1, 1, B456, 1, 1	FALSE
	5073__200720	5073_	200720	None	0	FALSE	A311, 1, 1, B097, 1, 1, B509, 1, 1, B627, 1, 1, B856, 1, 1	FALSE
	5074__200720	5074_	200720	None	0	FALSE	A311, 1, 1, B097, 1, 1, B509, 1, 1, B627, 1, 1, B856, 1, 1	FALSE
	5075__200720	5075_	200720	None	0	FALSE	B100, 1, 2, B181, 1, 1	FALSE
WGS_controle_200706		WGS_controle	200706	None	0	FALSE	B860, 3, 79	FALSE
WGS_controle_200713		WGS_controle	200713	None	0	FALSE	B860, 3, 79	FALSE
WGS_controle_200720		WGS_controle	200720	None	0	FALSE	B860, 3, 79	FALSE
WGS_controle_200724		WGS_controle	200724	None	0	FALSE	B860, 3, 79	FALSE
WGS_controle_200725		WGS_controle	200725	None	0	FALSE	B860, 3, 79	FALSE
WGS_controle_200727		WGS_controle	200727	None	0	FALSE	B860, 3, 79	FALSE

Appendix C

ID	isolate	run	assigned clu	assigned c	assigned V	found csSNP's + its size + csSNP overlap	csSNP matche
00113_200724	00113	200724	B746	15	TRUE	B746, 15, 11	TRUE
00113_200725	00113	200725	B746	15	TRUE	B746, 15, 11	TRUE
00123_200724	00123	200724	B764	12	TRUE	B764, 12, 185	TRUE
00123_200725	00123	200725	B764	12	TRUE	B764, 12, 185	TRUE
00130_200724	00130	200724	B757	9	TRUE	B757, 9, 131	TRUE
00130_200725	00130	200725	B757	9	TRUE	B757, 9, 131	TRUE
00377_200713	00377	200713	A169	8	TRUE	A169, 8, 21	TRUE
00100_200724	00100	200724	B752	7	FALSE	None	FALSE
00100_200725	00100	200725	B752	7	FALSE	None	FALSE
00392_200706	00392	200706	B477	5	TRUE	B477, 5, 26	TRUE
00392_200720	00392	200720	B477	5	TRUE	B477, 5, 26	TRUE
00407_200713	00407	200713	A938	4	TRUE	A938, 4, 114	TRUE
00414_200713	00414	200713	A371	4	TRUE	A371, 4, 48	TRUE
00426_200727	00426	200727	A081	4	TRUE	A081, 4, 72	TRUE
00387_200706	00387	200706	B358	3	TRUE	B358, 3, 34	TRUE
00394_200706	00394	200706	B537	3	TRUE	B537, 3, 89	TRUE
00431_200727	00431	200727	B848	3	TRUE	B848, 3, 111	TRUE
00438_200727	00438	200727	A018	3	TRUE	A018, 3, 82	TRUE
00376_200720	00376	200720	B875	2	FALSE	None	FALSE
00384_200720	00384	200720	B878	2	FALSE	A859, 2, 1, B846, 5, 3	FALSE
00416_200713	00416	200713	B895	2	FALSE	B540, 2, 1	FALSE
00416_200727	00416	200727	B895	2	FALSE	B540, 2, 1	FALSE
00424_200720	00424	200720	B902	2	FALSE	A067, 2, 2	FALSE
00424_200727	00424	200727	B902	2	FALSE	A067, 2, 2	FALSE
00360_200706	00360	200706	B872	2	TRUE	B872, 2, 85	TRUE
00395_200706	00395	200706	A540	2	TRUE	A540, 2, 2	TRUE
00396_200706	00396	200706	B868	2	TRUE	B868, 2, 11	TRUE
00450_200727	00450	200727	B292	2	TRUE	B292, 2, 13	TRUE
00358_200706	00358	200706	B881	1	FALSE	None	FALSE
00375_200706	00375	200706	B882	1	FALSE	A160, 2, 1	FALSE
00391_200706	00391	200706	B883	1	FALSE	None	FALSE
00393_200706	00393	200706	B884	1	FALSE	None	FALSE
00398_200706	00398	200706	B885	1	FALSE	None	FALSE
00399_200706	00399	200706	B886	1	FALSE	None	FALSE
00400_200706	00400	200706	B887	1	FALSE	None	FALSE
00401_200713	00401	200713	B891	1	FALSE	None	FALSE
00402_200706	00402	200706	B888	1	FALSE	None	FALSE
00403_200706	00403	200706	B889	1	FALSE	A020, 16, 1, A021, 2, 1, A150, 2, 1, B273, 2, 1	FALSE
00404_200706	00404	200706	B890	1	FALSE	A049, 2, 1	FALSE
00406_200727	00406	200727	B906	1	FALSE	None	FALSE

						A011, 3, 1, A285, 2, 1, A060, 6, 1, A067, 2, 1, A073, 5, 1, A088, 5, 1, A097, 2, 1, A098, 5, 1, A739, 2, 1, A781, 2, 1, A135, 3, 1, A136, 2, 1, A138, 2, 1, A932, 2, 1, A965, 2, 1, A966, 6, 1, A147, 2, 1, A155, 2, 1, A165, 4, 1, A167, 2, 1, B330, 2, 1, B363, 2, 1, B430, 2, 1, B461, 3, 1, B510, 2, 1, B532, 2, 1, B534, 2, 1, B539, 3, 1, B548, 2, 1, B577, 2, 1, B590, 2, 1, B632, 2, 1, B655, 2, 1, B702, 2, 1, B728, 2, 1, B754, 4, 1, B767, 4, 1, B786, 2, 1, B787, 2, 2, B826, 2, 1, B828, 2, 1, B848, 3, 1,	
00409_200713		00409	200713 B892	1	FALSE	B867, 2, 1	FALSE
00410_200713		00410	200713 B893	1	FALSE	None	FALSE
00412_200713		00412	200713 B894	1	FALSE	None	FALSE
00418_200720		00418	200720 B899	1	FALSE	None	FALSE
00419_200720		00419	200720 B900	1	FALSE	None	FALSE
00420_200720		00420	200720 B901	1	FALSE	B173, 3, 1	FALSE
00422_200713		00422	200713 B896	1	FALSE	A059, 3, 13	FALSE
00423_200713		00423	200713 B897	1	FALSE	None	FALSE
00425_200713		00425	200713 B898	1	FALSE	None	FALSE
00429_200727		00429	200727 B907	1	FALSE	None	FALSE
00430_200727		00430	200727 B908	1	FALSE	None	FALSE
00434_200720		00434	200720 B903	1	FALSE	B540, 2, 1	FALSE
00436_200727		00436	200727 B909	1	FALSE	None	FALSE
00439_200727		00439	200727 B910	1	FALSE	None	FALSE
00441_200727		00441	200727 B911	1	FALSE	None	FALSE
00442_200727		00442	200727 B912	1	FALSE	B510, 2, 1	FALSE
00443_200727		00443	200727 B913	1	FALSE	None	FALSE
00444_200727		00444	200727 B914	1	FALSE	None	FALSE
00447_200720		00447	200720 B904	1	FALSE	None	FALSE
00448_200720		00448	200720 B905	1	FALSE	None	FALSE
00449_200727		00449	200727 B915	1	FALSE	None	FALSE
00454_200713		00454	200713 None	0	FALSE	None	FALSE
00440_200720		00440	200720 None	0	FALSE	A868, 2, 12	FALSE
5071__200720		5071_	200720 None	0	FALSE	None	FALSE
5072__200720		5072_	200720 None	0	FALSE	None	FALSE
5073__200720		5073_	200720 None	0	FALSE	None	FALSE
5074__200720		5074_	200720 None	0	FALSE	None	FALSE
5075__200720		5075_	200720 None	0	FALSE	None	FALSE
WGS_controle_200706	WGS_controle	200706 None		0	FALSE	B860, 3, 79	FALSE
WGS_controle_200713	WGS_controle	200713 None		0	FALSE	B860, 3, 79	FALSE
WGS_controle_200720	WGS_controle	200720 None		0	FALSE	B860, 3, 79	FALSE
WGS_controle_200724	WGS_controle	200724 None		0	FALSE	B860, 3, 79	FALSE
WGS_controle_200725	WGS_controle	200725 None		0	FALSE	B860, 3, 79	FALSE
WGS_controle_200727	WGS_controle	200727 None		0	FALSE	B860, 3, 79	FALSE

Appendix D

wgsid	loci	allele	pza resistance
A046	2289203	C	R
A046	2289203	C	R
A058	2289068	C	S
A058	2289068	C	S
A058	2289068	C	S
A058	2289068	C	
A058	2289068	C	I
A086	2288982	A	S
A086	2288982	A	
A117	2289168	A	
A117	2289168	A	
A117	2289168	A	
A117	2289168	A	S
A232	2289222	C	
A232	2289222	C	
A232	2289222	C	
A232	2289222	C	
A232	2289222	C	
A254	2289043	G	R
A254	2289043	G	
A301	2289040	C	R
A301	2289040	C	
A301	2289040	C	
A386	2288757	T	R
A527	2289137	T	S
A527	2289137	T	S
A625	2288868	C	R
A647	2289091	C	R
A657	2289231	C	R
A665	2289016	G	R
A814	2289213	C	R
A814	2289213	C	R
A886	2289090	C	R
A933	2289105	A	R
A953	2288718	G	R
A953	2288718	G	R
B095	2288820	G	R
B103	2289030	G	R
B141	2288800	A	
B156	2288832	C	R
B156	2288832	C	
B234	2288967	A	S
B254	2289070	C	S
B254	2289070	C	
B254	2289070	C	
B254	2289070	C	
B254	2289070	C	
B562	2288928	A	R
B562	2288928	A	
B562	2288928	A	
B688	2288773	A	S
B756	2288754	G	
B756	2288754	G	
B756	2288754	G	
B756	2288754	G	
B756	2288754	G	
B833	2289150	C	
B858	2289186	G	
B865	2289007	T	