Pervasive contaminations in sequencing experiments are a major source of false genetic variability: a *Mycobacterium tuberculosis* meta-analysis

Galo A. Goig[1], Silvia Blanco[2], Alberto L. Garcia-Basteiro[2,3,4], and Iñaki Comas[1,5] *

[1]Institute of Biomedicine of Valencia, IBV-CSIC, Valencia, Spain
[2]Centro de Investigaçao em Saúde de Manhiça (CISM), Maputo, Moçambique
[3]Global Health Institute Barcelona (ISGlobal), Barcelona, Spain
[4]Amsterdam Institute for Global Health and Development, Amsterdam University Medical Centers, Amsterdam, Netherlands
[5]CIBER in Epidemiology and Public Health, Madrid, Spain

Whole genome sequencing (WGS) is now playing a central role in the control and study of tuberculosis. Factors hindering genomic data interpretation can difficult our understanding of the pathogen biology and lead to incorrect clinical predictions. Analyzing an extensive *Mycobacterium tuberculosis* dataset comprising more than 1,500 sequencing samples from different published works, with additional samples sequenced in our laboratory, we find that contamination with non-target DNA is a common phenomenon among WGS studies. By using this data and *in-silico* simulations, we show that even subtle contaminations can produce dozens of false variants and large miscalculations of allele frequencies, often leading to errors that are very hard to detect and propagate through the analysis. In our dataset, 94% of the polymorphic positions were incorrectly identified due to contaminations. We exemplify the consequences of these errors in the context of clinical predictions for all the studies analyzed, and demonstrate that unexpected contaminations suppose a major pitfall in WGS studies. In addition, we present an approach based on the removal of contaminant reads that shows an outstanding performance analyzing both clean and contaminated data. Based in our findings, applicable to most of organisms, we urge for the implementation of contamination-aware analysis pipelines.

Tuberculosis | Whole-genome sequencing | Contamination | Variant calling
Correspondence: ***icomas@ibv.csic.es***

Introduction

Whole genome sequencing (WGS) has enhanced the study of complex biological phenomena in bacteria, such as population dynamics, host adaptation or outbreaks of microbial infections(1, 2). Democratization of high-throughput sequencing technologies and continuous improvements in laboratory procedures are also turning WGS into a promising alternative for the clinical diagnosis and surveillance of several pathogenic species (3–6). Since most genomic studies are based on the identification of nucleotide polymorphisms and calculation of allele frequencies, the accuracy of the variant calling process is pivotal to draw biological conclusions from WGS data. Notwithstanding, the performance of the analysis pipelines is seldomly evaluated, and possible factors hindering high-throughput sequencing data interpretation are usually overlooked.

Over the last years, groups working in different areas have shown that DNA contaminations is one factor potentially misleading biological conclusions (7–9). However, most bacterial genomic studies are still neglecting the impact of potential contaminations with non-target DNA in their analyses. This stems from the general assumption that pure culture sequencings only yield reads from the target organism and, if any, contaminating reads can hardly map to the target reference genome. As the decentralization of high-throughput sequencing and its associated analysis allow more groups to conduct WGS studies, controlling the factors that compromise the outcomes accuracy becomes crucial. This can be particularly relevant when studying bacterial organisms with low genetic diversity, for which evolutionary inferences and diagnostic may depend on few genetic variants.

Mycobacterium tuberculosis (MTB) represents a perfect example of such an organism. Tuberculosis is now recognized as the deadliest infectious disease in the world and one quarter of the global population is estimated to be infected (10). Due to its long culture-based diagnostic turnarounds, WGS is being extensively used not only to study its biology, but also in the clinical context(11, 12). Particularly, the ability of sequencing the complete genome of the pathogen from complex clinical samples will suppose a major breakthrough, now pursued by several groups (13, 14).

Here we use an extensive dataset comprising more than 1500 MTB sequencing samples coming from different WGS studies along with *in-silico* simulated data to assess the impact of contamination in WGS analysis. We show that presence of non-target reads is a common phenomenon among WGS studies, even when sequencing is performed from pure culture isolates. Importantly, we demonstrate that subtle contaminations represent a major pitfall, since they are prevalent, hard to detect, and can introduce dozens of false variants. In our dataset, a great proportion of the estimated genetic variability was due to contamination, and allele frequency calculations, the foundation of most genomic studies, were

largely altered. These alterations can lead to misinterpretation of WGS data as we demonstrate for two relevant applications: epidemiology and drug resistance prediction. In addition, we provide a contamination-aware workflow based on a taxonomic filter that exhibits an outstanding performance, allowing a robust analysis even for highly contaminated samples, such as sequencings from complex clinical specimens.

Results

DNA contamination is pervasive across WGS MTB studies and sample types. To look for contaminant DNA in WGS datasets, we used Kraken (15) to taxonomically classify the sequencing reads of 1553 sequencing runs from eight different MTB studies (here referred as the *experimental dataset*, see Methods). According to Kraken classifications, contamination is present to some extent in all studies analyzed (Figure 1). As expected, sequencings from direct clinical samples and early positive mycobacterial growth indicator tubes (MGIT), which are inoculated with primary clinical samples, present higher levels of contamination in terms of both the number of samples contaminated and the proportion of non-MTB DNA within them. Common contaminants for these samples comprise human DNA and bacteria usually found in oral and respiratory cavities like *Pseudomonas, Rothia, Streptococcus* or *Actinomyces* (16, 17), and can constitute virtually all reads in some samples. Importantly, contamination was also detected in studies in which the DNA sequenced came from pure culture isolates. For instance, *Bacillus*, *Negativicoccus* and *Enterococcus* represented up to the 68%, 58% and 32% respectively of different samples from the KwaZulu study. Strikingly, 17 out of 73 samples from Nigeria were identified as *Staphylococcus aureus* (92% to 99% of reads), probably due to a mistake during data uploading or mislabeling. The deep sequencing dataset was mostly free of contamination, with the exception of two samples for which a 3.32% of *Acinetobacter baumannii* and a 2.83% of non-tuberculosis mycobacteria (NTM) was identified (representing 795,887 and 920,379 reads respectively). Remarkably, Kraken left a high proportion of reads unclassified in many instances. This could be mainly due to either the absence of the organism from the database, or hard to classify sequences. Indeed, when using the NCBI blastn (18) to search a random subset of unclassified reads in the non-redundant database (nr), we observed three main patterns. Reads that either did not produce significant matches with any organism, or came from eukaryotes not present in our Kraken database; reads that produced partial alignments with many different taxa; and reads that produced good alignments, even with *Mycobacterium tuberculosis*, but having alignment identities around 90%, what makes Kraken unable to find exact matches of 31 base pairs.

Evaluating the impact of contamination in WGS analysis with simulated data. Firstly, we evaluated how non-MTB sequencing reads map to the MTB reference genome. In order to do this, we performed alignments of simulated sequencings for 45 organisms, including oral and respiratory microbiota, clinically common NTM and human. As expected, conserved genes like the 16S, rpoB or the tRNAs, constitute hotspots where contaminant sequences are easily aligned to. However, non-MTB alignments are produced across the reference genome (Figure 2a). Naturally, this is dependent on the phylogenetic relationship of the contaminant organism to the one being studied. For example, many more sequencing reads from *Nocardia* are aligned to the *Mycobacterium tuberculosis* complex (MTBC) reference genome than from distantly related organisms such as *Streptococcus* . Non-tuberculosis mycobacteria represent the best example of this as their read mappings can produce high sequencing depths along the MTBC reference genome. Importantly, human reads, which are usually the major concern in clinical samples, did not produced alignments at all.

Next, we assessed the performance of two contamination-aware approaches to avoid errors introduced by non-target reads: a filter removing low quality alignments, and a taxonomic filter removing non-MTBC reads. By measuring the mean sequencing depth obtained along the genome in 100-bp windows, we observed that both methods greatly reduced the number of non-MTB mappings (see Additional file 1). Remarkably, whereas the mapping filter was unable to eliminate contaminant alignments in several conserved regions, specially the 16S gene, the taxonomic filter displayed an outstanding performance eliminating all non-MTB mappings with the only exception of some *M. avium* reads (Additional file 2).

In order to evaluate how these contaminant alignments can affect WGS analysis, we simulated mock samples where sequencing reads from the reference genome were mixed with different proportions of non-MTBC organisms. In addition, we generated a version of the reference genome where random mutations were introduced. This allowed us to quantify both false positive and negative SNPs (see Methods). As shown in Figure 2b, slight contaminations can be responsible for a high number of false positive and negative variant calls. Predictably, the vast majority of the false positives corresponded to variable SNPs (vSNPs) and false positive fixed SNPs (fSNPs) only originated from extreme contaminations. Reduction of non-MTBC alignments mediated by the mapping filter led to a corresponding reduction in the number of false positive SNPs, although many erroneous variants still being called. In some cases, the number of false negatives was higher denoting the elimination of correct MTBC alignments by this filter. By contrast, the taxonomic filter eliminated nearly all erroneous SNP calls. Its performance was only compromised by contamination with *Mycobacterium avium*. In this case, the taxonomic filter was not able to avoid all false positive and negative SNPs introduced by contaminant *M. avium* reads. For example, when a 5% of *M. avium* was present, the 4,353 false positive SNPs and 71 false negative SNPs identified with a conventional pipeline, were reduced to 41 and 4 respectively applying the taxonomic filter. Importantly, applying the taxonomic filter to the uncontaminated mock sample of the reference genome did not produce any false negative.

Since many of the contaminant alignments are produced in conserved genes, all these erroneous calls led to the prediction of multiple false antibiotic resistances when using a conventional pipeline (see Additional file 1). Importantly, many of these false predictions corresponded to first line drugs, classified as high confidence mutations according to the PhyResSE reference catalog(19) (Additional file 3).

Normalization of SNPs by mapping to an MTBC ancestral genome. Normalizing the number of SNPs between samples allowed the direct comparison of WGS outcomes between studies and a detailed analysis of the impact produced by subtle contaminations. This normalization was possible by mapping to an inferred MTBC ancestral genome (20). To illustrate this, we compared the outcomes of a standard pipeline in the *experimental dataset* using the aforementioned ancestral genome and the conventional reference strain for MTB, H37Rv. In order to avoid noise introduced by possible contaminations, we only analyzed the samples with more than the 99% of the reads classified as MTBC and producing at least 40X of median sequencing depth (n=983). As shown in Figure 3, when mapping to the inferred MTBC ancestor (Figure 3.a), the number of fSNPs for all samples stabilized around a very narrow range, in opposite to the results obtained when mapping to H37Rv (Figure 3.b). In fact, 99% of samples coming from human MTB strains (n=964) were within a range of 668-953 fSNPs (median=860 fSNPs, interquartile range (IQR)=87). In contrast, when mapping to the reference strain, this range is dramatically extended to 381-1802 fSNPs (median=794 fSNPs, IQR=503). In the latter case, the lineage 4.10 strains, which H37Rv belongs to, are much closer (408 fSNPs on average) than lineage 2 (1178 fSNPs), or lineage 1 (1769 fSNPs) strains. When mapping to the MTBC ancestral genome, the remaining 1% of samples were out of this range due to low sequencing depths, coinfections where many fSNP are lost due to mixed calls, and high contaminations as will be detailed in the following sections. Animal strains had also a higher number of fSNPs to the ancestral reference than human strains (895-1325 fSNPs, median=936 fSNPs, IQR=248, n=19), what is probably linked to longer branches in the phylogeny. Based on this analysis we set up an expected range of SNPs for any given MTB sample: 650-1400 fSNPs to include animal strains. Through the following sections, we used this tool to focus our analysis on subtle, hard-to-detect changes introduced by contaminations.

Contamination often leads to subtle errors in variant calling that are elusive to conventional pipelines. Having established that contamination can have a large impact in SNP detection, we compared the outcomes of WGS analysis in the *experimental dataset* when using a conventional workflow with the taxonomic filter approach described. Overall, we found two main types of samples according to their level of contamination. High proportions of contaminating reads originate odd results such as very low sequencing depths, uneven genome coverages, or extreme number of variants. This type of contamination is less likely to suppose a real threat, as a careful analysis would probably detect it. Therefore, we focused our analysis on the subtle changes produced by moderate to light contaminations. To do this, we analyzed only the samples that were within the range of 650-1400 fSNPs to the MTBC ancestral genome when using a conventional pipeline, and had at least the 50% of sequencing reads classified as MTBC (n=1381 samples; 89.55% of the *experimental dataset*). These represent the samples in which neither the contaminations nor the errors produced would be easily detected. Comparing with the taxonomic filter approach revealed that variation in the number of SNPs (vSNPs + fSNPs) is a common fact among studies, with no exception (Figure 4). Globally, around the 10% of samples differed in at least 10 SNPs. In samples with 50% to 90% of MTBC (n=63), the median change in the number of SNPs was 81 (IQR=186). Notably, these differences are observed in samples with slight contaminations as well, as in 66 samples with 90% to 97% of MTBC we observed a median change of 13 SNPs (IQR=51). In fact, as shown in Figure 4, dozens of false SNPs can be called even for samples that could be considered "pure".
Analyzing samples with a percentage of MTBC reads greater than the 99% (n=1009; 72%) allowed us to confirm, along with the simulated data, that the taxonomic filter itself did not impact the results presented above. In these samples, the median difference in SNPs was one (IQR=2) before and after eliminating contamination. Indeed, for 377 of these samples (27%) no change in SNPs was observed at all.

Variant frequencies are extremely sensitive to contaminant sequences. Due to the relevance of variable positions and their individual frequencies in genome-based studies, we compared all the vSNPs called for the conventional and the taxonomic filter workflows. A total of 269,938 non-redundant variable positions were predicted by the conventional analysis pipeline. When applying the taxonomic filter, this number was dramatically reduced to only 16,464 non-redundant variable positions and, consequently, the 94% of the variants detected were attributed to noise introduced by contaminants. Importantly, non-target reads produced large fluctuations in SNP frequencies. The median difference in frequency for all vSNPs was 26% (IQR=39%) in contaminated samples (MTBC < 99%; n=383; 28% of samples), and the magnitude of the frequency shift was directly correlated to the level of contamination (Pearson correlation coefficient=0.87). In contrast, for the remaining 1009 clean samples, no difference in frequency was observed (median=0%, IQR=0.6%), thus confirming again that removal of non-MTBC reads does not impact the analysis in non-contaminated samples.

Contamination can affect clinical predictions from WGS data. To exemplify the extent of these alterations in WGS data interpretation, we evaluated how contamination can impact drug resistance predictions and epidemiology. When following the conventional analysis workflow, incorrect drug resistance susceptibility profiles (WGS-DST) were predicted for 53 samples according to known mutations described in the PhyResSE catalog (Additional File 1). Among these, 26 were false streptomycin resistances, 6 showed false

resistance profiles to first line drugs and 1 sample showed an undetected resistance to both fluoroquinolones and streptomycin. It must be noted that our pipeline reports any vSNP in a frequency of at least the 10%. Since WGS-DST strongly depends on the frequency of the mutations and the cutoffs used to call variable SNPs, we evaluated how contamination introduced variant frequency fluctuations among resistance associated genes. Whereas no change was observed for clean samples (median=0%; IQR=0%), strong fluctuations were observed in the case of contaminated samples (median=18%; IQR=43%). Strikingly, a large fraction of the vSNPs called in these genes showed a frequency shift of at least the 20% as a consequence of contaminating reads (Figure 5a).

In order to assess the impact of contamination in epidemiological predictions, we used the fSNPs to calculate the pairwise genetic distances for all samples belonging to the same study. We observed that contamination strongly affected genetic distance calculations when comparing the conventional and the taxonomic filter pipelines (chi square < 0.01). Remarkably, the level of contamination did not strongly correlate with the amount of change (Pearson correlation coefficient=0.51), in agreement with the fact that slight contaminations can be responsible for significant changes in the number of SNPs, and vice versa. Contamination largely depends on sample source and laboratory procedures. This fact is clearly reflected in the amount of change for different types of sample (Figure 5b). Genetic distance calculations are subject to greater alterations in sputum sample sequencings than in early MGIT culture sequencings (chi square < 0.01), and to greater alterations in the former than in standard cultures (chi square < 0.01). This distortion in genetic distance can lead to define wrong epidemiological relationships. For instance, a sample from Mozambique was unlinked from a transmission cluster due to a high number of contaminant reads. In this regard, it is particularly interesting that some samples from the sputum capture-sequencing study showed several fSNPs of distance to their respective matching cultures, four of which were above typical transmission cutoffs (17, 19, 20 and 35 fSNPs respectively). When removing non-MTBC reads, all genetic distances between sputum capture-sequencings and their respective matching cultures were of 0 fSNPs.

Discussion

In this work we use a taxonomic read classifier to check for contamination in an extensive dataset of *M. tuberculosis* WGS samples, demonstrating that non-target reads can be found in any type of sample, regardless of the experiment being conducted or the specimen source. The assumption that WGS from pure cultures only generates reads from the target organism, and that non-target reads are barely mapped to the reference genome, has encouraged pipelines vulnerable to contamination. Only few works included bioinformatic steps aimed to deal with DNA contamination and, even in these cases, the particular approaches used have not been evaluated (13, 21, 22).

A recent report focused on the artifactual variation in WGS of *Mycobacterium tuberculosis* from MGIT cultures hypothesized that removal of non-mycobacterial reads prior to mapping might lead to the loss of several SNPs (23). Remarkably, the authors considered mycobacterial DNA as a whole, disregarding non-tuberculosis mycobacteria. Here we show that both NTM and non-mycobacterial DNA can be found in cultures and primary diagnostic samples (Figure 1), which sequencing reads can be aligned along the MTBC reference genome (Figure 2a), leading to many false positive and negative variant predictions (Figure 2b). We use both simulated and real sequencing experiments to demonstrate that removal of contaminant reads prior to mapping is a solid approach not affecting either the number of variants detected nor the frequencies calculated. Of course, the efficacy of this methodology will directly depend on both the accuracy of the taxonomic classifier and the reference database used. In the case of MTB, Kraken showed an outstanding performance, although it was unable to distinguish some reads of *M. avium*, a closely related organism to the MTBC. In these cases, better databases or slower alignment-based algorithms like BLAST could be used to improve the taxonomic filter accuracy.

While developing the pipeline to understand the impact of contamination, we realized about the importance of the reference genome chosen for mapping. Most bioinformatic pipelines for *Mycobacterium tuberculosis* align the sequencing reads to the reference strain H37Rv. This strain has been comprehensively studied since its isolation in 1905 and has been historically used as reference genome as it was the first *Mycobacterium tuberculosis* genome ever sequenced (24). H37Rv belongs to the lineage 4 of human MTBC and, as a result, the variation expected when mapping to it deeply depends on how phylogenetically close is the strain under study. In contrast, by mapping to an inferred ancestral MTBC genome we are able to establish an expected range of SNPs for any given MTB sample. We use this range as a tool for quality control, rapidly identifying extreme contaminations, low quality sequencings, or super-infections with different strains.

Using the taxonomic filter approach we estimate that around 94% of the variants identified in clinical samples from eight different WGS studies were introduced by contaminant reads. Importantly, we show that many false SNPs can be called regardless of the extent of contamination, with dozens of incorrect calls even in samples that would normally be considered as "pure". All these alterations can have a massive impact in population diversity estimates, distorting the biological conclusions that are drawn from WGS data. Studies trying to estimate a molecular clock, or the action of selective forces could be completely biased, since they are often based on the frequency spectrum derived from low frequency variants (25). We exemplify this issue by comparing the outcomes of a conventional pipeline with a contamination-aware approach based on a taxonomic filter for two relevant applications: epidemiology and prediction of drug-resistant phenotypes. When estimating transmission, we observed a considerable fSNPs-distance variation for many of the pairwise distances calculated. This is of particular relevance in slow-evolving organisms like MTB, for which transmission events

are determined based on few variants. Transmission estimation can be affected by either false positive or false negative fSNPs, since contaminant reads can make the frequencies fluctuate around the threshold used to call fixed variants (90% in this work). This holds true when performing WGS-DST, as predictions based on the observation of certain mutations will finally depend on the thresholds used to call variable positions (10% in this work).

Most of the contaminations and the errors they lead to are very likely to pass unnoticed to conventional pipelines. We claim that inclusion of analysis steps aimed to deal with contaminant data are indispensable in any bacterial WGS bioinformatic workflow. This will be particularly relevant in large sequencing projects directed towards the discovery of new genotype to phenotype associations and when performing direct sequencing from complex clinical samples.

Methods

Experimental and simulated datasets analyzed. In order to evaluate the extent of contaminations among different MTB WGS studies and their impact in the analysis outcomes, we analyzed tuberculosis WGS runs from eight different studies. Seven of these datasets were publicly available beforehand (13, 14, 21, 26–29). Another dataset was generated in our laboratory, corresponding to an Illumina MiSeq sequencing of 138 samples from Mozambique see Additional file 1. All these samples are referred as the *experimental dataset* and comprised a total of 1553 Illumina runs (Table 1).

To validate our analysis and quantify the impact of contamination under controlled conditions, we generated a simulated dataset. This dataset consisted of simulated Illumina MiSeq 250bp paired-end sequencing experiments using ART (30) for the MTBC reference genome, the human genome (GRCh38, Ensembl release 81) and 44 different non-MTBC bacterial species (see Additional file 1).

Contamination assessment using Kraken. In order to assess the contamination level in each sample, all sequencing reads were taxonomically classified using Kraken (15) with a custom database comprising all sequences of bacteria, archaea, virus, protozoa, plasmids and fungi in RefSeq (release 78), plus the human genome (GRCh38, Ensembl release 81).

Use of an ancestral genome to normalize the number of SNPs. In this work we normalize the number of variants for any given MTB sample by mapping to the inferred ancestral genome of the MTBC, establishing a range of expected SNPs that we use as a tool for quality control. To highlight the hazard that unexpected contaminations suppose to WGS outcomes, we focus our analysis in samples presenting variations within this range that could, therefore, pass completely unnoticed when using a conventional pipeline. To better illustrate this, we analyzed all the samples from the *experimental dataset* as described in the variant calling section mapping to both the conventional reference strain H37Rv (NC_000962.3) and the aforementioned MTBC ancestral genome (20). We compared these two approaches in terms of the number of fixed SNPs called for all non-contaminated samples, considering non-contaminated those with more than the 99% of the reads classified as MTBC and producing at least 40X of median sequencing depth.

Analysis pipeline: impact of contaminant DNA in variant calling outcomes. In order to evaluate the impact of DNA contamination in WGS analysis, each sample was analyzed following three different methods. Firstly, each sample was analyzed using a conventional tuberculosis WGS pipeline. In summary, reads were trimmed and filtered using Trimmomatic (31) to remove low-quality sequences and then mapped to the MTBC ancestral genome using bwa mem (32). Variants were then called and filtered using VarScan2 (33) with two different set of parameters. To study transmission we used high-stringent parameters to obtain high confidence fixed variants (fSNPs) (minimum depth of 20 reads, average base quality of 20, p-value cutoff 0.01, observed in both strands and minimum frequency of 90%), including removal of SNPs in repetitive and mobile regions, SNPs called within a 4bp window from deleted positions and high SNP density regions (allowing a maximum of 3 SNPs in 10bp windows). In order to predict drug resistance and study bacterial subpopulations, we used another set of parameters that allowed us to look into variable SNPs (vSNPs) (minimum depth of 10 reads and variant observed in at least 6 reads, average base quality of 20, p-value cutoff 0.01, observed in both strands and minimum frequency of 10%), including removal of SNPs in repetitive and mobile regions, and SNPs called near deleted positions within a window of 4bp. Sequencing depth was calculated for each sample using bedtools (34). Samples were then reanalyzed, following the same steps, but adding a previous taxonomic filter removing those reads classified by Kraken as any species other than *Mycobacterium tuberculosis* complex. Alternatively, samples were reanalyzed substituting this taxonomic filter by a custom alignment filter. Different parameter combinations were tested to maximize removal of non-MTB mappings, minimizing loss of MTB mappings (see Additional file 1). The filter finally consisted in the removal of alignments with length, identity and mapping quality below 40bp, 97% and 60 respectively.

Estimation of false positive and negative calls in the *simulated dataset*. In order to inspect which regions of the reference genome are susceptible of recruiting non-MTBC reads, we analyzed the samples of the *simulated dataset* as described in the analysis pipeline and then measured the mean sequencing depth along the genome in 100 bp windows. To assess whether false positive SNPs and drug resistance predictions are produced by these non-MTBC mappings, we generated mock contaminated samples by mixing simulated sequencing reads from the MTBC ancestor reference with different proportions (5%, 15%, 30% and 70%) of other organisms corresponding to 12 common contaminants identified in the *experimental dataset*. Therefore, any SNP identified when analyzing these samples, would be a false

positive SNP imputable to contamination. In addition, we mapped these mock samples to a modified version of the reference genome where we introduced two types of mutations. Random mutations each 100bp, and all the drug resistance conferring mutations described as "high confidence" in the PhyResSE catalog (19). Therefore, any of the introduced SNPs that were undetected when analyzing these samples, would be false negative SNPs attributable to contamination.

Pairwise genetic distances and drug resistance prediction in the *experimental dataset*. For each study, genetic distances between samples were calculated from multiple alignments of non-redundant variable positions across samples using the R library APE (35) as the fSNP distance, disregarding resistance-associated positions and using pairwise deletion for gap positions. Samples with at least 650 fixed SNPs to the MRCA genome and a median coverage of 20X were grouped in the same transmission cluster using a maximum distance threshold of 15 fSNPs to cover a range of recent and older transmission events. Drug resistance was predicted for samples with a median coverage of at least 20X according to the PhyResSE catalog of known resistance conferring mutations. Mutations in resistance associated genes that were not in this list and not described as phylogenetic polymorphisms in the literature, were annotated as "candidate resistance mutations" (see Additional file 1).

ACKNOWLEDGEMENTS
This work was funded by projects of the European Research Council (ERC) (638553-TB-ACCELERATE) and Ministerio de Economia y Competitividad (Spanish Government) research grant SAF2016-77346-R (to IC), and BES-2014-071066 (to GAG).

AUTHOR CONTRIBUTIONS
GAG and IC designed the study, analyzed the data and wrote the manuscript. AGB and SB provided the samples from Mozambique.

COMPETING FINANCIAL INTERESTS
The authors declare that they have no competing interests.

Bibliography

1. Paul R McAdam, Emily J Richardson, and J Ross Fitzgerald. High-throughput sequencing for the study of bacterial pathogen biology. *Current Opinion in Microbiology*, 19:106–113, June 2014. ISSN 1369-5274. doi: 10.1016/j.mib.2014.06.002.
2. Xavier Didelot, Rory Bowden, Daniel J. Wilson, Tim E. A. Peto, and Derrick W. Crook. Transforming clinical microbiology with bacterial genome sequencing. *Nature reviews. Genetics*, 13(9):601–612, September 2012. ISSN 1471-0056. doi: 10.1038/nrg3226.
3. David J. Roach, Joshua N. Burton, Choli Lee, Bethany Stackhouse, Susan M. Butler-Wu, Brad T. Cookson, Jay Shendure, and Stephen J. Salipante. A Year of Infection in the Intensive Care Unit: Prospective Whole Genome Sequencing of Bacterial Clinical Isolates Reveals Cryptic Transmissions and Novel Microbiota. *PLOS Genetics*, 11(7):e1005413, July 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005413.
4. Henk C. den Bakker, Marc W. Allard, Dianna Bopp, Eric W. Brown, John Fontana, Zamin Iqbal, Aristea Kinney, Ronald Limberger, Kimberlee A. Musser, Matthew Shudt, Errol Strain, Martin Wiedmann, and William J. Wolfgang. Rapid Whole-Genome Sequencing for Surveillance of Salmonella enterica Serovar Enteritidis. *Emerging Infectious Diseases*, 20 (8):1306–1314, August 2014. ISSN 1080-6040. doi: 10.3201/eid2008.131399.
5. Mette T. Christiansen, Amanda C. Brown, Samit Kundu, Helena J. Tutill, Rachel Williams, Julianne R. Brown, Jolyon Holdstock, Martin J. Holland, Simon Stevenson, Jayshree Dave, CY William Tong, Katja Einer-Jensen, Daniel P. Depledge, and Judith Breuer. Whole-genome enrichment and sequencing of Chlamydia trachomatisdirectly from clinical samples. *BMC Infectious Diseases*, 14:591, November 2014. ISSN 1471-2334. doi: 10.1186/s12879-014-0591-3.
6. Dhruba J. SenGupta, Lisa A. Cummings, Daniel R. Hoogestraat, Susan M. Butler-Wu, Jay Shendure, Brad T. Cookson, and Stephen J. Salipante. Whole-Genome Sequencing for High-Resolution Investigation of Methicillin-Resistant Staphylococcus aureus Epidemiology and Genome Plasticity. *Journal of Clinical Microbiology*, 52(8):2787–2796, January 2014. ISSN 0095-1137, 1098-660X. doi: 10.1128/JCM.00759-14.
7. Richard W. Lusk. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PloS One*, 9(10):e110808, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0110808.
8. Christopher G. Wilson, Reuben W. Nowell, and Timothy G. Barraclough. Cross-Contamination Explains "Inter and Intraspecific Horizontal Genetic Transfers" between Asexual Bdelloid Rotifers. *Current biology: CB*, July 2018. ISSN 1879-0445. doi: 10.1016/j.cub.2018.05.070.
9. Marion Ballenghien, Nicolas Faivre, and Nicolas Galtier. Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC biology*, 15(1):25, 2017. ISSN 1741-7007. doi: 10.1186/s12915-017-0366-6.
10. WORLD HEALTH ORGANIZATION. *GLOBAL TUBERCULOSIS REPORT 2017*. WORLD HEALTH ORGANIZATION, S.l., 2017. ISBN 978-92-4-156551-6. OCLC: 1017579392.
11. Timothy M Walker, Camilla LC Ip, Ruth H Harrell, Jason T Evans, Georgia Kapatai, Martin J Dedicoat, David W Eyre, Daniel J Wilson, Peter M Hawkey, Derrick W Crook, Julian Parkhill, David Harris, A Sarah Walker, Rory Bowden, Philip Monk, E Grace Smith, and Tim EA Peto. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious Diseases*, 13(2):137–146, February 2013. ISSN 1473-3099. doi: 10.1016/S1473-3099(12)70277-3.
12. Timothy M Walker, Thomas A Kohl, Shaheed V Omar, Jessica Hedge, Carlos Del Ojo Elias, Phelim Bradley, Zamin Iqbal, Silke Feuerriegel, Katherine E Niehaus, Daniel J Wilson, David A Clifton, Georgia Kapatai, Camilla L C Ip, Rory Bowden, Francis A Drobniewski, Caroline Allix-Béguec, Cyril Gaudin, Julian Parkhill, Roland Diel, Philip Supply, Derrick W Crook, E Grace Smith, A Sarah Walker, Nazir Ismail, Stefan Niemann, and Tim E A Peto. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. *The Lancet Infectious Diseases*, 15(10): 1193–1202, October 2015. ISSN 1473-3099. doi: 10.1016/S1473-3099(15)00062-6.
13. Amanda C. Brown, Josephine M. Bryant, Katja Einer-Jensen, Jolyon Holdstock, Darren T. Houniet, Jacqueline Z. M. Chan, Daniel P. Depledge, Vladyslav Nikolayevskyy, Agnieszka Broda, Madeline J. Stone, Mette T. Christiansen, Rachel Williams, Michael B. McAndrew, Helena Tutill, Julianne Brown, Mark Melzer, Caryn Rosmarin, Timothy D. McHugh, Robert J. Shorten, Francis Drobniewski, Graham Speight, and Judith Breuer. Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. *Journal of Clinical Microbiology*, 53(7):2230–2237, July 2015. ISSN 0095-1137. doi: 10.1128/JCM.00486-15.
14. Antonina A. Votintseva, Phelim Bradley, Louise Pankhurst, Carlos del Ojo Elias, Matthew Loose, Kayzad Nilgiriwala, Anirvan Chatterjee, E. Grace Smith, Nicolas Sanderson, Timothy M. Walker, Marcus R. Morgan, David H. Wyllie, A. Sarah Walker, Tim E. A. Peto, Derrick W. Crook, and Zamin Iqbal. Same-day diagnostic and surveillance data for tuberculosis via whole genome sequencing of direct respiratory samples. *Journal of Clinical Microbiology*, pages JCM.02483–16, March 2017. ISSN 0095-1137, 1098-660X. doi: 10.1128/JCM.02483-16.
15. Derrick E. Wood and Steven L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15:R46, March 2014. ISSN 1474-760X. doi: 10.1186/gb-2014-15-3-r46.
16. Jørn A. Aas, Bruce J. Paster, Lauren N. Stokes, Ingar Olsen, and Floyd E. Dewhirst. Defining the Normal Bacterial Flora of the Oral Cavity. *Journal of Clinical Microbiology*, 43(11):5721–5732, November 2005. ISSN 0095-1137. doi: 10.1128/JCM.43.11.5721-5732.2005.
17. Christine M. Bassis, John R. Erb-Downward, Robert P. Dickson, Christine M. Freeman, Thomas M. Schmidt, Vincent B. Young, James M. Beck, Jeffrey L. Curtis, and Gary B. Huffnagle. Analysis of the Upper Respiratory Tract Microbiotas as the Source of the Lung and Gastric Microbiotas in Healthy Individuals. *mBio*, 6(2), March 2015. ISSN 2150-7511. doi: 10.1128/mBio.00037-15.
18. Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2.
19. Silke Feuerriegel, Viola Schleusener, Patrick Beckert, Thomas A. Kohl, Paolo Miotto, Daniela M. Cirillo, Andrea M. Cabibbe, Stefan Niemann, and Kurt Fellenberg. PhyResSE: a Web Tool Delineating Mycobacterium tuberculosis Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. *Journal of Clinical Microbiology*, 53(6):1908–1914, June 2015. ISSN 0095-1137. doi: 10.1128/JCM.00025-15.
20. Iñaki Comas, Jaidip Chakravartti, Peter M. Small, James Galagan, Stefan Niemann, Kristin Kremer, Joel D. Ernst, and Sebastien Gagneux. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nature Genetics*, 42(6):498–503, June 2010. ISSN 1061-4036. doi: 10.1038/ng.590.
21. Louise J Pankhurst, Carlos del Ojo Elias, Antonina A Votintseva, Timothy M Walker, Kevin Cole, Jim Davies, Jilles M Fermont, Deborah M Gascoyne-Binzi, Thomas A Kohl, Clare Kong, Nadine Lemaitre, Stefan Niemann, John Paul, Thomas R Rogers, Emma Roycroft, E Grace Smith, Philip Supply, Patrick Tang, Mark H Wilcox, Sarah Wordsworth, David Wyllie, Li Xu, and Derrick W Crook. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *The Lancet. Respiratory Medicine*, 4 (1):49–58, January 2016. ISSN 2213-2600. doi: 10.1016/S2213-2600(15)00466-X.
22. José Afonso Guerra-Assunção, Rein M. G. J. Houben, Amelia C. Crampin, Themba Mzembe, Kim Mallard, Francesc Coll, Palwasha Khan, Louis Banda, Arthur Chiwaya, Rui P. A. Pereira, Ruth McNerney, David Harris, Julian Parkhill, Taane G. Clark, and Judith R. Glynn. Recurrence due to Relapse or Reinfection With Mycobacterium tuberculosis: A Whole-Genome Sequencing Approach in a Large, Population-Based Cohort With a High HIV Infection Prevalence and Active Follow-up. *The Journal of Infectious Diseases*, 211(7): 1154–1163, April 2015. ISSN 0022-1899. doi: 10.1093/infdis/jiu574.
23. David H. Wyllie, Nicholas Sanderson, Richard Myers, Tim Peto, Esther Robinson, Derrick W. Crook, E. Grace Smith, and A. Sarah Walker. Control of artefactual variation in reported inter-sample relatedness during clinical use of a Mycobacterium tuberculosis sequencing pipeline. *Journal of Clinical Microbiology*, pages JCM.00104–18, June 2018. ISSN 0095-1137, 1098-660X. doi: 10.1128/JCM.00104-18.
24. 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, and B. G. Barrell. Deciphering the bi-

ology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 393 (6685):537–544, June 1998. ISSN 1476-4687. doi: 10.1038/31159.

25. Xavier Didelot, A. Sarah Walker, Tim E. Peto, Derrick W. Crook, and Daniel J. Wilson. Within-host evolution of bacterial pathogens. *Nature reviews. Microbiology*, 14(3):150–162, March 2016. ISSN 1740-1526. doi: 10.1038/nrmicro.2015.13.
26. Keira A. Cohen, Thomas Abeel, Abigail Manson McGuire, Christopher A. Desjardins, Vanisha Munsamy, Terrance P. Shea, Bruce J. Walker, Nonkqubela Bantubani, Deepak V. Almeida, Lucia Alvarado, Sinéad B. Chapman, Nomonde R. Mvelase, Eamon Y. Duffy, Michael G. Fitzgerald, Pamla Govender, Sharvari Gujja, Susanna Hamilton, Clinton Howarth, Jeffrey D. Larimer, Kashmeel Maharaj, Matthew D. Pearson, Margaret E. Priest, Qiandong Zeng, Nesri Padayatchi, Jacques Grosset, Sarah K. Young, Jennifer Wortman, Koleka P. Mlisana, Max R. O'Donnell, Bruce W. Birren, William R. Bishai, Alexander S. Pym, and Ashlee M. Earl. Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. *PLOS Medicine*, 12(9):e1001880, September 2015. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001880.
27. Kurt R. Wollenberg, Christopher A. Desjardins, Aksana Zalutskaya, Vervara Slodovnikova, Andrew J. Oler, Mariam Quiñones, Thomas Abeel, Sinead B. Chapman, Michael Tartakovsky, Andrei Gabrielian, Sven Hoffner, Aliaksandr Skrahin, Bruce W. Birren, Alexander Rosenthal, Alena Skrahina, and Ashlee M. Earl. Whole-Genome Sequencing of Mycobacterium tuberculosis Provides Insight into the Evolution and Genetic Composition of Drug-Resistant Tuberculosis in Belarus. *Journal of Clinical Microbiology*, 55(2):457–469, February 2017. ISSN 0095-1137. doi: 10.1128/JCM.02116-16.
28. Andrej Trauner, Qingyun Liu, Laura E. Via, Xin Liu, Xianglin Ruan, Lili Liang, Huimin Shi, Ying Chen, Ziling Wang, Ruixia Liang, Wei Zhang, Wang Wei, Jingcai Gao, Gang Sun, Daniela Brites, Kathleen England, Guolong Zhang, Sebastien Gagneux, Clifton E. Barry, and Qian Gao. The within-host population dynamics of Mycobacterium tuberculosis vary with treatment efficacy. *Genome Biology*, 18:71, April 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1196-0.
29. Madikay Senghore, Jacob Otu, Adam Witney, Florian Gehre, Emma L. Doughty, Gemma L. Kay, Phillip Butcher, Kayode Salako, Aderemi Kehinde, Nneka Onyejepu, Emmanuel Idigbe, Tumani Corrah, Bouke de Jong, Mark J. Pallen, and Martin Antonio. Whole-genome sequencing illuminates the evolution and spread of multidrug-resistant tuberculosis in Southwest Nigeria. *PLOS ONE*, 12(9):e0184510, September 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0184510.
30. Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, February 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr708.
31. Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu170.
32. Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp324.
33. Daniel C. Koboldt, Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, March 2012. ISSN 1088-9051. doi: 10.1101/gr.129684.111.
34. Aaron R. Quinlan. BEDTools: the Swiss-army tool for genome feature analysis. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, 47:11.12.1–11.12.34, September 2014. ISSN 1934-3396. doi: 10.1002/0471250953.bi1112s47.
35. Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20(2):289–290, January 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg412.

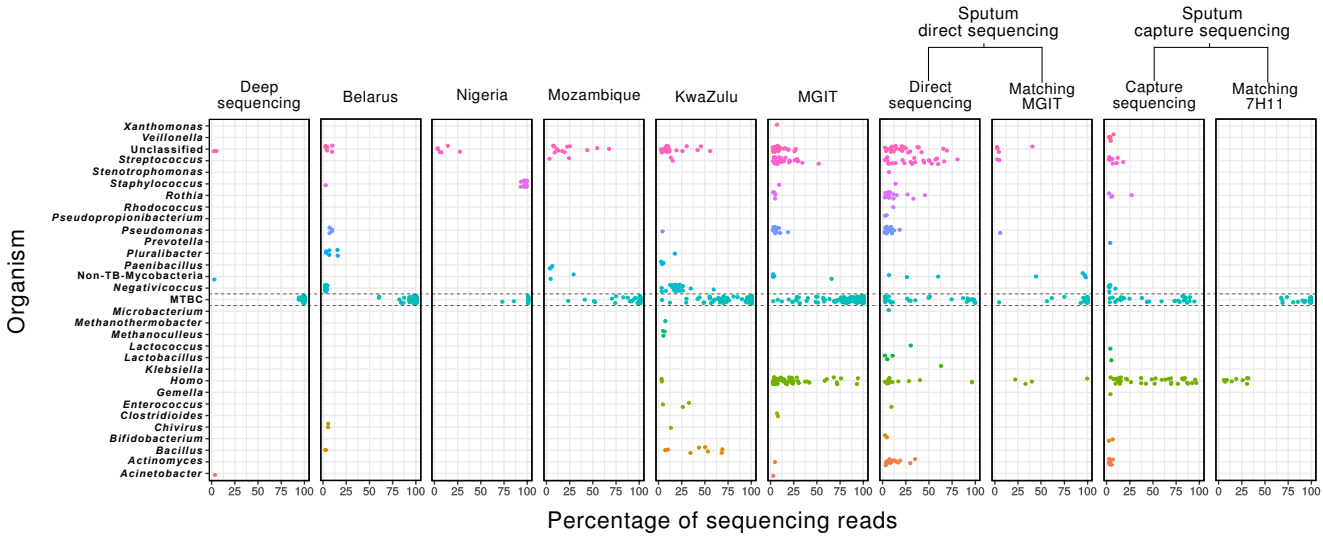

Fig. 1. Samples with contaminant sequencing reads are present in all the studies analyzed. Proportion of sequencing reads for different organisms across samples from the *experimental dataset*. Each dot represents a sample with a given percentage of sequencing reads coming from the genus indicated in the y-axis. Only organisms in a proportion above 2% are shown.

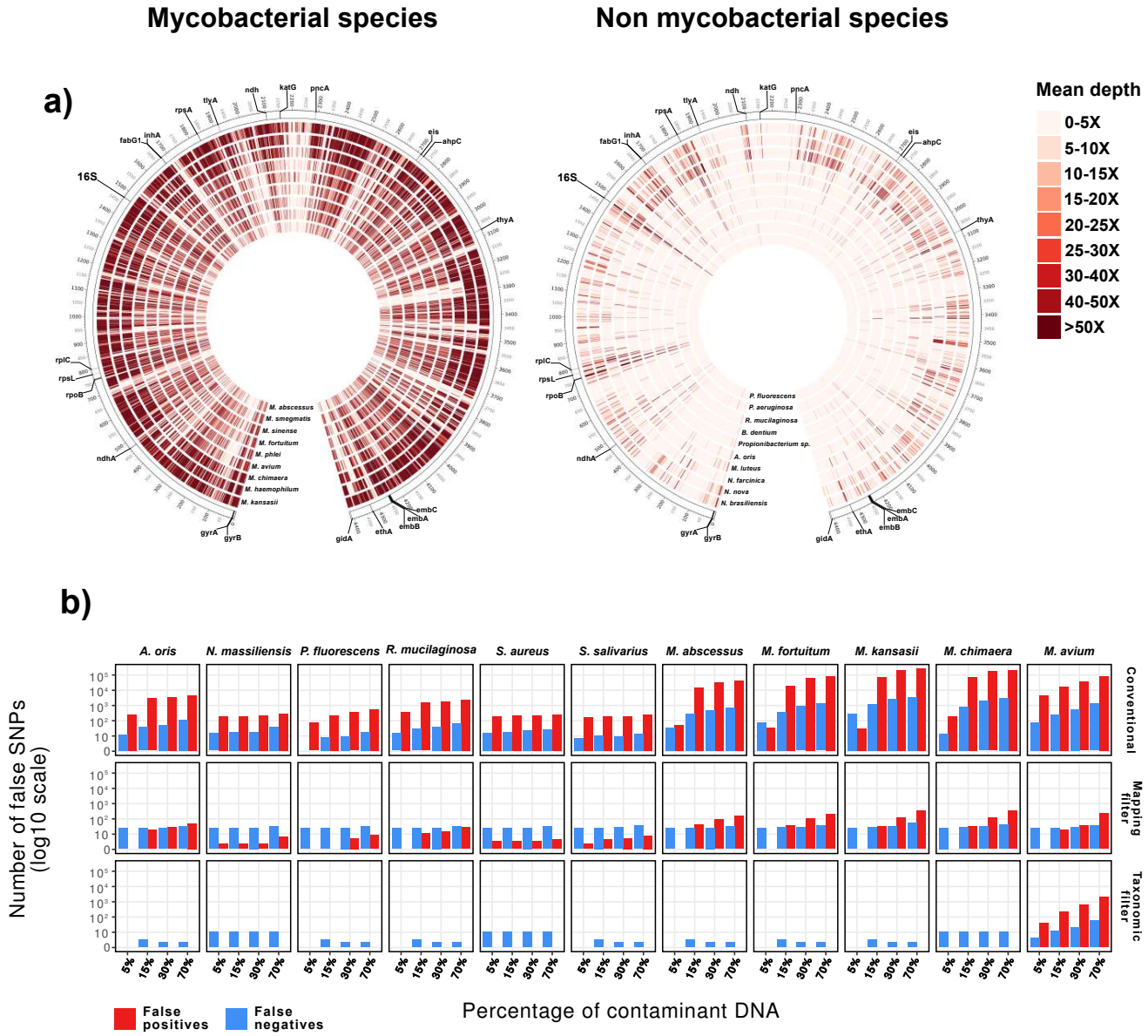

Fig. 2. Mapping of non-MTBC reads along the MTBC reference genome lead to false positive and negative calls. a) Mean sequencing depth obtained along the MTBC reference genome across 100 bp windows when mapping 1,500,000 simulated reads of non-tuberculosis mycobacteria and organisms other than mycobacteria (OTM). The top 10 organisms OTM that produced higher sequencing depths are shown. **b)** Number of false positive and negative SNPs (note log10 scale) called for mock samples *in-silico* contaminated with different proportion of non-MTB organisms when following three different analysis pipelines (conventional, mapping filter and taxonomic filter). In the case of the taxonomic filter, only contamination with *M. avium* produced false positive calls. Contamination with human DNA did not produce any false positive SNP

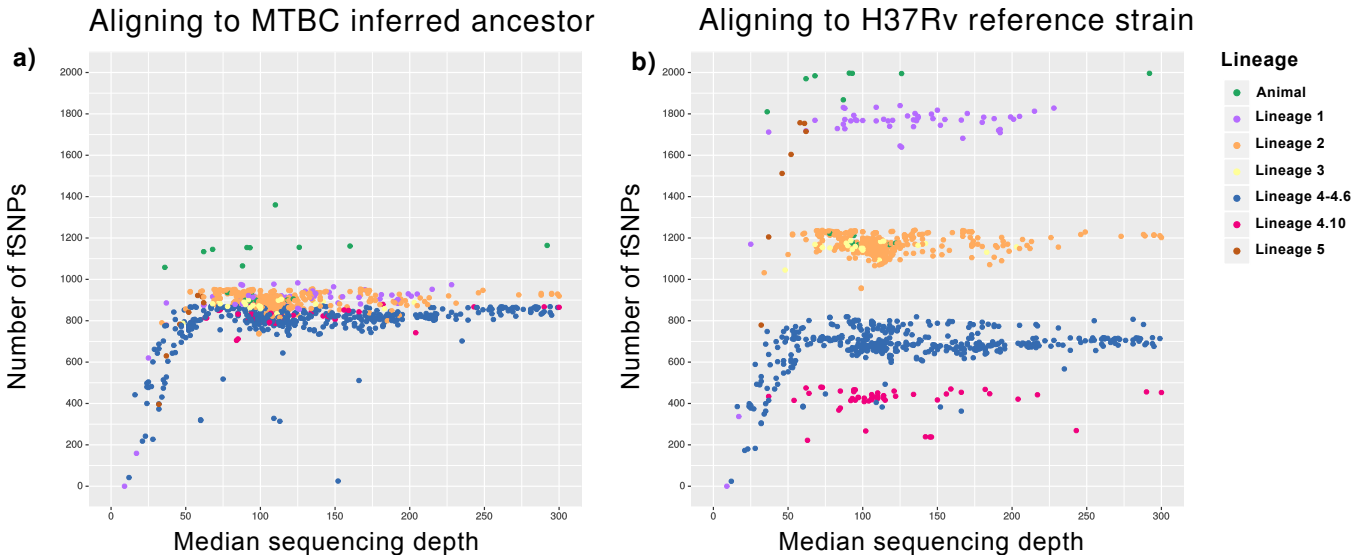

Fig. 3. Aligning to the inferred MTBC ancestor allows to define a range of expected SNPs for MTB samples. a) Number of fixed SNPs called for all samples of the *experimental dataset* when using the MTBC inferred ancestor as reference genome. **b)** Number of fSNPs called when using the H37Rv strain as reference genome. Only samples with more than 99% of MTBC reads are shown. Samples with median sequencing depths below 40X are shown only for illustrative purposes but were not considered in the analysis.

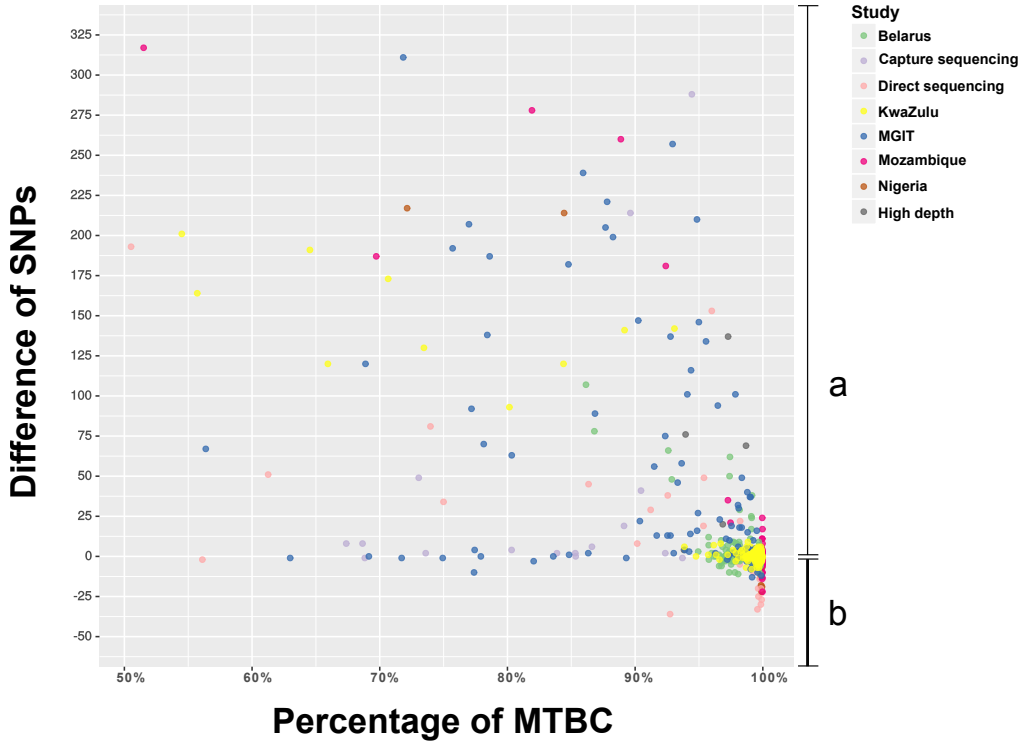

Fig. 4. Contamination is responsible for many false positive and negative SNPs in the *experimental dataset*. Difference between the number of SNPs called with the conventional and the taxonomic filter pipelines. Each dot, representing a sample, is positioned in the x-axis according to the percentage of sequencing reads classified as *Mycobacterium tuberculosis* Complex. **a)** False positive SNPs attributable to contamination. **b)** False negative SNPs attributable to contamination. Only samples with more than 50% of MTBC and 650-1400 fSNPs to the MTBC MRCA reference are shown.

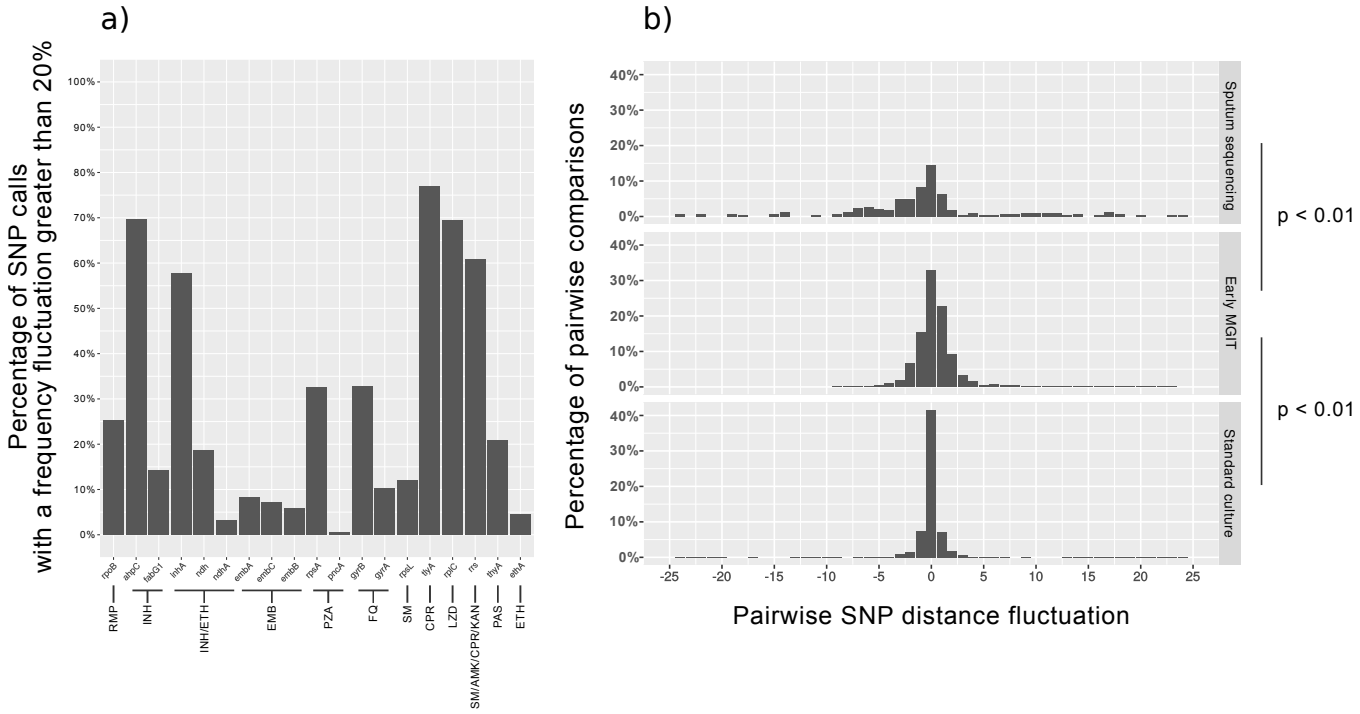

Fig. 5. Contaminant reads can lead to false clinical predictions. a) Fraction of vSNPs called within each drug-resistance associated gene that showed a frequency shift greater than the 20% as a consequence of contamination. AMK (Amikacin); CPR (Capreomycin); EMB (Ethambutol); ETH (Ethionamide); FQ (Fluoroquinolones); INH (Isoniazid); KAN (Kanamycin); LZD (Linezolid); PAS (Para-aminosalicylic acid); PZA (Pyrazinamide); RMP (Rifampicin); SM (Streptomycin). **b)**Distribution of fSNP distance shifts introduced by contamination for each type of sample. Negative fluctuations correspond to false negative fSNPs of distance as a consequence of contamination, whereas positive fluctuations correspond to false positive fSNPs

Table 1. Studies of the *experimental dataset*. *This study included sequencings from non-MTB organisms. We analyzed the 168 reported as MTB by the authors.

Study	Publication	Runs analyzed	Accession
Mozambique	Unpublished	138	PRJEB27421
Kwazulu-Natal	Cohen *et al.* 2015	433	PRJNA183624
Nigeria	Senghore *et al.* 2017	73	PRJEB15857
Belarus	Wollenberg *et al.* 2017	552	PRJNA200335
High depth	Trauner *et al.* 2017	63	PRJEB13325, PRJEB17864
Sputum capture-sequencing	Brown *et al.* 2015	58	PRJEB9206
Sputum direct-sequencing	Votintseva *et al.* 2017	68	SRP093599
MGIT sequencing	Pankhurst *et al.* 2016	168*	PRJNA268101, PRJNA302362