Supplemental material

Computational methods

jQuery[1]. Upload was implemented using the jQuery file upload plugin supporting multiple file selection, progress bars and chunked uploads. Sessions are implemented by using the CGI-Session Perl Module, with the 32-character session keys being transmitted over the internet only after an SSL encryption has been established, and being stored client-side as cookies.

FastqValidator [2]. The C++ program validates the FastQ format [3], e. g. verifying that the four mandatory lines per read (sequence identifier line, raw sequence line, plus line and quality string line) are formatted correctly, read IDs are unique and raw sequence line and quality string line have same lengths (http://genome.sph.umich.edu/wiki/FastQ_Validation_Criteria).

FastQC. FastQC is a quality control tool for NGS data implemented in Java [4]. It provides an overview of data quality and typical problems such as per base sequence quality, per sequence quality scores, per base GC content, per sequence GC content, sequence length distribution, and overrepresented sequences or kmers. The information is visualized in summary graphs, properly highlighting any problems or biases. At the beginning of the report, a brief summary outlines if the data meet the criteria (green tick), are slightly problematic (orange triangle) or failed the test (red cross).

BWA-MEM. The Burrwos Wheeler Alignment (BWA) Maximal Exact Matches (MEM) algorithm is a fast and accurate means to map NGS reads [5]. It is based on the Burrows-Wheeler transformation of the reference sequence, here *M. tuberculosis* H37Rv (NC_000962.3, http://www.ncbi.nlm.nih.gov/nuccore/NC_000962.3). The algorithm supports read lengths ranging from about 70bp to a few megabases, allows for permissive gaps and chimeric alignments and produces standard Sequence Alignment/Map format (SAM) output.

QualiMap. QualiMap is a quality control tool for evaluating mapping results in SAM or BAM format [6]. Amongst other criteria, genome coverage and mapping quality distribution across the genome as well as nucleotide distribution and mean and median insert size are assessed. As with FastQC, the output comprises proper visualization highlighting potential problems, here related to the mapping process or problems that become visible only after mapping.

SAMtools. SAMtools is a tool collection for further processing SAM files [7]. It enables important preprocessing steps. Here, we use it for sorting (samtools sort) and indexing (samtools index) the alignment, as well as for conversion into the binary form BAM (samtools view) and to remove duplicates originated by PCR artefacts (samtools rmdup). Furthermore, we use SAMtools to visualize the alignment (samtools tview).

Genome Analysis Toolkit. The Genome Analysis Toolkit (GATK) is implemented in Java [8, 9]. It comprises additional tools required before calling variants in order to considerably reduce the false discovery rate (FDR) such as base-recalibration and re-alignment around small insertions and deletions (indels). Furthermore, it also supplies different variant callers. Here, we used its functions RealignerTargetCreator, IndelRealigner, BaseRecalibrator, PrintReads, and UnifiedGenotyper.

The pipeline

From a user's point of view one page displaying bacterial strain/lineage and antibiotics resistance detected would be most convenient and sufficient in an ideal world. The raw WGS data must not be taken at face value, however. They are prone to artifacts, low base quality, or may simply be too few to reliably cover all positions in the genome where known resistance-mediating mutations occur. Hence, additionally emphasizing on QC and reliability measures, PhyResSE consists of six sequential pages (Fig. 1). The resulting pages, as well as the ones belonging to the remaining steps, will open in the same tab (or window, depending on browser preferences). This means more clarity when operating with multiple samples, with each sample being displayed in a separate tab. As an alternative to the progress page links, all pages belonging to the same sample are available via the top right icon panel above the documentation link. In the following the six steps are explained in detail:

Upload. While WGS read files can be uploaded both as FastQ (.fastq) and as GNU-zipped FastQ files (.fastq.gz), the latter format is encouraged in order to save time, space and bandwidth. Paired-end reads need to be uploaded in two separate files (_R1.fastq.gz and _R2.fastq.gz), single-end read files are required to end with _R0.fastq.gz. Preceding sample names can be arbitrary. It is, however, recommended not to use special characters and to replace comprised space characters with underscores. The maximum file size for uploads was set to 2 GB (extensively tested). Multiple file selection is supported. Users of Chrome, Firefox or Safari may drag and drop files to the webpage from their desktop. Alternatively, pressing a button "Add files" will produce a multiple file selection menu. Subject to later changes adapting to high traffic, the number of files uploaded sequentially or simultaneously by pressing “Start upload’’ is currently unlimited and files and results are kept for one month. Each file is checked for proper FastQ format (taking extra time after the progress bar reaches the right end). Also, one thousand randomly chosen reads are blasted against the reference sequence *M. tuberculosis* H37Rv. The system will accept samples with down to 50/1000 full-length perfect hits as may be the case for heavily contaminated MGIT cultures. However, it will reject completely unrelated or improperly formatted data files with an appropriate error message.

Data processing takes several minutes to few days depending on size and number of uploaded files as well as on the present workload of the system. While it is possible to inspect the first results after a few minutes, users can also choose to wait e.g. overnight until everything is completed, since no further intervention is required. As the upload of a considerable number of files may take considerable time in and of itself, processing all uploaded files can be invoked not only manually via a link “Process files’’, but also “automatically upon upload’’ by checking the respective box. The “Process files” link produces a progress report displaying five icons per file resembling the following five steps or result pages described below. Each icon switches color from gray to red and black, as soon as its computation is complete, now linking to a result page.

QC. The first result page displays the quality of the raw reads. Contradicting our expectations, test results can considerably differ between left and right sets of paired-end reads (data not shown) and are thus displayed in separate frames. Handling tips describing e. g. how to export the pure numbers for further processing in a spreadsheet are linked, here and in all further pages, by a gray question mark pointing at the respective part of documentation connected to the current subject.

FastQC [4] accounts for a compendium of systematic errors occurring in NGS data, independent from the experimental platform. As such, it often appears very strict. Many samples e.g. fail both the per base sequence content criteria and the per base CG content test due to considerable GC bias. Although *Mycobacterium tuberculosis* obviously comprises 65.6% GC, there are also (few) samples passing the per base GC content test. Furthermore, we observed the same problem with other organisms such as higher plants (e.g. *Arabidopis thaliana*, 36% GC). It may vanish only when PCR amplification becomes oblivious with 3rd generation sequencing. Another PCR-induced problem is frequently occurring duplicates. In some cases, high numbers of duplicates considerably diminish the sample complexity and thus the yielded genome coverage. However, duplicates can be avoided by emulsion PCR [10, 11]. A third common problem are low base qualities, which, if occurring in high frequencies, can lead to an increased FDR.

Mapping. Reads are mapped to the *M. tuberculosis* H37Rv genome (GenBank accession number NC_000962.3) with an algorithm of BWA called BWA MEM [5]. As one of several important benchmarks for mapping performance, the mean coverage is displayed on top of the page, when unreliably small (short of 50) in red. Even in this case, however, all remaining steps are performed in order to at least reveal all resistance-mediating mutations that may show up, nonetheless. All the remaining QualiMap [6] criteria are listed and plotted at the bottom of the page (button “STATS”). In between, the page shows all the reads at a certain position, color-coded by mapping quality (top) and base quality (bottom). Thus, the mapping can also be visually inspected in detail at any position (green triangle).

To reduce false positive variant calls, the initial mapping is corrected. After removing duplicates (samtools) [7], the reads are re-aligned around indels. This is necessary because gaps are penalized harder than mismatches by any mapper. As a result, correct insertions or deletions might be missed and, what is worse in our context, false positive SNPs are generated. We apply GATK’s “IndelRealigner” to correct this. As an example Figure 3 shows one case in which without re-alignment a deletion is missed and instead a false positive SNP is called. Further, it is known that sequencing devices overestimate the quality of called bases [12]. Therefore, quality scores are recalibrated with another algorithm of GATK’s tool collection (BaseRecalibrator). Without recalibration of quality scores a false positive variant Rv3795_918g>A would be called in one of the Sierra Leone samples (Fig. 2). This SNP is well known to mediate Ethambutol resistance so that this sample would be wrongly classified as Ethambutol resistant.

Variants. After preprocessing, variants are called by GATK UnifiedGenotyper. For *Mycobacterium tuberculosis*, the number of variants (page top) to expect depends on the underlying genotype (Table S1). Numbers exceeding 4,000 tend to indicate technical artifacts, however. Variants are provided in standard Variant Call Format (VCF) and as an HTML table carrying additional information such as the affected gene and amino acid (aa) exchange. Typically, variant qualities of several thousand indicate reliability whereas a quality smaller than 500 indicates a problematic case. A mouse click on the position allows to visually inspect all involved reads in detail, showing whether or not the variant call can be trusted. The Point Accepted Mutation 1 (PAM1) lists the probability (multiplied with 10,000 for clarity) for the particular aa exchange to occur, given that 1% of the aa are changed (99% similarity, i.e. for very similar proteins). In practice, transitions between aa equivalent in charge and size are more likely whereas a transition to a most dissimilar aa will yield a small score or even a zero (e.g. Arg →Asp).

The table lists all detected variants, each genome position being represented by one line. If multiple alleles have been detected (type=MUL), the sample column lists all nucleotides observed, delimited by commas. For the time being, aa exchanges as well as PAM1 probabilities reflect only the main allele which is listed first. However, in case a variant position is located within intersecting coding sequences (CDS), more than one region, aa transition, and PAM 1 probability are provided, separated by semicolons. The icon next to the "Region" column header adds information about start(s) and stop(s), product(s), and type(s) such as rRNA or CDS. In column "AA Exchange", potential start codons are appended an "s" in round brackets. Mutations that create or abolish a potential start are not considered silent, even if the amino acid does not change.

In order to import the variants information into a spreadsheet program, here and in the two remaining pages, the table(s) can be exported as comma-separated values (CSV).

Genotype. SNPs discriminating the different phylogenetic lineages [13, 14] offer the opportunity to deduce the phenotype when they occur in the sample. There are variants specific for just one genotype while others characterize a strain set or family. In case data are sufficient, a decision tree largely following the procedure described in [15] results in predicting a genotype (bacterial strain, e.g. "Beijing") heading the page. Also, a maximum likelihood tree of our reference collection (www.MIRU-VNTR*plus*.org) [16] is linked, putting the sample genotype into a larger context (tree icon below the heading). All genotypical ("phylo") SNPs occurring in the sample are listed below. Here, for the subset of lineage discriminating variants, more information is made available than in the previous table. Starting with the leftmost column, the position links to the detailed read alignment view as before. Here, however, a mouse-over additionally produces frequencies of all four nucleotides at that position, as well as the number of reads available on the forward and on the reverse strand. Also, the already described variant quality can be amended with typical criteria a reliable mutation call has to meet, Gene ID and name link more information within Tuberculist [17], and Reference Pubmed ID links the publication, the genotypical SNP entry in our list is based on, just to name some examples. More data on the occurring entries are available by listing the "skipped columns" and inserting those of interest into the table. Since too many columns tend to make a table visually unwieldy, each one can also be removed (skipped) via a small cross in the top right of the column header. The resulting table can be exported as comma-separated values (CSV).

In case of insufficient data or samples comprising multiple strains, it becomes increasingly important not only to assess the probability of a reported SNP to be a false positive (type I error), but also the probability to overlook something (type II error). Here, beyond the average read depth (coverage) across the entire genome, we assess the positions of all genotype variants in our list (ideally being covered by more than 10 reads, of both directions).

Resistance. PhyResSE is based on an extensive summary of generally known genetic polymorphisms determining a resistance phenotype. The set of single nucleotide polymorphisms (SNPs) includes polymorphisms from our own published analyses and from the literature. Together with above “phylo“ SNPs, they are collected in one single list which is linked at the page top, here and on the previous page, as tab-delimited text as well as in VCF format. Polymorphisms which have been confirmed as mediating phenotypic drug resistance by e.g. allelic exchange experiments [18] or which are included in commercially available line probe assays, such as the Hain test (Hain, Nehren, Germany), are displayed in bold as *high confidence* SNPs (e.g. *katG*315, *rpoB*526, *rpoB*531, rightmost column). Polymorphisms not characterized *high confidence* in such a way have been correlated to drug resistance by means of e.g. resistance association studies [19, 20]. In both cases, respective source publications are referenced. In addition to known resistance-mediating variants exactly matching the sample, also other entries showing differing (non-WT) bases, however at the same position as a sample variant, can be shown via an icon next to (“show other entries“) directly below the table.

In terms of handling, the resistance page provides all features already reported for the genotype page. Publications and gene descriptions are linked, columns can be switched on and off as desired, the table can be exported as comma-separated values (CSV), and base frequencies, detailed view as well as additional quality criteria may help when in doubt about the reliability of SNP in the sample (possible type I error). Here, however, it is even more important than in the previous (genotype) context also to account for the possibility to miss a resistance-mediating SNP (type II error). For the design of appropriate treatment regimens and thus the achievement of best possibly effecting a cure, diagnoses should avoid overlooking any resistance by all means possible. To this end, we amended the lower part headed by “Did we overlook something” with a second table. As before, any known resistance-related genome position should be covered by at least 5 (better 10) reads stemming from both forward and reverse strands. Additionally, the second table lists all the genes affected by known resistance-mediating mutations in our list. Affiliated genomic regions include 300 nucleotides upstream of the gene start in order to account for possible yet unknown mutations affecting the promoter. Since unknown mutations may affect expression or gene product of such a resistance gene to the same effect as already known mutations, coverage of their genomic regions is closely monitored. All positions of either poor or excessively large coverage (as may hint at a PCR artifact) are provided as links to the detailed alignment view (artifact becomes visible by largely identical read starts and ends). In case of multiple consecutive problematic positions, intervals are denoted by start and end position, delimited by a colon. In case of any problematic position that also fails visual inspection, we strongly recommend another sequencing run of the sample library. The reads of new runs may be added to the existing ones in case of insufficient coverage until the entire regions of all resistance genes are sufficiently covered. Sufficient data can be recognized at first glance by a lack of (blue-colored) links in the left part of the table. Frequently, however, the rightmost column shows a limited number of additional (i.e. unknown) variants requiring close inspection. While upstream variants are printed spaced and in italics, indels are enclosed in brackets. Point mutations (normal print) are listed, when resulting in a different aa or a potentially differential start site, and not already being known as a genotype variant. All genotype variants have been observed in sensitive (i.e. antibiotic susceptible) strains, but for the exception of *M. bovis*. Because of its additional resistance entry, it will be displayed (correctly) as mediating pyrazinamide resistance.

Table S1: SNP counts per genotype. Number of variants one can expect per genotype. Corresponds to the similarity with H37Rv laboratory strain. Out-group M. canettii.

Genotype	**SNP counts**
M. tuberculosis Beijing	1741
M. bovis Bovis	2699,5
M. tuberculosis Cameroon	1061
M. canettii Canettii	21169
M. caprae Caprae	2864
M. tuberculosis Delhi/CAS	1789
M. tuberculosis EAI	2562
M. tuberculosis Ghana	1310
M. tuberculosis Haarlem	1251
M. tuberculosis LAM	1127
M. microti Microti	2382
M. tuberculosis New-1	1081
M. pinnipedii Pinnipedii	2486
M. tuberculosis S-type	1070
M. tuberculosis Tur	1331
M. tuberculosis Uganda	1007,5
M. tuberculosis Ural	1397
M. africanum West African 1a	2471
M. africanum West African 1b	2471
M. africanum West African 2	2593
M. tuberculosis X-type	1224

References

1. **jQuery**. .

2. **FastQValidator - Genome Analysis Wiki** [http://genome.sph.umich.edu/wiki/FastQValidator]

3. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants**. *Nucleic Acids Res* 2010, **38**:1767–1771.

4. **FastQC A Quality Control tool for High Throughput Sequence Data** [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/]

5. Li H: *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. arXiv e-print*; 2013.

6. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A: **Qualimap: evaluating next-generation sequencing alignment data**. *Bioinforma Oxf Engl* 2012, **28**:2678–2679.

7. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**:2078–2079.

8. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. *Nat Genet* 2011, **43**:491–498.

9. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res* 2010, **20**:1297–1303.

10. Schütze T, Rubelt F, Repkow J, Greiner N, Erdmann VA, Lehrach H, Konthur Z, Glökler J: **A streamlined protocol for emulsion polymerase chain reaction and subsequent purification**. *Anal Biochem* 2011, **410**:155–157.

11. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P: **Library construction for next-generation sequencing: overviews and challenges**. *BioTechniques* 2014, **56**:61–64, 66, 68, passim.

12. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J: **SNP detection for massively parallel whole-genome resequencing**. *Genome Res* 2009, **19**:1124–1132.

13. Homolka S, Projahn M, Feuerriegel S, Ubben T, Diel R, Nübel U, Niemann S: **High Resolution Discrimination of Clinical Mycobacterium tuberculosis Complex Strains Based on Single Nucleotide Polymorphisms**. *PLoS ONE* 2012, **7**:e39855.

14. Feuerriegel S, Köser CU, Niemann S: **Phylogenetic polymorphisms in antibiotic resistance genes of the Mycobacterium tuberculosis complex**. *J Antimicrob Chemother* 2014, **69**:1205–1210.

15. Homolka S, Projahn M, Feuerriegel S, Ubben T, Diel R, Nübel U, Niemann S: **High Resolution Discrimination of Clinical Mycobacterium tuberculosis Complex Strains Based on Single Nucleotide Polymorphisms**. *PLoS ONE* 2012, **7**:e39855.

16. Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D: **MIRU-VNTRplus: a web tool for polyphasic genotyping of Mycobacterium tuberculosis complex bacteria**. *Nucleic Acids Res* 2010, **38**(suppl 2):W326–W331.

17. Lew JM, Kapopoulou A, Jones LM, Cole ST: **TubercuList--10 years after**. *Tuberc Edinb Scotl* 2011, **91**:1–7.

18. Nebenzahl-Guimaraes H, Borgdorff MW, Murray MB, van Soolingen D: **A Novel Approach - The Propensity to Propagate (PTP) Method for Controlling for Host Factors in Studying the Transmission of Mycobacterium Tuberculosis**. *PLoS ONE* 2014, **9**:e97816.

19. Von Groll A, Martin A, Jureen P, Hoffner S, Vandamme P, Portaels F, Palomino JC, da Silva PA: **Fluoroquinolone resistance in Mycobacterium tuberculosis and mutations in gyrA and gyrB**. *Antimicrob Agents Chemother* 2009, **53**:4498–4500.

20. Feuerriegel S, Cox HS, Zarkua N, Karimovich HA, Braker K, Rusch-Gerdes S, Niemann S: **Sequence Analyses of Just Four Genes To Detect Extensively Drug-Resistant Mycobacterium tuberculosis Strains in Multidrug-Resistant Tuberculosis Patients Undergoing Treatment**. *Antimicrob Agents Chemother* 2009, **53**:3353–3356.