JCM
Journals.ASM.org

PhyResSE: a Web Tool Delineating *Mycobacterium tuberculosis* Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data

Silke Feuerriegel,[a,b] Viola Schleusener,[c] Patrick Beckert,[a,b] Thomas A. Kohl,[a] Paolo Miotto,[d] Daniela M. Cirillo,[d] Andrea M. Cabibbe,[d] Stefan Niemann,[a,b] Kurt Fellenberg[c]

Molecular Mycobacteriology, Research Center Borstel, Borstel, Germany[a]; German Center for Infection Research (DZIF), Borstel Site, Borstel, Germany[b]; Bioinformatics, Research Center Borstel, Borstel, Germany[c]; Emerging Bacterial Pathogens Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Scientific Institute, Milan, Italy[d]

Antibiotic-resistant tuberculosis poses a global threat, causing the deaths of hundreds of thousands of people annually. While whole-genome sequencing (WGS), with its unprecedented level of detail, promises to play an increasingly important role in diagnosis, data analysis is a daunting challenge. Here, we present a simple-to-use web service (free for academic use at http://phyresse.org). Delineating both lineage and resistance, it provides state-of-the-art methodology to life scientists and physicians untrained in bioinformatics. It combines elaborate data processing and quality control, as befits human diagnostics, with a treasure trove of validated resistance data collected from well-characterized samples in-house and worldwide.

An estimated 9 million people developed tuberculosis (TB) worldwide in 2013. During the same period, 1.5 million died from the disease (1). Effective treatment and control of TB are complicated by the emergence and spread of drug-resistant, multidrug-resistant (MDR), or even extensively drug-resistant (XDR) strains (2). MDR *Mycobacterium tuberculosis* complex (MTBC) strains are resistant to at least isoniazid (INH) and rifampin (RIF); XDR strains carry additional resistance to at least one fluoroquinolone and one injectable drug (3). The treatment of these strains is expensive and often ineffective, which not only facilitates the spread of resistant strains but also enables the development of resistance to additional drugs (4, 5). Key to the control of this public health threat would be the timely diagnosis of drug resistance. Yet, owing to the low growth rate of MTBC strains, even the fastest conventional method of phenotypic drug susceptibility testing (DST) with the Bactec mycobacterial growth indicator tube (MGIT) 960 system takes at least 7 days after a positive culture has been obtained. Further, it shows poor reproducibility for some drugs, such as ethambutol (EMB) and pyrazinamide (PZA) (6). Despite being faster than phenotypic DST, the current generation of genotypic DST assays targets only the most frequent resistance-mediating mutations for a limited number of first- and second-line drugs (7, 8). Consequently, they have played only a limited role clinically. In contrast, rapid whole-genome sequencing (WGS) by next-generation sequencing (NGS) approaches can be used to detect known molecular markers associated with resistance to all drugs simultaneously directly from a primary culture (9). Moreover, WGS provides the ultimate molecular resolution to guide contact tracing (10, 11). In light of the ongoing improvements in ease of use, turnaround time, and cost, WGS is therefore destined to become routine in well-resourced countries for every case of culture-positive TB (12, 13).

This paradigm shift can occur only if the data analysis is fully automated (14, 15). Broadly speaking, data analysis consists of two parts. First, NGS data have to be processed and variants (mainly single-nucleotide polymorphisms [SNPs]) need to be determined relative to a reference genome. Second, the variants in question need to be interpreted as to whether they probably confer antibiotic resistance or not. A variety of advanced and reasonably well-tested software tools exist for the former task (16–20); however, they require bioinformatic analysis skills and hardware infrastructure.

To address these challenges, we developed an easy-to-use, web-based tool called the Phylo-Resistance Search Engine (PhyResSE). In contrast to other tools, such as MuBII (21) and KvarQ (22), it performs rigorous preprocessing of both raw FastQ files and mapping results. In comparison to ANNOVAR (23), PhyResSE is tailor made for MTBC antibiotic resistance diagnosis, linking the raw data processing and any problems becoming visible in this process with in-depth quality control (QC) of the results. We provide our Software as a Service (SaaS) freely available without registration at http://phyresse.org. Thus, relieving users of any local installation and system administration, we aim to make NGS-based resistance determination available to a larger community.

MATERIALS AND METHODS

Experimental and computational methods. (i) Strain collection. The WGS data used to evaluate our tool consisted of 92 strains from Sierra Leone. The DST results, Sanger sequencing data for selected drug resis-

Received 7 January 2015 **Returned for modification** 9 February 2015 **Accepted** 31 March 2015

Accepted manuscript posted online 8 April 2015

Citation Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, Cabibbe AM, Niemann S, Fellenberg K. 2015. PhyResSE: a Web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. J Clin Microbiol 53:1908–1914. doi:10.1128/JCM.00025-15.

Editor: K. C. Carroll

Address correspondence to Stefan Niemann, sniemann@fz-borstel.de, or Kurt Fellenberg, kfellenberg@fz-borstel.de.

S.F., V.S., and P.B. contributed equally to this work.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/JCM.00025-15.

doi:10.1128/JCM.00025-15

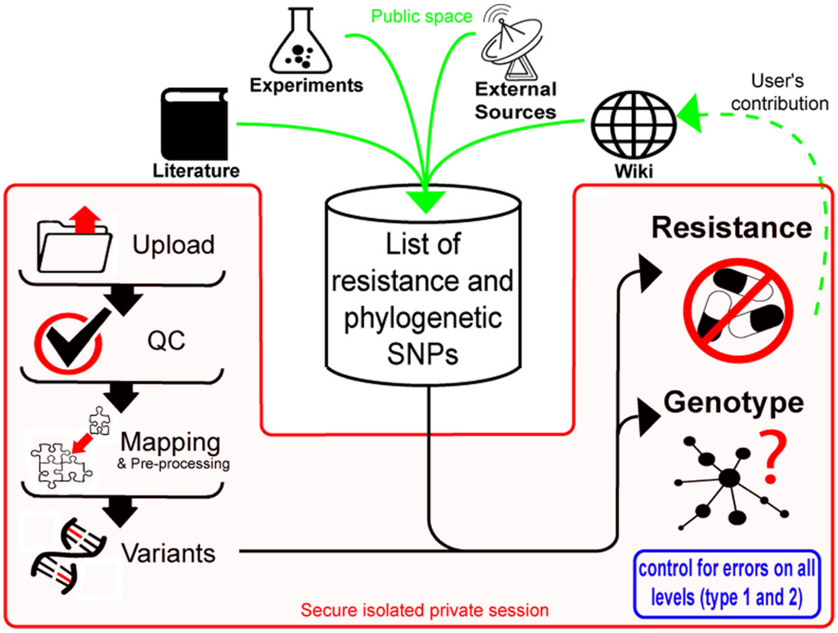


FIG 1 PhyResSE workflow. Sequential steps accessible to the users of one specific session are depicted in red, and input to the list of known mutations is in green. The list consists of common resistance and phylogenetic SNPs extracted from the literature. Users can contribute by adding new SNPs to the wiki.

tance-mediating genes (*katG*, *rpoB*, *embB*, *rrs*, *rpsL*, *gidB*, and *pncA* [for some strains also *inhA*, *ahpC*, *embA*, and *embC*]), as well as conventional typing data for this collection have been described previously (7, 24). Whenever differences between Sanger sequencing data and PhyResSE results were observed, Sanger sequencing was repeated.

(ii) NGS sequencing. DNA was extracted from cultured MTBC isolates as previously described (25). DNA was prepared for sequencing with the Illumina Nextera XT kit and sequenced on an Illumina MiSeq (251 and 301 bp, paired end) according to the manufacturer's instructions. Each strain was sequenced to obtain 99% coverage relative to the *M. tuberculosis* H37Rv reference genome, with an average depth of at least 50×.

(iii) Analysis pipeline. Data access is restricted via private sessions, which are implemented with the CGI-Session Perl Module. For stable and accurate uploading of files, jQuery is used (26). Before files are stored, the format is checked with FastQValidator (27) and a limited number of reads (1,000) are blasted against the reference genome. The system will accept samples with as few as 50 full-length perfect hits to the reference sequence (*M. tuberculosis* H37Rv, GenBank accession no. NC_000962.3). To start the analysis, all reads of one isolate are mapped to that reference with BWA-MEM (17). The quality of the FastQ and resulting BAM file is checked by FastQC (28) and Qualimap (18). The number of false-positive variant calls is decreased by preprocessing the BAM file. Duplicates are removed by SAMtools (16). Base quality score recalibration and realignment around small insertions or deletions (indels, 1 to 30 bp) are performed with the Genome Analysis Toolkit (GATK) (19, 29). Another GATK function is used to call both SNPs and small indels (UnifiedGenotyper). The quality of called variants is recalibrated by VariantQualityScoreRecalibrator (from GATK), and potentially incorrect calls are grayed out. For further details concerning the tools used, see Document S1 in the supplemental material.

Nucleotide sequence accession numbers. All NGS data have been submitted to the EMBL-EBI ENA sequence read archive (accession no. PRJEB7727 and ERR551412).

RESULTS

PhyResSE general handling and functionality. PhyResSE is designed to enable nonspecialized users such as microbiologists and clinicians to extract phylogenetic and resistance information from NGS data via an easy-to-use web interface. Entering the page as a public user yields a secured private session; i.e., uploaded data and results cannot be seen by any other user. Privacy is ensured by a 32-character key transferred after SSL encryption is established and stored at the client site as a cookie (cookies must be allowed). The analysis consists of five sequential steps, including QC (Fig. 1), which starts automatically after either single- or paired-end reads are uploaded in FastQ format. Multiple file selection for upload is supported. The "Process files" link leads to the progress page, where the respective icons of the five analysis steps per sample change to red once the steps are complete. The results of the earlier analysis steps can be viewed before the later steps are finished, or users can choose to wait until all samples are completed. Switching off the browser and/or the computer does not interrupt the computation of uploaded files. More details are provided in Document S1 in the supplemental material and by the gray question marks next to the reports linking to the online documentation. Files and results are kept for 1 month.

Data processing takes several minutes (for one sample) to a few days, depending on the size and number of uploaded files and the workload of the system. During the validation stage, we found that the time-consuming steps of recalibrating base quality scores and realigning around indels were essential to prevent false-positive variant calls, as demonstrated by pipelines that lacked these steps. Pipelines without realignment led to numerous positions that were incorrectly aligned, as shown in Fig. 2. Furthermore, skipping the recalibration of base quality scores resulted in erroneous SNPs being called. This even led to a false-positive EMB resistance result for one strain from our validation data set (Fig. 3).

We successfully tested PhyResSE with various web browsers, such as Firefox versions 11.0, 16.0, 23.0, and 31.0 (Mozilla Corporation, Mountain View, CA); Internet Explorer versions 9, 10, and 11 (Microsoft Corporation, Redmond, WA); Google Chrome ver-

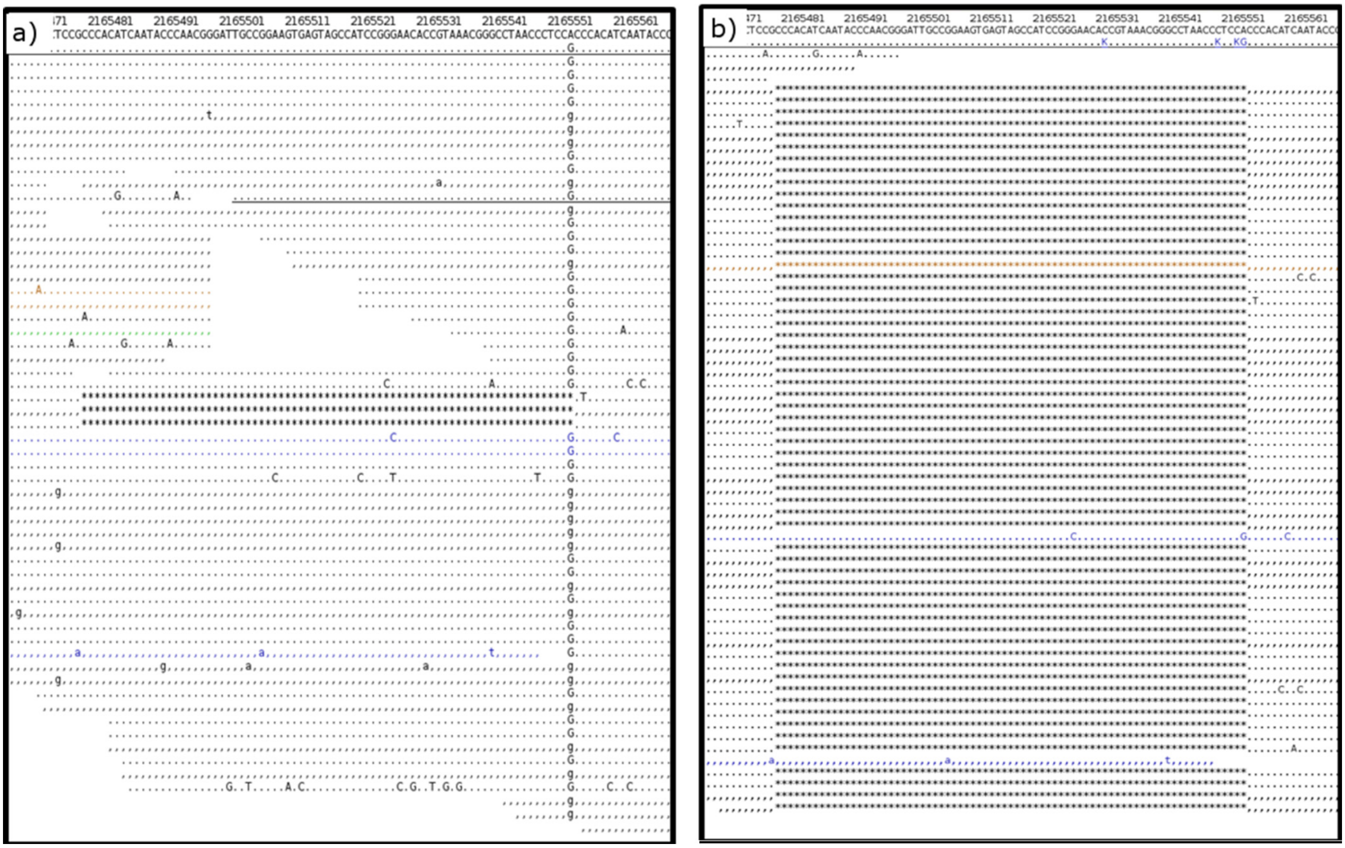

a)
71 2165481 2165491 2165501 2165511 2165521 2165531 2165541 2165551 2165561
TCCGCCCACATCAATACCCAACGGGATTGCCGGAAGTGAGTAGCCATCCGGGAACACCGTAAACGGGCCTAACCCTCCACCCACATCAATACC
b)
71 2165481 2165491 2165501 2165511 2165521 2165531 2165541 2165551 2165561
TCCGCCCACATCAATACCCAACGGGATTGCCGGAAGTGAGTAGCCATCCGGGAACACCGTAAACGGGCCTAACCCTCCACCCACATCAATACC

FIG 2 Evaluation of different pipelines. (a) Mapping result without realignment results in a false-positive SNP at position 2165551 (A→G) with the true deletion not being detected. (b) Mapping result after realignment. Correctly, no variant is called at position 2165551; instead, the true deletion is called. Reads are color coded according to mapping quality (blue, 0 to 9; green, 10 to 19; yellow, 20 to 29; black, ≥30). A dot indicates a match (identity with the reference) on the forward strand, and a comma indicates a match on the reverse strand. Mismatches are shown in capital or lowercase letters when detected on the forward or reverse strand, respectively.

sions 19.0, 29.0, and 30.0 (Google Inc., Mountain View, CA); Opera versions 16 and 17 (Opera Software ASA, Oslo, Norway); Safari 5 (Apple Inc., Cupertino, CA); and Konqueror 4 (KDE e.V., Berlin, Germany).

Mutation catalogue. PhyResSE uses a variant catalogue with well-described mutations that are known to confer antibiotic resistance. It further comprises lineage-specific mutations. This list was compiled on the basis of a review of the literature and our own laboratory data. Links to the appropriate studies are provided for each mutation in the plain-language report. This allows assessment of the evidence for the impact of each variant. To simplify this process, variants for which strong experimental evidence is available (e.g., from allelic-exchange experiments [30]) are highlighted in bold as high-confidence SNPs (e.g., *katG315*, *rpoB445*, *rpoB450*). The nomenclature used in the plain-language report is based on the *M. tuberculosis* H37Rv reference genome. In addition, the *Escherichia coli* nomenclature is provided for mutations in *rpoB*.

Given that homoplasies are rare in MTBC, strains can be classified by using phylogenetically informative SNPs as described earlier (31, 32). Here, phylogenetic classification is based on the decision tree described previously (31). Specifically, PhyResSE relies on a combination of SNPs that are markers for specific genotypes, as well as mutations that are more deeply rooted in the MTBC diversity and therefore characterize a larger set of strains (e.g., the Euro-American superlineage).

Because of the flexible data format, the variant list used can be extended to incorporate novel findings, including additional resistance mechanisms for drugs that are currently being evaluated in clinical trials.

Validation. PhyResSE was tested with 92 strains from a well-characterized strain collection from Sierra Leone that comprised 44 phenotypically susceptible strains and 48 strains that are either monoresistant (RIF, INH, or streptomycin [SM]) or polyresistant (RIF, INH, SM, EMB, and/or PZA) (7, 24). For the full data set, comprising DST, Sanger sequencing, and WGS data, as well as phylogenetic strain classification obtained by IS*6110* DNA fingerprinting, spoligotyping, and 24-locus mycobacterial interspersed repetitive-unit–variable-number tandem-repeat (MIRU-VNTR) typing, see Table S2 in the supplemental material. We observed 100% concordance for resistance SNPs in *katG*, *inhA*, *ahpC*, *rrs*, *rpsL*, *embA*, and *embC*; 98.91% concordance for those in *gidB* and *pncA*; and 97.83% concordance for those in *rpoB* and *embB* (Table 1). Discrepancies observed between Sanger sequencing and WGS data fall into only two categories. Five discrepancies are due to heteroresistant samples (a mixture of wild-type and mutant alleles) that were not detected by Sanger sequencing. For example, one isolate harbored an *rpoB* resistance-mediating mutation (CCG [Pro] at codon 452) in 19% of the total population, as deduced from our WGS data, which is below the limit of detection by Sanger sequencing (Fig. 4) (33). One discordant SNP is covered

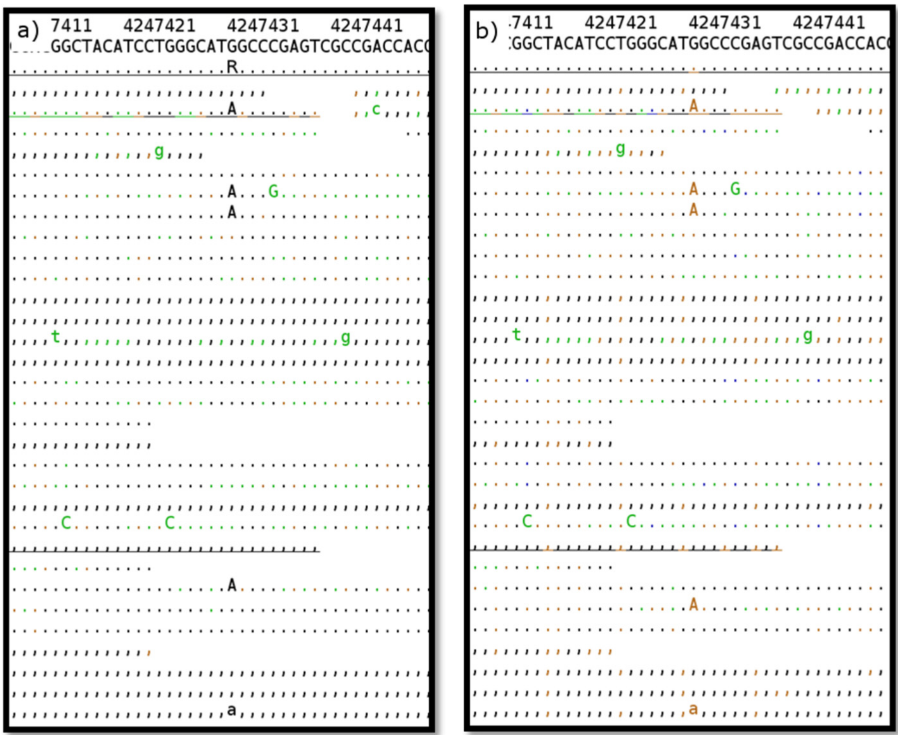

a)
7411 4247421 4247431 4247441
GGCTACATCCTGGGCATGGCCCGAGTCGCCGACCAC
b)
7411 4247421 4247431 4247441
GGCTACATCCTGGGCATGGCCCGAGTCGCCGACCAC

FIG 3 Evaluation of different pipelines. (a) Mapping result without base quality score recalibration. A false-positive variant at position 4247431 (G→A) leads to incorrect classification of the isolate as EMB resistant. (b) Mapping result after recalibration of base quality scores. Correctly, no variant is called at position 4247431. The layout corresponds to that of Fig. 2. Unlike in Fig. 2, however, the color code refers to base quality.

by the reverse primer used for Sanger sequencing, which leads to a wild-type call (data not shown).

The lineages detected by PhyResSE were concordant with the results determined by conventional genotyping for all isolates belonging to the previously well-defined *M. tuberculosis* Beijing, Cameroon, East African Indian, Ghana, Haarlem, Latin American Mediterranean, S-type, and *M. africanum* West African 1 and 2 lineages (Table 2). Isolates that belonged to the Sierra Leone 1 and 2 or X-type lineages were not differentiated from the Euro-American superlineage as a whole because the classification system used by PhyResSE did not include markers for these lineages. Conversely, it was possible to assign one previously undefined isolate to the Haarlem lineage and 10 to the Euro-American superlineage, respectively (Table 2).

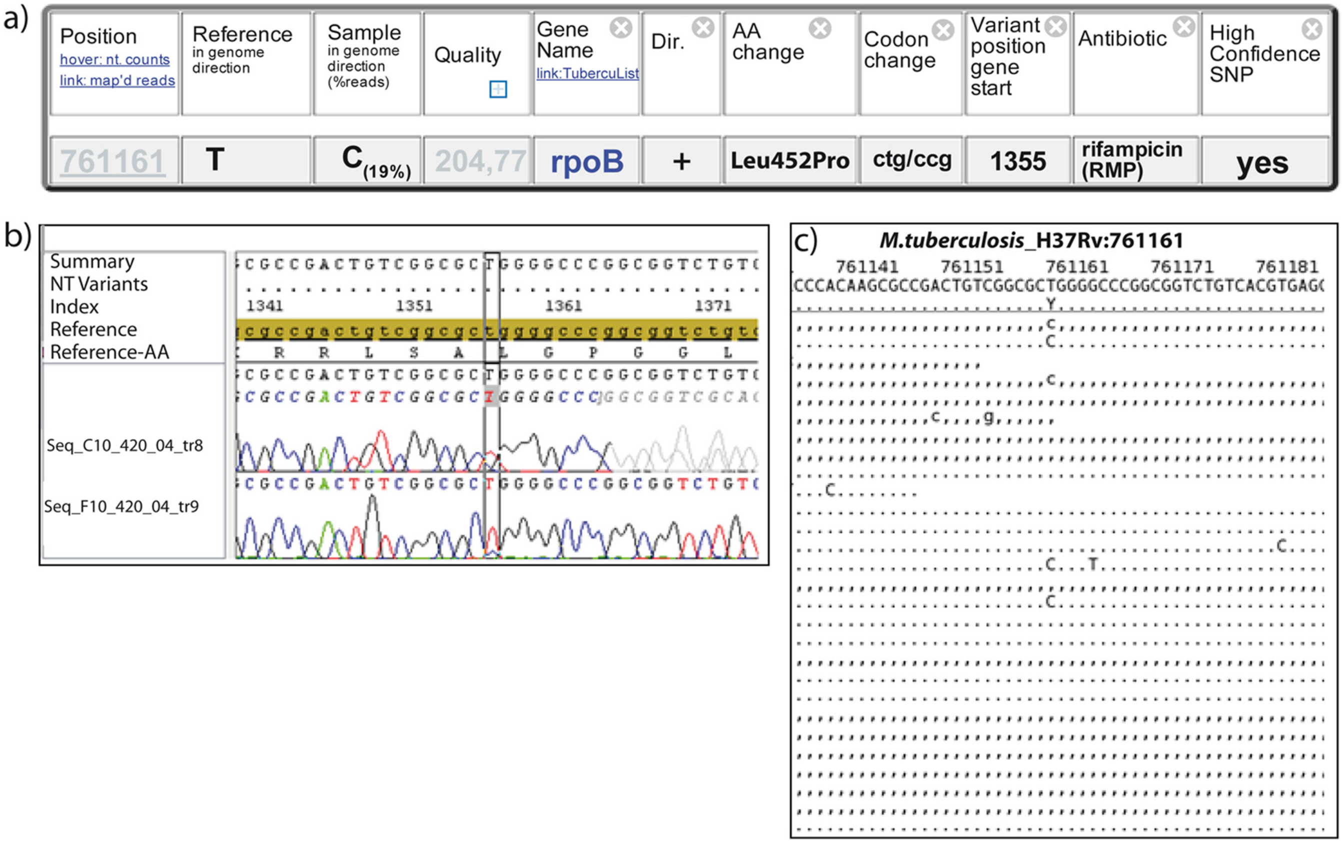
DISCUSSION

PhyResSE is the first web interface that enables the automated interpretation of MTBC WGS data for the identification of resis-

TABLE 1 Evaluation of PhyResSE with respect to resistance-mediating mutations[a]

Resistance-mediating gene sequenced	No. of SNPs detected by Sanger sequencing (total no. of strain sequences)	No. of SNPs detected by WGS/ interpreted by PhyResSE (total no. of strain sequences)	No. of discordant SNPs	No. of discordant strains	% Concordance[b]
rpoB	20 (92)	22 (92)	2	2	97.83[c]
katG	27 (92)	27 (92)	0	0	100
inhA	2 (3)	4 (92)	2	0[d]	100
ahpC	1 (3)	1 (92)	0	0	100
rrs	2 (92)	2 (92)	0	0	100
rpsL	20 (92)	20 (92)	0	0	100
gidB	72 (92)	73 (92)	1	1	98.91[c]
embB	16 (92)	18 (92)	2	2	97.83[c]
embA	0 (4)	1 (92)	1	0[d]	100
embC	1 (4)	1 (92)	0	0	100
pncA	11 (92)	12 (92)	1	1	97.83[e]

[a] Results obtained by Sanger sequencing and WGS data interpreted by PhyResSE are compared, and numbers of identical and differing results are listed.
[b] Related to the total number of strain sequences.
[c] Heteroresistant.
[d] Not Sanger sequenced.
[e] Covered by reverse primer.



Position hover: nt. counts link: map'd reads	Reference in genome direction	Sample in genome direction (%reads)	Quality	Gene Name link:TubercuList	Dir.	AA change	Codon change	Variant position gene start	Antibiotic	High Confidence SNP
761161	T	$C_{(19\%)}$	204,77	rpoB	+	Leu452Pro	ctg/ccg	1355	rifampicin (RMP)	yes

FIG 4 Detection of heteroresistant samples with PhyResSE. (a) Developing RIF resistance. Only 19% of the reads show CCG (Pro) instead of the wild-type codon CTG (Leu). This reflects the small proportion of bacteria already carrying RIF resistance at the time of sample preparation. Dir., direction; AA, amino acid. (b) Electropherogram of Sanger sequencing at position 761161. In the reverse sequence of 420/04 (Tr8), no additional peak is visible. In the forward sequence (Tr9), a small C peak is visible. NT, nucleotide. (c) Mapping result at position 761161. Approximately 19% of the reads show a C instead of the wild-type base T (independent of the strand). A representative subset of 26 out of 102 reads in total is shown for clarity.

tance-mediating variants and phylogenetic lineage classification. The data obtained in our study confirmed that NGS results are superior to classical Sanger sequencing and conventional typing methods, e.g., by detecting heterogeneous SNPs and improved phylogenetic lineage classification. Rigorous preprocessing and QC at the levels of raw data, mapping performance, and individual SNPs are unique characteristics of our system. While the system would be considerably faster if time-consuming preprocessing steps were omitted, we selected this more conservative analysis pipeline for its minimum level of false-positive variants. A pipeline lacking recalibration of base quality scores resulted in incorrect classification of EMB resistance in one strain. For the complete pipeline finally implemented in PhyResSE, however, results obtained were highly accurate and concordant with conventional genotyping and Sanger sequencing of resistance genes.

TABLE 2 Evaluation of PhyResSE regarding lineage-defining mutations

No. of strains	Genotype determined: Conventionally[a]	By WGS
4	Beijing	Beijing
4	Cameroon	Cameroon
4	EAI[b]	EAI
1	Ghana	Ghana
12	Haarlem	Haarlem
15	LAM[c]	LAM
3	S type	S type
6	West African 1	West African 1
14	West African 2	West African 2
1	None	Haarlem
10	None	Euro-American superlineage
7	Sierra Leone 1	Euro-American superlineage
9	Sierra Leone 2	Euro-American superlineage
2	X type	Euro-American superlineage

[a] By IS*6110* fingerprinting, spoligotyping, and 24-locus MIRU-VNTR typing.
[b] EAI, East African Indian.
[c] LAM, Latin American Mediterranean.

Moreover, in comparison with other tools allowing the fast detection of antibiotic resistance on the basis of NGS data (e.g., KvarQ [22]), PhyResSE provides a more complete analysis of resistance targets also including longer genes such as *pncA* or the *embABC* operon that have been excluded by KvarQ because of its specific analysis procedure (22). PhyResSE users are also provided more detailed information, e.g., on the variants by viewing all mapped reads at a critical position. It can already handle also heterogeneous SNP calls, as perfectly shown in our reference data set.

However, new data might comprise resistance-mediating SNPs not yet described. To account for variants and also algorithms to be added in the future, both are organized in standardized metadata files. Users can even work with private variant lists (superseding the one we provide and only visible within their session) amended with their own SNPs by using any spreadsheet program such as Microsoft Excel. In the same way, switching of the mapper or any other preprocessing algorithm can be achieved by editing one line in a metadata file that includes all of the parameters for all of the computational steps involved. This grants the flexibility to quickly adapt our system to new developments in the field. For obvious reasons (the new mapper would need to be installed on the machine to run), the latter is restricted to system administrators. However, the documentation accesses the same metadata file that also governs the processing, informing users exactly what is being done. To document any amendments and alterations, PhyResSE development versions are named. Also for the variant list (comprising both the genotype-discriminant and the antibiotic resistance-related mutations), versions are recorded by a release list, enabling full tracing back of the analysis flow that has yielded a certain output.

We thus aim to make our workflow both adequate and transparent. A third prerequisite for successful NGS workflows is to apply sufficient underlying computational resources. We made available 40 cores, half a terabyte of RAM, and 50 terabytes of storage capacity that will be further expanded according to system usage. In order to avoid losing any resources to processing of inadequate data, the system restrictively tests the integrity of the uploaded GNU zip archive, proper FastQ format, and the proportion of MTBC-specific reads. One thousand randomly chosen reads are blasted against the *M. tuberculosis* H37Rv reference sequence. While reads obtained by NGS of DNA isolated from pure cultures of MTBC strains generally show 98% perfect full-length hits to the *M. tuberculosis* H37Rv reference sequence, samples from MGIT cultures of direct specimens can contain larger fractions of human or other bacterial contaminants (data not shown). To allow analysis also of these data sets, we set the threshold to 50 out of 1,000 full-length perfect hits to the H37Rv reference genome.

Apart from problems that might arise from excessive contamination, the validity of the NGS data depends on numerous other factors, such as proper genomic library design and the use of high-fidelity DNA polymerases, to name just two examples. PhyResSE provides extensive QC and highlights systematic errors where possible. However, the user is responsible for inspecting these and for taking them into account for proper data interpretation. It should be noted that final diagnostic decisions in clinical microbiology cannot be provided by even high-quality software, as these require the integration of genetic, microbiological, and clinical data by specialized personnel for the appropriate management of patients. Also, the use of PhyResSE for diagnostic purposes and the clinical management of patients is still limited by an insufficient understanding of the role of some SNPs in determining drug resistance and/or an unfavorable clinical outcome. However, the increasing use of NGS in association with computational tools such as PhyResSE will help us to gain crucial knowledge in this field.

To support this process, a computational tool should be simple to use. As a preinstalled web-based service, PhyResSE does not require any installation procedure at the user site. NGS software often needs high-performance computers, as well as staff skilled in installing and operating bioinformatic analysis tools. Being accessible via the Internet (SaaS), a running network connection remains the only requirement. If the user desires it, processing starts automatically after upload, enabling the uploading and analysis of hundreds of files even via slow network connections in one go. After the files to be uploaded are selected, one mouse click invokes the entire workflow, with no further action being required of the user. Afterwards, results are available as plain-language (HTML) reports in which we invested a considerable share of development time to produce a clear, self-explanatory, and simple layout.

Thus, intended to make NGS-based resistance diagnosis available to a larger community, PhyResSE opens the way for a wider application of WGS in the mycobacteriological laboratory for day-to-day use. In addition to the wider availability of NGS that is facilitated by smaller benchtop systems that can be integrated into a normal laboratory workflow, it represents the crucial analysis platform development needed for large-scale routine NGS application.

ACKNOWLEDGMENTS

We thank I. Razio, J. Zallet, T. Struve-Sonnenschein, A. Luedemann, and T. Ubben of Research Center Borstel, Borstel, Germany, for technical assistance. We thank Claudio Köser of the Department of Medicine, University of Cambridge, Cambridge, United Kingdom, for very valuable discussions and the opportunity to test MGIT samples.

Parts of this work were supported by the European Union TB-PAN-NET (FP7-223681) and PathoNgenTrace (FP7-278864-2) projects and the German Center for Infection Research (DZIF). The funders had no role in study design, data collection and analysis, the decision to publish, or preparation of the manuscript.

REFERENCES

1. **World Health Organization.** 2014. Global tuberculosis report 2014. World Health Organization, Geneva, Switzerland.
2. **Shah NS, Wright A, Bai G-H, Barrera L, Boulahbal F, Martin-Casabona N, Drobniewski F, Gilpin C, Havelkova M, Lepe R, Lumb R, Metchock B, Portaels F, Rodrigues MF, Rusch-Gerdes S, Van Deun A, Vincent V, Laserson K, Wells C, Cegielski JP.** 2007. Worldwide emergence of extensively drug-resistant tuberculosis. Emerg Infect Dis **13:**380–387. http://dx.doi.org/10.3201/eid1303.061400.
3. **Chang K-C, Yew W-W.** 2013. Management of difficult multidrug-resistant tuberculosis and extensively drug-resistant tuberculosis: update 2012. Respirology **18**(1):8–21. http://dx.doi.org/10.1111/j.1440-1843.2012.02257.x.
4. **Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, Corander J, Bryant J, Parkhill J, Nejentsev S, Horstmann RD, Brown T, Drobniewski F.** 2014. Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet **46:**279–286. http://dx.doi.org/10.1038/ng.2878.
5. **Cox HS, Sibilia C, Feuerriegel S, Kalon S, Polonsky J, Khamraev AK, Rüsch-Gerdes S, Mills C, Niemann S.** 2008. Emergence of extensive drug resistance during treatment for multidrug-resistant tuberculosis. N Engl J Med **359:**2398–2400. http://dx.doi.org/10.1056/NEJMc0805644.
6. **Piersimoni C, Olivieri A, Benacchio L, Scarparo C.** 2006. Current perspectives on drug susceptibility testing of *Mycobacterium tuberculosis* complex: the automated nonradiometric systems. J Clin Microbiol **44:**20–28. http://dx.doi.org/10.1128/JCM.44.1.20-28.2006.
7. **Feuerriegel S, Oberhauser B, George AG, Dafae F, Richter E, Rüsch-Gerdes S, Niemann S.** 2012. Sequence analysis for detection of first-line drug resistance in Mycobacterium tuberculosis strains from a high-incidence setting. BMC Microbiol **12:**90. http://dx.doi.org/10.1186/1471-2180-12-90.
8. **Feuerriegel S, Cox HS, Zarkua N, Karimovich HA, Braker K, Rusch-Gerdes S, Niemann S.** 2009. Sequence analyses of just four genes to detect extensively drug-resistant *Mycobacterium tuberculosis* strains in multidrug-resistant tuberculosis patients undergoing treatment. Antimicrob Agents Chemother **53:**3353–3356. http://dx.doi.org/10.1128/AAC.00050-09.
9. **Köser CU, Bryant JM, Becq J, Török ME, Ellington MJ, Marti-Renom MA, Carmichael AJ, Parkhill J, Smith GP, Peacock SJ.** 2013. Whole-genome sequencing for rapid susceptibility testing of M. tuberculosis. N Engl J Med **369:**290–292. http://dx.doi.org/10.1056/NEJMc1215305.

10. **Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE.** 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis **13:**137–146. http://dx.doi.org/10.1016/S1473-3099(12)70277-3.
11. **Merker M, Kohl TA, Roetzer A, Truebe L, Richter E, Rüsch-Gerdes S, Fattorini L, Oggioni MR, Cox H, Varaine F, Niemann S.** 2013. Whole genome sequencing reveals complex evolution patterns of multidrug-resistant Mycobacterium tuberculosis Beijing strains in patients. PLoS One **8:**e82551. http://dx.doi.org/10.1371/journal.pone.0082551.
12. **Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J, Peacock SJ.** 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog **8:**e1002824. http://dx.doi.org/10.1371/journal.ppat.1002824.
13. **Köser CU, Ellington MJ, Peacock SJ.** 2014. Whole-genome sequencing to control antimicrobial resistance. Trends Genet **30:**401–407. http://dx.doi.org/10.1016/j.tig.2014.07.003.
14. **Wyres K, Conway T, Garg S, Queiroz C, Reumann M, Holt K, Rusu L.** 2014. WGS analysis and interpretation in clinical and public health microbiology laboratories: what are the requirements and how do existing tools compare? Pathogens **3:**437–458. http://dx.doi.org/10.3390/pathogens3020437.
15. **Scholz MB, Lo C-C, Chain PS.** 2012. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Curr Opin Biotechnol **23:**9–15. http://dx.doi.org/10.1016/j.copbio.2011.11.013.
16. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The sequence alignment/map format and SAMtools. Bioinformatics **25:**2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352.
17. **Li H.** 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN]. http://arxiv.org/abs/1303.3997.
18. **García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A.** 2012. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics **28:**2678–2679. http://dx.doi.org/10.1093/bioinformatics/bts503.
19. **McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA.** 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res **20:**1297–1303. http://dx.doi.org/10.1101/gr.107524.110.
20. **Altmann A, Weber P, Bader D, Preuß M, Binder EB, Müller-Myhsok B.** 2012. A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet **131:**1541–1554. http://dx.doi.org/10.1007/s00439-012-1213-z.
21. **Flandrois J-P, Lina G, Dumitrescu O.** 2014. MUBII-TB-DB: a database of mutations associated with antibiotic resistance in Mycobacterium tuberculosis. BMC Bioinformatics **15:**107. http://dx.doi.org/10.1186/1471-2105-15-107.
22. **Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S.** 2014. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. BMC Genomics **15:**881. http://dx.doi.org/10.1186/1471-2164-15-881.
23. **Wang K, Li M, Hakonarson H.** 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res **38:**e164. http://dx.doi.org/10.1093/nar/gkq603.
24. **Homolka S, Post E, Oberhauser B, George AG, Westman L, Dafae F, Rüsch-Gerdes S, Niemann S.** 2008. High genetic diversity among Mycobacterium tuberculosis complex strains from Sierra Leone. BMC Microbiol **8:**103. http://dx.doi.org/10.1186/1471-2180-8-103.
25. **van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD.** 1991. Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. J Clin Microbiol **29:**2578–2586.
26. **Resig J.** 2014. jQuery. http://jquery.com/.
27. **Anonymous.** 2015. Statgen/fastQValidator. https://github.com/statgen/fastQValidator.
28. **Andrews S.** FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
29. **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.** 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet **43:**491–498. http://dx.doi.org/10.1038/ng.806.
30. **Nebenzahl-Guimaraes H, Borgdorff MW, Murray MB, van Soolingen D.** 2014. A novel approach—the propensity to propagate (PTP) method for controlling for host factors in studying the transmission of Mycobacterium tuberculosis. PLoS One **9:**e97816. http://dx.doi.org/10.1371/journal.pone.0097816.
31. **Feuerriegel S, Köser CU, Niemann S.** 2014. Phylogenetic polymorphisms in antibiotic resistance genes of the Mycobacterium tuberculosis complex. J Antimicrob Chemother **69:**1205–1210. http://dx.doi.org/10.1093/jac/dkt535.
32. **Homolka S, Projahn M, Feuerriegel S, Ubben T, Diel R, Nübel U, Niemann S.** 2012. High resolution discrimination of clinical Mycobacterium tuberculosis complex strains based on single nucleotide polymorphisms. PLoS One **7:**e39855. http://dx.doi.org/10.1371/journal.pone.0039855.
33. **Pholwat S, Stroup S, Foongladda S, Houpt E.** 2013. Digital PCR to detect and quantify heteroresistance in drug resistant Mycobacterium tuberculosis. PLoS One **8:**e57238. http://dx.doi.org/10.1371/journal.pone.0057238.