Proteogenomic Investigation of Strain Variation in Clinical *Mycobacterium tuberculosis* Isolates

Tiaan Heunis, Anzaan Dippenaar, Robin M. Warren, Paul D. van Helden, Ruben G. van der Merwe, Nicolaas C. Gey van Pittius, Arnab Pain, Samantha L. Sampson, and David L. Tabb

J. Proteome Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.jproteome.7b00483 • Publication Date (Web): 18 Aug 2017

Downloaded from <http://pubs.acs.org> on August 22, 2017

Just Accepted

"Just Accepted" manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides "Just Accepted" as a free service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. "Just Accepted" manuscripts appear in full in PDF format accompanied by an HTML abstract. "Just Accepted" manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are accessible to all readers and citable by the Digital Object Identifier (DOI®). "Just Accepted" is an optional service offered to authors. Therefore, the "Just Accepted" Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the "Just Accepted" Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these "Just Accepted" manuscripts.

Proteogenomic Investigation of Strain Variation in Clinical *Mycobacterium tuberculosis* Isolates

Tiaan Heunis †§*, Anzaan Dippenaar †§, Robin M. Warren †, Paul D. van Helden †, Ruben G. van der Merwe †, Nicolaas C. Gey van Pittius †, Arnab Pain ‡, Samantha L. Sampson † and David L. Tabb †

† DST/NRF Centre of Excellence for Biomedical Tuberculosis Research, SAMRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

‡ Pathogen Genomics Laboratory, BESE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

KEYWORDS

Mycobacterium tuberculosis, clinical microbial proteogenomics, variant peptides, strain variation.

ABSTRACT

Mycobacterium tuberculosis consists of a large number of different strains that display unique virulence characteristics. Whole-genome sequencing has revealed substantial genetic diversity among clinical *M. tuberculosis* isolates, and elucidating the phenotypic variation encoded by this

genetic diversity will be of utmost importance to fully understand *M. tuberculosis* biology and pathogenicity. In this study we integrated whole-genome sequencing and mass spectrometry (GeLC-MS/MS) to reveal strain-specific characteristics in the proteomes of two clinical *M. tuberculosis* Latin American-Mediterranean isolates. Using this approach we identified 59 peptides containing single amino acid variants, which covered ~9% of all total coding nonsynonymous single nucleotide variants detected by whole-genome sequencing. Furthermore, we identified 29 distinct peptides that mapped to a hypothetical protein not present in the *M. tuberculosis* H37Rv reference proteome. Here we provide evidence for the expression of this protein in the clinical *M. tuberculosis* SAWC3651 isolate. The strain-specific databases enabled confirmation of genomic differences (i.e. large genomic regions of difference and nonsynonymous single nucleotide variants) in these two clinical *M. tuberculosis* isolates and allowed strain differentiation at the proteome level. Our results contribute to the growing field of clinical microbial proteogenomics and can improve our understanding of phenotypic variation in clinical *M. tuberculosis* isolates.

INTRODUCTION

Mycobacterium tuberculosis can be considered as the most successful bacterial pathogen in human history and is the causative agent of tuberculosis (TB), a chronic granulomatous lung disease. The complexity of the disease, coupled with increased drug resistance in *M. tuberculosis*, has ensured that TB remains one of the deadliest infectious diseases worldwide. This is exemplified by the fact that 6.1 million new TB cases and 1.4 million TB-related deaths were reported globally to the WHO in 2015.¹ Alarming, this seems to be an underestimation, as the actual number of new TB cases is estimated to be 10.4 million.¹ *M. tuberculosis* strains can

be grouped into at least 40 known strain families, with many of these strains presenting different pathogenic characteristics and host adaptations.² One example is the Latin American-Mediterranean (LAM) genotype of *M. tuberculosis*, shown to cause ~15% of TB cases worldwide.³ Differences in virulence characteristics have been observed for members of this genotype.⁴

It was envisaged that *M. tuberculosis* phenotypic variation would readily be explained by underlying genetic features when the first whole-genome sequence of *M. tuberculosis* (of the laboratory strain H37Rv) was made available.⁵ However, the genome sequence of *M. tuberculosis* H37Rv has not provided an unambiguous understanding of the mechanisms of pathogenicity. Furthermore, recent advances in next-generation sequencing technologies have enabled whole-genome sequencing of several members of the *M. tuberculosis* complex (MTBC), revealing substantial genetic diversity among clinical *M. tuberculosis* isolates.^{6,7} Unravelling the phenotypic significance of this genetic diversity will be of utmost importance to fully understand the biology and pathogenicity of *M. tuberculosis*.

Shotgun proteomics provides direct protein-level evidence for gene products by matching peptide tandem mass spectra, obtained by high-resolution tandem mass spectrometry, to predicted spectra from a proteome database or to entries from spectral reference libraries. Traditional proteome databases are derived from a reference strain and do not reflect genomic variants and strain differences (nonsynonymous single nucleotide variants (nsSNVs) and insertions and deletions). These could impact strain phenotype and consequently disease outcome. Proteogenomics has thus emerged as a powerful tool that integrates high-throughput next-generation sequencing with high-resolution tandem mass spectrometry to generate bespoke databases for peptide-spectrum matching. Proteogenomics has been used to refine genome

1
2
3 annotation, define translational start sites and to discover novel protein-coding regions.^{8–15}
4
5 Predicted protein sequences are commonly obtained by six-frame translation of a reference
6
7 genome; however, this leads to an explosion in search space and results in decreased sensitivity
8
9 of identifications using standard target-decoy proteomic database search strategies.¹⁶ This
10
11 approach has most extensively been used for proteogenomic analysis of *M. tuberculosis* H37Rv
12
13 and recently of *Mycobacterium smegmatis* mc²155.^{8,17,18} It is important to note that the above
14
15 mentioned proteogenomic studies were performed on mycobacterial strains that have complete
16
17 genome sequences available, which is not the case for most clinical *M. tuberculosis* isolates. De
18
19 Souza et al. (2011) generated a proteomic search database to characterize laboratory and clinical
20
21 *M. tuberculosis* isolates by using their Multistrain Mass Spectrometry prokaryote database
22
23 builder (MSMSpddb) and protein sequences derived from five *M. tuberculosis* strains as well as
24
25 three *M. bovis* strains.¹⁹ Single amino acid variants (SAAVs) were introduced into this database
26
27 by concatenating variant peptide sequences to the canonical protein sequences using an approach
28
29 described by Schandorff et al. (2007).²⁰ This strategy, however, does not accommodate all strain-
30
31 specific SAAVs or strain differences and is limited to those present in the reference proteomes
32
33 used for database construction. Customized databases for clinical *M. tuberculosis* isolates can
34
35 easily be generated by supplementing reference proteomes with data obtained from whole-
36
37 genome sequencing and/or RNA sequencing of the same isolates, to directly capture sample-
38
39 specific SAAVs and other genomic differences, as has been reported for other organisms.^{13–15,21}
40
41
42
43
44
45
46
47
48

49 In this study we used a proteogenomic approach to elucidate strain-specific
50
51 characteristics in the proteomes of two clinical *M. tuberculosis* LAM (lineage 4.3) isolates. We
52
53 were able to detect variant peptides as well as novel peptides and proteins not present in the *M.*
54
55 *tuberculosis* H37Rv reference proteome. Furthermore, we were able to verify genomic
56
57
58
59
60

differences in these two clinical *M. tuberculosis* isolates at the proteome level. Our proteogenomic analysis contributes to the growing field of clinical microbial proteogenomics and facilitates elucidation of strain variation in clinical *M. tuberculosis* isolates.

MATERIALS AND METHODS

Mycobacterial sample selection and growth conditions

This study was approved by the Health Research Ethics Committee of Stellenbosch University (N10/04/126) and all culturing of *M. tuberculosis* was conducted in a Biosafety level 3 facility. Clinical isolates were made available from an extensive longitudinal collection of *M. tuberculosis* isolates circulating in the Western Cape of South Africa. Isolates were initially cultured in mycobacterial growth indicator tubes (MGITs) at 37 °C until positive growth was detected by means of the BACTEC 460 instrument (BD Biosciences, Sparks, MD, USA). The MGIT tubes were briefly centrifuged to collect mycobacterial cells and the isolates were routinely characterized by IS6110 RFLP and spoligotyping using internationally standardized techniques.^{22,23} Clinical isolates SAWC3651 and SAWC3517, belonging to the LAM (lineage 4.3) genotype and IS6110 family 14 and 9, respectively, and *M. tuberculosis* H37Rv were selected for analysis.²⁴ Cultures were subsequently maintained at -80 °C. The strains were cultured in 7H9 Middlebrook medium supplemented with 0.2% (g/L) dextrose, 0.0004% (v/v) catalase, 0.2% (v/v) glycerol and 0.1% (v/v) Tween 80.

DNA extraction and whole-genome sequencing

DNA was extracted as previously described.²⁵ The DNA library was sequenced on the Illumina HiSeq2000 platform (Illumina, Inc, San Diego, CA) according to the manufacturer's specifications. One microgram of DNA was used to prepare the library for sequencing, according

to the manufacturer's instructions using the Nextera DNA sample preparation kit. Paired-end sequencing (2 x 100 bp) was performed on ~500 bp fragment sizes.

Whole-genome sequence analysis

An in-house automated pipeline for *M. tuberculosis* whole-genome sequencing analysis was developed to routinely analyze Illumina sequencing data, as described previously.²⁶ Briefly, quality assessment of the sequencing data (in FASTQ format) was performed using FASTQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), followed by trimming of adapters and low-quality bases with a Phred quality score of less than 20 and filtering for a minimum read length of 36 using Trimmomatic.²⁷ Reads were subsequently mapped to the *M. tuberculosis* H37Rv genome (Genbank: AL123456) using three different alignment algorithms, namely the Burrows-Wheeler Alignment Tool (BWA)²⁸, Novoalign (<http://www.novocraft.com/main/downloadpage.php>) and SMALT.²⁹ The alignment files were subjected to local realignment and de-duplication using the Genome Analysis Toolkit (GATK)³⁰ and Picard Tools (<http://picard.sourceforge.net>). Single nucleotide variants (SNVs) in coding as well as non-coding regions and insertions and deletions (indels) were identified from each alignment file using GATK, and the overlap of these variants identified from all three alignment files was used for further analysis. Variants were annotated using annotation data from TubercuList.³¹

Phylogenomic analysis

The raw sequencing data generated for the purpose of this study and 21 selected mycobacterial genomes published previously or available in public databases (Supplementary Table S1) were assessed for evolutionary relationships. Concatenated sequences containing high-confidence

variable sites (coding and non-coding SNVs) with respect to *M. tuberculosis* H37Rv were used to infer the phylogenetic relationship among the various strains analyzed, as previously described.³² The general time reversal (GTR) model of nucleotide substitution was applied to construct a maximum likelihood phylogeny of the isolates included in this analysis with Randomized Axelerated Maximum Likelihood (RaxML) with 1000 bootstrap pseudo-replicates.^{33,34} Positions containing gaps or missing data were not considered for the analysis.

Sample preparation and protein extraction

Protein extraction was conducted as previously described, with slight modifications.³⁵ Briefly, *M. tuberculosis* H37Rv and the clinical isolates were cultured in biological duplicate at 37 °C until mid-log phase, OD_{600nm} ~ 0.6-0.7, and mycobacterial cells were collected by centrifugation (2500g, 10 min) at 4 °C. The cell pellets were washed with ice-cold lysis buffer (10 mM Tris-HCl pH 7.4, 0.1% Tween 80), supplemented with Complete protease inhibitor cocktail mix (Roche, Mannheim, Germany). Cell pellets were collected after centrifugation (13 000g, 2 min) and suspended in ice-cold lysis buffer containing RNase-free DNase I. Cells were mechanically lysed with glass beads, 0.1mm, for 20 seconds at a speed of 4.0 m.s⁻¹ using a Ribolyser (Bio101 Savant, Vista, CA, USA). Cells were cooled on ice for one minute after lysis, and this process was repeated for a total of six lysis steps. The lysate was subsequently centrifuged for one minute at 13 000g and cooled on ice for one minute. The supernatant was transferred to a new Eppendorf tube, and this process was repeated twice to collect cell debris and clarify the cell lysate. The supernatant, containing the whole cell lysate proteins, was retained and filter sterilized using a 0.22 µm pore Acrodisc® 25 mm PF syringe filter (Sigma-Aldrich, St Louis, USA) before removal from the Biosafety level 3 facility.

Protein fractionation and in-gel trypsin digestion

Protein concentration of the whole cell lysates were determined using the RC DC Protein Assay (Bio-Rad, Hercules, CA, USA). Each biological replicate was analyzed in technical duplicate from this point onwards. We employed a GeLC-MS/MS approach for proteome analysis. Briefly, a total amount of 60 μ g of each sample was mixed with 10 μ l reducing sample buffer and incubated at 95 °C for 5 minutes. Proteins were fractionated by SDS-PAGE, using a 4-12% gradient, 1.0 mm NuPage gel (Invitrogen, Carlsbad, CA, USA), under reducing conditions for one hour at 150 V. Each SDS-PAGE gel was stained using Imperial™ Protein Stain (Invitrogen, Carlsbad, CA, USA) for one hour. The gels were subsequently destained overnight in ultra-pure water with continuous shaking. Each gel lane was divided into 10 fractions and each fraction was prepared for analysis by mass spectrometry. In brief, fractions were washed sequentially with 50% acetonitrile (ACN) and 50 mM ammonium bicarbonate (ABC) to remove excess stain and SDS. Proteins were reduced by the addition of 10 mM dithiothreitol (DTT) and incubation at 56 °C for one hour. Proteins were subsequently alkylated with 55 mM iodoacetamide (IAA) for one hour in the dark at room temperature. The reduced and alkylated proteins were subjected to in-gel trypsin digestion for 17 hours at 37 °C using sequencing grade, modified trypsin (Promega, Madison, USA) in 50 mM ABC. The enzyme to substrate ratio was 1:50 (w/w). The reaction was stopped and peptides were eluted from the gel pieces with 70% ACN containing 0.1% formic acid (FA). The peptides were desalted using STAGE-tips, prepared with Empore™ C18 SPE Disks (Sigma-Aldrich, St Louis, USA), and dried before storage at -20 °C.

Mass spectrometry

Dried peptides were dissolved in 5% ACN, containing 0.1% FA, and 10 μ l injections were made for nano-LC chromatography. All experiments were performed using a Thermo Scientific EASY-nLC II connected to a LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) equipped with a nano-electrospray source. For liquid chromatography, separation was achieved on an EASY-Column (2 cm, ID 100 μ m, 5 μ m, C18) pre-column followed by separation on a XBridge BEH130 NanoEase column (15 cm, ID 75 μ m, 3.5 μ m, C18) with a flow rate of 300 nl/min. Solvent A was 100% water, containing 0.1% FA, and solvent B was 100% ACN, containing 0.1% FA. The gradient used was 5-17% B in 5 min, 17-25% B in 90 min, 25-60% B in 10 min, 60-80% B in 5 min and maintained at 80% B for 10 min. The mass spectrometer was operated in data-dependent mode to automatically switch between Orbitrap-MS and LTQ-MS/MS acquisition. Data were acquired using the Xcalibur software package (Thermo Fisher Scientific, Bremen, Germany). Precursor ion scan MS spectra (m/z 400 - 2000) were acquired in the Orbitrap with resolution $R = 60\,000$ with the number of accumulated ions being 1×10^6 . The 20 most intense ions were serially isolated and fragmented in the linear ion trap (number of accumulated ions 1.5×10^4) using collision induced dissociation (30% normalized collision energy). The lock mass option (polydimethylcyclsiloxane; $m/z = 445.120025$) enabled internal calibration of mass measurements in both the MS and MS/MS modes. In data-dependent LC-MS/MS experiments, dynamic exclusion was used with 60 s exclusion duration. Mass spectrometry conditions were 1.8 kV, capillary temperature of 250 $^{\circ}$ C, with no sheath and auxiliary gas flow. The ion selection threshold was 500 counts for MS/MS and an activation Q-value of 0.25 and activation time of 10 ms was applied.

Generation of strain-specific proteome databases

The consensus sequence for coding DNA sequences (CDSs) of each strain, which contained nsSNVs, was generated and the strain-specific versions of CDSs were translated and appended to the *M. tuberculosis* H37Rv reference proteome. No indels in CDSs were considered during database construction. All reads that failed to align to the *M. tuberculosis* H37Rv reference genome were assembled with the *de novo* assembly software package for short reads, Velvet, version 1.2.07.³⁶ A script published by the Victorian Bioinformatics Consortium (velvetk.pl) was used to determine the optimal kmer length to be used. Velvet was run with recommended parameters according to the software documentation, and a kmer length of 61 was used. A six-frame translation of the unmapped-read assemblies was generated for each clinical isolate and appended to the above mentioned databases to generate strain-specific proteome databases for the SAWC3517 and SAWC3651 clinical *M. tuberculosis* isolates, respectively.

Six-frame translation of *M. tuberculosis* H37Rv

A six-frame translation of the *M. tuberculosis* (Genbank: AL123456) reference genome was generated using an in-house Python script, applying a threshold of no less than seven putative amino acids to discard small putative open reading frames.

Peptide identification and protein assembly

We used the *M. tuberculosis* H37Rv reference proteome (downloaded from TubercuList on 20 November 2015) as well as customized strain-specific proteome databases, containing 72 known contaminants, for peptide identification by searching all tandem mass spectra using MS-GF+ (version 10089).³⁷ Peak picking was performed using MSConvert from ProteoWizard by converting Thermo RAW files to mzML peaklists.³⁸ A precursor mass tolerance of 20 ppm and a fragment mass tolerance of 0.5 Da was implemented in MS-GF+, and enzyme specificity was set

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

to consider semi-tryptic peptides. Carbamidomethylation of cysteines was set as fixed modification, and oxidation of methionine was allowed as variable modification. MS-GF+ mzIdentML files were loaded into IDPicker 3.1.9288.0 for protein assembly.³⁹ Stringent criteria were used for both peptide and protein identifications; spectral counts per protein were required to meet an increasing threshold until an empirical protein false discovery rate (FDR) of < 1% was achieved. Proteins were only considered when they contained a minimum of 2 distinct peptides that met these stringent criteria. Charge state differences and modifications on the same peptide sequence were not reported as distinct peptides and were regarded as duplicates. Finally, proteins that contained similar peptides and that could not be differentiated based on tandem mass spectrometry analysis alone were grouped to satisfy the principles of parsimony. Reverse hits and contaminants were removed before downstream analysis.

RESULTS AND DISCUSSION

Strain information and genomic characterization of clinical *M. tuberculosis* LAM isolates

Differences in virulence characteristics have been observed for clinical *M. tuberculosis* isolates belonging to the LAM genotype.⁴ Additionally, the LAM genotype has been associated with outbreaks of extensively drug-resistant TB in South Africa.⁴⁰ A sublineage of LAM strains, RD^{Rio}, was identified as the major cause of TB in Rio de Janeiro, Brazil, and these isolates all contained a 26 kb deletion affecting ten genes.⁴¹ Members of the RD^{Rio} LAM genotype show effective transmissibility and are thought to be hypervirulent, compared to non-RD^{Rio} strains.^{3,4,41} However, the molecular basis for this is not known. We therefore performed proteogenomic analysis on an RD^{Rio} (SAWC3517) and a non-RD^{Rio} (SAWC3651) clinical *M. tuberculosis* isolate to reveal strain-specific differences at the proteome level. Table 1 summarizes the

1
2
3 statistics obtained from whole-genome sequencing of SAWC3517 and SAWC3651, respectively.
4
5
6 Phylogenetic and SNV analysis revealed that these clinical isolates fall within lineage 4.3 of the
7
8 *M. tuberculosis* complex⁴² (Supplementary Figure S1), and spoligotype patterns confirm that
9
10 these strains belong to the LAM genotype (Supplementary Table S2). A total of 956 and 984
11
12 SNVs were detected in SAWC3517 and SAWC3651 respectively, as compared to *M.*
13
14 *tuberculosis* H37Rv that served as reference genome for read mapping and variant identification
15
16 (see Materials and Methods, Supplementary Table S3). One hundred and one and 105 SNVs
17
18 were located in intergenic regions in SAWC3517 and SAWC3651, respectively. Five hundred
19
20 and six of the 956 SNVs in SAWC3517 were nonsynonymous SNVs (nsSNVs). A total of 538
21
22 nsSNVs were identified in SAWC3651. Interestingly, 137 and 169 nsSNVs were specific to
23
24 either SAWC3517 or SAWC3651, respectively. The nsSNVs were found in the coding regions
25
26 of 422 and 430 genes in SAWC3517 and SAWC3651, respectively. Further genomic variants
27
28 identified in SAWC3517 and SAWC3651 included 15 indels that were solely detected in
29
30 SAWC3517 and 12 indels that were solely detected in SAWC3651. Forty eight indels were
31
32 identified in both of these isolates (Supplementary Table S4). Finally, 10 regions of difference
33
34 (RD), or large genomic deletions, were detected between SAWC3517 and SAWC3651
35
36 (Supplementary Table S5). Four novel regions of difference were identified in these two isolates
37
38 as well as an RD^{Rio} deletion (Supplementary Table S5). We identified a 26.416 kb deletion in
39
40 SAWC3517, using DELLY⁴³, confirming that this isolate belongs to the RD^{Rio} sublineage of the
41
42 LAM genotype. Whole-genome sequencing thus revealed genomic differences in these clinical
43
44 *M. tuberculosis* LAM isolates that can facilitate analysis of strain differentiation at the proteome
45
46 level, using a proteogenomic approach.
47
48
49
50
51
52
53
54

55 Preliminary quality control of proteomic data

56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

We compared our GeLC-MS/MS proteomic analysis, using *M. tuberculosis* H37Rv whole-cell lysates, to extensively fractionated *M. tuberculosis* H37Rv whole-cell lysates generated in independent laboratories. Publicly available *M. tuberculosis* H37Rv data was retrieved from ProteomeXchange or MassIVE and analyzed concurrently with our own *M. tuberculosis* H37Rv data.⁴⁴ Spectral counts per protein were required to meet an increasing threshold until an empirical protein FDR of < 1% was achieved for the combined datasets. In our dataset of four 10-fraction experiments from H37Rv, we identified 13 510 distinct peptides that lead to the identification of 2 041 protein groups at a combined empirical protein FDR of 0.74% (Supplementary Table S6). Fractionation of a single *M. tuberculosis* H37Rv whole-cell lysate into 24 fractions by reverse-phase chromatography (PXD001722), followed by subsequent analysis by nLC-MS/MS, enabled the identification of 32 416 distinct peptides and led to the identification of 2 877 protein groups at the same empirical protein FDR. Finally, a total of 35 839 distinct peptides and 2 906 protein groups were identified in a sample set consisting of 24 peptide fractions, comprised of peptides from both the exponential and stationary growth phases of *M. tuberculosis* H37Rv and fractionated through off-gel isoelectric focusing.¹⁸ Our dataset contained ~ 40% of the *M. tuberculosis* H37Rv peptide diversity of these other two datasets. This is not entirely unexpected as we performed limited protein-level fractionation in-gel, whereas these other workflows performed extensive peptide-level fractionation on single or combined (different growth phases) *M. tuberculosis* H37Rv whole-cell lysates. The protein contents compatible with each of the different fractionation methods can be a limiting factor that will significantly influence peptide and proteome coverage. Furthermore, peptide extraction efficiencies from polyacrylamide gels can vary.^{45,46} Even though our GeLC-MS/MS workflow did not provide as comprehensive proteome coverage as these other extensively fractionated *M.*

tuberculosis H37Rv whole-cell lysate datasets, we were still able to identify 13 510 distinct peptides that allowed identification of 51% of the predicted *M. tuberculosis* H37Rv proteome. This enables us to assess the recovery of isolate-specific peptides in our clinical *M. tuberculosis* proteomic data.

Proteome database construction and search space evaluation

We generated bespoke databases for automated searching of tandem mass spectra from whole-cell lysates of two clinical *M. tuberculosis* LAM isolates. The custom sequences were included alongside the *M. tuberculosis* H37Rv reference proteome, increasing the total search space with a minimum diminishment of sensitivity (Figure 1A). As anticipated, a database containing a six-frame translation of the *M. tuberculosis* H37Rv reference genome yielded the highest number of database entries (141 930) as compared to the *M. tuberculosis* H37Rv reference proteome (4 183), even when we removed spurious database entries that contained less than seven putative amino acids (Figure 1A). It is well known that searching large databases, such as six-frame translations of whole-genome sequences, decreases the sensitivity of peptide identification^{16,47} and leads to a large number of false-positive “novel” peptide identifications at a fixed global FDR.^{21,48} It is worth noting that these six-frame translation databases do not accommodate strain-specific nsSNVs that could be of importance for the detection of novel variant peptides in proteome datasets from clinical *M. tuberculosis* isolates. Our approach appended variant proteins, obtained from translation of the coding sequences in which an nsSNV was identified using *M. tuberculosis* H37Rv as reference genome, to the *M. tuberculosis* H37Rv reference proteome for a more tailored database. This resulted in the addition of 422 and 430 variant proteins to the *M. tuberculosis* H37Rv reference proteome, increasing the database entries from 4 183 in the *M. tuberculosis* H37Rv reference proteome to 4 605 entries in the variant SAWC3517

and 4 613 entries in the variant SAWC3651 proteome databases, respectively (Figure 1A). The approach of only introducing nsSNVs into the reference database has one drawback, however, in that protein coding sequences not present in the reference strain are discarded when using a reference genome for genome-guided assembly of sequencing reads. We therefore argued that a more complete strain-specific database can be generated by performing *de novo* assembly on the unmapped reads, to obtain possible coding regions not present in the *M. tuberculosis* H37Rv reference genome. Addition of the six-frame translation of the *de novo* assembled unmapped reads to the variant SAWC3517 and variant SAWC3651 proteome databases resulted in 5 890 entries in the strain-specific SAWC3517 and 5 880 entries in the strain-specific SAWC3651 proteome databases, respectively (Figure 1A). Although this added ~1300 entries to each strain database, the expansion of the search space was much less drastic than a whole genome six-frame translation. This highlights the effectiveness of this strategy to minimize search space inflation for possible detection of novel peptides and proteins in clinical *M. tuberculosis* isolates (Figure 1B, C). The *M. tuberculosis* H37Rv reference proteome contributed most of the search space for peptide-spectral matching in the bespoke databases for SAWC3517 and SAWC3651 (Figure 1B, C).

Peptide and protein identification in clinical *M. tuberculosis* isolates using the H37Rv reference proteome database

We identified 18.87% and 19.92% of the filtered tandem mass spectra using our search parameters and the canonical *M. tuberculosis* H37Rv reference proteome as database for searching tandem mass spectra of peptides derived from the clinical *M. tuberculosis* isolates SAWC3517 and SAWC3561, respectively. This resulted in the identification of 14 181 and 12 580 distinct peptides in SAWC3517 and SAWC3561, respectively (Figure 2, Supplementary

Table S7). A total of 1 766 and 1 588 protein groups were identified at an empirical protein FDR of < 1% in SAWC3517 and SAWC3561, respectively (Figure 2, Supplementary Table S7).

Peptide and protein identification in clinical *M. tuberculosis* isolates using bespoke databases

We identified 18.69% and 19.65% of the filtered tandem mass spectra when using strain-specific proteome databases in SAWC3517 and SAWC3561, respectively. This resulted in the identification of 13 931 and 12 541 distinct peptides in SAWC3517 and SAWC3561, respectively (Figure 2, Supplementary Table S8). These peptide identifications included a total of 59 distinct, high-confidence, variant peptides that are not present in the *M. tuberculosis* H37Rv reference proteome. Forty one and 38 variant peptides were identified in SAWC3517 and SAWC3561, respectively (Supplementary Figure S2, Supplementary Table S9). Distinct peptide identifications from the filtered tandem mass spectra slightly decreased from 14.55% to 14.41% in SAWC3517 and from 15.74% to 15.47% in SAWC3561, respectively, when using strain-specific proteome databases. A total of 1 733 and 1 555 protein groups, at an empirical protein FDR < 1%, were identified in SAWC3517 and SAWC3561 respectively (Figure 2, Supplementary Table S8). Slightly decreased identifications in the strain-specific proteome databases were most likely a result of the reduced sensitivity associated with larger proteome databases.⁴⁷ However, this reduction in peptide and protein group identifications is minimal and demonstrates the value of this approach to minimize search space inflation during proteogenomic analysis of clinical *M. tuberculosis* isolates.

Importantly, we identified peptides that mapped to proteins encoded in the *de novo* assembled unmapped reads when using the strain-specific proteome databases. Some of these

coding sequences have possibly resulted from genomic rearrangements or gene duplications not present in the *M. tuberculosis* H37Rv genome, as the peptides also mapped to proteins present in the *M. tuberculosis* H37Rv reference proteome. These included a putative transposase required for the transposition of IS6110, an insertion element found at multiple and diverse locations in the genomes of members of the MTBC.⁴⁹ Other protein group identifications found in both the SAWC3517 *de novo* assembled unmapped reads as well as *M. tuberculosis* H37Rv reference proteome included a hypothetical protein SseC1, predicted to be involved in sulfur metabolism, and a thiosulfate sulfurtransferase, CysA2, predicted to be involved in the formation of thiosulfates. Further analysis revealed that the hypothetical protein SseC1 was identical to the SseC1 protein in *M. tuberculosis* H37Rv, and the thiosulfate sulfurtransferase was identical to the thiosulfate sulfurtransferase CysA2 (rhodanese-like protein) in *M. tuberculosis* H37Rv. Protein group identifications found in the SAWC3651 *de novo* assembled unmapped reads also included proteins identical to SseC1 and CysA2 in SAWC3517.

Encouragingly, we identified distinct peptides that mapped to six-frame translations of contigs, derived from the SAWC3561 *de novo* assembled unmapped reads, not present in the *M. tuberculosis* H37Rv reference proteome. We identified 29 distinct peptides, derived from 94 filtered tandem mass spectra, that mapped to a protein encoded in the first reverse strand of NODE_8 of the SAWC3651 *de novo* assembled unmapped reads (Supplementary Figure S3). Further analysis revealed that this protein is a predicted helicase as it contains a helicase superfamily C-terminal domain and a P-loop-containing nucleoside triphosphate hydrolase domain. Furthermore, this protein is identical to a predicted helicase present in four other sequenced clinical *M. tuberculosis* isolates. We thus provide the first direct protein-level evidence for the expression of this predicted helicase in our clinical *M. tuberculosis* SAWC3651

1
2
3 isolate using a proteogenomic approach. Finally, we also identified 2 distinct peptides that
4 mapped to the first forward strand of NODE_5 of the SAWC3651 *de novo* assembled unmapped
5 reads. Manual inspection of the 3 filtered peptide tandem mass spectra for this entry, however,
6 revealed that two of these spectra were likely false positive identifications as there are numerous
7 unassigned peaks and the spectra contained a high signal-to-noise ratio.⁵⁰ Appending six-frame
8 translations of *de novo* assembled unmapped reads thus enabled identification of peptides
9 mapping to proteins that are either not present in *M. tuberculosis* H37Rv or where coding regions
10 differ in the genomes of clinical *M. tuberculosis* isolates, possibly as a result of genomic
11 rearrangements or gene duplications.
12
13
14
15
16
17
18
19
20
21
22
23

24 25 **Identification of single amino acid variants and variant proteins in clinical *M. tuberculosis*** 26 **isolates** 27 28 29 30

31 We detected 72 variant peptides in the proteomes of the two clinical *M. tuberculosis* LAM
32 isolates when using our strain-specific proteomic databases (Supplementary Table S8). However,
33 the FDR of novel peptides can be substantially underestimated during proteogenomic analysis.⁴⁷
34 We therefore manually verified the automated peptide-spectral matches and this resulted in the
35 identification of 59 high-confidence variant peptides, which covered 8% and 7% of the total
36 coding nsSNVs detected in SAWC3517 and SAWC3651, respectively (Supplementary Figure
37 S2, Supplementary Table S9). It is important to note that we only analyzed proteome samples
38 from mid-log phase cultures and we might have missed variants in conditionally expressed
39 proteins. Analyzing samples from multiple time-points and/or different conditions could thus
40 increase the number of variant peptides identified. Interestingly, our “SAAV recovery rates” are
41 comparable to those obtained in two published studies aimed at the identification of variant
42 peptides in cancer proteomes.^{51,52} Only five of the 59 variant peptides identified in this study
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 have previously been reported.¹⁹ Twenty of the 59 variant peptides were identified in both
4
5 SAWC3517 and SAWC3561, while 21 and 18 of the variant peptides were solely detected in
6
7 SAWC3517 and SAWC3651, respectively. Nine of these 21 variant peptides were strain-specific
8
9 to SAWC3517 and 12 of the 18 variant peptides were strain-specific to SAWC3561. These
10
11 peptides can thus be used to distinguish these two closely-related clinical *M. tuberculosis* isolates
12
13 from each other. One example of a strain-specific variant peptide is EPVDSAVLANGDEIQIGK
14
15 that is unique to SAWC3517 (Figure 3 A). This peptide originates from the glycogen
16
17 accumulation regulator GarA in SAWC3517 and contains a valine to isoleucine substitution at
18
19 position 137, compared to the canonical sequence in *M. tuberculosis* H37Rv and SAWC3651
20
21 (Figure 3 A, B).
22
23
24
25
26
27

28 It is important to consider that certain chemical modifications can introduce the same
29
30 mass shift as an amino acid substitution, and care should be taken when reporting these as
31
32 variant peptides.⁴⁷ We thus screened our dataset for possible chemical modifications that can
33
34 result in false positive variant peptide identifications (Supplementary Table S9). We identified 5
35
36 “variant” peptides that could possibly be false positive identifications due to chemical
37
38 modification. Deamidation of asparagine causes a +0.9848 Da mass shift that mimics an
39
40 asparagine to aspartic acid substitution and peptides with a similar mass shift were identified in
41
42 Rv3521 (possible N295D substitution) and Rv0292 (possible N217D substitution). Similarly,
43
44 methylation causes a +14.0266 mass shift that can mimic a variety of amino acid substitutions
45
46 including a serine to threonine and an aspartic acid to glutamic acid substitution, respectively.
47
48 We identified peptides in Rv0638 (possible S21T substitution), Rv2931 (possible D624E
49
50 substitution) and Rv2178c (possible D265E substitution) with a mass shift suggestive of possible
51
52 serine and aspartic acid methylation, respectively (Supplementary Table S9). These peptides can
53
54
55
56
57
58
59
60

thus not definitively be identified as variant peptides. However, proteins containing a N295D substitution in Rv3521, a N217D substitution in Rv0292, a D624E substitution in Rv2931 and a D265E substitution in Rv2178c are present in members of the MTBC, and one could argue that our identifications could indeed be true variants. This is supported by the fact that the VSADAELPDTLPAIR variant peptide (with a N217D substitution) has previously been reported in a clinical *M. tuberculosis* Beijing isolate.¹⁹

The 59 variant peptides mapped to 51 variant proteins, from a total of 540 variant proteins in both strain-specific databases. These variant proteins were involved in a variety of metabolic processes that can influence the physiological state of the mycobacterial cells. Slight differences were observed in the TubercuList functional categories of proteins in which the variant peptides were detected (Figure 4, Supplementary Table S9). The clinical *M. tuberculosis* SAWC3517 isolate expressed a higher number of variant proteins involved in “cell wall and cell processes” and “conserved hypotheticals” as compared to SAWC3651 (Figure 4, Supplementary Table S9). On the other hand, the clinical *M. tuberculosis* SAWC3651 isolate expressed a higher number of variant proteins involved “intermediary metabolism and respiration” and “regulatory proteins” as compared to SAWC3517 (Figure 4, Supplementary Table S9). Interestingly, a number of proteins that form part of mycobacterial type VII secretion systems, or which serve as substrates of this secretion system, were identified as variant proteins in SAWC3517 and included Rv1797, Rv0292 and Rv3864.

Identification of peptides in large genomic regions of difference (RDs)

Next we examined if we could detect peptides and proteins encoded in the RDs, i.e. large genomic deletions, present in each of the clinical *M. tuberculosis* isolates, as these would still be

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

present in the databases. No peptides were detected in regions deleted from SAWC3517, such as RD174, RD149 and SAWC3517 RD1. Peptides were identified that mapped to RD152 and the SAWC3517 RD4 in SAWC3517, however these peptides map to a putative transposase required for the transposition of IS6110. There are numerous IS6110 insertion sites in the *M. tuberculosis* genome and these peptides could not be unambiguously assigned to these specific RD regions.⁴⁹ No peptides were detected from proteins encoded in the RD^{Rio} deletion region in SAWC3517.⁴¹ We also did not detect any peptides in regions deleted from SAWC3651, such as RD115, RD149, RD152 and the SAWC3651 RDs 2 and 3. The deletion detected in the *pks15/l* gene in our clinical *M. tuberculosis* LAM isolates is a well-described 7 bp deletion in Euro-American lineages that results in a loss of function of the phenolphthiocerol synthesis polyketide synthase responsible for the production of the well-known phenolic glycolipid virulence factor.^{53,54} We were able to detect peptides that mapped to this gene product in both SAWC3517 and SAWC3651. Finally, we detected peptides that mapped to RDs that can be used to differentiate these two clinical isolates. We detected two distinct peptides that mapped to Rv0145 in SAWC3651, but no peptides were identified that mapped to this protein in SAWC3517, which exhibits an Rv0145 deletion (SAWC3517 RD2). Similarly, we detected two distinct peptides that mapped to Rv0137 in SAWC3651 but none that mapped to Rv0137 in SAWC3517, which exhibits a deletion spanning Rv0137-Rv0138 (SAWC3517 RD1). Peptides were detected in SAWC3651 that mapped to Rv1993c, Rv1996 and Rv1997, whereas no peptides were detected in SAWC3517, where the RD174 deletion stretches from Rv1992c-Rv1997. Twenty distinct peptides mapping to Rv2280 were detected in SAWC3651 whereas no peptides were detected that mapped to Rv2280 in SAWC3517, where the SAWC3517 RD4 deletion spans Rv2277c-Rv2281. Our data are thus consistent with the absence of protein products when a deletion was

identified during genomic analysis of these two clinical *M. tuberculosis* LAM isolates. This information can thus be used to distinguish closely-related *M. tuberculosis* isolates on a proteome level and help to elucidate strain differences in clinical *M. tuberculosis* isolates.

CONCLUSION

In this study we combined whole-genome sequencing and proteomic analysis of two clinical *M. tuberculosis* LAM isolates to identify strain-specific features at the proteome level. Our database construction approach minimized the increase in theoretical search space and enabled the identification of novel variant peptides with high confidence. Databases can possibly be compacted even more to reduce search space by using the approach described by Schandorff et al. (2007).²⁰ Nevertheless, we were able to identify ~9% of all total coding nsSNVs, detected by whole-genome sequencing, in the proteomes of these two clinical *M. tuberculosis* LAM isolates. Importantly, the majority of these variant peptides are not present in publicly available *M. tuberculosis* protein sequences. This further supports this approach for general use in bespoke database construction for clinical *M. tuberculosis* proteogenomics. Furthermore, the approach of appending six-frame translations of *de novo* assembled unmapped reads enabled the detection of a protein not present in the *M. tuberculosis* H37Rv reference proteome. This allowed more in-depth proteome characterization of two clinical *M. tuberculosis* isolates. The approach applied in this study will help to expand the growing field of *M. tuberculosis* proteogenomics and has the potential to advance our understanding of strain variation in circulating *M. tuberculosis* strains. This could provide novel insights into biological mechanisms and pathogenesis of clinical *M. tuberculosis* isolates.

FIGURES

Figure 1. Proteome database construction and impact on search space. (A) Number of database entries in different bespoke clinical *M. tuberculosis* proteome databases as compared to the *M. tuberculosis* H37Rv reference from TubercuList. (B) Database search space contribution of appended variant proteins (indicated as variant proteins in the label of Figure 1 B and C) and six-frame translations of *de novo* assembled unmapped reads (indicated as strain-specific in the label of Figure 1 B and C) from the clinical *M. tuberculosis* isolate SAWC3517 and (C) the clinical *M. tuberculosis* isolate SAWC3651, compared to the *M. tuberculosis* H37Rv reference proteome.

Figure 2. Influence of strain-specific proteomic search databases on peptide and protein group identifications in clinical *M. tuberculosis* isolates (SAWC3517 and SAWC3651) belonging to the Latin-American Mediterranean (LAM) genotype. Identifications obtained using strain-specific proteome databases was compared to those obtained when using the canonical *M. tuberculosis* H37Rv reference proteome database.

Figure 4. TubercuList functional categories of variant proteins identified in the clinical *M. tuberculosis* LAM isolates.

Table 1. Statistics from whole-genome sequencing and variant identification of clinical *M. tuberculosis* isolates, when using *M. tuberculosis* H37Rv as reference genome.

	SAWC3517	SAWC3651
Total number of reads	4 010 380 (x2)	3 666 637 (x2)
Percentage of mapped reads	99.05%	98.93%
Mean depth of coverage	136.12	125.37
Genome coverage $\geq 50\times$ (%)	95.33%	95.03%
Total SNVs [#]	956 (260)	984 (288)
Intergenic SNVs*	101 (19)	105 (23)
Synonymous SNVs*	855 (241)	879 (265)
Nonsynonymous SNVs*	506 (137)	538 (169)
Indels ^{##}	63 (15)	60 (12)

[#]SNVs = single nucleotide variants, ^{##}Indels = insertions and deletions. *Value in brackets indicate SNVs and indels specific to each isolate.

ASSOCIATED CONTENT

Supporting Information

Supplementary Figure S1. Phylogenetic relationship of two clinical *M. tuberculosis* isolates, SAWC3517 and SAWC3651, inferred from high-confidence variable sites (coding and non-coding SNVs) as compared to *M. tuberculosis* H37Rv, with 21 other mycobacterial strains.

Supplementary Figure S2. Peptide-spectral matches of the variant peptides identified in this study.

Supplementary Figure S3. Peptide-spectral matches of the peptides mapping to a predicted helicase identified in the *M. tuberculosis* SAWC3651 *de novo* assembled unmapped reads.

Supplementary Table S1. Twenty one selected mycobacterial genomes used for Phylogenomic analysis in combination with the two clinical *M. tuberculosis* Latin-American Mediterranean isolates used in this study.

Supplementary Table S2. Spoligotype patterns of the two clinical *M. tuberculosis* isolates, SAWC3517 and SAWC3651, indicating that they belong to the Latin-American Mediterranean genotype.

Supplementary Table S3. Single nucleotide variants (SNVs) identified in the two clinical *M. tuberculosis* isolates, SAWC3517 and SAWC3651, when compared to *M. tuberculosis* H37Rv.

Supplementary Table S4. Insertions and deletions (indels) identified in the two clinical *M. tuberculosis* isolates, SAWC3517 and SAWC3651, when compared to *M. tuberculosis* H37Rv.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Supplementary Table S5. Large genomic regions of difference (RD) identified in the two clinical *M. tuberculosis* isolates, SAWC3517 and SAWC3651, when compared to *M. tuberculosis* H37Rv.

Supplementary Table S6. Peptides and protein groups identified in *M. tuberculosis* H37Rv from a multisite comparison.

Supplementary Table S7. Peptides and protein groups identified in *M. tuberculosis* SAWC3517 and SAWC3651 when using the canonical *M. tuberculosis* H37Rv proteome database.

Supplementary Table S8. Peptides and protein groups identified in *M. tuberculosis* SAWC3517 and SAWC3651 when using strain-specific proteome databases.

Supplementary Table S9. Single amino acid variant peptides identified in the clinical *M. tuberculosis* isolates, SAWC3517 and SAWC3651, when using strain-specific databases for peptide and protein identification.

AUTHOR INFORMATION

Corresponding Author

*Tiaan Heunis. E-mail: heunistd@sun.ac.za.

Author Contributions

T.H., A.D., D.L.T., R.M.W., and N.C.G. v P. conceived this study. P.D. v H., N.C.G. v P., R.M.W., S.L.S., and D.L.T. directed the project. T.H., A.D., R.G. vd M., A.P., and D.L.T. generated and/or analyzed the data. T.H., A.D., S.L.S., R.M.W., and D.L.T. drafted the manuscript or revised it critically for important intellectual content. The manuscript was written

through contributions of all authors. All authors approved the final version of the manuscript.

Any opinion, finding and conclusion or recommendation expressed in this material is that of the authors and the NRF does not accept any liability in this regard. The authors declare no conflict of interest.

§ These authors contributed equally to this work.

Notes

Whole genome sequence data have been deposited at the European Nucleotide Archive under accession number PRJEB21206 (ERS1779972 and ERS1779973). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD006843.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Salome Smit for performing the mass spectrometric analysis at the Proteomics Laboratory of the Central Analytical Facility of Stellenbosch University.

FUNDING SOURCES

This work was supported by the South African National Research Foundation (NRF), South African Medical Research Council (SAMRC), Harry Crossley Foundation, Claude Leon Foundation, the Department of Biomedical Sciences, Stellenbosch University and King Abdullah University of Science and Technology (KAUST award number BAS/1/1020-01-01). S.L. Sampson is funded by the South African Research Chairs Initiative of the Department of Science and Technology and NRF of South Africa, award number UID 86539. D.L. Tabb is funded by the SAMRC under the South African Tuberculosis Bioinformatics Initiative.

REFERENCES

(1) World Health Organization. Global tuberculosis report, **2016**. Geneva: WHO; 2016. (<http://apps.who.int/iris/bitstream/10665/250441/1/9789241565394-eng.pdf?ua=1>).

(2) Gagneux, S.; DeRiemer, K.; van, T.; Kato-Maeda, M.; de Jong, B. C.; Narayanan, S.; Nicol, M.; Niemann, S.; Kremer, K.; Gutierrez, M. C.; et al. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (8), 2869–2873.

(3) Gibson, A. L.; Huard, R. C.; Gey van Pittius, N. C.; Lazzarini, L. C. O.; Driscoll, J.; Kurepina, N.; Zozio, T.; Sola, C.; Spindola, S. M.; Kritski, A. L.; et al. Application of sensitive and specific molecular methods to uncover global dissemination of the major RD^{Rio} sublineage of the Latin American-Mediterranean *Mycobacterium tuberculosis* spoligotype family. *J. Clin. Microbiol.* **2008**, *46* (4), 1259–1267.

(4) Lazzarini, L. C. O.; Spindola, S. M.; Bang, H.; Gibson, A. L.; Weisenberg, S.; Carvalho, W. D. S.; Augusto, C. J.; Huard, R. C.; Kritski, A. L.; Ho, J. L. RD^{Rio} *Mycobacterium tuberculosis* infection is associated with a higher frequency of cavitary pulmonary disease. *J. Clin. Microbiol.* **2008**, *46* (7), 2175–2183.

(5) Cole, S. T.; Brosch, R.; Parkhill, J.; Garnier, T.; Churcher, C.; Harris, D.; Gordon, S. V.; Eiglmeier, K.; Gas, S.; Barry, C. E.; et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **1998**, *393* (6685), 537–544.

(6) Zhang, H.; Li, D.; Zhao, L.; Fleming, J.; Lin, N.; Wang, T.; Liu, Z.; Li, C.; Galwey, N.; Deng, J.; et al. Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat. Genet.*

- 2013**, *45* (10), 1255–1260.
- (7) Farhat, M. R.; Shapiro, B. J.; Kieser, K. J.; Sultana, R.; Jacobson, K. R.; Victor, T. C.; Warren, R. M.; Streicher, E. M.; Calver, A.; Sloutsky, A.; et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* **2013**, *45* (10), 1183–1189.
- (8) Potgieter, M. G.; Nakedi, K. C.; Ambler, J. M.; Nel, A. J. M.; Garnett, S.; Soares, N. C.; Mulder, N.; Blackburn, J. M. Proteogenomic analysis of *Mycobacterium smegmatis* using high resolution mass spectrometry. *Front. Microbiol.* **2016**, *7*, 1–15.
- (9) Yang, M.; Yang, Y.; Chen, Z.; Zhang, J.; Lin, Y.; Wang, Y.; Xiong, Q.; Li, T.; Ge, F.; Bryant, D. A.; et al. Proteogenomic analysis and global discovery of posttranslational modifications in prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (52), E5633–5642.
- (10) Koch, A.; Gawron, D.; Steyaert, S.; Ndah, E.; Crappé, J.; De Keulenaer, S.; De Meester, E.; Ma, M.; Shen, B.; Gevaert, K.; et al. A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics* **2014**, *14* (23–24), 2688–2698.
- (11) Yagoub, D.; Tay, A. P.; Chen, Z.; Hamey, J. J.; Cai, C.; Chia, S. Z.; Hart-Smith, G.; Wilkins, M. R. Proteogenomic discovery of a small, novel protein in yeast reveals a strategy for the detection of unannotated short open reading frames. *J. Proteome Res.* **2015**, *14* (12), 5038–5047.
- (12) Bitton, D. A.; Smith, D. L.; Connolly, Y.; Scutt, P. J.; Miller, C. J. An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome. *PLoS One* **2010**, *5* (1), 1–10.

- (13) Wang, X.; Slebos, R. J. C.; Wang, D.; Halvey, P. J.; Tabb, D. L.; Liebler, D. C.; Zhang, B. Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* **2012**, *11* (2), 1009–1017.
- (14) Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Scalf, M.; Smith, L. M. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J. Proteome Res.* **2014**, *13* (1), 228–240.
- (15) Krug, K.; Popic, S.; Carpy, A.; Taumer, C.; Macek, B. Construction and assessment of individualized proteogenomic databases for large-scale analysis of nonsynonymous single nucleotide variants. *Proteomics* **2014**, *14* (23–24), 2699–2708.
- (16) Blakeley, P.; Overton, I. M.; Hubbard, S. J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J. Proteome Res.* **2012**, *11* (11), 5221–5234.
- (17) Kelkar, D. S.; Kumar, D.; Kumar, P.; Balakrishnan, L.; Muthusamy, B.; Yadav, A. K.; Shrivastava, P.; Marimuthu, A.; Anand, S.; Sundaram, H.; et al. Proteogenomic analysis of *Mycobacterium tuberculosis* by high resolution mass spectrometry. *Mol. Cell. Proteomics* **2011**, *10* (12), M111.011627.
- (18) Schubert, O. T.; Mouritsen, J.; Ludwig, C.; Röst, H. L.; Rosenberger, G.; Arthur, P. K.; Claassen, M.; Campbell, D. S.; Sun, Z.; Farrah, T.; et al. The Mtb proteome library: A resource of assays to quantify the complete proteome of *Mycobacterium tuberculosis*. *Cell Host Microbe* **2013**, *13* (5), 602–612.
- (19) de Souza, G. A.; Arntzen, M. Ø.; Fortuin, S.; Schürch, A. C.; Målen, H.; McEvoy, C. R. E.; van Soolingen, D.; Thiede, B.; Warren, R. M.; Wiker, H. G. Proteogenomic analysis of polymorphisms and gene annotation divergences in prokaryotes using a clustered mass

- spectrometry-friendly database. *Mol. Cell. Proteomics* **2011**, *10* (1), M110.002527.
- (20) Schandorff, S.; Olsen, J. V.; Bunkenborg, J.; Blagoev, B.; Zhang, Y.; Andersen, J. S.; Mann, M. A mass spectrometry-friendly database for cSNP identification. *Nat. Methods* **2007**, *4* (6), 465–466.
- (21) Zhang, K.; Fu, Y.; Zeng, W.-F.; He, K.; Chi, H.; Liu, C.; Li, Y.-C.; Gao, Y.; Xu, P.; He, S.-M. A note on the false discovery rate of novel peptides in proteogenomics. *Bioinformatics* **2015**, *31* (20), 3249–3253.
- (22) Kamerbeek, J.; Schouls, L.; Kolk, A.; van Agterveld, M.; van Soolingen, D.; Kuijper, S.; Bunschoten, A.; Molhuizen, H.; Shaw, R.; Goyal, M.; et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.* **1997**, *35* (4), 907–914.
- (23) Warren, R. M.; van Helden, P. D.; Gey van Pittius, N. C. Insertion element IS6110-based restriction fragment length polymorphism genotyping of *Mycobacterium tuberculosis*. *Mycobact. Protoc. Second Ed.* **2009**, 353–370.
- (24) Warren, R. M.; Sampson, S. L.; Richardson, M.; van Der Spuy, G. D.; Lombard, C. J.; Victor, T. C.; van Helden, P. D. Mapping of IS6110 flanking regions in clinical isolates of *Mycobacterium tuberculosis* demonstrates genome plasticity. *Mol. Microbiol.* **2000**, *37* (6), 1405–1416.
- (25) Warren, R.; de Kock, M.; Engelke, E.; Myburgh, R.; Gey van Pittius, N.; Victor, T.; van Helden, P. Safe *Mycobacterium tuberculosis* DNA extraction method that does not compromise integrity. *J. Clin. Microbiol.* **2006**, *44* (1), 254–256.
- (26) Black, P. A.; de Vos, M.; Louw, G. E.; van der Merwe, R. G.; Dippenaar, A.; Streicher, E. M.; Abdallah, A. M.; Sampson, S. L.; Victor, T. C.; Dolby, T.; et al. Whole genome

- sequencing reveals genomic heterogeneity and antibiotic purification in *Mycobacterium tuberculosis* isolates. *BMC Genomics* **2015**, *16* (857), 1–14.
- (27) Bolger, A. M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30* (15), 2114–2120.
- (28) Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr. arXiv* **2013**, *1303* (3997), 1–3.
- (29) Ponstingl, H.; Ning, Z. SMALT - A new mapper for DNA sequencing reads. *Fl000Posters* **2010**, No. 313, 1.
- (30) McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernysky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2009**, *20* (9), 1297–1303.
- (31) Lew, J. M.; Kapopoulou, A.; Jones, L. M.; Cole, S. T. TubercuList - 10 years after. *Tuberculosis* **2011**, *91* (1), 1–7.
- (32) Dippenaar, A.; Parsons, S. D. C.; Sampson, S. L.; van Der Merwe, R. G.; Drewe, J. A.; Abdallah, A. M.; Siame, K. K.; Gey van Pittius, N. C.; van Helden, P. D.; Pain, A.; et al. Whole genome sequence analysis of *Mycobacterium suricattae*. *Tuberculosis* **2015**, *95* (6), 682–688.
- (33) Stamatakis, A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **2006**, *22* (21), 2688–2690.
- (34) Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30* (9), 1312–1313.
- (35) de Souza, G. A.; Fortuin, S.; Aguilar, D.; Pando, R. H.; McEvoy, C. R. E.; van Helden, P.

- D.; Koehler, C. J.; Thiede, B.; Warren, R. M.; Wiker, H. G. Using a label-free proteomics method to identify differentially abundant proteins in closely related hypo- and hypervirulent clinical *Mycobacterium tuberculosis* Beijing isolates. *Mol. Cell. Proteomics* **2010**, *9* (11), 2414–2423.
- (36) Zerbino, D. R.; Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **2008**, *18* (5), 821–829.
- (37) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5* (5277), 1–10.
- (38) Chambers, M.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918–920.
- (39) Ma, Z.-Q.; Dasari, S.; Chambers, M. C.; Litton, M. D.; Scott, M.; Zimmerman, L. J.; Halvey, P. J.; Schilling, B.; Penelope, M.; Gibson, B. W.; et al. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* **2010**, *8* (8), 3872–3881.
- (40) Pillay, M.; Sturm, A. W. Evolution of the extensively drug-resistant F15/LAM4/KZN strain of *Mycobacterium tuberculosis* in KwaZulu-Natal, South Africa. *Clin. Infect. Dis.* **2007**, *45* (11), 1409–1414.
- (41) Lazzarini, L. C. O.; Huard, R. C.; Boechat, N. L.; Gomes, H. M.; Oelemann, M. C.; Kurepina, N.; Shashkina, E.; Mello, F. C. Q.; Gibson, A. L.; Virginio, M. J.; et al. Discovery of a novel *Mycobacterium tuberculosis* lineage that is a major cause of tuberculosis in Rio de Janeiro, Brazil. *J. Clin. Microbiol.* **2007**, *45* (12), 3891–3902.
- (42) Stucki, D.; Brites, D.; Jeljeli, L.; Coscolla, M.; Liu, Q.; Trauner, A.; Fenner, L.; Rutaihwa,

- L.; Borrell, S.; Luo, T.; et al. *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat. Genet.* **2016**, *48* (12), 1535–1543.
- (43) Rausch, T.; Zichner, T.; Schlattl, A.; Stütz, A. M.; Benes, V.; Korbel, J. O. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **2012**, *28* (18), i333–i339.
- (44) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dienes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223–226.
- (45) Shevchenko, A.; Tomas, H.; Havlis, J.; Olsen, J. V.; Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **2006**, *1* (6), 2856–2860.
- (46) Speicher, K. D.; Kolbas, O.; Harper, S.; Speicher, D. W. Systematic analysis of peptide recoveries from in-gel digestions for protein identifications in proteome studies. *J. Biomol. Tech.* **2000**, *11* (2), 74–86.
- (47) Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **2014**, *11* (11), 1114–1125.
- (48) Krug, K.; Carpy, A.; Behrends, G.; Matic, K.; Soares, N. C.; Macek, B. Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol. Cell. Proteomics* **2013**, *12* (11), 3420–3430.
- (49) Thierry, D.; Brisson-Noël, A.; Vincent-Lévy-Frébault, V.; Nguyen, S.; Guesdon, J. L.; Gicquel, B. Characterization of a *Mycobacterium tuberculosis* insertion sequence, IS6110,

- and its application in diagnosis. *J. Clin. Microbiol.* **1990**, 28 (12), 2668–2673.
- (50) Tabb, D. L.; Friedman, D. B.; Ham, A. L. Verification of automated peptide identifications from proteomic tandem mass spectra. *Nat. Protoc.* **2006**, 1 (5), 2213–2222.
- (51) Lobas, A. A.; Karpov, D. S.; Kopylov, A. T.; Solovyeva, E. M.; Ivanov, M. V.; Ilina, I. Y.; Lazarev, V. N.; Kuznetsova, K. G.; Ilgisonis, E. V.; Zgoda, V. G.; et al. Exome-based proteogenomics of HEK-293 human cell line: Coding genomic variants identified at the level of shotgun proteome. *Proteomics* **2016**, 16 (14), 1980–1991.
- (52) Ruggles, K. V.; Tang, Z.; Wang, X.; Grover, H.; Askenazi, M.; Teubl, J.; Cao, S.; McLellan, M. D.; Clauser, K. R.; Tabb, D. L.; et al. An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Mol. Cell. Proteomics* **2015**, 15 (3), 1060–1071.
- (53) Constant, P.; Perez, E.; Malaga, W.; Lanelle, M. A.; Saurel, O.; Daff, M.; Guilhot, C. Role of the *pks15/1* gene in the biosynthesis of phenolglycolipids in the *Mycobacterium tuberculosis* complex: Evidence that all strains synthesize glycosylated p-hydroxybenzoic methyl esters and that strains devoid of phenolglycolipids harbor a frameshift. *J. Biol. Chem.* **2002**, 277 (41), 38148–38158.
- (54) Reed, M. B.; Domenech, P.; Manca, C.; Su, H.; Barczak, A. K.; Kreiswirth, B. N.; Kaplan, G.; Barry, C. E. A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. *Nature* **2004**, 431 (7004), 84–87.

For TOC only