Accepted Manuscript

Pool-seq driven proteogenomic database for Group G Streptococcus

R.G. Weldatsadik, N. Datta, C. Kolmeder, J. Vuopio, J. Kere, S.V. Wilkman, J.W. Flatt, R. Vuento, K.J. Haapasalo, S. Keskitalo, M. Varjosalo, T.S. Jokiranta

PII: S1874-3919(19)30128-9
DOI: <https://doi.org/10.1016/j.jprot.2019.04.015>
Reference: JPROT 3365

To appear in: *Journal of Proteomics*

Received date: 24 January 2019
Revised date: 29 March 2019
Accepted date: 17 April 2019

Please cite this article as: R.G. Weldatsadik, N. Datta, C. Kolmeder, et al., Pool-seq driven proteogenomic database for Group G Streptococcus, Journal of Proteomics, <https://doi.org/10.1016/j.jprot.2019.04.015>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Pool-seq driven proteogenomic database for Group G Streptococcus

R.G. Weldatsadik^{1,2,3}, N. Datta^{1,12*}, C. Kolmeder^{2,11*}, J. Vuopio^{4,5}, J. Kere^{8,9,10}, S.V. Wilkman¹, J.W. Flatt⁶, R. Vuento⁷, K.J. Haapasalo¹, S. Keskitalo^{2,3}, M. Varjosalo^{2,3*§} and T. S. Jokiranta^{1*}

¹ Research Programs Unit, Immunobiology, Univ. of Helsinki, Helsinki, Finland

² Institute of Biotechnology, Molecular Systems Biology Research Group, HiLIFE, Univ. of Helsinki, Helsinki, Finland

³ Protein-Proteome Research Program, HiLIFE, Univ. of Helsinki, Helsinki, Finland

⁴ Institute of Biomedicine, Medical Microbiology and Immunology, Univ. of Turku, Turku, Finland

⁵ Department of Health Security, National Institute for Health and Welfare, Helsinki, Finland

⁶ Faculty of Biological and Environmental Sciences, Univ. of Helsinki, Helsinki, Finland

⁷ Department of Clinical Microbiology, Fimlab Laboratories, Tampere, Finland

⁸ Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden

⁹ Molecular Neurology Research Program, University of Helsinki, and Folkhälsan Institute of Genetics, Biomedicum Helsinki, Helsinki, Finland

¹⁰ Department of Genetics and Molecular Medicine, King's College London, London, UK

¹¹ Current address; Max Planck Institute for Developmental Biology, Tübingen, Germany

¹² Current address: Minerva Foundation Institute for Medical Research, Helsinki, Finland

* equal contribution

§To whom correspondence should be addressed:

Dr. Markku Varjosalo,

Viikinkaari 1, FIN-00014,

University of Helsinki, Finland;

tel. +358 2 941 59413, fax. +358 2 941 59366

e-mail: markku.varjosalo@helsinki.fi.

Keywords: Proteogenomics, Variant protein database, Group G Streptococcus, Pool-seq

Abstract

Proteogenomic databases use genomic and transcriptomic information for improved identification of peptides and proteins from mass spectrometry analyses. One application of such databases is in the discovery of variants/mutations. In this study, we created a proteogenomic database that contained sequences with variants derived from Pooled sequencing experiments (137 Group G Streptococcus strains sequenced in 3 pools) and used tandem mass spectrometry (MS/MS) to analyse eight protein samples from randomly selected strains sequenced in the pools. Using the proteogenomic variant database, we identified 385 variant peptides from the eight samples, none of which could be identified from the single genome conventional database utilized, while 71.2% and 93.5% of them were identified from the databases that contained 4 complete genomes and 26 assemblies, respectively. The proteogenomic variant databases exhibited the same properties as the conventional databases in terms of the Andromeda score distributions and the posterior error probability (PEP) values of the identified peptides.

Significance

For bacterial populations, such as Group G Streptococcus (GGS), with substantial intra-species diversity, simultaneous sequencing of large numbers of strains and generation of proteogenomic databases from those aids in improving the discovery of peptides in mass spectrometric analyses. Therefore, generation of proteogenomic variant protein databases from Pooled sequencing experiments can be a cost-effective method to complement conventional databases and discover subtle strain wise differences.

INTRODUCTION

Proteomics is shifting from the analysis of model organisms, with a defined sequenced genome and the mere goal of identifying the most abundant proteins, to the analysis of less defined systems [1]. Some examples of this new trend include the detection of variant proteins due to DNA variations [2], [3] and discovery of RNA-splice and -editing variants thereby improving gene annotations [4], [5]. In bacteria, the study of genomic variations in protein coding regions can shed light on phenotypic differences and disease pathogenesis such as virulence and antibiotic resistance [6]. This in turn enables the discovery of new targets for diagnostic use and vaccine research.

For large-scale studies, shotgun tandem mass spectrometry (MS/MS) has emerged as an effective method for peptide and protein identification from various types of samples. Identification of protein variants in mass spectrometry (MS) based proteomics is not, however, trivial since the protein databases often do not contain variant sequences. Certain search engines such as X! Tandem[7] and Mascot[8] offer exhaustive search options that take into account all possible amino acid substitutions resulting from single nucleotide substitutions. However, these methods lead to the dramatic explosion of the search space, loss of sensitivity and increased database search time. Another approach is proteogenomics[9], [10] in which custom protein sequence or spectral databases that contain the sequence variants are used for searching the spectra. Such databases have been created from genomic or transcriptomic data sources such as based on six-frame translation[11], [12].

Technical advances in genome sequencing during the past decade has led to the tremendous increase in the number of sequenced bacterial strains, thereby allowing the

study of the genomic diversity between different isolates. This is especially important for certain bacterial species or serogroups, such as Group A and G Streptococcus[13], [14], with high diversity (i.e. open pan-genomes). Even so, sequencing of large number of strains of each species incurs relatively high-cost and -labour associated with genome sequencing, assembly and annotation.

In this study we used pooled sequencing (Pool-seq) that we have previously shown to be a cost-effective and rapid alternative to whole genome sequencing of single bacterial strains [15]. We sequenced 137 Group G Streptococcus clinical strains in three pools and used SNVs and short INDELs discovered from the Pool-seq experiment to create custom mass spectrometry search database for identifying variant peptides in GGS. These bacteria have long been considered less pathogenic than GAS in humans, but over the past few decades GGSs has been recognised as a significant cause of a variety of infections [16].

We used MS/MS to analyse eight protein samples from strains that were used in the Pool-seq experiments, and utilizing the variant protein database we identified 385 variant peptides (90 of which were due to missed cleavages). Most of these variant peptides (71.2% and 93.5% respectively) could also be identified from conventional databases that contained 4 complete genomes and 26 assemblies from RefSeq. Our results reinforce the importance of proteogenomic approaches, particularly the incorporation of genomic driven variations to protein databases, for improving the identification of variant peptides and proteins from MS/MS analyses that might potentially be associated with certain characteristic of strains, such as invasiveness.

MATERIALS AND METHODS

Bacterial strains and culture conditions

For genomic sequencing, 90 *emm*-typed strains and 47 other clinical GGS strains were used. The *emm*-typed strains, originally collected in Pirkanmaa, Finland during years 1995-2004 and 2008-2009 [17], [18], derived from the bacterial culture collection of the National Institute of Health and Welfare. The strains which were not *emm*-typed were acquired from United Medix Laboratories Oy, and contained a random selection of strains extracted in the years 2014-2015 from all over Finland.

All strains were cultured over one or two nights at +37 °C on blood agar plates in 5% CO₂. DNA was isolated using UltraClean Microbial DNA Isolation Kit (MoBio) according to manufacturer's instructions except for a few modifications. Briefly a 10 µl loopful of bacteria was scraped from the blood agar culture plate as the starting material. 6 µl mutanolysin (1 mg/ml) was mixed additionally with the MicroBead solution and incubated for 60 min at +37 °C with gentle agitation of 850 rpm. The solution was transferred to a MicroBead tube, and 2 µl of RNase A (1 mg/ml) was added. At the final stage, DNA was eluted with 35 µl of solution MD5. Before pooling, the quality and the integrity of the DNA were checked using Nanodrop equipment (Thermo Scientific) and agarose gel electrophoresis. The sample concentrations were measured using Qubit 2.0 Fluorometer (Life Technologies). Subsequently, 3 pools were made out of the 137 strains, pools 1 and 2 contained 47 and 43 *emm*-typed strains and pool 3 consisted of 47 strains that had not been *emm*-typed, respectively. Each pool contained 200 ng of DNA from each strain. Sequencing was done with Illumina HiSeq high output mode with 2*125bp cycles at National Genomics Infrastructure hosted by SciLife Lab, Stockholm.

Samples for MS/MS

Eight randomly selected strains of GGS were used in the MS/MS analysis, and from each, three different protein fractions (one fraction of exoproteome proteins, one fraction of shaved surface proteome and one fraction with the proteins extracted from the entire cellular proteome the intact cells) were generated according to [19] with minor modifications. The eight strains were taken from -70°C milk/glycerol suspensions and grown at 37°C under an atmosphere of 5% CO_2 on blood agar plates. Individual colonies were taken for growth in 10 ml Todd–Hewitt broth at 37°C under 5% CO_2 atmosphere. The next day, the over-night cultures were diluted 1:20 in 10 ml of prewarmed Todd–Hewitt broth and they grew under the same conditions as before until they reached the exponential phase (OD 0.4-0.5).

Preparation of different protein fractions, protein digestion and peptide clean-up

For preparing the protein fractions, the mid-exponential cultures were split into twice 5ml and centrifuged at $2500 \times g$ for 10 min. The exoproteome proteins were isolated from the combined growth supernatants while one bacterial pellet was used for the entire cellular proteome and the other for shaved surface proteome isolations.

To generate the shaved surface proteome fraction the bacterial pellet was treated with $5\mu\text{g}$ of trypsin for 15 min at 37°C . The exoproteome fraction was recovered from the growth supernatant by 30kDa filtration (Amicon Ultra-15 Centrifugal Filter Units, 30 MWCO, Millipore). Proteins from the the entire cellular proteome were extracted by bead beating of the washed bacterial pellet using a FastPrep 24 (MP Biomedicals) 200 mg mm Zirconia/Silica Bead, Biospec without Triton as compared to the original publication.

Proteins were reduced with 50 mM dithiothreitol for 45 min at 37 °C, and alkylated with 25 mM iodoacetamide for 45 min at room temperature. An ammonium bicarbonate solution was used to dilute the urea to a concentration below 1.5 M (whole cell digests) and to adjust the pH to 8. Tryptic digestion was carried out at 37 °C over night. Peptide solutions were cleaned by C18 spin column purification (The Nest Group, Southborough, MA, USA) including an intensive washing step of ten repeats.

LC-MS/MS analysis

RP-HPLC online coupled to MS/MS (LC–MS/MS) analysis of peptides was carried out on a nano-HPLC system EASY-nLC 1000 (Thermo Fisher Scientific, Germany) connected to an Orbitrap Elite hybrid mass spectrometer (Thermo Fisher Scientific) with a nano-electrospray ion source (Thermo Fisher Scientific). Peptides were automatically loaded from the autosampler into a C18-packed precolumn (Acclaim pepmapTM 100, 2 cm × 75 µm × 2 cm, 3 µm, 100 Å, nanoViper; Thermo Scientific) and transferred onward to a C18-packed analytical column (Acclaim pepmapTM RSLC, 50 µm × 15cm, 2 µm, 100 Å, nanoViper; Thermo Scientific). Peptide separation started at 5% buffer B (98% acetonitrile, 0.1% formic acid and 0.01% trifluoroacetic acid in HPLC-grade water) for 5 min, followed by a 60-min linear gradient from 5 to 35% of buffer B at a flow rate of 300 nL/min. This gradient was followed by a 5-min gradient from 35 to 80% of buffer B, a 1-min gradient from 80 to 100% of buffer B, and a 9-min column wash with 100% buffer B, at a constant flow rate of 300 nL/min. Full MS scan (MS) was acquired in positive ion mode, with a resolution of 60 000 at normal mass range, in the high-resolution orbitrap mass analyzer using profile mode. The method was set to fragment the 20 most intense precursor ions with CID and the fragment ion scan (MS/MS) was acquired using centroid mode.

Data analysis

Raw MS files were analysed using Andromeda search engine of the MaxQuant software suite, version 1.6.0.16 [20], [21]. Four types of databases were used for searching: 'Variant_Pools' constructed by appending Pool-seq driven variant containing sequences to the reference genome ATCC 12394, a 'Single_Genome' database that contained protein sequences of the ATCC 12394 reference genome, and 'RefSeq_Complete' and 'RefSeq_Assembly' that contained 4 complete and 26 genome assemblies (including the 4 complete genomes) of *Streptococcus dysgalactiae* subsp. *equisimilis* (SDSE) downloaded from RefSeq (see Table 1). The default contaminants file and reverted decoy sequences were used and all 'Potential contaminant' and 'Reverse' hits were discarded. Specific digestion mode with Trypsin/p was used, with a maximum of 2 missed cleavages allowed. Match between runs was disabled. Carbamidomethyl (C) fixed modification, and Oxidation (M) and Acetyl (Protein N-term) variable modifications were set. First and main search tolerances of 20 ppm and 4.5 ppm respectively were applied. A minimum peptide length of seven amino acids was used. The default score cutoffs of 0 and 40 were used for unmodified and modified peptides respectively. Minimum number of razor plus unique peptide for protein group identification were set to 1 while minimum number of unique peptide was set to 0. FDR cut-off values of 0.01 were applied at the PSM, protein and site level. Protein groups only identified by PTMs and the associated peptides were not used in the downstream analyses.

Pool-seq based variant protein database creation

Coding region SNPs and INDELs that cause a non-synonymous effect, e.g. non-synonymous coding, missense, nonsense and frameshift mutations were chosen from the variants that

were concordantly mined using three variant calling tools, GATK's UnifiedGenotyper [22], SNVer[23] and Freebayes[24]. For each protein, the aligned sequence reads obtained from a Pool-seq analysis (of 137 in-house GGS strains) that span the protein were analysed to determine which of these final set of variants could be found in the reads. Unique set of variants from these reads were then inserted to the nucleotide sequence of the protein with an aim to capture the strain-wise differences and the protein was *in silico* digested (i.e. cleaved after each R and K, when not followed by P) after translating to amino acid sequence. The resulting variant peptides which were unique and more than 3 amino acids long together with their right and left flanking peptides (for cases of missed cleavage) were written as a separate entry. The unique order of the variants in a particular entry was included in the headers of the fasta databases using bitwise flags. The 'wild type' protein sequences were also written to the databases (see Figure 1 for the workflow of the variant database creation from Pool-Seq). The source code for the python script used for creating the variant protein database can be accessed from the following github link:

https://github.com/RigbeGW/poolseq_variantdb

Minimizing false positives in variant peptide discovery

Since the matches to variant peptides in the variant databases could contain more false positives than matches to the 'wild type' peptides[2], besides the global FDR filtering, we applied stricter filtering on the variant matches using PEP based local fdrs (<0.01)[25]. A custom python script was used to perform the local fdr filtering. Initially, the target variant peptide and decoy variant hits were selected from the identified peptides, and they were sorted by their PEP values in ascending order. Then the proportion of the decoy and target variant peptide hits was iteratively calculated, at each iteration checking if the proportion

was less than or equal to the cut off value of 0.01. Variant hits beyond the cut off value were discarded. In addition, we used variants identified from analysis of a few publicly available SDSE sequences from the European Nucleotide Archive (ENA) to determine whether further filtering is required based on, for instance, confounding post-translational or chemical modifications.

ACCEPTED MANUSCRIPT

RESULTS

Andromeda score and PEP distributions of the protein databases

Various previous studies [26]–[29] have established that inflated database size and sequence redundancy leads to decreased sensitivity, for instance, by resulting in larger Posterior Error Probability (PEP) values for six-frame translated databases where 5 out of the 6 translations will contribute to larger decoy database sizes [26]. The variant protein databases contain multiple, almost identical, peptide sequence entries per protein that could confound identification similar to the six-frame translated databases, even though the extra sequences are not essentially wrong sequences. Therefore, to investigate if the Pool-seq based variant databases exhibit the same characteristics as the six-frame translated databases, we compared the distributions of the scores and the PEPs of the variant database with the conventional databases.

In terms of the search scores of the peptides identified, there was not a notable difference among the databases (Figure 2A), indicating that the type (those containing 'variant' versus 'wild type' peptides) and size of the databases (ranging from thousands to ten thousand sequences) used did not affect the search score of the peptides identified. The introduction of variants did not lead to the identification of peptides with extremely high or low search scores. Moreover, the target-decoy score distributions were almost identical in all the databases and target scores had right-tailed distributions in all cases (Figure 2B). With regard to the PEP values, unlike the six-frame translated databases in the study by Blakeley et al. [26], the values were not inflated in the variant databases (Figure 2C).

Peptide/protein identifications

The proteomics data from the three different fractions were combined for this project. In addition to the 'wild type' entries of the 1887 coding genes in the reference genome, 36725 variant entries were incorporated to the variant database, leading to ~18-fold increase in the database sizes. In total, 385 variant peptides and 296 variant protein groups (that is protein groups associated with one or more variant peptides) were identified from the three pools. This is after the filtering steps where all the variant peptides passed the local fdr filtering while two variant peptides, 'NLVNGNWK' and 'AEQAWNDLFAAYQAEYPELAAEYQK', containing single mutations ('D12N' and 'E320Q' respectively) not identified from the ENA sequences were discarded as the aberrations could potentially be due the deamidation of Asn(N) and Gln(Q) residues in the peptides resulting in the mass shift of 0.98402. Furthermore, 13 variant peptides that were associated with protein groups that were only identified by PTMs ('only identified by site') were removed.

Majority of these variant peptides (71.2%; 93.5%) and protein groups (55.4%; 80.7%) were also identified from the 'RefSeq_Complete' and 'RefSeq_Assembly' databases respectively (Figure 3). On the other hand, more 'wild type' peptides and protein groups were identified from these conventional databases than the variant databases (see Supplementary Figure S1 of the supplementary data). The False Discovery Rates(FDRs) at the peptide and protein group level were similar between the variant and conventional databases, and were below the threshold value of 0.01 (Figure 3), even though they were somewhat inflated at the protein level due to the naïve target-decoy method employed[30].

At the genomic level there were 17711 missense and nonsense mutations, relative to the reference strain, that were identified from the 137 GGS strains used in the three Pool-seq

samples. With the exception of a few of the GGS genes (145/1887), all genes had one or more of these missense or nonsense variants, with some genes exhibiting very high or very low variant rates (Figure 4A, second innermost circle). The genes with the highest variant rates were mostly hypothetical proteins (see Supplementary Table S1 of the supplementary data). Out of the 951 protein groups identified from the 'Variant_Pools' database, 296 of them were associated with one or more variant peptides (and of these 260 of them constituted of single sequence entries). The 260 variant protein groups containing only one sequence entry were from 166 unique GGS proteins of the reference ATCC 12394 genome (this is because we incorporated more than one variant sequence entry per protein). Majority (153/166) of these proteins were core genes, as identified from the 4 complete genomes using PanOCT (v 3.23)[31] in which a cluster was assumed to be core if it contained at least one gene from each of the 4 complete genomes, and had a mean variation rate of ~ 0.03 (Figure 4A, see the two outermost circles). The genes with the largest number of variant peptides (without considering missed cleavages) included those that code for potential virulence factors such as the M protein and the streptokinase G and also those that produce important enzymes such as transketolase and class I and II aminoacyl-tRNA synthetases (Figure 4A, the outermost histogram) (see Supplementary Table S2 for a list of these proteins and their associated variant peptides; the proteins with the largest number of variant peptides are highlighted with orange).

The proteins containing variant peptides (referred to as variant proteins henceforth) were involved in various cellular functions such as translation, biogenesis and the transportation and metabolism of various molecules including carbohydrates and amino acids (Figure 4B). Majority (50.9%) of the proteins of the ATCC 12394 GGS strain were predicted to be cytoplasmic; this was also true for majority of the variant proteins (Figure 4C).

Variant rate of proteins identified from the variant database

When searching the MS/MS data of the 8 strains using the proteogenomic variant database, 12 of the proteins had only variant peptide hits (and only one of them, 'WP_003061134.1' was identified when using only the Single_genome database) while 481 proteins had only 'wild type' hits, and 170 of the proteins had both 'wild type' and variant hits. The 8 samples and the 3 fractions contained different proportions of these proteins (see Table2). The 12 proteins had higher variant rates (with a mean of 0.040) compared to the 481 proteins (the mean variant rate of a random sample of 12 proteins done 100 times ranged from 0.007 to 0.023). The proteins with variant hits were those that had a statistically significantly higher variant rate than those that only had 'wild type' hits (Figures 5A-C). The 12 variant proteins had various important functions, such as regulation of DNA-binding response and transcription repression (highlighted in green in Supplementary Table S2).

DISCUSSION

Protein databases play a central role in most mass spectrometry based proteomic experiments since peptide/protein identifications depend on the sequences that are available in the databases. In a typical liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis, the extracted proteins are enzymatically (e.g. trypsin) digested into peptides, which are then separated and analysed. The peptides are identified by matching the obtained MS/MS spectra against theoretical peptide spectra from *in silico* digestion of the protein sequences in the used database and scored by database search engines. This results in peptide-spectrum matches (PSMs), with the highest scoring peptide taken as a match to the spectrum. Therefore, if the database is not complete, i.e., a sequence is not present in the database, the peptide that contains variants or mutations will not be detected[32]. Incomplete databases containing only 'wild type' sequences from a few genomes fail to identify variant peptides that might be associated with extracellular immune evasion or invasiveness, especially in highly diverse species such as beta-hemolytic streptococci. Also, the availability of sequences of representative strains of a certain taxonomic group in databases is paramount for the correct identification and differentiation of closely related species[33].

Databases that contain mutations from existing databases[2], [3], [34] as well as from RNA-seq [35], [36] have been created to mine variant peptides, for instance related to diseases such as cancer in humans[37], [38], and recently in *M. tuberculosis* bacteria to investigate strain variation between two isolates[39]. A proteogenomic approach that utilized spectral libraries from a six-frame translation of the composite genome of 34 *S. pyogenes* M1 bacterial strains was used to identify proteins affected by mutations in a DIA-MS study[40].

But, to our knowledge, our study is the first to create variant databases from a relatively large number of pooled GGS strains.

Our approach of incorporating more than one version of a variant peptide from the Pool-seq enabled us to identify the same tryptic variant peptides but with different mutations, for instance from the gene that codes for the M protein, which is known to be variable. In total, there were 10 proteins that had two such variant peptides identified from and two of the proteins also had a 'wild type' identified from besides the two variant peptides (without considering missed cleavages). Even though this approach will increase the database size, the increment will not be drastic since the variations included are the observed ones (and not from six-frame translations for instance) and most bacterial genomes are small in size. Furthermore, the increase in the database size and the redundancy of the sequences does not appear to affect the identification sensitivity unlike, for instance, in the six-frame translation based databases [26]. Indeed, the variant peptides combined to the 26 assemblies resulted in almost equal number of total peptide identifications as the 26 assemblies (7944 vs 7970) even if the addition of the variant peptides resulted in ~2-fold (200%) increase in the database size of the 26 assemblies. Moreover, the PEP values were not inflated unlike in the six-frame translation case. This indicates that having the wrong sequences in the database has a more extreme effect on the sensitivity of the searches than the database size or the redundancy of the sequences in the database.

Another evidence on the importance of variant databases was that certain proteins that only had variant peptide hits were those that had higher variant rates than the proteins that only had 'wild type' peptide hits. Most of the variant peptides clearly existed in other strains since we identified them from the four complete genomes and the 26-assemblies as well;

these peptides would not have been discovered if we had utilized only a single reference genome that does not contain variant peptide sequences. Among the 12 proteins that only had variant peptide hits, only one protein (WP_003061134.1) had 'wild type' peptides identified from when the 'Single_Genome' database was used for searching. These 12 proteins included various protein families including Cupin which is known to be a diverse superfamily and proteins that could play a role in adaptive mechanisms and thus virulence of the bacteria, such as a DNA-binding response regulator and a ribosome assembly RNA binding protein (Supplementary Table S2, highlighted in green).

CONCLUSION

Even though for the eight samples analysed we managed to identify majority of the variant peptides from the composite genome conventional databases, the increase in the variant peptide identification ranged from 71.2% to 93.5% as the number of genomes increased from 4 to 26. This is an indication of the intrinsic strain diversity. We therefore anticipate that large number of strains need to be sequenced for optimal identification of peptides/proteins from large samples of highly diverse bacterial species. For species with good quality reference genomes available, Pool-seq can be used to sequence large number of strains comparatively inexpensively and incorporate the variations into custom databases for improved identification. Use of variant protein databases with conventional genomes might enable the identification of novel variant peptides that could for instance be responsible for virulence in certain strains.

ACKNOWLEDGMENTS

We would like to thank Fitsum Tamene and Dawit A. Yohannes for their valuable discussions. We also thank Dr. Jaana Syrjänen for her input during the early steps of the study, and Susanna Vähäkuopus for *emm*-typing and Kati Räisänen and Anna Muotiala for GGS strain selection. This study was partially supported by Tekes funding for SalWe GetItDone program (to TSJ), and the Academy of Finland (Terveystieteiden tutkimuksen Toimikunta, Grant Numbers: 288475, 294173 to MV, Grant 255636 to JV, and Grant 255985 to JK). We acknowledge Finnish Centre for Scientific Computing (CSC) for the needed CPU and data storage quotas.

TSJ and MV are equally contributing last authors while ND and CK are equally contributing second authors.

REFERENCES

- [1] J. Armengaud, J. Trapp, O. Pible, O. Geffard, A. Chaumot, and E. M. Hartmann, "Non-model organisms, a species endangered by proteogenomics," *J. Proteomics*, vol. 105, pp. 5–18, Jun. 2014.
- [2] M. K. Bunger *et al.*, "Detection and Validation of Non-synonymous Coding SNPs from Orthogonal Analysis of Shotgun Proteomics Data," *J. Proteome Res.*, vol. 6, no. 6, pp. 2331–2340, Jun. 2007.
- [3] S. Schandorff *et al.*, "A mass spectrometry-friendly database for cSNP identification," *Nat. Methods*, vol. 4, no. 6, pp. 465–466, Jun. 2007.
- [4] M. Brosch *et al.*, "Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and 'resurrected' pseudogenes in the mouse genome," *Genome Res.*, vol. 21, no. 5, pp. 756–767, May 2011.
- [5] S. Tanner *et al.*, "Improving gene annotation using peptide mass spectrometry," *Genome Res.*, vol. 17, no. 2, pp. 231–239, Feb. 2007.
- [6] R. J. Olsen, S. W. Long, and J. M. Musser, "Bacterial genomics in infectious disease and the clinical pathology laboratory," *Arch. Pathol. Lab. Med.*, vol. 136, no. 11, pp. 1414–1422, Nov. 2012.
- [7] R. Craig and R. C. Beavis, "TANDEM: matching proteins with tandem mass spectra," *Bioinformatics*, vol. 20, no. 9, pp. 1466–1467, Jun. 2004.
- [8] D. M. Creasy and J. S. Cottrell, "Error tolerant searching of uninterpreted tandem mass spectrometry data," *PROTEOMICS*, vol. 2, no. 10, pp. 1426–1434, Oct. 2002.
- [9] J. D. Jaffe, H. C. Berg, and G. M. Church, "Proteogenomic mapping as a complementary method to perform genome annotation," *PROTEOMICS*, vol. 4, no. 1, pp. 59–77, Jan. 2004.
- [10] A. I. Nesvizhskii, "Proteogenomics: concepts, applications and computational strategies," *Nat. Methods*, vol. 11, no. 11, pp. 1114–1125, Oct. 2014.
- [11] E. Brunner *et al.*, "A high-quality catalog of the *Drosophila melanogaster* proteome," *Nat. Biotechnol.*, vol. 25, no. 5, pp. 576–583, May 2007.
- [12] N. E. Castellana *et al.*, "An automated proteogenomic method uses mass spectrometry to reveal novel genes in *Zea mays*," *Mol. Cell. Proteomics MCP*, vol. 13, no. 1, pp. 157–167, Jan. 2014.
- [13] H. Tettelin, D. Riley, C. Cattuto, and D. Medini, "Comparative genomics: the bacterial pan-genome," *Curr. Opin. Microbiol.*, vol. 11, no. 5, pp. 472–477, Oct. 2008.
- [14] H. Tettelin *et al.*, "Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial 'pan-genome,'" *Proc. Natl. Acad. Sci.*, vol. 102, no. 39, pp. 13950–13955, Sep. 2005.
- [15] R. G. Weldatsadik *et al.*, "Sequence analysis of pooled bacterial samples enables identification of strain variation in group A streptococcus," *Sci. Rep.*, vol. 7, p. 45771, Mar. 2017.
- [16] S. Rantala, "Streptococcus dysgalactiae subsp. equisimilis bacteremia: an emerging infection," *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.*, vol. 33, no. 8, pp. 1303–1310, Aug. 2014.
- [17] S. Rantala, J. Vuopio-Varkila, R. Vuento, H. Huhtala, and J. Syrjänen, "Predictors of mortality in beta-hemolytic streptococcal bacteremia: a population-based study," *J. Infect.*, vol. 58, no. 4, pp. 266–272, Apr. 2009.
- [18] S. Vähäkuopus, R. Vuento, T. Siljander, J. Syrjänen, and J. Vuopio, "Distribution of emm types in invasive and non-invasive group A and G streptococci," *Eur. J. Clin. Microbiol. Infect. Dis.*, vol. 31, no. 6, pp. 1251–1256, Jun. 2012.
- [19] R. Karlsson *et al.*, "Strain-level typing and identification of bacteria using mass spectrometry-based proteomics," *J. Proteome Res.*, vol. 11, no. 5, pp. 2710–2720, May 2012.

- [20] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann, "Andromeda: a peptide search engine integrated into the MaxQuant environment," *J. Proteome Res.*, vol. 10, no. 4, pp. 1794–1805, Apr. 2011.
- [21] J. Cox and M. Mann, "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification," *Nat. Biotechnol.*, vol. 26, no. 12, pp. 1367–1372, Dec. 2008.
- [22] M. A. DePristo *et al.*, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nat. Genet.*, vol. 43, no. 5, pp. 491–498, 2011.
- [23] Z. Wei, W. Wang, P. Hu, G. J. Lyon, and H. Hakonarson, "SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data," *Nucleic Acids Res.*, vol. 39, no. 19, pp. e132–e132, 2011.
- [24] E. Garrison and G. Marth, "Haplotype-based variant detection from short-read sequencing," *ArXiv Prepr. ArXiv12073907*, 2012.
- [25] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble, "Posterior error probabilities and false discovery rates: two sides of the same coin," *J. Proteome Res.*, vol. 7, no. 1, pp. 40–44, Jan. 2008.
- [26] P. Blakeley, I. M. Overton, and S. J. Hubbard, "Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies," *J. Proteome Res.*, vol. 11, no. 11, pp. 5221–5234, Nov. 2012.
- [27] M. Fitzgibbon, Q. Li, and M. McIntosh, "Modes of inference for evaluating the confidence of peptide identifications," *J. Proteome Res.*, vol. 7, no. 1, pp. 35–39, Jan. 2008.
- [28] V. Granholm and L. Käll, "Quality assessments of peptide-spectrum matches in shotgun proteomics," *Proteomics*, vol. 11, no. 6, pp. 1086–1093, Mar. 2011.
- [29] K. Krug, A. Carpy, G. Behrends, K. Matic, N. C. Soares, and B. Macek, "Deep Coverage of the *Escherichia coli* Proteome Enables the Assessment of False Discovery Rates in Simple Proteogenomic Experiments," *Mol. Cell. Proteomics*, vol. 12, no. 11, pp. 3420–3430, Nov. 2013.
- [30] L. Reiter *et al.*, "Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry," *Mol. Cell. Proteomics MCP*, vol. 8, no. 11, pp. 2405–2417, Nov. 2009.
- [31] D. E. Fouts, L. Brinkac, E. Beck, J. Inman, and G. Sutton, "PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species," *Nucleic Acids Res.*, vol. 40, no. 22, p. e172, Dec. 2012.
- [32] A. I. Nesvizhskii, "Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data: Toward More Efficient Identification of Post-translational Modifications, Sequence Polymorphisms, and Novel Peptides," *Mol. Cell. Proteomics*, vol. 5, no. 4, pp. 652–670, Dec. 2005.
- [33] R. Karlsson *et al.*, "Proteotyping bacteria: Characterization, differentiation and identification of pneumococcus and other species within the Mitis Group of the genus *Streptococcus* by tandem mass spectrometry proteomics," *PLOS ONE*, vol. 13, no. 12, p. e0208804, Dec. 2018.
- [34] J. Li *et al.*, "A Bioinformatics Workflow for Variant Peptide Detection in Shotgun Proteomics," *Mol. Cell. Proteomics*, vol. 10, no. 5, p. M110.006536-M110.006536, May 2011.
- [35] G. M. Sheynkman, M. R. Shortreed, B. L. Frey, and L. M. Smith, "Discovery and Mass Spectrometric Analysis of Novel Splice-junction Peptides Using RNA-Seq," *Mol. Cell. Proteomics*, vol. 12, no. 8, pp. 2341–2353, Aug. 2013.
- [36] X. Wang *et al.*, "Protein Identification Using Customized Protein Sequence Databases Derived from RNA-Seq Data," *J. Proteome Res.*, vol. 11, no. 2, pp. 1009–1017, Feb. 2012.
- [37] J. Li, D. T. Duncan, and B. Zhang, "CanProVar: a human cancer proteome variation database," *Hum. Mutat.*, vol. 31, no. 3, pp. 219–228, Mar. 2010.
- [38] H. Xi, J. Park, G. Ding, Y.-H. Lee, and Y. Li, "SysPIMP: the web-based systematical platform for identifying human disease-related mutated sequences from mass spectrometry," *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D913–920, Jan. 2009.

- [39] T. Heunis *et al.*, "Proteogenomic Investigation of Strain Variation in Clinical Mycobacterium tuberculosis Isolates," *J. Proteome Res.*, vol. 16, no. 10, pp. 3841–3851, Oct. 2017.
- [40] L. Malmström *et al.*, "Quantitative proteogenomics of human pathogens using DIA-MS," *J. Proteomics*, vol. 129, pp. 98–107, Nov. 2015.
- [41] J. Huerta-Cepas *et al.*, "eggNOG4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D286–D293, Jan. 2016.
- [42] N. Y. Yu *et al.*, "PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes," *Bioinforma. Oxf. Engl.*, vol. 26, no. 13, pp. 1608–1615, Jul. 2010.

FIGURE LEGENDS

Figure 1. Workflow for creating the variant protein databases from Pool-seq data. The alignment files were interrogated to include all possible combinations of variants identified from the different strains in the pools, i.e., those variants in a protein that were concordantly identified by all four variant calling tools, and were found in the reads in a unique order were inserted in to the database as individual entries.

Figure 2. The Andromeda score and PEP distributions of the four GGS databases used in this study. A) The distribution of the Andromeda score of the four databases (from top to bottom: 'Single_genome', 'RefSeq_Complete', 'RefSeq_Assembly', 'Variant_Pools') is shown with the respective median score (dotted lines). B) The decoy (grey) vs target score distributions of the four databases and the respective median scores C) the PEP vs score plot of the four databases

Figure 3. Number of identified variant peptides and protein groups and their respective FDRs. The number of variant peptides and protein-groups identified (after local fdr filtering and discarding those identified 'only by site') from the three GGS protein databases for the eight GGS samples analyzed are shown. For the 'RefSeq_Complete' and 'RefSeq_Assembly' databases, the 'variant' peptides are those that are common between these databases and the variant peptides from the 'Variant_Pools' database. The FDR of the peptides and protein groups identified from each database were calculated using the naïve target-decoy strategy (which in the protein groups case could lead to inflated values [30]).

Figure 4. The ATCC 12394 GGS reference genome genes/proteins and variant peptides identified from them. We refer to proteins that had one or more variant peptide hits as 'variant proteins'. **A)** The outermost circle is made of histograms of the number of variant

peptides (without considering missed cleavages) discovered from each GGS protein, the 182 single proteins that mapped to unique peptides are considered. The proteins are colored based on their subcellular location (the names of proteins with the highest number of variant peptides, i.e. ≥ 4 are shown). The next circle is core GGS proteins identified from the 4 complete genomes using the PanOCT tool. The core proteins that had a variant peptide hit are colored by COG categories. The line graphs show the variant rates of all the GGS proteins based on the pooled GGS strains (black) and the average depth of coverage of sequence reads per each protein (light purple). Names of proteins with the highest variants rate (≥ 0.20) are depicted. **B)** The COG functional categories [41] of the core and non-core variant proteins. The colors correspond to the color codes of the second outermost circle in A). The core proteins are marked with the hatch patterns. **C)** The predicted subcellular location (using PSORTb v3.0.2 [42]) of the variant proteins. Most of the variant core as well as non-core genes inferred from the MS/MS analysis were predicted to be cytoplasmic. The colors correspond to the color codes of the outermost circle in A) and the core variant proteins are depicted with hatch patterns.

Figure 5. The variant rate of GGS proteins (calculated based on the 137 pooled GGS strains against the ATCC 12394 reference GGS strain) that had a match to the 'wild type' or variant peptides. A) The variant rate of GGS proteins using the 'Variant_Pools' databases and the results aggregated for the whole eight protein samples **B)** The variant rate of GGS proteins in the eight samples individually. **C)** The variant rate of GGS proteins in the extracts that targeted proteins in different cellular compartments, only four of the samples for which all the extracts could successfully be analyzed with MS/MS were used here to avoid bias. In all instances, the proteins that had a variant peptide hit (they could also have a 'wild type'

hit) were those that had a statistically significant higher variants rate than the proteins that only had 'wild type' peptide hits.

ACCEPTED MANUSCRIPT

Table 1. The protein databases used in this study. The entries are counts of the protein sequences included in the databases; in the 'Variant_Pools' case, in addition to the whole protein sequences of the ATCC 12394 genome, the variant peptides with their flanking 'wild type' peptides are included.

Database	Description	Entries	MS/MS	PSMs
Single_Genome	The ATCC 12394 Group G Streptococcus reference strain downloaded from RefSeq	1887	246738	56814
Variant_Pools	Variant sequences from in-house 137 pooled Group G Streptococcus strains besides sequences from the ATCC 12394 genome	38612	246738	57332
RefSeq_Complete	4 complete SDSE genomes from RefSeq (downloaded on 06/11/2017). Only unique sequences were used as there were many proteins with 100% sequence similarity	5100	246738	61006
RefSeq_Assembly	26 SDSE assemblies including the 4 complete genomes from RefSeq (downloaded on 06/11/2017). Only unique sequences were used	16496	246738	60109

Table 2. The number of proteins identified from the eight samples and the three fractions that had variant only, 'wild type' only and variant + 'wild type' hits.

	Samples								Fractions		
	1	2	3	4	5	6	7	8	exoproteome	shaved proteome	entire cellular proteome
Variant only	5	0	0	1	1	0	2	9	2	3	11
'Wild type' only	391	247	217	231	279	179	293	490	333	300	426
Variant + 'Wild type'	167	132	121	131	144	111	140	167	150	148	158

Significance

For bacterial populations, such as Group G Streptococcus (GGS), with substantial intra-species diversity, simultaneous sequencing of large numbers of strains and generation of proteogenomic databases from those aids in improving the discovery of peptides in mass spectrometric analyses. Therefore, generation of proteogenomic variant protein databases from Pooled sequencing experiments can be a cost-effective method to complement conventional databases and discover subtle strain wise differences.

Highlights

- Variants mined from Pool-seq experiments are used for creating proteogenomic database.
- By including more than one version of variant peptide sequences of a protein that contain different mutations, the completeness of the proteogenomic database was improved without extremely inflating the database size.
- Certain proteins with high variant rates could be identified only after variant peptides derived from the proteogenomic approach or multiple genomes were included in the search databases.
- Andromeda score distributions and posterior error probabilities of the proteogenomic database are similar to conventional databases.
- Pool-seq driven proteogenomic databases are cost efficient methods to discover novel variant peptides that can complement conventional databases.

Figure 1

A) **B)** **C)** **Figure 2**

Figure 3

A)

B)

C)

Figure 4

A)

B)

C)

Figure 5