# Quantitative proteogenomics of human pathogens using DIA-MS.

Malmström, Lars; Bakochi, Anahita; Svensson Birkedal, Gabriel; Kilsgård, Ola; Lantz, Henrik; Petersson, Ann Cathrine; Hauri, Simon; Karlsson, Christofer; Malmström, Johan

*Published in:*
Journal of Proteomics

*DOI:*
10.1016/j.jprot.2015.09.012

2015

*Document Version:*
Peer reviewed version (aka post-print)

# Quantitative proteogenomics of human pathogens using DIA-MS

Lars Malmström[1], Anahita Bakochi[2], Gabriel Svensson[2], Ola Kilsgård[2], Henrik Lantz[3], Ann Cathrine Petersson[4], Simon Hauri[2], Christofer Karlsson[2] and Johan Malmström[2]

[1]S3IT, University of Zurich, Zurich, Switzerland

[2]Division of Infection Medicine, Department of Clinical Sciences Lund, Lund University, Lund, Sweden

[3]Department of Medical Biochemistry and Microbiology/BILS, Uppsala University, Uppsala, Sweden

[4]Department of Clinical Microbiology, Division of Laboratory Medicine, Region Skåne, Lund, Sweden

Corresponding author: lars.malmstroem@uzh.ch

## Abstract

The increasing number of bacterial genomes in combination with reproducible quantitative proteome measurements provides new opportunities to explore how genetic differences modulate proteome composition and virulence. It is challenging to combine genome and proteome data as the underlying genome influences the proteome. We present a strategy to facilitate the integration of genome data from several genetically similar bacterial strains with data-independent analysis mass spectrometry (DIA-MS) for rapid interrogation of the combined data sets. The strategy relies on the construction of a composite genome combining all genetic data in a compact format, which can accommodate the fusion with quantitative peptide and protein information determined via DIA-MS. We demonstrate the method by combining data sets from whole genome sequencing, shotgun MS and DIA-MS from 34 clinical isolates of *Streptococcus pyogenes*. The data structure allows for fast exploration of the data showing that undetected proteins are on average more amenable to amino acid substitution than expressed proteins. We identified several significantly differentially expressed proteins between invasive and non-invasive strains. The work underlines how integration of whole genome sequencing with accurately quantified proteomes can further advance the interpretation of the relationship between genomes, proteomes and virulence.


## Highlights

- 34 sequenced genomes and corresponding shotgun and DIA-MS measurements
- Construction of a composite genome for fast data integration
- Quantitative DIA-MS of the conserved and non-conserved peptide pool across all strains

## Significance

This paper outlines a novel strategy for combining genomics and quantitative DIA-MS proteomics data. We demonstrate a DIA-MS-based proteogenomics strategy for quantifying conserved and non-conserved peptides across clinical isolates of *Streptococcus pyogenes* from non-invasive and invasive infections. We suggest a strategy for constructing a composite genome that is optimal for MS data integration and querying. The work demonstrates how biological insight can be gained from the integration of the different data types.


## Keywords

quantitative mass spectrometry, proteogenomics, data integration, DIA, *Streptococcus pyogenes*


## Abbreviations

FDR, false discovery rate; WGS, whole genome sequencing; SNP, single nucleotide polymorphism; DIA, data-independent analysis; DDA, data-dependent acquisition

## Introduction

In proteogenomics, mass spectrometry (MS)-based proteomics is used as a supplement to genomic data by adding a level of information to the interpretation of genomic sequences[1]. In this context, MS is particularly relevant in microbiology where a large number of genomes are sequenced regularly[1]. Comparative genomic analysis of microbial genomes has revealed compelling evidence that some pathogens undergo rapid genomic adaption to increase fitness in their host[2]. The influence of single nucleotide polymorphisms (SNPs) on the molecular phenotype may be substantial, leading to increased virulence or the ability to survive and thereby cause disease[3]. Other events such as DNA methylation[4] and phosphorylation[5] can modify how the genome is translated, leading to increased virulence. Small genomic changes can influence survival and virulence in several ways, for example by activating/inactivating regulatory systems controlling part of the proteome expression[3], disrupting protein-protein interactions[6] or by increasing or decreasing the affinity between transcription factors and their target promoters[7]. The rapid increase in the number of genomes provides the opportunity to use matching genotype and strain to investigate how sets of SNPs alter proteome homeostasis. However, matching genotype and strain information in MS-based proteomics presents considerable challenges.

MS-based proteomics experiments rely on a protein database to provide the ground truth, i.e. information on all the possible tryptic peptides that can be derived from a given genome. The ideal protein database should contain all required information while remaining as small as possible. In the case of proteogenomics, this problem becomes amplified if approached naively by concatenating the protein database from each genome as it becomes challenging to select a particular protein if many similar proteins exist in the database[8]. On the other hand, searching each MS data file against its appropriate genome is standard procedure; the challenge here is to combine the independent searches without increasing the false discovery rate (FDR) dramatically[9]. The reason for the increase in FDR is that the correct proteins are, to a large extent, the same

across the different searches, whereas false hits are not and will ultimately represent a larger fraction in the combined list. Another related challenge is the mapping of all identified peptides to a set of orthologous proteins. For a given ortholog there may be peptides that are completely conserved whereas other peptides may differ in one or more amino acids. The challenge in mapping identified peptides to a set of orthologs introduces problems with accurate protein quantification if non-conserved peptide species are included for quantification. In theory, the conserved peptide sequences can be used to reference peptides necessary for protein quantification whereas the non-conserved peptides provide an opportunity to relatively quantify the presence of a certain protein species in a complex mixture.

In contrast to shotgun MS and traditional database searches, DIA-MS provides new opportunities to use the differential degree of peptide conservation to further explore the rapid increase in sequenced genomes. DIA-MS was originally developed to expand the detectable dynamic range and does not use real-time ion selection-based precursor scans[10]. This can be accomplished by interrogating predetermined *m/z* ranges by either fragmenting all ions entering the mass spectrometer[11-14] or by dividing the full *m/z* range into fixed smaller isolation windows[15-18]. Several of the developed DIA methods differ in how subsequent data analysis is performed[10]. In 2012, Gillet et al showed that the identification of peptides from DIA experiments can be accomplished via spectral libraries constructed from previously acquired shotgun MS[17], nowadays implemented in search algorithms[19]. In general, the DIA methods are associated with increased signal-to-noise ratios, increased sensitivity and increased specificity based on peptide fragmentation[15], and have shown improved reproducibility compared to a data-dependent acquisition (DDA) counterpart[20,21]. Importantly for proteogenomic strategies, the spectral libraries can easily include all observed SNPs in a given strain and thereby remove the problem with large FASTA databases or difficulties with controlling FDR resulting from concatenating several individual searches, provided that the peptides are represented in the spectral library. Spectral libraries can be constructed based on the level of peptide conservation and this enables quantitative analysis of both conserved and non-conserved peptides, which can be used to determine protein abundance

or for quantitative monitoring of specific SNPs across several strains. In the work presented here we aimed at providing a general quantitative proteogenomics strategy for exploring the consequences of genome adaptation at the proteome level using the important Gram-positive bacterium *Streptococcus pyogenes* as a model system.

*S. pyogenes* is one of the most common and important human pathogens[22,23] and is responsible for mild diseases such as pharyngitis, erysipelas and impetigo as well as severe diseases such as streptococcal toxic shock syndrome and necrotizing fasciitis[24]. Annually, *S. pyogenes* causes over 616 million cases of pharyngitis and 111 million cases of impetigo[24]. It encodes many well-characterized virulence factors, including surface-bound M protein and M-like proteins, hyaluronic acid capsules, adhesins, surface-bound collagen-like proteins, superantigenic exotoxins, and numerous secreted and extracellular proteins[25]. Antigenic differences in the hypervariable region of the M protein are the basis for the Lancefield serological classification of *S. pyogenes* with over 200 identified serotypes to date[26]. Strains of certain serotypes are epidemiologically associated with particular clinical syndromes where serotype M1 and M3 have frequently, but not exclusively, been isolated from patients with severe invasive disorders and infections with these serotypes are associated with increased mortality[27]. The extent to which genomic adaptation observed in invasive *S. pyogenes* strains results in altered proteome composition and increased virulence remains unclear.

In this study, we collected 34 clinical strains of *S. pyogenes* serotype M1, sequenced all the genomes and then analysed full proteome digests of all strains with DDA-MS and DIA-MS. We generated a so-called composite genome that contains all the genetic information of the strains and derived all potential tryptic peptides containing between 7 and 50 amino acids that this composite genome could theoretically encode. We constructed a spectral library by searching the shotgun MS data against the peptide database. The spectral library was then used to analyse the DIA-MS data to generate a quantitative expression matrix. We constructed a data structure that allowed us to analyse the three different data sets in light of each other, highlighting the relevance of several known and putative virulence factors. The proposed workflow can be extended

154 to other bacterial species, demonstrating how DIA-MS can further facilitate the
155 interpretation of proteome changes based on genomic information.

# Methods

## Isolates

Emm1 GAS were isolated between April and May 2012 at the accredited diagnostic laboratories of clinical microbiology, Division of Laboratory Medicine, Lund, Sweden. Isolates from sterile sites were sent to the laboratories as part of routine health care whereas isolates from throat swabs were collected as a part of a surveillance programme from selected geographically scattered primary care units in southern Sweden. Isolates were characterized as group A streptococci through agglutination and were typed through PCR and sequencing essentially as described[28]. The modified primers *emm* for 5´-GCT TAG AAA ATT AAA AAM MGG-3´[28] and CDC-R 5´-GCA AGT TCT TCA GCT TGT-3´ (http://www.cdc.gov/streplab/protocol-emm-type.html) were used. *Emm* types were assigned through the type-specific database at http://www2a.cdc.gov/ncidod/biotech/strepblast.asp. In total, 34 *S. pyogenes* M1 strains were subdivided into strains responsible for non-invasive conditions, in this case tonsillitis (n=18), and invasive conditions such as necrotizing fasciitis, toxic shock syndrome and/or endomyometritis (n=16).

## Whole genome sequencing

Genomic DNA was extracted from the *Streptococcus pyogenes* isolates using a silica-membrane spin column kit (Macherey-Nagel). In brief, overnight cultures (3.5 mL) were harvested by centrifugation at 3500 x g, resuspended in ice-cold 70% ethanol and incubated at -20 °C for 20 minutes. The cell wall was digested by resuspending the bacteria in 25 mM Tris-HCl, 2 mM EDTA, 1% (v/v) Triton X-100 containing 20 mg/mL lysozyme and 250 units/mL mutanolysin (both enzymes from Sigma-Aldrich) followed by incubation at 37 °C for 2 hours. Genomic DNA was released from the bacteria by resuspending the bacteria in a buffer containing SDS and 20 mg/mL proteinase K and overnight incubation at 56 °C. Subsequent DNA purification was performed according to the manufacturer's protocol for the silica-membrane spin column kit. Preheated elution buffer (70 °C, 5 mM Tris-HCl, pH 8.5) was applied to the spin column

followed by incubation of the spin column at 70 °C for 10 minutes prior to elution of the DNA. The quantity and quality of the extracted genomic DNA were assessed using agarose gel electrophoresis, a microvolume spectrophotometer (Thermo Scientific) and a fluorescence-based quantification kit (Life Technologies). The purified genomic DNA was sent to GATC (Germany) for genomic library construction and sequencing on a HiSeq 2000 (Illumina) with 50 bp single reads.

## Whole genome assembly and annotation

Several assemblers were tried, and based on comparisons using Quast[29], Abyss 1.3.7 was chosen with a kmer size of 39[30]. This gave a good balance of a low number of misassemblies compared to the reference genome of strain MGAS5005 together with a high continuity of the genome assemblies. Annotation was performed using Prokka 1.10 with the rfam option[31].

## Sample preparation for mass spectrometry

The clinically isolated *S. pyogenes* strains were grown overnight on blood agar plates (37 °C, 5% CO2), after which single colonies were grown to mid-exponential phase in Todd-Hewitt broth (30 g/l) (Difco Laboratories) supplemented with yeast extract (6 g/l) (Difco Laboratories). The cells were harvested by centrifugation and resuspended in 50 mM Tris-HCl and 150 mM NaCl (Medicago) wash buffer, pH 7.6, to a final concentration of 2 x $10^9$ CFU/mL. After several washes the bacterial pellets were spun down and dissolved in ice-cold LC-grade water and heat-inactivated by incubation on a heat block for 5 min at 80 °C. The cells were transferred to lysing matrix tubes (Nordic Biolabs) containing 90 mg of 0.1 mm silica beads and homogenized using a cell disruptor (Beadbeater, FastPrep 96, MP Biomedicals). The cell debris was removed and the supernatants were denaturated in 10 M urea (Sigma-Aldrich) and 50 mM ammonium bicarbonate (ABC) (Fluka Analytical), followed by incubation with 1 µg trypsin (Sequencing Grade Modified Trypsin, Porcine, Promega, Madison, WI, USA) for 30 min at 37 °C for protein digestion. The samples were reduced using 500 mM Tris(2-carboxyethyl)phosphine (TCEP) (Sigma-Aldrich) for 60 minutes at 37 °C, and alkylated with 500 mM 2-Iodoacetamide (IAA) (AppliChem) for 30

min at room temperature in the dark. The samples were diluted in 250 µl 100 mM ABC and further digested with 1 µg trypsin (Sequencing Grade Modified Trypsin, Porcine, Promega) overnight. The trypsin was inactivated by adding formic acid (FA) until the pH was 2–3. In accordance with the manufacturer's instructions, C18 columns (Vydac UltraMicro Spin™ Silica C18 300Å Columns, #SUM SS18V, The Nest Group, Inc., Southborough, MA, USA) were used to clean up, desalt and concentrate the peptides in the samples. The solvents were removed in a SpeedVac and the peptides were resuspended in 50 µl buffer A (2% acetonitrile, 0.2% FA in LC-H20).

**LC-MS/MS analysis**

All peptide measurements were acquired on a Q Exactive Plus mass spectrometer (Thermo Scientific) coupled to an EASY-nLC 1000 ultra-high pressure liquid chromatography system (Thermo Scientific). Peptides were trapped on an Acclaim PepMap® 100 pre-column (Thermo Scientific, C18, 3 µm, 100 Å; ID 75 µm x 2 cm) and separated with a PepMap® RSLC EASY-Spray column (Thermo Scientific; C18 2 µm, 100 Å; ID 75 µm x 25 cm; heated to 45° C), using intelligent flow control for column equilibration and sample load at 800 bars. A linear gradient of between 5% and 35% acetonitrile in aqueous 0.1% formic acid was run for 120 min at a flow rate of 300 nl/min.

For shotgun MS, one full scan (resolution 70,000 @ 200 m/z; mass range 400–1600 m/z) was followed by 15 MS/MS scans (resolution 17,500 @ 200 m/z) of the most abundant ion signals (TOP15). Precursor ions were fragmented using HCD at a normalized collision energy of 30. Charge state screening was set to reject unassigned or singly charged ions. The dynamic exclusion time was set to 15 s and limited to 300 entries. AGC was set to 1e6 for both MS and MS/MS with ion accumulation times of 100 ms (MS) and 60 ms (MS/MS). The intensity threshold for precursor ion selection was $1.7e^4$.

For data-independent SWATH-like analysis, a full MS scan (resolution 70,000 @ 200 m/z; mass range 400–1200 m/z) was followed by 32 MS/MS fragmentation scans (resolution 35,000 @ 200 m/z) using an isolation window of 26 m/z (including 1 m/z overlap between windows). The precursor ions within each isolation window were fragmented using high-energy collision-induced

dissociation (HCD) at a normalized collision energy of 30. The automatic gain control (AGC) was set to 1e6 for both MS and MS/MS with ion accumulation times of 100 ms (MS) and 120 ms (MS/MS).

All samples injected contained a peptide standard for retention time calibration. The obtained raw files were converted to mzXML using the software tool ProteoWizard[32].

## Database searching and bioinformatics

The shotgun MS data was searched as described by Quandt et al[33]. In short, we used X! Tandem[34] and MyriMatch[35] with a precursor ion mass tolerance of 30 ppm and a fragment ion mass tolerance of 10 ppm allowing no miscleavages. The search results were statistically validated using Peptide Prophet[9]. The spectral library was created[36] and the resulting TraML file was used to analyse the DIA-MS data as described by Röst et al[19]. Both WGS and MS data were stored in openBIS[37] and processing related to MS was carried out using iPortal[38]. The DIA-MS data was statistically evaluated using pyProphet[39]. All data integration was carried out under the DDB framework[40,41], using non-normalized analytical tables[42].

# Results and discussion

## Workflow overview

The integration of several highly similar, but not identical genomes can result in complex data structures due to SNPs, insertions and deletions. This prohibits accurate fusion of peptide and protein information and results in long query times. At the same time, quantifying proteomes relying on a diversified peptide pool is not straightforward. To address these open computational challenges, we constructed an analysis workflow based on DIA-MS for improved integration of whole genome sequencing (WGS) and DIA-MS data as shown in Figure 1. The workflow contains seven distinct steps in which four of the steps in particular are highlighted – C) generation of a composite genome; D) search the shotgun data against the composite genome; E) construction of a spectral library; and F) generation of a quantitative peptide expression matrix – to detect consistent differences in trends in expressed and non-expressed proteins and regulated proteins between non-invasive and invasive strains.

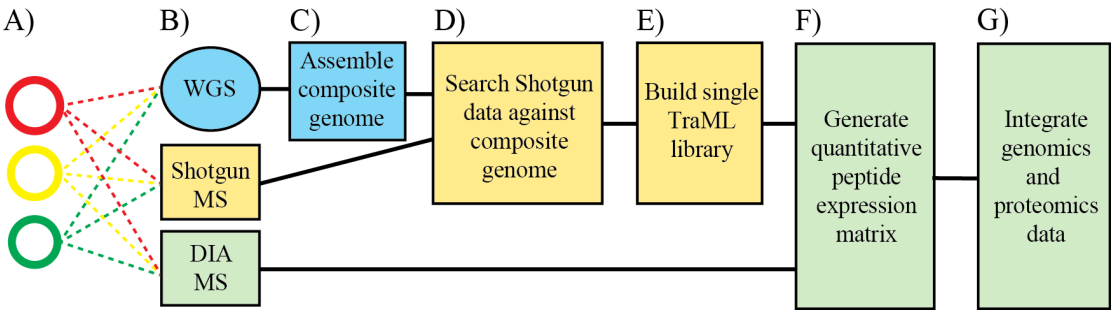



**Figure 1. Schematic overview of the outlined strategy** A) Genetically distinct clinical isolates, represented by coloured spheres, were B) digitized using genome sequencing, shotgun MS and DIA-MS. C) The individual genomes were assembled and aligned to create a composite genome, which was D) used to infer peptides from the shotgun MS data. E) A TraML spectral library file was created and F) the TraML file was then used to quantify peptides in all DIA-MS maps producing a nearly complete expression matrix. G) Peptides were mapped back to groups of orthologous proteins and integrated with the composite genome data.

## Generation of a composite genome

A particularly relevant feature when combining quantitative proteome data with genome data is information regarding the conserved and non-conserved peptides for a given open reading frame (ORF) used to assess protein quantification. Other, related information is the total number of silent and

expressed SNPs for that given ORF and whether the ORF is preceded by strain-specific changes in the intragenic region, which may influence the abundance level of that protein. To integrate all genetic information from the 34 genome sequences, we constructed a composite genome as follows: the Illumina reads from the 34 strains were assembled into contigs (see Figure 2 for a summary). The number of reads per strain varied from 6,513,248 to 50,054,680 with an average of 30,124,134 and resulted in 273 contigs on average (number of contigs range: 161–2632). We included the two poorly assembled genomes (g14 and g26) since the number of identified peptides from these genomes was similar to the others (7442 and 6835 peptides respectively, ranking 11 and 30 of 34, the range is 6213-8288, median 7174). This indicates that the assemblies over the expressed ORFs were of similar quality to other genome assemblies despite the high number of contigs. We choose NC_002737.1, a complete *S. pyogenes* genome of serotype M1, as reference and we refer to it as $M1_{ref}$ in the text below[43]. The contigs were ordered according to $M1_{ref}$ using Abacas[44] and we used Mugsy[45] to align the ordered contigs onto the $M1_{ref}$. The alignment was used to build a composite genome that contains all the genetic information from all strains (Fig. 2), stored in a denormalized analytical table for fast querying[42]. The consensus genome was 1,994,567 BP, only slightly larger than the average 1.8 MB member genomes, indicating a high degree of genomic similarity between the strains. Importantly, a consensus sequence was generated by a majority vote with random selection in cases of equal counts. We estimated the sequence conservation identically to Crooks et al[46]. The resulting composite genome is displayed in Figure 2b using CGView[47]. The composite genome is represented as the black line in the middle, and tracks on the inside represent features on the reverse strand and tracks on the outside features on the direct strand. Closest to the genome are the open reading frames (red and blue) followed by a track indicating all detected SNPs (purple and navy). The third track shows SNPs that lead to an amino acid substitution (fuchsia and lime). The zoom-in panel on the left shows the genomics region between 1953500 and 1953900 where the ORF coding for *hasA* is located (Fig. 1A). *hasA* has been implicated in the virulence mechanisms previously and its primary function is in the biosynthesis of the capsule[48].

Two additional tracks are shown on the global CGView panel to the right in Figure 2: the outermost track in orange represents conservedness and higher bars means less conserved. The track in green and lime represents the number of genomes that parts of the consensus genome are missing. The composite genome displays five larger regions of lower genome conservation (Fig. 2). The regions with a high degree of genome conservation are covered by all 34 member genomes and referred to as the core genome, corresponding to 85.6% of the composite genome. In total, 667 (0.039%) SNPs were detected in the core genome, whereas only 8.5% of the composite genome was exclusively present in a single member genome. The SNP rate was almost 22 times higher in the 5.9% of the composite genome that was outside the core but present in more than one genome. In these regions, 998 (0.85%) SNPs were detected in 117,119 base pairs, as can be visually detected in two high-density regions of SNPs in Figure 2. These two regions are associated with two of the regions with a lower level of genome conservation. Importantly, the composite genome data structure can allow faster and better integration with quantitative MS data, providing improved accessibility for the relationship between expressed proteins and the underlying genetic information.

The composite genome further supports the exploration of how the observed genomic alters the proteome homeostasis by providing an improved data structure for annotating the genome with both identified peptides and putative proteins found by Prokka[31]. This allowed us to separate the SNPs that are found within an ORF from SNPs found elsewhere. The ones found within an ORF were further divided into synonymous and non-synonymous. Figure 2 shows that SNPs that lead to amino acid substitution are rare compared to the total number of observed SNPs. As previously demonstrated, invasive strains tend to accumulate specific SNPs of relevance for invasive disease[3]. This system represents a suitable model system for establishing the DIA-MS-based proteogenomic strategy described next.

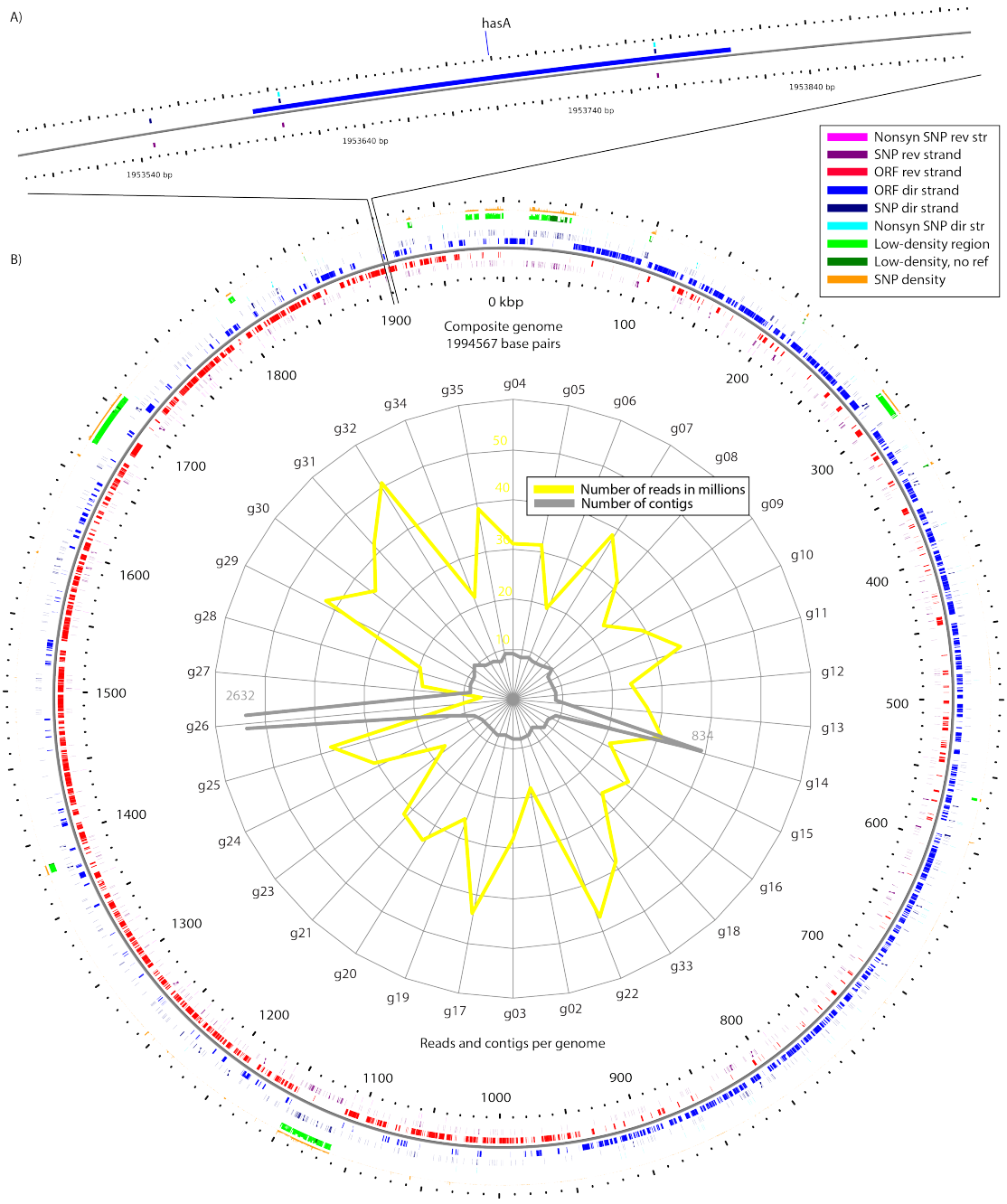



**Figure 2. Genome assembly and analysis**. A) A zoom in of the *hasA* loci located in the composite genome region 1,953,500–1,953,900. *hasA* has two non-synonymous SNPs. There is also one SNP in the intergenic region preceding *hasA*. B) The genomes were assembled individually and the quality of each assembly was assessed as displayed by the spider plot in the centre. The number of reads for each genome is displayed in yellow and the number of contigs is displayed in grey. One strain has a significantly lower number of reads and was difficult to assemble leading to 2632 contigs. Another genome had an average number of reads but still resulted in a poor assembly with 834 contigs. A composite genome was constructed by globally aligning the genomes. Each position in the meta-genome is represented in the CGView with the following tracks, from the inside out: fuchsia, non-synonymous SNPs; purple, SNPs; red, annotated genes all on the reverse strand. Blue, annotated genes on the direct strand; navy, SNPs; lime non-synonymous SNPs. Green and light green is the 1-density where a thicker line means fewer genomes are aligned at this position. Darker green indicates that the $M1_{ref}$ genome is present. The orange track indicates 1-conservedness. A thicker line means less conserved.

## Generation of a spectral library for DIA-MS analysis

One important step of the proposed quantitative proteogenomics strategy is the construction of a spectral library that contains all detectable peptides including peptide sequences conserved across all strains as well as the non-conserved peptides. Here, we constructed the peptide spectral library by translating all members of the composite genome in six frames and selecting all fully tryptic peptides between 7 and 50 amino acids in length resulting in a total of 223,952 unique peptide sequences[49]. These unique peptide sequences were used to search the 34 strains grown in duplicate resulting in 68 shotgun MS experiments using X! tandem[34], Myrimatch[35] and peptideProphet[9] on a previously published portal[33]. The search results were used to construct a spectral library in the TraML format as previously described[50] (Fig. 1c–d). In total, this effort generated a spectral library for *S. pyogenes* containing 14,633 precursors corresponding to 11,552 unique peptide sequences at 1% peptide-level FDR, representing 5.1% of the total 223,952 unique peptide sequences that can be potentially produced from all the 34 genomes. The relatively low coverage is not surprising since the vast majority of the putative peptides are never expressed. For example, only one out of six reading frames is actually used for any stretch of DNA. Of course, intergenic DNA and proteins not expressed under the tested condition cannot be detected either for obvious reasons.

## Generation of a quantitative expression matrix

One of the biological replicates from the 34 SWATH-like MS DIA-MS data sets was analysed with OpenSWATH[19] using the spectral library as the source of precursors to consider. The resulting expression matrix contained quantitative values for 6880 peptides over the 34 genomes. Figure 1e shows a schematic overview of how the expression matrix was produced. We generated profiles for all 223,952 unique peptide sequences derived from the composite genome based on their presence or absence in the 34 member genomes, resulting in a total of 680 profiles. Out of the 680 profiles, 37 were associated with at least one detected peptide as shown in the heat map in Figure 3. The histogram to the left shows the number of peptides associated with each profile; the bar graph to the right displays the fraction of the identified peptides for the profile. The vast

majority of the peptides are conserved across all strains. However, only four per cent of these peptides were identified. The most abundant profiles were followed by a decreasing number of peptides associated with the remaining profiles. The heat map (Fig. 3) reveals that the two genomes with high numbers of contigs (Fig. 2b) make a considerable contribution to the expression matrix. The columns in the heat map are ordered so that the invasive strains are to the left and the non-invasive ones to the right. No obvious trends of peptides that distinguish the two groups can be observed, indicating that detection of a coding SNP has a low correlation with virulence. In contrast, the quantitative peptide data is more discriminative (Fig. 4), showing that there are two main groups of bacteria; one of these groups is divided into two sub-groups and the other main group is divided into four sub-groups for a total of six sub-groups. Non-invasive bacteria make up three of these sub-groups up to 100% and only invasive bacteria make up two groups. The last group contains one non-invasive bacterial isolate among the five invasive isolates. We used pvclust, an algorithm using multiscale bootstrap resampling (n=1000, default clustering method=average, default distance measure=correlation) to assess significance of a hierarchical clustering, to indicate clusters with an approximate unbiased p-value of 0.01 as indicated by the asterisks in Figure 4. As these strains are grown under identical conditions, the observation that, on average, invasive strains are more similar to each other than non-invasive strains indicates that the underlying genomes are driving these differences. On the other hand, the classification of the strains is not perfectly subdivided into the two groups. These results show that in some cases proteome expression patterns for some invasive strains are more similar to non-invasive strains than other invasive stains. The absence of a clear trend in the heat map in Figure 3a indicates that it is not sufficient to measure the abundance level of the non-synonymous SNPs to make assessments on whether or not a strain is invasive. Genetic differences outside the coding regions, like for example in promoter regions, can influence protein abundance level, which may explain why the abundance levels can improve strain classification.

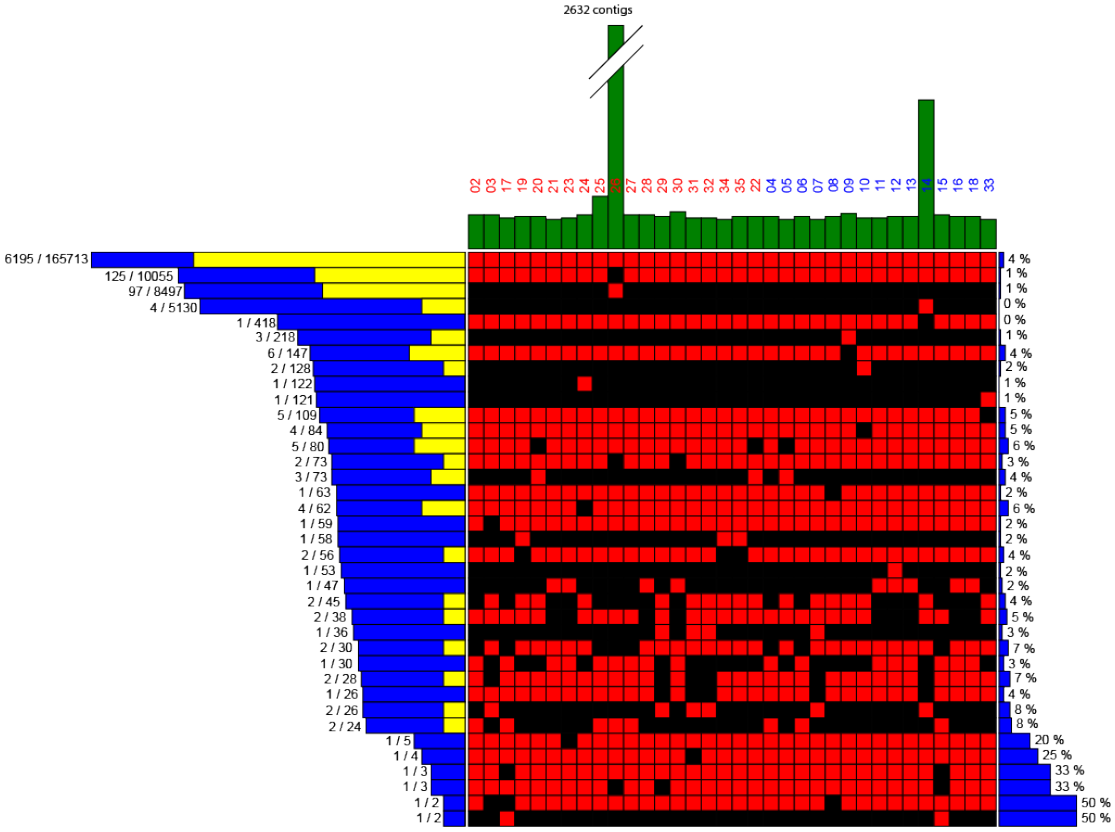



**Figure 3. Peptide-centric view of the coding potential of the genomes**. All peptides were mapped to the composite genome and the individual genomes. Six hundred and eighty conservation profiles were constructed from this data by mapping peptides to genomes and the 37 profiles with at least one detected peptide are shown. Each row corresponds to a profile, and presence of the peptide in the given genome is indicated by a red box, absence by black. The total number of peptides for each profile is shown in the blue histogram to the left and the number of displayed in yellow (log scale); the fraction of peptides in each profile that was detected is displayed in the bar graph to the right, calculated by dividing the total number of peptides by the number of observed ones. The histogram at the top indicates the number of contigs for the genome in question. The top histogram is organized according to virulence where red text indicates invasive and blue text non-invasive.

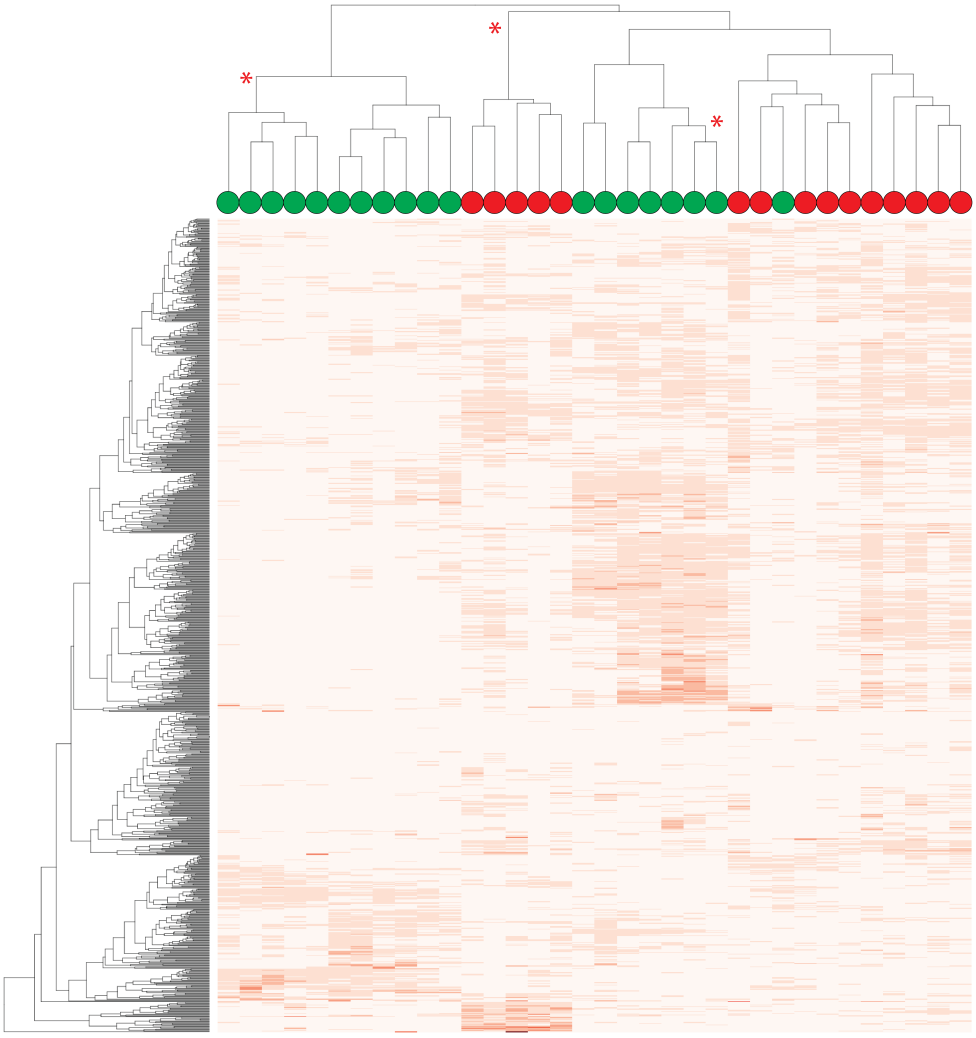


**Figure 4. Quantitative peptide expression matrix**. Construction of a relative abundance matrix using DIA-MS. The DIA-MS data was processed through OpenSWATH using the TraML spectral library. A heat map and unsupervised hierarchical clustering of strains and peptides were simultaneously created using the pvclust algorithm from the R package pvclust. The peptides are coloured according to intensity with darker colours indicating a higher level of expression. The asterisks at the top of the dendrogram indicate statistical significance. The coloured spheres indicate if the strain was invasive (red) or non-invasive (green).

## Small but consistent differences in SNP frequencies in expressed and non-expressed proteins

A total of 1665 SNPs were detected among the 34 genomes and the $M1_{ref}$ genome. These can be divided up into three groups: non-synonymous SNPs that cause amino acid substitutions, synonymous SNPs in the coding regions that do not cause amino acid substitutions and SNPs in the intergenic regions. Proteins that are not expressed might on average be more amenable to SNPs since they presumably would not cause deleterious phenotypes if mutated. This

presumption is supported as seen in Figure 5. Both the number of non-synonymous SNPs and synonymous SNPs in the coding regions are statistically more common in the proteins that were not quantified (Fig. 5a–b). In contrast, there is no difference in the number of SNPs in the promoter regions between quantified and unquantified proteins (Fig. 5c). The three types of mutations are represented in Figure 5d as follows: red/pink lines are the number of non-synonymous SNPs in quantified versus unquantified proteins. The proteins are ordered in a descending order in respect to the number of mutations. There are more unquantified proteins with a higher number of SNPs as the pink line is above the red. The same holds true for synonymous SNPs (green/lime lines) and SNPs in the intergenic regions (blue/cyan lines). These results partly explain why relatively few of the total of 223,952 unique peptide sequences were quantified. While speculative, one possible explanation is that unquantified proteins reflect the background mutation rates as fewer of these mutations will have a negative impact on the fitness of the individual strain. Mutation rates in highly expressed constituent proteins are more likely to have an impact on the fitness. The logical extension of this is that proteins that are both expressed and affected by mutations are more likely to be involved in increasing the fitness of the individual. These proteins are likely candidates to hold the key in what differs between an individual that is fit in a hostile environment and ones that were never exposed to this environment.

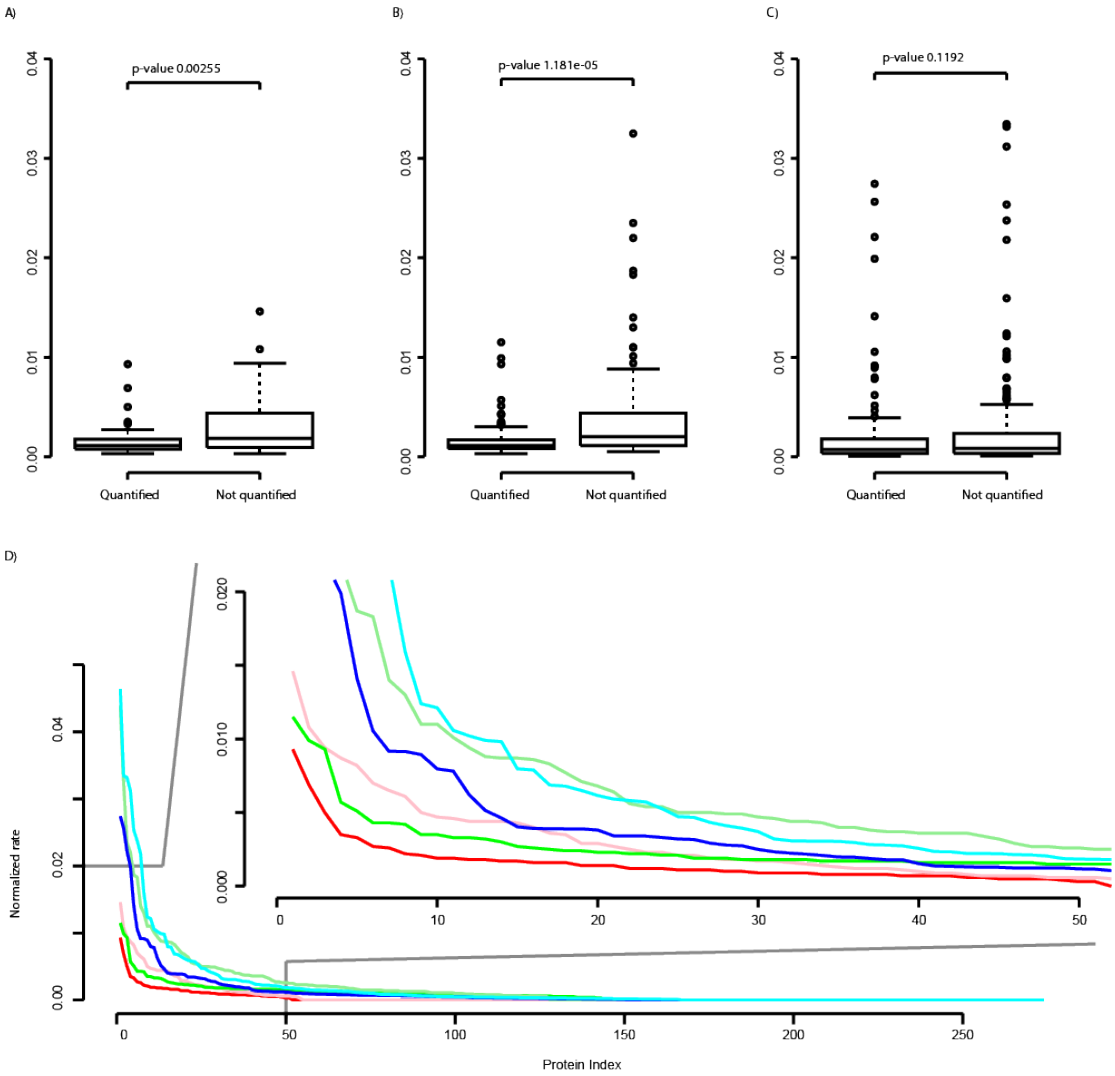
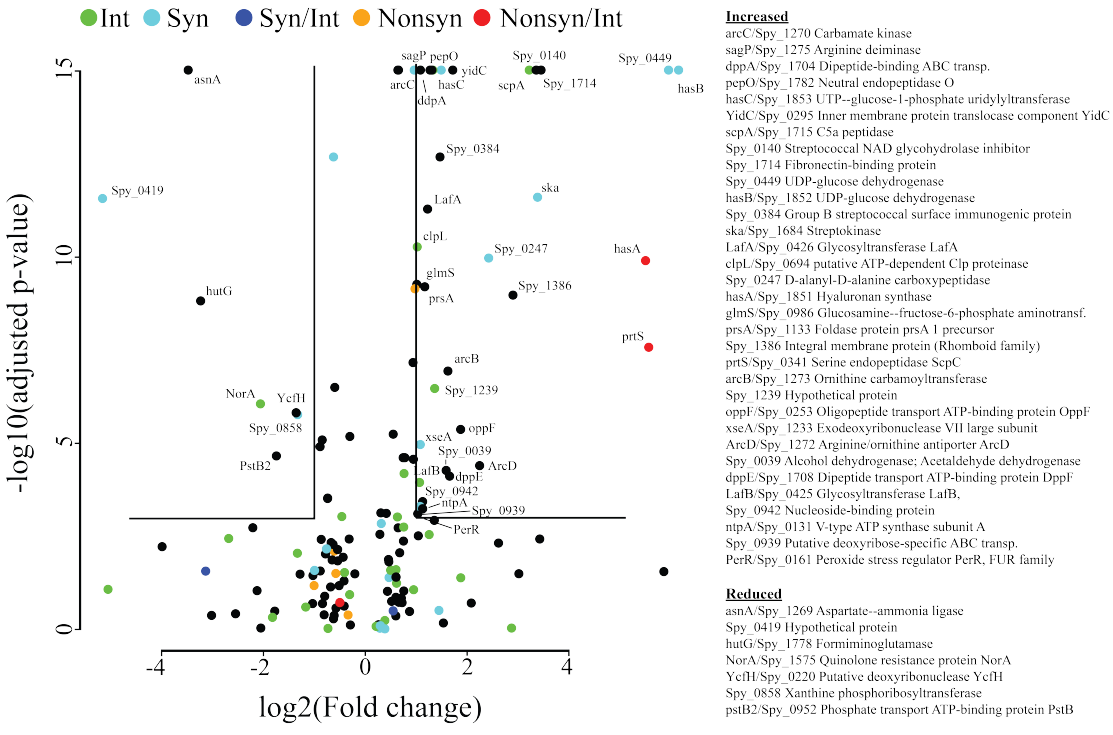



**Figure 5. Small but consistent differences between detected and undetected ORFs**. Box plots of number of A) non-synonymous SNPs per ORF length, B) SNPs per ORF length, and C) SNPs in the intergenic region normalized for length. Proteins with quantified peptides have significantly fewer SNPs than non-quantified proteins. D) Three pairs of lines; the red (quantified)/pink (unquantified) lines are the non-synonymous SNPs per ORF length, the green (quantified)/light-green (unquantified) lines are for synonymous SNPs and the blue (quantified)/cyan (unquantified) lines are the number of SNPs in the preceding intergenic region.

The composite genome data structure allows for fast exploration of the data, especially an explorative interrogation of the relationship between differentially expressed proteins and the SNPs that affect the amino acid composition and/or abundance. We performed statistical analysis of the significant clusters from Figure 4 to find discriminatory proteins. Figure 6 displays 40 proteins with significantly changed abundance levels in the significant cluster containing the invasive strains (adjusted p-value <0.001). In total, 33 of these proteins were

significantly increased and are significantly enriched for the protein functions arginine deiminase pathway, streptococcus pyogenes virulome and sucrose metabolism. A subset of these was also affected by SNPs in the coding region or in the preceding intergenic region or both as indicated by the coloured dots (Fig. 6). Some of the proteins with statistically increased protein abundance levels impact the virulence grade or the general fitness, as shown previously[51]. It is plausible that proteins that are both differentially expressed and affected by mutations are of significance for the virulence grade of the pathogen. The most prominent example is *hasA*, which is also highlighted in Figure 2. *hasA* is a known virulence factor and its gene has two SNPs, both of which cause amino acid substitutions. There is also an SNP in the preceding intergenic region. *hasA* is significantly induced several fold (p-value < $1x10^{-10}$) among these invasive strains compared to all the non-invasive strains. The SNP data and protein expression differences observed between non-invasive and invasive strains may indicate that these proteins have a role in the development of severe invasive disease, and represent interesting targets for additional future experiments.




**Increased**
arcC/Spy_1270 Carbamate kinase
sagP/Spy_1275 Arginine deiminase
dppA/Spy_1704 Dipeptide-binding ABC transp.
pepO/Spy_1782 Neutral endopeptidase O
hasC/Spy_1853 UTP--glucose-1-phosphate uridylyltransferase
YidC/Spy_0295 Inner membrane protein translocase component YidC
scpA/Spy_1715 C5a peptidase
Spy_0140 Streptococcal NAD glycohydrolase inhibitor
Spy_1714 Fibronectin-binding protein
Spy_0449 UDP-glucose dehydrogenase
hasB/Spy_1852 UDP-glucose dehydrogenase
Spy_0384 Group B streptococcal surface immunogenic protein
ska/Spy_1684 Streptokinase
LafA/Spy_0426 Glycosyltransferase LafA
clpL/Spy_0694 putative ATP-dependent Clp proteinase
Spy_0247 D-alanyl-D-alanine carboxypeptidase
hasA/Spy_1851 Hyaluronan synthase
glmS/Spy_0986 Glucosamine--fructose-6-phosphate aminotransf.
prsA/Spy_1133 Foldase protein prsA 1 precursor
Spy_1386 Integral membrane protein (Rhomboid family)
prtS/Spy_0341 Serine endopeptidase ScpC
arcB/Spy_1273 Ornithine carbamoyltransferase
Spy_1239 Hypothetical protein
oppF/Spy_0253 Oligopeptide transport ATP-binding protein OppF
xseA/Spy_1233 Exodeoxyribonuclease VII large subunit
ArcD/Spy_1272 Arginine/ornithine antiporter ArcD
Spy_0039 Alcohol dehydrogenase; Acetaldehyde dehydrogenase
dppL/Spy_1708 Dipeptide transport ATP-binding protein DppF
LafB/Spy_0425 Glycosyltransferase LafB,
Spy_0942 Nucleoside-binding protein
ntpA/Spy_0131 V-type ATP synthase subunit A
Spy_0939 Putative deoxyribose-specific ABC transp.
PerR/Spy_0161 Peroxide stress regulator PerR, FUR family

**Reduced**
asnA/Spy_1269 Aspartate--ammonia ligase
Spy_0419 Hypothetical protein
hutG/Spy_1778 Formiminoglutamase
NorA/Spy_1575 Quinolone resistance protein NorA
YcfH/Spy_0220 Putative deoxyribonuclease YcfH
Spy_0858 Xanthine phosphoribosyltransferase
pstB2/Spy_0952 Phosphate transport ATP-binding protein PstB

**Figure 6.** The cluster with invasive strains was evaluated using a Hochberg-adjusted two-sample Welch t-test as implemented in the R-package multi-test. Proteins regulated at least two-fold with an adjusted p-value cutoff of at least 0.001 are listed to the right and the corresponding protein names are marked in the volcano plot.

## Conclusion

In this work we present a generic data strategy for integrating genome and proteome data. The strategy relies on the construction of a composite genome to integrate peptide and protein information. The composite genome provides the basis for the construction of a spectral library based on shotgun MS analysis of the strains followed by DIA-MS. The spectral library is subsequently used to monitor the expression of peptides that could be quantified from the DIA maps using the spectral library. The work demonstrates how DIA can accomplish quantification of both conserved and non-conserved peptides and that DIA-MS is a promising technology for proteogenomics research. We applied the strategy to shed light on the comparatively few genetic differences that can be identified between non-invasive and invasive *S. pyogenes* strains. Several factors influence the fitness of a pathogen inside the host and to help to avoid detection by the host immune defence system. Evasion of the immune system may depend on the types and amounts of proteins exposed outside the cell wall and the affinity of these proteins to host molecules. Some interactions are beneficial for survival, such as the ability to bind blood plasma proteins to cover potential epitopes and others. Proteins that are expressed at detectable levels have fewer SNPs in the coding region than ones that are not expressed or are expressed below the limit of detection. This study revealed several proteins that are both affected by mutations and differentially expressed between invasive and non-invasive strains. There are many aspects of this data set that remain unexplored and we are confident that more insights into the interaction between pathogens and their hosts can be extracted from this data set and future proteogenomics data sets.

## Author contribution

LM and JM designed the study and wrote the manuscript. LM carried out most of the data analysis. ACP selected and collected the clinical isolates. AB and OK grew the strains and prepared them for mass spectrometry analysis. CK and GS prepared the DNA for sequencing. HL assembled and annotated the genomes. SH carried out the MS measurements.

## Acknowledgements

The Swedish Foundation for Strategic Research financially supports this project for strategic research. JM was supported by the Swedish Research Council (project 621-2012-3559), the Swedish Foundation for Strategic Research (grant FFL4), the Crafoord Foundation (grant 20100892), Stiftelsen Olle Engkvist Byggmästare, the Wallenberg Academy Fellow KAW (2012.0178) and a European Research Council starting grant (ERC-2012-StG-309831). We would also like to extend our thanks to the SyBIT project of the SystemsX.ch initiative and the Brutus and Crick system administrators for their support with computing infrastructure and other IT-related resources. We also thank Niclas Winqvist and Magnus Rasmussen for their assistance in collecting the isolates and performing the initial characterizations of the clinical isolates.

## References

(1) Kucharova, V.; Wiker, H. G. Proteogenomics in microbiology: taking the right turn at the junction of genomics and proteomics. *Proteomics* **2014**, *14*, 2360–2675.

(2) Nasser, W.; Beres, S. B.; Olsen, R. J.; Dean, M. A.; Rice, K. A.; Long, S. W.; Kristinsson, K. G.; Gottfredsson, M.; Vuopio, J.; Raisanen, K.; et al. Evolutionary pathway to increased virulence and epidemic group A Streptococcus disease derived from 3,615 genome sequences. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, E1768–E1776.

(3) Olsen, R. J.; Laucirica, D. R.; Watkins, M. E.; Feske, M. L.; Garcia-Bustillos, J. R.; Vu, C.; Cantu, C.; Shelburne, S. A.; Fittipaldi, N.; Kumaraswami, M.; et al. Polymorphisms in regulator of protease B (RopB) alter disease phenotype and strain virulence of serotype M3 group A Streptococcus. *J. Infect. Dis.* **2012**, *205*, 1719–1729.

(4) Euler, C. W.; Ryan, P. A.; Martin, J. M.; Fischetti, V. A. M.SpyI, a DNA methyltransferase encoded on a mefA chimeric element, modifies the genome of Streptococcus pyogenes. *J. Bacteriol.* **2007**, *189*, 1044–1054.

(5) Kant, S.; Agarwal, S.; Pancholi, P.; Pancholi, V. The Streptococcus pyogenes orphan protein tyrosine phosphatase, SP-PTP, possesses dual specificity and essential virulence regulatory functions. *Mol. Microbiol.* **2015**, n/a–n/a.

(6) Yates, C. M.; Sternberg, M. J. E. The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *J. Mol. Biol.* **2013**, *425*, 3949–3963.

(7) Horstmann, N.; Sahasrabhojane, P.; Suber, B.; Kumaraswami, M.; Olsen, R. J.; Flores, A.; Musser, J. M.; Brennan, R. G.; Shelburne, S. A. Distinct single amino acid replacements in the control of virulence regulator protein differentially impact streptococcal pathogenesis. *PLoS Pathog.* **2011**, *7*,

e1002311.
(8) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 4646–4658.
(9) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383–5392.
(10) Chapman, J. D.; Goodlett, D. R.; Masselon, C. D. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom Rev* **2014**, *33*, 452–470.
(11) Purvine, S.; Eppel, J.-T.; Yi, E. C.; Goodlett, D. R. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* **2003**, *3*, 847–850.
(12) Ramos, A. A.; Yang, H.; Rosen, L. E.; Yao, X. Tandem parallel fragmentation of peptides for mass spectrometry. *Anal. Chem.* **2006**, *78*, 6391–6397.
(13) Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G.-Z.; McKenna, T.; Nold, M. J.; Richardson, K.; Young, P.; et al. Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* **2005**, *77*, 2187–2200.
(14) Geiger, T.; Cox, J.; Mann, M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol. Cell Proteomics* **2010**, *9*, 2252–2261.
(15) Venable, J. D.; Dong, M.-Q.; Wohlschlegel, J.; Dillin, A.; Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **2004**, *1*, 39–45.
(16) Carvalho, P. C.; Han, X.; Xu, T.; Cociorva, D.; Carvalho, M. D. G.; Barbosa, V. C.; Yates, J. R. XDIA: improving on the label-free data-independent analysis. *Bioinformatics* **2010**, *26*, 847–848.
(17) Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell Proteomics* **2012**, *11*, O111.016717–O111.016717.
(18) Weisbrod, C. R.; Eng, J. K.; Hoopmann, M. R.; Baker, T.; Bruce, J. E. Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification. *J. Proteome Res.* **2012**, *11*, 1621–1632.
(19) Röst, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinović, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmstrom, J.; Malmstrom, L.; et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **2014**, *32*, 219–223.
(20) Blackburn, K.; Mbeunkui, F.; Mitra, S. K.; Mentzel, T.; Goshe, M. B. Improving protein and proteome coverage through data-independent multiplexed peptide fragmentation. *J. Proteome Res.* **2010**, *9*, 3621–3637.
(21) Geromanos, S. J.; Vissers, J. P. C.; Silva, J. C.; Dorschel, C. A.; Li, G.-Z.; Gorenstein, M. V.; Bateman, R. H.; Langridge, J. I. The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics* **2009**, *9*, 1683–1695.
(22) Molecular basis of group A streptococcal virulence. **2003**, *3*, 191–200.

(23) Mitchell, T. J. The pathogenesis of streptococcal infections: from tooth decay to meningitis. *Nat. Rev. Microbiol.* **2003**, *1*, 219–230.
(24) Cunningham, M. W. Pathogenesis of group A streptococcal infections. *Clin. Microbiol. Rev.* **2000**, *13*, 470–511.
(25) Bisno, A. L.; Brito, M. O.; Collins, C. M. Molecular basis of group A streptococcal virulence. *Lancet Infect Dis* **2003**, *3*, 191–200.
(26) Steer, A. C.; Law, I.; Matatolu, L.; Beall, B. W.; Carapetis, J. R. Global emm type distribution of group A streptococci: systematic review and implications for vaccine development. *Lancet Infect Dis* **2009**, *9*, 611–616.
(27) O'Brien, K. L.; Beall, B.; Barrett, N. L.; Cieslak, P. R.; Reingold, A.; Farley, M. M.; Danila, R.; Zell, E. R.; Facklam, R.; Schwartz, B.; et al. Epidemiology of invasive group a streptococcus disease in the United States, 1995-1999. *Clin. Infect. Dis.* **2002**, *35*, 268–276.
(28) Kahn, F.; Linder, A.; Petersson, A. C.; Christensson, B.; Rasmussen, M. Axillary abscess complicated by venous thrombosis: identification of Streptococcus pyogenes by 16S PCR. *J. Clin. Microbiol.* **2010**, *48*, 3435–3437.
(29) Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **2013**, *29*, 1072–1075.
(30) Simpson, J. T.; Wong, K.; Jackman, S. D.; Schein, J. E.; Jones, S. J. M.; Birol, I. ABySS: a parallel assembler for short read sequence data. *Genome Res.* **2009**, *19*, 1117–1123.
(31) Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **2014**, *30*, 2068–2069.
(32) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; 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*, 918–920.
(33) Quandt, A.; Espona, L.; Balasko, A.; Weisser, H.; Brusniak, M.-Y.; Kunszt, P.; Aebersold, R.; Malmstrom, L. Using synthetic peptides to benchmark peptide identification software and search parameters for MS/MS data analysis. *EuPA Open Proteomics* **2014**, *5*, 21–31.
(34) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466–1467.
(35) David L Tabb; Christopher G Fernando, A.; Chambers, M. C. MyriMatch: Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis. *J. Proteome Res.* **2007**, *6*, 654–661.
(36) Rosenberger, G.; Koh, C. C.; Guo, T.; Röst, H. L.; Kouvonen, P.; Ben C Collins; Heusel, M.; Liu, Y.; Caron, E.; Vichalkovski, A.; et al. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Scientific Data, Published online: 16 September 2014; | doi:10.1038/sdata.2014.31* **2014**, *1*, 140031.
(37) Bauch, A.; Adamczyk, I.; Buczek, P.; Elmer, F.-J.; Enimanev, K.; Glyzewski, P.; Kohler, M.; Pylak, T.; Quandt, A.; Ramakrishnan, C.; et al. openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics* **2011**, *12*, 468.
(38) Kunszt, P.; Blum, L.; Hullár, B.; Schmid, E.; Srebniak, A.; Wolski, W.; Rinn,

B.; Elmer, F.-J.; Ramakrishnan, C.; Quandt, A.; et al. iPortal: the swiss grid proteomics portal. *Concurrency and Computation: Practice and Experience* **2014**, n/a–n/a.

(39) Teleman, J.; Röst, H. L.; Rosenberger, G.; Schmitt, U.; Malmstrom, L.; Malmstrom, J.; Levander, F. DIANA-algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics* **2015**, *31*, 555–562.

(40) Malmstrom, L.; Marko-Varga, G.; Westergren-Thorsson, G.; Laurell, T.; Malmstrom, J. 2DDB - a bioinformatics solution for analysis of quantitative proteomics data. *BMC Bioinformatics* **2006**, *7*, 158.

(41) Malmstrom, L.; Malmstrom, J.; Marko-Varga, G.; Westergren-Thorsson, G. Proteomic 2DE database for spot selection, automated annotation, and data analysis. *J. Proteome Res.* **2002**, *1*, 135–138.

(42) Lars Malmström, P. N. J. M. Business intelligence strategies enables rapid analysis of quantitative proteomics data. *journal of Proteome Science and Computational Biology* **2012**, *1*, 5.

(43) Ferretti, J. J.; McShan, W. M.; Ajdic, D.; Savic, D. J.; Savic, G.; Lyon, K.; Primeaux, C.; Sezate, S.; Suvorov, A. N.; Kenton, S.; et al. Complete genome sequence of an M1 strain of Streptococcus pyogenes. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 4658–4663.

(44) Assefa, S.; Keane, T. M.; Otto, T. D.; Newbold, C.; Berriman, M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **2009**, *25*, 1968–1969.

(45) Angiuoli, S. V.; Salzberg, S. L. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* **2011**, *27*, 334–342.

(46) Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **2004**, *14*, 1188–1190.

(47) Stothard, P.; Wishart, D. S. Circular genome visualization and exploration using CGView. *Bioinformatics* **2005**, *21*, 537–539.

(48) DeAngelis, P. L.; Papaconstantinou, J.; Weigel, P. H. Molecular cloning, identification, and sequence of the hyaluronan synthase gene from group A Streptococcus pyogenes. *J. Biol. Chem.* **1993**, *268*, 19181–19184.

(49) Wang, X.; Zhang, B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **2013**, *29*, 3235–3237.

(50) Deutsch, E. W.; Chambers, M.; Neumann, S.; Levander, F.; Binz, P.-A.; Shofstahl, J.; Campbell, D. S.; Mendoza, L.; Ovelleiro, D.; Helsens, K.; et al. TraML--a standard format for exchange of selected reaction monitoring transition lists. *Mol. Cell Proteomics* **2012**, *11*, R111.015040–R111.015040.

(51) Malmstrom, J.; Karlsson, C.; Nordenfelt, P.; Ossola, R.; Weisser, H.; Quandt, A.; Hansson, K.; Aebersold, R.; Malmstrom, L.; Björck, L. Streptococcus pyogenes in human plasma: adaptive mechanisms analyzed by mass spectrometry-based proteomics. *J. Biol. Chem.* **2012**, *287*, 1415–1425.