



LUND UNIVERSITY

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](https://doi.org/10.1016/j.jprot.2015.09.012)

2015

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

[Link to publication](#)

Citation for published version (APA):
Malmström, L., Bakochi, A., Svensson Birkedal, G., Kilsgård, O., Lantz, H., Petersson, A. C., Hauri, S., Karlsson, C., & Malmström, J. (2015). Quantitative proteogenomics of human pathogens using DIA-MS. *Journal of Proteomics*, 129, 98-107. <https://doi.org/10.1016/j.jprot.2015.09.012>

Total number of authors:
9

Creative Commons License:
CC BY-NC-ND

General rights

Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: <https://creativecommons.org/licenses/>

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

1 Quantitative proteogenomics of human
2 pathogens using DIA-MS

3 Lars Malmström¹, Anahita Bakochi², Gabriel Svensson², Ola Kilsgård², Henrik
4 Lantz³, Ann Cathrine Petersson⁴, Simon Hauri², Christofer Karlsson² and Johan
5 Malmström²

6

7

8 ¹S3IT, University of Zurich, Zurich, Switzerland

9 ²Division of Infection Medicine, Department of Clinical Sciences Lund, Lund
10 University, Lund, Sweden

11 ³Department of Medical Biochemistry and Microbiology/BILS, Uppsala
12 University, Uppsala, Sweden

13 ⁴Department of Clinical Microbiology, Division of Laboratory Medicine, Region
14 Skåne, Lund, Sweden

15 Corresponding author: lars.malmstroem@uzh.ch

16 **Abstract**

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

36 **Highlights**

- 37 • 34 sequenced genomes and corresponding shotgun and DIA-MS
38 measurements
- 39 • Construction of a composite genome for fast data integration
- 40 • Quantitative DIA-MS of the conserved and non-conserved peptide pool
41 across all strains

42 **Significance**

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

50 **Keywords**

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

53 **Abbreviations**

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

57 **Introduction**

58 In proteogenomics, mass spectrometry (MS)-based proteomics is used as a
59 supplement to genomic data by adding a level of information to the
60 interpretation of genomic sequences¹. In this context, MS is particularly relevant
61 in microbiology where a large number of genomes are sequenced regularly¹.
62 Comparative genomic analysis of microbial genomes has revealed compelling
63 evidence that some pathogens undergo rapid genomic adaption to increase
64 fitness in their host². The influence of single nucleotide polymorphisms (SNPs)
65 on the molecular phenotype may be substantial, leading to increased virulence
66 or the ability to survive and thereby cause disease³. Other events such as DNA
67 methylation⁴ and phosphorylation⁵ can modify how the genome is translated,
68 leading to increased virulence. Small genomic changes can influence survival and
69 virulence in several ways, for example by activating/inactivating regulatory
70 systems controlling part of the proteome expression³, disrupting protein-protein
71 interactions⁶ or by increasing or decreasing the affinity between transcription
72 factors and their target promoters⁷. The rapid increase in the number of
73 genomes provides the opportunity to use matching genotype and strain to
74 investigate how sets of SNPs alter proteome homeostasis. However, matching
75 genotype and strain information in MS-based proteomics presents considerable
76 challenges.

77 MS-based proteomics experiments rely on a protein database to provide the
78 ground truth, i.e. information on all the possible tryptic peptides that can be
79 derived from a given genome. The ideal protein database should contain all
80 required information while remaining as small as possible. In the case of
81 proteogenomics, this problem becomes amplified if approached naively by
82 concatenating the protein database from each genome as it becomes challenging
83 to select a particular protein if many similar proteins exist in the database⁸. On
84 the other hand, searching each MS data file against its appropriate genome is
85 standard procedure; the challenge here is to combine the independent searches
86 without increasing the false discovery rate (FDR) dramatically⁹. The reason for
87 the increase in FDR is that the correct proteins are, to a large extent, the same

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

99 In contrast to shotgun MS and traditional database searches, DIA-MS provides
100 new opportunities to use the differential degree of peptide conservation to
101 further explore the rapid increase in sequenced genomes. DIA-MS was originally
102 developed to expand the detectable dynamic range and does not use real-time
103 ion selection-based precursor scans¹⁰. This can be accomplished by interrogating
104 predetermined m/z ranges by either fragmenting all ions entering the mass
105 spectrometer¹¹⁻¹⁴ or by dividing the full m/z range into fixed smaller isolation
106 windows¹⁵⁻¹⁸. Several of the developed DIA methods differ in how subsequent
107 data analysis is performed¹⁰. In 2012, Gillet et al showed that the identification of
108 peptides from DIA experiments can be accomplished via spectral libraries
109 constructed from previously acquired shotgun MS¹⁷, nowadays implemented in
110 search algorithms¹⁹. In general, the DIA methods are associated with increased
111 signal-to-noise ratios, increased sensitivity and increased specificity based on
112 peptide fragmentation¹⁵, and have shown improved reproducibility compared to
113 a data-dependent acquisition (DDA) counterpart^{20,21}. Importantly for
114 proteogenomic strategies, the spectral libraries can easily include all observed
115 SNPs in a given strain and thereby remove the problem with large FASTA
116 databases or difficulties with controlling FDR resulting from concatenating
117 several individual searches, provided that the peptides are represented in the
118 spectral library. Spectral libraries can be constructed based on the level of
119 peptide conservation and this enables quantitative analysis of both conserved
120 and non-conserved peptides, which can be used to determine protein abundance

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

126 *S. pyogenes* is one of the most common and important human pathogens^{22,23} and
127 is responsible for mild diseases such as pharyngitis, erysipelas and impetigo as
128 well as severe diseases such as streptococcal toxic shock syndrome and
129 necrotizing fasciitis²⁴. Annually, *S. pyogenes* causes over 616 million cases of
130 pharyngitis and 111 million cases of impetigo²⁴. It encodes many well-
131 characterized virulence factors, including surface-bound M protein and M-like
132 proteins, hyaluronic acid capsules, adhesins, surface-bound collagen-like
133 proteins, superantigenic exotoxins, and numerous secreted and extracellular
134 proteins²⁵. Antigenic differences in the hypervariable region of the M protein are
135 the basis for the Lancefield serological classification of *S. pyogenes* with over 200
136 identified serotypes to date²⁶. Strains of certain serotypes are epidemiologically
137 associated with particular clinical syndromes where serotype M1 and M3 have
138 frequently, but not exclusively, been isolated from patients with severe invasive
139 disorders and infections with these serotypes are associated with increased
140 mortality²⁷. The extent to which genomic adaptation observed in invasive *S.*
141 *pyogenes* strains results in altered proteome composition and increased
142 virulence remains unclear.

143 In this study, we collected 34 clinical strains of *S. pyogenes* serotype M1,
144 sequenced all the genomes and then analysed full proteome digests of all strains
145 with DDA-MS and DIA-MS. We generated a so-called composite genome that
146 contains all the genetic information of the strains and derived all potential
147 tryptic peptides containing between 7 and 50 amino acids that this composite
148 genome could theoretically encode. We constructed a spectral library by
149 searching the shotgun MS data against the peptide database. The spectral library
150 was then used to analyse the DIA-MS data to generate a quantitative expression
151 matrix. We constructed a data structure that allowed us to analyse the three
152 different data sets in light of each other, highlighting the relevance of several
153 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.

156 **Methods**

157 **Isolates**

158 Emm1 GAS were isolated between April and May 2012 at the accredited
159 diagnostic laboratories of clinical microbiology, Division of Laboratory Medicine,
160 Lund, Sweden. Isolates from sterile sites were sent to the laboratories as part of
161 routine health care whereas isolates from throat swabs were collected as a part
162 of a surveillance programme from selected geographically scattered primary
163 care units in southern Sweden. Isolates were characterized as group A
164 streptococci through agglutination and were typed through PCR and sequencing
165 essentially as described²⁸. The modified primers *emm* for 5'-GCT TAG AAA ATT
166 AAA AAM MGG-3'²⁸ and CDC-R 5'-GCA AGT TCT TCA GCT TGT-3'
167 (<http://www.cdc.gov/streplab/protocol-emm-type.html>) were used. *Emm* types
168 were assigned through the type-specific database at
169 <http://www2a.cdc.gov/ncidod/biotech/strepblast.asp>. In total, 34 *S. pyogenes*
170 M1 strains were subdivided into strains responsible for non-invasive conditions,
171 in this case tonsillitis (n=18), and invasive conditions such as necrotizing
172 fasciitis, toxic shock syndrome and/or endomyometritis (n=16).

173 **Whole genome sequencing**

174 Genomic DNA was extracted from the *Streptococcus pyogenes* isolates using a
175 silica-membrane spin column kit (Macherey-Nagel). In brief, overnight cultures
176 (3.5 mL) were harvested by centrifugation at 3500 x g, resuspended in ice-cold
177 70% ethanol and incubated at -20 °C for 20 minutes. The cell wall was digested
178 by resuspending the bacteria in 25 mM Tris-HCl, 2 mM EDTA, 1% (v/v) Triton X-
179 100 containing 20 mg/mL lysozyme and 250 units/mL mutanolysin (both
180 enzymes from Sigma-Aldrich) followed by incubation at 37 °C for 2 hours.
181 Genomic DNA was released from the bacteria by resuspending the bacteria in a
182 buffer containing SDS and 20 mg/mL proteinase K and overnight incubation at
183 56 °C. Subsequent DNA purification was performed according to the
184 manufacturer's protocol for the silica-membrane spin column kit. Preheated
185 elution buffer (70 °C, 5 mM Tris-HCl, pH 8.5) was applied to the spin column

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

193 **Whole genome assembly and annotation**

194 Several assemblers were tried, and based on comparisons using Quast²⁹, Abyss
195 1.3.7 was chosen with a kmer size of 39³⁰. This gave a good balance of a low
196 number of misassemblies compared to the reference genome of strain
197 MGAS5005 together with a high continuity of the genome assemblies. Annotation
198 was performed using Prokka 1.10 with the rfam option³¹.

199 **Sample preparation for mass spectrometry**

200 The clinically isolated *S. pyogenes* strains were grown overnight on blood agar
201 plates (37 °C, 5% CO₂), after which single colonies were grown to mid-
202 exponential phase in Todd-Hewitt broth (30 g/l) (Difco Laboratories)
203 supplemented with yeast extract (6 g/l) (Difco Laboratories). The cells were
204 harvested by centrifugation and resuspended in 50 mM Tris-HCl and 150 mM
205 NaCl (Medicago) wash buffer, pH 7.6, to a final concentration of 2 x 10⁹ CFU/mL.
206 After several washes the bacterial pellets were spun down and dissolved in ice-
207 cold LC-grade water and heat-inactivated by incubation on a heat block for 5 min
208 at 80 °C. The cells were transferred to lysing matrix tubes (Nordic Biolabs)
209 containing 90 mg of 0.1 mm silica beads and homogenized using a cell disruptor
210 (Beadbeater, FastPrep 96, MP Biomedicals). The cell debris was removed and the
211 supernatants were denatured in 10 M urea (Sigma-Aldrich) and 50 mM
212 ammonium bicarbonate (ABC) (Fluka Analytical), followed by incubation with 1
213 µg trypsin (Sequencing Grade Modified Trypsin, Porcine, Promega, Madison, WI,
214 USA) for 30 min at 37 °C for protein digestion. The samples were reduced using
215 500 mM Tris(2-carboxyethyl)phosphine (TCEP) (Sigma-Aldrich) for 60 minutes
216 at 37 °C, and alkylated with 500 mM 2-Iodoacetamide (IAA) (AppliChem) for 30

217 min at room temperature in the dark. The samples were diluted in 250 μ l 100
218 mM ABC and further digested with 1 μ g trypsin (Sequencing Grade Modified
219 Trypsin, Porcine, Promega) overnight. The trypsin was inactivated by adding
220 formic acid (FA) until the pH was 2-3. In accordance with the manufacturer's
221 instructions, C18 columns (Vydac UltraMicro Spin™ Silica C18 300Å Columns,
222 #SUM SS18V, The Nest Group, Inc., Southborough, MA, USA) were used to clean
223 up, desalt and concentrate the peptides in the samples. The solvents were
224 removed in a SpeedVac and the peptides were resuspended in 50 μ l buffer A (2%
225 acetonitrile, 0.2% FA in LC-H2O).

226 **LC-MS/MS analysis**

227 All peptide measurements were acquired on a Q Exactive Plus mass
228 spectrometer (Thermo Scientific) coupled to an EASY-nLC 1000 ultra-high
229 pressure liquid chromatography system (Thermo Scientific). Peptides were
230 trapped on an Acclaim PepMap® 100 pre-column (Thermo Scientific, C18, 3 μ m,
231 100 Å; ID 75 μ m x 2 cm) and separated with a PepMap® RSLC EASY-Spray
232 column (Thermo Scientific; C18 2 μ m, 100 Å; ID 75 μ m x 25 cm; heated to 45° C),
233 using intelligent flow control for column equilibration and sample load at 800
234 bars. A linear gradient of between 5% and 35% acetonitrile in aqueous 0.1%
235 formic acid was run for 120 min at a flow rate of 300 nl/min.

236 For shotgun MS, one full scan (resolution 70,000 @ 200 m/z; mass range 400–
237 1600 m/z) was followed by 15 MS/MS scans (resolution 17,500 @ 200 m/z) of
238 the most abundant ion signals (TOP15). Precursor ions were fragmented using
239 HCD at a normalized collision energy of 30. Charge state screening was set to
240 reject unassigned or singly charged ions. The dynamic exclusion time was set to
241 15 s and limited to 300 entries. AGC was set to 1e6 for both MS and MS/MS with
242 ion accumulation times of 100 ms (MS) and 60 ms (MS/MS). The intensity
243 threshold for precursor ion selection was 1.7e⁴.

244 For data-independent SWATH-like analysis, a full MS scan (resolution 70,000 @
245 200 m/z; mass range 400–1200 m/z) was followed by 32 MS/MS fragmentation
246 scans (resolution 35,000 @ 200 m/z) using an isolation window of 26 m/z
247 (including 1 m/z overlap between windows). The precursor ions within each
248 isolation window were fragmented using high-energy collision-induced

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

252 All samples injected contained a peptide standard for retention time calibration.
253 The obtained raw files were converted to mzXML using the software tool
254 ProteoWizard³².

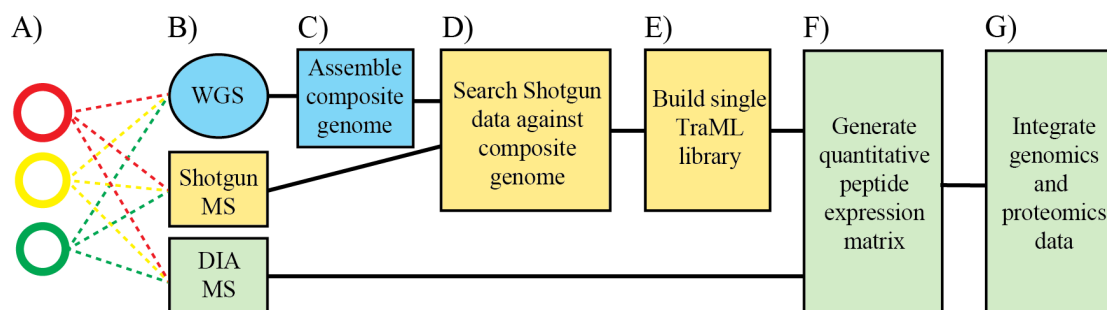
255 **Database searching and bioinformatics**

256 The shotgun MS data was searched as described by Quandt et al³³. In short, we
257 used X! Tandem³⁴ and MyriMatch³⁵ with a precursor ion mass tolerance of 30
258 ppm and a fragment ion mass tolerance of 10 ppm allowing no miscleavages. The
259 search results were statistically validated using Peptide Prophet⁹. The spectral
260 library was created³⁶ and the resulting TraML file was used to analyse the DIA-
261 MS data as described by Röst et al¹⁹. Both WGS and MS data were stored in
262 openBIS³⁷ and processing related to MS was carried out using iPortal³⁸. The DIA-
263 MS data was statistically evaluated using pyProphet³⁹. All data integration was
264 carried out under the DDB framework^{40,41}, using non-normalized analytical
265 tables⁴².

266 Results and discussion

267 Workflow overview

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



281

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

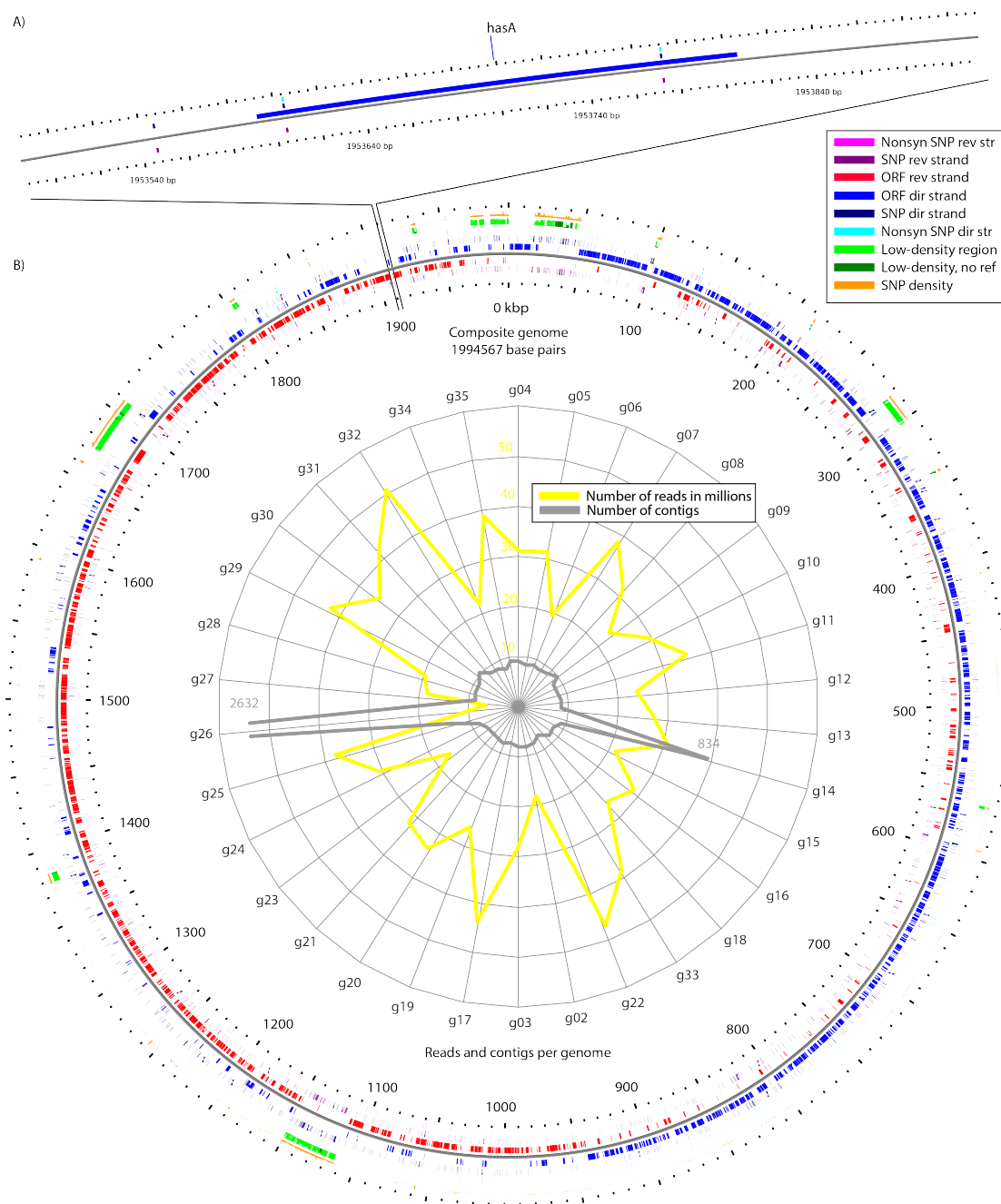
289 Generation of a composite genome

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

294 expressed SNPs for that given ORF and whether the ORF is preceded by strain-
295 specific changes in the intragenic region, which may influence the abundance
296 level of that protein. To integrate all genetic information from the 34 genome
297 sequences, we constructed a composite genome as follows: the Illumina reads
298 from the 34 strains were assembled into contigs (see Figure 2 for a summary).
299 The number of reads per strain varied from 6,513,248 to 50,054,680 with an
300 average of 30,124,134 and resulted in 273 contigs on average (number of contigs
301 range: 161–2632). We included the two poorly assembled genomes (g14 and
302 g26) since the number of identified peptides from these genomes was similar to
303 the others (7442 and 6835 peptides respectively, ranking 11 and 30 of 34, the
304 range is 6213-8288, median 7174). This indicates that the assemblies over the
305 expressed ORFs were of similar quality to other genome assemblies despite the
306 high number of contigs. We choose NC_002737.1, a complete *S. pyogenes* genome
307 of serotype M1, as reference and we refer to it as M1_{ref} in the text below⁴³. The
308 contigs were ordered according to M1_{ref} using Abacas⁴⁴ and we used Mugsy⁴⁵
309 to align the ordered contigs onto the M1_{ref}. The alignment was used to build a
310 composite genome that contains all the genetic information from all strains (Fig.
311 2), stored in a denormalized analytical table for fast querying⁴². The consensus
312 genome was 1,994,567 BP, only slightly larger than the average 1.8 MB member
313 genomes, indicating a high degree of genomic similarity between the strains.
314 Importantly, a consensus sequence was generated by a majority vote with
315 random selection in cases of equal counts. We estimated the sequence
316 conservation identically to Crooks et al⁴⁶. The resulting composite genome is
317 displayed in Figure 2b using CGView⁴⁷. The composite genome is represented as
318 the black line in the middle, and tracks on the inside represent features on the
319 reverse strand and tracks on the outside features on the direct strand. Closest to
320 the genome are the open reading frames (red and blue) followed by a track
321 indicating all detected SNPs (purple and navy). The third track shows SNPs that
322 lead to an amino acid substitution (fuchsia and lime). The zoom-in panel on the
323 left shows the genomics region between 1953500 and 1953900 where the ORF
324 coding for *hasA* is located (Fig. 1A). *hasA* has been implicated in the virulence
325 mechanisms previously and its primary function is in the biosynthesis of the
326 capsule⁴⁸.

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

345 The composite genome further supports the exploration of how the observed
346 genomic alters the proteome homeostasis by providing an improved data
347 structure for annotating the genome with both identified peptides and putative
348 proteins found by Prokka³¹. This allowed us to separate the SNPs that are found
349 within an ORF from SNPs found elsewhere. The ones found within an ORF were
350 further divided into synonymous and non-synonymous. Figure 2 shows that
351 SNPs that lead to amino acid substitution are rare compared to the total number
352 of observed SNPs. As previously demonstrated, invasive strains tend to
353 accumulate specific SNPs of relevance for invasive disease³. This system
354 represents a suitable model system for establishing the DIA-MS-based
355 proteogenomic strategy described next.



357

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

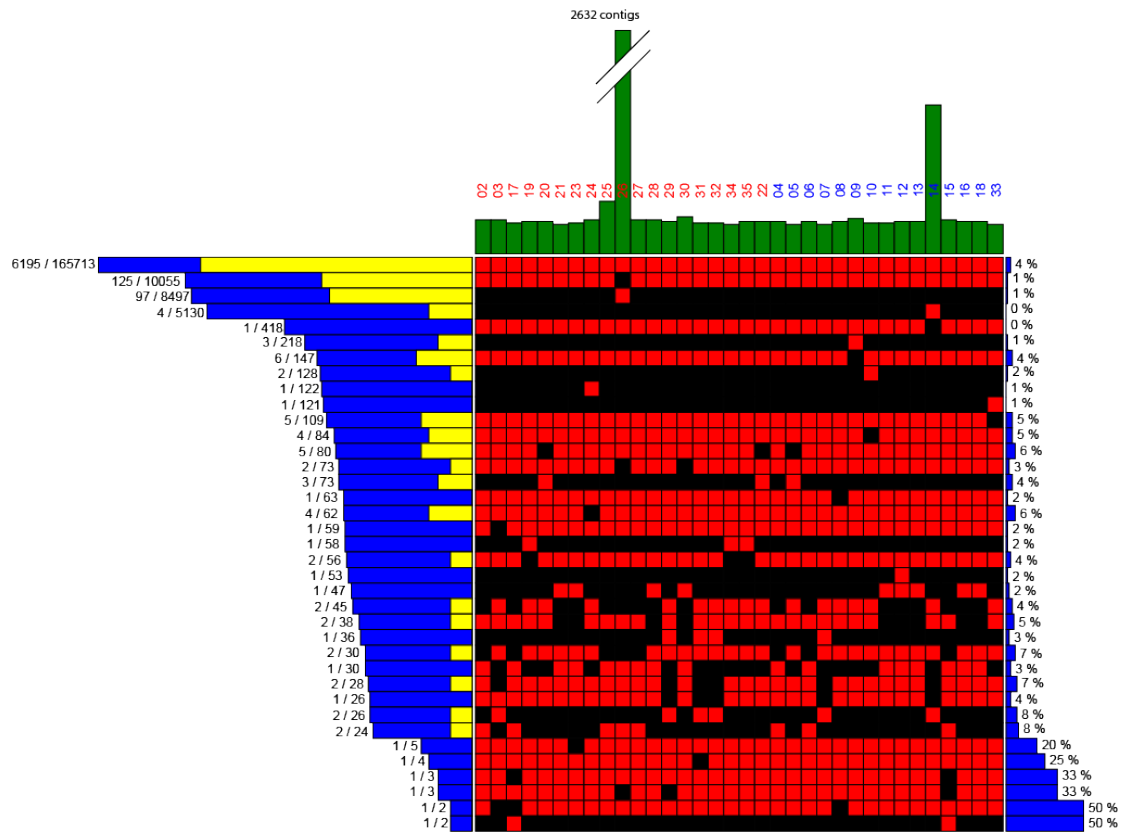
372 **Generation of a spectral library for DIA-MS analysis**

373 One important step of the proposed quantitative proteogenomics strategy is the
374 construction of a spectral library that contains all detectable peptides including
375 peptide sequences conserved across all strains as well as the non-conserved
376 peptides. Here, we constructed the peptide spectral library by translating all
377 members of the composite genome in six frames and selecting all fully tryptic
378 peptides between 7 and 50 amino acids in length resulting in a total of 223,952
379 unique peptide sequences⁴⁹. These unique peptide sequences were used to
380 search the 34 strains grown in duplicate resulting in 68 shotgun MS experiments
381 using X! tandem³⁴, Myrimatch³⁵ and peptideProphet⁹ on a previously published
382 portal³³. The search results were used to construct a spectral library in the
383 TraML format as previously described⁵⁰ (Fig. 1c-d). In total, this effort generated
384 a spectral library for *S. pyogenes* containing 14,633 precursors corresponding to
385 11,552 unique peptide sequences at 1% peptide-level FDR, representing 5.1% of
386 the total 223,952 unique peptide sequences that can be potentially produced
387 from all the 34 genomes. The relatively low coverage is not surprising since the
388 vast majority of the putative peptides are never expressed. For example, only
389 one out of six reading frames is actually used for any stretch of DNA. Of course,
390 intergenic DNA and proteins not expressed under the tested condition cannot be
391 detected either for obvious reasons.

392 **Generation of a quantitative expression matrix**

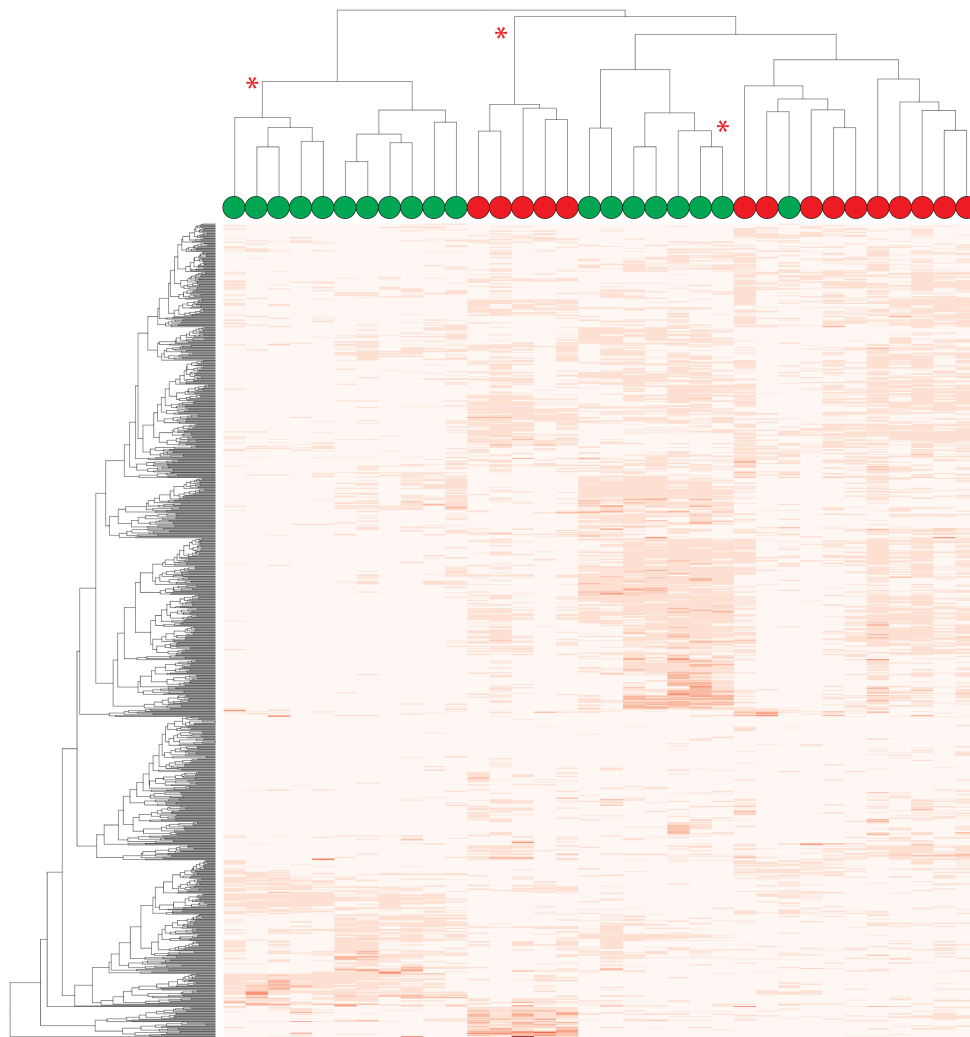
393 One of the biological replicates from the 34 SWATH-like MS DIA-MS data sets
394 was analysed with OpenSWATH¹⁹ using the spectral library as the source of
395 precursors to consider. The resulting expression matrix contained quantitative
396 values for 6880 peptides over the 34 genomes. Figure 1e shows a schematic
397 overview of how the expression matrix was produced. We generated profiles for
398 all 223,952 unique peptide sequences derived from the composite genome based
399 on their presence or absence in the 34 member genomes, resulting in a total of
400 680 profiles. Out of the 680 profiles, 37 were associated with at least one
401 detected peptide as shown in the heat map in Figure 3. The histogram to the left
402 shows the number of peptides associated with each profile; the bar graph to the
403 right displays the fraction of the identified peptides for the profile. The vast

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



434

435 **Figure 3. Peptide-centric view of the coding potential of the genomes.** All peptides were
 436 mapped to the composite genome and the individual genomes. Six hundred and eighty
 437 conservation profiles were constructed from this data by mapping peptides to genomes and the
 438 37 profiles with at least one detected peptide are shown. Each row corresponds to a profile, and
 439 presence of the peptide in the given genome is indicated by a red box, absence by black. The total
 440 number of peptides for each profile is shown in the blue histogram to the left and the number of
 441 displayed in yellow (log scale); the fraction of peptides in each profile that was detected is
 442 displayed in the bar graph to the right, calculated by dividing the total number of peptides by the
 443 number of observed ones. The histogram at the top indicates the number of contigs for the
 444 genome in question. The top histogram is organized according to virulence where red text
 445 indicates invasive and blue text non-invasive.



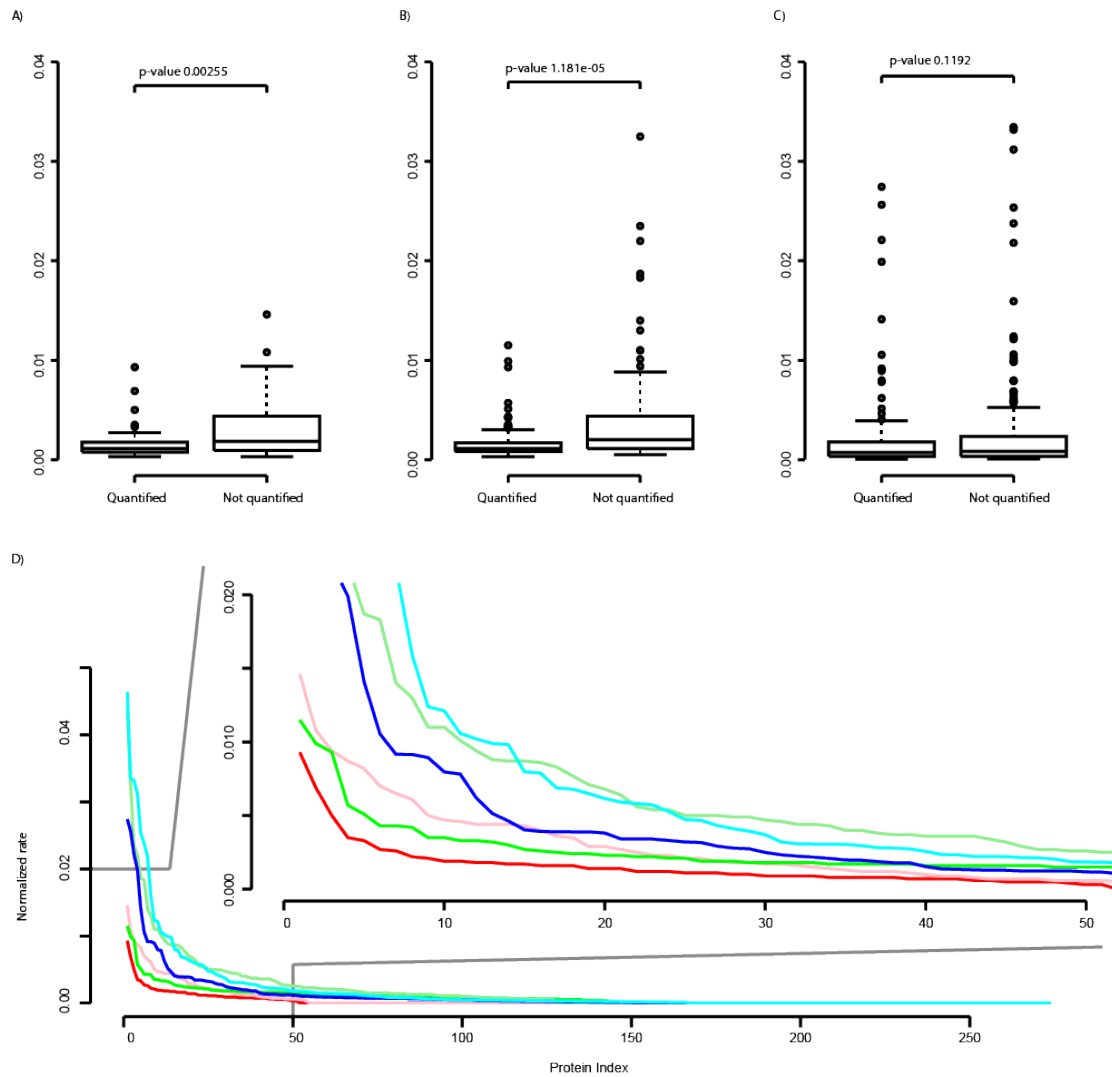
446

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

454 **Small but consistent differences in SNP frequencies in expressed and non-**
 455 **expressed proteins**

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

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



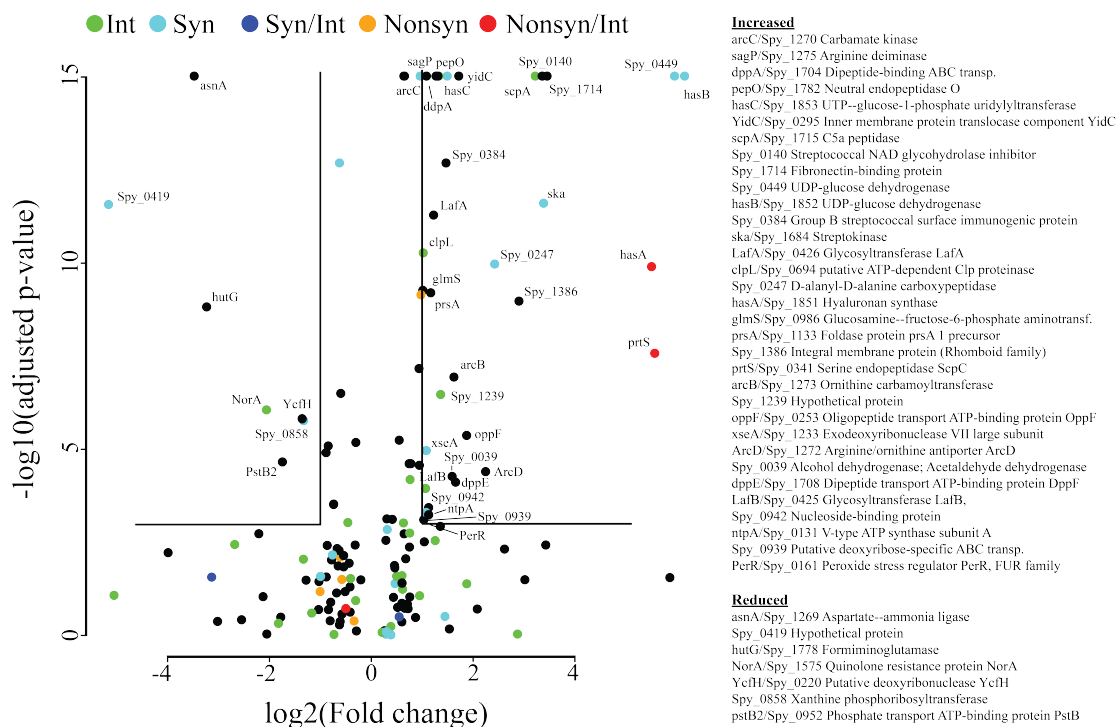
483

484

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

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

500 significantly increased and are significantly enriched for the protein functions
 501 arginine deiminase pathway, streptococcus pyogenes virulome and sucrose
 502 metabolism. A subset of these was also affected by SNPs in the coding region or
 503 in the preceding intergenic region or both as indicated by the coloured dots (Fig.
 504 6). Some of the proteins with statistically increased protein abundance levels
 505 impact the virulence grade or the general fitness, as shown previously⁵¹. It is
 506 plausible that proteins that are both differentially expressed and affected by
 507 mutations are of significance for the virulence grade of the pathogen. The most
 508 prominent example is *hasA*, which is also highlighted in Figure 2. *hasA* is a known
 509 virulence factor and its gene has two SNPs, both of which cause amino acid
 510 substitutions. There is also an SNP in the preceding intergenic region. *hasA*
 511 is significantly induced several fold (p-value < 1x10⁻¹⁰) among these invasive
 512 strains compared to all the non-invasive strains. The SNP data and protein
 513 expression differences observed between non-invasive and invasive strains may
 514 indicate that these proteins have a role in the development of severe invasive
 515 disease, and represent interesting targets for additional future experiments.



516
 517
 518
 519
 520
 521

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.

522 **Conclusion**

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

547 **Author contribution**

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

553 Acknowledgements

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

565 References

- 566 (1) Kucharova, V.; Wiker, H. G. Proteogenomics in microbiology: taking the
567 right turn at the junction of genomics and proteomics. *Proteomics* **2014**,
568 *14*, 2360–2675.
- 569 (2) Nasser, W.; Beres, S. B.; Olsen, R. J.; Dean, M. A.; Rice, K. A.; Long, S. W.;
570 Kristinsson, K. G.; Gottfredsson, M.; Vuopio, J.; Raisanen, K.; et al.
571 Evolutionary pathway to increased virulence and epidemic group A
572 Streptococcus disease derived from 3,615 genome sequences. *Proc. Natl.*
573 *Acad. Sci. U.S.A.* **2014**, *111*, E1768–E1776.
- 574 (3) Olsen, R. J.; Laucirica, D. R.; Watkins, M. E.; Feske, M. L.; Garcia-Bustillos, J.
575 R.; Vu, C.; Cantu, C.; Shelburne, S. A.; Fittipaldi, N.; Kumaraswami, M.; et al.
576 Polymorphisms in regulator of protease B (RopB) alter disease
577 phenotype and strain virulence of serotype M3 group A Streptococcus. *J.*
578 *Infect. Dis.* **2012**, *205*, 1719–1729.
- 579 (4) Euler, C. W.; Ryan, P. A.; Martin, J. M.; Fischetti, V. A. M. Sply, a DNA
580 methyltransferase encoded on a mefA chimeric element, modifies the
581 genome of Streptococcus pyogenes. *J. Bacteriol.* **2007**, *189*, 1044–1054.
- 582 (5) Kant, S.; Agarwal, S.; Pancholi, P.; Pancholi, V. The Streptococcus
583 pyogenes orphan protein tyrosine phosphatase, SP-PTP, possesses dual
584 specificity and essential virulence regulatory functions. *Mol. Microbiol.*
585 **2015**, n/a–n/a.
- 586 (6) Yates, C. M.; Sternberg, M. J. E. The effects of non-synonymous single
587 nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *J.*
588 *Mol. Biol.* **2013**, *425*, 3949–3963.
- 589 (7) Horstmann, N.; Sahasrabhojane, P.; Suber, B.; Kumaraswami, M.; Olsen, R.
590 J.; Flores, A.; Musser, J. M.; Brennan, R. G.; Shelburne, S. A. Distinct single
591 amino acid replacements in the control of virulence regulator protein
592 differentially impact streptococcal pathogenesis. *PLoS Pathog.* **2011**, *7*,

- 593 e1002311.
- 594 (8) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model
595 for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**,
596 *75*, 4646–4658.
- 597 (9) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical
598 model to estimate the accuracy of peptide identifications made by
599 MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383–5392.
- 600 (10) Chapman, J. D.; Goodlett, D. R.; Masselon, C. D. Multiplexed and data-
601 independent tandem mass spectrometry for global proteome profiling.
602 *Mass Spectrom Rev* **2014**, *33*, 452–470.
- 603 (11) Purvine, S.; Eppel, J.-T.; Yi, E. C.; Goodlett, D. R. Shotgun collision-induced
604 dissociation of peptides using a time of flight mass analyzer. *Proteomics*
605 **2003**, *3*, 847–850.
- 606 (12) Ramos, A. A.; Yang, H.; Rosen, L. E.; Yao, X. Tandem parallel fragmentation
607 of peptides for mass spectrometry. *Anal. Chem.* **2006**, *78*, 6391–6397.
- 608 (13) Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G.-Z.;
609 McKenna, T.; Nold, M. J.; Richardson, K.; Young, P.; et al. Quantitative
610 proteomic analysis by accurate mass retention time pairs. *Anal. Chem.*
611 **2005**, *77*, 2187–2200.
- 612 (14) Geiger, T.; Cox, J.; Mann, M. Proteomics on an Orbitrap benchtop mass
613 spectrometer using all-ion fragmentation. *Mol. Cell Proteomics* **2010**, *9*,
614 2252–2261.
- 615 (15) Venable, J. D.; Dong, M.-Q.; Wohlschlegel, J.; Dillin, A.; Yates, J. R.
616 Automated approach for quantitative analysis of complex peptide
617 mixtures from tandem mass spectra. *Nat. Methods* **2004**, *1*, 39–45.
- 618 (16) Carvalho, P. C.; Han, X.; Xu, T.; Cociorva, D.; Carvalho, M. D. G.; Barbosa, V.
619 C.; Yates, J. R. XDIA: improving on the label-free data-independent
620 analysis. *Bioinformatics* **2010**, *26*, 847–848.
- 621 (17) Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner,
622 R.; Aebersold, R. Targeted data extraction of the MS/MS spectra
623 generated by data-independent acquisition: a new concept for consistent
624 and accurate proteome analysis. *Mol. Cell Proteomics* **2012**, *11*,
625 0111.016717–0111.016717.
- 626 (18) Weisbrod, C. R.; Eng, J. K.; Hoopmann, M. R.; Baker, T.; Bruce, J. E.
627 Accurate peptide fragment mass analysis: multiplexed peptide
628 identification and quantification. *J. Proteome Res.* **2012**, *11*, 1621–1632.
- 629 (19) Röst, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinović, S. M.;
630 Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmstrom, J.; Malmstrom, L.; et
631 al. OpenSWATH enables automated, targeted analysis of data-
632 independent acquisition MS data. *Nat. Biotechnol.* **2014**, *32*, 219–223.
- 633 (20) Blackburn, K.; Mbeunkui, F.; Mitra, S. K.; Mentzel, T.; Goshe, M. B.
634 Improving protein and proteome coverage through data-independent
635 multiplexed peptide fragmentation. *J. Proteome Res.* **2010**, *9*, 3621–3637.
- 636 (21) Geromanos, S. J.; Vissers, J. P. C.; Silva, J. C.; Dorschel, C. A.; Li, G.-Z.;
637 Gorenstein, M. V.; Bateman, R. H.; Langridge, J. I. The detection,
638 correlation, and comparison of peptide precursor and product ions from
639 data independent LC-MS with data dependant LC-MS/MS. *Proteomics*
640 **2009**, *9*, 1683–1695.
- 641 (22) Molecular basis of group A streptococcal virulence. **2003**, *3*, 191–200.

- 642 (23) Mitchell, T. J. The pathogenesis of streptococcal infections: from tooth
643 decay to meningitis. *Nat. Rev. Microbiol.* **2003**, *1*, 219–230.
- 644 (24) Cunningham, M. W. Pathogenesis of group A streptococcal infections.
645 *Clin. Microbiol. Rev.* **2000**, *13*, 470–511.
- 646 (25) Bisno, A. L.; Brito, M. O.; Collins, C. M. Molecular basis of group A
647 streptococcal virulence. *Lancet Infect Dis* **2003**, *3*, 191–200.
- 648 (26) Steer, A. C.; Law, I.; Matatolu, L.; Beall, B. W.; Carapetis, J. R. Global emm
649 type distribution of group A streptococci: systematic review and
650 implications for vaccine development. *Lancet Infect Dis* **2009**, *9*, 611–
651 616.
- 652 (27) O'Brien, K. L.; Beall, B.; Barrett, N. L.; Cieslak, P. R.; Reingold, A.; Farley, M.
653 M.; Danila, R.; Zell, E. R.; Facklam, R.; Schwartz, B.; et al. Epidemiology of
654 invasive group a streptococcus disease in the United States, 1995-1999.
655 *Clin. Infect. Dis.* **2002**, *35*, 268–276.
- 656 (28) Kahn, F.; Linder, A.; Petersson, A. C.; Christensson, B.; Rasmussen, M.
657 Axillary abscess complicated by venous thrombosis: identification of
658 *Streptococcus pyogenes* by 16S PCR. *J. Clin. Microbiol.* **2010**, *48*, 3435–
659 3437.
- 660 (29) Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUASt: quality
661 assessment tool for genome assemblies. *Bioinformatics* **2013**, *29*, 1072–
662 1075.
- 663 (30) Simpson, J. T.; Wong, K.; Jackman, S. D.; Schein, J. E.; Jones, S. J. M.; Birol, I.
664 ABySS: a parallel assembler for short read sequence data. *Genome Res.*
665 **2009**, *19*, 1117–1123.
- 666 (31) Seemann, T. Prokka: rapid prokaryotic genome annotation.
667 *Bioinformatics* **2014**, *30*, 2068–2069.
- 668 (32) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.;
669 Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; et al. A cross-
670 platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*
671 **2012**, *30*, 918–920.
- 672 (33) Quandt, A.; Espona, L.; Balasko, A.; Weisser, H.; Brusniak, M.-Y.; Kunszt,
673 P.; Aebersold, R.; Malmstrom, L. Using synthetic peptides to benchmark
674 peptide identification software and search parameters for MS/MS data
675 analysis. *EuPA Open Proteomics* **2014**, *5*, 21–31.
- 676 (34) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass
677 spectra. *Bioinformatics* **2004**, *20*, 1466–1467.
- 678 (35) David L Tabb; Christopher G Fernando, A.; Chambers, M. C. MyriMatch:
679 Highly Accurate Tandem Mass Spectral Peptide Identification by
680 Multivariate Hypergeometric Analysis. *J. Proteome Res.* **2007**, *6*, 654–661.
- 681 (36) Rosenberger, G.; Koh, C. C.; Guo, T.; Röst, H. L.; Kouvonen, P.; Ben C
682 Collins; Heusel, M.; Liu, Y.; Caron, E.; Vichalkovski, A.; et al. A repository
683 of assays to quantify 10,000 human proteins by SWATH-MS. *Scientific*
684 *Data, Published online: 16 September 2014; | doi:10.1038/sdata.2014.31*
685 **2014**, *1*, 140031.
- 686 (37) Bauch, A.; Adamczyk, I.; Buczek, P.; Elmer, F.-J.; Enimanev, K.; Glyzowski,
687 P.; Kohler, M.; Pylak, T.; Quandt, A.; Ramakrishnan, C.; et al. openBIS: a
688 flexible framework for managing and analyzing complex data in biology
689 research. *BMC Bioinformatics* **2011**, *12*, 468.
- 690 (38) Kunszt, P.; Blum, L.; Hullár, B.; Schmid, E.; Srebniak, A.; Wolski, W.; Rinn,

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

694 (39) Teleman, J.; Röst, H. L.; Rosenberger, G.; Schmitt, U.; Malmstrom, L.;
695 Malmstrom, J.; Levander, F. DIANA-algorithmic improvements for
696 analysis of data-independent acquisition MS data. *Bioinformatics* **2015**,
697 *31*, 555–562.

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

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

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

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

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

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

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

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

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

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

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

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

735
736