Proteome-wide selected reaction monitoring assays for the human pathogen Streptococcus pyogenes.

Karlsson, Christofer; Malmström, Lars; Aebersold, Ruedi; Malmström, Johan

Published in:
Nature Communications

DOI:
10.1038/ncomms2297

2012

Proteome-wide selected reaction monitoring assays for the human pathogen *Streptococcus pyogenes*

Christofer Karlsson[1, #], Lars Malmström[2, #], Ruedi Aebersold[2, 3], Johan Malmström[1,4] *

[1)] Department of Immunotechnology, Lund University, Lund SE-221 84, Sweden
[2)] Institute of Molecular Systems Biology, ETH Zurich, Zurich CH-8093, Switzerland.
[3)] Faculty of Science, University of Zurich, Zurich CH-8006, Switzerland
[4)] Biognosys AG, Schlieren CH-8952, Switzerland
[#)] Equal author contribution

[*)] Corresponding author:
Johan Malmström, Ph.D
Department of Immunotechnology
BMC, D13
SE-221 84 Lund, Sweden

Phone: +4646-2220830
Fax: +4646-2224200
Johan.Malmstrom@immun.lth.se

Keywords: targeted proteomics, *Streptococcus pyogenes*, mass spectrometry, selected reaction monitoring (SRM), multiple reaction monitoring (MRM), SRM assay, proteome-wide

Running title: Proteome wide analysis of *Streptococcus pyogenes*

Abstract

Selected reaction monitoring (SRM) mass spectrometry (MS) is a targeted proteomics technology used to identify and quantify proteins with high sensitivity, specificity and high reproducibility. Execution of SRM-MS relies on protein-specific SRM assays, a set of experimental parameters that requires considerable effort to develop. Here we present a proteome-wide SRM assay repository for the important gram-positive human pathogen group A Streptococcus (GAS). Using a multi-layered approach we generated SRM assays for 10412 distinct GAS peptides followed by extensive testing of the SRM assays in more than 200 different GAS protein pools. Based on the number of SRM assay observations we created a rule-based SRM assay scoring model to select the most suitable assays per protein for a given cellular compartment and bacterial state. The resource described here represents an important tool for deciphering the GAS proteome using SRM and we anticipate that concepts described here can be extended to other pathogens.

Bacterial infections are a major cause of disease and mortality aggravated by the emerging resistance to antibiotics. During an infection, pathogenic bacteria can rapidly alter their proteome composition to adapt to hostile environments and evade immune response[1-4]. How the bacteria regulate their proteome composition *in vivo* to accomplish host environment adaption and immune response evasion is, however, still unclear. Quantitative and comprehensive *in vivo* proteome-wide analysis of large cohorts of clinically isolated bacterial strains would considerably improve our understanding of how these processes are accomplished and how they are influenced by underlying genetic differences and environmental factors. For example advances in understanding underlying genetic differences between clinical GAS strains have revealed that mutations in the regulatory system *covRS* is linked to a severe disease outcome and as reviewed by Cole et al[5].

Modern proteomics technologies allow quantitative measurement of the vast majority of proteins in bacterial proteomes as recently reviewed[6]. The conceptual advance of directed mass spectrometry technologies using liquid chromatography coupled to tandem MS (LC-MS/MS)[7] has resulted in several proteome-wide absolute quantification studies of how bacteria adapt to new environments *in vitro*[8-10]. The development of SRM-MS analysis has recently become a viable complement to data dependent and directed MS analysis because data sets with unprecedented reproducibility across multiple samples and a large dynamic range can be achieved. SRM is a targeted MS technology where preselected pairs of peptide precursor ion and fragment ion mass masses, also known as transitions, are explicitly monitored over time in a triple quadrupole (QQQ) MS instrument. The non-scanning mode of measurement of the most intense peptides and peptide fragments for each protein results in the lowest limit of detection of any LC-based MS technique. Using SRM to study pathogen virulence mechanisms is attractive as bacterial proteomes have an estimated dynamic range of 4-5 orders of magnitude[10] which is smaller than the linear dynamic range of SRM[11] A key characteristic of SRM-MS analysis is the accurate protein quantification capability, where the quantification variance is similar to ELISA in bacterial proteomes[12]. The accurate protein quantification capability along with the reproducible mode of analysis results in comprehensive data matrices (protein quantity vs. sample) with very few missing values, as the same peptide species are measured in all samples [13,14]. The consistency and completeness of such data sets is important for the analysis of, for example, large collections of clinically isolated strains or for studying small genetic differences resulting in single amino acid substitutions.

The execution of SRM experiments is dependent on *a priori* knowledge regarding which peptides and transitions to target. This knowledge is typically obtained by creating deep proteome maps using multidimensional peptide fractionation strategies followed by data dependent LC-MS/MS analysis. From such proteome maps, proteotypic peptides (PTP's) uniquely identifying proteins of interest, and suitable transitions are selected and optimized[15].

The transitions are subsequently used by QQQ mass spectrometers, where peptide ions are isolated in the first quadrupole, fragmented in the second and the resulting peptide-fragments are isolated and monitored in the third quadrupole, providing a high degree of selectivity and sensitivity for the detection of the targeted peptides. Several transitions per peptide are commonly used to increase the confidence level that the targeted peptide is in fact identified and accurately quantified. Sets of transitions for a single precursor peptide along with the precise retention time are collectively referred to as an SRM assay.

The limited availability of SRM assays is a prohibitive bottleneck for carrying out SRM-MS analyses and necessitates time consuming and expensive SRM assay development. Although there is currently a considerable amount of effort put into the construction of large-scale transitions atlases to facilitate the step from selecting target proteins to actually measuring them[16], a proteome-wide repository for a bacterial pathogen has not been reported to date.

In the work described here we demonstrate the construction of a proteome-wide SRM assay repository for the important human pathogen Group A Streptococcus (GAS). GAS is a gram-positive bacterium responsible for common and relatively mild clinical conditions such as pharyngitis and streptococcal skin infections[17,18]. GAS can also cause severe and potentially life-threatening conditions such as septic shock and necrotizing fasciitis, resulting in more than 500000 deaths every year, thus making GAS one of the more important human pathogens.

The work described here outlines a multi-layered approach to generate SRM assays for 10412 distinct GAS peptides. To improve the usability of the repository we performed extensive testing of all SRM assays in different bacterial states and cellular compartments. Based on the performance of the individual assays as a function of biological matrix, we calculated an assay score based on a rule based assay-scoring model. This score ranks individual assays based on their detectability. The assay score ranked, proteome-wide SRM assay repository presented here provides an important resource for understanding GAS proteome spatial distribution,

organization of related protein functions and protein abundance range. Furthermore, we define a transportability index indicating the portability of individual SRM assays across related genomes. We anticipate that the resource described here will become an important resource for understanding GAS biology and that it can be used as a basis for the construction of SRM-wide assay repositories for other pathogens, emerging pathogens and commensal bacteria.

Results

Construction of a proteome-wide GAS SF370 SRM assay repository

We selected GAS SF370 as the model strain for the construction of a proteome-wide GAS SRM assay repository. Previous LC-MS/MS analysis on GAS SF370 resulted in the identification of 946 of the 1905 GAS SF370 open reading frames (ORFs)[7]. The data resulting from these measurements was stored in a publically available instance of PeptideAtlas[19]. In this study we expanded the available PeptideAtlas instance by resorting to sub-cellular fractionation from several GAS strains grown under various environmental conditions (Figure 1a). In total 433 high-resolution LC-MS/MS measurements using 231 unique protein pools resulted in the identification of 8320 proteotypic peptides (PTP's) for GAS. The PTP's were ranked according to decreasing extracted ion chromatogram (XIC) intensities, estimating protein abundance as previously described[10], and served as the basis for the construction of the proteome-wide SRM assay repository.

The construction of the proteome-wide SRM assay repository relied on a two-legged strategy. The first leg, outlined in Figure 1b, involved the construction of SRM assays based on a MS/MS spectral library. We constructed the spectral library for high abundant PTP's identified by several MS/MS spectra from the large-scale proteome inventory as previously described[20]. For PTP's without a sufficient number of fragment ion spectra to create a reliable spectral library, the corresponding PTP's were chemically synthesized. For proteins that remained undetected in the proteome mapping data sets we predicted the most suitable PTP's using APEX[21] and synthesized corresponding peptides, generating in total 2489 synthetic peptides. The synthesized peptides were analyzed by shotgun MS/MS and the resulting data were amended to the spectral library. The strongest conserved transitions and the retention time (RT) were extracted from the spectral library and stored, enabling RT normalized SRM assays to be downloaded directly into

the SRM methods used by the MS[22]. The efforts resulted in 10412 SRM assays and the transitions were ranked according to intensity as described earlier[20].

The second leg included iterative testing of the assays with SRM using a QQQ instrument to increase the confidence of individual SRM assays (Figure 1c). We tested 7621 distinct peptide sequences with their corresponding SRM assays, represented by a total of 79277 transitions, in cell lysates from GAS grown under different conditions. The conditions included different growth phases, oxidative stress, exposure to human plasma supplement, or antibiotics (Figure 1a). All SRM assays were tested at least two times and several more than hundred times (Figure 2a), resulting in 957850 individual ion chromatograms. The most frequently observed SRM assays and the most intense transitions associated with them were ranked as described previously[20]. We used this information to build an SRM assay score using a rule-based scoring model. The model divides the assays into three categories, low-, medium- and high-scoring. The scoring indicates the ability of an SRM assay to detect the corresponding peptide in tryptic GAS digests from cellular compartments and different bacterial states (Figure 1c). The major assay score parameter is based on the SRM false discovery rate (FDR) thresholds of peptide identification in complex biological peptide mixtures. The high scoring assays represent cases where the peptide was detected with high confidence in complex GAS peptide mixtures (FDR of ≤1%). The medium scoring assays represent cases where the peptide was detected with lower confidence (FDR 1>2%). These SRM assay score categories received and arbitrary score of 100 and 50 respectively. SRM assays developed on synthetic peptides were included in the medium scoring SRM assays. The fine-tuning of the assay score within these two categories was based on the number of times the peptides were observed, minus the number of attempted observations divided by two. Thus, the higher the frequency with which an SRM assay was observed with high probability, the higher the assay score. The low scoring SRM assays represent cases where the peptide remained undetected in complex GAS peptide mixtures. These SRM assays were scored based on the number of transitions per SRM assay, which positively influences the low-

scoring SRM assay scores to a moderate extent. Hence, peptides represented by low scoring SRM assays were never observed in streptococcal protein extracts. High- and medium-scoring assays were measured in good agreement with the synthetic SRM assays, but with different FDR thresholds. Figure 2b-c shows the current distribution of GAS SRM assays in the three categories, demonstrating that we can preferentially select suitable assays per protein using the rule-based assay-scoring model. As the number of observations increases over time, we anticipate continued improvement of the assay score prediction. In total we have defined 2731 SRM assays with a medium- or high-score, targeting 1332 proteins. Out of these 1201 (63.0% of total ORF's) were detected in complex mixtures (high-scoring SRM assays). 23% of the predicted ORF's were associated with low scoring SRM assays. For the majority of applications, the complete set of SRM assays provided in this study are typically not included for analysis. The developed assays score provides means to select the suitable SRM assays depending on targeted proteins and type of experiment and represent a useful strategy for prioritizing among the SRM assays.

SRM assay score biases

As the dynamic range of SRM is higher than the estimated dynamic range for microbial proteomes, we anticipate that the undetected proteins are not related to limitations in sensitivity of the method or the dynamic range of the sample. To estimate method biases between detected and undetected proteins we compared enrichment of the proteins associated with the different assay score categories to a number of parameters such as protein length, protein conservation and functional classification. Apart from three major tendencies between the scoring categories we discovered surprisingly few biases.

Firstly, we compared the relationship between proteins associated with high SRM assay scores and protein abundance estimated from extracted XIC from the shotgun proteome inventory established in the beginning of the study. XIC of the identified proteins with 1% FDR were

extracted and summed up and associated to the SRM assay score categories (Figure 3a). On average 91.6 % of the identified XIC were associated with proteins with at least one high- or medium-scoring SRM assay. Figure 3a does not include the proteins that were exclusively identified using SRM. The overlap between proteins with high assays score and proportion of XIC indicates that SRM can recapture the vast majority of identified proteins from the large shotgun proteome inventory analysis and in addition identify additional proteins. In the SRM experiments however, no additional offline peptide separation was performed, as was the case when the shotgun proteome inventory was constructed.

The second tendency relates to protein length where we observed that proteins with low-scoring SRM assays were predominately short compared to proteins with medium- and high-scoring SRM assays (Figure 3b). In contrast, proteins with high-scoring SRM assays were predominately long, indicating, as expected, that a larger number of PTP's to choose from per protein is beneficial for SRM analysis (Figure 3b).

The third tendency relates to protein conservation among the GAS genomes where relatively many proteins with low-scoring SRM assays are less conserved or unique compared to other GAS strains (Figure 3c). Around 58% of the protein sequences associated with low-scoring SRM assays are conserved across all 13 genomes tested (see Supplementary Table S1). In contrast 88% of the high-scoring SRM assays were associated to conserved proteins conserved across all 13 genomes, indicating that conserved proteins are expressed at higher frequency than non conserved protein sequences under the tested conditions (Figure 3c).

Collectively, we note that the extensive MS analysis of GAS described here using two different MS platforms provide a large degree of overlap. Proteins with low-scoring assays tend to be less suitable for MS analysis by any method, as they were short and thus contain fewer suitable PTP's. In addition, we observe an overrepresentation of proteins with high scoring assays among conserved protein sequences within the GAS protein universe. It is likely that the protein

sequences with exclusively low scoring SRM assays are not expressed under the tested growth conditions. Membrane proteins were not specifically addressed in the sub-cellular fractionation, however, we observe no overall bias against membrane proteins. Nevertheless, detection of certain membrane protein species may require specific digestion/extraction methods[23,24].

Spatial distribution of proteins with high scoring assays

The iterative testing of the developed SRM assays on more than 540 LC-SRM-MS measurements represents a comprehensive proteome-wide targeted measurement of the GAS SF370 encoding genome. As we tested the SRM assays in enriched fractions of the intracellular, surface associated and secreted protein pools we could also estimate the predominant localization for proteins associated with high scoring SRM assays (Figure 4). The majority of the proteins, 934 proteins were predominately present in the intracellular pool (Figure 4a and b). A smaller fraction, 115 and 28 proteins, were predominately found in the surface associated and secreted protein pools respectively (Figure 4c and d) leaving 124 proteins that were relatively evenly split between two or more compartments (Figure 4e and f).

To visualize the proteome distribution we used Cystoscape[25] with the Cerebral[26] plugin (Figure 4g, Supplementary Figure S1). The proteins were grouped according to cellular functions using the National pathogen microbe data resource (NMPDR) subsystem information[27] (now part of PATRICs Bioinformatics Resource Center[28]) and selected cellular location for the individual functional categories based on the protein expression profiles shown in Figure 4a-d, represented as circles in Figure 4g. Cellular functional categories containing either proteins with contradicting cellular location from cluster Figure 4e-f or different proteins with contradicting cellular location were considered to have unknown localization and are represented as rectangles in Figure 4g. The size of the circles/rectangles indicates the number of member proteins ranging from 1 to 34. The edges between the circles/rectangles represent protein members that belong to more than one NMPDR subsystem, whereas the location of the rectangles within the network

view is influenced by the edges. The color scheme represents continuous decreasing average SRM assay score, where red indicates NMPDR subsystems with high average SRM assay score (>119) and black indicating proteins with predominately low scoring SRM assays (<10). It can be noted that the majority of circles and rectangles are red, demonstrating the coverage of high-scoring assays across the GAS proteome. Several black protein groups were not detectable under the tested growth conditions. In general these subsystems contain relatively few members. In summary, the iterative testing of the SRM assays in three subcellular compartments provides an overview of the subcellular protein distribution for GAS strain SF370. The majority of the nodes have a relatively high proportion of high-scoring SRM assays. More information regarding the subcellular localization for individual proteins can be found in Supplementary Data 1.

Transportability of SRM assays across the GAS pan-genome and related species.

We used GAS SF370 as a model strain when developing the proteome-wide SRM assay repository. However, as there is substantial genetic variation within and between genomes from different GAS serotypes[29-32], it is important to know which SRM assays target proteins in other GAS strains. To explore the transportability of this resource to other GAS strains and closely or distantly related species, we selected in total 75 taxa of low GC Gram-positive bacteria, Firmictutes, (see Supplementary Figure S2) and mapped medium- and high-scoring SF370 assays onto respective genome. To estimate the taxon evolutionary relationship, a phylogenetic tree was constructed based on respective *rpoB* gene sequence (Supplementary Figure S2). There was a large attrition of assays with increasing evolutionary distance (Figure 5a) and depending taxonomic rank (Figure 5b). Nevertheless, transportability within the species rank was high (Fig 5b, c) with average genome coverage's of 59-70% (1167-1332 ORFs) demonstrating that the developed SRM-assays will target a majority of the currently defined GAS pan-genome products independent on serotype or strain (Figure 5c). Transportability in the genus rank was the most diverse (8-31% genome coverage) (Figure 5b) with Group C streptococci genomes having the

highest degree of average coverage and also being the closest related taxa based on *ropB* homologies (Figure 5a).

To address the identity of proteins with transferable assays we calculated the number of species with at least a single high scoring peptide belonging NMPDR subsystem. We note that the most frequent NMPDR functions with SRM assays with high-degree of transportability are as expected ribosomal proteins, universal GTPases and proteins involved in central metabolism. In contrast the NMPDR subsytems with lowest level of transportability are Phage capsid proteins, Custered, Regularly Interspaced Short Palindromic Repeats (CRISPR) associated proteins and *Streptococcus pyogenes* Virulome (see Supplementary Figure S3a, b for more information). The SRM assay transportability is an important parameter when targeting GAS proteins in microbial communities as for example in the oral cavity where many bacterial species are present[33].

SRM assay repository and availability

The availability of the SRM assays (transitions, retention time and collision energy), their NMPDR subsystems, measured subcellular localization and degree of transferability is found in Supplementary Data 2. The full list of SRM assays can be downloaded from the PeptideAtlas Public SRM Transition Lists at

https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/GetTransitionLists under the accession number PATR00014.

Discussion

SRM is a targeted proteomics technology capable of accurate and reproducible quantification of proteins in complex samples. It has a dynamic range that is consistent with the analysis of microbial proteomes. In this paper we provide a comprehensive proteome-wide SRM assay repository resource for the important human pathogen group A Streptococcus (GAS), to remove a considerable bottleneck in the SRM workflow. We used a new proof-of-concept assays score to rank assays associated to a particular protein, based on the ability to detect assays in the tested conditions, resulting in 2731 medium- or high-scoring assays covering 1332 of 1905 ORFs in the reference strain SF370. Any laboratory with access to triple quadrupole mass spectrometers can use the SRM assays described in this manuscript to in a multiplexed manner quantify GAS proteins of interest. The additional use of stable isotope labeled reference peptides enables absolute protein copy number per cell estimations as previously described[10].

Searching for biases between proteins with low-, medium- and high-scoring SRM assays reveal that the majority of proteins with exclusively low-scoring assays tend to be shorter, less well annotated and less conserved among GAS strains. In contrast, the proteins associated with high-scoring SRM assays are longer, well annotated and conserved. As the dynamic range of SRM is higher than the dynamic range of the sample protein abundance we estimate that the missing proteins are due to lack of expression or proteins lost in the protein enrichment, digest or MS analysis. In addition, as the elution profile for tryptic peptides is not evenly distributed across the LC-chromatograms, certain peptides in particularly crowded areas are harder to confidently identify due to presence of interfering signals from other high abundant peptides[20]. Although the outlined work resulted in considerable coverage over the GAS SF370 genome, still a considerable proportion of predicted GAS genes remained elusive.

The iterative testing of more than 7500 distinct peptide sequences with corresponding SRM assays provides an indication of the spatial distribution of the GAS proteome. This exemplified

by high enrichments of cell walled anchored proteins in the surface compartment, and virulence factors in the secreted fraction. However, a substantial number of ribosomal proteins and central metabolic enzymes are detected in the extracellular fractions. In fact, around 10% of the detected proteins displayed similar abundance levels in more than one cellular compartment. There are probably several reasons for this, which are difficult to distinguish. Similar phenomena has been described earlier[34-36], and could be explained by spontaneous cell lysis or artifacts related to experimental procedures. However, it is also likely that certain proteins are present in more than one cellular compartment natively. This has been described in detail for several GAS proteins as examples, glycolytic proteins implicated as virulence factors located on the surface[37-39] or cell wall anchored proteins proteolytically released into the growth media [40-42].

We believe the SRM assay repository will become a useful resource for addressing central medical and molecular microbiological related questions regarding GAS in general as the transportability of the SRM assays across the known GAS protein universe was high. Relatively little effort is required to also cover other strain-specific SRM assays in the respiratory. Defining GAS proteome composition differences between clinical isolates and mutant strains *in vitro* and *in vivo* are examples of how SRM assay repository could be used. Awareness of SRM assay transportability to closely related species is essential if targeting GAS proteins in microbial ecologies, such as in pharyngitis *in vivo*.

In conclusion we have in this work provided a proteome-wide SRM assay repository resource for one of the most important human pathogens to facilitate SRM-MS analysis for this bacterium. As several assays can be transported to other species we expect that the reach of the resource extend beyond GAS. We believe that the iterative testing of all SRM assays and the construction of a novel SRM assay score model along with estimating protein specific biases for the differential scoring SRM assays increases the usefulness of the described resource.

Methods

Bacterial culture conditions

S. pyogenes M1 strains SF370, MGAS5005 and AP1 (strain 40/58 from the WHO Collaborating Centre for Reference and Research on Streptococci, Prague, Czech Republic) was cultured (37°C; 5% CO_2) in C-medium[43], Todd-Hewitt (TH) broth (Difco Laboratories) or in TH with supplements as indicated below or in Protein-reduced TH (PR-TH) broth[36] for secreted protein isolation. Supplements were added to TH as indicated at the following concentrations: 1, 5, 10, 20 or 50 % (V/V) citrate treated human plasma (Skåne University Hospital, SUS) 50 % (V/V) citrate treated mouse plasma from CD1 mice (SeraLab), 4 mg/ml human serum albumin (Sigma-Aldrich), 4 mg/ml essentially fatty acid free (~0.005%) human serum albumin (Sigma-Aldrich), 0.3 mg/ml human fibrinogen (Sigma-Aldrich), 1.2 mg/ml human IgG (Sigma-Aldrich), rifampicin at the following concentrations 0.25, 1.25, 2.5, 12.5 or 25 ng/ul, erythromycin at the following concentrations 0.1, 0.5, 1, 5 or 10 (µg/ml), hydrogen peroxide at the following concentrations 0.5 mM or 5 mM. Cultures were also grown at the following conditions: strict anaerobically (Elektrotek Workstation), room atmosphere, or pH at levels 5.5, 6.4, 7.3, 8.1 or 9

Subcellular protein isolation

Bacteria were generally harvested at exponential ($OD_{620\ nm}$=0.4-0.5) or at stationary phase ($OD_{620\ nm}$=0.7-0.8) by centrifugation 10 minutes at 2500 x g. To isolate intracellular proteins samples were treated as earlier described[1]. For surface-associated protein isolation, TBS washed cells were re-suspended in 20 mM Tris-HCl, 150 mM NaCl, 10 mM $CaCl_2$, 1 M D-arabinose, pH 7.6 to a concentration of 1.6 x 10^9 colony forming units (CFU) per ml. Samples were treated with 10 µg sequencing grade trypsin (Promega) per ml for 15 min at 37 °C[44,45]. Cells were removed by centrifugation at 1000 x g for 15 min at 4 °C and the resulting supernatant was treated as described below in the ’Protein digestion & peptide cleaning’ section except for more extensive washes during peptide cleaning for arabinose removal. Secreted proteins were isolated

from 22 µm filtered culture supernatants that were concentrated with Amicon Ultra-15 Centrifugal Filter Units, 30 MWCO (Millipore). The resulting concentrate was diafiltrated in the same filter unit type twice with 50 mM Tris-HCl, pH 8.35 and then once with 6 M Urea, 0.2 M Tris-HCl, pH 8.35.

Protein digestion & peptide cleaning

The proteins were reduced with 5 mM dithiothreitol (DTT) for 45 min at 37 °C, and alkylated with 25 mM iodoacetamide for 45 min before diluting the sample with 100 mM ammonium bicarbonate to a final urea concentration below 1.5M. Proteins were digested by incubation with trypsin (1/100, w/w) for at least 6 h at 37 °C. The peptides were cleaned up by C18 reversed-phase spin columns according to the manufacturer's instructions (Harvard Apparatus).

Shotgun tandem mass spectrometry analysis

The shotgun tandem and targeted mass spectrometry analysis was performed as previously described[1]. Briefly, the hybrid Orbitrap-LTQ XL mass spectrometer (Thermo Electron, Bremen, Germany) was coupled online to a split-less Eksigent 2D NanoLC system (Eksigent technologies, Dublin, CA, USA). Peptides were loaded with a constant flow rate of 10 µl/min onto a pre-column (Zorbax 300SB-C18 5 x 0.3 mm, 5 µm, Agilent technologies, Wilmington, DE, USA) and subsequently separated on a RP-LC analytical column (Zorbax 300SB-C18 150 mm x 75 µm, 3.5 µm, Agilent technologies) with a flow rate of 350 nl/min. The peptides were eluted with a linear gradient from 95% solvent A (0.1% formic acid in water) and 5% solvent B (0.1% formic acid in acetonitrile) to 40% solvent B over 55 minutes. The mass spectrometer was operated in data-dependent acquisition mode to automatically switch between Orbitrap-MS (from m/z 400 to 2000) and LTQ-MS/MS. Four MS/MS spectra were acquired in the linear ion trap per each Fourier Transform-MS scan which was acquired at 60,000 FWHM nominal resolution settings using the lock mass option (m/z 445.120025) for internal calibration. The dynamic exclusion list was restricted to 500 entries using a repeat count of two with a repeat

duration of 20 seconds and with a maximum retention period of 120 seconds. Precursor ion charge state screening was enabled to select for ions with at least two charges and rejecting ions with undetermined charge state. The normalized collision energy was set to 30%, and one microscan was acquired for each spectrum.

The data analysis was performed as previously described [1]. Briefly, the resulting MS2 data were searched with X! Tandem search engine, version 2009.04.01.1 with the k-score plugin[46], a common peptide and protein list was generated using the Trans-Proteomic pipeline, version 4.4.0[47]. All searches were performed with full-tryptic cleavage specificity, up to 2 allowed missed cleavages, a precursor mass error of 15 ppm and an error tolerance of 0.5 Da for the fragment ions. Because of the sample preparation cysteine carbamidomethylation was defined as fixed modification in the search parameters. A protein database with sequences for GAS SF370 (Genome ID 79812 from PATRIC) was used to match the individual spectra to certain peptides. The database was extended by decoy sequences to validate the resulting peptide-spectrum matches (PSMs)[48]. A 1% false-discovery rate (FDR) was then used to generate the final protein list with ProteinProphet. MS1-based quantification was done using SuperHirn[49]. Features were detected using SuperHirn using a retention time tolerance of 1, MS1 m/z tolerance of 10, MS2 PPM m/z tolerance of 30. Only features with charge 1-5 were included. Any feature for which more than one peptide could be identified at the 1% FDR, hence mapping to more than one protein, were discarded.

Generation of proteome-wide SRM assays

We generated experimentally validated SRM assay for three proteotypic peptides for each protein in the SF370 proteome. Transitions were generated from experimental MS2 spectra either from off gel electrophoresis fractionated cell lysates for a pool of GAS SF370 grown under various conditions or from crude synthetic peptides purchased from JPT (Berlin, Germany). The transitions were scored in a scoring scheme that favored y-ions over b-ions,

required both Q1/Q3 to be between 400 and 1500 M/Z, Q3 larger than Q1 was favored and precursor charge of 2 was preferred over other charge states. The four best transitions for each peptide were measured in none-scheduled SRM mode against the sample where the peptide was identified by MS2.

SRM analysis

SRM transition assays were constructed by testing the twenty most abundant peptide fragments for selected proteotypic peptides identified with high confidence in the LC-MS/MS experiments. Spiked in the RT-peptides (Biognosys AG, Zurich, Switzerland) allowed normalization of the retention time as previously described[20]. The SRM measurements were performed on a TSQ Vantage triple quadropole mass spectrometer (Thermo Electron, Bremen, Germany) equipped with a nanoelectrospray ion source (Thermo Electron). Chromatographic separations of peptides were performed on an Eksigent 1D NanoLC system (Eksigent technologies) using the same chromatographic conditions as described above for the Eksigent 2D NanoLC system connected to the hybrid Orbitrap-LTQ XL mass spectrometer. The LC was operated with a flow rate of 400 nl/min. The mass spectrometer was operated in SRM mode, with both Q1 and Q3 settings at unit resolution (FWHM 0.7 Da). A spray voltage of +1700 V was used with a heated ion transfer setting of 270ºC for desolvation. Data were acquired using the Xcalibur software (version 2.1.0). The dwell time was set to 10 ms and the scan width to 0.01 m/z. All collision energies were calculated using the formula: CE = (Parent m/z) x 0.034 + 3.314.

The data analysis was performed as previously described[20] using a 1% FDR. The resulting peptide abundances were exported into a database, where protein abundances were inferred by summing up the abundances for the peptides uniquely mapping to each protein[22].

References

1. Malmström, J. A. *et al.* Streptococcus pyogenes in human plasma: adaptive mechanisms analyzed by mass spectrometry based proteomics. *J Biol Chem* **287**, 1415–1425 (2011).
2. Hecker, M., Becher, D., Fuchs, S. & Engelmann, S. A proteomic view of cell physiology and virulence of Staphylococcus aureus. *Int J Med Microbiol* **300**, 76–87 (2010).
3. Poetsch, A., Haussmann, U. & Burkovski, A. Proteomics of corynebacteria: From biotechnology workhorses to pathogens. *Proteomics* **11**, 3244–3255 (2011).
4. Chao, T.-C. & Hansmeier, N. The current state of microbial proteomics: Where we are and where we want to go. *Proteomics* **12**, 638–650 (2012).
5. Cole, J. N., Barnett, T. C., Nizet, V. & Walker, M. J. Molecular insight into invasive group A streptococcal disease. *Nat Rev Microbiol* **9**, 724–736 (2011).
6. Malmström, L., Malmström, J. A. & Aebersold, R. Quantitative proteomics of microbes: Principles and applications to virulence. *Proteomics* **11**, 2947–2956 (2011).
7. Lange, V. *et al.* Targeted quantitative analysis of Streptococcus pyogenes virulence factors by multiple reaction monitoring. *Mol Cell Proteomics* **7**, 1489–1500 (2008).
8. Kuhner, S. *et al.* Proteome Organization in a Genome-Reduced Bacterium. *Science* **326**, 1235–1240 (2009).
9. Schmidt, A. *et al.* Absolute quantification of microbial proteomes at different states by directed mass spectrometry. *Mol Syst Biol* **7**, 510 (2011).
10. Malmström, J. A. *et al.* Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. *Nature* **460**, 762–765 (2009).
11. Stahl-Zeng, J. *et al.* High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites. *Mol Cell Proteomics* **6**, 1809–1817 (2007).
12. Teleman, J. *et al.* Automated selected reaction monitoring software for accurate label-free protein quantification. *J Proteome Res* **11**, 3766–3773 (2012).
13. Malmström, J. A., Lee, H. & Aebersold, R. Advances in proteomic workflows for systems biology. *Curr Opin Biotechnol* **18**, 378–384 (2007).
14. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol* **4**, 222 (2008).
15. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. Scoring proteomes with proteotypic peptide probes. *Nat Rev Mol Cell Biol* **6**, 577–583 (2005).
16. Picotti, P. *et al.* A database of mass spectrometric assays for the yeast proteome. *Nat Methods* **5**, 913–914 (2008).
17. Carapetis, J. R., Steer, A. C., Mulholland, E. K. & Weber, M. The global burden of group A streptococcal diseases. *The Lancet infectious diseases* **5**, 685–694 (2005).
18. Cunningham, M. W. Pathogenesis of group A streptococcal infections. *Clin Microbiol Rev* **13**, 470–511 (2000).
19. Deutsch, E. W., Lam, H. & Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* **9**, 429–434 (2008).
20. Malmström, L., Malmström, J. A., Selevsek, N., Rosenberger, G. & Aebersold, R. Automated Workflow for Large-Scale Selected Reaction Monitoring Experiments. *J Proteome Res* **11**, 1644–1653 (2012).
21. Vogel, C. & Marcotte, E. M. Calculating absolute and relative protein abundance from mass spectrometry-based protein expression data. *Nat Protoc* **3**, 1444–1451 (2008).
22. Malmström, L., Marko-Varga, G., Westergren-Thorsson, G., Laurell, T. & Malmström, J. A. 2DDB - a bioinformatics solution for analysis of quantitative proteomics data. *BMC Bioinformatics* **7**, 158 (2006).
23. Wu, C. C. & Yates, J. R. The application of mass spectrometry to membrane proteomics.

Nat Biotechnol **21**, 262–267 (2003).
24. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat Methods* **6**, 359–362 (2009).
25. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003).
26. Barsky, A., Gardy, J. L., Hancock, R. E. W. & Munzner, T. Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics* **23**, 1040–1042 (2007).
27. McNeil, L. K. *et al.* The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. *Nucleic Acids Res* **35**, D347–53 (2007).
28. Gillespie, J. J. *et al.* PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect Immun* **79**, 4286–4298 (2011).
29. Beres, S. B. *et al.* Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. *Proc Natl Acad Sci USA* **107**, 4371–4376 (2010).
30. Beres, S. B. *et al.* Molecular genetic anatomy of inter- and intraserotype variation in the human bacterial pathogen group A Streptococcus. *Proc Natl Acad Sci USA* **103**, 7059–7064 (2006).
31. Sumby, P. *et al.* Evolutionary origin and emergence of a highly successful clone of serotype M1 group a Streptococcus involved multiple horizontal gene transfer events. *J Infect Dis* **192**, 771–782 (2005).
32. McShan, W. M. *et al.* Genome sequence of a nephritogenic and highly transformable M49 strain of Streptococcus pyogenes. *J Bacteriol* **190**, 7773–7785 (2008).
33. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
34. Chhatwal, G. S. Anchorless adhesins and invasins of Gram-positive bacteria: a new class of virulence factors. *Trends Microbiol* **10**, 205–208 (2002).
35. Barinov, A. *et al.* Prediction of surface exposed proteins in Streptococcus pyogenes, with a potential application to other Gram-positive bacteria. *Proteomics* **9**, 61–73 (2009).
36. Lei, B., Mackie, S., Lukomski, S. & Musser, J. M. Identification and immunogenicity of group A Streptococcus culture supernatant proteins. *Infect Immun* **68**, 6807–6818 (2000).
37. Boël, G., Jin, H. & Pancholi, V. Inhibition of cell surface export of group A streptococcal anchorless surface dehydrogenase affects bacterial adherence and antiphagocytic properties. *Infect Immun* **73**, 6237–6248 (2005).
38. Lottenberg, R. *et al.* Cloning, sequence analysis, and expression in Escherichia coli of a streptococcal plasmin receptor. *J Bacteriol* **174**, 5204–5210 (1992).
39. Cork, A. J. *et al.* Defining the structural basis of human plasminogen binding by streptococcal surface enolase. *J Biol Chem* **284**, 17129–17137 (2009).
40. Raeder, R., Woischnik, M., Podbielski, A. & Boyle, M. D. P. A secreted streptococcal cysteine protease can cleave a surface-expressed M1 protein and alter the immunoglobulin binding properties. *Research in Microbiology* **149**, 539–548 (1998).
41. Nelson, D. C., Garbe, J. & Collin, M. Cysteine proteinase SpeB from Streptococcus pyogenes - a potent modifier of immunologically important host and bacterial proteins. *Biol Chem* **392**, 1077–1088 (2011).
42. Berge, A. & Björck, L. Streptococcal cysteine proteinase releases biologically active fragments of streptococcal surface proteins. *J Biol Chem* **270**, 9862–9867 (1995).
43. Collin, M. & Olsén, A. EndoS, a novel secreted protein from Streptococcus pyogenes with endoglycosidase activity on human IgG. *EMBO J* **20**, 3046–3055 (2001).
44. Solis, N., Larsen, M. R. & Cordwell, S. J. Improved accuracy of cell surface shaving proteomics in Staphylococcus aureus using a false-positive control. *Proteomics* **10**, 2037–2049 (2010).

45. Severin, A. *et al.* Proteomic analysis and identification of Streptococcus pyogenes surface-associated proteins. *J Bacteriol* **189**, 1514–1522 (2007).
46. Craig, R. & Beavis, R. C. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom* **17**, 2310–2316 (2003).
47. Keller, A., Eng, J., Zhang, N., Li, X.-J. & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* **1**, 2005.0017 (2005).
48. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**, 207–214 (2007).
49. Mueller, L. N. *et al.* SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7**, 3470–3480 (2007).

Acknowledgements

JM is funded by the Swedish Research Council (project 2008-3356), the Crafoord Foundation (ref nr 20100892) and the Swedish Foundation for Swedish Research (FFL4). CK is supported by the Swedish Research Council postdoctoral fellowship (project 2010-996). RA is supported by the European Research Council (grant #ERC-2008-AdG 233226), SystemsX.ch, the Swiss initiative for systems biology, the Swiss National Science Foundation grant Nr. 3100A0-130530 and by the European Union Seventh Framework Program PROSPECTS (Proteomics Specification in Space and Time, Grant HEALTH-F4-2008). We thank Karin M. Hansson and Mats Mågård for excellent technical assistance

Author Contributions

Conceived and designed the experiments: CK LM RA JM. Performed the experiments: CK JM. Bioinformatic analysis: LM. Analyzed the data: CK LM RA JM. Wrote the paper: CK LM RA JM.

Additional information

The authors declare no competing financial interest.

Figures Legends

Figure 1: Construction of a proteome-wide SRM assay repository. a) Graphical representation of the enriched Group A Streptococci (GAS) cellular compartments. Repeated enrichment of protein pools from the cellular compartments and bacterial states were digested using trypsin and analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS). b) Outline of the strategy used to construct a spectral library from where the low scoring selected reaction monitoring (SRM) assays were extracted. For high-abundant proteotypic peptides (PTP's) the SRM assays were determined directly in biological samples, whereas for medium- and low-abundant PTP peptides were synthesized and analyzed with LC-MS/MS. c) To increase the confidence of the individual SRM assays the low-scoring SRM assays were tested extensively in complex mixture of GAS tryptic digest using SRM.

Figure 2: Rule based assay score modeling can select high performing assays depending on cellular compartment and cellular state. SRM assays were tested repeatedly in mixtures of GAS tryptic digests from different subcellular compartments and bacterial states. a) Statistics over the repetitive testing of all SRM assays. b) The assays were divided up into three categories, low-, medium-, and high-scoring SRM assays based on a rule-based assay score model. c) Assay score distribution for high scoring SRM assays.

Figure 3: Protein identification biases across functional categories and ORF properties. Iterative testing of the developed SRM assays in complex biological mixtures of GAS tryptic digests resulted in subdivision of the proteins into three SRM assays-score categories; proteins with low-, medium- or high-scoring SRM assays. Using the SRM assay score categories we determined biases among associated proteins within the three categories. a) Proportion of

extracted ion chromatogram (XIC) intensities associated with proteins with at least one high or medium-scoring SRM assay or proteins with low-scoring SRM assays. NMPDR was used to categorize proteins; Protein Metabolism includes categories Amino Acids and Derivatives and Protein Metabolism; Miscellaneous includes categories Clustering-based subsystems, Miscellaneous, Phages, Prophages, Transposable elements, Plasmids, Regulation and Cell signaling, Respiration, Stress Response, Cell Division and Cell Cycle and Cell Wall and Capsule; Carbohydrate Metabolism includes category Carbohydrates; RNA & DNA Metabolism includes categories Nucleosides and Nucleotides, DNA Metabolism and RNA Metabolism; Other Metabolism includes categories Phosphorus Metabolism, Potassium metabolism, Fatty Acids, Lipids, and Isoprenoids, Cofactors, Vitamins, Prosthetic Groups and Pigments, Sulfur Metabolism, Iron acquisition and metabolism, Nitrogen Metabolism, Membrane Transport and Metabolism of Aromatic Compounds Virulence includes categories Virulence and Virulence, Disease and Defense.

b) Genome-wide correlation between SRM assay-score and ORF length. c) Correlation between SRM assay score and relative degree of protein conservation across 13 GAS strains as determined with TOP-BLAST hits with SF370 as reference.

Figure 4: Spatial distribution of proteins with high scoring SRM assays. Testing of all SRM assays in three subcellular compartments enabled the construction of a subcellular distribution map for the proteins with high scoring assays. a-b) predominately intracellular proteins, c) surface-associated proteins, d) secreted proteins and e-f) proteins with split subcellular compartmentalization. Red lines in panels a-f represent the average distribution of the clusters. Subsequently the identified proteins were grouped into NMPDR subsystems and visualized using Cytoscape. g) Outline of the GAS proteome network topology, where circles represent NMPDR subsystems where all proteins predominantly have the same subcellular location,

secreted, surface associated or intracellular, according to the subcellular protein profiles in Figure 4a-d. Rectangles represent NMPDR subsystems where an equal number of members have opposing subcellular location profiles. The localization of the rectangles in the network is influenced by the edges, which represent protein members that belong to more than one NMPDR subsystem. Increasing node size represents increasing number of member proteins. The color represents average SRM assay score, where red indicates NMPDR subsystems with high-average SRM assay score and black indicating NMPDR subsystems with low average SRM assays score. For full details of NMPDR subsystems see Supplementary Figure S1.

Figure 5: SRM assay transportability to related species. All SRM assays were developed on basis of the GAS strain SF370 genome. The degree of SRM assay transportability within selected species was determined phylogeny clustering of respective *rpoB* gene and mapping medium- and high-scoring SRM assays on to respective genome. a) Transportability for the SRM assays across 75 genomes within the Firmicutes phylum. b) Average ORF genome coverage of high- and medium-scoring SRM assays within taxonomic ranks. Boxes extend from the 25th to 75th percentiles and error bars represent minimum to maximum values. c) View of SRM assay transportability for 13 GAS genomes deposited in the public domain.

Figure 1

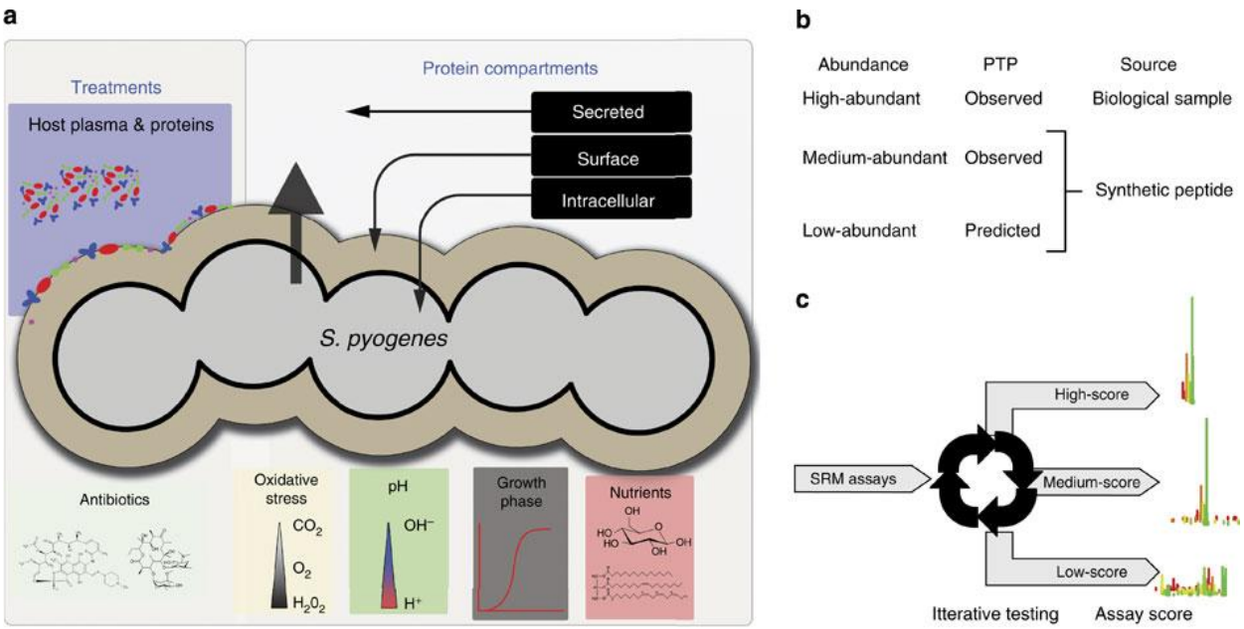

a
Treatments
Host plasma & proteins
Protein compartments
Secreted
Surface
Intracellular
S. pyogenes
Antibiotics
Oxidative stress
CO2
O2
H2O2
pH
OH−
H+
Growth phase
Nutrients
b
Abundance
PTP
Source
High-abundant
Observed
Biological sample
Medium-abundant
Observed
Synthetic peptide
Low-abundant
Predicted
c
SRM assays
High-score
Medium-score
Low-score
Itterative testing
Assay score

Figure 2

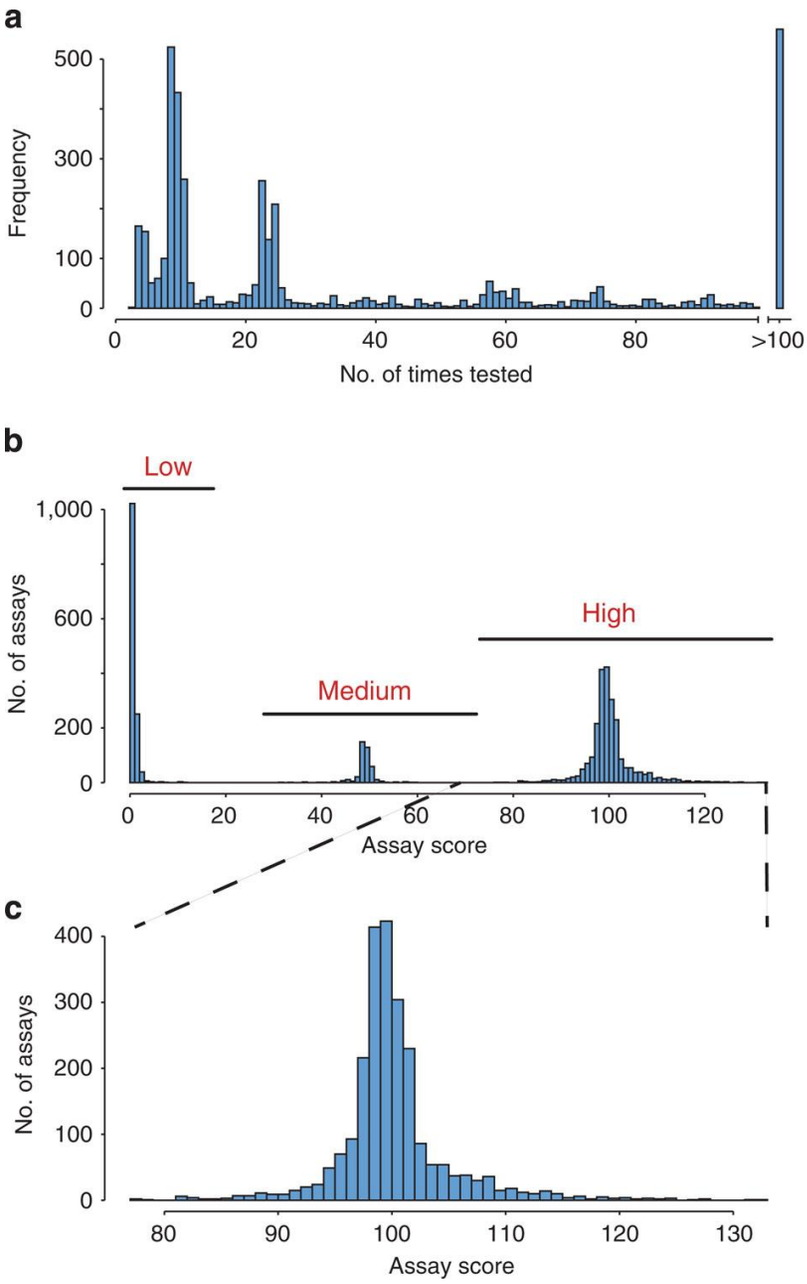

a
Frequency
500
300
100
0
0
20
40
60
80
>100
No. of times tested
b
Low
1,000
600
200
0
No. of assays
High
Medium
0
20
40
60
80
100
120
Assay score
c
400
300
200
100
0
No. of assays
80
90
100
110
120
130
Assay score

Figure 3

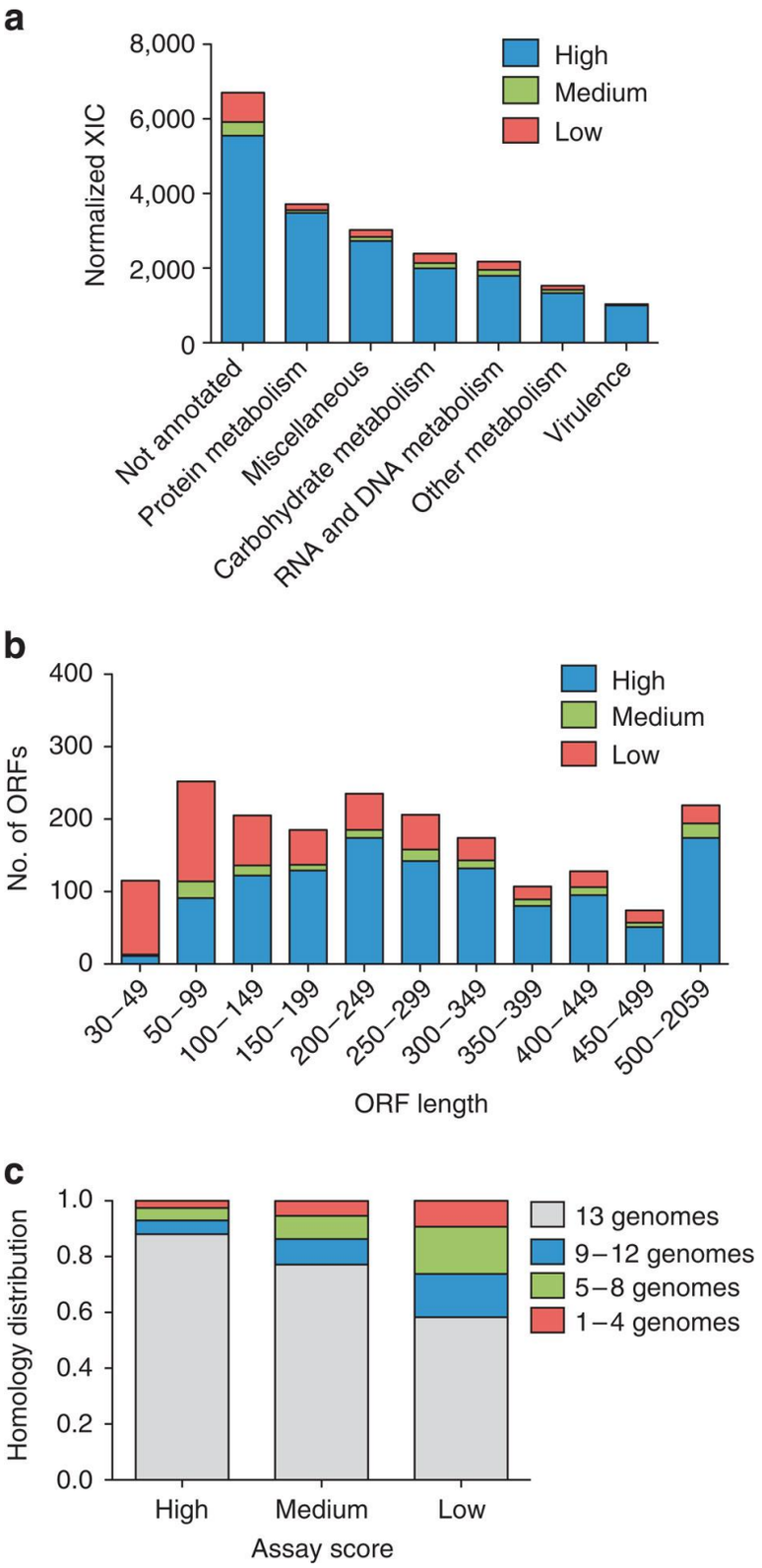

a
8,000
6,000
4,000
2,000
0
Normalized XIC
High
Medium
Low
Not annotated
Protein metabolism
Miscellaneous
Carbohydrate metabolism
RNA and DNA metabolism
Other metabolism
Virulence
b
400
300
200
100
0
No. of ORFs
High
Medium
Low
30–49
50–99
100–149
150–199
200–249
250–299
300–349
350–399
400–449
450–499
500–2059
ORF length
c
1.0
0.8
0.6
0.4
0.2
0.0
Homology distribution
13 genomes
9–12 genomes
5–8 genomes
1–4 genomes
High
Medium
Low
Assay score

Figure 4

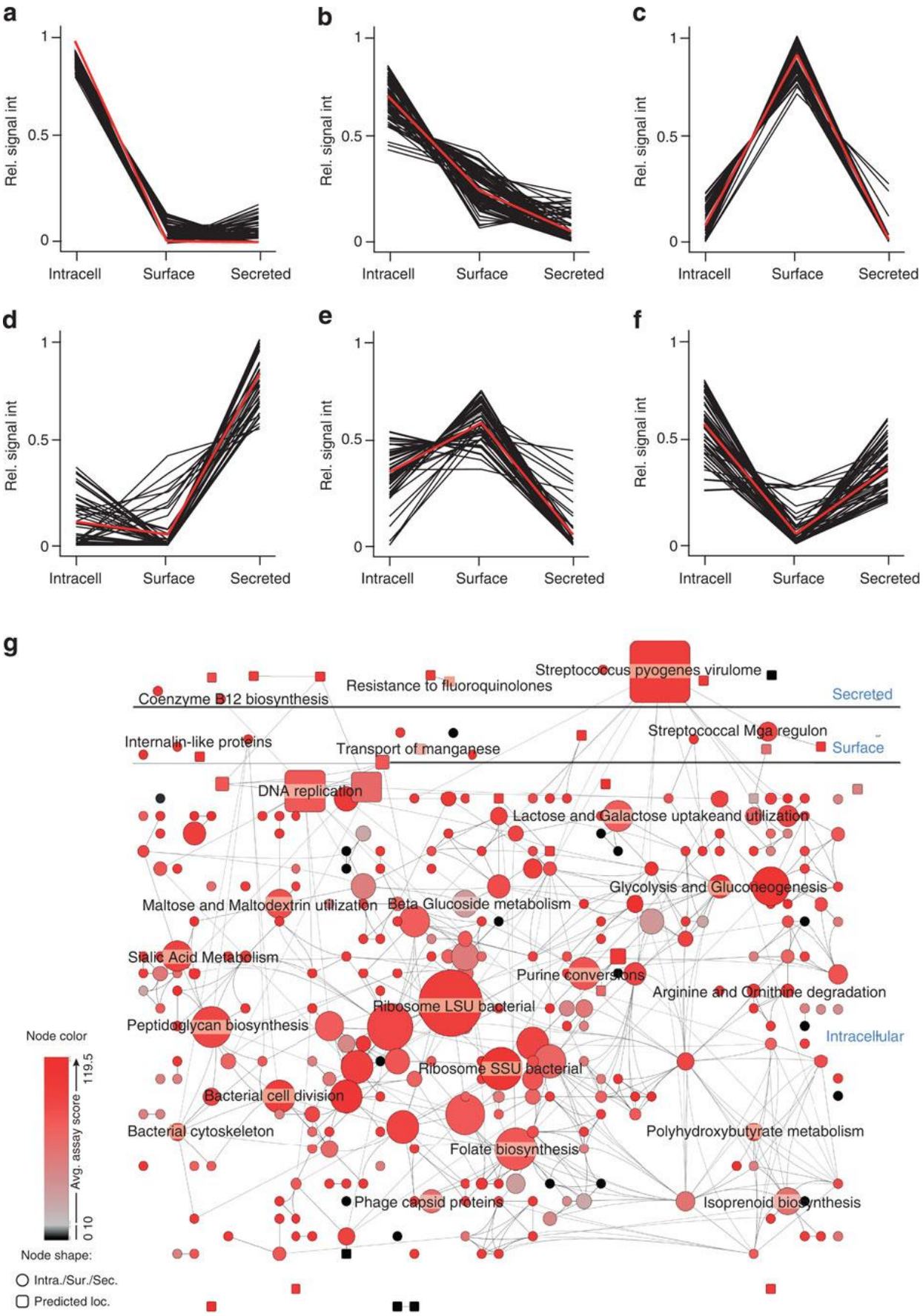
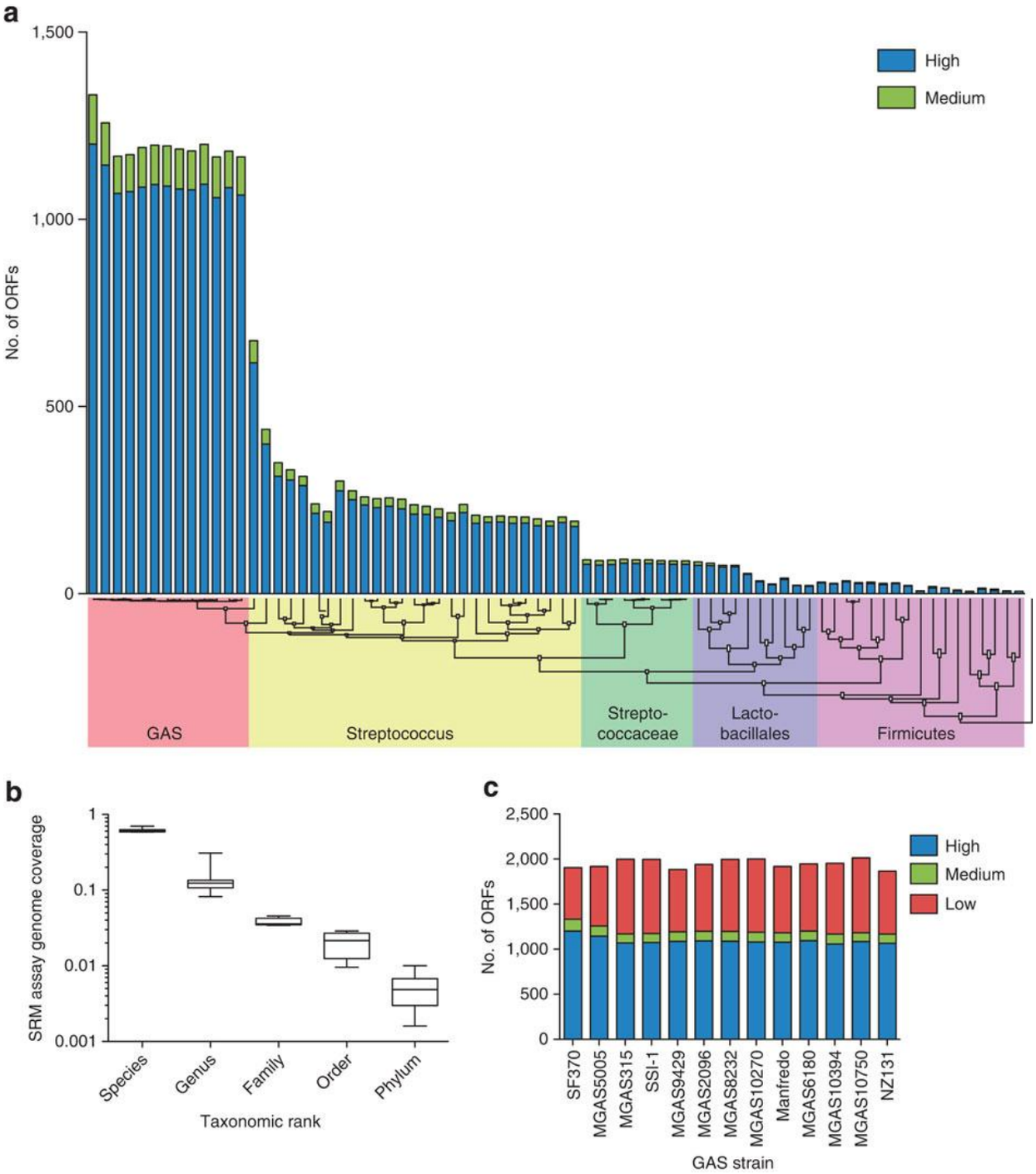

a
b
c
d
e
f
Rel. signal int
1
0.5
0
Intracell
Surface
Secreted
g
Streptococcus pyogenes virulome
Resistance to fluoroquinolones
Coenzyme B12 biosynthesis
Secreted
Streptococcal Mga regulon
Internalin-like proteins
Transport of manganese
Surface
DNA replication
Lactose and Galactose uptakeand utilization
Glycolysis and Gluconeogenesis
Maltose and Maltodextrin utilization
Beta Glucoside metabolism
Sialic Acid Metabolism
Purine conversions
Arginine and Ornithine degradation
Ribosome LSU bacterial
Peptidoglycan biosynthesis
Intracellular
Ribosome SSU bacterial
Bacterial cell division
Bacterial cytoskeleton
Polyhydroxybutyrate metabolism
Folate biosynthesis
Phage capsid proteins
Isoprenoid biosynthesis
Node color
Avg. assay score
119.5
0
10
Node shape:
Intra./Sur./Sec.
Predicted loc.

Figure 5


a
1,500
1,000
500
0
No. of ORFs
High
Medium
GAS
Streptococcus
Strepto-
coccaceae
Lacto-
bacillales
Firmicutes
b
SRM assay genome coverage
1
0.1
0.01
0.001
Species
Genus
Family
Order
Phylum
Taxonomic rank
c
2,500
2,000
1,500
1,000
500
0
No. of ORFs
High
Medium
Low
SF370
MGAS5005
MGAS315
SSI-1
MGAS9429
MGAS2096
MGAS8232
MGAS10270
Manfredo
MGAS6180
MGAS10394
MGAS10750
NZ131
GAS strain