RIGBE GEBREMICHAEL WELDATSADIK

APPLICATION OF POOL-SEQ FOR VARIATION DETECTION AND PROTEOGENOMIC DATABASE CREATION IN ß-HEMOLYTIC STREPTOCOCCI

INSTITUTE OF BIOTECHNOLOGY
HELSINKI INSTITUTE OF LIFE SCIENCE HILIFE AND
DEPARTMENT OF BIOSCIENCES
FACULTY OF BIOLOGICAL AND ENVIRONMENTAL SCIENCES
DOCTORAL PROGRAMME IN BIOMEDICINE
UNIVERSITY OF HELSINKI
Application of Pool-seq for variation detection and proteogenomic database creation in β-hemolytic streptococci

Rigbe Gebremichael Weldatsadik

Institute of Biotechnology
Helsinki Institute of Life Sciences (HiLIFE)
Faculty of Biological and Environmental Sciences
University of Helsinki
Helsinki, Finland

Medicum
Department of Bacteriology and Immunology
Faculty of Medicine
University of Helsinki
Helsinki, Finland

Doctoral Programme in Biomedicine (DPBM)
University of Helsinki
Helsinki, Finland

Doctoral thesis

To be presented for public examination with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki, in Auditorium 1041, Biocenter 2, on the 16th of April, 2021 at 1 o’clock
ABSTRACT

Proteogenomics is an emerging field that combines genomic (transcriptomic) and proteomic data with the aim of improving gene models and identification of proteins. Technological advances in each domain increase the potential of the field in fostering further understanding of organisms. For instance, the current low cost and fast sequencing technologies have made it possible to sequence multiple representative samples of organisms thus improving the comprehensiveness of the organisms’ reference proteomes. At the same time, improvements in mass spectrometry techniques have led to an increase in the quality and quantity of proteomics data produced, which are utilized to update the annotation of coding sequences in genomes.

Sequencing of pooled individual DNAs (Pool-seq) is one method for sequencing large numbers of samples cost effectively. It is a robust method that can accurately identify variations that exist between samples. Similar to other proteogenomics methods such as the sample specific databases derived from RNA-seq data, the variants from Pool-seq experiments can be utilized to create variant protein databases and improve the completeness of protein reference databases used in mass spectrometry (MS)-based proteomics analysis. In this thesis work, the efficiency of Pool-seq in identifying variants and estimating allele frequencies from strains of three β-hemolytic bacteria (GAS, GGS and GBS) is investigated. Moreover, in this work a novel Python package (‘PoolSeqProGen’) for creating variant protein databases from the Pool-seq experiments was developed. To our knowledge, this was the first work to use Pool-seq for sequencing large numbers of β-hemolytic bacteria and assess its efficiency on such genetically polymorphic bacteria. The ‘PoolSeqProGen’ tool is also the first and only tool available to create proteogenomic databases from Pool-seq data.

The accuracy of the Pool-seq method in variation identification and allele frequency estimation was evaluated by comparing with the variants identified from individual sequencing of samples in the pools. Pool-seq had a high sensitivity (>90%) in identifying variants from all the 6 GAS and majority of the 40 GBS individually sequenced genome samples. The allele frequency estimates from the Pool-seq and individual sequencing of the GBS data also showed high correlation (R=0.96). The effect of sequencing coverage on the sensitivity of Pool-seq was assessed by downsampling analyses.

The Pool-seq driven proteogenomic databases created were compared to conventional reference protein databases, and ~5% more and ~10% less peptides
were identified than the single genome and multi-genome based protein databases respectively. In comparison to other proteogenomic based databases created from ab initio gene predictions by Prodigal (after assembling with MEGAHIT) and 6-frame translation based ‘Metapeptides’ created by the Sixgill tool, respectively ~4% and 19% more peptides were identified.

For organisms such as the β-hemolytic bacteria GAS, GBS and GGS that have open pangenomes, the sequencing and annotation of multiple representative strains is paramount in advancing our understanding of these human pathogens and in developing mass spectrometry databases. Due to the increasing use of MS in diagnostics of infectious diseases, this in turn translates to better diagnosis and treatment of the diseases caused by the pathogens and alleviating their devastating burdens on the human population. In this thesis, it is demonstrated that Pool-seq can be used to cost effectively and accurately identify variations that exist among strains of these polymorphic bacteria. In addition, the utility of the tool developed to extend single genome based databases and thereby improve the completeness of the databases and peptide/protein identifications by using variants identified from Pool-seq experiments is illustrated.


Pool-sekvensoinnilla luotuja proteogenomisia tietokantoja verrattiin lisäksi tavanomaisiin vertailuproteiinitietokantoihin. Tunnistimme noin 5% enemmän peptidejä kuin yksittäisiin genomeihin perustuva menetelmä ja noin 10% vähemmän peptidejä kuin useisiin eri genomeihin perustuvissa tietokannoissa. Verrattuna muilhin proteogenomisiin tietokantoihin, jotka on luotu Prodigalin ab initio -geeniennusteista (MEGAHIT) ja Sixgill-työkalun luomiin 6-kehykisiin käännöspohjaisiin 'Metapeptideihin', tunnistettiin vastaavasti noin 4% ja 19% enemmän peptidejä.

# CONTENTS

1 Literature review ................................................................................................................................. 11  
1.1 From DNA to proteins ....................................................................................................................... 11  
1.2 DNA sequencing .............................................................................................................................. 11  
1.3 Pooled sequencing (Pool-seq) ......................................................................................................... 13  
1.4 Pool-seq variant detection .............................................................................................................. 16  
1.5 MS-based bottom up proteomics workflow .................................................................................... 18  
1.6 Protein databases ............................................................................................................................ 21  
1.7 β-hemolytic streptococci ................................................................................................................ 23  
1.8 Application of MS for bacterial characterization ........................................................................... 24  

2 Study aims ........................................................................................................................................... 26  

3 Materials and methods ........................................................................................................................ 27  
3.1 DNA isolation and pooling .............................................................................................................. 27  
3.2 DNA sequencing and analysis ....................................................................................................... 28  
3.3 Protein extraction and digestion .................................................................................................... 28  
3.4 LC-MS/MS ....................................................................................................................................... 29  
3.5 MS/MS analysis .............................................................................................................................. 29  

4 Results and discussion ........................................................................................................................ 31  
4.1 A sequencing coverage of ~200-250X is optimal for accurately identifying variants from bacterial Pool-seq data .................................................................................................................. 31  
4.2 Pool-seq incompatible variant calling results in substantial reduction of sensitivity ......................... 35  
4.3 Pool-seq is robust for accurate SNP detection and allele frequency estimation ................................ 36  
4.4 Pool-seq driven proteogenomics database generation ...................................................................... 41  
4.5 The Pool-seq driven proteogenomic databases allow identification of identical tryptic peptides with different variant profiles ......................................................................................... 44  
4.6 The Pool-seq driven proteogenomic databases result in more peptide identification compared to single genome databases ......................................................................................................................... 46  

5 Concluding remarks and future perspectives ...................................................................................... 54  

6 References .......................................................................................................................................... 58
# ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF</td>
<td>Allele Frequency</td>
</tr>
<tr>
<td>API</td>
<td>Analytical profile index</td>
</tr>
<tr>
<td>BWA</td>
<td>Burrows-Wheeler Aligner</td>
</tr>
<tr>
<td>CID</td>
<td>Collision Induced Dissociation</td>
</tr>
<tr>
<td>DDA</td>
<td>Data Dependent Accusation</td>
</tr>
<tr>
<td>ddNTP</td>
<td>Dideoxynucleotide triphosphate</td>
</tr>
<tr>
<td>DIA</td>
<td>Data Independent Accusation</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>Deoxyribonucleotide triphosphate</td>
</tr>
<tr>
<td>E&amp;R</td>
<td>Evolve and Resequence</td>
</tr>
<tr>
<td>ELISA</td>
<td>Enzyme-linked immunosorbent assay</td>
</tr>
<tr>
<td>ENA</td>
<td>European Nucleotide Archive</td>
</tr>
<tr>
<td>ESI</td>
<td>Electrospray ionization</td>
</tr>
<tr>
<td>FDR</td>
<td>False Discovery Rate</td>
</tr>
<tr>
<td>GAS</td>
<td>Group A streptococcus</td>
</tr>
<tr>
<td>GBS</td>
<td>Group B streptococcus</td>
</tr>
<tr>
<td>GGS</td>
<td>Group G streptococcus</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome Wide Association Studies</td>
</tr>
<tr>
<td>HPLC</td>
<td>High Pressure Liquid Chromatography</td>
</tr>
<tr>
<td>INDEL</td>
<td>Insertions and Deletions</td>
</tr>
<tr>
<td>LC</td>
<td>Liquid Chromatography</td>
</tr>
<tr>
<td>MALDI-TOF</td>
<td>Matrix-assisted laser desorption ionization time-of-flight</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>MS</td>
<td>Mass spectrometry</td>
</tr>
<tr>
<td>MS/MS</td>
<td>Tandem MS</td>
</tr>
<tr>
<td>NGS</td>
<td>Next Generation Sequencing</td>
</tr>
<tr>
<td>PEP</td>
<td>Posterior Error Probability</td>
</tr>
<tr>
<td>Pool-seq</td>
<td>Pooled sequencing</td>
</tr>
<tr>
<td>PSM</td>
<td>Peptide Spectrum Match</td>
</tr>
<tr>
<td>RAD</td>
<td>Restriction site Associated DNA</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>RNA sequencing</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
</tbody>
</table>
ORIGiNAL PUBLiCATiONS


Contribution: RGW participated in the study design together with TSJ and the SalWe GetItDone consortium. RGW designed and performed all the data analysis, prepared all the figures and wrote the manuscript.


Contribution: RGW developed the Python package and wrote the manuscript.


Contribution: RGW designed the study together with TSJ and MV, performed all the data analysis, prepared all the figures and wrote the manuscript.
1 LITERATURE REVIEW

1.1 From DNA to proteins

Deoxyribonucleic acid (DNA) is a molecule that holds the genetic instructions which specify the structure and function of cells in all living things and are passed down from one generation to the next. These instructions are embedded in the arrangements of the 4 nucleotide bases A (Adenine), T (Thymine), G (Guanine) and C (Cytosine), and are first transcribed into messenger RNA (mRNA) and then translated to proteins which carry out most of the task in the organism’s body. This one-way flow of information from DNA to mRNA and finally to protein proposed by Crick in 1958 (Crick, 1958) is known as the central dogma of molecular biology (Figure 1).

![Figure 1.](https://example.com/central-dogma.png)  

**Figure 1.** The central dogma of molecular biology that depicts the one way genetic information flow from DNA to mRNA and finally to protein (which are composed of 20 amino acids specified by codons). Created with BioRender.com

1.2 DNA sequencing

The ability to read what is in a DNA sequence is a prerequisite for decoding the instructions contained in it. DNA sequencing is determining the exact order of the nucleotides in a certain segment of a DNA molecule. It was almost after two decades,
in 1977, since the molecular structure of DNA was resolved (Watson and Crick, 1953), that the first DNA genome of the bacteriophage PhiX174 was sequenced (Sanger et al., 1977) using one of the first generation sequencing methods, i.e. Sanger’s chain termination method (Sanger and Coulson, 1975; Sanger, Nicklen, and Coulson, 1977). Prior to this, there were sequencing endeavors that targeted proteins (Edman et al., 1950; Sanger and Thompson, 1953; Sanger and Tuppy, 1951) and the comparatively simpler RNA molecules (Adams et al., 1969; Brownlee and Sanger, 1967; Wu and Kaiser, 1968) and also shorter stretches of bacteriophage DNA using DNA polymerase (Padmanabhan, Jay, and Wu, 1974; Sanger et al., 1973). Besides Sanger’s chain termination method, Maxam–Gilbert’s chemical cleavage method (Maxam and Gilbert, 1977) was extensively used at the time.

In Sanger’s chain termination method, four reactions that contained template DNA strand hybridized to a primer and chain terminating nucleotides (ddNTPs) as well as normal nucleotides (dNTPs) were used to synthesize different sized fragments of the complementary DNA using DNA polymerase. Meanwhile, in Maxam–Gilbert’s chemical cleavage approach, chemical treatments (acids, hydrazines and salts) were applied on the DNA that is radioactively labeled on the 5’ end to create a break in a small proportion of bases which allow the creation of different sized fragments when the DNA is cleaved at the chemically modified site. In both methods polyacrylamide gel electrophoresis was used to separate the fragments based on size.

Various technical improvements in the laboratory as well as in software technologies such as the adaptation of a single reaction containing all ddNTPS each tagged with a unique fluorescent label (Prober et al., 1987; Smith et al., 1986), capillary electrophoresis (Jorgenson and Lukacs, 1981; Kasper et al., 1988), double strand DNA sequencing (Zhang et al., 1988), the Phred quality metrics (Ewing et al., 1998; Ewing and Green, 1998) and Celera assembler (Myers, 2000), enabled the sequencing of larger and more complicated genomes using Sanger sequencing in the 1990s and early 2000s including, *Saccharomyces cerevisiae* (Goffeau et al., 1996), *Caenorhabditis elegans* (The C.elegans Sequencing Consortium, 1998) and the human genome (International Human Genome Sequencing Consortium, 2001).

After the completion of the Human Genome project, the next wave of sequencing methodologies (known as next generation sequencing, NGS) took over Sanger sequencing and continue to be the prominent methods of sequencing to date. These technologies are massively parallel with DNA fragment templates immobilized on solid surfaces or on beads and clonally amplified in vitro before sequencing by synthesis (using DNA polymerase) or by ligation (using DNA ligase)
where the fluorescent signal or pH change introduced by the incorporation of a base is recorded to infer the type and number of bases added (Margulies et al., 2005; McKernan et al., 2009; Ronaghi et al., 1996). They include the 454, ABI SOLiD, Illumina and Ion Torrent platforms that differ in the kind of amplification, sequencing and detection methods they employ (reviewed in Heather and Chain 2016; Mardis 2009; Metzker 2010; Shendure et al. 2017; Voelkerding, Dames, and Durtschi 2009). The introduction of a commercial NGS instrument, the Genome Sequencer 20 (GS20) in 2005 by 454, and afterwards by different competitor companies resulted in the ‘democratization’ of sequencing and also in the plummeting of sequencing cost. NGS technologies produce millions to billions of short sequence reads compared to Sanger sequencing which produces a single, albeit longer, read per run.

A couple of third generation real time, single molecule sequencing platforms (Clarke et al., 2009; Levene, 2003) are also available that aim to remedy errors and biases introduced by the amplification stage in NGS technologies and produce longer reads. These are especially important in genome assembly, haplotype phasing and structural variant identification (Giani et al., 2020). The base calling accuracy of these technologies used to be much lower than NGS technologies but recent advancements such as the HiFi (‘High Fidelity’) long-reads introduced by PacBio are reported to be highly accurate (Wenger et al., 2019).

A genome can be sequenced in its entirety (termed whole genome sequencing) or part of it can be targeted and sequenced such as the protein coding regions in exome sequencing. Over the past decade, sequencing cost decreased at an unprecedented rate leading to whole genome sequencing of large numbers of model and non-model organisms and enabling various kinds of studies including comparative genomics such as between eukaryotes (Batzoglou, 2000; Rubin, 2000), prokaryotes (Abby and Daubin, 2007) and different strains of the same species (Tettelin et al., 2005).

### 1.3 Pooled sequencing (Pool-seq)

Population level studies require the sequencing of large numbers of individuals to elucidate the genetic basis of complex diseases and traits through the identification of polymorphic loci and allele frequency differences. And although sequencing cost is plummeting at a high rate, the $1000 per genome goal is not yet achieved (Schwarze et al., 2020) and especially for population level studies, the cost of
sequencing is still steep for many labs. Therefore there is a need for cost-effective sequencing approaches such as pooled sequencing (Pool-seq).

Pool-seq involves constructing a single DNA library from several individual genomes and the sequencing of that library usually with NGS. This is unlike individual sequencing that requires individual libraries to be prepared for every genome (Figure 2).

![Figure 2. The steps involved in Pool-seq. After DNA is isolated from multiple samples, it is pooled before library preparation, unlike in individual sequencing in which a library is prepared from each DNA genome isolated. The prepared library is then sequenced usually with NGS platforms such as Illumina. Created with BioRender.com.](image)

Pool-seq is mainly used to cut back on the cost and effort of individual sequencing in population genomics studies. It has been utilized in numerous study types including GWAS, Evolve & Resequence and Bulk segregant analysis in organisms of varying genome sizes (Amaral et al., 2011; Boitard et al., 2012; Burke et al., 2010; Calvo et al., 2010; Cheng et al., 2012; Kaartokallio et al., 2016; Kolaczkowski et al., 2011; Lamichhaney et al., 2012; Micheletti and Narum, 2018; Nejentsev et al., 2009; Turner et al., 2010, 2011; Van Tassell et al., 2008; Zhu et al., 2012). To identify the
samples in the pools, the DNA fragments can be barcoded before pooling (Smith et al., 2010), or a combinatorial/overlapping pooling strategy can be adopted (Patterson and Gabriel, 2009; Wang et al., 2013) but both methods incur additional cost and effort. Pooling can also be used with other cost reducing sequencing techniques such as exome-sequencing (Calvo et al., 2010; Kaartokallio et al., 2016) and RAD-sequencing (Amaral et al., 2011; Van Tassell et al., 2008), methods that mainly achieve the economic benefits through targeted sequencing.

The cost reduction in Pool-seq is as a result of the lower chance of sequencing the same reads repeatedly (unlike in individual sequencing) since usually large number of samples are pooled (Gautier et al., 2013; Schlötterer et al., 2014). This also reduces sampling variance leading to accurate population-wide allele frequency estimates (Anand et al., 2016; Futschik and Schlötterer, 2010). The accuracy of the allele frequency estimates and variation discovery improves with increasing sequencing coverage and pool size, while higher sequencing error rates and unequal representation of individual genomes (especially in small sized pools) have the opposite effect (Anand et al., 2016; Futschik and Schlötterer, 2010; Gautier et al., 2013; Pérez-Enciso and Ferretti, 2010; Schlötterer et al., 2014). Besides such common factors, others such as the potential with-in and across-sample amplification bias, and reference allele preferential bias have been shown to have an impact on the accuracy of allele frequency estimates especially in disease association studies (Chen et al., 2012). In addition to re-sequencing of organisms that were previously whole genome sequenced, recently Pool-seq data was utilized for de-novo rough draft assemblies (with Transcriptome guided scaffolding of contigs) to detect genomic diversity in species that lack good quality reference genomes (Kurland et al., 2019; Neethiraj et al., 2017).

Pool-seq has been successfully applied to study low frequency variants in complex diseases such as IBD and type1 diabetes (Calvo et al., 2010; Momozawa et al., 2011; National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC) et al., 2011; Nejentsev et al., 2009; Out et al., 2009). However, it is not suitable for studying rare variants (AF < 0.01) as it is difficult to distinguish such variants from sequencing errors (Anand et al., 2016; Druley et al., 2009; Schlötterer et al., 2014). Moreover, current Pool-seq setups that use NGS technologies are not favorable for studies that rely on linkage disequilibrium and haplotype information since the sequence reads are short making it difficult to associate which of the reads derive from the same haplotypes (Micheletti and Narum, 2018; Schlötterer et al., 2014). Nonetheless there are tools that can predict haplotype frequencies from Pool-seq data when the
haplotypes are already known or can be predicted from existing databases (Cao and Sun, 2015; Kessner, Turner, and Novembre, 2013; Long et al., 2011).

Population genetics studies rely on accurate allele frequency estimates. Given large enough sequencing depth and pool sizes, Pool-seq based allele frequency estimates have been reported to be robust by several studies that assessed their accuracy using individual sequencing/genotyping, allele frequencies reported in publicly available databases and simulations (Amaral et al., 2011; Anand et al., 2016; Bansal et al., 2010; Hajirasaouliha et al., 2008; Holt et al., 2009; Ingman and Gyllensten, 2009; Margraf et al., 2011; Out et al., 2009; Rellstab et al., 2013; Ryu et al., 2018; Shaw et al., 1998; Van Tassell et al., 2008; Wang et al., 2016; Zhu et al., 2012). On the other hand, with smaller size pools and low sequencing coverage, Pool-seq was shown to be inadequate for estimating allele frequencies of especially low frequency variants in association studies of complex diseases (Chen et al., 2012; Day-Williams et al., 2011; Guo et al., 2013; Harakalova et al., 2011). The classical estimators in population genetics have been modified/adopted for Pool-seq data and tools for inferring patterns of variability and differentiation at the population level from these data have been developed (Boitard et al., 2013; Ferretti, Ramos-Onsins, and Pérez-Enciso, 2013; Futschik and Schlötterer, 2010; Kofler et al., 2011; Kofler, Pandey, and Schlotterer, 2011; Pérez-Enciso and Ferretti, 2010).

1.4 Pool-seq variant detection

After the DNA sequences of two or more samples are determined through sequencing, it is possible to compare them to identify genetic variants. Accurate identification of the genetic differences that exist at different levels, for instance between populations of the same species, is vital in understanding the genetic basis for the existing phenotypic variations. Variant calling is the process of identifying loci in a given genome that exhibit differences (such as SNPs and INDELS) compared to a representative (‘reference’) genome. It involves quality checking and preprocessing steps before usually aligning the sequence reads to a reference genome using tools such as BWA (Li and Durbin, 2009) and Bowtie (Langmead and Salzberg, 2012), even though there are also tools capable of calling variants without using reference genomes (Chan et al., 2016; Lopez-Maestre et al., 2016; Peterlongo et al., 2010; Ratan et al., 2010). One of the main challenges in this process is the distinguishing of errors from true variant calls. Most variant callers account for this error probability in the likelihood calculations of their Bayesian model, which can be
summarized as $\Pr(G \mid R) \propto \Pr(R \mid G) \times \Pr(G)$, where $G$ represents the genotypes at the variant sites and $R$ the reads.

As the ploidy (or the number of pooled samples) increases, distinguishing variants from sequencing errors becomes more complicated unlike in individual sequencing where evidence from multiple reads can be used to distinguish between rare alleles and sequencing errors (Futschik and Schlötterer, 2010; Wei et al., 2011). There are a number of variant calling tools suitable for pooled sequence data (Albers et al., 2011; Altmann et al., 2011; Bansal, 2010; Chen and Sun, 2013; Druley et al., 2009; Garrison and Marth, 2012; Koboldt et al., 2009; McKenna et al., 2010; Raineri et al., 2012; Vallania et al., 2010; Wei et al., 2011; Zhou, 2012) and among these, GATK’s Unified Genotyper (McKenna et al., 2010), SNVer (Wei et al., 2011) and FreeBayes (Garrison and Marth, 2012), were applied in this thesis work and below is a brief description of these tools. Besides these three Pool-seq suitable callers, SAMtools (Li et al., 2009), which is a popular variant calling software for diploid organisms, was used to illustrate the consequences of using polyploidy unaware tools on pooled data.

**GATK’s unified Genotyper**, which is currently retired and is replaced by HaplotypeCaller, uses a Bayesian model with a binomial likelihood to infer the posterior probability of all the possible genotypes. For reads traversing an INDEL which may as a result be misaligned, a local multiple sequence realignment is carried out to minimize false positive SNP calls. Additionally, base qualities are re-calibrated taking into account issues such as machine cycle and dinucleotide context which are known to result in inaccurate base qualities. There is also a post-processing step termed variant quality recalibration that uses information from high quality variants provided by users to assign a new quality score to the called variants. Multi-sample calling is also available that increases the sensitivity of the caller even when the sequencing depth is very low.

**FreeBayes**, is a Bayesian caller that infers the genotypes of small stretches of sequences (haplotypes) rather than single positions to avoid issues that arise during the alignment stage such as misalignment of INDEL spanning reads. As the length of the haplotype increases, the chance of the artifact occurring in repeated instances (and therefore taken as evidence of true variation) decreases. It first identifies candidate polymorphic regions and then performs a local de-novo assembly of reads from these candidate regions. It also incorporates estimates of other sources of errors such as strand, cycle, placement and allele bias in its Bayesian model (which other callers usually handle in a post-processing step such as GATK’s variant quality recalibration). It also uses evidence from multiple samples to make high-quality calls. Post variant calling filtering is recommended to remove low quality calls.
SNVer, compares the observed allele frequency against sequencing error using a binomial-binomial model. It computes an overall P-value for each candidate site which can be used in deciding the false positive rate threshold by the user. It accepts different thresholds for filtering the variants based on for instance allele imbalance and reference bias.

Various guidelines have been proposed for minimizing false positive variant calls in Pool-seq experiments. These include decisions during and after the alignment stage such as, using paired-end reads and trimming low quality bases towards the 3' end in Illumina reads, avoiding the use of heterologous reference genomes, filtering by a quality threshold deduced from a comparison of the quality distributions of known variant sets, using replicated pools and removing variants from hard to align regions and also those that display strand bias (Anand et al., 2016; Schlötterer et al., 2014).

1.5 MS-based bottom up proteomics workflow

The availability of databases that contain protein sequences (usually in silico translated from the genome or transcriptome sequences of organisms), has enabled large scale MS-based proteomics analyses of various organisms. To identify the proteins in samples, the experimental masses (from MS analysis) are often compared with that of the theoretical masses calculated from the sequences in the databases. Mass spectrometry (MS) is an analytical tool that can be used for identifying (and quantifying) the molecules that exist in a sample based on their mass. It can also be used to elucidate their structural and chemical properties.

In the most commonly used bottom up shotgun MS proteomics analysis paradigm (Figure 3), proteins (protein mixtures) are first broken down into peptides (which are chains of amino acids linked by what are known as peptide bonds) in a digestion protocol using proteases. This is in contrast to the top-down approach which involves the analysis of whole proteins directly without first digesting them into peptides (Chait, 2006; Demirev et al., 2005; McLafferty et al., 2007). While one MS analysis is enough to measure the mass and thereby identify a peptide in very simple mixtures, known as peptide mass fingerprinting (PMF) (Mann, Hojrup, and Roepstorff, 1993; Pappin, Hojrup, and Bleasby, 1993; Yates et al., 1993), for extracting the sequence of a peptide especially in complex mixtures (where the same mass can match different peptides even from different proteins), most often two/tandem MS analyses are carried out to increase the sensitivity of the identification.
Figure 3. The commonly used bottom-up shotgun proteomics analysis workflow. The proteins are first digested with a protease such as Trypsin and analyzed by LC-MS/MS. Peptide/protein identification is carried out using database searching by comparing the experimental spectra to the theoretical spectra produced based on the fragmentation technique employed after in silico digestion of the sequences in the protein database.

For complex proteins, different separation methods such as gel electrophoresis and liquid chromatography (at the protein or peptide level) can be used to reduce the sample complexity and increase the sample coverage. In the most commonly used high pressure liquid chromatography (HPLC) method in proteomics (Domon, 2006), peptide mixtures are dissolved in a liquid mobile phase and passed through a column stationary phase using high pressure pumps to be separated from the liquid mixture based on for instance hydrophobicity differences (known as reversed-phase chromatography). When these are coupled to a mass spectrometer online, components that are eluted from the column will first move through the ionization source where they will be ionized using soft ionization methods such as electrospray ionization (ESI) (Gaskell, 1997; Yamashita and Fenn, 1984). The first MS analysis is then used for determining the accurate mass of precursor ions and to select those that fall within a certain m/z ratio from the MS1 spectrum, say the n most intense peaks as in Data Dependent Accusation (DDA) or all the peaks as in Data Independent Accusation (DIA). These parent ions will then be fragmented by the second MS analysis usually via collision induced dissociation (CID) (Mitchell Wells and McLuckey, 2005) producing an MS2 spectrum and are finally detected by the detector component of the mass spectrometer which is connected to a computer that displays the spectrum (which is a graph of the ion intensity as a function of the m/z ratio).
Identification of peptides usually proceeds by supplying the experimental peak list to special software programs called search engines (Cox et al., 2011; Craig and Beavis, 2004; Diamant and Noble, 2011; Dorfer et al., 2014; Eng, Jahan, and Hoopmann, 2013; Geer et al., 2004; Kim and Pevzner, 2014; Kong et al., 2017; Tabb, Fernando, and Chambers, 2007) that compare them to the theoretical spectra produced from sequences found in the database. The sequences in the databases are first in silico digested based on the specificity rules of the experimental protease utilized followed by filtering based on the precursor mass of the peptide that resulted in the MS2 spectra (taking into account the precursor mass accuracy). The theoretical fragmentation spectra is then produced from these filtered peptides based on the fragmentation rules of the dissociation technique employed during the experiment (for instance a list of b and y-ions if CID was used). The candidate peptides are subsequently scored (Bafna and Edwards, 2001; Colinge et al., 2003; Elias et al., 2004; Eng, McCormack, and Yates, 1994; Havilio, Haddad, and Smilansky, 2003; Zhang, Aebersold, and Schwikowski, 2002) and ranked based on how similar the produced theoretical spectra is to the experimental spectra, and the top ranking sequence is taken as a match (known as peptide to spectrum match, PSM). The Andromeda score (Cox et al., 2011) employed in the MaxQuant software for instance is a probabilistic score that determines the probability of at least k matches out of n theoretical masses to the experimental masses by chance; it is the logarithm of this binomial distribution probability multiplied by -10 such that the higher the score the more confident the match is.

The scores calculated by the different search engines are not comparable in their initial form and need to be converted into valid statistical measures that assess the confidence of PSMs. Such statistical scores (reviewed in (Granholm and Käll, 2011; Nesvizhskii, 2010) include those that consider individual PSMs such as posterior error probabilities (PEP) and q-values as well as those that take in to account sets of PSMs (and therefore the multiple testing effect) (Bern and Kil, 2011; Cerqueira et al., 2010; Joo et al., 2010; Käll et al., 2008a, 2008b; Navarro and Vázquez, 2009) such as the most widely used False Discovery Rates (FDRs) based on target-decoy search setups (Elias and Gygi, 2007). In these setups, equal number of decoy sequences that imitate the amino acid composition and the peptide lengths of the target sequences are created usually by reversing the target peptide sequences and searched (either separately or appended to target sequences) to estimate the false positive matches. MaxQuant uses Bayesian statistics to calculate PEP of a peptide hit being a false hit given the score and the peptide length. There are also post-search validation tools such as PeptideProphet (Choi and Nesvizhskii, 2008; Ding, Choi,
and Nesvizhskii, 2008; Keller et al., 2002) and Percolator (Käll et al., 2008a; Spivak et al., 2009) that use statistical modeling and machine learning techniques to distinguish correct and incorrect identifications and compute probabilities such as q-values and PEPs based on initial scores/p-values and other additional features such as precursor mass accuracy and the delta mass. The utilization of these statistical measures instead of the original scores by the search engines allows for the merging of results from different search tools which could lead to improved identification and reliability (Higgs et al., 2007; Price et al., 2007; Resing et al., 2004; Yu et al., 2010). Once confident PSMs are obtained, the search engines map the peptides to the proteins they originated from using various methods that aim to address the problem of the same peptide mapping to multiple proteins (Huang et al., 2012; Noble and Serang, 2012).

Instead of protein sequences, the search databases can also contain spectra (Yates et al., 1998) (from a previous protein sequence database search) which result in a faster processing time. There are also methods that identify the peptide sequences de novo, merely from the spectra (Frank and Pevzner, 2005; Ma et al., 2003; Taylor and Johnson, 1997) which can be especially useful for organisms with no or incomplete protein sequence databases (Ma and Johnson, 2012; Menschaert et al., 2010) even though they are limited by the lack of the complete ladder of fragment ion peaks, which happens more often than not. It is also possible to combine these different methods for improved identification (Thomas and Shevchenko, 2008).

1.6 Protein databases

Protein database search tools are by far the most popular methods employed in MS-based proteomics studies owing to their simplicity and performance especially with today’s high throughput mass spectrometers. Since the introduction of the first database search program SEQUEST (Eng, McCormack, and Yates, 1994; Yates et al., 1995) in 1994/95, protein databases outplaced de novo methods that preceded them and have become the prominent approach for large scale proteomics analysis. The most commonly used public protein databases include Swissprot (Gasteiger, Jung, and Bairoch, 2001), UniProtKB (Apweiler, 2004), Ensembl (Cunningham et al., 2019), NCBI RefSeq (O’Leary et al., 2016), Entrez (Maglott, 2004), and UniRef (Suzek et al., 2015). Some of these databases, such as Swissprot, are highly curated and annotated and contain non-redundant entries while others are repository style databases that contain in silico translated sequences from reference
genomes/transcriptomes of organisms with redundant (e.g. Entrez) or non-redundant (e.g. RefSeq) sequence collections (Apweiler, Bairoch, and Wu, 2004).

The completeness and accuracy of protein sequence databases plays a decisive role in peptide/protein identifications since spectra matching will fail or produce false identifications if the correct peptide sequence is not available in the search database. For instance, peptide sequences that contain disease related proteomics data such as oncogenic mutations would not be identified from standard protein sequence databases that are derived from few representative samples. Moreover, unaccounted chemical/post-translational modifications also add to the ‘dark matter’ of bottom-up proteomics (Skinner and Kelleher, 2015). For this reason, various methods have been adopted to improve identifications of the unassignable spectra including, but not limited to

- Proteogenomic approaches that utilize data from genomic/transcriptomic sequence sources (Albertsen et al., 2013; Edwards, 2007; Evans et al., 2012; Hu et al., 2015; Krishna et al., 2015; Kumar et al., 2016; Li, Duncan, and Zhang, 2010; May et al., 2016; Tanner et al., 2007; Zhu et al., 2018; Zickmann and Renard, 2015). The extension/creation of protein databases using proteogenomics approaches can be based on de novo predicted genes/exon-graphs, 6 frame translation of raw (for prokaryotes whose genome contains mostly coding sequences) or assembled sequence reads, and known SNPs from public databases.
- The consideration of all possible linear combinations of amino acids during database searches (Yates et al., 1995).
- Using “error tolerant” methods that allow mismatches in spectra matching to account for modifications (Hughes, Ma, and Lajoie, 2010; Mann and Wilm, 1994; Tabb, Saraf, and Yates, 2003; Shilov et al., 2007; Creasy and Cottrell, 2002; Starkweather et al., 2007; DiMaggio, Jr. et al., 2008; Yonghua Han, Bin Ma, and Kaizhong Zhang, 2004; Searle et al., 2004; Wang et al., 2014; Renard et al., 2012).
- Blind searches (also referred to by the names unrestricted, open or mass tolerant searches) that allow for large mass tolerances to identify all possible post-translational and chemical modifications (Bittremieux et al., 2018; Chen et al., 2009; Chick et al., 2015; Devabhaktuni et al., 2019; Kong et al., 2017; Na, Bandeira, and Pack, 2012; Tanner, Pevzner, and Bafna, 2006).

Most of the above approaches extend the search spaces and as a result suffer from runtime and memory complexity and more importantly decreased identification sensitivity due to an increase of high scoring decoys which can
Literature Review

confound the true peptide targets. The increased redundancy can also affect protein inference. (Blakeley, Overton, and Hubbard, 2012; Muth et al., 2015; Nesvizhskii, 2014; Reiter et al., 2009; Renard et al., 2012; Schiebenhoefer et al., 2019; Shanmugam and Nesvizhskii, 2015) Therefore the best approach for striking a balance between the completeness and the size of protein databases has been a subject of much discussion (Noble, 2015; Noble and Keich, 2017; Sticker, Martens, and Clement, 2017). Moreover, Colaert et al. (Colaert et al., 2011) showed that in a target-decoy setup, when very similar isobaric (“targeted”) decoys were introduced to the database, for 95% of the target matches the decoy hits had equal or better score. As such, the appropriateness of target-decoy strategies and the associated statistical significance measures need to be carefully investigated when querying an extended search space. For the case of proteogenomic approaches, Nesvizhskii (Nesvizhskii, 2014) recommends the application of “class-specific” FDRs, that is, FDRs calculated separately for the novel (and ideally for all the different types of the novel peptides) and known peptides.

1.7 β-hemolytic streptococci

Bacteria are the most widely sequenced organisms and currently more than 150 thousand of bacterial strains have been whole genome sequenced, of which more than 12 thousand are streptococci (a spherically shaped gram-positive, facultative anaerobe bacteria) (Reddy et al., 2015). The β-hemolytic streptococci belong to the streptococcus genus and form a broad and entirely transparent zone around colonies when grown in blood agar due to the complete lysis of the red blood cells. They are also one of the most clinically relevant species in this genus besides Streptococcus pneumoniae. They are divided into 20 groups (A to H and K to V) based on the Lancefield classification (Lancefield, 1933) and among those, serogroups A, B, C and G (GAS, GBS, GCS, GGS) are known to cause variety of diseases in humans including mild ones such as pharyngitis and impetigo, and the more invasive ones for instance toxic shock syndrome, sepsis and necrotizing fasciitis, in addition to sequelae of infections such as acute rheumatic fever (ARF), post streptococcal glomerulonephritis, and post streptococcal reactive arthritis (PSRA).

The spectrum of diseases associated with these groups largely overlap (Haslam and St. Geme, 2018) especially with groups A, C and G, which are usually found in the upper respiratory tract. Group B mostly cause sepsis, meningitis and pneumonia in neonates and are often found in the vagina of women which could lead to transmission to infants through the amniotic fluid or during delivery (Heath and
Jardine, 2014). There is a high disease burden associated with these bacteria, for instance GAS is estimated to cause at least 517,000 deaths from invasive infections every year worldwide (Carapetis et al., 2005). Some serotypes of these bacteria are associated with certain disease types and severity of the diseases. For instance, in GAS, strains with serotypes M1, M3, M12 and M18 are usually associated with invasive infections; the M3 and M18 were prominently isolated during the rheumatic fever outbreak of the mid 1980s in the US (Johnson, Stevens, and Kaplan, 1992). In GBS, the serotypes Ia, II, III, and V are known to cause majority of the infections in the US and Europe, with serotypes Ia and III causing most of the neonatal infections (Hickman et al., 1999; Kieran et al., 1998). There is also geographical and temporal variations; certain dominant serotypes in a region or at a certain point in time could be rare in other regions or disappear gradually (Colman et al., 1993; Gaworzewska and Colman, 1988).

GAS and GBS are the best studied serogroups with a few hundred strains whole genome sequenced currently (170 group A, 101 group B) (Reddy et al., 2015) as well as having their proteomic profiles investigated in several studies using various techniques such as mass spectrometry and 2D-PAGE (Campeau et al., 2020; Hughes et al., 2002; Johri et al., 2007; Malmström et al., 2012; Nakamura et al., 2004; Nordenfelt et al., 2012; Papasergi et al., 2013; Wen et al., 2011; Wilk et al., 2018; Yang et al., 2010; Zhang et al., 2007).

1.8 Application of MS for bacterial characterization

Accurate identification is vital for characterizing and classifying microorganisms. In the case of bacterial pathogens, it enables improved diagnosis, treatment and tracing of infectious outbreaks. Traditionally, mainly immunochemical (e.g. ELISA) and phenotypic based methods (e.g. the API system) that require the isolation and culturing of microbes were employed to identify and characterize microbes. Although inexpensive, these methods have limited identification sensitivity and specificity, are laborious, and could be slow in identifying certain organisms (Franco-Duarte et al., 2019). With the advent of molecular methods such as PCR and DNA sequencing in the past few decades, nucleotide based approaches including the sequencing of the 16s rRNA and other conserved genes have been complementarily used with these traditional approaches. This facilitated the accurate identification and characterization of a wide variety of microbes including uncultivable bacteria (Prakash et al., 2007). Even though the molecular based methods have high
discriminatory power, they are not routinely employed in clinical microbiology since they are time consuming and expensive.

In recent years, MS based approaches for characterization of bacteria are growing in popularity owing to their simplicity, reliability, speed, and cost effectiveness (Bizzini and Greub, 2010; Croxatto, Prod’hom, and Greub, 2012; Seng et al., 2009, 2010). Most notably, the Matrix-assisted Laser Desorption Ionization technique coupled to the Time-of-flight analyzer (MALDI-TOF) MS-based method has been successfully applied for characterizing large numbers of gram positive (Barbuddhe et al., 2008; Boggs, Cazares, and Drake, 2012; Lartigue et al., 2011; Lasch et al., 2009; Moura et al., 2008; Reil et al., 2011; Williamson et al., 2008) as well as gram-negative bacteria (Berrazeg et al., 2013; Christner et al., 2014; Clark et al., 2013; Kuhns et al., 2012), at the species and sub species levels. The spectra produced by this method is compared to the available spectral library to identify the bacteria, and thus the accuracy of the method is affected by the completeness of the reference database.
2 STUDY AIMS

The aims of this thesis work are:

1. Assessing the accuracy of Pool-seq for identifying genomic variants and estimating allele frequencies in the three β-hemolytic bacteria utilized in our studies.
   ❖ The effect of sequencing coverage and the choice of variant calling tools and approaches on the accuracy of Pool-seq are investigated.
2. Developing a bioinformatics method that utilizes variants mined from the Pool-seq experiments to create proteogenomic databases.
3. Applying the bioinformatics method on Pool-seq data of the three β-hemolytic bacteria to evaluate the performance of the tool.
   ❖ The protein databases created by this method are compared to conventional as well as other proteogenomic databases.
3 MATERIALS AND METHODS

This study used 100 GAS, 137 GGS and 80 GBS bacterial strains that were mostly selected from the bacterial culture collections of the National Institute of Health and Welfare and United Medix Laboratories Ltd, Finland. Pool-seq was used to sequence these strains and variant protein databases based on the variants identified from the Pool-seq experiments were created for mass spectrometry analysis of a small number of strains randomly selected from the pools.

The GAS study establishing the efficiency of Pool-seq was used in publication I, while the Pool-seq driven variants from GGS were used in publication III. For GBS, we individually whole genome sequenced the 40 strains that were in one of the pools (unlike in GAS with only 6 strains individually sequenced) and the pool was sequenced at lower sequencing depth than GAS and GGS. Thus, I have included results from the GBS Pool-seq experiment in this thesis to further extrapolate on our findings from the two published articles. In addition, for GAS we had previously analyzed with MS 10 other strains that were not in the pools and we reanalyzed those strains using our Pool-seq driven database and the result from the re-analysis is included in this thesis. The materials and methods employed in our studies is summarized below.

3.1 DNA isolation and pooling

DNA was isolated from the bacteria that were cultured on blood agar plates in 5% CO2 over one or two nights at +37 ° by using the UltraClean Microbial DNA Isolation Kit (QIAGEN) according to the instructions by the manufacturer except at the start, 6 μl mutanolysin (1 mg/ml) was mixed with the MicroBead solution provided and 10ul of loopful bacteria scraped from the culture plate and incubated for 60 min at +37 °C. The mixture was then transferred to another MicroBead tube, and 2 μl of RNAse A (1 mg/ml) was added. Also at step 18, 35 μl of solution MD5 (instead of 50 μl) was added and incubated for 2 min. The Nanodrop equipment (ThermoFisher) and agarose gel electrophoresis were used to check the quality and integrity of the isolated DNA. The final per sample DNA concentrations of the pooled GAS, GGS, and GBS were respectively 400ng, 200ng and 500ng (as measured by Qubit 2.0 Fluorometer). Two pools each containing 50 and 40 strains for GAS and GBS respectively, while for GGS 3 pools, two of them containing 47 strains and the third one containing 43 strains were prepared.
3.2 DNA sequencing and analysis

The pooled GAS and GGS samples were sequenced using Illumina Hiseq2500, 2 × 125 bp in 2 and 5 lanes respectively at the National Genomics Infrastructure hosted by SciLife Lab, Stockholm. Illumina MiSeq 2×150 bp was used for sequencing the two pools of GBS, the 40 individual strains from one of the GBS pools (at GATC Biotech AB, Sweden) as well as the 6 individual strains from GAS (at Turku Centre of Biotechnology, Finland).

For pre-processing of the raw reads FastQC (v0.11.2) and Trimmomatic (v0.33) were used for quality checking and adapter trimming before aligning them with Bwa mem (version0.7.10) using default parameters to their respective reference genomes (AE004092, NEM316_III and ATCC12394 for GAS, GBS and GGS respectively). Variant calling of the GAS and GGS pools was then performed using 4 different tools with default parameters except those listed in parenthesis that were used to make results comparable, namely SAMtools v1.1 (–q 20 -d 10000), UnifiedGenotyper v3.2–2 (stand_call_conf 20 -stand_emit_conf 20), FreeBayes v0.9.18-1 (m 20 -q 13 -F 0.02) and SNVer v0.5.3 (bq 13 -t 0.02). SAMtools (v1.1) was also used to call variants from the 6 individual GAS strains and publicly available sequencing runs from ENA while only FreeBayes (v0.9.18-1) was used for calling variants from one of the GBS pools and also the 40 individual samples in that pool.

Other tools that were utilized for variant processing include, C-Sibelia (v3.0.5) for variant identification of the 45 complete GAS genomes, Bedtools (v2.17.0) and SAMtools (v1.1) for sequencing coverage calculation, Picard (v1.122) for duplicate marking, Vt (v0.57721) for variant normalization, BCFtools (v 1.10) for variant concatenation and SnpEff and SnpSift (v4.0e) for variant annotation and filtering.

3.3 Protein extraction and digestion

From the strains that were pooled for sequencing, for GAS and GBS 7, and for GGS 8 strains were randomly selected for MS-based proteomics analysis. The strains were extracted from −70 °C milk/glycerol suspensions and grown on blood agar plates at 37 °C under an atmosphere of 5% CO2. Individual colonies were taken for growth in 10 ml Todd–Hewitt broth at 37 °C under 5% CO2 atmosphere. The over-night cultures were diluted 1:20 in 10 ml of prewarmed Todd–Hewitt broth the next day and grew under the same conditions as before until they reached the exponential phase (OD 0.4–0.5). Three types of protein fractions i.e., exoproteome, shaved surface proteome and proteins extracted from the entire cellular proteome were
prepared from the mid-exponential cultures by splitting into twice 5 ml and centrifuging at 2500×g for 10 min. From the combined growth supernatants, the exoproteome proteins were extracted by 30 kDa filtration (Amicon Ultra-15 Centrifugal Filter Units, 30 MWCO, Millipore). While one of the pellets was treated with 5 μg of trypsin for 15 min at 37 °C to produce the shaved surface proteome fraction, the other pellet was used for isolating the proteins from the entire cellular proteome by bead beating with 200 mg mm Zirconia/Silica Bead, Biospec without Triton. The fractions were analyzed together for the purposes of this thesis work.

The extracted proteins were reduced and alkylated with 25 mM dithiothreitol (45 min at 37°C) and 50 mM iodoacetamide (45 min at room temperature) respectively before tryptic digestion (at 37 °C). The peptides were then cleaned using C18 spin column purification (The Nest Group, Southborough, MA, USA).

3.4 LC-MS/MS

An EASY-nLC 1000 (Thermo Fisher Scientific) LC coupled online to an Orbitrap Elite (Thermo Fisher Scientific) mass spectrometer with a nano-electrospray ionsource was used. The RPLC started with a 5-min gradient at 5% of buffer B (98% acetonitrile, 0.1% formic acid and 0.01% trifluoroacetic acid) followed by a 60-min linear gradient (300 nl/min) at 5 to 35% of buffer B, a 5-min gradient at 35 to 80% of buffer B, a 1-min gradient from 80 to 100% of buffer B and finally a 9-min column wash with 100% of buffer B.

MS was acquired in positive ion profile mode with a resolution of 60,000 at normal mass range (m/z 50–2000) and for MS/MS the 20 most intense peaks were chosen to be fragmented by CID.

3.5 MS/MS analysis

The Andromeda search engine in MaxQuant (v1.6.0.16) was used for peptide/protein identification of the MS/MS spectra. The parameters used for the searches include reversed decoy, specific Trypsin/P digestion, Carbamidomethyl (C) as a fixed modification, and Oxidation (M) and Acetyl (N-term) as variable modifications, and FDR (Peptide, Protein, Site) set at 0.01.

The databases used for the searches (and their unique tryptic peptide sizes) are listed in the table below.
The different types of databases employed in this study together with their tryptic peptide size. The Pool-seq driven in-house databases refer to the proteogenomic protein databases that were created using the Python package ‘PoolSeqProGen’ developed in this study which is discussed in the results section.

<table>
<thead>
<tr>
<th>Databases</th>
<th>Tryptic peptide size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single genome</strong></td>
<td></td>
</tr>
<tr>
<td>GAS (AE004092)</td>
<td>72272</td>
</tr>
<tr>
<td>GBS (NEM316_III)</td>
<td>43047</td>
</tr>
<tr>
<td>GGS (ATCC12394)</td>
<td>38173</td>
</tr>
<tr>
<td><strong>Multi-genome</strong></td>
<td></td>
</tr>
<tr>
<td>GAS (49complete/294assembly)</td>
<td>89705/20469</td>
</tr>
<tr>
<td>GGS (4complete/26assembly)</td>
<td>57898/113208</td>
</tr>
<tr>
<td><strong>Pool-seq driven</strong></td>
<td></td>
</tr>
<tr>
<td>GAS</td>
<td>91650</td>
</tr>
<tr>
<td>GBS</td>
<td>62256</td>
</tr>
<tr>
<td>GGS</td>
<td>72050</td>
</tr>
<tr>
<td><strong>Megahit + Prodigal ab initio</strong></td>
<td>153248</td>
</tr>
<tr>
<td><strong>Sixgill Metapeptides</strong></td>
<td>251946</td>
</tr>
</tbody>
</table>
4 RESULTS AND DISCUSSION

In this chapter, the findings of the three publications used in the current thesis work will be summarized and discussed. Based on our Pool-seq studies of the three β-hemolytic bacteria, the first three results offer guidelines regarding the sequencing coverage, choice of variant calling tools and methods for Pool-seq experiments that involve highly diverse bacterial species and establish the robustness of the Pool-seq approach for SNP discovery and allele frequency estimation. The next result then discusses the PoolSeqProGen software that utilizes the variants mined from Pool-seq experiments to create proteogenomic protein databases that are used during MS searches. The last two results compare the performance of the variant protein databases created using PoolSeqProGen with conventional databases as well proteogenomic databases created by other tools.

4.1 A sequencing coverage of ~200-250X is optimal for accurately identifying variants from bacterial Pool-seq data

Sequencing coverage has a direct effect on the accuracy of SNP detection both in individual and Pool-seq experiments, high coverage resulting in highly confident base calls. The average sequencing coverage for our Pool-seq experiments were very high (Figure 4A), which allowed us to set a 2% threshold for reads supporting variant alleles, compared to the error rate of Illumina sequence reads which is at most 1% (Glenn, 2011), to minimize false positive calls. We individually sequenced all 40 of the samples in one of the GBS pools and we downsampled reads from this pool to determine the optimal sequencing coverage for our experiments. For most of the 40 samples, the sensitivity of the Pool-seq approach stabilizes at around 30% of the downsampled reads (~20X), although few samples required higher coverage to achieve increased sensitivity (Figure 4B). Similarly, in the GAS Pool-seq experiment, a downsampling to only 1% (~230X) resulted in maximum 1% loss in sensitivity for the 6 samples for which we had individual sequencing information. This suggests that we could have adopted lesser sequencing depth for GAS and GGS to sequence higher number of samples without substantially increasing our initial budget and compromising the accuracy of the variant detection using Pool-seq.

Holt et al. (2009) showed that at a coverage of 40X they could obtain an 83% SNP detection sensitivity when their pool containing 6 samples of the pathogen *Salmonella paratyphi* A was compared to the individual sequence data. However, the SNPs they considered were those with high frequency (found in at least 2 strains)
and the sensitivity declined to 37% for SNPs from only a single strain indicating that identifying low frequency alleles would require a much higher coverage. Moreover, the Salmonella Paratyphi A bacteria is highly monomorphic and as a result such a low coverage could result in inaccurate SNP discovery for more polymorphic bacteria. In our GBS experiment for instance, at ~60X (10% downsampling), the sensitivity decreased by 8%.

Even though in Pool-seq a coverage of 50X is deemed sufficient for SNP calling (Schlötterer et al., 2014), it has been demonstrated that higher coverages might be required for certain study types. For instance, Kofler et al. (2011) noted that even a coverage of 90X was small for individual SNP based population estimators and analysis based on larger window size was recommended for studies with such low coverages (e.g. 40X with a window size of 1 kb). Likewise, Kofler & Schlötterer (2014) showed in a simulation analysis that a coverage of at least 200X was needed to detect weakly selected loci in E&R studies. We also found that at a coverage of ~200-250X, the Pool-seq approach results in accurate variant identification and the results from our studies can be taken as a guideline for similar studies that aim to identify variants from other genetically polymorphic bacteria.
Results and discussion

A)

![Graph showing read depth per base for GGS, GAS, and GBS](image)

- **GGS**: Constant read depth across the range.
- **GAS**: Constant read depth across the range.
- **GBS**: Peaks at specific depths, indicating a higher variability or specific types of reads.

**Read depth per base**

0 2000 4000 6000 8000
Figure 4. Sequencing coverage and its effect on accuracy of Pool-seq. A) The aligned read depth per base of the three Pool-seq experiments included in this thesis. While most of the bases in the GAS and GGS Pool-seq studies were covered by ~8000 reads, the GBS experiment had an average of ~600 reads covering most of the bases. B) Downsampling of aligned reads from the GBS Pool-seq experiment to analyze the fraction of variants that
could be identified from the 40 individual samples that were pooled. For most of the samples the increase in sensitivity was not that high above 200X (30% of the reads).

### 4.2 Pool-seq incompatible variant calling results in substantial reduction of sensitivity

The efficiency of 4 variant calling tools for Pool-seq data was assessed by the overlap of variants identified by 2 or more of the tools as well as by the number of variants they could identify from the individually sequenced strains. While FreeBayes, UG and SNVer had comparable recall and precision, SAMtools had the least sensitivity when applied to Pool-seq data, identifying ~25% less variants from the 6 individually sequenced GAS strains (Table 2 and Figure 5). Furthermore, when SAMtools was used on the de-duplicated Pool-seq GAS data its recall decreased by ~20% more (see Table 1 of Publication I).

Mullen et al. (2012), found that a large number of the dbSNP variants could not be discovered from their Pool-seq data and based on our studies we anticipate that might partly be because they used SAMtools on de-duplicated data even though they attributed the false negative results to factors such as low coverage and erroneous dbSNP data. On the other hand, de-duplication might be necessary for certain study types such as those involving RADseq; for instance Gautier et al. (Gautier et al., 2013) showed that the overall experimental error increases when duplicates are not removed. Our results suggest that utilizing polyploidy aware variant calling tools and avoiding de-duping (for non-RADseq data) in Pool-seq experiments can improve the sensitivity of the approach.

<table>
<thead>
<tr>
<th></th>
<th>GAS pool 1</th>
<th></th>
<th>GAS pool 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>FreeBayes</td>
<td>0.98</td>
<td>0.86</td>
<td>0.99</td>
<td>0.85</td>
</tr>
<tr>
<td>GATK UG</td>
<td>1.00</td>
<td>0.94</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>SAMtools</td>
<td>0.38</td>
<td>1.00</td>
<td>0.38</td>
<td>0.99</td>
</tr>
<tr>
<td>SNVer</td>
<td>0.98</td>
<td>0.90</td>
<td>0.98</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Results and discussion

Figure 5. The sensitivity of the 4 variant calling tools when used with the GAS Pool-seq samples.

4.3 Pool-seq is robust for accurate SNP detection and allele frequency estimation

Except in few of the GBS samples, Pool-seq was capable of calling more than 90% of the SNPs from the individually sequenced samples of GBS and GAS (Figures 6A and 6B). Publicly available sequence data analysis also showed that most of the variants mined from the GAS pools could also be identified from the 44 GAS
complete genomes (>70%) and 3407 GAS sequences from ENA (>90%), which were analyzed against the same reference genome (Figure 6C).
Results and discussion

Figure 6. Pool-seq’s variant identification sensitivity. The sensitivity of Pool-seq to identify variants from A) 6 GAS and B) 40 GBS individually sequenced samples. C) The overlap of variants mined from the GAS Pool-seq and publicly available sequence data. In A, FreeBayes, SNVer and GATK’s UG were used and the variants called by 2 or more tools were used, while in B, only FreeBayes was used for calling variants.

Moreover, the correlation of the allele frequency estimates between the individual and pooled sequencing approaches for the GBS experiment was high (R=0.96, p<2-2e-16) (Figures 7A and 7B). This is in accordance with other studies that showed high correlation in the frequency estimates (R values >0.9) between individually sequenced/genotyped and pooled samples (Bansal et al., 2010; Druley et al., 2009; Gautier et al., 2013; Holt et al., 2009; Van Tassell et al., 2008; Zhu et al., 2012). The accuracy of allele frequency estimates from Pool-seq experiments increases with the increase in sequencing coverage and pool size (Day-Williams et al., 2011; Gautier et al., 2013; Rellstab et al., 2013). In general, from the analysis of publicly available data and individual sequencing, we have demonstrated that Pool-seq is a robust method for variant identification and allele frequency estimation of large bacterial samples.
Results and discussion

A)

$R = 0.96, p < 2.2e-16$
Figure 7. Allele frequency estimates from GBS pools and individually sequenced samples of the A) reference (major) allele and B) the variant (minor) alleles. From the 33716 total number of variants, 1000 were randomly chosen for easier display. The frequencies were estimated by using SAMtools mpileup (which shows the pileup of reads at a position).

4.4 Pool-seq driven proteogenomics database generation

The general workflow for variant discovery and annotation including, quality checking/filtering and mapping to the reference genome, followed by variant identification and annotation was applied to generate the variants that are incorporated to the protein databases. The Python package ‘PoolSeqProGen’ that implements the variant protein database creation from Pool-seq driven variants was
then developed (publication II) and applied to GAS, GBS, and GGS (publication III). The tool performs the following steps to create the variant protein databases (Figure 8): 1) Choosing the coding variants with non-synonyms effects (both SNPs and INDELs) from the variant annotations file, 2) Extracting reads that span those variant positions to identify unique combinations (assortments) of the variants, 3) Inserting these unique set of variants to the nucleotide sequence of the proteins, 4) Translating and in silico digesting the sequences, and 5) Writing to the fasta databases the variant peptide sequences (which are unique and >4 amino acids long) together with flanking sequences and also the wild type protein sequences. The fasta header holds information of the particular variants that resulted in the variant peptide in the form of bitwise flags that have been converted to decimal numbers to minimize the space required. For instance if 3 non-synonyms SNPs were identified in a protein and some reads contained 2 of the first SNPs and others contained all 3, then the bitwise flags will respectively be 110 and 111 which are then converted to 6 and 7 and are appended to the fasta header of the respective variant peptides.

Variant protein databases from known SNPs found in databases such as dbSNP and TCGA have been created (Ahn et al., 2014; Cao et al., 2017; Li, Duncan, and Zhang, 2010). There are also tools and pipelines for the generation and visualization of proteogenomic data (Ahn et al., 2014; Krasnov et al., 2015a, 2015b; Nagaraj et al., 2015; Peterson et al., 2012; Sheynkman et al., 2014; Wang and Zhang, 2013; Wingo et al., 2017; Zickmann and Renard, 2015). Our tool is different as it is designed to take in to account the pooled nature of the samples. By interrogating the aligned reads for unique patterns of variant combinations, we include these ‘seen’ combinations of variants in the variant databases to capture the sample wise variations that exist in our Pool-seq data (Figure 9). This way, the database’s completeness is improved while avoiding unnecessarily inflating the size, unlike if for instance every combination of the variants were considered.
**Figure 8.** Workflow of the Pool-seq driven proteogenomic database creation. After variants are identified and annotated, those with non-synonymous effects are inserted to the nucleotide sequence of the proteins they span and the sequence is then in silico translated and digested. The variant peptides and the flanking peptides are written to the database together with the wild type reference protein sequences.
Results and discussion

Figure 9. How the different combinations of Non synonymous variants that are ‘seen’ in the reads spanning the variants are incorporated to the variant peptides.

4.5 The Pool-seq driven proteogenomic databases allow identification of identical tryptic peptides with different variant profiles

The main distinguishing feature of our tool is the incorporation of combinations of observed variants so that the individual strains in the pools can be represented to a certain level even though it is not possible to differentiate the samples based on the sequence reads. This approach has allowed us to identify homologous tryptic peptides that contain different variants from 10 GGS proteins which are shown in Table 3. These are interesting findings as they confirm our tool’s ability to capture the strain wise differences to a certain extent. For instance the two variant peptides (IPVIAYGVCPECQAK, IPVIAYGVCPECQVK') that were identified from the transcriptional repressor protein ‘WP_014612608.1’, were identified uniquely from transcriptional repressor proteins of 13 and 4 of the assembly genomes respectively. This implies that these variant peptides do possibly exist in the different strains that were pooled and we have been able to identify them as a result of the read interrogation strategy we adopted in our tool and any tool that considers only a single version of these tryptic peptide would be unable to identify all of them.
### Table 3.

The 10 GGS proteins that had same tryptic peptides but with different variant profiles identified from (these tryptic peptides and the AA changes that produced them are shown in red). Other variant peptides identified from these proteins are also listed. The total number of peptides identified from the proteins including ‘wild type’ and those with missed cleavages are shown in the last column.

<table>
<thead>
<tr>
<th>Id</th>
<th>Length</th>
<th>Name</th>
<th>AA</th>
<th>Variant peptides</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP_003056081.1</td>
<td>304</td>
<td>Protein jag</td>
<td>S55N&amp;DK65G</td>
<td>KPAVQVDIEGINK</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>D56G</td>
<td>KPAVQVDIEGISGK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>I82V</td>
<td>QNAPVVPADVELEEMK</td>
<td></td>
</tr>
<tr>
<td>WP_014612608.1</td>
<td>155</td>
<td>Transcriptional repressor</td>
<td>I143V</td>
<td>IPVIAYGVCPECQAK</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>I143V&amp;A149</td>
<td>IPVIAYGVCPECQVK*</td>
<td></td>
</tr>
<tr>
<td>WP_014611973.1</td>
<td>511</td>
<td>M protein</td>
<td>E89D&amp;L94F</td>
<td>EVADYNLFDK</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A113T&amp;E115K</td>
<td>VVNDGLQTTK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E105K&amp;A115T&amp;E115K</td>
<td>MKVNDLQTTK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>D141N&amp;Y1455&amp;T149A</td>
<td>NKEFSLGALR</td>
<td></td>
</tr>
<tr>
<td>WP_003059396.1</td>
<td>145</td>
<td>YtxH domain-containing protein</td>
<td>A40S</td>
<td>AYQSYKENGPDYHYQLAK</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N44S&amp;L51F</td>
<td>AYQAKEYSPDDYHQFAK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V104A</td>
<td>TEKTELAEVEAK</td>
<td></td>
</tr>
<tr>
<td>WP_014612218.1</td>
<td>699</td>
<td>ATP-dependent Clp protease ATP-binding subunit</td>
<td>T335A</td>
<td>SLEAMATQK</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E415G</td>
<td>GHVGQDGADAVEVAR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E415G&amp;V427A</td>
<td>GHVGQDGAVEAAAR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T582A</td>
<td>NTVIATSNAQFGHGEDENTDQAIMDR</td>
<td></td>
</tr>
<tr>
<td>WP_014611976.1</td>
<td>440</td>
<td>Kinase</td>
<td>L88P</td>
<td>SKPfatDSGAMPHKLEK</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>I155T</td>
<td>VYFADKDSVTLPQVQEFLLK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>I155T&amp;K164S</td>
<td>VYFADKDSVTLPQVQEFLLSFGHR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R202K&amp;G204V</td>
<td>SVDVEYTVQFPLNPDDDFKPVLDT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G204V</td>
<td>SVDVEYTVQFPLNPDDFRPVLKDT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>I220V</td>
<td>TLAIGLTVTSQELLAQASILNK</td>
<td></td>
</tr>
</tbody>
</table>
Results and discussion

<table>
<thead>
<tr>
<th>Protein Accession</th>
<th>Position</th>
<th>Description</th>
<th>Variant 1</th>
<th>Variant 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP_014612152.1</td>
<td>269</td>
<td>Cof-type HAD-IIB family hydrolase</td>
<td>I29V&amp;D31E</td>
<td>ITDDVFQAVQEAK</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>D31E</td>
<td>ITDDVFQAIQEAK</td>
</tr>
<tr>
<td>WP_014611901.1</td>
<td>371</td>
<td>Redox-regulated ATPase YchF</td>
<td>D38N&amp;I41V</td>
<td>AGAEAANYPFATIPNVGM VEVPDER</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>I41V</td>
<td>AGAEAANYPFATIDPNVGM VEVPDER</td>
</tr>
<tr>
<td>WP_014612591.1</td>
<td>873</td>
<td>SEC10/PgrA surface exclusion domain-containing protein</td>
<td>A63T</td>
<td>ASNTSEESLPKTETCEETK</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A129G&amp;K133E</td>
<td>ALTSQEIYTNTLASSEETLL GOGAEYQR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>K133E</td>
<td>ALTSQEIYTNTLASSEETLL AOQAEGYQR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T207A</td>
<td>AAQTANDNTKALSELEK</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V774I</td>
<td>IDTTLPGQEMIK</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P752L</td>
<td>HLDEDIATVPDLQVARLLTG VKPLSYSK</td>
</tr>
<tr>
<td>WP_003058857.1</td>
<td>350</td>
<td>BMP family ABC transporter substrate-binding protein</td>
<td>A332D</td>
<td>EALKDIEEAK</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A332D&amp;S338A</td>
<td>EALKDIEEAKAK</td>
</tr>
</tbody>
</table>

4.6 The Pool-seq driven proteogenomic databases result in more peptide identification compared to single genome databases

The Pool-seq driven proteogenomic databases identified ~600 (Figure 10A) and ~200 more peptides than the single reference genome based databases of GAS and GGS respectively. The improvement in identification is also apparent when the samples analyzed were not part of the pools (Figure 10B). This indicates that in situations where multiple individual whole genome sequences are not available (which is the situation for most non model organisms), augmenting databases with variant sequences identified from pooled sequencing could be a better approach in terms of performance as well as cost. On the other hand, multi-genome databases (that contain 49 GAS and 4 GGS complete genomes, and 294 GAS and 26 GGS assemblies) identified ~1000 (Figure 10A) and ~500 more peptides compared to the Pool-seq derived proteogenomic databases, emphasizing the importance of sequencing and annotating several representative strains especially for organisms with high intra species diversity. Our results indicate that augmenting databases with
variants identified from Pool-seq data improves identification in MS based proteomic analyses.

**Figure 10.** The number of peptides identified when using the Pool-seq driven variant databases (Variant_Pools) compared to conventional single genome and multi-genome (from 49
results and discussion

complete genomes and 294 assemblies of GAS) A) 7 GAS samples that were in the pools and B) 10 GAS samples that were not in the pools were analyzed by MS.

Other options for constructing protein databases from genome sequence information include ab initio gene prediction after assembly and predicting genes directly from the sequence reads for prokaryotes without assembly since most part of their genomes is protein coding. Both methods do not require a reference genome and therefore could also be applied in Pool-seq experiments lacking good quality reference genomes. We used the GBS Pool-seq experiment to compare the peptide identifications from such methods to our Pool-seq driven variant database approach. For the assembly we used the metagenome assembler MEGAHIT (Li et al., 2015), followed by Prodigal (a gene prediction tool for prokaryotic genomes) for predicting genes from the contigs. For the assembly free approach we used the Sixgill (May et al., 2016) tool that produces ‘Metapeptides’, which are tryptic peptide sequences derived from open reading frames without stop codons either considering all 6-frames or after gene prediction by MetaGeneAnnotator (Noguchi, Taniguchi, and Itoh, 2008) and are expected to be identified in a MS analysis.

Compared to the ab initio gene prediction based and the ‘Metapeptides’ databases, the GBS Pool-seq driven databases identified ~200 and ~1000 more peptides respectively (Figure 11A). May et al. noted that peptide identifications from assembly based approaches may be low owing to the challenges of reliably assembling real genes in to contigs from metagenomics data. But contrary to this observation, the assembly based methods identified more peptides than the ‘Metapeptides’ when applied to our GBS Pool-seq data (Figure 11A). One reason for this could be because our Pool-seq data does not contain metagenomes but rather different strains of the same species and therefore the sequences could be more alike leading to better assembly yield. And so, if the tools were used on real metagenomic data, the assembly free method may perform better. The other and more likely reason has to do with the size and content of the databases. Large, redundant databases that especially contain “wrong” sequences from other organisms or 6-frame translations result in lower number of identifications as a result of the expanded search space (Blakeley, Overton, and Hubbard, 2012; Muth et al., 2015; Nesvizhskii, 2014). Due to the 6-frame translation, the ‘Metapeptides’ database size is inflated (by ~64% compared to the assembly based database) and contains ‘wrong’ sequences leading to the low sensitivity as a result of high scoring decoys as can be seen in Figure 11C. The target peptide hits were also marginally of higher scores in the ‘Metapeptides’ database even though in general their score and PEP distributions were similar to the other two databases (Figure 11D).
Open search tools are other alternative approaches that are mainly applied for the identification of unaccounted for chemical and post translational modifications and could help fill the gap of identifying unassigned spectra. In proteogenomics studies it has been recommended to rule out that novel peptides are not as a result of such modifications (Nesvizhskii, 2014) and therefore these tools can be applied towards that purpose. In addition, great strides have been achieved in alleviating the computational bottlenecks associated with such tools with the development of fast tools such as MSFagger. However, these approaches could lead to loss of sensitivity as a result of the expanded search space. This was evident in our GBS Pool-seq data where an open search using the MSFagger tool resulted in the identification of the least number of peptides (~2000 peptides less than the ‘Metapeptides’ approach) (Figure 11A) and also had higher numbers of exclusive peptides (Figure 11B) albeit the authors demonstrating that open searches do not lead to a huge reduction in the number of identified peptides (see Table 1 in Kong et al. (Kong et al., 2017)). It is therefore important to consider various options to maximize identification while avoiding pitfalls such as inflated FDRs and search spaces that could result in increased false positives and false negatives. Our studies demonstrate that one way to achieving such a goal is to utilize the data obtained from cost effective Pool-seq experiments to expand existing protein databases.
Results and discussion

A)
Results and discussion

B)
Results and discussion

C)

Metapeptides

PoolSeqProGen

Assembly-based

Andromeda score of decoy hits
Figure 11. Comparison of the Pool-seq driven approach with other proteogenomics methods i.e., assembly based ab initio prediction and the non-assembly ‘Metapeptides’ approaches as well as an open search method MSFragger. A) The number of unique peptides identified from these approaches B) The overlap of the unique peptides among these approaches C) The Andromeda score distribution of the decoy sequences from the Pool-seq, Metapeptides and Assembly driven databases and D) The distribution of the Andromeda Score and PEP of the target hits from the Pool-seq, Metapeptides and Assembly driven databases. Since MSFragger uses different scoring and statistical measures than MaxQuant which was used for searching the Pool-seq, Metapeptides and Assembly driven databases, only the latter three databases are compared in C&D.
5 CONCLUDING REMARKS AND FUTURE PERSPECTIVES

Protein databases have been playing a major role in bottom up shotgun mass spectrometry analyses as they are the source of the theoretical spectra that is compared with the experimental spectra to identify the proteins that exist in a sample. The databases are usually constructed from translation of genomic sequences of organisms and certain databases such as Swissprot contain curated protein sequences and are usually used in such analyses if they exist for the organism of interest. Two major issues with protein database based identifications are: 1) If the sequences are not available in the databases, the proteins will not be identified and 2) Large databases create problems for search engines in scoring candidate hits and hence result in suboptimal identifications due to the extended search space.

Therefore there is a need to consider the tradeoff between creating a complete database that contains all possible sequences and keeping the database size small enough. One way to increase the completeness of databases is to expand existing protein databases with evidence from genomic/transcriptomic sources in what are known as proteogenomics approaches. However such approaches will increase the database size and do not allow the searching of unknown variants. Instead, error and mass tolerant searches could be adopted to identify novel peptides. But these methods will also expand the search space resulting in decreased sensitivity on top of requiring longer run times. The extended search space also warrants reexamination of the suitability of routinely used statistical measures of confidence in these scenarios.

The ideal solution would have been to de novo sequence the peptides but current de novo methods are not capable of sequencing the entire peptide accurately mostly due to the incomplete fragment peak ladder. With the continuous improvements in the quality of data produced by mass spectrometers and also the software tools, de novo methods are increasingly being used in MS-based proteomics analyses. However, currently no single tool/approach is able to identify all the spectra that is available in shotgun proteomics with high confidence. Therefore, integrating different approaches appears to be the best option to characterize much of the unidentified spectra and getting us closer to uncovering the ‘dark’ matter of shot gun mass spectrometry.

When large number of samples are sequenced comparatively inexpensively using the Pool-seq approach, our method can be utilized to incorporate variants from the samples in a way that reflects their composition in the pooled samples improving the
completeness of protein databases without greatly inflating the search space. By utilizing this method novel peptides that have important implications for instance in the virulence of certain bacterial strains could be investigated. In addition, the databases produced by the method can be utilized for novel diagnostic marker screening which are underway at the moment for the β-hemolytic bacteria. Additionally, since in certain bacteria the accessory genome could also be a major source of variation, alignment free methods, such as the ab initio prediction of genes from reads that do not align to reference genomes, could be integrated with our method to further enhance the comprehensiveness of the protein databases. Utilizing MS based approaches for characterizing bacteria will continue to grow in the years to come enabling a plethora of applications that are aimed towards typing, novel virulence/antibiotic resistance antigen identification, disease biomarker discovery, and vaccine development, to mention but a few, and the role of comprehensive protein databases in these endeavors is indispensable.
ACKNOWLEDGEMENTS

This thesis work was carried out in the Bacteriology and Immunology department at the Haartman Institute, as well as the Molecular Systems Biology research group and Proteomics Unit at the institute of Biotechnology, University of Helsinki. In addition, the data storage, and high performance and cloud computing services vastly utilized in this research work were provided by CSC (Center for Scientific Computing), Finland.

Many people were instrumental for the completion of my PhD thesis and I am indebted to all. First and foremost, I would like to express my deepest gratitude to my supervisors Dr. Sakari Jokiranta and Dr. Markku Varjosalo for giving me the opportunity to work on this thesis project and for their valuable guidance and support throughout the years. I consider myself fortunate to have had such great supervisors who are very knowledgeable, goal oriented, flexible and most of all, always looking out for their student.

I would also like to extend my sincere thanks to my thesis committee members, Professors Jaana Vuopio and Sampsa Hautaniemi for their constructive advices, and to the pre-examiners, Doctors Maija Vihinen-Ranta and Ulrich Bergmann for thoroughly reviewing the thesis and for their valued feedbacks. Special thanks goes to Professor Laura Elo for agreeing to serve as an opponent in my thesis defence. My sincere gratitude to Professor Kari Keinänen for all his help in the course of the thesis review and examination and for serving as Custos in the defence. I also gratefully acknowledge the assistance of Susanna Puusniekka from Viikki PhD Study Services.

I also wish to thank all my co-authors especially Professors Jaana Vuopio and Juha Kere, Doctors Neeta Datta, Carolin Colmeder, Karita Hapasalo, Salla Keskitalo and Xiaonan Liu, Kari Salokas, Kai Puhakainen, and all members of the SalWe GetItDone consortium for the successful collaborations and research outputs.

I had great pleasure working with all the previous and current members of the Jokiranta and Varjosalo groups, including Aino Koskinen, Antti Nykänen, Arnab Bhattacharjee, Awel Eshetu, Carolin Colmeder, Derek Ho, Eija Nissilä, Fitsum Tamene, Hanne Amdahl, Helka Göös, Iftekhar Chowdhury, Jaakko Teppo, Kari Salokas, Karita Haapasalo, Leena Yadav, Lisa Gawriyski, Matias Kinnunen, Neeta Datta, Saara Laulumaa, Salla Keskitalo, Satu Hyvärinen, Sini Huuskonen, Sini Miettinen, Tiina Öhman and Xiaonan Liu. Thank you all for the supportive and pleasant working environment and especially Sini and Liu for your kindness and readiness to offer help whenever it is needed. Liu, you have been a godsend during
the defence process and I am very thankful for that. Tiina, Leena and Lisa, thank you for being great office roommates.

Finally, I want to extend my deepest gratitude to my family and friends whose love and support has been invaluable in my PhD journey. I am indebted to my mom Tsega and my dad Gebremichael for their love and nurture. My sisters Selamawit and Dr. Eyerusalem, my number one fans always saying jokingly “there isn’t anything you can’t do” after listening patiently to my complaints, thank you for being the best sisters. My lovely nieces and nephews (Selina, Maereg, Senay and Tiya) thank you for all the curious questions and discussions; you always brighten my days and inspire me to do more to be a good role model, I love you all dearly. The loves of my life, my husband Dawit and our son Fitih, words are not enough to express how grateful I am that you are part of my life. Dawito, thank you for the unwavering support and love, all of this is possible because of you. My darling Fitih, in the words of your current favorite song, you are my shining star, you indeed light up my life. My dear friends, Bezawit, Eyerusalem, Mihret, Betelhem, Hiwot, Hewan, Hanna and Elrom, I very much appreciate all your support and love, thank you.

Helsinki, February 2021

Rigbe G.Weldatsdik
6 REFERENCES


66


68


Käll, L., Storey, J.D., MacCoss, M.J., Noble, W.S., 2008b. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. J. Proteome Res. 7, 40–44. https://doi.org/10.1021/pr700739d


Automated Pipeline for SNP, SAP, and Splice Variant Detection in the Context of Proteogenomics. J. Proteome Res. 14, 3729–3737. https://doi.org/10.1021/acs.jproteome.5b00490


Lopez-Maestre, H., Brinza, L., Marchet, C., Kielbassa, J., Bastien, S., Boutigny, M., Monnin, D., Filali, A.E., Carareto, C.M., Vieira, C., Picard, F., Kremer, N.,


Nejentsev, S., Walker, N., Riches, D., Egholm, M., Todd, J.A., 2009. Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type


Noble, W.S., Keich, U., 2017. Response to “Mass spectrometrists should search for all peptides, but assess only the ones they care about.” Nat Methods 14, 644. https://doi.org/10.1038/nmeth.4339


Yang, Q., Zhang, M., Harrington, D.J., Black, G.W., Sutcliffe, I.C., 2010. A proteomic investigation of *Streptococcus agalactiae* grown under conditions
associated with neonatal exposure reveals the upregulation of the putative virulence factor C protein β antigen. International Journal of Medical Microbiology 300, 331–337. https://doi.org/10.1016/j.ijmm.2010.01.001


APPLICATION OF POOL-SEQ FOR VARIATION DETECTION AND PROTEOGENOMIC DATABASE CREATION IN ß-HEMOLYTIC STREPTOCOCCI

RIGBE GEBREMICHAEL WELDATSADIK

20/2021

HELSINKI 2021