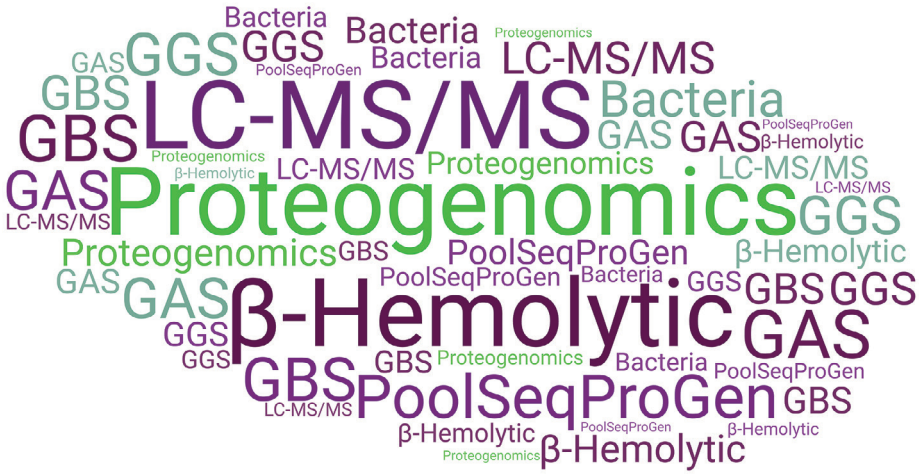
DISSERTATIONES SCHOLAE DOCTORALIS AD SANITATEM INVESTIGANDAM
UNIVERSITATIS HELSINKIENSIS

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

Supervisors

T.Sakari Jokiranta, Docent (Adjunct Professor)
Faculty of Medicine
University of Helsinki
Helsinki, Finland

Markku Varjosalo, Docent, Research director
Institute of Biotechnology
University of Helsinki
Helsinki, Finland

Thesis committee members

Jaana Vuopio, Professor
Institute of Biomedicine
University of Turku
Turku, Finland

Sampsa Hautaniemi, Professor, Research director
Faculty of Medicine
University of Helsinki
Helsinki, Finland

Thesis reviewers

Maija Vihinen-Ranta, Docent
Faculty of Mathematics and Science
University of Jyväskylä
Jyväskylä, Finland

Ulrich Bergmann, Docent
Faculty of Biochemistry and Molecular Medicine
University of Oulu
Oulu, Finland

Opponent

Laura Elo, Professor, Research director
Turku Bioscience Centre
University of Turku
Turku, Finland

Custos

Kari P. Keinänen, Professor
Institute of Biotechnology
University of Helsinki
Helsinki, Finland

Dissertationes scholae doctoralis ad sanitatem investigandam Universitatis Helsinkiensis
ISBN 978-951-51-7168-9 (paperback)
ISBN 978-951-51-7169-6 (PDF)
ISSN 2342-3161 (print)
ISSN 2342-317X (online)
https://ethesis.helsinki.fi/
Cover layout by Anita Tienhaara

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 down-sampling 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.

TIIVISTELMÄ

Proteogenomiikka on kehittyvä tieteenala, joka yhdistää genomiikkaa ja proteomiikkaa geenimallien parantamiseksi ja proteiinien tunnistamiseksi. Molempien alojen tekninen kehitys lisää tämän yhdistetyn tieteenalan mahdollisuuksia eri eliöiden toimintojen ymmärtämiseksi. Esimerkiksi nykyiset edulliset ja nopeat sekvensointitekniikat ovat mahdollistaneet useiden eri organismien kattavan sekvensoinnin, mikä luonnollisesti parantaa myös näiden organismien vertailuproteomien kattavuutta. Samanaikaisesti massaspektrometriatekniikan kehitys on johtanut proteomiikka-analyysien laadun paranemiseen ja syvyyden lisääntymiseen. Tämä mahdollistaa ennustettujen sekvenssialueiden (esim. uusien geenien) validoinnin.

Yhdistettyjen yksittäisten DNA-näytteiden sekvensointi (Pool-sekvensointi) mahdollistaa suurten näytemäärien sekvensoinnin erittäin kustannustehokkaasti. Se on luotettava menetelmä, jolla voidaan tunnistaa tarkasti eri näytteiden väliset vaihtelut. Pool-sekvensointikokeiden muunnelmia voidaan käyttää luomaan variantti-proteiinitietokantoja ja parantamaan massaspektrometriaan perustuvien proteiinitietokantojen kattavuutta. Tässä väitöskirjassa tutkittiin Pool-sekvensoinnin tehokkuutta eri varianttien tunnistamisessa ja alleelitaajuuksien arvioimisessa kolmen β-hemolyyttisen streptokokki-bakteerin (GAS, GGS ja GBS) kannoista. Lisäksi työssä kehitettiin uusi Python-ohjelmointikielellä kirjoitettu ohjelmisto ('PoolSeqProGen') proteiinivariantitietokantojen luomiseksi Pool-sekvensointi -kokeista. Tämä on ensimmäinen työ, jossa Pool-sekvensointia käytettiin sekvensoimaan suuri määrä streptokokkeja ja arvioimaan menetelmän tehokkuutta geneettisesti polymorfisissa bakteereissa. "PoolSeqProGen" -työkalu on myös ensimmäinen ja ainoa saatavilla oleva työkalu proteogenomisten tietokantojen luomiseen Pool-sekvensoinnilla tuotetusta datasta.

Pool-sekvensointimenetelmän tarkkuus variaation tunnistamisessa ja alleelitaajuuden estimoinnissa arvioitiin vertaamalla variantteja, jotka tunnistettiin poolien näytteiden yksittäisistä sekvensoinneista. Pool-sekvensoinnin herkkyys oli erittäin hyvä (> 90%) kaikkien testattujen kuuden GAS-kannan geenivariaatioiden tunnistamisessa. Sillä havaittiin myös suurin osa neljänkymmenen GBS-kannan erikseen sekvensoiduista genominäytteistä. Pool-sekvensoinnin alleelitaajuusestimaatit ja GBS-datan yksittäiset sekvenssit korreloivat myös erittäin hyvin ($R = 0{,}96$). Sekvensoinnin kattavuuden vaikutus Pool-sekvensoinnin herkkyyteen arvioitiin alinäytteenotto-analyyseillä.

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 muihin proteogenomisiin tietokantoihin, jotka on luotu Prodigalin ab initio -geeniennusteista (MEGAHIT) ja Sixgill-työkalun luomiin 6-kehyksisiin käännöspohjaisiin 'Metapeptideihin', tunnistettiin vastaavasti noin 4% ja 19% enemmän peptidejä.

Kehitettäessä massaspektrometria tietokantoja avoimiin pangenomeihin perustuville organismeille, kutenβ-hemolyyttisille streptokokeille GAS, GBS ja GGS, useiden edustavien kantojen sekvensointi ja annotointi on ensiarvoisen tärkeää. Massaspekrometrian lisääntynyt käyttö tartuntatautien diagnosoinnissa parantaa näiden mikrobien aiheuttamien sairauksien diagnosointia ja mahdollistaa siten myös hoidon paremman kohdentamisen. Tässä väitöskirjatyössä osoitetaan, että Pool-sekvensointia voi käyttää kustannustehokkaasti ja tarkasti polymorfisten bakteerikantojen välillä esiintyvien variaatioiden tunnistamiseen. Lisäksi havainnollistamme yhteen genomiin pohjautuvien tietokantojen laajentamiseksi kehitetyn työkalun hyödyllisyyttä, jolla voidaan parantaa tietokantojen kattavuutta ja peptidi- ja proteiinitunnistusta käyttämällä Pool-sekvensointikokeissa tunnistettuja variantteja.

CONTENTS

ABBREVIATIONS

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

ORIGINAL PUBLICATIONS

Publication I Weldatsadik, RG, Wang, J., Puhakainen, K., Jiao, H., Jalava, J., Raisanen, K., Datta, N., Skoog, T., Vuopio, J., Jokiranta, TS, & Kere, J. (2017). Sequence analysis of pooled bacterial samples enables identification of strain variation in group A Streptococcus. Sci. Rep., 7, [45771]. https://doi.org/10.1038/srep45771

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.

Publication II Rigbe G Weldatsadik, Markku Varjosalo, Sakari T Jokiranta. (Submitted for publication). PoolSeqProGen: a Python package for the generation of proteogenomic protein databases from Pool-seq derived variants.

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

Publication III Weldatsadik, RG, Datta, N., Kolmeder, C., Vuopio, J., Kere, J., Wilkman, SV, Flatt, JW, Vuento, R., Haapasalo, KJ, Keskitalo, S., Varjosalo, M., & Jokiranta, TS (2019). Pool-seq driven proteogenomic database for Group G Streptococcus. J Proteomics, 201, 84-92. https://doi.org/10.1016/j.jprot.2019.04.015

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).

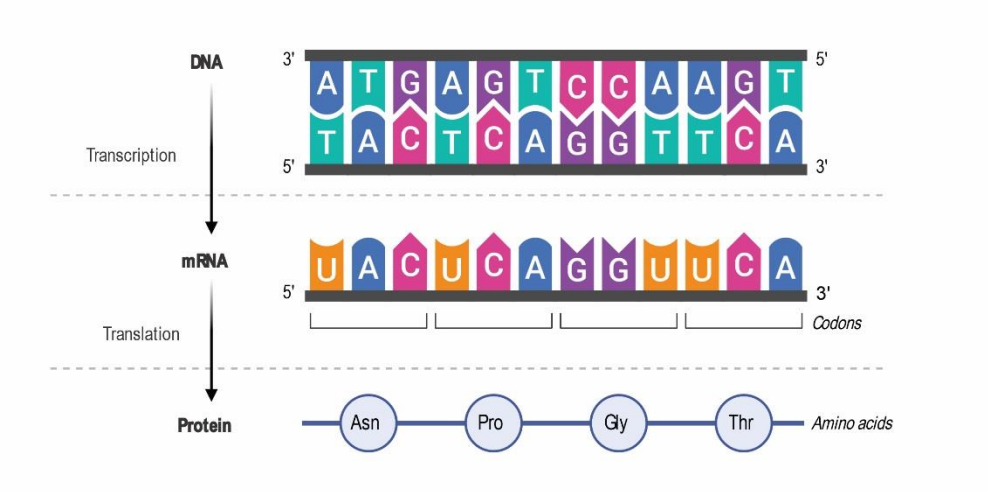


Figure 1. The central dogma of molecular biology that depicts the one way genetic information flow from DNA to mRNA and finally to protein (which are composed of 20 amino acids specified by codons). Created with BioRender.com

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).

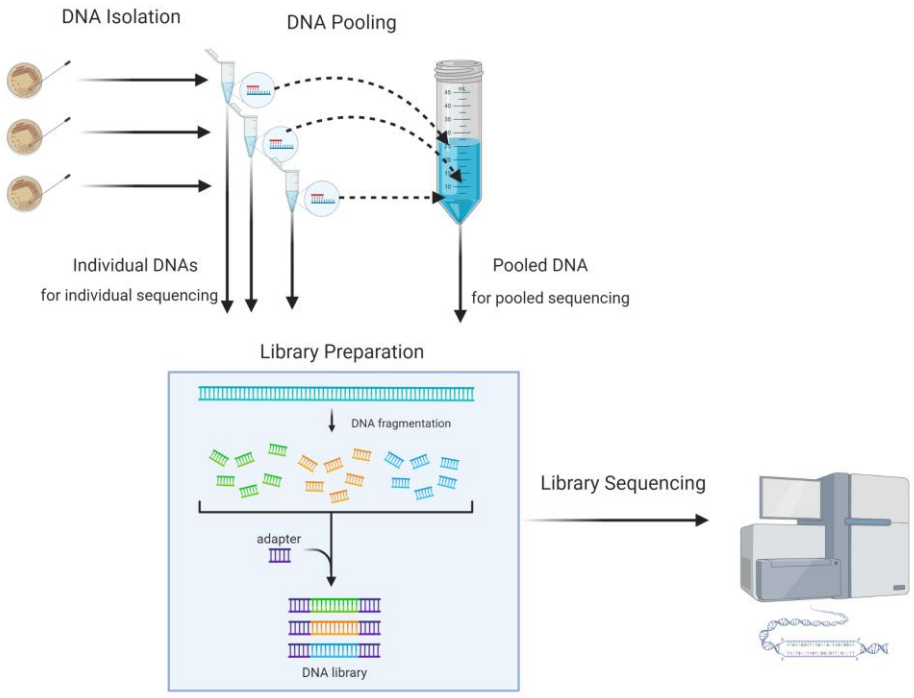


Figure 2. The steps involved in Pool-seq. After DNA is isolated from multiple samples, it is pooled before library preparation, unlike in individual sequencing in which a library is prepared from each DNA genome isolated. The prepared library is then sequenced usually with NGS platforms such as Illumina.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

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; Hajirasouliha 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 $Pr(G|R) \propto Pr(R|G) * Pr(G)$, 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 in to 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, Højrup, 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.

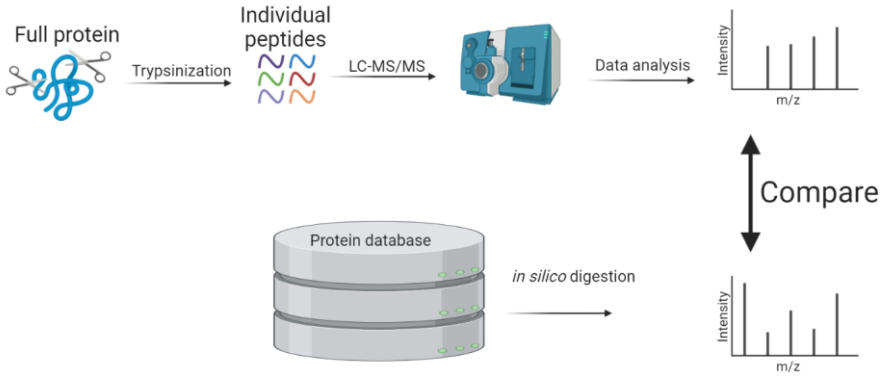


Figure 3. The commonly used bottom-up shotgun proteomics analysis workflow. The proteins are first digested with a protease such as Trypsin and analyzed by LC-MS/MS. Peptide/protein identification is carried out using database searching by comparing the experimental spectra to the theoretical spectra 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; Diament 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 in to 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 unrestrictive, 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 Paek, 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

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 the16s 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 25mM dithiothreitol (45 min at 37°C) and 50mM 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.

Table 1. 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.

Databases		Tryptic peptide size
Single genome	GAS (AE004092)	72272
	GBS (NEM316_III)	43047
	GGS (ATCC12394)	38173
Multi-genome	GAS (49complete/294assembly)	89705/20469
	GGS (4complete/26assembly)	57898/113208
Pool-seq driven	GAS	91650
	GBS	62256
	GGS	72050
Megahit + Prodigal ab initio		153248
Sixgill Metapeptides		251946

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.

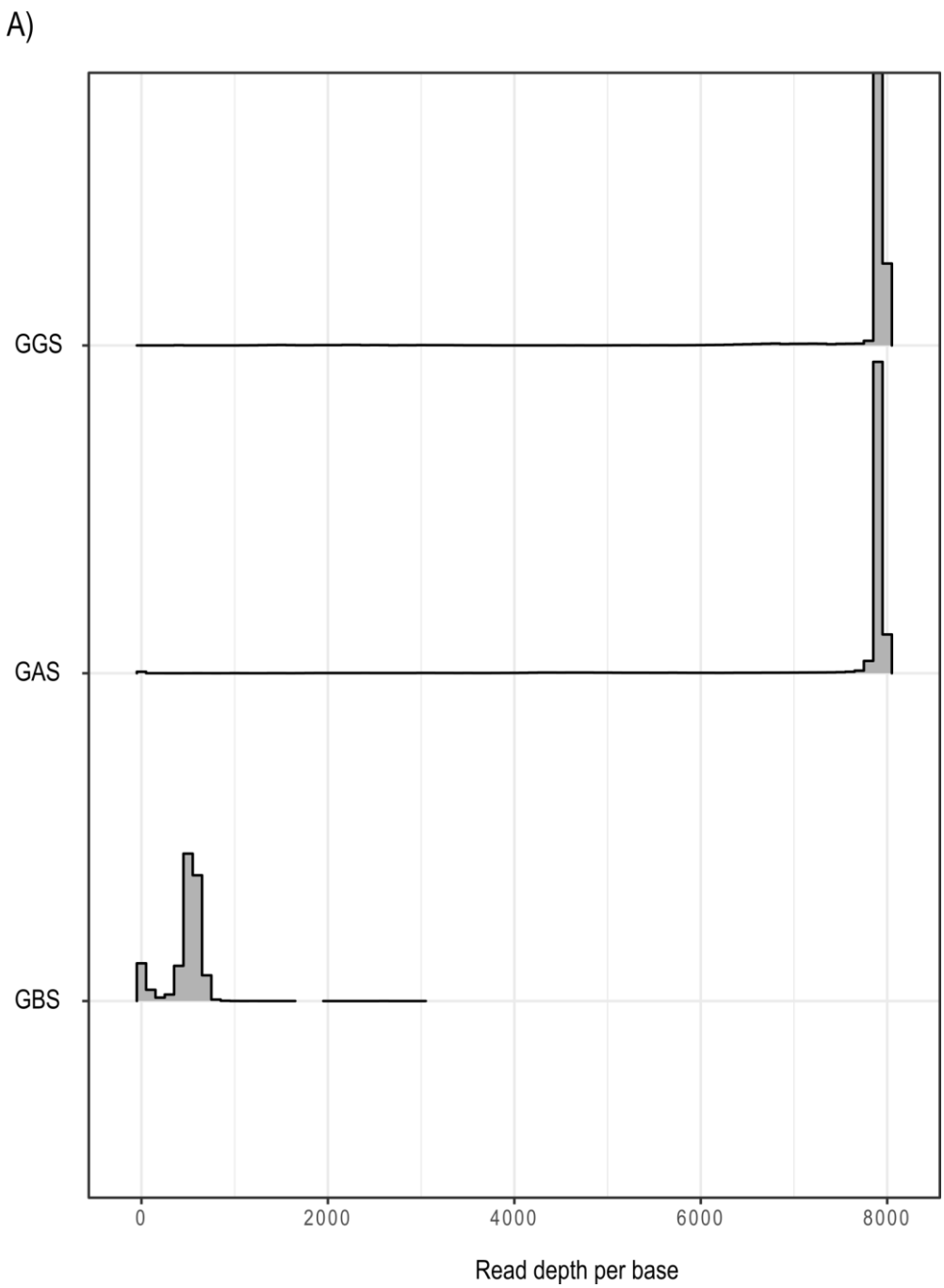
4.1 A sequencing coverage of ~200-250X is optimal for accurately identifying variants from bacterial Pool-seq data

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 (~200X), 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.


A)
GGS
GAS
GBS
0
2000
4000
6000
8000
Read depth per base

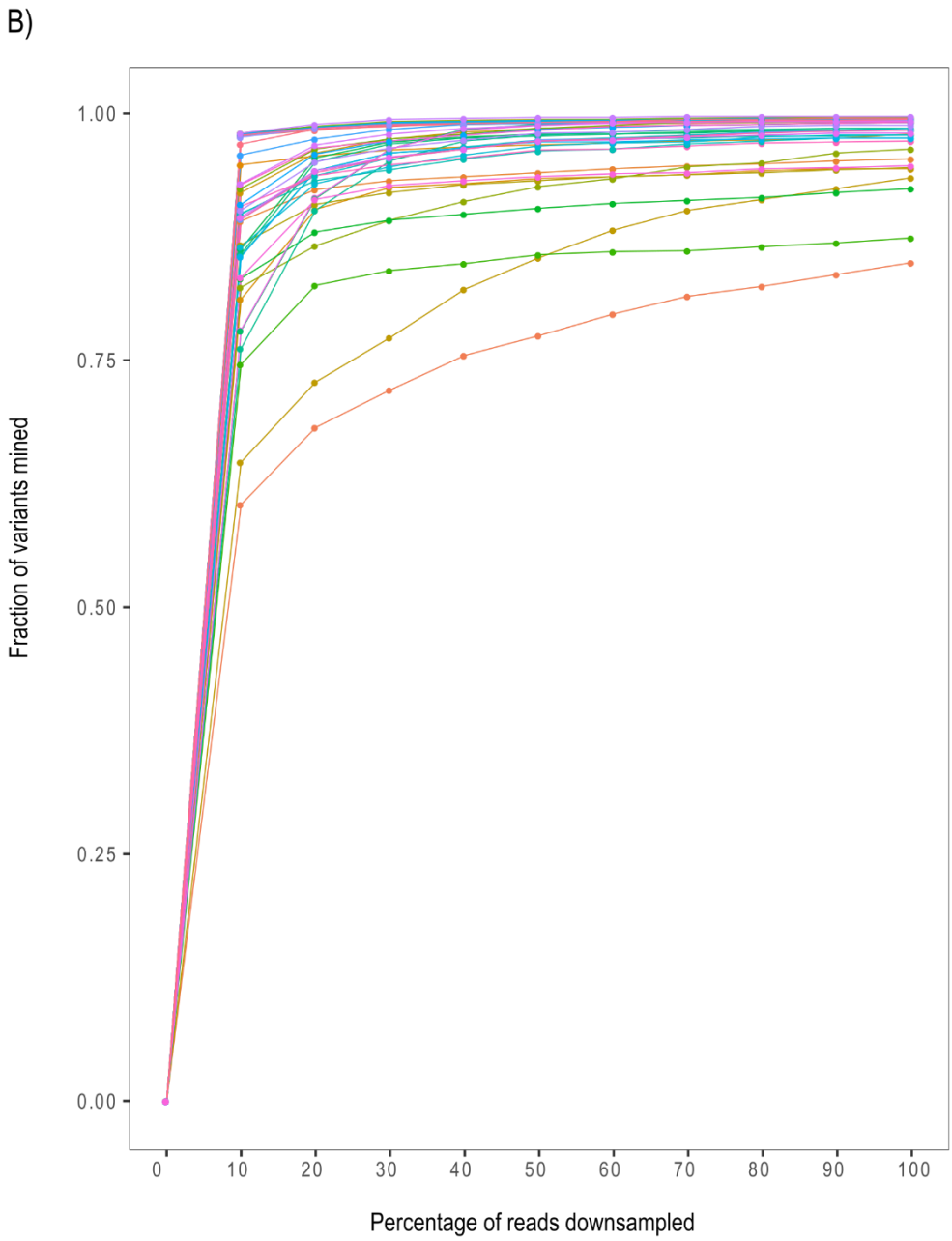


Figure 4. Sequencing coverage and its effect on accuracy of Pool-seq. A) The aligned read depth per base of the three Pool-seq experiments included in this thesis. While most of the bases in the GAS and GGS Pool-seq studies were covered by ~8000 reads, the GBS experiment had an average of ~600 reads 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 2. The recall (sensitivity) and precision of the four variant calling tools for GAS Pool-seq data as assessed by the overlap of calls in 2 or more tools.

	GAS pool 1		GAS pool 2	
	Recall	Precision	Recall	Precision
FreeBayes	0.98	0.86	0.99	0.85
GATK UG	1.00	0.94	1.00	0.96
SAMtools	0.38	1.00	0.38	0.99
SNVer	0.98	0.90	0.98	0.90

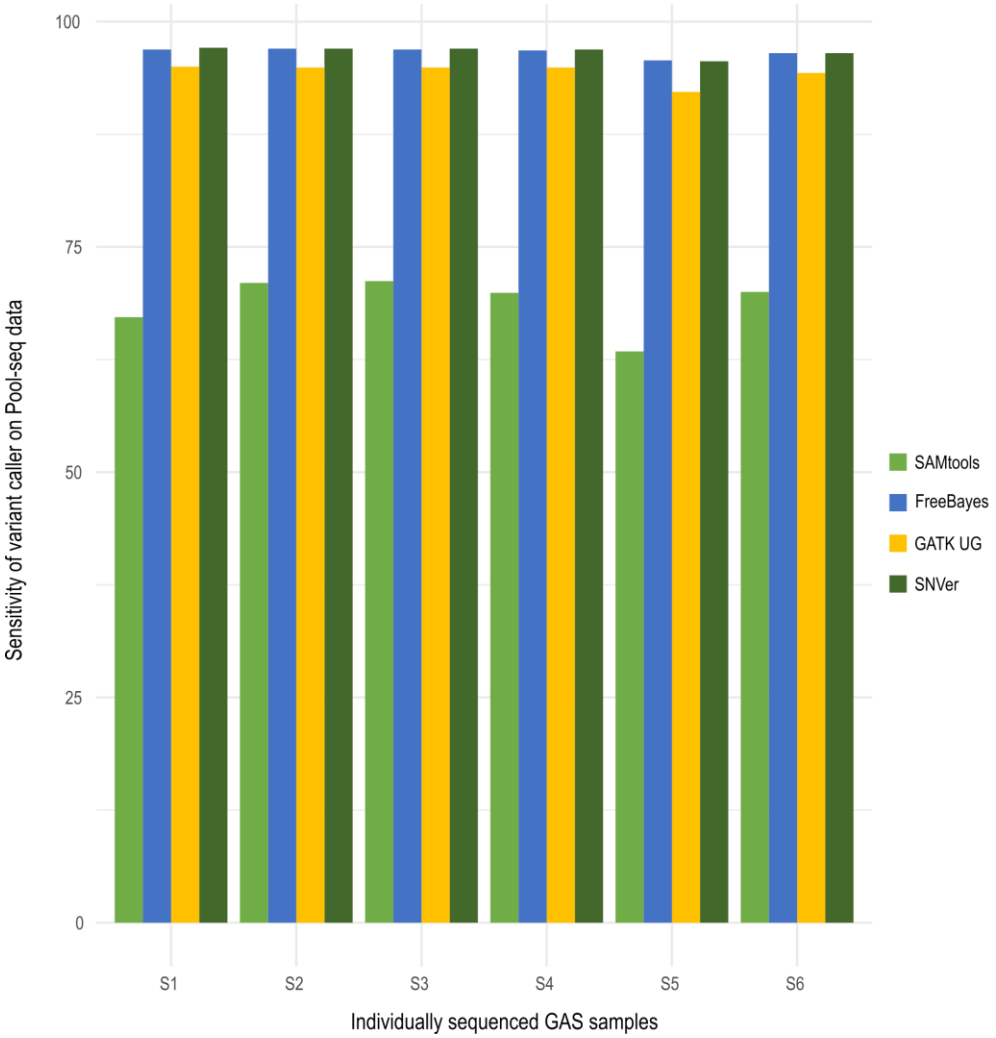
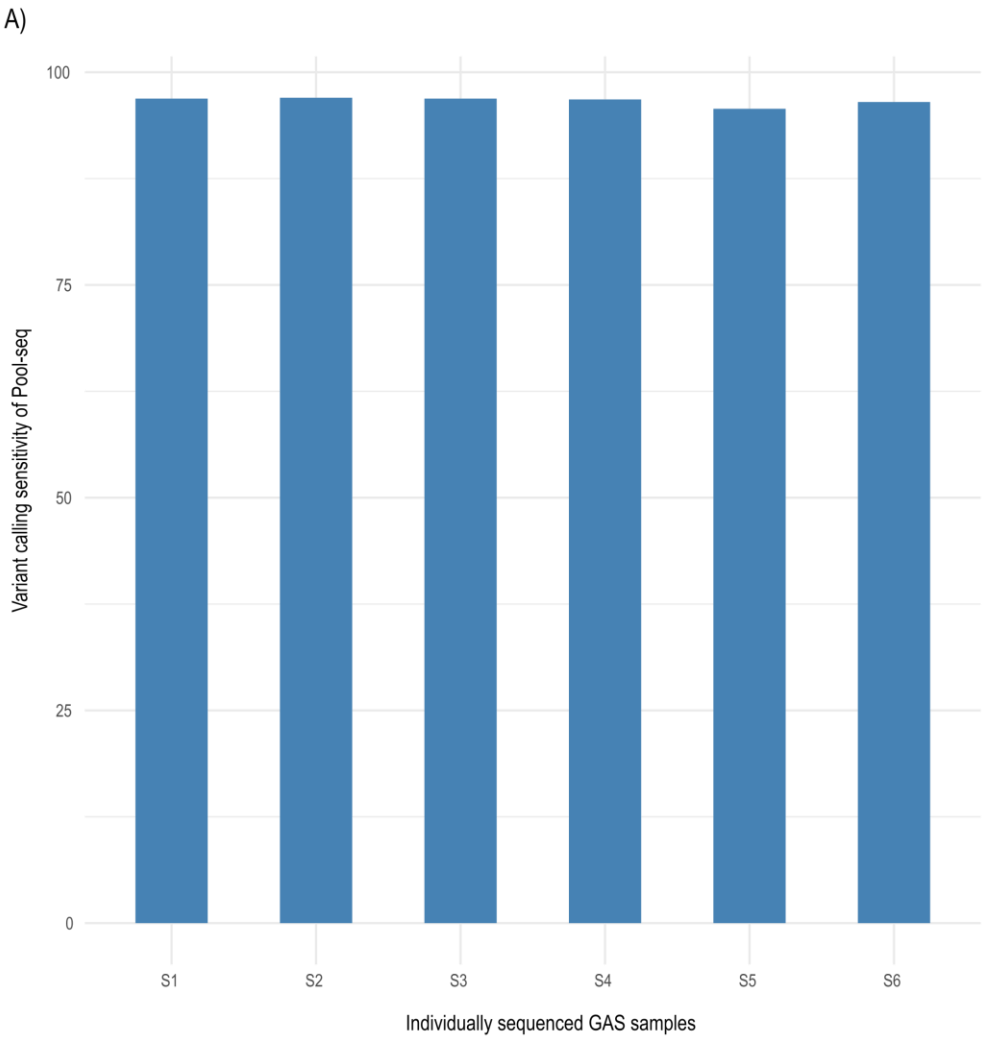


Figure 5. The sensitivity of the 4 variant calling tools when used with the GAS Pool-seq samples.

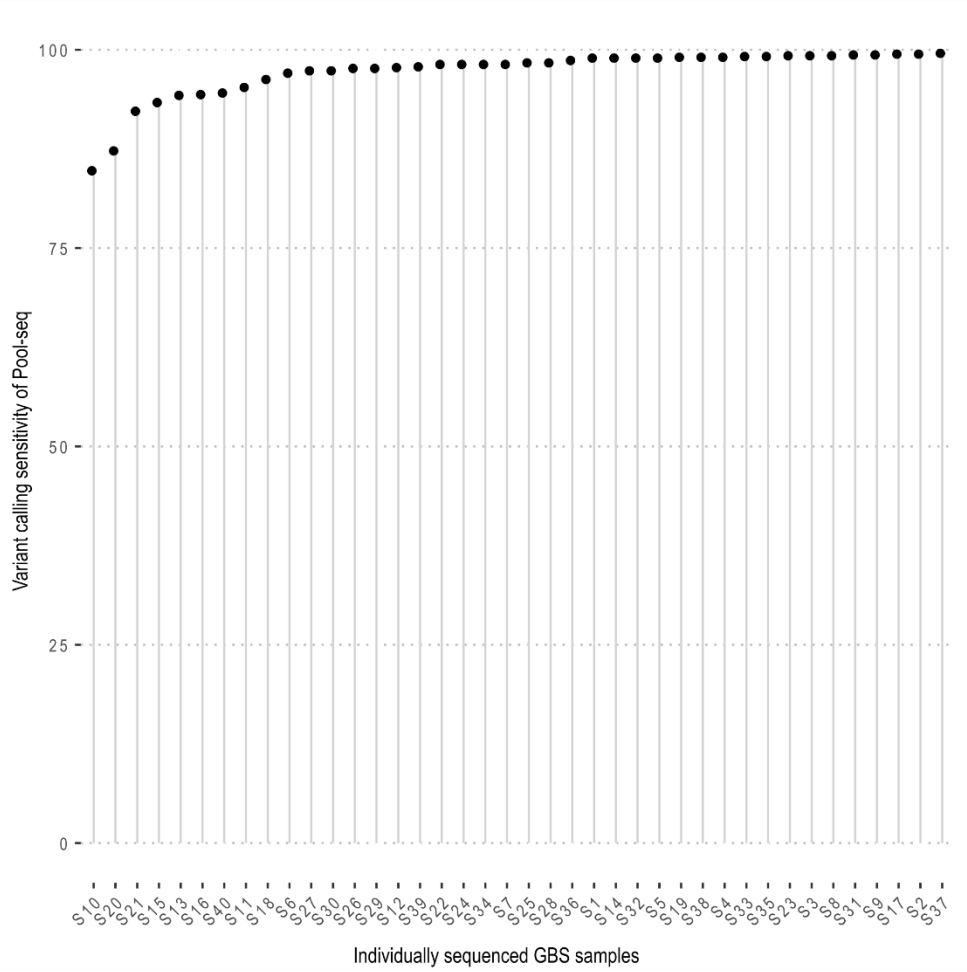
4.3 Pool-seq is robust for accurate SNP detection and allele frequency estimation

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).



B)


100
75
50
25
0
Variant calling sensitivity of Pool-seq
S10 S20 S21 S15 S13 S16 S40 S11 S18 S6 S27 S30 S26 S29 S12 S39 S22 S24 S34 S7 S25 S28 S36 S1 S14 S32 S5 S19 S38 S4 S33 S35 S23 S3 S8 S31 S9 S17 S2 S37
Individually sequenced GBS samples

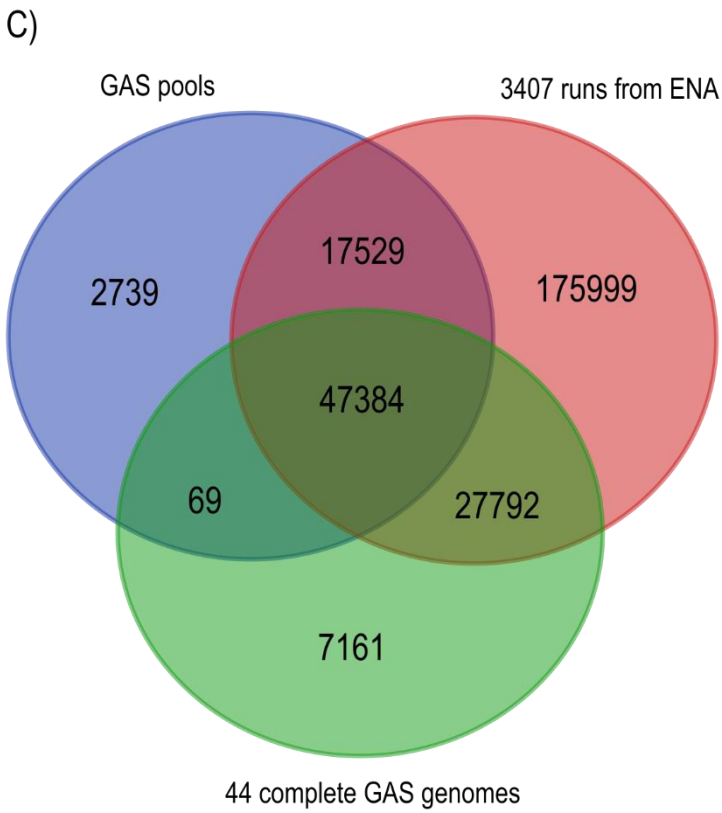
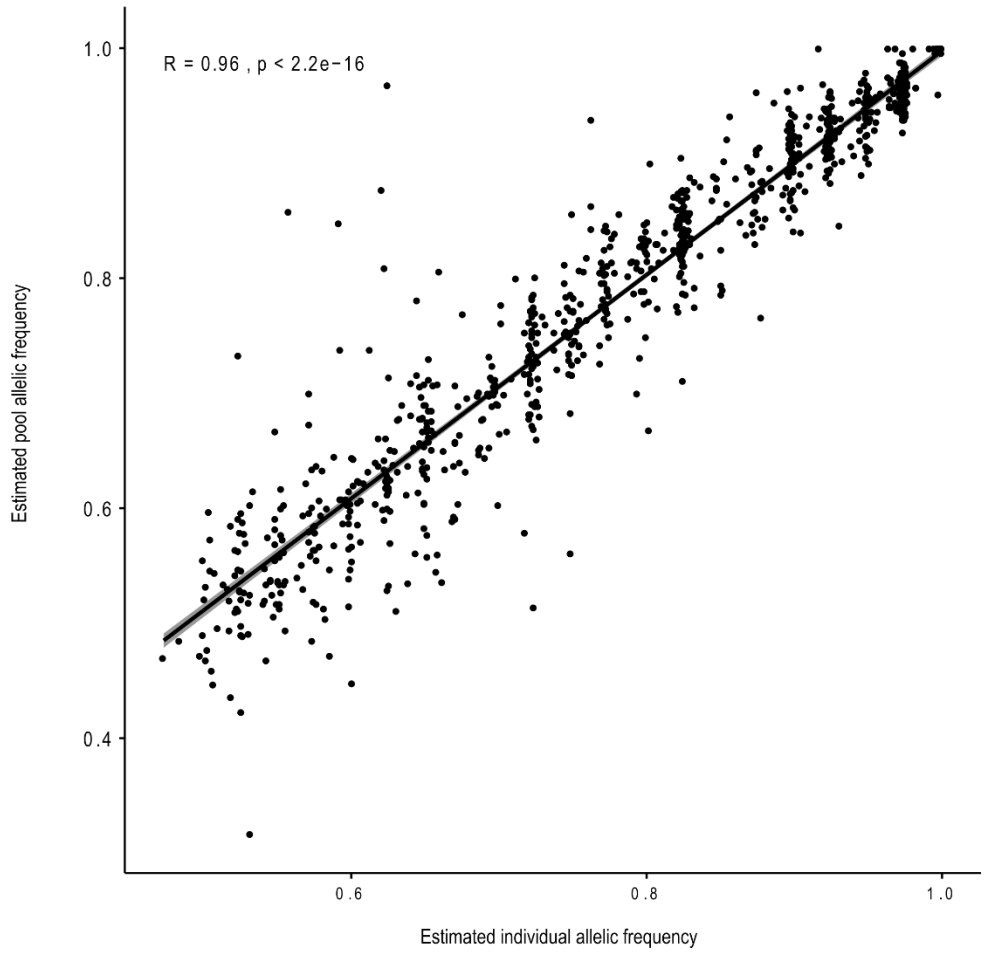


Figure 6. Pool-seq's variant identification sensitivity. The sensitivity of Pool-seq to identify variants from A) 6 GAS and B) 40 GBS individually sequenced samples. C) The overlap of variants mined from the GAS Pool-seq and publicly available sequence data. In A, FreeBayes, SNVer and GATK's UG were 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 ($R=0.96$, $p<2\text{-}2e\text{-}16$) (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.

A)


R = 0.96 , p < 2.2e−16
1.0
0.8
0.6
0.4
0.6
0.8
1.0
Estimated pool allelic frequency
Estimated individual allelic frequency

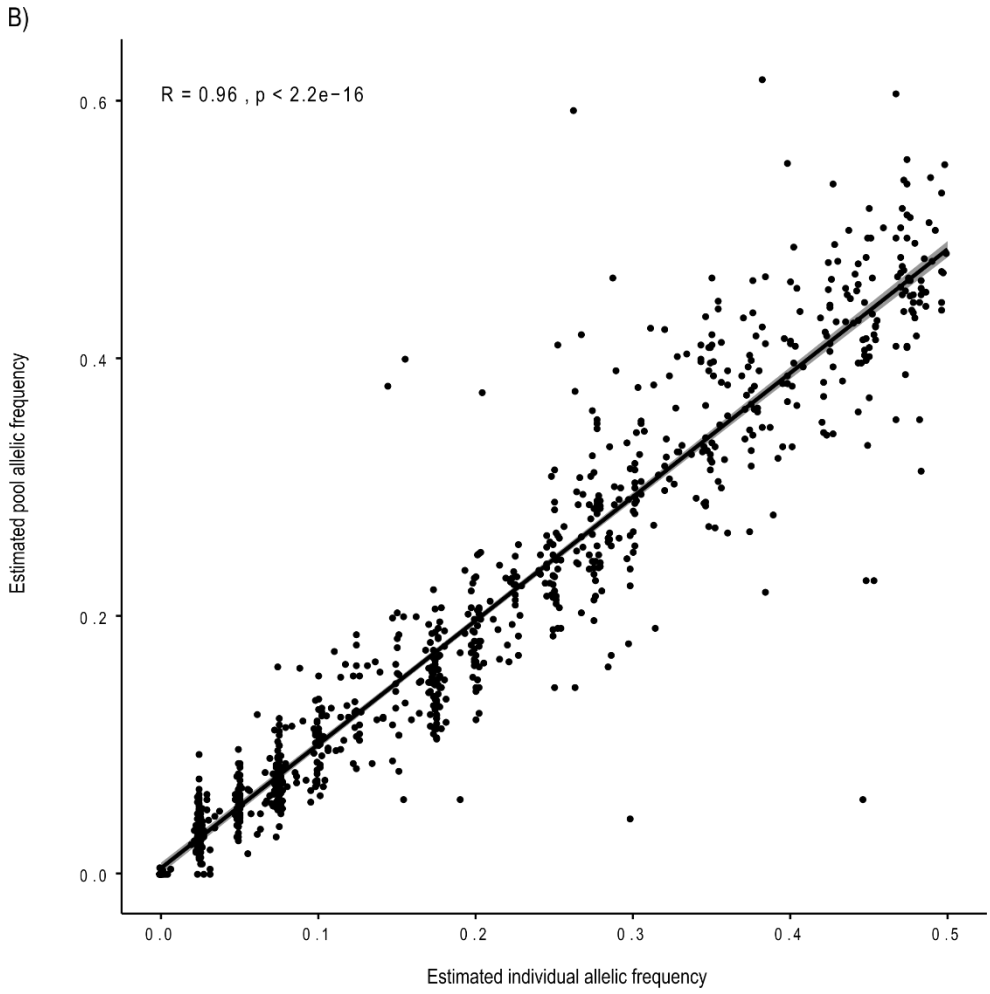


Figure 7. Allele frequency estimates from GBS pools and individually sequenced samples of the A) reference (major) allele and B) the variant (minor) alleles. From the 33716 total number of variants, 1000 were randomly chosen for easier display. The frequencies were estimated by using SAMtools mpileup (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.

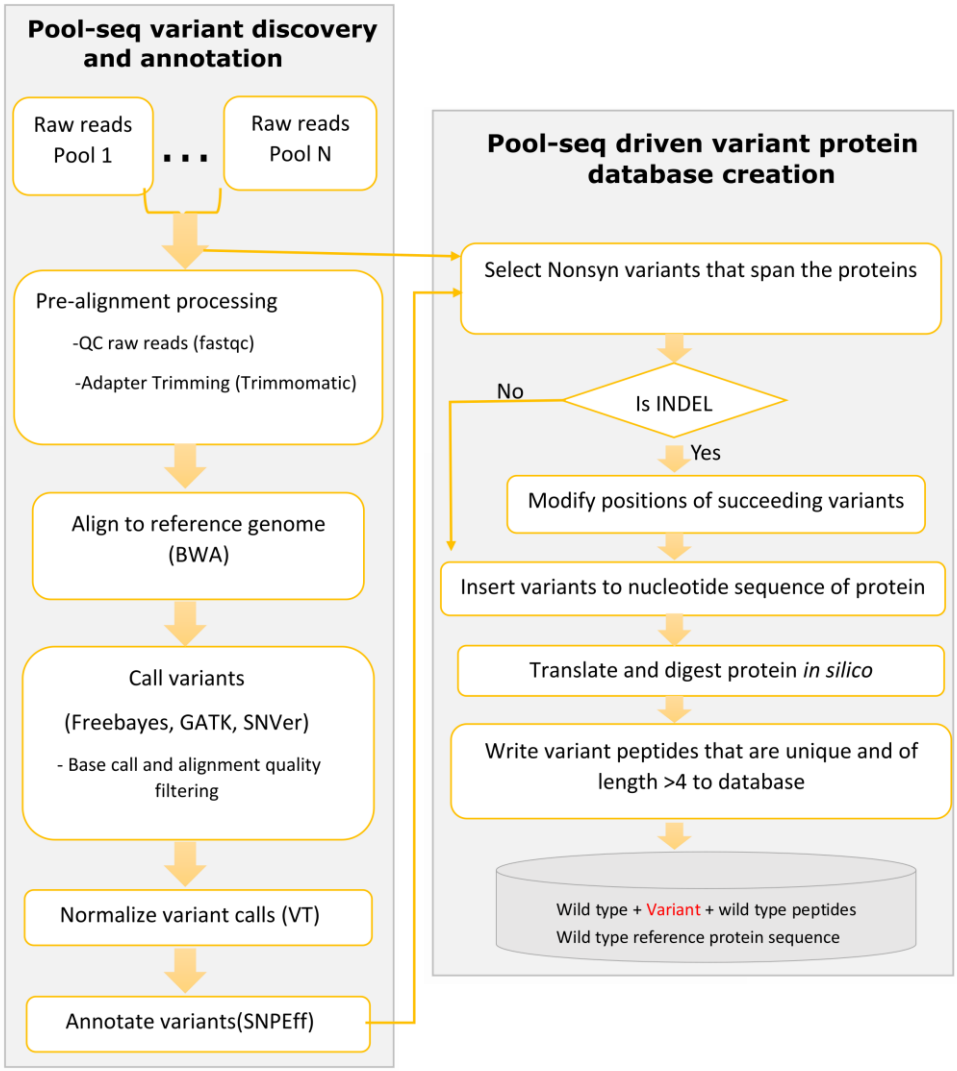


Figure 8. Workflow of the Pool-seq driven proteogenomic database creation. After variants are identified and annotated, those with non-synonymous effects are inserted to the nucleotide sequence of the proteins they span and the sequence is then in silico translated and digested. The variant peptides and the flanking peptides are written to the database together with the wild type reference protein sequences.

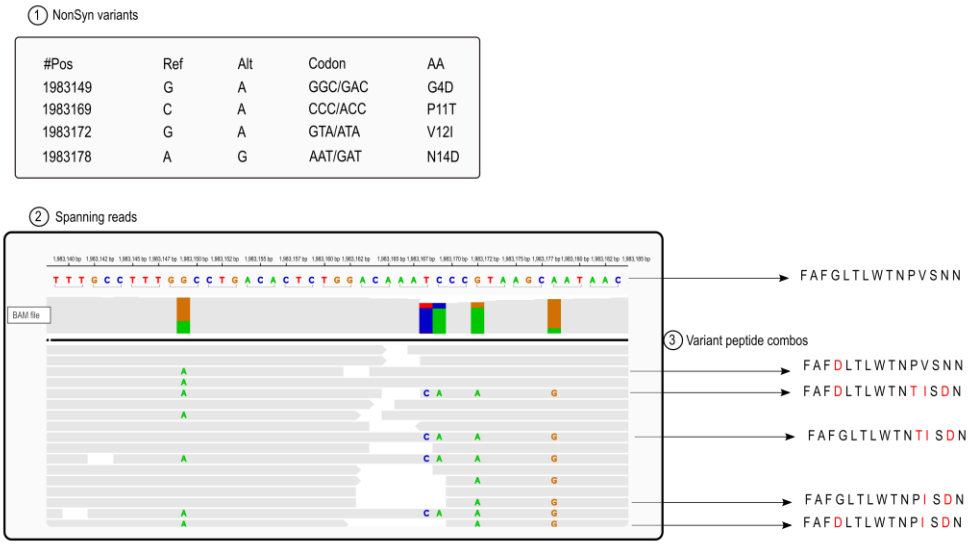


Figure 9. How the different combinations of Non synonymous variants that are 'seen' in the reads spanning the variants are incorporated to the variant peptides.

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.

Id	Length	Name	AA	Variant peptides	Total
WP_003056081.1	304	Protein jag	S55N&D56G	KPAQVDIEGINGK	17
			D56G	KPAQVDIEGISGK	
			I82V	QNAPVVNPADVELEEMK	
WP_014612608.1	155	Transcriptional repressor	I143V	IPVIAYGVCPECQAK	2
			I143V&A149	IPVIAYGVCPECQVK'	
WP_014611973.1	511	M protein	E89D&L94F	EVADYNSLFDK	54
			A113T&E115K	VVNDSLQTTK	
			E105K&A113T&E115K	MKVVNDSLQTTK	
			D141N&Y145S&T149A	NKEFSLGEALR	
			Y145S&T149A	EFSLGEALR	
			K192Y&A193Q	QTLEAEYQK	
			K192Y&A193E	QTLEAEYEK	
			D184Y&K192Y&A193Q	AEAYRQTLEAEYQKLEEEK	
			E273D	LEEQNKISDASR	
			E442D	ASDSQTPDATPGNKVVPGK	
			E442D&T444K	AGKASDSQTPDAKPGNK	
WP_003059396.1	145	YtxH domain-containing protein	A40S	AYQSYKENPDDYHQLAK	11
			N44S&L51F	AYQAYKESPDDYHQFAK	
			V104A	TKETLAEVEAK	
WP_014612218.1	699	ATP-dependent Clp protease ATP-binding subunit	T335A	SLEAEMATQK	48
			E415G	GHVIGQDGAVEAVAR	
			E415G&V427A	GHVIGQDGAVEAAAR	
			T582A	NTVIIATSNAGFGHQEDENTDQPAIMDR	
WP_014611976.1	440	Kinase	L88P	SKPFATDSGAMPHKLEK	32
			I155T	VYFADKDGSVTLPTQPVQEFLLK	
			I155T&K164S	VYFADKDGSVTLPTQPVQEFLLSGHVR	
			R202K&G204V	SVDVEYTVQFTPLNPDDDFKPVLKDTK	
			G204V	SVDVEYTVQFTPLNPDDDFRPVLKDTK	
			I220V	TLAIGDTVTSQELLAQAQSILNK	

			N254K	DSSIVTHDKDIFR	
			H270R	TILPMDQEFTYR	
WP_014612152.1	269	Cof-type HAD-IIB family hydrolase	I29V&D31E	ITDDVFQAVQEAK	5
			D31E	ITDDVFQAIQEAK	
WP_014611901.1	371	Redox-regulated ATPase YchF	D38N&I41V	AGAEAANYPFATINPNVGM VEVPDER	22
			I41V	AGAEAANYPFATIDPNVGM VEVPDER	
WP_014612591.1	873	SEC10/PgrA surface exclusion domain-containing protein	A63T	ASNTSEESLPKTETCEETK	105
			A129G&K133E	ALTSAQEIYTNTLASSEETLL GQGAEYQR	
			K133E	ALTSAQEIYTNTLASSEETLL AQGAEYQR	
			T207A	AAQTANDNTKALSSELEK	
			V774I	IDTTPLVQEMIK	
			P752L	HLDEDIATVPDLQVAPLLTG VKPLSYSK	
WP_003058857.1	350	BMP family ABC transporter substrate-binding protein	A332D	EALKDIEEAK	15
			A332D&S338A	EALKDIEEAKAK	

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.

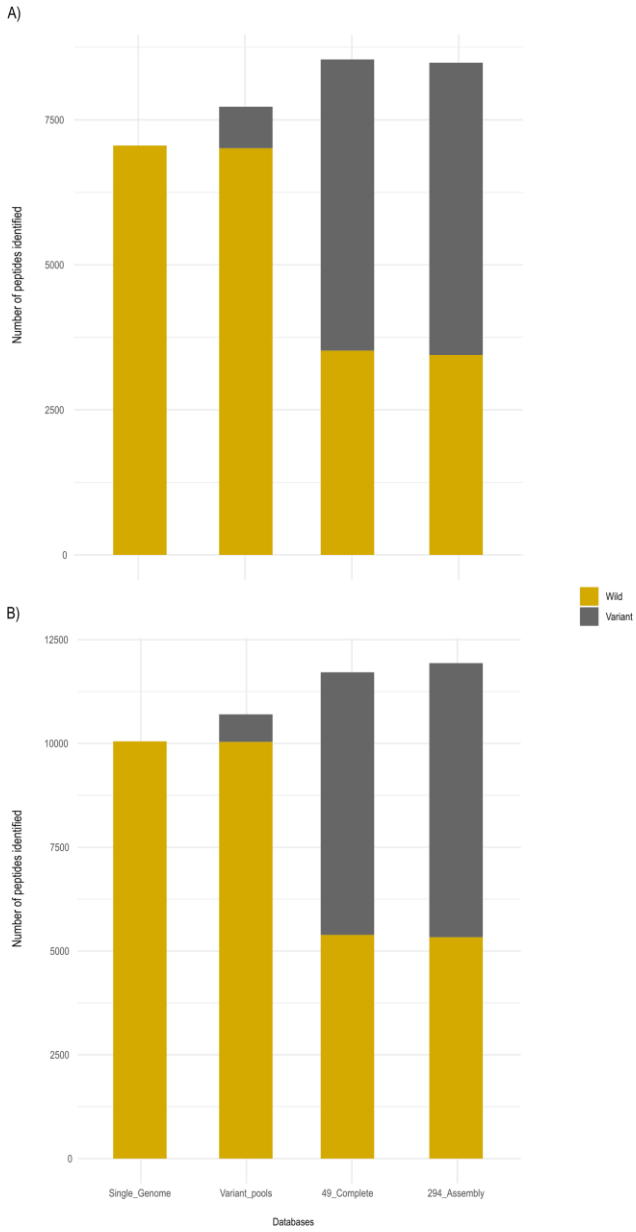


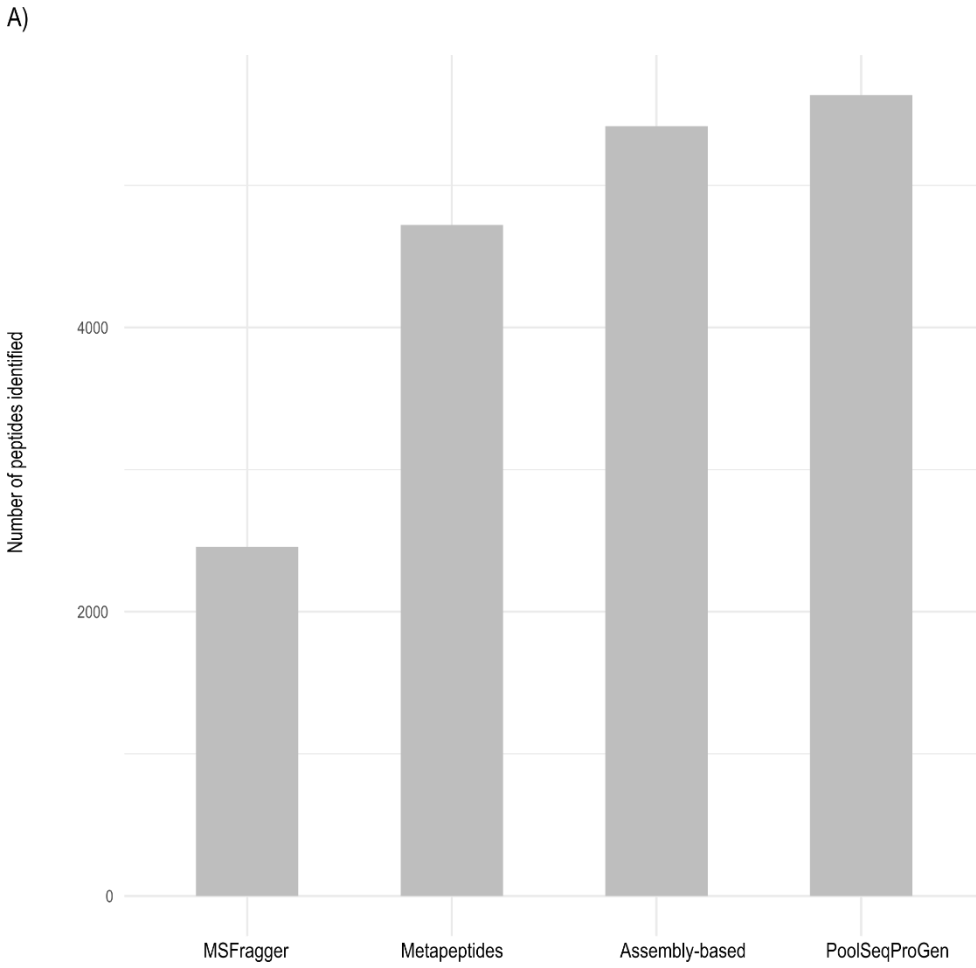
Figure 10. The number of peptides identified when using the Pool-seq driven variant databases (Variant_Pools) compared to conventional single genome and multi-genome (from 49

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.

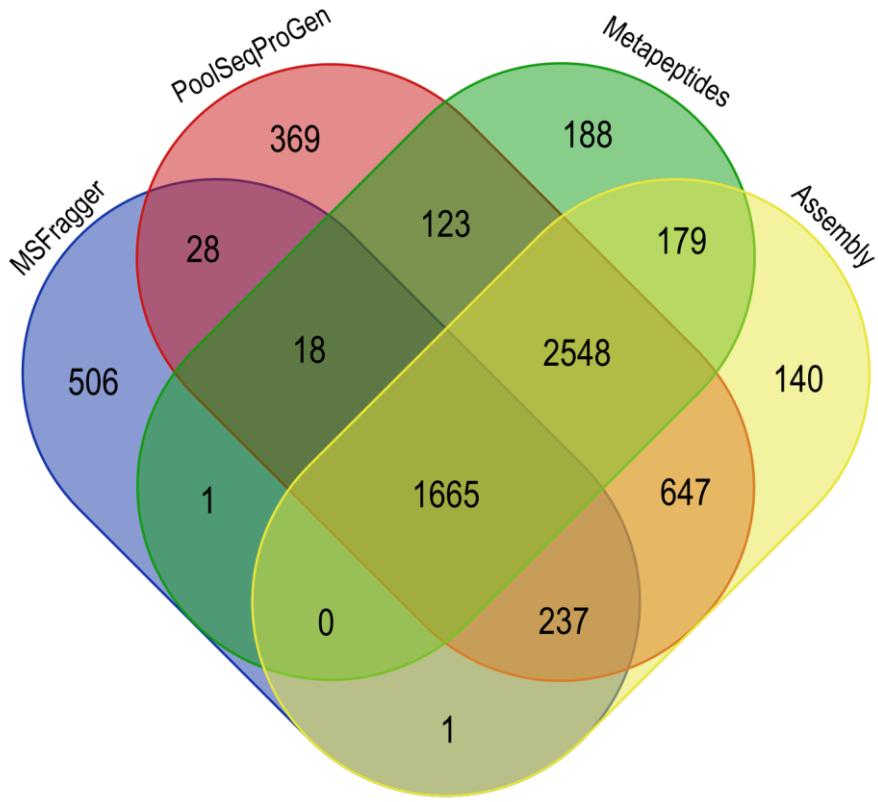
Other options for constructing protein databases from genome sequence information include ab initio gene prediction after assembly and predicting genes directly from the sequence reads for prokaryotes without assembly since most part of their genomes is protein coding. Both methods do not require a reference 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).

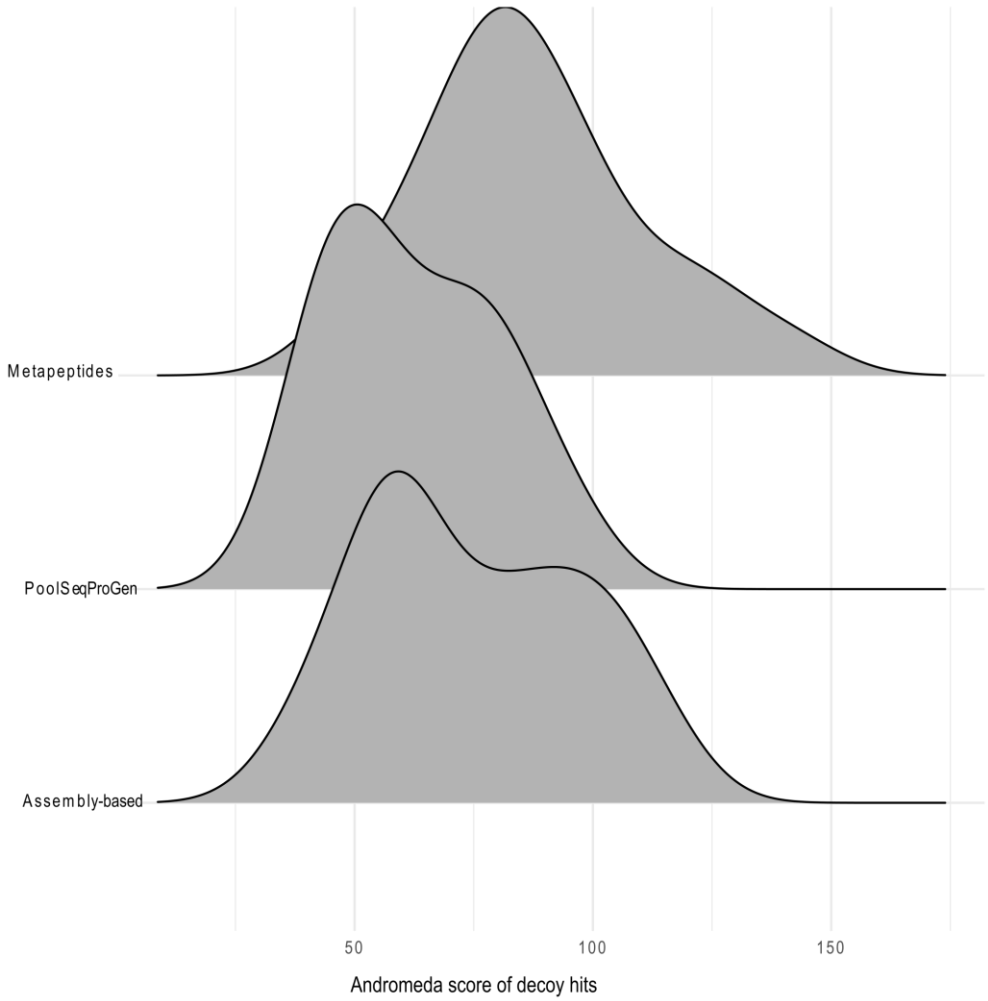
Open search tools are other alternative approaches that are mainly applied for the identification of unaccounted for chemical and post translational modifications and could help fill the gap of identifying unassigned spectra. In proteogenomics studies it has been recommended to rule out that novel peptides 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 MSFragger. 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 MSFragger 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.


A)
Number of peptides identified
4000
2000
0
MSFragger
Metapeptides
Assembly-based
PoolSeqProGen

B)


PoolSeqProGen
Metapeptides
MSFragger
Assembly
369
188
28
123
179
506
18
2548
140
1
1665
647
0
237
1

C)


Metapeptides
PoolSeqProGen
Assembly-based
50
100
150
Andromeda score of decoy hits

D)

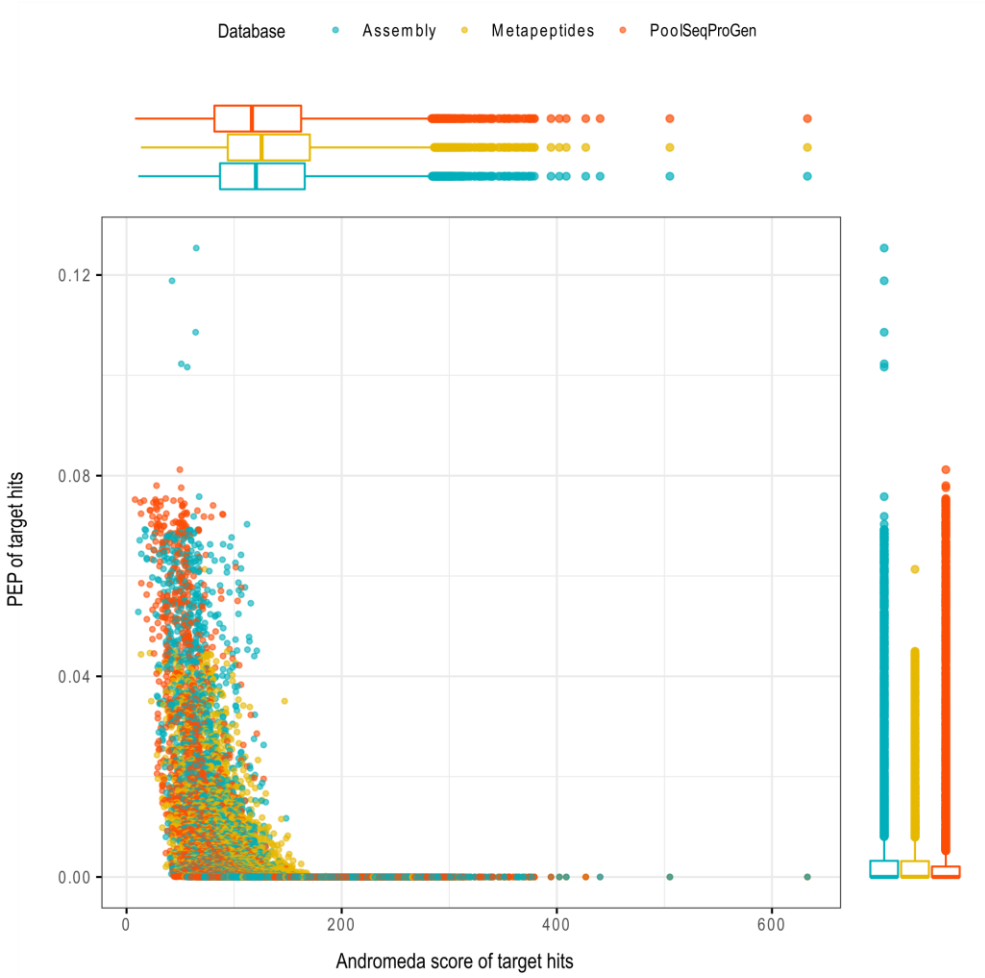


Figure 11. Comparison of the Pool-seq driven approach with other proteogenomics methods i.e., assembly based ab initio prediction and the non-assembly 'Metapeptides' approaches as well as an open search method MSFragger. A) The number of unique peptides identified from these approaches B) The overlap 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 statisitcal 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 Moelcular 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, Betelehem, Hiwot, Hewan, Hanna and Elrom, I very much appreciate all your support and love, thank you.

Helsinki, February 2021

Rigbe G.Weldatsdik

6 REFERENCES

Abby, S., Daubin, V., 2007. Comparative genomics and the evolution of prokaryotes. Trends in Microbiology 15, 135–141. https://doi.org/10.1016/j.tim.2007.01.007

Adams, J.M., Jeppesen, P.G.N., Sanger, F., Barrell, B.G., 1969. Nucleotide Sequence from the Coat Protein Cistron of R17 Bacteriophage RNA. Nature 223, 1009–1014. https://doi.org/10.1038/2231009a0

Ahn, J.-M., Kim, M.-S., Kim, Y.-I., Jeong, S.-K., Lee, H.-J., Lee, S.H., Paik, Y.-K., Pandey, A., Cho, J.-Y., 2014. Proteogenomic Analysis of Human Chromosome 9-Encoded Genes from Human Samples and Lung Cancer Tissues. J. Proteome Res. 13, 137–146. https://doi.org/10.1021/pr400792p

Albers, C.A., Lunter, G., MacArthur, D.G., McVean, G., Ouwehand, W.H., Durbin, R., 2011. Dindel: Accurate indel calls from short-read data. Genome Research 21, 961–973. https://doi.org/10.1101/gr.112326.110

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., Nielsen, P.H., 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 31, 533–538. https://doi.org/10.1038/nbt.2579

Altmann, A., Weber, P., Quast, C., Rex-Haffner, M., Binder, E.B., Muller-Myhsok, B., 2011. vipR: variant identification in pooled DNA using R. Bioinformatics 27, i77–i84. https://doi.org/10.1093/bioinformatics/btr205

Amaral, A.J., Ferretti, L., Megens, H.-J., Crooijmans, R.P.M.A., Nie, H., Ramos-Onsins, S.E., Perez-Enciso, M., Schook, L.B., Groenen, M.A.M., 2011. Genome-Wide Footprints of Pig Domestication and Selection Revealed through Massive Parallel Sequencing of Pooled DNA. PLoS ONE 6, e14782. https://doi.org/10.1371/journal.pone.0014782

Anand, S., Mangano, E., Barizzone, N., Bordoni, R., Sorosina, M., Clarelli, F., Corrado, L., Martinelli Boneschi, F., D'Alfonso, S., De Bellis, G., 2016. Next Generation Sequencing of Pooled Samples: Guideline for Variants' Filtering. Sci Rep 6, 33735. https://doi.org/10.1038/srep33735

Apweiler, R., 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Research 32, 115D – 119. https://doi.org/10.1093/nar/gkh131

Apweiler, R., Bairoch, A., Wu, C.H., 2004. Protein sequence databases. Current Opinion in Chemical Biology 8, 76–80. https://doi.org/10.1016/j.cbpa.2003.12.004

Bafna, V., Edwards, N., 2001. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics 17, S13–S21. https://doi.org/10.1093/bioinformatics/17.suppl_1.S13

Bansal, V., Harismendy, O., Tewhey, R., Murray, S.S., Schork, N.J., Topol, E.J., Frazer, K.A., 2010. Accurate detection and genotyping of SNPs utilizing population sequencing data. Genome Research 20, 537–545. https://doi.org/10.1101/gr.100040.109

Bansal, V., 2010. A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics 26, i318–i324. https://doi.org/10.1093/bioinformatics/btq214

Barbuddhe, S.B., Maier, T., Schwarz, G., Kostrzewa, M., Hof, H., Domann, E., Chakraborty, T., Hain, T., 2008. Rapid Identification and Typing of Listeria Species by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. AEM 74, 5402–5407. https://doi.org/10.1128/AEM.02689-07

Batzoglou, S., 2000. Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. Genome Research 10, 950–958. https://doi.org/10.1101/gr.10.7.950

Bern, M.W., Kil, Y.J., 2011. Two-Dimensional Target Decoy Strategy for Shotgun Proteomics. J. Proteome Res. 10, 5296–5301. https://doi.org/10.1021/pr200780j

Berrazeg, M., Diene, S.M., Drissi, M., Kempf, M., Richet, H., Landraud, L., Rolain, J.-M., 2013. Biotyping of multidrug-resistant Klebsiella pneumoniae clinical isolates from France and Algeria using MALDI-TOF MS. PLoS One 8, e61428. https://doi.org/10.1371/journal.pone.0061428

Bittremieux, W., Meysman, P., Noble, W.S., Laukens, K., 2018. Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing. J. Proteome Res. 17, 3463–3474. https://doi.org/10.1021/acs.jproteome.8b00359

Bizzini, A., Greub, G., 2010. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry, a revolution in clinical microbial identification. Clin Microbiol Infect 16, 1614–1619. https://doi.org/10.1111/j.1469-0691.2010.03311.x

Blakeley, P., Overton, I.M., Hubbard, S.J., 2012. Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies. J. Proteome Res. 11, 5221–5234. https://doi.org/10.1021/pr300411q

Boggs, S.R., Cazares, L.H., Drake, R., 2012. Characterization of a Staphylococcus aureus USA300 protein signature using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. J Med Microbiol 61, 640–644. https://doi.org/10.1099/jmm.0.037978-0

Boitard, S., Schlotterer, C., Nolte, V., Pandey, R.V., Futschik, A., 2012. Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples. Molecular Biology and Evolution 29, 2177–2186. https://doi.org/10.1093/molbev/mss090

Boitard, S., Kofler, R., Françoise, P., Robelin, D., Schlötterer, C., Futschik, A., 2013. Pool-hmm: a Python program for estimating the allele frequency spectrum and detecting selective sweeps from next generation sequencing of pooled samples. Mol Ecol Resour 13, 337–340. https://doi.org/10.1111/1755-0998.12063

Brownlee, G.G., Sanger, F., 1967. Nucleotide sequences from the low molecular weight ribosomal RNA of Escherichia coli. Journal of Molecular Biology 23, 337-IN9. https://doi.org/10.1016/S0022-2836(67)80109-8

Burke, M.K., Dunham, J.P., Shahrestani, P., Thornton, K.R., Rose, M.R., Long, A.D., 2010. Genome-wide analysis of a long-term evolution experiment with Drosophila. Nature 467, 587–590. https://doi.org/10.1038/nature09352

Calvo, S.E., Tucker, E.J., Compton, A.G., Kirby, D.M., Crawford, G., Burtt, N.P., Rivas, M., Guiducci, C., Bruno, D.L., Goldberger, O.A., Redman, M.C., Wiltshire, E., Wilson, C.J., Altshuler, D., Gabriel, S.B., Daly, M.J., Thorburn, D.R., Mootha, V.K., 2010. High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. Nat Genet 42, 851–858. https://doi.org/10.1038/ng.659

Campeau, A., Mills, R.H., Blanchette, M., Bajc, K., Malfavon, M., Munji, R.N., Deng, L., Hancock, B., Patras, K.A., Olson, J., Nizet, V., Daneman, R., Doran, K., Gonzalez, D.J., 2020. Multidimensional Proteome Profiling of Blood-Brain Barrier Perturbation by Group B *Streptococcus*. mSystems 5, e00368-20, /msystems/5/4/msys.00368-20.atom. https://doi.org/10.1128/mSystems.00368-20

Cao, C.-C., Sun, X., 2015. Accurate estimation of haplotype frequency from pooled sequencing data and cost-effective identification of rare haplotype carriers by overlapping pool sequencing. Bioinformatics 31, 515–522. https://doi.org/10.1093/bioinformatics/btu670

Cao, R., Shi, Y., Chen, S., Ma, Y., Chen, J., Yang, J., Chen, G., Shi, T., 2017. dbSAP: single amino-acid polymorphism database for protein variation detection. Nucleic Acids Res 45, D827–D832. https://doi.org/10.1093/nar/gkw1096

Carapetis, J.R., Steer, A.C., Mulholland, E.K., Weber, M., 2005. The global burden of group A streptococcal diseases. The Lancet Infectious Diseases 5, 685–694. https://doi.org/10.1016/S1473-3099(05)70267-X

Cerqueira, F.R., Graber, A., Schwikowski, B., Baumgartner, C., 2010. MUDE: A New Approach for Optimizing Sensitivity in the Target-Decoy Search Strategy for Large-Scale Peptide/Protein Identification. J. Proteome Res. 9, 2265–2277. https://doi.org/10.1021/pr901023v

Chait, B.T., 2006. CHEMISTRY: Mass Spectrometry: Bottom-Up or Top-Down? Science 314, 65–66. https://doi.org/10.1126/science.1133987

Chan, C.H.S., Octavia, S., Sintchenko, V., Lan, R., 2016. SnpFilt: A pipeline for reference-free assembly-based identification of SNPs in bacterial genomes. Computational Biology and Chemistry 65, 178–184. https://doi.org/10.1016/j.compbiolchem.2016.09.004

Chen, Q., Sun, F., 2013. A unified approach for allele frequency estimation, SNP detection and association studies based on pooled sequencing data using EM algorithms. BMC Genomics 14, S1. https://doi.org/10.1186/1471-2164-14-S1-S1

Chen, X., Listman, J.B., Slack, F.J., Gelernter, J., Zhao, H., 2012. Biases and Errors on Allele Frequency Estimation and Disease Association Tests of Next-Generation Sequencing of Pooled Samples. Genetic epidemiology 36, 549–560.

Chen, Y., Chen, W., Cobb, M.H., Zhao, Y., 2009. PTMap—A sequence alignment software for unrestricted, accurate, and full-spectrum identification of post-translational modification sites. PNAS 106, 761–766. https://doi.org/10.1073/pnas.0811739106

Cheng, C., White, B.J., Kamdem, C., Mockaitis, K., Costantini, C., Hahn, M.W., Besansky, N.J., 2012. Ecological Genomics of *Anopheles gambiae* Along a Latitudinal Cline: A Population-Resequencing Approach. Genetics 190, 1417–1432. https://doi.org/10.1534/genetics.111.137794

Chick, J.M., Kolippakkam, D., Nusinow, D.P., Zhai, B., Rad, R., Huttlin, E.L., Gygi, S.P., 2015. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol 33, 743–749. https://doi.org/10.1038/nbt.3267

Choi, H., Nesvizhskii, A.I., 2008. Semisupervised Model-Based Validation of Peptide Identifications in Mass Spectrometry-Based Proteomics. J. Proteome Res. 7, 254–265. https://doi.org/10.1021/pr070542g

Christner, M., Trusch, M., Rohde, H., Kwiatkowski, M., Schlüter, H., Wolters, M., Aepfelbacher, M., Hentschke, M., 2014. Rapid MALDI-TOF mass spectrometry strain typing during a large outbreak of Shiga-Toxigenic Escherichia coli. PLoS One 9, e101924. https://doi.org/10.1371/journal.pone.0101924

Clark, C.G., Kruczkiewicz, P., Guan, C., McCorrister, S.J., Chong, P., Wylie, J., van Caeseele, P., Tabor, H.A., Snarr, P., Gilmour, M.W., Taboada, E.N., Westmacott, G.R., 2013. Evaluation of MALDI-TOF mass spectroscopy methods for determination of Escherichia coli pathotypes. J Microbiol Methods 94, 180–191. https://doi.org/10.1016/j.mimet.2013.06.020

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., Bayley, H., 2009. Continuous base identification for single-molecule nanopore DNA sequencing. Nature Nanotech 4, 265–270. https://doi.org/10.1038/nnano.2009.12

Colaert, N., Degroeve, S., Helsens, K., Martens, L., 2011. Analysis of the Resolution Limitations of Peptide Identification Algorithms. J. Proteome Res. 10, 5555–5561. https://doi.org/10.1021/pr200913a

Colinge, J., Masselot, A., Giron, M., Dessingy, T., Magnin, J., 2003. OLAV: Towards high-throughput tandem mass spectrometry data identification. Proteomics 3, 1454–1463. https://doi.org/10.1002/pmic.200300485

Colman, G., Tanna, A., Efstratiou, A., Gaworzewska, E.T., 1993. The serotypes of Streptococcus pyogenes present in Britain during 1980-1990 and their association with disease. Journal of Medical Microbiology 39, 165–178. https://doi.org/10.1099/00222615-39-3-165

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., Mann, M., 2011. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. J. Proteome Res. 10, 1794–1805. https://doi.org/10.1021/pr101065j

Craig, R., Beavis, R.C., 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20, 1466–1467. https://doi.org/10.1093/bioinformatics/bth092

Creasy, D.M., Cottrell, J.S., 2002. Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics 2, 1426–1434. https://doi.org/10.1002/1615-9861(200210)2:10<1426::AID-PROT1426>3.0.CO;2-5

Crick, F.H., 1958. On protein synthesis. Symp Soc Exp Biol 12, 138–163.

Croxatto, A., Prod'hom, G., Greub, G., 2012. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. FEMS Microbiol Rev 36, 380–407. https://doi.org/10.1111/j.1574-6976.2011.00298.x

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M.R., Armean, I.M., Bennett, R., Bhai, J., Billis, K., Boddu, S., Cummins, C., Davidson, C., Dodiya, K.J., Gall, A., Girón, C.G., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O.G., Janacek, S.H., Juettemann, T., Kay, M., Laird, M.R., Lavidas, I., Liu, Z., Loveland, J.E., Marugán, J.C., Maurel, T., McMahon, A.C., Moore, B., Morales, J., Mudge, J.M., Nuhn, M., Ogeh, D., Parker, A., Parton, A., Patricio, M., Abdul Salam, A.I., Schmitt, B.M., Schuilenburg, H., Sheppard, D., Sparrow, H., Stapleton, E., Szuba, M., Taylor, K., Threadgold, G., Thormann, A., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Frankish, A., Hunt, S.E., Kostadima, M., Langridge, N., Martin, F.J., Muffato, M., Perry, E., Ruffier, M., Staines, D.M., Trevanion, S.J., Aken, B.L., Yates, A.D., Zerbino, D.R., Flicek, P., 2019. Ensembl 2019. Nucleic Acids Research 47, D745–D751. https://doi.org/10.1093/nar/gky1113

Day-Williams, A.G., McLay, K., Drury, E., Edkins, S., Coffey, A.J., Palotie, A., Zeggini, E., 2011. An Evaluation of Different Target Enrichment Methods

in Pooled Sequencing Designs for Complex Disease Association Studies. PLoS ONE 6, e26279. https://doi.org/10.1371/journal.pone.0026279

Demirev, P.A., Feldman, A.B., Kowalski, P., Lin, J.S., 2005. Top-Down Proteomics for Rapid Identification of Intact Microorganisms. Anal. Chem. 77, 7455–7461. https://doi.org/10.1021/ac051419g

Devabhaktuni, A., Lin, S., Zhang, L., Swaminathan, K., Gonzalez, C.G., Olsson, N., Pearlman, S.M., Rawson, K., Elias, J.E., 2019. TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. Nat Biotechnol 37, 469–479. https://doi.org/10.1038/s41587-019-0067-5

Diament, B.J., Noble, W.S., 2011. Faster SEQUEST Searching for Peptide Identification from Tandem Mass Spectra. J. Proteome Res. 10, 3871–3879. https://doi.org/10.1021/pr101196n

DiMaggio, Jr., P.A., Floudas, C.A., Lu, B., Yates, III, J.R., 2008. A Hybrid Method for Peptide Identification Using Integer Linear Optimization, Local Database Search, and Quadrupole Time-of-Flight or OrbiTrap Tandem Mass Spectrometry. J. Proteome Res. 7, 1584–1593. https://doi.org/10.1021/pr700577z

Ding, Y., Choi, H., Nesvizhskii, A.I., 2008. Adaptive discriminant function analysis and reranking of MS/MS database search results for improved peptide identification in shotgun proteomics. J Proteome Res 7, 4878–4889. https://doi.org/10.1021/pr800484x

Domon, B., 2006. Mass Spectrometry and Protein Analysis. Science 312, 212–217. https://doi.org/10.1126/science.1124619

Dorfer, V., Pichler, P., Stranzl, T., Stadlmann, J., Taus, T., Winkler, S., Mechtler, K., 2014. MS Amanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra. J. Proteome Res. 13, 3679–3684. https://doi.org/10.1021/pr500202e

Druley, T.E., Vallania, F.L.M., Wegner, D.J., Varley, K.E., Knowles, O.L., Bonds, J.A., Robison, S.W., Doniger, S.W., Hamvas, A., Cole, F.S., Fay, J.C., Mitra, R.D., 2009. Quantification of rare allelic variants from pooled genomic DNA. Nat Methods 6, 263–265. https://doi.org/10.1038/nmeth.1307

Edman, P., Högfeldt, E., Sillén, L.G., Kinell, P.-O., 1950. Method for Determination of the Amino Acid Sequence in Peptides. Acta Chemica Scandinavica 4, 283–293. https://doi.org/10.3891/acta.chem.scand.04-0283

Edwards, N.J., 2007. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. Mol Syst Biol 3, 102. https://doi.org/10.1038/msb4100142

Elias, J.E., Gibbons, F.D., King, O.D., Roth, F.P., Gygi, S.P., 2004. Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotechnol 22, 214–219. https://doi.org/10.1038/nbt930

Elias, J.E., Gygi, S.P., 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4, 207–214. https://doi.org/10.1038/nmeth1019

Eng, J.K., McCormack, A.L., Yates, J.R., 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5, 976–989. https://doi.org/10.1016/1044-0305(94)80016-2

Eng, J.K., Jahan, T.A., Hoopmann, M.R., 2013. Comet: An open-source MS/MS sequence database search tool. Proteomics 13, 22–24. https://doi.org/10.1002/pmic.201200439

Evans, V.C., Barker, G., Heesom, K.J., Fan, J., Bessant, C., Matthews, D.A., 2012. De novo derivation of proteomes from transcriptomes for transcript and protein identification. Nat Methods 9, 1207–1211. https://doi.org/10.1038/nmeth.2227

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-Calling of Automated Sequencer Traces Using *Phred.* I. Accuracy Assessment. Genome Research 8, 175–185. https://doi.org/10.1101/gr.8.3.175

Ewing, B., Green, P., 1998. Base-Calling of Automated Sequencer Traces Using *Phred.* II. Error Probabilities. Genome Research 8, 186–194. https://doi.org/10.1101/gr.8.3.186

Ferretti, L., Ramos-Onsins, S.E., Pérez-Enciso, M., 2013. Population genomics from pool sequencing. Mol Ecol 22, 5561–5576. https://doi.org/10.1111/mec.12522

Franco-Duarte, R., Černáková, L., Kadam, S., Kaushik, K.S., Salehi, B., Bevilacqua, A., Corbo, M.R., Antolak, H., Dybka-Stępień, K., Leszczewicz, M., Relison Tintino, S., Alexandrino de Souza, V.C., Sharifi-Rad, J., Coutinho, H.D.M., Martins, N., Rodrigues, C.F., 2019. Advances in Chemical and Biological Methods to Identify Microorganisms-From Past to Present. Microorganisms 7. https://doi.org/10.3390/microorganisms7050130

Frank, A., Pevzner, P., 2005. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. Anal. Chem. 77, 964–973. https://doi.org/10.1021/ac048788h

Futschik, A., Schlötterer, C., 2010. The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. Genetics 186, 207–218. https://doi.org/10.1534/genetics.110.114397

Garrison, E., Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 [q-bio].

Gaskell, S.J., 1997. Electrospray: Principles and practice. J. Mass Spectrom 677–688.

Gasteiger, E., Jung, E., Bairoch, A., 2001. SWISS-PROT: connecting biomolecular knowledge via a protein database. Curr Issues Mol Biol 3, 47–55.

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C., Estoup, A., 2013. Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. Mol Ecol 22, 3766–3779. https://doi.org/10.1111/mec.12360

Gaworzewska, E., Colman, G., 1988. Changes in the pattern of infection caused by Streptococcus pyogenes. Epidemiol Infect 100, 257–269. https://doi.org/10.1017/s095026880006739x

Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., Bryant, S.H., 2004. Open Mass Spectrometry Search Algorithm. J. Proteome Res. 3, 958–964. https://doi.org/10.1021/pr0499491

Giani, A.M., Gallo, G.R., Gianfranceschi, L., Formenti, G., 2020. Long walk to genomics: History and current approaches to genome sequencing and assembly. Computational and Structural Biotechnology Journal 18, 9–19. https://doi.org/10.1016/j.csbj.2019.11.002

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G., 1996. Life with 6000 Genes. Science 274, 546–567. https://doi.org/10.1126/science.274.5287.546

Granholm, V., Käll, L., 2011. Quality assessments of peptide-spectrum matches in shotgun proteomics. Proteomics 11, 1086–1093. https://doi.org/10.1002/pmic.201000432

Guo, Y., Samuels, D.C., Li, J., Clark, T., Li, C.-I., Shyr, Y., 2013. Evaluation of Allele Frequency Estimation Using Pooled Sequencing Data Simulation. The Scientific World Journal 2013, 1–9. https://doi.org/10.1155/2013/895496

Hajirasouliha, I., Hormozdiari, F., Sahinalp, S.C., Birol, I., 2008. Optimal pooling for genome re-sequencing with ultra-high-throughput short-read technologies. Bioinformatics 24, i32–i40. https://doi.org/10.1093/bioinformatics/btn173

Harakalova, M., Nijman, I.J., Medic, J., Mokry, M., Renkens, I., Blankensteijn, J.D., Kloosterman, W., Baas, A.F., Cuppen, E., 2011. Genomic DNA Pooling Strategy for Next-Generation Sequencing-Based Rare Variant Discovery in Abdominal Aortic Aneurysm Regions of Interest—Challenges and Limitations. J. of Cardiovasc. Trans. Res. 4, 271–280. https://doi.org/10.1007/s12265-011-9263-5

Haslam, D.B., St. Geme, J.W., 2018. Groups C and G Streptococci, in: Principles and Practice of Pediatric Infectious Diseases. Elsevier, pp. 736-737.e1. https://doi.org/10.1016/B978-0-323-40181-4.00122-5

Havilio, M., Haddad, Y., Smilansky, Z., 2003. Intensity-Based Statistical Scorer for Tandem Mass Spectrometry. Anal. Chem. 75, 435–444. https://doi.org/10.1021/ac0258913

Heath, P.T., Jardine, L.A., 2014. Neonatal infections: group B streptococcus. BMJ Clin Evid 2014.

Heather, J.M., Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hickman, M.E., Rench, M.A., Ferrieri, P., Baker, C.J., 1999. Changing epidemiology of group B streptococcal colonization. Pediatrics 104, 203–209. https://doi.org/10.1542/peds.104.2.203

Higgs, R.E., Knierman, M.D., Bonner Freeman, A., Gelbert, L.M., Patil, S.T., Hale, J.E., 2007. Estimating the Statistical Significance of Peptide Identifications from Shotgun Proteomics Experiments. J. Proteome Res. 6, 1758–1767. https://doi.org/10.1021/pr0605320

Holt, K.E., Teo, Y.Y., Li, H., Nair, S., Dougan, G., Wain, J., Parkhill, J., 2009. Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA. Bioinformatics 25, 2074–2075. https://doi.org/10.1093/bioinformatics/btp344

Hu, Z., Scott, H.S., Qin, G., Zheng, G., Chu, X., Xie, L., Adelson, D.L., Oftedal, B.E., Venugopal, P., Babic, M., Hahn, C.N., Zhang, B., Wang, X., Li, N., Wei, C., 2015. Revealing Missing Human Protein Isoforms Based on Ab Initio Prediction, RNA-seq and Proteomics. Sci Rep 5, 10940. https://doi.org/10.1038/srep10940

Huang, T., Wang, J., Yu, W., He, Z., 2012. Protein inference: a review. Briefings in Bioinformatics 13, 586–614. https://doi.org/10.1093/bib/bbs004

Hughes, C., Ma, B., Lajoie, G.A., 2010. De novo sequencing methods in proteomics. Methods Mol Biol 604, 105–121. https://doi.org/10.1007/978-1-60761-444-9_8

Hughes, M.J.G., Moore, J.C., Lane, J.D., Wilson, R., Pribul, P.K., Younes, Z.N., Dobson, R.J., Everest, P., Reason, A.J., Redfern, J.M., Greer, F.M., Paxton, T., Panico, M., Morris, H.R., Feldman, R.G., Santangelo, J.D., 2002. Identification of Major Outer Surface Proteins of Streptococcus agalactiae. IAI 70, 1254–1259. https://doi.org/10.1128/IAI.70.3.1254-1259.2002

Ingman, M., Gyllensten, U., 2009. SNP frequency estimation using massively parallel sequencing of pooled DNA. Eur J Hum Genet 17, 383–386. https://doi.org/10.1038/ejhg.2008.182

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921. https://doi.org/10.1038/35057062

Johnson, D.R., Stevens, D.L., Kaplan, E.L., 1992. Epidemiologic Analysis of Group A Streptococcal Serotypes Associated with Severe Systemic Infections, Rheumatic Fever, or Uncomplicated Pharyngitis. Journal of Infectious Diseases 166, 374–382. https://doi.org/10.1093/infdis/166.2.374

Johri, A.K., Margarit, I., Broenstrup, M., Brettoni, C., Hua, L., Gygi, S.P., Telford, J.L., Grandi, G., Paoletti, L.C., 2007. Transcriptional and proteomic profiles of group B Streptococcus type V reveal potential adherence proteins associated with high-level invasion. Infect Immun 75, 1473–1483. https://doi.org/10.1128/IAI.00638-06

Joo, J.W.J., Na, S., Baek, J.-H., Lee, C., Paek, E., 2010. Target-Decoy with Mass Binning: A Simple and Effective Validation Method for Shotgun Proteomics Using High Resolution Mass Spectrometry. J. Proteome Res. 9, 1150–1156. https://doi.org/10.1021/pr9006377

Jorgenson, J.W., Lukacs, K.D., 1981. Free-zone electrophoresis in glass capillaries. Clinical Chemistry 27, 1551–1553. https://doi.org/10.1093/clinchem/27.9.1551

Kaartokallio, T., Wang, J., Heinonen, S., Kajantie, E., Kivinen, K., Pouta, A., Gerdhem, P., Jiao, H., Kere, J., Laivuori, H., 2016. Exome sequencing in pooled DNA samples to identify maternal pre-eclampsia risk variants. Sci Rep 6, 29085. https://doi.org/10.1038/srep29085

Käll, L., Storey, J.D., MacCoss, M.J., Noble, W.S., 2008a. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. J. Proteome Res. 7, 29–34. https://doi.org/10.1021/pr700600n

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

Kasper, T.J., Melera, M., Gozel, P., Brownlee, R.G., 1988. Separation and detection of DNA by capillary electrophoresis. Journal of Chromatography A 458, 303–312. https://doi.org/10.1016/S0021-9673(00)90574-0

Keller, A., Nesvizhskii, A.I., Kolker, E., Aebersold, R., 2002. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. Anal. Chem. 74, 5383–5392. https://doi.org/10.1021/ac025747h

Kessner, D., Turner, T.L., Novembre, J., 2013. Maximum Likelihood Estimation of Frequencies of Known Haplotypes from Pooled Sequence Data. Molecular Biology and Evolution 30, 1145–1158. https://doi.org/10.1093/molbev/mst016

Kieran, E., Matheson, M., Mann, A.G., Efstratiou, A.A., Butler, K., Gorman, W., 1998. Group B streptococcus (GBS) colonisation among expectant Irish mothers. Ir Med J 91, 21–22.

Kim, S., Pevzner, P.A., 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5, 5277. https://doi.org/10.1038/ncomms6277

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Ding, L., 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25, 2283–2285. https://doi.org/10.1093/bioinformatics/btp373

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R.V., Nolte, V., Futschik, A., Kosiol, C., Schlötterer, C., 2011. PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. PLoS ONE 6, e15925. https://doi.org/10.1371/journal.pone.0015925

Kofler, R., Pandey, R.V., Schlotterer, C., 2011. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27, 3435–3436. https://doi.org/10.1093/bioinformatics/btr589

Kolaczkowski, B., Kern, A.D., Holloway, A.K., Begun, D.J., 2011. Genomic Differentiation Between Temperate and Tropical Australian Populations of *Drosophila melanogaster*. Genetics 187, 245–260. https://doi.org/10.1534/genetics.110.123059

Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D., Nesvizhskii, A.I., 2017. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. Nat Methods 14, 513–520. https://doi.org/10.1038/nmeth.4256

Krasnov, G.S., Dmitriev, A.A., Kudryavtseva, A.V., Shargunov, A.V., Karpov, D.S., Uroshlev, L.A., Melnikova, N.V., Blinov, V.M., Poverennaya, E.V., Archakov, A.I., Lisitsa, A.V., Ponomarenko, E.A., 2015a. PPLine: An

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

Krasnov, G.S., Dmitriev, A.A., Kudryavtseva, A.V., Shargunov, A.V., Karpov, D.S., Uroshlev, L.A., Melnikova, N.V., Blinov, V.M., Poverennaya, E.V., Archakov, A.I., Lisitsa, A.V., Ponomarenko, E.A., 2015b. PPLine: An 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

Krishna, R., Xia, D., Sanderson, S., Shanmugasundram, A., Vermont, S., Bernal, A., Daniel-Naguib, G., Ghali, F., Brunk, B.P., Roos, D.S., Wastling, J.M., Jones, A.R., 2015. A large-scale proteogenomics study of apicomplexan pathogens-*Toxoplasma gondii* and *Neospora caninum*. Proteomics 15, 2618–2628. https://doi.org/10.1002/pmic.201400553

Kuhns, M., Zautner, A.E., Rabsch, W., Zimmermann, O., Weig, M., Bader, O., Groß, U., 2012. Rapid discrimination of Salmonella enterica serovar Typhi from other serovars by MALDI-TOF mass spectrometry. PLoS One 7, e40004. https://doi.org/10.1371/journal.pone.0040004

Kumar, D., Yadav, A.K., Jia, X., Mulvenna, J., Dash, D., 2016. Integrated Transcriptomic-Proteomic Analysis Using a Proteogenomic Workflow Refines Rat Genome Annotation. Mol Cell Proteomics 15, 329–339. https://doi.org/10.1074/mcp.M114.047126

Kurland, S., Wheat, C.W., Paz Celorio Mancera, M., Kutschera, V.E., Hill, J., Andersson, A., Rubin, C., Andersson, L., Ryman, N., Laikre, L., 2019. Exploring a Pool-seq-only approach for gaining population genomic insights in nonmodel species. Ecol Evol 9, 11448–11463. https://doi.org/10.1002/ece3.5646

Lamichhaney, S., Barrio, A.M., Rafati, N., Sundstrom, G., Rubin, C.-J., Gilbert, E.R., Berglund, J., Wetterbom, A., Laikre, L., Webster, M.T., Grabherr, M., Ryman, N., Andersson, L., 2012. Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proceedings of the National Academy of Sciences 109, 19345–19350. https://doi.org/10.1073/pnas.1216128109

Lancefield, R.C., 1933. A SEROLOGICAL DIFFERENTIATION OF HUMAN AND OTHER GROUPS OF HEMOLYTIC STREPTOCOCCI. J Exp Med 57, 571–595. https://doi.org/10.1084/jem.57.4.571

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Lartigue, M.-F., Kostrzewa, M., Salloum, M., Haguenoer, E., Héry-Arnaud, G., Domelier, A.-S., Stumpf, S., Quentin, R., 2011. Rapid detection of "highly virulent" Group B Streptococcus ST-17 and emerging ST-1 clones by MALDI-TOF mass spectrometry. J Microbiol Methods 86, 262–265. https://doi.org/10.1016/j.mimet.2011.05.017

Lasch, P., Beyer, W., Nattermann, H., Stämmler, M., Siegbrecht, E., Grunow, R., Naumann, D., 2009. Identification of Bacillus anthracis by using matrix-assisted laser desorption ionization-time of flight mass spectrometry and artificial neural networks. Appl Environ Microbiol 75, 7229–7242. https://doi.org/10.1128/AEM.00857-09

Levene, M.J., 2003. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. Science 299, 682–686. https://doi.org/10.1126/science.1079700

Li, D., Liu, C.-M., Luo, R., Sadakane, K., Lam, T.-W., 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., others, 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.

Li, J., Duncan, D.T., Zhang, B., 2010. CanProVar: a human cancer proteome variation database. Hum. Mutat. 31, 219–228. https://doi.org/10.1002/humu.21176

Long, Q., Jeffares, D.C., Zhang, Q., Ye, K., Nizhynska, V., Ning, Z., Tyler-Smith, C., Nordborg, M., 2011. PoolHap: Inferring Haplotype Frequencies from Pooled Samples by Next Generation Sequencing. PLoS ONE 6, e15292. https://doi.org/10.1371/journal.pone.0015292

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.,

Vavre, F., Sagot, M.-F., Lacroix, V., 2016. SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. Nucleic Acids Res gkw655. https://doi.org/10.1093/nar/gkw655

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G., 2003. PEAKS: powerful software for peptidede novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. 17, 2337–2342. https://doi.org/10.1002/rcm.1196

Ma, B., Johnson, R., 2012. *De Novo* Sequencing and Homology Searching. Mol Cell Proteomics 11, O111.014902. https://doi.org/10.1074/mcp.O111.014902

Maglott, D., 2004. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research 33, D54–D58. https://doi.org/10.1093/nar/gki031

Malmström, J., Karlsson, C., Nordenfelt, P., Ossola, R., Weisser, H., Quandt, A., Hansson, K., Aebersold, R., Malmström, L., Björck, L., 2012. *Streptococcus pyogenes* in Human Plasma: ADAPTIVE MECHANISMS ANALYZED BY MASS SPECTROMETRY-BASED PROTEOMICS. J. Biol. Chem. 287, 1415–1425. https://doi.org/10.1074/jbc.M111.267674

Mann, M., Højrup, P., Roepstorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol. Mass Spectrom. 22, 338–345. https://doi.org/10.1002/bms.1200220605

Mann, M., Wilm, M., 1994. Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. Anal. Chem. 66, 4390–4399. https://doi.org/10.1021/ac00096a002

Mardis, E.R., 2009. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. Genome Med 1, 40. https://doi.org/10.1186/gm40

Margraf, R.L., Durtschi, J.D., Dames, S., Pattison, D.C., Stephens, J.E., Voelkerding, K.V., 2011. Variant identification in multi-sample pools by illumina genome analyzer sequencing. J Biomol Tech 22, 74–84.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-

B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380. https://doi.org/10.1038/nature03959

Maxam, A.M., Gilbert, W., 1977. A new method for sequencing DNA. Proceedings of the National Academy of Sciences 74, 560–564. https://doi.org/10.1073/pnas.74.2.560

May, D.H., Timmins-Schiffman, E., Mikan, M.P., Harvey, H.R., Borenstein, E., Nunn, B.L., Noble, W.S., 2016. An Alignment-Free "Metapeptide" Strategy for Metaproteomic Characterization of Microbiome Samples Using Shotgun Metagenomic Sequencing. J. Proteome Res. 15, 2697–2705. https://doi.org/10.1021/acs.jproteome.6b00239

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20, 1297–1303. https://doi.org/10.1101/gr.107524.110

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., Zhang, Z., Ranade, S.S., Dimalanta, E.T., Hyland, F.C., Sokolsky, T.D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C.L., Li, B., Kotler, L., Stuart, J.R., Malek, J.A., Manning, J.M., Antipova, A.A., Perez, D.S., Moore, M.P., Hayashibara, K.C., Lyons, M.R., Beaudoin, R.E., Coleman, B.E., Laptewicz, M.W., Sannicandro, A.E., Rhodes, M.D., Gottimukkala, R.K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J.M., Eichler, E.E., Reese, M.G., De La Vega, F.M., Blanchard, A.P., 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Research 19, 1527–1541. https://doi.org/10.1101/gr.091868.109

McLafferty, F.W., Breuker, K., Jin, M., Han, X., Infusini, G., Jiang, H., Kong, X., Begley, T.P., 2007. Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics: Top-down MS of proteins. FEBS Journal 274, 6256–6268. https://doi.org/10.1111/j.1742-4658.2007.06147.x

Menschaert, G., Vandekerckhove, T.T.M., Baggerman, G., Landuyt, B., Sweedler, J.V., Schoofs, L., Luyten, W., Van Criekinge, W., 2010. A Hybrid, *de Novo* Based, Genome-Wide Database Search Approach Applied to the Sea Urchin Neuropeptidome. J. Proteome Res. 9, 990–996. https://doi.org/10.1021/pr900885k

Metzker, M.L., 2010. Sequencing technologies — the next generation. Nat Rev Genet 11, 31–46. https://doi.org/10.1038/nrg2626

Micheletti, S.J., Narum, S.R., 2018. Utility of pooled sequencing for association mapping in nonmodel organisms. Molecular Ecology Resources 18, 825–837. https://doi.org/10.1111/1755-0998.12784

Mitchell Wells, J., McLuckey, S.A., 2005. Collision-Induced Dissociation (CID) of Peptides and Proteins, in: Methods in Enzymology. Elsevier, pp. 148–185. https://doi.org/10.1016/S0076-6879(05)02005-7

Momozawa, Y., Mni, M., Nakamura, K., Coppieters, W., Almer, S., Amininejad, L., Cleynen, I., Colombel, J.-F., de Rijk, P., Dewit, O., Finkel, Y., Gassull, M.A., Goossens, D., Laukens, D., Lémann, M., Libioulle, C., O'Morain, C., Reenaers, C., Rutgeerts, P., Tysk, C., Zelenika, D., Lathrop, M., Del-Favero, J., Hugot, J.-P., de Vos, M., Franchimont, D., Vermeire, S., Louis, E., Georges, M., 2011. Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. Nat Genet 43, 43–47. https://doi.org/10.1038/ng.733

Moura, H., Woolfitt, A.R., Carvalho, M.G., Pavlopoulos, A., Teixeira, L.M., Satten, G.A., Barr, J.R., 2008. MALDI-TOF mass spectrometry as a tool for differentiation of invasive and noninvasive Streptococcus pyogenes isolates. FEMS Immunol Med Microbiol 53, 333–342. https://doi.org/10.1111/j.1574-695X.2008.00428.x

Mullen, M.P., Creevey, C.J., Berry, D.P., McCabe, M.S., Magee, D.A., Howard, D.J., Killeen, A.P., Park, S.D., McGettigan, P.A., Lucy, M.C., Machugh, D.E., Waters, S.M., 2012. Polymorphism discovery and allele frequency estimation using high-throughput DNA sequencing of target-enriched pooled DNA samples. BMC Genomics 13, 16. https://doi.org/10.1186/1471-2164-13-16

Muth, T., Kolmeder, C.A., Salojärvi, J., Keskitalo, S., Varjosalo, M., Verdam, F.J., Rensen, S.S., Reichl, U., de Vos, W.M., Rapp, E., Martens, L., 2015. Navigating through metaproteomics data: A logbook of database searching. Proteomics 15, 3439–3453. https://doi.org/10.1002/pmic.201400560

Myers, E.W., 2000. A Whole-Genome Assembly of Drosophila. Science 287, 2196–2204. https://doi.org/10.1126/science.287.5461.2196

Na, S., Bandeira, N., Paek, E., 2012. Fast multi-blind modification search through tandem mass spectrometry. Mol Cell Proteomics 11, M111.010199. https://doi.org/10.1074/mcp.M111.010199

Nagaraj, S.H., Waddell, N., Madugundu, A.K., Wood, S., Jones, A., Mandyam, R.A., Nones, K., Pearson, J.V., Grimmond, S.M., 2015. PGTools: A Software Suite for Proteogenomic Data Analysis and Visualization. J. Proteome Res. 14, 2255–2266. https://doi.org/10.1021/acs.jproteome.5b00029

Nakamura, T., Hasegawa, T., Torii, K., Hasegawa, Y., Shimokata, K., Ohta, M., 2004. Two-dimensional gel electrophoresis analysis of the abundance of virulent exoproteins of group A streptococcus caused by environmental changes. Archives of Microbiology 181, 74–81. https://doi.org/10.1007/s00203-003-0632-6

National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC), United Kingdom Inflammatory Bowel Disease Genetics Consortium, International Inflammatory Bowel Disease Genetics Consortium, Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., Fennell, T., Kirby, A., Latiano, A., Goyette, P., Green, T., Halfvarson, J., Haritunians, T., Korn, J.M., Kuruvilla, F., Lagacé, C., Neale, B., Lo, K.S., Schumm, P., Törkvist, L., Dubinsky, M.C., Brant, S.R., Silverberg, M.S., Duerr, R.H., Altshuler, D., Gabriel, S., Lettre, G., Franke, A., D'Amato, M., McGovern, D.P.B., Cho, J.H., Rioux, J.D., Xavier, R.J., Daly, M.J., 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet 43, 1066–1073. https://doi.org/10.1038/ng.952

Navarro, P., Vázquez, J., 2009. A Refined Method To Calculate False Discovery Rates for Peptide Identification Using Decoy Databases. J. Proteome Res. 8, 1792–1796. https://doi.org/10.1021/pr800362h

Neethiraj, R., Hornett, E.A., Hill, J.A., Wheat, C.W., 2017. Investigating the genomic basis of discrete phenotypes using a Pool-Seq-only approach: New insights into the genetics underlying colour variation in diverse taxa. Mol Ecol 26, 4990–5002. https://doi.org/10.1111/mec.14205

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

1 Diabetes. Science 324, 387–389. https://doi.org/10.1126/science.1167728

Nesvizhskii, A.I., 2010. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. Journal of Proteomics 73, 2092–2123. https://doi.org/10.1016/j.jprot.2010.08.009

Nesvizhskii, A.I., 2014. Proteogenomics: concepts, applications and computational strategies. Nat Methods 11, 1114–1125. https://doi.org/10.1038/nmeth.3144

Noble, W., Serang, O., 2012. A review of statistical methods for protein identification using tandem mass spectrometry. Statistics and Its Interface 5, 3–20. https://doi.org/10.4310/SII.2012.v5.n1.a2

Noble, W.S., 2015. Mass spectrometrists should search only for peptides they care about. Nat Methods 12, 605–608. https://doi.org/10.1038/nmeth.3450

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

Noguchi, H., Taniguchi, T., Itoh, T., 2008. MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. DNA Research 15, 387–396. https://doi.org/10.1093/dnares/dsn027

Nordenfelt, P., Waldemarson, S., Linder, A., Mörgelin, M., Karlsson, C., Malmström, J., Björck, L., 2012. Antibody orientation at bacterial surfaces is related to invasive infection. Journal of Experimental Medicine 209, 2367–2381. https://doi.org/10.1084/jem.20120325

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference

sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44, D733–D745. https://doi.org/10.1093/nar/gkv1189

Out, A.A., van Minderhout, I.J.H.M., Goeman, J.J., Ariyurek, Y., Ossowski, S., Schneeberger, K., Weigel, D., van Galen, M., Taschner, P.E.M., Tops, C.M.J., Breuning, M.H., van Ommen, G.-J.B., den Dunnen, J.T., Devilee, P., Hes, F.J., 2009. Deep sequencing to reveal new variants in pooled DNA samples. Hum. Mutat. 30, 1703–1712. https://doi.org/10.1002/humu.21122

Padmanabhan, R., Jay, E., Wu, R., 1974. Chemical Synthesis of a Primer and Its Use in the Sequence Analysis of the Lysozyme Gene of Bacteriophage T4. Proceedings of the National Academy of Sciences 71, 2510–2514. https://doi.org/10.1073/pnas.71.6.2510

Papasergi, S., Galbo, R., Lanza-Cariccio, V., Domina, M., Signorino, G., Biondo, C., Pernice, I., Poyart, C., Trieu-Cuot, P., Teti, G., Beninati, C., 2013. Analysis of the Streptococcus agalactiae exoproteome. Journal of Proteomics 89, 154–164. https://doi.org/10.1016/j.jprot.2013.06.003

Pappin, D.J.C., Hojrup, P., Bleasby, A.J., 1993. Rapid identification of proteins by peptide-mass fingerprinting. Current Biology 3, 327–332. https://doi.org/10.1016/0960-9822(93)90195-T

Patterson, N., Gabriel, S., 2009. Combinatorics and next-generation sequencing. Nat Biotechnol 27, 826–827. https://doi.org/10.1038/nbt0909-826

Pérez-Enciso, M., Ferretti, L., 2010. Massive parallel sequencing in animal genetics: wherefroms and wheretos: Massive parallel sequencing in animal genetics. Animal Genetics 41, 561–569. https://doi.org/10.1111/j.1365-2052.2010.02057.x

Peterlongo, P., Schnel, N., Pisanti, N., Sagot, M.-F., Lacroix, V., 2010. Identifying SNPs without a Reference Genome by Comparing Raw Reads, in: Chavez, E., Lonardi, S. (Eds.), String Processing and Information Retrieval, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 147–158. https://doi.org/10.1007/978-3-642-16321-0_14

Peterson, E.S., McCue, L.A., Schrimpe-Rutledge, A.C., Jensen, J.L., Walker, H., Kobold, M.A., Webb, S.R., Payne, S.H., Ansong, C.K., Adkins, J.N., Cannon, W.R., Webb-Robertson, B.-J.M., 2012. VESPA: software to facilitate genomic annotation of prokaryotic organisms through integration

of proteomic and transcriptomic data. BMC Genomics 13, 131. https://doi.org/10.1186/1471-2164-13-131

Prakash, O., Verma, M., Sharma, P., Kumar, M., Kumari, K., Singh, A., Kumari, H., Jit, S., Gupta, S.K., Khanna, M., Lal, R., 2007. Polyphasic approach of bacterial classification - An overview of recent advances. Indian J Microbiol 47, 98–108. https://doi.org/10.1007/s12088-007-0022-x

Price, T.S., Lucitt, M.B., Wu, W., Austin, D.J., Pizarro, A., Yocum, A.K., Blair, I.A., FitzGerald, G.A., Grosser, T., 2007. EBP, a Program for Protein Identification Using Multiple Tandem Mass Spectrometry Datasets. Mol Cell Proteomics 6, 527–536. https://doi.org/10.1074/mcp.T600049-MCP200

Prober, J., Trainor, G., Dam, R., Hobbs, F., Robertson, C., Zagursky, R., Cocuzza, A., Jensen, M., Baumeister, K., 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238, 336–341. https://doi.org/10.1126/science.2443975

Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S., Pérez-Enciso, M., 2012. SNP calling by sequencing pooled samples. BMC Bioinformatics 13, 239. https://doi.org/10.1186/1471-2105-13-239

Ratan, A., Zhang, Y., Hayes, V.M., Schuster, S.C., Miller, W., 2010. Calling SNPs without a reference sequence. BMC Bioinformatics 11, 130. https://doi.org/10.1186/1471-2105-11-130

Reddy, T.B.K., Thomas, A.D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E.A., Kyrpides, N.C., 2015. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res. 43, D1099-1106. https://doi.org/10.1093/nar/gku950

Reil, M., Erhard, M., Kuijper, E.J., Kist, M., Zaiss, H., Witte, W., Gruber, H., Borgmann, S., 2011. Recognition of Clostridium difficile PCR-ribotypes 001, 027 and 126/078 using an extended MALDI-TOF MS system. Eur J Clin Microbiol Infect Dis 30, 1431–1436. https://doi.org/10.1007/s10096-011-1238-6

Reiter, L., Claassen, M., Schrimpf, S.P., Jovanovic, M., Schmidt, A., Buhmann, J.M., Hengartner, M.O., Aebersold, R., 2009. Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem

Mass Spectrometry. Mol Cell Proteomics 8, 2405–2417. https://doi.org/10.1074/mcp.M900317-MCP200

Rellstab, C., Zoller, S., Tedder, A., Gugerli, F., Fischer, M.C., 2013. Validation of SNP Allele Frequencies Determined by Pooled Next-Generation Sequencing in Natural Populations of a Non-Model Plant Species. PLoS ONE 8, e80422. https://doi.org/10.1371/journal.pone.0080422

Renard, B.Y., Xu, B., Kirchner, M., Zickmann, F., Winter, D., Korten, S., Brattig, N.W., Tzur, A., Hamprecht, F.A., Steen, H., 2012. Overcoming species boundaries in peptide identification with Bayesian information criterion-driven error-tolerant peptide search (BICEPS). Mol Cell Proteomics 11, M111.014167. https://doi.org/10.1074/mcp.M111.014167

Resing, K.A., Meyer-Arendt, K., Mendoza, A.M., Aveline-Wolf, L.D., Jonscher, K.R., Pierce, K.G., Old, W.M., Cheung, H.T., Russell, S., Wattawa, J.L., Goehle, G.R., Knight, R.D., Ahn, N.G., 2004. Improving Reproducibility and Sensitivity in Identifying Human Proteins by Shotgun Proteomics. Anal. Chem. 76, 3556–3568. https://doi.org/10.1021/ac035229m

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., Nyrén, P., 1996. Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. Analytical Biochemistry 242, 84–89. https://doi.org/10.1006/abio.1996.0432

Rubin, G.M., 2000. Comparative Genomics of the Eukaryotes. Science 287, 2204–2215. https://doi.org/10.1126/science.287.5461.2204

Ryu, S., Han, J., Norden-Krichmar, T.M., Schork, N.J., Suh, Y., 2018. Effective discovery of rare variants by pooled target capture sequencing: A comparative analysis with individually indexed target capture sequencing. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 809, 24–31. https://doi.org/10.1016/j.mrfmmm.2018.03.007

Sanger, F., Tuppy, H., 1951. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem J 49, 481–490. https://doi.org/10.1042/bj0490481

Sanger, F., Thompson, E.O.P., 1953. The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. Biochem J 53, 366–374. https://doi.org/10.1042/bj0530366

Sanger, F., Donelson, J.E., Coulson, A.R., Kossel, H., Fischer, D., 1973. Use of DNA Polymerase I Primed by a Synthetic Oligonucleotide to Determine a Nucleotide Sequence in Phage f1 DNA. Proceedings of the National Academy of Sciences 70, 1209–1213. https://doi.org/10.1073/pnas.70.4.1209

Sanger, F., Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology 94, 441–448. https://doi.org/10.1016/0022-2836(75)90213-2

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M., Smith, M., 1977. Nucleotide sequence of bacteriophage φX174 DNA. Nature 265, 687–695. https://doi.org/10.1038/265687a0

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74, 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Schiebenhoefer, H., Van Den Bossche, T., Fuchs, S., Renard, B.Y., Muth, T., Martens, L., 2019. Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. Expert Review of Proteomics 16, 375–390. https://doi.org/10.1080/14789450.2019.1609944

Schlötterer, C., Tobler, R., Kofler, R., Nolte, V., 2014. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. Nat Rev Genet 15, 749–763. https://doi.org/10.1038/nrg3803

Schwarze, K., Buchanan, J., Fermont, J.M., Dreau, H., Tilley, M.W., Taylor, J.M., Antoniou, P., Knight, S.J.L., Camps, C., Pentony, M.M., Kvikstad, E.M., Harris, S., Popitsch, N., Pagnamenta, A.T., Schuh, A., Taylor, J.C., Wordsworth, S., 2020. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. Genet Med 22, 85–94. https://doi.org/10.1038/s41436-019-0618-7

Searle, B.C., Dasari, S., Turner, M., Reddy, A.P., Choi, D., Wilmarth, P.A., McCormack, A.L., David, L.L., Nagalla, S.R., 2004. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. Anal Chem 76, 2220–2230. https://doi.org/10.1021/ac035258x

Seng, P., Drancourt, M., Gouriet, F., La Scola, B., Fournier, P., Rolain, J.M., Raoult, D., 2009. Ongoing Revolution in Bacteriology: Routine Identification of Bacteria by Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry. CLIN INFECT DIS 49, 543–551. https://doi.org/10.1086/600885

Seng, P., Rolain, J.-M., Fournier, P.E., La Scola, B., Drancourt, M., Raoult, D., 2010. MALDI-TOF-mass spectrometry applications in clinical microbiology. Future Microbiol 5, 1733–1754. https://doi.org/10.2217/fmb.10.127

Shanmugam, A.K., Nesvizhskii, A.I., 2015. Effective Leveraging of Targeted Search Spaces for Improving Peptide Identification in Tandem Mass Spectrometry Based Proteomics. J. Proteome Res. 14, 5169–5178. https://doi.org/10.1021/acs.jproteome.5b00504

Shaw, S.H., Carrasquillo, M.M., Kashuk, C., Puffenberger, E.G., Chakravarti, A., 1998. Allele Frequency Distributions in Pooled DNA Samples: Applications to Mapping Complex Disease Genes. Genome Res. 8, 111–123. https://doi.org/10.1101/gr.8.2.111

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., Waterston, R.H., 2017. DNA sequencing at 40: past, present and future. Nature 550, 345–353. https://doi.org/10.1038/nature24286

Sheynkman, G.M., Johnson, J.E., Jagtap, P.D., Shortreed, M.R., Onsongo, G., Frey, B.L., Griffin, T.J., Smith, L.M., 2014. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. BMC Genomics 15, 703. https://doi.org/10.1186/1471-2164-15-703

Shilov, I.V., Seymour, S.L., Patel, A.A., Loboda, A., Tang, W.H., Keating, S.P., Hunter, C.L., Nuwaysir, L.M., Schaeffer, D.A., 2007. The Paragon Algorithm, a Next Generation Search Engine That Uses Sequence Temperature Values and Feature Probabilities to Identify Peptides from Tandem Mass Spectra. Mol Cell Proteomics 6, 1638–1655. https://doi.org/10.1074/mcp.T600050-MCP200

Skinner, O.S., Kelleher, N.L., 2015. Illuminating the dark matter of shotgun proteomics. Nat Biotechnol 33, 717–718. https://doi.org/10.1038/nbt.3287

Smith, A.M., Heisler, L.E., St.Onge, R.P., Farias-Hesson, E., Wallace, I.M., Bodeau, J., Harris, A.N., Perry, K.M., Giaever, G., Pourmand, N., Nislow, C., 2010.

Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Research 38, e142–e142. https://doi.org/10.1093/nar/gkq368

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B.H., Hood, L.E., 1986. Fluorescence detection in automated DNA sequence analysis. Nature 321, 674–679. https://doi.org/10.1038/321674a0

Spivak, M., Weston, J., Bottou, L., Käll, L., Noble, W.S., 2009. Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets. J Proteome Res 8, 3737–3745. https://doi.org/10.1021/pr801109k

Starkweather, R., Barnes, C.S., Wyckoff, G.J., Keightley, J.A., 2007. Virtual Polymorphism: Finding Divergent Peptide Matches in Mass Spectrometry Data. Anal. Chem. 79, 5030–5039. https://doi.org/10.1021/ac0703496

Sticker, A., Martens, L., Clement, L., 2017. Mass spectrometrists should search for all peptides, but assess only the ones they care about. Nat Methods 14, 643–644. https://doi.org/10.1038/nmeth.4338

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., UniProt Consortium, 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31, 926–932. https://doi.org/10.1093/bioinformatics/btu739

Tabb, D.L., Saraf, A., Yates, J.R., 2003. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem 75, 6415–6421. https://doi.org/10.1021/ac0347462

Tabb, D.L., Fernando, C.G., Chambers, M.C., 2007. MyriMatch: Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis. J. Proteome Res. 6, 654–661. https://doi.org/10.1021/pr0604054

Tanner, S., Pevzner, P.A., Bafna, V., 2006. Unrestrictive identification of post-translational modifications through peptide mass spectrometry. Nat Protoc 1, 67–72. https://doi.org/10.1038/nprot.2006.10

Tanner, S., Shen, Z., Ng, J., Florea, L., Guigo, R., Briggs, S.P., Bafna, V., 2007. Improving gene annotation using peptide mass spectrometry. Genome Research 17, 231–239. https://doi.org/10.1101/gr.5646507

Taylor, J.A., Johnson, R.S., 1997. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 11, 1067–1075. https://doi.org/10.1002/(SICI)1097-0231(19970615)11:9<1067::AID-RCM953>3.0.CO;2-L

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., DeBoy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J.B., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., Fraser, C.M., 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." Proceedings of the National Academy of Sciences 102, 13950–13955. https://doi.org/10.1073/pnas.0506758102

The C.elegans Sequencing Consortium, 1998. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science 282, 2012–2018. https://doi.org/10.1126/science.282.5396.2012

Thomas, H., Shevchenko, A., 2008. Simplified validation of borderline hits of database searches. Proteomics 8, 4173–4177. https://doi.org/10.1002/pmic.200800250

Turner, T.L., Bourne, E.C., Von Wettberg, E.J., Hu, T.T., Nuzhdin, S.V., 2010. Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. Nat Genet 42, 260–263. https://doi.org/10.1038/ng.515

Turner, T.L., Stewart, A.D., Fields, A.T., Rice, W.R., Tarone, A.M., 2011. Population-Based Resequencing of Experimentally Evolved Populations Reveals the Genetic Basis of Body Size Variation in Drosophila melanogaster. PLoS Genet 7, e1001336. https://doi.org/10.1371/journal.pgen.1001336

Vallania, F.L.M., Druley, T.E., Ramos, E., Wang, J., Borecki, I., Province, M., Mitra, R.D., 2010. High-throughput discovery of rare insertions and deletions in

large cohorts. Genome Research 20, 1711–1718. https://doi.org/10.1101/gr.109157.110

Van Tassell, C.P., Smith, T.P.L., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., Sonstegard, T.S., 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods 5, 247–252. https://doi.org/10.1038/nmeth.1185

Voelkerding, K.V., Dames, S.A., Durtschi, J.D., 2009. Next-Generation Sequencing: From Basic Research to Diagnostics. Clinical Chemistry 55, 641–658. https://doi.org/10.1373/clinchem.2008.112789

Wang, J., Skoog, T., Einarsdottir, E., Kaartokallio, T., Laivuori, H., Grauers, A., Gerdhem, P., Hytönen, M., Lohi, H., Kere, J., Jiao, H., 2016. Investigation of rare and low-frequency variants using high-throughput sequencing with pooled DNA samples. Sci Rep 6, 33256. https://doi.org/10.1038/srep33256

Wang, W., Yin, X., Soo Pyon, Y., Hayes, M., Li, J., 2013. Rare variant discovery and calling by sequencing pooled samples with overlaps. Bioinformatics 29, 29–38. https://doi.org/10.1093/bioinformatics/bts645

Wang, X., Zhang, B., 2013. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics 29, 3235–3237. https://doi.org/10.1093/bioinformatics/btt543

Wang, X., Li, Y., Wu, Z., Wang, H., Tan, H., Peng, J., 2014. JUMP: A Tag-based Database Search Tool for Peptide Identification with High Sensitivity and Accuracy. Mol Cell Proteomics 13, 3663–3673. https://doi.org/10.1074/mcp.O114.039586

Watson, J.D., Crick, F.H.C., 1953. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature 171, 737–738. https://doi.org/10.1038/171737a0

Wei, Z., Wang, W., Hu, P., Lyon, G.J., Hakonarson, H., 2011. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Research 39, e132–e132. https://doi.org/10.1093/nar/gkr599

Wen, Y.-T., Tsou, C.-C., Kuo, H.-T., Wang, J.-S., Wu, J.-J., Liao, P.-C., 2011. Differential Secretomics of S *treptococcus pyogenes* Reveals a Novel Peroxide Regulator (PerR)-regulated Extracellular Virulence Factor Mitogen Factor3 (MF3). Mol Cell Proteomics 10, M110.007013. https://doi.org/10.1074/mcp.M110.007013

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology 37, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wilk, L., Happonen, L., Malmström, J., Herwald, H., 2018. Comprehensive Mass Spectrometric Survey of *Streptococcus pyogenes* Subcellular Proteomes. J. Proteome Res. 17, 600–617. https://doi.org/10.1021/acs.jproteome.7b00701

Williamson, Y.M., Moura, H., Woolfitt, A.R., Pirkle, J.L., Barr, J.R., Carvalho, M.D.G., Ades, E.P., Carlone, G.M., Sampson, J.S., 2008. Differentiation of Streptococcus pneumoniae conjunctivitis outbreak isolates by matrix-assisted laser desorption ionization-time of flight mass spectrometry. Appl Environ Microbiol 74, 5891–5897. https://doi.org/10.1128/AEM.00791-08

Wingo, T.S., Duong, D.M., Zhou, M., Dammer, E.B., Wu, H., Cutler, D.J., Lah, J.J., Levey, A.I., Seyfried, N.T., 2017. Integrating Next-Generation Genomic Sequencing and Mass Spectrometry To Estimate Allele-Specific Protein Abundance in Human Brain. J Proteome Res 16, 3336–3347. https://doi.org/10.1021/acs.jproteome.7b00324

Wu, R., Kaiser, A.D., 1968. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. Journal of Molecular Biology 35, 523–537. https://doi.org/10.1016/S0022-2836(68)80012-9

Yamashita, M., Fenn, J.B., 1984. Electrospray ion source. Another variation on the free-jet theme. J. Phys. Chem. 88, 4451–4459. https://doi.org/10.1021/j150664a002

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

Yates, J.R., Speicher, S., Griffin, P.R., Hunkapiller, T., 1993. Peptide Mass Maps: A Highly Informative Approach to Protein Identification. Analytical Biochemistry 214, 397–408. https://doi.org/10.1006/abio.1993.1514

Yates, J.R., Eng, J.K., McCormack, A.L., Schieltz, David., 1995. Method to Correlate Tandem Mass Spectra of Modified Peptides to Amino Acid Sequences in the Protein Database. Anal. Chem. 67, 1426–1436. https://doi.org/10.1021/ac00104a020

Yates, J.R., Morgan, S.F., Gatlin, C.L., Griffin, P.R., Eng, J.K., 1998. Method To Compare Collision-Induced Dissociation Spectra of Peptides: Potential for Library Searching and Subtractive Analysis. Anal. Chem. 70, 3557–3565. https://doi.org/10.1021/ac980122y

Yonghua Han, Bin Ma, Kaizhong Zhang, 2004. SPIDER: software for protein identification from sequence tags with de novo sequencing error, in: Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004. Presented at the Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004., IEEE, Stanford, CA, USA, pp. 198–207. https://doi.org/10.1109/CSB.2004.1332434

Yu, W., Taylor, J.A., Davis, M.T., Bonilla, L.E., Lee, K.A., Auger, P.L., Farnsworth, C.C., Welcher, A.A., Patterson, S.D., 2010. Maximizing the sensitivity and reliability of peptide identification in large-scale proteomic experiments by harnessing multiple search engines. Proteomics 10, 1172–1189. https://doi.org/10.1002/pmic.200900074

Zhang, H., Scholl, R., Browse, J., Somerville, C., 1988. Double stranded DNA sequencing as a choice for DNA sequencing. Nucleic Acids Research 16, 1220–1220. https://doi.org/10.1093/nar/16.3.1220

Zhang, M., McDonald, F.M., Sturrock, S.S., Charnock, S.J., Humphery-Smith, I., Black, G.W., 2007. Group A streptococcus cell-associated pathogenic proteins as revealed by growth in hyaluronic acid-enriched media. Proteomics 7, 1379–1390. https://doi.org/10.1002/pmic.200600578

Zhang, N., Aebersold, R., Schwikowski, B., 2002. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass

spectral data. Proteomics 2, 1406–1412. https://doi.org/10.1002/1615-9861(200210)2:10<1406::AID-PROT1406>3.0.CO;2-9

Zhou, B., 2012. An empirical Bayes mixture model for SNP detection in pooled sequencing data. Bioinformatics 28, 2569–2575. https://doi.org/10.1093/bioinformatics/bts501

Zhu, Y., Bergland, A.O., González, J., Petrov, D.A., 2012. Empirical Validation of Pooled Whole Genome Population Re-Sequencing in Drosophila melanogaster. PLoS ONE 7, e41901. https://doi.org/10.1371/journal.pone.0041901

Zhu, Y., Orre, L.M., Johansson, H.J., Huss, M., Boekel, J., Vesterlund, M., Fernandez-Woodbridge, A., Branca, R.M.M., Lehtiö, J., 2018. Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nat Commun 9, 903. https://doi.org/10.1038/s41467-018-03311-y

Zickmann, F., Renard, B.Y., 2015. MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. Bioinformatics 31, i106–i115. https://doi.org/10.1093/bioinformatics/btv236

ISBN 978-951-51-7168-9 (PRINT)
ISBN 978-951-51-7169-6 (ONLINE)
ISSN 2342-3161 (PRINT)
ISSN 2342-317X (ONLINE)
http://ethesis.helsinki.fi

HELSINKI 2021