Systems Biology

A systems level approach to characterize the phenotypic effects of SNPs in closely related bacterial genomes

Joseph Mullen*

School of Computing Science, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom

ABSTRACT

Motivation: Mutations such as SNPs account for much variation present between closely related bacterial strains. A plethora of sequencing data is available for closely related strains of bacteria. With this data comes the need to analyse the effects of these mutations on phenotype is. SNP databases exist, such as dbSNP which store SNP data in flat-file format thus limiting analysis. To truly understand the influence these SNPs play on the phenotype of an organism a systems level approach is beneficial, allowing for the visualisation of these sequence variations within large biological networks. Through the integration of numerous open source databases semantically rich datasets can be developed for an organism of interest. One such example is BacillusOndex, produced for the gram positive model organism B. subtilis. BacillusOndex can be visualised as a graphical knowledgebase using the open source visualisation platform, Ondex. SNPs may be filtered into this dataset using CatSNP, an Ondex plug-in; essentially creating a new ‘SNP’ concept within the graph. Analysis of sequence variations can then be performed at systems level, following SNP from genotype to phenotype. Evaluation of this systems level approach was achieved by applying it to two application projects; B. subtilis 168 (against Marburg 6051) and a Protease Deficient strain of B. subtilis 168 Trpc2 (against 168).

Results: Methodology proved successful in identifying the phenotypic effects of known SNPs within 168. Cumulative phenotypic effects of SNPs within ppsC, degQ and sipF were proposed, yet contradicted previous knowledge. Approach also allowed for a greater understanding of the Protease Deficient strain. A nsSNP within the regulatory region of a putative protease, spoIIGA, identified potential overexpression of the SpoIIGA protein as means of maintaining a healthy organism in the absence of 10 proteases.

Availability: Ondex platform is available from www.ondex.org, with BacillusOndex knowledgebase, (along with the Plug-In allowing for its manual creation) downloadable from www.bacillondex.org CatSNP plug-in is available on request from Dr. Katherine James at Newcastle University and all sub-graphs produced during this project available for download from http://www.students.ncl.ac.uk/j.mullen.

Contact: j.mullen@ncl.ac.uk

1 INTRODUCTION

Mutations such as single nucleotide polymorphisms (SNPs) are often responsible for the subtle differences in phenotype between closely related bacterial strains as well as being associated with disease in humans. As more and more closely related bacterial genomes are being sequenced the need to analyze the effects of SNPs becomes ever more necessary. A systems level approach to this analysis allows for the visualization of a SNP within a biological network. Through identifying interactions we can see how the presence of a SNP may affect a large biological system, allowing for a greater understanding of how the phenotype may be altered (Weile et al, 2011).

Strategies have previously been described for the visualization of SNPs and their subsequent analysis in the dynamics of a network model, in cell signaling pathways during human disease state (Bauer-Mehren et al, 2010). Bayesian network models have also been applied to a pediatric acute lymphoblastic leukemia dataset, including SNPs, as means of phenotypic prediction (Chang & McGeachie, 2011). In order to analyze these SNPs at a systems level successful integration of data from numerous resources is a pre-requisite.

This need for data integration is a consequence of the plethora of heterogeneous datasets freely available online; a direct result of high throughput technologies in the post-genomic era (Weile et al, 2011). Successful data integration allows for the production of semantically rich datasets, which can be visualized and explored interactively via systems level biology approaches. Tools such as Cytoscape (Shannon et al, 2003) allow for integrated datasets to be visualized at network level, but are constrained. There is no ability to apply common semantics to relations. The open source visualization platform Ondex allows for the integration of unprecedented amounts of data and can be easily manipulated via the production of datasets and plug-ins. CatSNP for example (developed by Dr Katherine James at Newcastle University) is a filter plug-in and allows for the integration (filtration) of SNP data into a dataset. BacillusOndex, is an integrated semantically rich B. subtilis knowledgebase (Mirisli et al, 2011) and used concomitantly with CatSNP allows for SNP visualization and analysis at a systems level.

B. subtilis was chosen for this project due the fact it has many parent wild type as well as daughter strains, varying greatly in phenotype. Many genomes for the organism have been sequenced and, along with numerous data bases regarding the organism, freely available online. B. subtilis 168 (from herein referred to as 168) is a mutated TrpA form of its parent wild type, Marburg. Sequence variations (v 168) during domestication (prior to mutagenesis) identified SNPs in two genes required for swarming: swrA and

*To whom correspondence should be addressed.
sfP (Kearns et al, 2004). Post mutagenesis analysis within the trp region of 168 and Marburg identified SNPs in the trpC and aroH genes (Ziegler et al, 2008). SNPs of this strain have yet to be analyzed using a systems level approach. The work presented here is an evaluation of the use of systems levels approaches in characterizing the phenotypic effects of closely related bacterial strains and the functionality provided by the CatSNP plug in. Two application projects were used:

1) Reads were provided from a Marburg ATCC 6051 sequencing project. 6051 is the parent wild type of 168. Reads were assembled using 168 as a reference and a systems level SNP analysis was performed on 168 (against Marburg ATCC 6051). This analysis had been previously carried out by Dr. Katherine James who had identified 100 SNPs in 168 yet SNP type had not been identified and ‘cleaning’ methodology had not been applied.

2) Reads from a Protease Deficient strain of 168 were also provided by Colin Harwood and were assembled again using 168 as a reference. This strain was analysed for large deletions and insertions as well as having systems level SNP analysis performed.

2 BACKGROUND
2.1 The need for a systems view of SNPs

Strain differentiation may be the result of mutations. Examples of mutations include SNPs and single INDELs (insertions/deletions). These mutations represent the most abundant form of genetic variation amongst closely related microbial species. SNP is a term used to denote point nucleotide changes, that is, when a single nucleotide within the genome is altered potentially conferring change at translation. Occurring in structural genes SNPs can have a dramatic effect on the biology of whole organisms, from bacteria and viruses to mammals (Weissman et al, 2003).

Although SNPs usually occur in non-coding (intergenic) regions, distribution within the genome is not homogenous. Occurring in the coding DNA sequence (CDS) synonymous/silent SNPs (sSNPs) do not change protein sequence as genetic code is preserved. It is argued they sSNPs may still subtly alter structural and functional properties of a protein (Komar, 2007). Non-synonymous SNPs (nsSNPs) alter amino sequence and thus impact the protein level. nsSNPs may cause a missense (altered amino acid produced) or nonsense mutation (premature stop codon). In order to accurately decide whether a SNP is a sSNP or a nsSNP a codon usage table for the organism in question must be referred to (Bauer-Mehren et al, 2009). Expression SNPs (eSNPs) occur in regions that modulate gene expression (i.e. promoters, introns, splice sites, transcription factor binding sites). eSNPs can drastically alter the expression of a protein encoding gene and as a consequence lead to disease (Bauer-Mehren et al, 2009).

Single INDELs are a common and functionally important sequence variation type. When occurring in coding regions INDELs cause frame shift mutations. They can be used as genetic markers in natural populations; especially in phylogenetic studies (Albers et al, 2010). Herein, like most articles, the term SNP will include SNPs and single INDELs (as defined).

SNPs are conserved over evolution and play a part in many human diseases, becoming a prevalent point of interest in pharmacogenetics. In bacterial (and viral) isolates they can be used to aid in phylogenetic characterisation. This allows for strains to be ‘tracked’ during an epidemic (in bacterial pathogens a variety of SNPs have been identified that belong to the category of pathogenicity-enhancing mutations) as well as forensic investigations. Of great interest during this evaluation study is the ability of SNPs to provide means of identifying a correlation between genotype and phenotype in closely related strains (Gardner & Slezak, 2010).

Many tools/methods exist for the identification of SNPs. NGS platforms concomitantly with improved assembly software have flooded databases with sequence data (for a review on NGS platforms, assembly techniques and SNP/INDEL detection please see (Magi et al, 2010). Whole genome sequences of many similar organisms are widely available, making it easier to compare and produce genome-wide SNP analysis for closely related strains. During SNP prediction factors that must be considered include the error rate of sequencing platform used, the probability of bad mapping and coverage of the variation (Magi et al, 2010). Many tools are available for the identification of SNPs and are often included in assembly software (during reference assemblies). These tools tend to follow 3 basic steps; (1) identification of homologous regions amongst genomes being compared (2) elimination of paralogous sequences / regions minimizes false positive SNP identification (3) generation of multiple sequence alignment for detection of SNPs, thereafter ‘cleaning’ methods must be employed to ensure only SNPs with the highest quality are analysed (Magi et al, 2010). Data retrieved may then be deposited in databases such as dbSNP (Sherry et al, 2001) provided by NCBI.

Once a SNP ‘report’ has been created, in order to fully understand the functional consequences of the SNPs, analysis at a systems level is highly advantageous. This approach allows for the visualization of complex interactions of SNPs within biological systems, as opposed to the DNA level or indeed the single molecule level. Systems level analysis allows for the pathways in which such proteins play a valuable role to be visualized and associated ‘concepts’ to be mapped (Bauer-Mehren et al, 2009). Through identifying the location of SNPs used concomitantly with data from other heterogeneous sources (databases and scientific literature), one can derive biological knowledge, studying the association between sequence variation and inheritable phenotype via network visualization (Koehler et al, 2005) (Kohler et al, 2006).

2.2 Ondex: In silico network analysis

One way of visualising biological relevance of SNPs in differing strains is via systems level analysis using biological networks. These show integrated views of data from the endless repositories of bioinformatics data combined with SNP data. Data integration is
not as simple as one may first think, with the biological data in particular, showing a wide variety of both syntactic and semantic heterogeneity; this must be addressed with the need for a common semantics (Cockell et al., 2010). Biomart\(^1\) is a project that has attempted to provide a solution to this issue through the transformation of disparate data schema into a unified Mart format. This may then be accessed through a query interface (Heider et al., 2009). Graph-based analysis software that allow data to be imported from several sources and visualized as a network, with enrichment of additional information, include Cytoscape (Shannon et al., 2003), MAPMAN (Thimm et al., 2004) and Osprey (Breitkreutz et al., 2003). These tools are, however limited in their ability to support for automated linking and mapping of data from many heterogeneous data sources as well as in their visualisation capabilities (Kohler et al., 2006).

Ondex is an open source systems level platform that addresses the problem of large-scale database integration as well as graph visualization (Kohler et al., 2006). The Ondex framework consists of:

1. **Data Input:** from databases using parsers or in standards such as BioPax, SBML, XGMML or the more reliable OXL (Taubert, 2007)
2. **Data integration:** achieved via their conversion to a common graph-based data structure using integrated ontologies, in two fully automated steps (1) import of databases and ontologies and (2) the alignment of data from different sources, which produces a graph-based data structure containing internally represented data that is semantically consistent via the usage of the same ontology-based data structure (Kohler, 2004),
   - (i) **Sequence Analysis:** may be applied to genes and proteins that are integrated into Ondex especially useful for high quality functional annotations of genes and genomes (Kohler et al., 2006),
   - (ii) **Text mining:** for the retrieval data stored in scientific literature,
3. **Data Analysis:** OVTK from the GUI, where exporters are used to convert data between formats,
   - (i) **Graph Analysis:** used to exploit the data integrated in an Ondex system where data is represented as ‘concepts’ (nodes) and ‘relations’ (edges).

Graph analysis allows for the representation of the integrated heterogeneous data as a network of interconnected nodes, where the nodes (or concepts) and edges (or relations) are annotated with semantically rich metadata (Cockell et al., 2010). Ultimately this allows multiple sources of information to be brought together meaningfully in the same graph. Within an Ondex graph each concept has a Concept Class and each relation has a Relation Type. This means that complex biological relationships can be visualised within the network (see Fig 1). Ondex also allows for both concepts and relations to have attributes, accessions and names (Weil et al, 2011) (Cockell et al., 2010).

The Ondex Integrator, a utility that is accessible from Ondex’s graphical user interface (GUI) processes and integrates the data sources in a sequence of concatenated steps; these steps, when called, take the form of .xml ‘workflows’. Workflows allow for unique pipelines to be developed, including components such as parsers, mappers, filters and plugins, depending on user preference. Ondex may be customized by the development of plug-ins, using an open API and datasets produced and parsed directly into the platform (Taubert et al., 2007).

### 2.3 BacillOndex; a *B. subtilis* knowledgebase

Semantically rich datasets to be visualized in Ondex have been created for organisms including *S. cerevisaea* (Weile et al., 2011) and *A. Thalilana*. These datasets integrate data from a large repository of databases and apply common semantics; resulting in large detailed networks of the organisms and their respective concepts/ biological entities. BacillOndex is a semantically rich integrated knowledgebase, developed for *B. subtilis* (168) (Misirli et al., 2011). It was created via the integration of data from BacilliScope, DBTBS, STRING, KEGG, KEGG EXPRESSION and GO and represents these data in a systems level visualisation of the *B. subtilis* genome when visualised using the Ondex platform. Annotations were taken directly from BacilliScope, which contains data achieved during the second *B. subtilis* sequencing project (Antoine et al., 2007), with intergenic regions being annotated using dbTBS\(^2\) (version 5) (Sierro et al., 2008). A plugin was also developed that allows for the manual mining of these databases, as the databases vast information swells further (Misirli et al., 2011).

---

\(^1\) http://www.biomart.org/

\(^2\) http://dbtbs.hgc.jp/

---

Fig 1. An overview of the Concept Classes and Relation Types contained within the graph produced by the BacillOndex dataset; Showing how certain concept classes may be related to others.

---

2.4 CatSNP; a SNP filtering plugin

The Cellular Analysis Tool for SNPs (CatSNP) is a SNP filter plug-in for developed to integrate further data into the Bacil-
lOndex dataset. It is currently run in the integrator as part of a workflow. Once the location of SNPs has been found, CatSNP takes these locations in flat file format and filters them into the BacillOndex dataset. The resulting graph includes the location of the SNP within the genome and all entities from the BacillOndex dataset associated to it. In essence CatSNP creates a new set of SNP ‘concepts’ for Ondex and displays all concepts and relations associated with these in a graphical output. CatSNP workflow contains an OXL parser (takes in BacillOndex integrated dataset), a SNP parser (takes in a SNP.txt file containing SNP locations as an input), a filter (filters SNPs into BacillOndex dataset) and an OXL exporter (resulting graph is exported to any location on the local machine).

Using CatSNP and the BacillOndex dataset we can study SNPs that occur within 168 or any other strain where 168 is the reference.

### 2.5 B. subtilis

*B. subtilis*, the rod-shaped endospore-forming aerobe, is a model organism for gram positive bacteria (Stein, 2005) (Barbe *et al.*, 2009). Generally recognized as safe it is used in food supplies and is an A+T rich Fermicute commonly found in soil environments, where it competes with numerous other organisms for nutrients it is also found in water sources and in association with plants (Kunst *et al.*, 1997) (Westers *et al.*, 2004). Several hundred wild type strains have been collected and sequenced with well documented biochemical and physiological data (Stein, 2005), yet gaps in the understanding of genomic heritage and provenance of many widely used strains still exists (Ziegler *et al.*, 2008). One of the original parent wild-types is the Marburg strain, which remains elusive (Ziegler *et al.*, 2008). The Marburg strain was distributed upon its’ identification and is thought to be the parent of two separate strains; the NCIB 3610 strain and the ATCC 6051 strain. Marburg is the parent of the highly transformable, tryptophan auxotroph strain 168, isolated after the Marburg strain, was mutagenized with X-rays (Burkholder & Giles, 1947) and is particularly prevalent amongst laboratories due to its impressive levels of growth. Although reads from Marburg 6051 were provided for the first application process SNPs identified will be analyzed at a systems level to see how they affect the phenotype of 168.

168 has one circular chromosome of 4.2 Mb containing 4354 genes (4176 protein coding genes and 178 RNAs) and is believed to have 192 indispensable, as well as 79 essential genes (Kobayashi *et al.*, 2003). The first sequenced genome of the strain was completed in 1997 as a paradigm for gram-positive bacteria (Kunst *et al.*, 1997) by a consortium that made up the *B. subtilis* Genome Sequencing Project. The project was coordinated largely by Frank Kunst with the French geneticist Antoine Danchin leading the bioinformatics. It was one of two European based sequencing consortia, with the other focusing on *S. cerevisiae* (Barbe *et al.*, 2009). The project ultimately had contributions from 25 European laboratories seven Japanese, two US and one Korean laboratory (Moszer *et al.*, 1996) (Kunst *et al.*, 1997). As a result the first reference database of the bacteria, SubList1, was created. This was not updated until 2002 (Moszer *et al.*, 2002) and at present is on data release R16.1 (April 26, 2001). Bacterial strains evolve fast in laboratory environments and re-sequencing of the genome took place in 2009, again influenced by Kunst & Danchin. This project utilized faster, more accurate sequencing techniques and a recently developed high level annotation platform, MaGe (Barbe *et al.*, 2009) (Vallenet *et al.*, 2006). Sequencing culminated in an updated version of Sublist (GenoList) (Lechat *et al.*, 2008) which holds a genome information for >700 prokaryotic organisms) and the production of BaciluScope (providing an overview of the annotation platform MaGe), used for annotations in the BacillOndex dataset (Misirli *et al.*, 2011). Interestingly a total of 407 previously unknown ‘y’ genes were given biologically significant names following this project (Barbe *et al.*, 2009).

Versions of the DNA sequence representing the whole 168 chromosome are available as a GenBank files with the EMBL/GenBank/DDBJ accession number AL009126. The transcriptome has also been recently re-mapped by a large consortium (Nicolas *et al.*, 2012).

Many phenotypic differences exist between 168 and its wild-type parent, with knowledge of its genetic heritage remaining, to this day, incomplete (Ziegler *et al.*, 2008). Used in basic laboratory research, where focus lies on its well documented sporulation and its use as a reference model for cell differentiation (Barbe *et al.*, 2009), as well as industry. Of industrially relevance is the ability of the organism to:

1. Produce numerous homologous enzymes at gram per litre concentrations (≥ 20 g l⁻¹) (Kunst *et al.*, 1997). Such as alkaline proteases as washing agents or amylases for the starch industry (Westers *et al.*, 2004).

2. Produce, in commercially attractive levels; small metabolites such as antibiotics.

---

2. [http://genodb.pasteur.fr/cgi-bin/WebObjects/GenoList](http://genodb.pasteur.fr/cgi-bin/WebObjects/GenoList)
(3) Be used as a host for heterologous protein production.

It is for these reasons that B. subtilis has been made the focus of extensive metabolic, genetic and protein engineering in order to maximize its yield of proteins and metabolites alike (Stein, 2005).

2.5.1. Antibiotic production

As mentioned above B. subtilis, like many Streptomyces species (Kunst et al., 1997), synthesizes many secondary metabolites. ~5% of genome is dedicated to antibiotic production, producing over two dozen antibiotics (concomitantly during spore production). Antibiotics are used to fight off other bacteria in such a competitive microbial community, increasing its chances of survival whilst also playing an important role in its morphology and physiology (Stein, 2005). Antibiotics may be harvested and used in treatment of infectious diseases due to innate anti-microbial actions (Davies & Davies, 2010).

B. subtilis antibiotics may be of the non-peptidic class, such as polyketides, or, the predominant peptide class which can be either;

(1) ribosomally synthesised and post-transationally modified i.e. lantibiotics
(2) non-ribosomally synthesised (Stein, 2005) including tyrocidine, gramicidin S, bacitracin, all well characterized at the genetic level. Two large operons; the pps operon and the surfactin operon. srfA are also responsible for the production of non-ribosomal peptide synthetases (Tsuge, et al 1999).

2.5.2. As a host for heterologous protein production

In recent years, considerable effort has been aimed at developing B. subtilis as a host for the production of heterologous proteins (Harwood & Cranenburgh, 2008). The high level of homologous proteins secreted by B. subtilis is apparent due to its naturally high secretory capacity and exportation directly to the growth medium (Simonsen & Palva, 1993). This causes separation of the product from cell components; making downstream processing a lot simpler and thus reducing costs (Harwood & Cranenburgh, 2008) (Kodama et al, 2012). Attempts to use Bacillus as a host for heterologous protein production have been hampered somewhat by properties of the secretory pathway, including; poor targeting to the translocace and incorrect/inefficient folding. Another problemmatic area of the pathway is degradation of the secretory protein due to the proteolytic characteristics of quality control peptidases and proteases (Westers et al, 2004). B. subtilis can be very genetically manipulated relatively easily so research into improving the organism as a host for heterologous protein production focusses largely on the inactivation (via mutagenesis or even total deletion) of protease producing genes (Barbe et al, 2009).

To date 8 extracellular proteases have been identified which are responsible for protein degradation; aprE (Stahl et al, 1984), hpr (Sloma et al, 1990), epr (Bruckner et al, 1990), mpr (Rufo et al, 1990), nprB (Tran et al, 1984), nprE (Yang et al, 1984), vpr (Sloma et al, 1990) and wprA (Margot et al, 1996) (Kodama et al, 2012). Strains showing deletions within the aprE and nprE genes with said mutants showing lower activities of extracellular proteases (Sloma et al, 1991). Another strain has been developed showing a deletion in the epr gene and resulted in low protease activity in the culture supernatant (Wang et al, 1989). Deletion mutations in multiple extracellular proteases have also been created displaying extracellular activities of <0.5% (in comparison to the parent strain) (Wu et al, 1991). WB800, an eight extracellular protein deficient strain proved to be a useful host for various heterologous proteins. KA8AX, deficient in the eight extracellular proteases as well as aprX, htrA and htrB looks promising (Kodama et al, 2012) Finding a strain that maximizes the yield of heterologous proteins is an essential area of research if B. subtilis is to achieve its potential as a host for industrial protein production (Kodama et al, 2012).

2.6 Strains to be analysed

Marburg ATCC 6051 was sequenced by Newcastle University (see previous for description).

A Protease Deficient strain was provided by Colin Harwood at Newcastle University. The strain was a manipulated B. subtilis 168 (trpC2), obtained from the Pasteur Culture Collection (CIP106309). Colin Harwood currently focusses on work involving increasing the yield of heterologous proteins using B. subtilis as a host; attempting to ‘knock out’ as many proteases as possible.

3 METHODS

Two fastaq files containing were provided containing: (1) reads from B. subtilis Marburg ATCC sequenced on behalf of Newcastle University, which will be referred to as the ‘Marburg strain’ and (2) from the Protease Deficient strain sequenced at Newcastle University on behalf of Prof. Colin Harwood this will referred to as the ‘Protease Deficient’ strain.

Table 1. Summary of reads provided for both application projects

<table>
<thead>
<tr>
<th>Strain</th>
<th>Platform Used</th>
<th>Type</th>
<th>No. Reads</th>
<th>Min/Max Length (bp)</th>
<th>Read Length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marburg</td>
<td>Illumina</td>
<td>Single</td>
<td>5,092,107</td>
<td>36/36</td>
<td>36</td>
</tr>
<tr>
<td>PD</td>
<td>Ion Torrent</td>
<td>Single</td>
<td>503,015</td>
<td>5/396</td>
<td>256</td>
</tr>
</tbody>
</table>

Details reads provided for both Marburg 6051 and the Protease Deficient (PD) strain; bp= base pairs.
3.1 Genome Assembly

Lasergene Core Suite, the commercially available genome and assembly software suite from DNASTar (including SeqNGen, SeqMan Pro & SeqBuilder), was downloaded\(^6\) and installed on Windows 7 (v 6.1) OS.

Command line operated GS Data Analysis Software package (including the GS De Novo Assembler and GS Reference Mapper), provided by 454 sequencing, was requested and downloaded\(^7\) to the Linux image on VM. FastQC v.10.1 for Win/Linux was downloaded\(^8\).

*Reads were run through fastqc prior to any analysis to test quality

3.1.1 Reference Assembly

3.1.1.1 Marburg In order to ensure assembly was done using same version of 168 genome retrieved from BacillusScope (used to populate BacillOndex dataset), Nummer was carried out on versions of the 168 genome from Genbank (accession number AL009126) against BacillusScope genome .fa files of the sequences were aligned in Nummer using commands: mummer - r mum [referenceseq location.fa] [queryseq location.fa] > [output location reference-query.mums] and then mummerplot -x “[0,4300000]“ -y “[0,4300000]” ref_query.mums.

3.1.1.2 Protease Deficient Strain DeNovo assembly was completed on reads of Protease Deficient Strain using SeqNGen (DNASTar suite) as well as GS DeNovo Assembler (GS Data Analysis Software Package).

In SeqNGen a de novo genome assembly project type was created and Ion Torrent selected as read technology. .fastaq file was entered in unpaired reads. Parameters were left as default with expected genome length entered as 4300000.

In GS De Novo Assembler the following commands were used; newAssembly [Project Directory Name]. addRun [Project Directory Name] [Read Files.fastq]. runProject -r [Project Directory Name].

10 largest contigs produced from each assembly were run through nucleotide Blast\(^9\), using Nucleotide collection database.

3.1.2 Reference Assembly

3.1.2.1 Marburg reference assembly was completed using SeqMan NGen; a template genome assembly was created and project folder identified for storage of output files. Reference genome.gb was added as input template file, with parameters left as default save Illumina <50 nt was selected as read technology, reads were added as unpaired and haploid genome ploidy selected.

After assembly project was opened in the SeqMan Pro tool for downstream analysis.

3.1.2.2 Protease Deficient reference assembly was completed using both SeqMan NGen as well as GS Mapper.

In SeqMan NGen everything was identical to that used for Marburg Strain (3.1.2.1), apart from Ion Torrent being selected as read technology.

In GS Mapper the following commands were used; newMapping [Project Directory name]. setRef [Project Directory] [Ref Seq.fa]. addRun [Project Directory Name] [Read data files]. runProject -r [Project Directory Name].

In order to identify large insertions that may be present within Protease Deficient Strain unassembled reads were re-assembled in the respective software using a de novo assembly and contigs produced Blasted.

3.2 Gap Analysis of Protease Deficient strain

Assemblies in both Seqman Pro and GS Mapper were analysed for gaps. To view areas of assembly contained gaps in SeqMan Pro, a coverage report was created showing depth of coverage throughout assembly. Areas with coverage of <2 were analyzed as potential gaps in the assembled genome. Gaps were also manually searched for as some were missed from coverage report.

GS Mapper allowed us to visualize the assembled contigs against the reference, points of interest were searched simply using the allocated search bar and any gaps that were observed noted.

Gaps present in both assembly tools were considered to be deleted regions, unless they occurred in areas of known assembly conflict (i.e. repetitive regions).

3.3 SNP Analysis for both strains

Ondex June 2012release (v. 0.4.0) was downloaded\(^11\)as a tar.gz and extracted onto Linux image within VM (using tar xzvf); the runme shell was made executable by applying chmod 755 runme.sh command.

BacillusOndex was downloaded\(^12\) as a .xml.gz folder, containing workflows (.xml), an ondex_metadata.xml file as well as the bacillondeX.PLUGIN.jar. Plugin was placed in the Onex plugin directory.

CatSNP files were downloaded and Plugin (Snpparse-0.2.0-snapshot.jar-withdependencies.jar) placed into Ondex plugin directory with a new ondex_metadata.xml replacing existing metadata file, ensuring required metadata required for the CatSNP graph was available. CatSNP.xml workflow and BacillOndex.xml (dataset) were placed in a new folder.

3.3.1 SNP Identification & Cleaning

For each strain SNPs were identified using the DNAStar SeqMan Pro tool and a SNP report created, which identified the SNP type of those occurring in CDS (ssSNP/ssSNP/INDEL). SNPs were filtered at 90-100% with coverage of 20 and manually visualized, with coverage, location on read (i.e. were the SNPs at read boundaries) and accuracy manually evaluated before being ‘confirmed’. Confirmed SNPs were exported in a tab delimited SNP.txt file. If for any reason SNPs occurring in the CDS of a protein were not identified as ‘type’ the B. subtilis codon usage table was referred to and type was manually added.

3.3.2 Systems Level Visualization of SNPs

Ondex was launched (hash runme.sh) in terminal, and CatSNP.xml workflow opened from integrator. BacillusOndex dataset was selected as input file and SNP.txt selected. Output type was identified as FULL (meaning all integrated concepts were present within the graph) and workflow run.

In order to visualise each SNP within a separate sub-graph SNPs were searched for individually within graph and neighbouring concepts filtered to a depth of 1. SNP associated concepts were selected and immediate neighbourhood shown; this process was iterated until all concepts relating to an individual SNP were visualized. Labels for both concepts and relations were called and a ‘Gem’ layout chosen. Once sub-graphs were created, annotations associated to each concept were viewed and noted.

\(^5\) http://virtualbox.so932.com/
\(^6\) http://www.ubuntu.com/
\(^7\) http://www.dnastar.com/
\(^8\) http://454.com/products/analysis-software/index.asp
\(^9\) http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc
\(^11\) http://www.ondex.org/
\(^12\) http://www.bacillondeX.org
4 RESULTS
4.1 Marburg Strain

4.1.1 Assembly

After running reads through fastaQC all modules of the report were deemed ‘normal’, apart from per base sequence content and per base GC content deemed ‘slightly abnormal’ and sequence duplication levels apparently ‘very unusual’. Per base sequence quality remained ‘high’ for position 1-36, with quality only falling slightly toward the end of the reads.

Nummer (using all the versions of AL009126 available from GenBank against the sequence from BacillusScope (used for the BacillOndex dataset), found a perfect alignment with GenBank entry AL009126.3 (Submitted March 17, 2009); this was used as the reference sequence of 168 for assembly.

Reference assembly was achieved using >99% of reads (only 21,131 reads were not aligned against reference), 27 short gaps (<64 nt) and 3 longer gaps (between 276-733 nt) were present.

1.1.2 SNP Identification & Cleaning

Fig 3. Breakdown of SNPs and INDELS at different % filters of SNPS produced in DNAStar SeqMan Pro post-assembly (SNPs not found in graph during Systems Level visualization were not included in this graph)

Increasing the level of SNP quality (calculated as a percentage of reads covering that base showing SNP). This does not cause the loss of SNPs from 70% to 90% with only a reduction in INDELS present. It was noted that as SNP quality was raised to 90% we lost the known INDEL associated with SwrA (with a score of 85.6%).

4.1.2 Systems Level SNP Visualization

Although, after cleaning, we had 65 SNPs in the .txt file to be filtered into the BacillusScope dataset using CatSNP, of which only 57 were successfully filtered into overall graph (44 nsSNPs/INDELS, 5 sSNPs and 8 eSNPs) with 8 not located.

The Overall Graph produce using CatSNP for Marburg Strain containing the following concepts included: SNP (57), Pathway (18), Reaction (19), Enzyme Classification (11), Enzyme (10), Protein (40), CDS (39), Protein Feature (55), Promoter (3), Shim(1), Ribosomal Binding Site (1), Operator (1), and of the GO concepts; Cellular Component (8), Molecular function (59), Biological Process (48)

Table 2. Summary of proteins within 168 with associated SNPs

<table>
<thead>
<tr>
<th>Protein</th>
<th>Systems Level Involvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsSNPs</td>
<td></td>
</tr>
<tr>
<td>TrpC</td>
<td>Phenylalanine, tyrosine and tryptophan biosynthesis*</td>
</tr>
<tr>
<td>AroH</td>
<td>Phenylalanine, tyrosine and tryptophan biosynthesis *</td>
</tr>
<tr>
<td>SIP</td>
<td>Antibiotic anabolism</td>
</tr>
<tr>
<td>PpsC</td>
<td>Antibiotic anabolism</td>
</tr>
<tr>
<td>GerAA</td>
<td>Spore germination</td>
</tr>
<tr>
<td>SpoG</td>
<td>Cellular spore formation</td>
</tr>
<tr>
<td>SigH</td>
<td>Cellular spore formation</td>
</tr>
<tr>
<td>MbL</td>
<td>Cellular morphogenesis</td>
</tr>
<tr>
<td>PIt</td>
<td>Phosphate transport</td>
</tr>
<tr>
<td>AmyD</td>
<td>Transport</td>
</tr>
<tr>
<td>PhoD</td>
<td>Folate biosynthesis *</td>
</tr>
<tr>
<td>SacA</td>
<td>Starch &amp; sucrose metabolism *</td>
</tr>
<tr>
<td>Glt</td>
<td>Alanine, aspartate and glutamate metabolism*</td>
</tr>
<tr>
<td>HemA</td>
<td>Porphyrin and chlorophyll metabolism *</td>
</tr>
<tr>
<td>RbsR</td>
<td>Regulation of transcription</td>
</tr>
<tr>
<td>TrmA</td>
<td>Methylation</td>
</tr>
<tr>
<td>Rlb</td>
<td>RNA modification</td>
</tr>
<tr>
<td>eSNPs</td>
<td></td>
</tr>
<tr>
<td>YesS</td>
<td>Transcription, DNA dependent</td>
</tr>
<tr>
<td>YisR</td>
<td>Transcription, DNA dependent</td>
</tr>
<tr>
<td>PerR</td>
<td>Regulation of transcription</td>
</tr>
<tr>
<td>FglM</td>
<td>Regulation of transcription</td>
</tr>
<tr>
<td>MinR</td>
<td>Cysteine and methionine metabolism*</td>
</tr>
<tr>
<td>DefQ</td>
<td>-</td>
</tr>
<tr>
<td>sSNPs</td>
<td></td>
</tr>
<tr>
<td>PpsC</td>
<td>Antibiotic anabolism</td>
</tr>
<tr>
<td>KipA</td>
<td>Cellular spore formation</td>
</tr>
<tr>
<td>PgdS</td>
<td>Hydrolase **</td>
</tr>
<tr>
<td>IvIC</td>
<td>Valine, leucine and isoleucine biosynthesis*</td>
</tr>
</tbody>
</table>

Details all proteins with SNPs associated (either in regulatory regions or CDS) in 168 successfully located within the graph. Please note some proteins have more than one SNP associated. SNPs are broken down by type systems level involvement extracted from the graph as a pathway* or in the absence of a pathway the most descriptive biological process was used. Does not include the uncharacterized/putative proteins (YjxJ, XldJ, YjgA).

After applying systems level analysis to all SNPs found (Table 2) via the production of sub-graphs (see Fig 4) we find that a few proteins have >1 SNP associated, including the plipistatin synthase sub-unit C, PpsC (15), the transcriptional regulator Ribose operon repressor, RbsR (3), the uncharacterized protein YmK (2) and the putative enzyme, YjpA (2) with the promotor region of the citrate synthase protein, CitZ also containing eSNPs (2).

Looking at Table 3 we can see a summary of the most frequently occurring pathways and GO terms associated with proteins related to a SNP.

No nsSNPs were found in the essential genes of B. subtilis (as defined by Kobayashi et al (2003)), apart from within the protein sequence of the tRNA maturation associated protein (TrmA).
Fig 4. Shows an example of each SNP type in sub-graphs. A - displays 16 SNPs (1sSNP & 15 nSNPs- of which 9 are INDELs) associated with the protein chain of the plipistatin synthase subunit C, PpsC. 6 SNPs are found within the adenylation 1 region of the protein chain with 9 found within the epimerization region (please not that for the purpose of this graph the SNP_Labels as well as SNP_relation labels (is_subfeature_of) have been hidden). B- shows an eSNP within the promoter region of the pleiotropic regulator, DegQ. C- graphical visualisation of the insertion within the protein chain of the Surfactin synthetase-activating enzyme, SfP. D- we can see that both SfP and PpsC are involved in the formation of antibiotics.
4.2 Protease Deficient Strain

After running reads through fastQC a few modules of the report were deemed ‘very unusual’, including: per base sequence quality, per base GC content and Kmer Content. No reads were filtered but it was shown that per base sequence content was ‘slightly abnormal’. For assembly results please see Table 5 (below). Note that after reference assembly unassembled reads failed to produce a contig.

4.2.1 Finding Reference

10 largest contigs from DNASTar (ranging from 17136-27077 bp in length) were blasted and all showed E-values of 0.0 and max identities of >98% when aligned against the *B. subtilis* subsp, subtilis strain 168 complete genome. This has an accession number of AL009126.3 and was the top ‘hit’. This was also the case when the 10 largest Newbler contigs were ‘Blasted’ (ranging from 164941-406546 nt in length). 168 could be used as the reference genome for assembly of the Protease Deficient strain. As mentioned in 4.1.1 we needed to ensure the correct version of the genome was used to allow for accurate analysis in Ondex and so the GenBank entry AL009126.3 (Submitted March 17, 2009) was used.

4.2.2 Gap Analysis

Looking at Table 5 we can see Newbler identified 28 gaps (of which 15 were repeat regions) and that DNASTar identified 14 (of which 3 were repeat regions). Only gaps that were apparent in both assemblies and did not occur in repeat regions were considered to be deleted genes. Repeat regions included BSU_rRNAs 1,2,3,4,5,6,8,9,14,15,16,17,18, 21,22, 26,27,28, 29, 30 and 80. 10 genes were found to be deleted, encoding eight extracellular proteins; *nprE* (Bacillocylin), *bpr* (bacillopeptidase F), *wprA* (cell wall associate protease A), *aprE* (subtilisin), *mpr* (extracellular metalllopeptase), *vpr* (extracellular serine protease) and *epr* (extracellular serine protease) and two membrane bound proteases, *htrA* (Serine Protease Do-like HtrA) and *htrB* (Serine protease Do-like HtrB).

* Encoding proteins produced by genes were taken from Uniprot13

4.2.3 Systems Level SNP Visualization

SNP filtering was originally carried out using a coverage of >20, producing 19 SNPs, this was changed to a coverage of >10 with 36 SNPs then located. No nsSNPs were found to be associated with any of the essential genes as defined by Kobayashi et al (2003).

Although, after cleaning, we had 36 SNPs in the .txt file to be filtered into the BacilIOndex dataset, only 25 were successfully filtered into overall Protease Deficient Strain graph (20 nsSNPs/INDELS, 3 sSNPs and 3 eSNPs) with 10 not located, other concepts (frequency) included; Operon (1), Terminator (1), CDS (20), RBS (1), Operator (1), Shim (1), Promoter (2), Protein Feature (28), Protein (21), Reaction (18), Enzyme Classification (6), Enzyme (5), Path (14) and of the GO concepts; Cellular Component (6), Biological Process (27), Molecular function (41).

25 sub-graphs were created, such as that for *SpoIIGA* (Fig 5) allowing for visualization of SNPs related to the proteins in Table 3 (see below).

Table 4. Summary of proteins within 168 with associated SNPs

<table>
<thead>
<tr>
<th>Protein</th>
<th>Systems Level Involvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsSNPs</td>
<td></td>
</tr>
<tr>
<td>AmyE</td>
<td>Starch and sucrose metabolism *</td>
</tr>
<tr>
<td>YifG</td>
<td>Transmembrane transport</td>
</tr>
<tr>
<td>YitF</td>
<td>Metabolic processes</td>
</tr>
<tr>
<td>UvrX</td>
<td>DNA repair</td>
</tr>
<tr>
<td>VpR</td>
<td>Peptidolysis</td>
</tr>
<tr>
<td>NprE</td>
<td>Peptidolysis</td>
</tr>
<tr>
<td>GerAA</td>
<td>Spore germination</td>
</tr>
<tr>
<td>IpdV</td>
<td>Valine, leucine &amp; isoleucine degradation*</td>
</tr>
<tr>
<td>YkoW</td>
<td>Signalling cascade</td>
</tr>
<tr>
<td>YskK</td>
<td>Fatty acid biosynthesis *</td>
</tr>
<tr>
<td>EpsH</td>
<td>Polysaccharide biosynthetic process*</td>
</tr>
<tr>
<td>eSNPs</td>
<td></td>
</tr>
<tr>
<td>SacX</td>
<td>Starch &amp; sucrose metabolism*</td>
</tr>
<tr>
<td>SpoIIGA</td>
<td>Peptidolysis</td>
</tr>
<tr>
<td>sSNPs</td>
<td></td>
</tr>
<tr>
<td>UvrX</td>
<td>DNA repair</td>
</tr>
<tr>
<td>ArgB</td>
<td>Arginine &amp; proline metabolism*</td>
</tr>
</tbody>
</table>

Details all proteins with SNPs associated (either in regulatory regions or the protein sequence) in the Protease Deficient strain successfully located within the graph. Please note some proteins have more than one SNP associated. SNPs are broken down by type systems level involvement extracted from the graph as a pathway* or in the absence of a pathway the most descriptive biological. Does not include the uncharacterized/putative proteins (YwiB, YwcH, YitS, YxbD).

Fig 5. Systems level visualization of the eSNP between the operator and promoter of the sigma-E factor-processing peptidase.
Table 5. Assembly Statistics

<table>
<thead>
<tr>
<th>Tool</th>
<th>Contigs/ &gt;2k nt</th>
<th>Contig</th>
<th>Unassembled Reads</th>
<th>% Coverage of ref</th>
<th>Unassembled Reads</th>
<th>Gaps</th>
<th>Repetitive Region Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNAStar</td>
<td>6298/637</td>
<td>1113</td>
<td>110736</td>
<td>99.2</td>
<td>21535</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Newbler</td>
<td>77/44</td>
<td>93867</td>
<td>13330</td>
<td>98.5</td>
<td>32827</td>
<td>28</td>
<td>15</td>
</tr>
</tbody>
</table>

Shows assembly statistics of the Protease Deficient strain in both de novo and reference assemblies using the DNAStar and Newbler tools.

Table 6. Summary of terms extracted from sub-graphs of SNPs in Protease Deficient Strain

<table>
<thead>
<tr>
<th>Pathways</th>
<th>Biological Process (GO term)</th>
<th>Molecular Function (GO term)</th>
<th>Cellular Component (GO term)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolic (4)</td>
<td>Metabolic Process (4)</td>
<td>Metal ion binding (7)</td>
<td>Membrane (7)</td>
</tr>
<tr>
<td>Starch and sucrose metabolism (3)</td>
<td>Peptidolysis (4)</td>
<td>Transferase (7)</td>
<td>Plasma membrane (7)</td>
</tr>
<tr>
<td>Biosynthesis of secondary metabolites (1)</td>
<td>Transport (3)</td>
<td>Hydrolase (5)</td>
<td>Integral to membrane (7)</td>
</tr>
<tr>
<td>Glycine, serine and threonine metabolism (1)</td>
<td>Phosphorylation (3)</td>
<td>Kinase (3)</td>
<td>Cytoplasm (2)</td>
</tr>
<tr>
<td>Microbial metabolism in diverse environments (1)</td>
<td>DNA repair (3)</td>
<td>Nucleotidyltransferase (3)</td>
<td>Protoplasm (1)</td>
</tr>
</tbody>
</table>

After all sub-graphs were created for Protease Deficient Strain, the information attached to each concept was removed. This table summarizes the most frequently occurring pathways as well as GO terms (Biological Pathways, Metabolic Functions and Cellular Components). Not all SNPs were associated to proteins that had pathway involved.

Fig 6. Shows the gaps in the reference assembly of the Protease Deficient Strain (against B. subtilis 168). Newbler assembly is represented by the red circle and DNAStar by green. Gaps are shown as intersects with consensus gaps considered deleted regions; genes that occur in these deleted regions are identified as blue intersects on the reference genome. SNPs in the Protease Deficient strain are shown as an orange triangle.
5 DISCUSSION

5.1 Project Application 1: Analysis of Marburg v 168

Using a similar methodology, minus the SNP ‘cleaning’ step, Katherine James previously identified 100 SNPs associated with 168 (v Marburg). Previous analysis failed to identify SNP type i.e. which were sSNPs, nsSNPs and eSNPs. Using the cleaning methods implemented it was possible to reduce this number, concomitantly with an increase in quality, to 65 whilst also identifying the type (44 nsSNP/INDELS occurring in protein CDS). Systems level analysis allowed for the analysis of these SNPs at phenotypic level including those that have previously been identified within the 168 strain; srF, trpC and aroH (Kearns et al, 2004) (Ziegler et al, 2008).

5.1.1 Assembly

fastQC identified a high level of A&T for each position of sequences (~30%) from reads, not unexpected for an organism known to be AT rich (Westers et al, 2004). The report also identified a sequence duplication level of 70.5% (expected to be ~40%) which may indicate some kind of enrichment bias (e.g. PCR over amplification). This is most likely due to the platform that was used.

5.1.2 SNP Identification

Looking at SNP percent filter (Fig 3) an increase in this from 70% to 90% only INDELS (>50%) were lost and no SNPs. Using the four known SNPs identified as reference (swrA, sfp, trpC and aroH), the INDEL in swrA, was not present yet the other three remained. It has also been identified that the 168 trpC locus is inactivated by a three-base deletion, induced during original x-ray treatment, with the Trp+ Marburg ATCC 6051 showing a short repeat [ATG] (Albertini, & Galazzi, 1999). This is apparent in the form of three INDELS that occur at position 2,374,564, however two of the INDELS were not included after SNP cleaning as they fell below the threshold but explains the presence of the third (83.2% & 85.2 % respectively) (Fig. 7).

Fig 7. Shows three INDELS associated with protein sequence of TrpC and SNP quality % score

It was decided, in spite of this, that a level of filtering of 90% would be used as this matched filtering used in similar studies (Ziegler et al, 2008) as well as the fact that it is known that accurate inference of INDELS from short read data is challenging (relatively to the identification of SNPs) for a number of reasons. INDELS are thought to occur at a rate of ~eight fold lower than that of SNPs, making them more difficult to detect, this coincides with results obtained (see Fig 3) with reads arising from INDEL sequence proving a lot more difficult to map to the correct location within a sequence (Li et al, 2008) (Albers et al, 2010).

5.1.3 Systems Level SNP Analysis

Upon initial of the most frequent GO terms associated with proteins related to SNPs it is apparent they are heavily involved in metabolism (Table 3). This is a rather ambiguous term to use in order to relate these SNPs to phenotype. It is also apparent that a number of proteins relating to cellular spore formation have SNPs associated (Table 2). This may be due to the fact that the strain is experiencing nutrient deficiency at lower levels than its wild type parent and thus binary fission (aka the vegetative life cycle which utilizes FtsZ) is the ‘preferred’ life cycle for the laboratory strain 168, as opposed to sporulation (where a small spore is produced by the mother cell) (Angert, 2005). Although these summary tables allow for a general overview of SNP effects on the focusing on an individual entity allows for greater understanding to be realized. The pathways phenylalanine, tyrosine and tryptophan biosynthesis and the biological process antibiotic anabolism for example are noticeable in their frequency; we will now look closer at SNPs associated to these terms.

5.1.3.1 AroH; within the Phenylalanine, tyrosine and tryptophan biosynthesis pathway DNAStar identified a nsSNP in the CDS of AroH (Marburg has a codon sequencing Alanine –GCT- and 168 a Valine –GTT- @ codon 112) and CatSNP allows us to identify the exact location of the sequence variation within the protein, ultimately allowing for the analysis of genomic effects at phenotypic level.

The AroH nsSNP is located within a helix of the chorismate mutase aroH-type domain of the chorismate mutase protein. AroH. CatSNP identifies gene ontologies associated with the protein/enzyme; it is located in the cytoplasm and has isomerase and a chorismate pyruvate mutase molecular functions; which are very specific to the protein. This is very well but we need to look at the biological processes which describe processes that molecular functions are involved in to gain a real insight into the protein at systems level; with the cellular amino acid biosynthesis process and aromatic amino acid family anabolism being associated. We can begin to see how a nsSNP at protein level may affect the organism as a whole. Finally pathways associated include biosynthesis of secondary metabolites and phenylalanine, tyrosine and tryptophan biosynthesis.
One can assume, looking purely at results achieved using CatSNP that a SNP affecting the AroH proteins structure and function will have a detrimental effect on the organism’s ability to successfully synthesise the amino acids, tryptophan, phenylalanine and tyrosine. This claim is supported with AroH (Chorismate mutase AroH) being the catalyst for the Claisen rearrangement of chorismate to prephenate. This is an important step involved in the biosynthesis of tyrosine and phenylalanine in B. subtilis (Choik et al., 1993). Phenotypic effect? 168 has an inability to synthesise tryptophan.

CatSNP produced analysis matches literature work on the same sequence variation (Ziegler et al., 2008). Accuracy of the methodology used for systems level analysis of SNPs in 168 accurately identified the effects of a known SNP on the organism as a whole; identifying its inability to produce tryptophan and phenylalanine (Ziegler et al., 2008). This allowed for a further assumption regarding the phenotype of 168 to be made based on results from this study; it has an inability to produce plipistatin.

5.1.3.2 PpsC: associated to the Antibiotic Anabolism Biological Process Looking at Fig 6 (A) we can see the 16 SNPs, of which 15 are non-synonymous, identified within the protein sequence of PpsC, occurring within two regions of interest within the protein chain; (1) an adenylation region (2) epimerization region.

The graph tells us plipistatin synthetase, PpsC is an enzyme with the role of antibiotic production. The sub-graph informs us that PpsC is a transferase (transfer of a functional group from one molecule to another) and a ligase (catalyses the joining of two large molecules) enzyme, that it is an acyl carrier (important component in both fatty acid and polyketide biosynthesis) involved in phosphopantetheine binding as well as cofactor binding. PpsC is associated with the rather ambiguous metabolic process as well as the more specific antibiotic anabolism.

Without any prior knowledge of PpsC; using the BacilliOndex dataset as well as CatSNP plugin we can see that we have 16 SNPs associated with a protein involved in the synthesis of the polyketide antibiotic plipistatin. At this stage it must be said that the systems level analysis of PpsC was hindered by the fact that none of the concepts that appear in Fig 6 are associated with any further concepts; we cannot see which other proteins PpsC interacts with and to what extent, or indeed any pathways PpsC is involved in, using CatSNP and in the Ondex platform alone, but we can assume an inability to produce plipistatin will be apparent at phenotypic level.

PpsC makes up one of five enzymes produced by the pps operon, a 38kb segment that carries genes ppsA,B,C,D & E for the biosynthesis of the non-ribosomally synthesized lipopeptidic antibiotic, plipistatin (Tsuge et al., 2007). It has been shown that the pps operon, as well as DegQ and SfP are essential in order for the production of plipistatin in B. subtilis (Tsuge et al., 1999). Analysis has successfully identified an eSNP in the promoter of the pleiotropic regulator, DegQ (see Fig 6 (B)) as well as a nsSNP in the chain of the surfactin synthetase activating enzyme which has been proven to inactivate it (Tsuge et al., 1996), SfP (see Fig 6 (C)). SfP acts both as an antibiotic as well as a surfactant used for swimming and swarming (Julkowska et al., 2005). Although this provides strong evidence that either Marburg or 168 are incapable of the production of plipistatin this is not the case (Tsuge et al., 2007) (Tomita et al., 2004). It is understood that 168 is capable of producing plipistatin at a very low yield in the absence of the pleiotropic regulator degQ, with its presence increasing production 10-fold (Tsuge et al., 1999).

Although nsSNPs were located in both protein chains of PpsC and SfP as well as an eSNP in the DegQ regulator the phenotypic effects are not as one would expect; raising an interesting limitation to the methodology employed.

5.1.4 Further Work for Project Application 1

Using the Marburg ATCC strain in order to look at SNPs in 168 has its limitations. Although it is believed that there is little to no difference in the sequence make-up of both Marburg ATCC and Marburg NCIB 3610 (Ziegler et al., 2008), one cannot presume this to be the case (ATCC is yet to be sequenced and published). We do not know the true parent strain to our 168. It would be useful to carry out the analysis using reads from NCIB 3610 and to compare the results to those obtained during this work.

B. subtilis 168 does not swarm on motility agar plates; as the genes responsible for this sfp and swrA are known to be inactivated in 168 by INDELs (Ziegler et al., 2008). As mentioned we ‘lost’ the swrA mutation during cleaning. This may be due to many factors and re-sequencing of this site is proposed via PCR and utilization of platforms that produce longer reads than Illumina; such as Sanger (Magi et al., 2010). A similar method was proposed for areas of inaccuracy during the 1997 sequencing project of 168 and other sequence verification procedures (Barbe et al., 2009) (Srivatsan et al., 2002). This will allow for re-annotation if necessary.

5.2 Project Application 2: Protease Deficient Strain v B.subtilis 168

5.2.1 Assembly

fastQC identified a large reduction in per base sequence quality as reads reached >250 nt (with quality score depreciating by 50% between position 100 and 299). This is to be expected as it is known that this is a common flaw with NGS platforms (Magi et al., 2010). Like the Marburg strain the sequence content across all bases identified T&A being present at ~30% in each position of
reads; dealing with *B. subtilis* strains, this is expected. Reads were deemed of high enough quality to continue with the analysis.

### 5.2.2 Deleted Proteases

10 genes encoding proteases were identified as having been deleted. This potentially has huge implications on the usage of *B. subtilis* as a host for the production of heterologous proteins. As mentioned, quality control methods are in place within the secretory pathway of *B. subtilis* monitoring proteins at the membrane and wall interface, degrading those that appear to be misfolded (Harwood & Cranenburgh, 2008). The three serine proteases WprA, HtrA and HtrB are the main proteases involved in this process.

htrA and htrB genes are transcribed in response to secretion stress in the transition phase of the growth cycle (i.e. over production of AmyQ) or during response to heat shock during exponential growth, controlled by CssR two component system which senses accumulation of misfolded proteins in the membrane-call (Hyrylainen *et al.*, 2001) (Harwood & Cranenburgh, 2008). HtrA and HtrB proteins rescue cells from lethal accumulations of misfolded proteins in the cell envelope. Expression is negatively auto regulated and reciprocally cross-regulated; the absence of HtrA leads to the increased synthesis of HtrB and vice versa. Inactivation of both genes causes a marked sensitivity to thermal or oxidative stress. Such mutants are proven to show a noticeably reduced growth rate and yield of secretory proteins (Noone *et al.*, 2001). This ‘sick’ organism is not a characteristic observed by the Protease Deficient strain (informed by Colin Harwood). This poses an interesting question: do SNPs affect the phenotype of the organism in such a manner as to counteract the effects of deleting these proteases?

### 5.2.3 Systems Level SNP Analysis

A coverage of 10 (as opposed to the proposed 20) was used for the production of our SNP report. This was due to the relatively low number of reads (Fig 1) which gave a median coverage of 18.42 (from DNASTar assembly).

#### 5.2.3.1 SNPs associated with flanking regions of deleted proteases

Deletions in the Protease Deficient strain were achieved using insertion cassettes and deletion plasmids. Insertion cassettes contain an antibiotic resistance gene flanked by *dif* sites (used to delete *nprB, aprE, expr* and *bpr*). This involves the insertion (into genome) of a cassette consisting of an antibiotic resistant gene flanked by 28-bp *dif* sites as well as regions that are homologous to the chromosomal target locus. *dif* is a recognition sequence for native Xer site specific recombinases responsible for chromosome and plasmid dimer resolution (Ripx/CodV in *B. subtilis*). Homologous recombination allows for the integration of the insertion cassette to the target site of the target chromosome; with recombinases then acting on the integrant to resolve the two directly repeated *dif* sites, thus excising the antibiotic resistant gene, leaving behind a *dif* site where the target gene once was (Bloor & Cranenburgh, 2006). The deletion plasmid used was ORI (used to delete *nprB, aprE, expr* and *bpr*). This uses a similar approach whilst utilising Xhol sites and leaving 6 bp flanks behind.

This is important when it comes to the analysis of our SNPs as it appears flanking regions that were left incorporated into the genome may explain the presence of 12 SNPs within genes neighbouring deleted proteases;

1. *ywCH* (neighbours *vpR*) encodes an uncharacterised protein. This protein has a sSNP and two nsSNPs in its protein chain; it is unknown how these will affect phenotype as no concepts are related to it.

2. *sacX* (neighbours *epr*) encodes a sucrose synthesis operon antiterminator; a negative regulator of SacY. It has one sSNP within its Ribosomal Binding Site (RBS) and an eSNP upstream of RBS. Phenotypic effect? An increase in the presence of the SacB anti-terminator, SacY. Resulting in decreased levels of SacB and a decreased sensitivity to sucrose (Idelson & Amster Choder, 1998).

3. *yitS* (neighbours *nprB*) encodes a DegV-domain containing. This protein has a nsSNP within its protein chain; it is unknown how this will affect phenotype as it is a homologue of previously reported proteins of unknown function.

4. *yitF* (neighbours *nprB*) encoding a putative isomerase, has a nsSNP within a strand of the protein chain of the putative isomerase *YitF* (4).

These SNPs cause the results observed in Table 4 and Table 6 to be quite misleading. SNPs in these regions could not be ignored and associated information dismissed from the tables as they are still have potential to effect phenotype; hence remaining. Interestingly SNPs were also found in the deleted genes *vpR* and *aprE* (Table 4). This is down to the gene deleted regions that were chosen during strain manipulation which differed between genes (some genes were completely removed, some not). Phenotypic effect? A reduction in peptidolysis, in exactly the same manner as the deleted proteases.

#### 5.2.3.2 Remaining SNPs

Upon initial of the most frequent GO terms associated with proteins related to SNPs it is apparent they are heavily involved in metabolism (Table 4). This is a rather ambiguous term to use in order to relate these SNPs to phenotype. It is also apparent that a number of proteins relating to peptidolysis (Table 6). Peptidolysis is expected to be apparent due to the SNPs associated deleted to the proteases. What is interesting is the mutation associated with SpolIga, the sporeulation sigma-E factor-processing peptidase (Fig 5); also related to the biological process peptidolysis.
5.2.3.3 SpoIIGA; Associated to the Peptidolysis Biological Process
Looking at the systems level view of this SNP we see it is found between the operator and the promoter of the sporulation sigma-E factor- processing peptidase SpoIIGA. This enzyme is a hydrolase acting on peptide bonds and has carboxyl protease characteristics. Involved in mitotic sporulation, cellular spore formation and peptidolysis molecular functions it is located in and is integral to the plasma membrane.

With the location of the SNP identified as being within the regulatory region of the gene encoding this protein, one can assume that expression is altered; in a negative or positive fashion is a detail that remains elusive. One may predict that, due to the peptidolysis characteristics of SpoIIGA its up-regulation has occurred in the absence of other proteases; aiding in the organisms ability to remain healthy.

Research into the functionality of SpoIIGA has led to suggestions it is a membrane protein with five regions spanning the membrane and a putative protease. Of particular interest is the characteristic of its C-terminal, of which it shares with serine protease (Masuda et al, 1990). It has also been suggested that the protein is a novel type of aspartic protease (Imamura, et al 2008).

Taking this into consideration it is proposed that the phenotypic effect of this SNP in the Protease Deficient strain is a higher expression of spoIIGA. Characteristics that it shares with the serine like proteases go some way to compensate for the absence of the HtrA-like proteases.

5.2.4 Areas of Interest

5.2.4.1 Assembly Looking at Table 5 we can see some of the statistics associated with assemblies. It is worth looking at the differences between the length of the contigs produced by DNAStar and Newbler which both utilise overlap-layout consensus (OLC) assembly strategies (Kumar & Blaxter, 2010). Newbler managed to incorporate more reads producing ~10^7 contigs, with an average length tenfold those produced by the commercially available (and costly) DNAStar. It has to be said that both tools produced contigs that aligned as a top hit with the reference sequence 168, and that DNAStar did perform better in the reference assembly, just; with a greater coverage of reference achieved (0.7%) and fewer unassembled reads. Gaps were more accurately identified using the DNAStar suite, with only 3 occurring in repetitive regions (compared to the 15 in Newbler), however extraction of the data was a lot more complicated and time consuming than that of Newbler, with many gaps having to be manually identified.

5.2.5 Further Work for Project Application 2

A third, more distantly related HtrA like protein has been identified, HtrC (yyck/ yyxa). It is part of the wallRKyychHIJK 6-cistron operon and is expressed during exponential growth by the main operon promoter, but is also expressed separately by a Sig-type promoter positioned immediately upstream of yyck that is active during (stage III) of sporulation (Kenney & Moran, 1987) (Fabret &Hoch, 1998) (de Jong, et al 2012). The sigma type responsible for the up regulation (solely) of the protease is yet to be identified. sigma-E is co-transcribed in an operon of which spoIIGA is the promoter-proximal gene, essential during sporulation (Kenney & Morgan, 1987). Of interest would be the interaction if any, that this sigma-E has with htrC in light of the identification of this SNP.

5.3 Limitations to systems level approach described

The success of CatSNP depends largely on the knowledgebase with which it is working on, which in turn relies on the data sources from where it integrates its data. Systems levels analysis is only as valuable as the data that it contains. Some concepts had very few concepts related to them and thus proved very difficult to analyze (see Fig 6-B, sub-graph for DegQ). It is not the genome sequence that is of particular interest, rather the identification of the genomic objects (Barbe et al, 2009). During the production of sub-graphs we found that 8 SNPs were absent from the .txt file in the Marburg strain and 11 from the Protease Deficient strain, this is likely due to the areas they are related to remaining un annotated. BacilluScope and dbTBS were used to annotate 168 in Bacil-Londex and annotations may need updating. After reassessing the transcriptome of B. subtilis it was found that some of the promoter regions may be different to what we currently believe (Nicolas et al, 2012). In order to ensure that analysis is as accurate as possible this new data should have probably been integrated into the dataset, as opposed to using the information provided by dbTBS with the transcriptome, for example, having been recently re-mapped by a large consortium Graphs are produced in such a way that concepts are not included if they have no relations to other concepts. This means that SNPs that are not included must be manually located within the genome and annotations relating to that region of the genome searched for.

Using Ondex allowed us to view all concepts related directly with another, i.e. all the information regarding a certain protein. A true systems level analysis was not as easy to visualize as one would have liked. If a concept was not associated with a pathway then it was not possible to reach other proteins/ genes, and hence not possible to see which proteins your concept of interest interacts with. If however a protein was indeed associated with a pathway it was possible to see which other proteins were related to that pathway and a truer systems level analysis could be carried out; telling us a lot more about how a SNP may affect the organism as a whole. A few problems were also apparent in that some concepts were not related to the concepts that they should be (e.g. the CitZ promoter
region was not linked to the CitZ protein). This meant that two sub-graphs needed to be created.

When we have identified a SNP and looked at its presence within a network we still need to know the exact manner in which the SNP is affecting the protein. Just knowing a nsSNP is present within a CDS is not enough to claim the protein is not functioning in the same manner as before, or indeed if the functionality has changed focus slightly from the pre mutated form. Systems level analysis allows us to visualize where the SNP falls into the organism, but without biological data regarding the definitive effect of a SNP no conclusive findings can be drawn.

5.4 Further work for CatSNP approach

From a users’ perspective CatSNP is not as easy to use as one would hope. This is not so much a problem with the plug-in but rather the platform on which it works. The Ondex integrator is not something that is easy to work with. If CatSNP could be ran without having to implement it via a workflow ease of use would be drastically improved.

The amount of time taken to manually search through graphs created by CatSNP is intense. It takes around 20 minutes to extract the required information into a tabular format which is easier to manipulate/ evaluate outside the Ondex platform. A plug-in that automatically creates this table would be extremely advantageous for quick analysis and is something that is currently being considered by my colleague Warispreet Singh. The introduction of different concept types for different SNP types would also be beneficial for quicker analysis.

CatSNP is simply a filter that filters a SNP dataset into an already existing dataset. Adapting the plug-in to allow for the integration of any kind of concept (not just SNP) data would be extremely beneficial.

CONCLUSION

CatSNP proved to be an extremely beneficial tool to aid in the study of SNPs at a systems level. Systems level analysis, however, essentially relies on a high level of assumption as well as the presence of highly in depth and accurate data sources. Phenotypic effects of known SNPs within 168 were successfully visualized and their phenotypic effects predicted correctly using this method. However, when an inability to produce plipistatin was proposed, existing literature disproved this. A systems level approach of SNP analysis gives you a greater insight into the world of a biological system and where these SNPs fall. It allows for solutions to biological problems to be proposed, but not definitively answered. A better understanding of the two, biological systems has been achieved, and the manner in which SNPs interact within them. It has not conclusively answered any problem in this study but has allowed for the proposal of further work, of particular interest the compensatory mechanisms within the Protease Deficient strain.

ACKNOWLEDGEMENTS

I would initially like to extend my thanks to my supervisors at Newcastle University, Prof Anil Wipat and Dr Jennifer Hallinan. I would also like to thank Dr Katherine James for the initial analysis of 168, providing the CatSNP plugin and the support she provided in using Ondex.

REFERENCES


