UNIVERSITY OF CALIFORNIA, SAN DIEGO

ALE Analytics: A Software Pipeline and Web Platform for the Analysis of Microbial Genomic Data from Adaptive Laboratory Evolution Experiments

A Thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Computer Science and Engineering

by

Patrick Phaneuf

Committee in charge:

Professor Bernhard Palsson, Chair
Professor Vineet Bafna
Professor Pavel Pevzner

2016

The Thesis of Patrick Phaneuf is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California, San Diego

2016

DEDICATION

I dedicate this work to my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I came to University of California, San Diego, pursuing a dream to apply computer science and engineering towards better understanding and leveraging the fundamentals of biology. I greatly appreciate the opportunities that the university, the department of computer science and engineering and their faculty have offered to enrich this endeavor.

I would like to thank Professor Bernhard Palsson for his support as adviser and the chair of my thesis committee. His ideas, optimism and direction have continuously inspired our broader ambitions for this work. With the Systems Biology Research Group, he has cultivated a group of kind, enthusiastic and driven researchers that continue to impress and innovate.

I would like to thank Professor Vineet Bafna and Professor Pavel Pevzner for their time as committee members of this thesis and for their work as professors to the many courses that have fueled my passions in bioinformatics and systems biology.

I would like to thank Dr. Adam Feist for his mentorship and support. Along with his enormous contributions to this work, he has done everything in his power to enable my success at the University of California, San Diego.

I would like to thank Dennis Gosting; his hard work, initiative and keen insights have greatly enriched the feature set and usability of the products of this thesis.

I would like to thank Dr. Adam Feist, Dr. Ryan LaCroix, Troy Sandberg, Gabriela Guzman, Joon Ho Park, Colton Llyod, Douglas McCloskey and the many others who have contributed their experiment data to enable the work of this thesis.

I would like to thank Dr. Ali Ebrahim, Dr. Ryan LaCroix, Kaiwen Zhang and Dylan Skola; their work lay the foundation for the products of this thesis.

I would like to thank Nadyne Nawar for all of her effort as my master's program

adviser in helping me build a personalized academic curriculum that was aligned with my interests and passions.

This work was supported by a grant from the Novo Nordisk Foundation Center for Biosustainability (NNF16CC0021858).

ABSTRACT OF THE THESIS

ALE Analytics: A Software Pipeline and Web Platform for the Analysis of Microbial Genomic Data from Adaptive Laboratory Evolution Experiments

by

Patrick Phaneuf

Master of Science in Computer Science and Engineering

University of California, San Diego, 2016

Professor Bernhard Palsson, Chair

Adaptive Laboratory Evolution (ALE) methodologies are used for studying microbial adaptive mutations that optimize host metabolism. The Systems Biology Research Group (SBRG) at the University of California, San Diego, has implemented high-throughput ALE experiment automation that enables the group to expand their experimental evolutions to scales previously infeasible with manual workflows. The data generated by the high-throughput automation now requires a post-processing, content management and analysis framework that can operate on the same scale. We developed a software system which solves the SBRG's specific ALE big data to knowledge challenges. The software system

is comprised of a post-processing protocol for quality control, a software framework and database for data consolidation and a web platform named ALE Analytics for report generation and automated key mutation analysis. The automated key mutation analysis is evaluated against published ALE experiment key mutation results from the SBRG and maintains an average recall of 89.6% and an average precision of 71.2%. The consolidation of all ALE experiments into a unified resource has enabled the development of web applications that compare key mutations across multiple experiments. These features find the genomic regions *rph*, *hns/tdk*, *rpoB*, *rpoC* and *pykF* mutated in more than one ALE experiment published by the SBRG. We reason that leveraging this software system relieves the bottleneck in ALE experiment analysis and generates new data mining opportunities for research in understanding system-level mechanisms that govern adaptive evolution.

Chapter 1

Introduction

Adaptive Laboratory Evolution (*ALE*) is a tool for studying biological molecular mechanisms of evolutionary adaptation through coupling with *whole genome sequencing* WGS [8]. Researchers involved in the study of microbial evolution and metabolic engineering use *ALE experiment* methodologies to explore adaptive mutations that optimize system level functions [17]. The *Systems Biology Research Group* (*SBRG*) at the University of California, San Diego, has implemented high-throughput ALE experiment automation that enables the group to expand their experimental evolutions to scales previously infeasible with manual workflows. The data generated by the high-throughput automation now requires a post-processing, content management and analysis framework that can operate on the same scale; in other words, the SBRG's ALE operations need a big data to knowledge solution, a circumstance common in biomedical fields [16]. The ALE big data to knowledge solution described herein is defined by a set of challenges; the goal of the work described by this thesis is to provide solutions for each of these challenges in the form of a cohesive system that can be leveraged for the SBRG's ALE mutation analysis.

Raw data often contains artifacts related to the methodologies used in its acquisition.

During analysis and interpretation, these artifacts can disrupt the process of finding meaningful information. It is therefore crucial to leverage a *post-processing protocol* to report on and refine the relative quality of acquired data. We have developed a quality reporting and control protocol for ALE sample sequencing data that provides stage dependent feedback for experiment data refinement.

Data consolidation consistently challenges efforts in providing comprehensive information on an ALE. High-throughput experimentation exponentially escalates this challenge. This thesis describes a framework we have developed to automate the consolidation of ALE mutation data. Automating consolidation will additionally enable experimentalists to process more samples into an ALE's mutation data set, therefore providing more resolution on an evolution. Consolidation of all experimental data into a unified resource will additionally allow for analysis and research across multiple ALE experiments.

Challenges that coincide with data consolidation are result reporting and accessibility; users require an accessible medium to review the reports on their consolidated ALE data and interpret results. We automate ALE experiment mutation report generation and deployment through a web platform named *ALE Analytics*. This platform makes all ALE experiment reports available for review to researchers and collaborators via web access.

The manual execution of analysis on ALE experiments can be inconsistent between researchers, is prone to human error and is often impractical to re-execute with updated protocols. Leveraging the ALE Analytics platform, we automate common ALE analysis to provide solutions to the challenges of consistency, accuracy and amendment of analysis results. In this thesis, we also evaluate the automated analysis according to a set of ALE experiments whose data and result have been published by the SBRG.

This thesis will therefore describe these challenges, their solutions and the culmina-

tion of their deployment as an ALE big data to knowledge solution. Being that the product of this work is a system meant to service users on a consistent basis through a web interface, we discuss an overview of the ALE Analytics web platform and its features. This thesis will also describe current features that leverage the database of multiple ALE experiments' mutations to identify significant mutations shared across experiments. With the features implemented to address the SBRG's ALE big data to knowledge challenges, it is the goal of the work described by this thesis to enable the post-processing, content management and analysis of ALE experiments at a rate that matches the automated high-throughput ALE experiment execution and ultimately reduce the effort and time necessary to understand system-level genetic mechanisms that govern adaptive evolution.

Chapter 2

Motivation and Specific Aims

2.1 Background

2.1.1 ALE Experiment

The foundational work to which this thesis builds on are the ALE experimental wet lab methodologies. Beginning with a well known starting strain, an ALE experiment is commonly executed by serially passing a selected cell culture to a fresh flask of media (Figure 2.1), enabling the particular strain passed to continue evolving under the experimental conditions. The selection of cells to pass on to the next flask is often based on growth rate since we assume that the most adapted strain population will outgrow their competition. ALE experiments can also involve replicate ALEs, which are identical experimental evolutions that execute in parallel. This approach reveals additional mutational data on the dynamics of adaptation and evolution for an organism and can describe a convergence of adaptations across multiple ALEs [18]. As shown with Figure 2.1, ALE experiments can involve multiple ALEs with multiple *flasks* and *isolates*. In addition, each isolate can have one or

more *technical replicates*, which are used to serialize duplicate isolates with the goal of increasing the quality of a previous isolate. The work described by this thesis serializes each sample with an abbreviation for each ALE layer and a count value representing a sample's identity within an ALE layer's sequence. For example, the serialization *A6 F21 I2 R1* represents the first technical replicate of the second isolate from the 21st flask of the 6th ALE in an ALE experiment. The output of an ALE experiment are populations or clonal

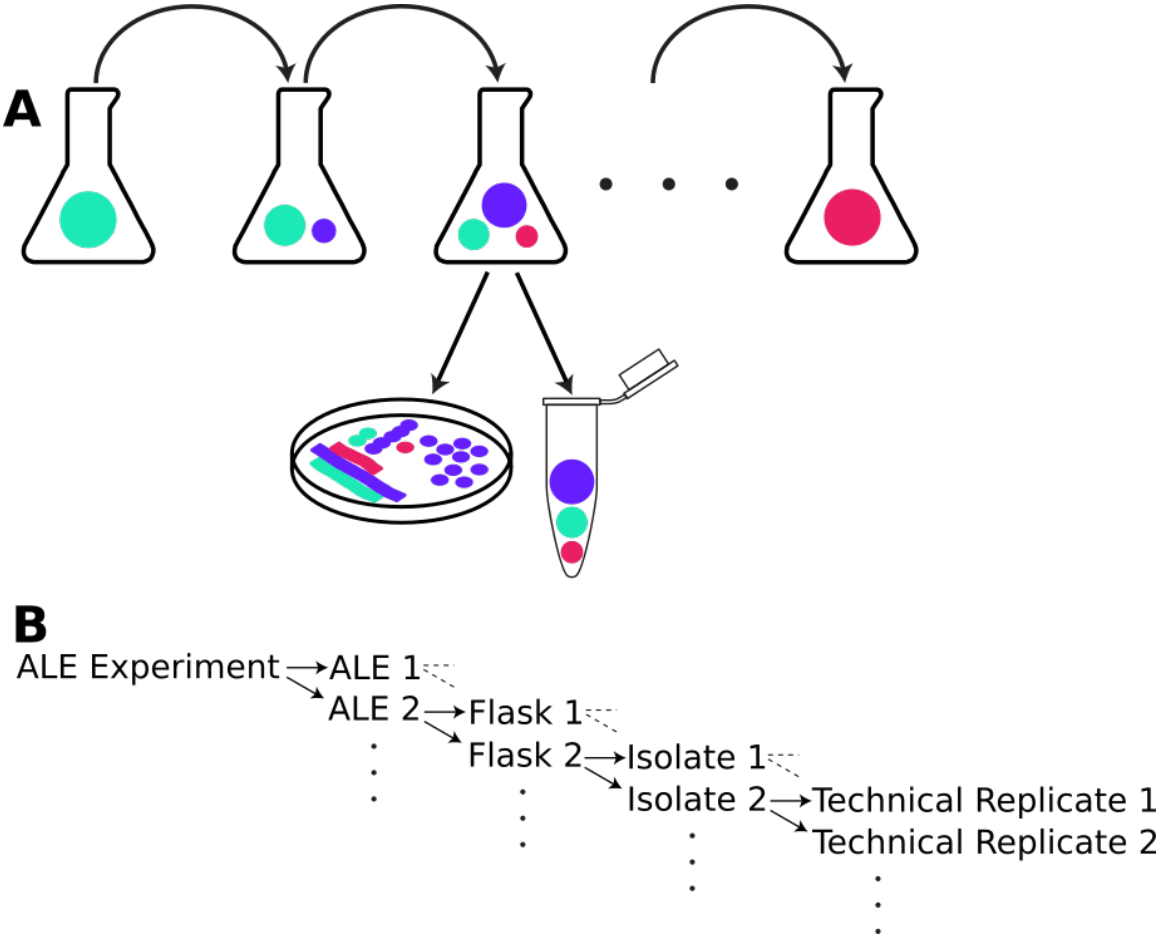


Figure 2.1: **A** An illustration of an ALE from an ALE experiment where both a clonal and population sample are isolated from a midpoint flask. The petri dish illustrates the streaking out of a colony for capturing clonal samples. The Eppendorf tube illustrates a population sample of possible heterogeneous strains. **B** An illustration of how an ALE experiment can have many ALEs, flasks, isolates and technical replicates and how they are associated with each other.

isolates which were derived from the original strain and additionally include the mutations in the cells according to the environmental stresses introduced in the ALE experiment. This output is then processed by additional foundational technologies leveraged with ALE: the DNA sequencer and mutation identifying software. The output sample to be analyzed is sequenced, typically by an Illumina platform, and resequenced primarily using the *Breseq* computational pipeline [9]. Breseq's resequencing process aligns all of the DNA reads to a provided genome reference and reports single-nucleotide mutations, point insertions and deletions, large deletions, missing coverage and new junctions according to the differences between the aligned reads and the reference genome.

Generally, researchers compare the mutations from each ALE's endpoint samples to identify genomic regions with many alleles. If a researcher sequenced ALE midpoint or intermediate samples, an ALE's mutations can be organized in chronological order to identify mutations that fix. The mutations that are involved in multiple alleles of a genomic region or are shown to fix from starting or midpoint to endpoint samples are considered the *key mutation* sets and are investigated for their possible fitness benefit according to the genomic region they affect.

2.1.2 ALE Machine

ALE methodologies are becoming increasingly popular for their potential in revealing novel discoveries on evolution and designing organisms, though their wet lab execution is often labor intensive and requires significant run-time. To impose a balance on the required labor, many ALE experiments are designed to allow for approximately 24 hours between sessions. Executing manual ALE experiments often additionally restricts the possible experimental parameters according to the feasible amount of experiment monitoring.

Current technologies in automation can be leveraged to automate many of the ALE processes, therefore alleviating these restrictions. Automation additionally can contribute to the consistency of the ALE experiments, better ensuring that artifacts in results are due to experimental conditions and not inconsistencies in experiment protocol execution. Finally, automation can enable the scale of an experiment to be greatly expanded at a much lower cost of effort to the experimentalists. This results in the potential for a larger amount of data to be generated on an ALE experiment, providing more resolution on the evolutions. These possible benefits have lead to the Systems Biology Research Group to develop an ALE automation platform, referred to as an *ALE Machine* [15].

The ALE Machine eliminates many of the constraints that manual ALE experiments are subjected to, therefore enabling larger ALE experiments with more consistent data. These outcomes are a boon to those studying evolution and encourage experimentalists to further leverage ALE methodologies to explore the dynamics of evolution.

2.2 Challenges

The potential for more diverse and greater scale ALE experiments enabled by the ALE Machine's automation exacerbates existing challenges and introduces new challenges. The following is an itemization and description of each challenge considered for this thesis.

2.2.1 Post-Processing Protocol

Many tools exist for the quality control of sequencing data and the identification of mutations. Each tool comes with an inherent set of strengths and weaknesses. For the case of the SBRG's ALE operations, consistency in data quality is a primary priority that

has been seen to vary between experiments. Additionally, a tool-set capable of identifying the majority of important genomic artifacts describing the difference between a reference genome and that of an evolved strain is necessary for the analysis of an ALE.

2.2.2 Experiment Data Consolidation Effort

As the scale of an ALE experiment grows, so too does the effort necessary to curate the data of the experiment's samples into a report that describes the ALE experiment's mutation lineages. These manual curations are additionally prone to human error, which is more likely to occur with larger experiments.

2.2.3 Result Accessibility

ALE experiments often involve multiple experimentalists and collaborators. The input of many collaborators may be necessary to fully capture and understand the results of an experiment. These collaborators may be locally or remotely located. The logistics of sharing results is often challenging between local collaborators and more problematic between remote collaborators. This is especially true when comparing multiple experiment results from both current and past ALE experiments.

2.2.4 Experiment Analysis Effort, Consistency and Accuracy

The issues of data consolidation with larger scale experiment also manifest with the analysis of the experiment results. The key mutation analysis of experiments may be inconsistent between investigators. Key mutations may be excluded from the analysis of the results depending on the amount of experience of an ALE experimentalist. Different

methods of reporting experiment results may be used between experimentalists, therefore making the comparison of different ALE experiments more difficult.

2.3 Specific Aims

We have defined a specific set of aims that would lead to solutions to these challenges. The aims are as follows:

2.3.1 Aim 1: Establish a Post-Processing Protocol

To ensure the consistency of quality and format of ALE experiment data, we aim to establish a post-processing protocol. This pipeline will take as input ALE experiment sample sequence data and output reports on each sample's mutation set with regards to a reference genome. Before the mutation reports are generated, reports on the sequencing data quality are generated and inspected; if necessary, measures will be taken to improve the quality of the sequencing data.

2.3.2 Aim 2: Automate ALE Experiment Data Consolidation and Accessible Report Generation

To reduce the effort in generating consistent ALE experiment mutation lineage reports that combine all appropriate ALE experiment samples, we must develop software to automate the parsing and databasing of ALE experiment mutation data. This database will then be leveraged by a reporting application that will produce mutation lineage reports. To address the challenge of results accessibility, the application used to generate ALE experiment reports will be developed as a web application which can be made accessible

to collaborators over the web. This application will be able to present all available ALE experiment mutation and results reports.

2.3.3 Aim 3: Automate Common ALE Experiment Analysis

The common methods of identifying an ALE experiment's key mutations will be defined and automated; their results will be presented by the reporting features alongside the mutation lineage reports.

Chapter 3

Aim 1: Establish a Post-Processing Pipeline

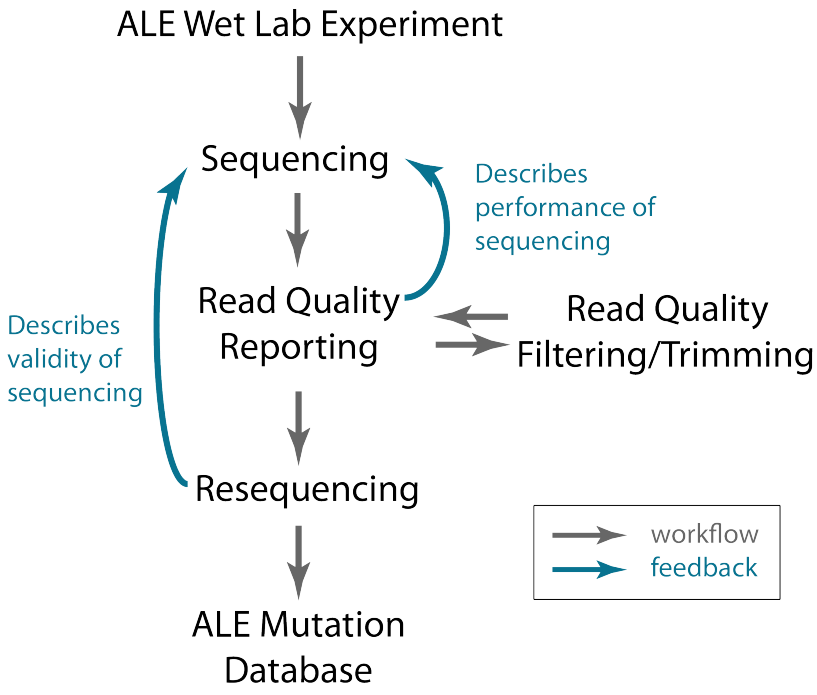


Figure 3.1: The term *resequencing* describes the alignment and variant identification process with given reads relative to a reference genome.

We have established a post-processing protocol for the quality control of the se-

quencing data and to capture each sample's mutations. The protocol uses the following software: *FastQC* [2] for sequenced reads quality analysis, the *FASTX-Toolkit* [13] for trimming sequenced reads, and Breseq [9] to align reads to a reference genome and identify mutations relative to a reference genome.

The protocol first requires that the reads for all sequenced ALE samples are inspected for their *per base sequence quality* and *per base sequence content* using FastQC. The per base sequence quality report presents the cumulative quality score of the bases in specific read positions as having good, reasonable or bad quality. From these results, we can understand which 3' and 5' read end positions have the lowest cumulative quality and trim these positions out using the FASTX-Toolkit. We additionally inspect the per base sequence content for any abnormalities such as the biasing for particular bases in specific positions. In a random set of sequenced reads, you would expect that each read position has approximately equal counts of each base, though in our more recent samples, we often experience a biasing of base types in the first 17 positions and the final 25 positions of each read. At this point, we additionally noticed that many alignment artifacts being identified as mutations in our subsequent variant calling were located in these read regions. This base type biasing is likely due to the sequencing library preparation chemistry that inherits an intrinsic DNA fragmentation sequence bias [10]. Though this bias does occur, it should theoretically not result in inappropriate base calling. Even so, trimming these biased regions out of our reads has lead to less mutation calling due to sporadic non-consensus bases. The process of quality inspection and trimming may iterate multiple times before moving on. Once the reads report acceptable quality, Breseq is used to align each sample's reads to a reference genome and identify mutations. In addition to mutations, Breseq reports return the alignment statistics of *mean coverage* and *mapped read* count. In combination with the

amount of *unassigned missing coverage* evidence, regions where no reads could be aligned to the reference, these statistics are used to evaluation the alignment performance, where we compare each statistic and unassigned missing coverage artifact count to an empirically derived threshold. If the mean coverage or mapped read values fall below our thresholds or the unassigned missing coverage artifact count exceeds our thresholds, we consider the sequencing library of the sample they belong to as potentially problematic in our analysis of the evolution and may again sequence the sample or discard them from the analysis.

Besides the actual generation of mutation data, the key functionality of our post-processing protocol is the stage dependent feedback provided. This feedback offers the opportunity for experimental protocol and data acquisition refinement that can be acted on by the many individuals involved in the sample preparation and processing pipeline.

Chapter 4

Aim 2: Automate ALE Experiment Data Consolidation and Generate Accessible Reports

The products of this thesis have been primarily driven by the need to consolidate and report on large amounts of ALE data in such a way as to describe adaptive mutations in evolved strains. We have done so by leveraging a full stack of industry standard technologies that enable the parsing and databasing of experiment data and the generation of reports on said data.

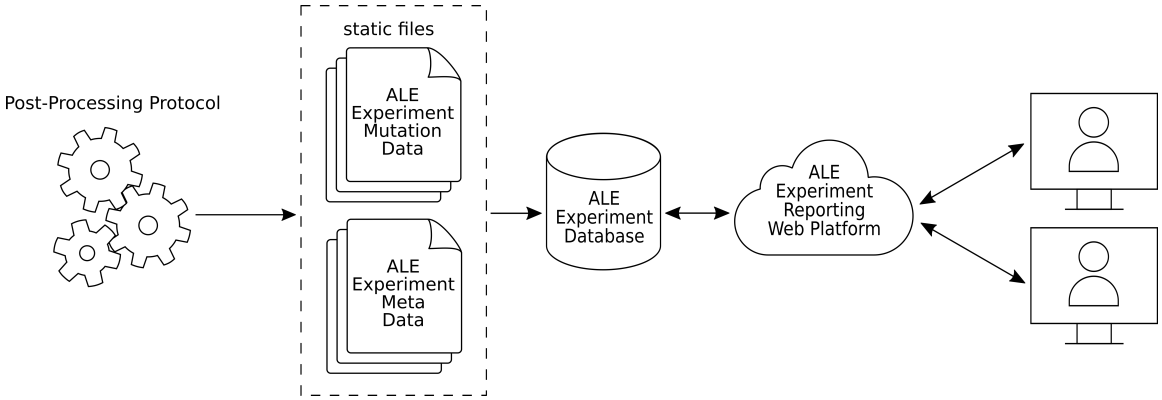


Figure 4.1: An illustration of the flow of ALE experiment data to the deployment of result report generation for end users.

The automation manifests as the programmatic parsing of the mutation reports generated by Breseq for all ALE experiment samples and with the mutational data subsequently loaded into an ALE mutation database. During the initiation of an ALE experiment, experimentalists generate meta data which can be additionally parsed and uploaded to the ALE mutation database by the automated parsing of ALE experiment data. This process consolidates all experiment mutation data and meta data into one resource. This database is used for all ALE experiments and therefore has resulted in all available ALE experiments being consolidated into one database, enabling further cross-experiment analysis.

The ALE mutation database is leveraged by a web application to automate the generation of reports for each ALE experiment that describe the mutational lineages of their ALEs. This application is deployed on the web, allowing for ubiquitous accessibility to ALE experiment reports for all experimentalists and collaborators. We have named this web application ALE Analytics.

4.1 ALE Analytics Web Platform

The ALE Analytics web platform enables and automates much of the analysis necessary for interpreting ALE experiment mutational data. For this thesis, we primarily describe the automated mutation analysis and reporting features, though this platform additionally includes ALE *mutation filtering*, *mutation search*, *ALE experiment export* to CSV, *ALE experiment comparison*, sample sequencing data alignment *quality statistics* and *mutation database overview*. The fundamental focus of the ALE Analytics web platform is report generation. Each ALE experiment can be described as a series of samples which contain both new mutations and mutations from ancestors relative to a reference genome. Ordering the samples as columns from earliest to latest in an ALE, where each row describes the manifestation of a specific mutation among samples, can serve as a visualization to grant intuition on mutational trends. The occurrence of a mutation in a sample is annotated as a value between 0 and 1 within the cell of a mutation row for the sample. The annotated value represents the estimated frequency of this mutation among the sample population [9]. Among the many mutations that manifest within an ALE experiment, mutation rows that describe the alleles of a gene in the ALE experiment will cluster together due to the sorting of mutations according to their positions on the genome. Due to the chronological sorting of the sample columns per ALE, a mutation that fixes across samples will manifest as a sequence of cells in a mutation row annotated with the manifestation of that mutation. These patterns are obvious to an observer and serve well to describe the mutation trends in an ALE experiment.

Position	Mutation Type	Sequence Change	Gene	Protein change	GLU A4 F66 I1 R1	GLU A4 F149 I1 R1	GLU A4 F237 I1 R1	GLU A4 F403 I1 R1
1,292,255	MOB	IS1 (–) +8 bp	hns, tdk	intergenic (-110/-488)		1.00	1.00	1.00
1,551,658	SNP	G→A	adhP	P69S (CCA→TCA)		1.00		
2,130,811	SNP	A→C	wcaA	I204S (ATC→AGC)			1.00	
3,999,668	DEL	Δ5 bp	corA	coding (220-224/951 nt)		1.00	1.00	1.00
4,000,174	DEL	Δ3 bp	corA	coding (726-728/951 nt)	1.00			

Figure 4.2: Illustration of a mutation lineage report. The *hns, tdk* intergenic mutation and a *corA* mutation can be seen to fix in samples over time. The multiple mutation rows describing the *corA* alleles cluster together.

Before the ALE Analytics platform was deployed for use, experimentalists would work to identify ALE experiment *key mutations* by manually annotating all ALE experiment sample mutations within a spreadsheet and investigate for significant mutation patterns. The task of manually annotating mutations proved to be time consuming and error-prone, causing delay of ALE experiment result interpretation. This initial bottleneck in the post-processing of ALE experiment results lead to the efforts that began the ALE Analytics project. The time cost of this manual curation additionally set an implicit limit on how many samples experimentalists were willing to sequence and process into their output data set, therefore potentially limiting the resolution on an ALE experiment's mutational trends. Additionally, the type of sample obtained from a flask was limited by the cost of manual curation according to the amount of mutations annotated per sample, where population samples can include an order of magnitude more mutations than clonal samples; this can therefore cause an experimentalist to limit the inclusion of population samples into an ALE experiment's output data set. The primary function of the ALE Analytics platform is to automate the annotation and reporting of all ALE experiment sample mutations in a single location with a uniform format. This automated functionality enables experimentalists to more quickly

interpret results and include more sequenced samples into their mutation set, without fear of the effort necessary in processing and consolidating their experiment's mutational data. Alongside the ALE Machine, ALE Analytics further enables more resolution on evolution. This automation may additionally result in more consistent reporting and accurate results since it removes the potential for human error in data processing, consolidation and report generation.

Since August 2015, the ALE Analytics web application has deployed a production version for experimentalists and collaborators to use in the analysis of their ALE experiments. Though we have many development versions that are launched and shutdown for testing and demonstrating new features to target audience, our production version of ALE Analytics has constant uptime and is available online for access by local and remote experimentalists and collaborators. To end-users, ALE Analytics represents a web platform and database of experimental evolution results that is ubiquitously available and ever growing in features and content.

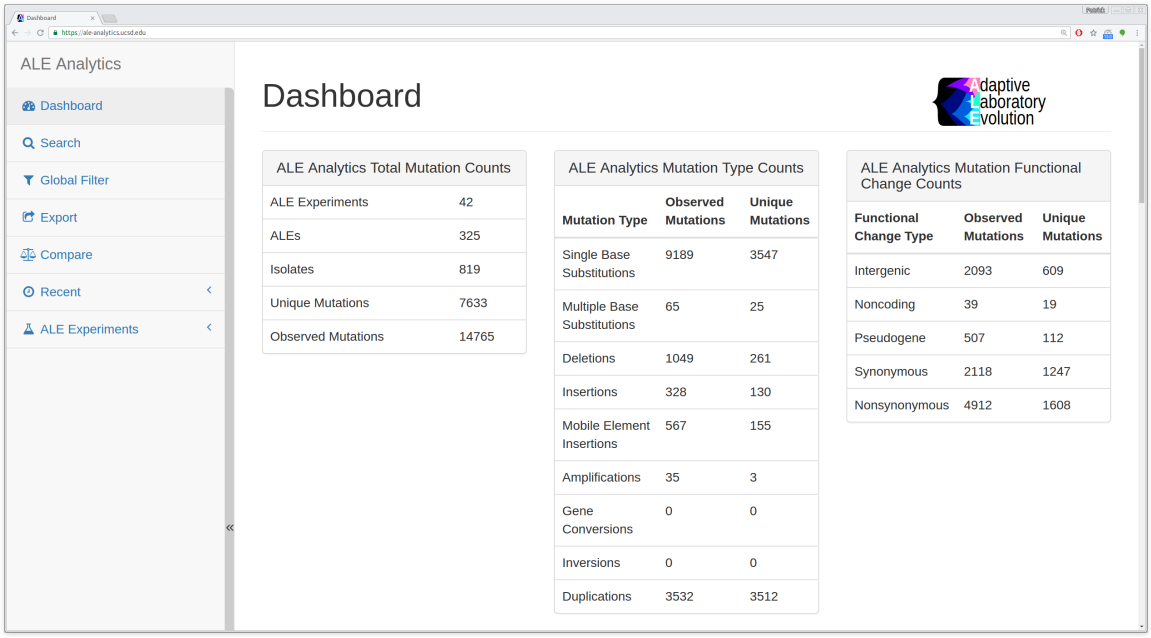


Figure 4.3: Screenshot of the ALE Analytics production version dashboard containing counts of significant ALE mutation database details.

The current production deployment of ALE Analytics contains 14,765 observed mutations and 7,633 unique mutations. These mutations are housed in 42 separate ALE experiments with a total of 325 ALEs and 819 isolates or samples. This large volume of mutations and samples is evidence to how the the automation of consolidating ALE experiment data has been leveraged to create a substantial resource on experimental evolutions. Being that ALE Analytics is being leveraged in investigating current experiments, full access hasn't yet been made available to the public. Our intent is to ultimately make public on ALE Analytics the results of the ALE experiments that have been published.

I would like to acknowledge the work accomplished by Dr. Ali Ebrahim, Dr. Ryan LaCroix, Kaiwen Zhang and Dylan Skola for their early prototyping on what we now know as ALE Analytics.

I would like to acknowledge Dennis Gosting for his work in consolidating ALE experiment data into the ALE mutation database.

Chapter 5

Aim 3: Automate Common ALE Experiment Analysis

Using the ALE Analytics mutation reporting mechanism, we have implemented features that automate the finding and reporting of mutations which describe significant mutational patterns within an ALE experiment. These features describe the significant mutation patterns as *enrichment* and *fixed* mutations. The methodologies encapsulated in the enrichment and fixed mutation analysis are those which have been developed to manually identify key mutations within the results of published ALE experiments ([5], [21], [20], [14]) and are therefore considered common ALE experiment key mutation analysis. As with the manual consolidation of ALE experiment mutational data, the analysis of key mutations can be prone to human error, inconsistent between researchers and time consuming. The automation of these common analyses will contribute to more accurate results, more consistent analysis and the shortening of turnaround time from ALE experiment execution to results interpretation.

5.1 Enrichment Mutation Analysis

Early ALE experiments only sequenced ALE endpoint samples. Key mutations were identified as those involved in the following two cases. If a mutation manifested in more than one sample, it signified that this mutation was likely correlated to the dominant phenotype. If a genomic region was mutated via different mutations across multiple samples, it was hypothesized that simply perturbing the gene may have rendered a benefit to the phenotype. Considering that these circumstances all involve populations with the same mutated genomic region, we can also consider a single sample with multiple alleles of the same region to represent multiple populations which benefit from this region's perturbation. The identification of these cases can be accomplished by finding genomic regions within an ALE experiment that have more than one observed mutation; we consider these enriched genomic regions and their mutations as enrichment key mutations. Previous to ALE Analytics, this analysis was accomplished as additional steps in manually curating the ALE experiment's mutation report such that mutations affecting the same genomic region would cluster together within a matrix of mutations and their samples. The *enrichment mutation* analysis automates this approach and reports the mutations in the same mutation reporting format given in Figure 4.2.

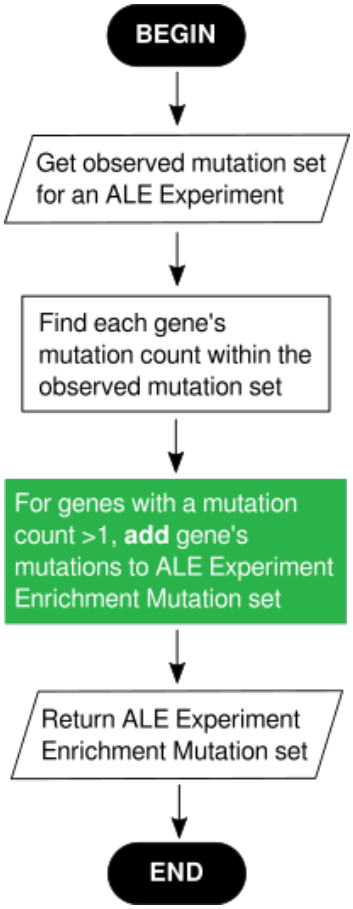


Figure 5.1: Enrichment mutation analysis flowchart

5.2 Fixed Mutation Analysis

A fixed mutation is one in which a mutation manifests in an ALE's midpoint, or intermediate sample, and is propagated to all following samples in the ALE. Fixed mutations are the strongest indicators of key mutations. This analysis is possible if an ALE experiment includes midpoint samples, so as to provide for the possibility of more than one data point per mutation. Previous to ALE Analytics, the identification of Fixed mutations was accomplished as an additional curation step in results reporting by manually organizing mutations according to their ALE's sample chronology and subsequently identifying mutations that

emerge in a midpoint and manifest in all remaining ALE samples. The *fixed mutation* analysis automates this approach and reports the mutations in the same mutation reporting format given in Figure 4.2.

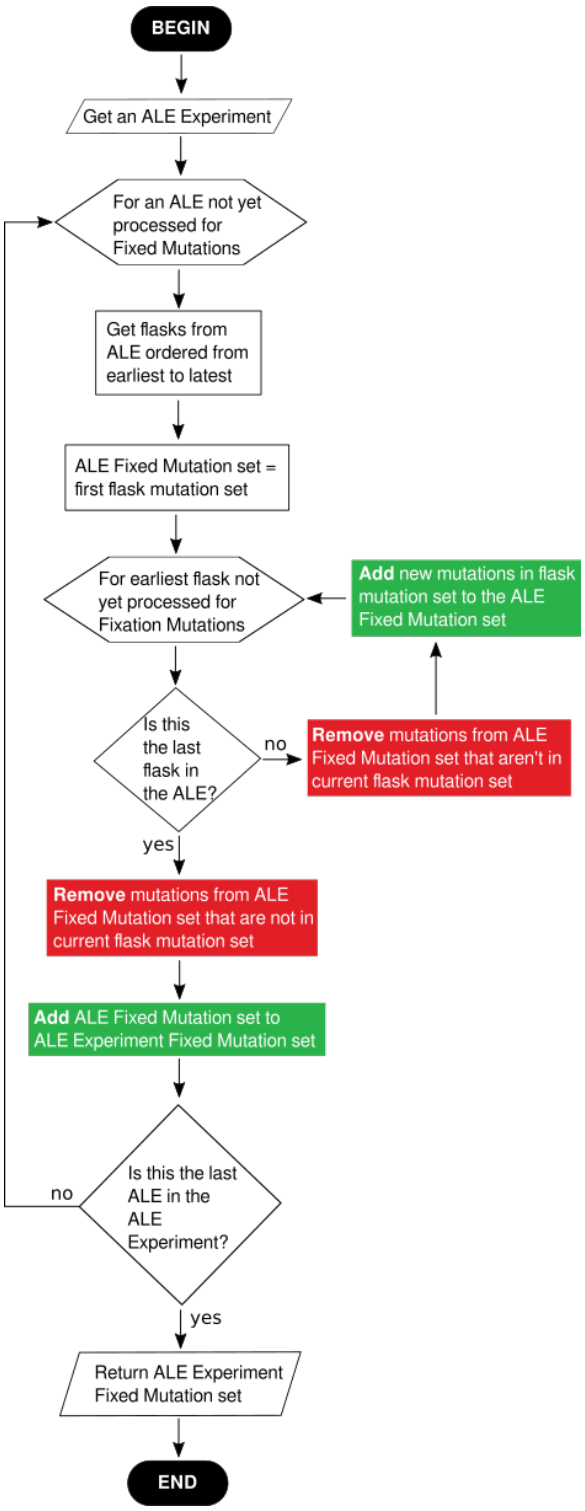


Figure 5.2: Fixed mutation analysis flowchart

Ideally, fixed mutations should have frequencies equal to or larger than their preced-

ing occurrences within an ALE. This trend demonstrates the most obvious key mutations within a ALE experiment. The fixed mutation analysis has a filtering option that results in only those fixed mutations with equal-to or ascending population frequency trends to be returned. This feature is referred to as the *ascending frequency* fixed mutation analysis and reports these mutations in the same mutation report format given in Figure 4.2.

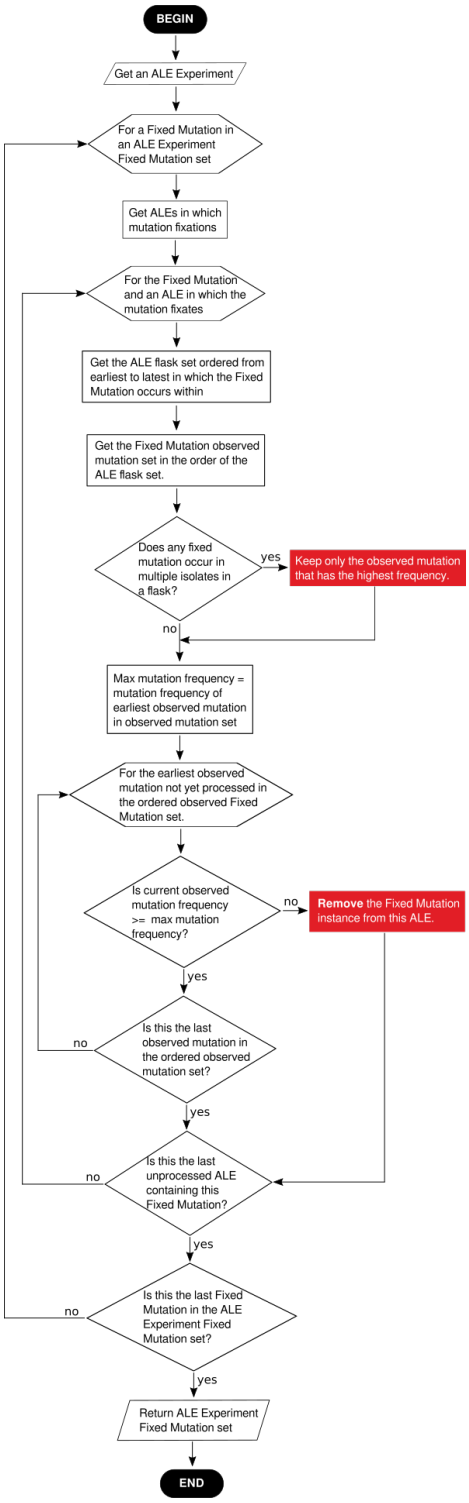


Figure 5.3: Ascending frequency fixed mutation analysis flowchart

5.3 Results

To ensure that our automated key mutation analysis returns key mutations based on published criteria, we evaluate the automated analysis results against those found in ALE experiments published by the SBRG. The automated analysis key mutations are generated by processing the sample reads available from these published ALE experiments through the post-processing protocol, ALE Analytics platform and key mutation automated analysis. Intuitively, our task can be described as finding all of the key mutations within the complete set of mutations in an ALE experiment and comparing the set for equality against the published key mutation set of the same experiment. The post-processing protocol will be configured to reproduce, as well as possible, the mutation set used by the authors in defining the published key mutations by using the same reference genome and Breseq version. This does not guarantee that the mutation set produced will be identical to those used by the authors and therefore introduces the issue of having different starting conditions between key mutation analysis.

We find a more practical comparison for our automated key mutation analysis evaluation to avoid issues due to possible different ALE experiment mutation sets between the published material and those generated by the post-processing protocol. Clarification of gene function is a primary result in ALE experiments. Organisms studied in ALE experiments are selected for their phenotype which is derived from their genotype. For ALE experiments, the areas of highest interest in a genotype are those genomic regions perturbed by mutations. With ALE experiment mutation sets, experimentalists investigate the known functions of the perturbed genomic regions along with the phenotypic results of the perturbation and infer additional functions of genomic regions. We can therefore reason that a better evaluation between the results of the published ALE experiments and the ALE

Analytics automated analysis is a comparison of the genomic regions found to be perturbed by key mutations. We discuss any false positives and false negatives and determine why they manifest and what can be implemented to handle their specific cases. The following is the set of publications that we use in our evaluation and their ALE experiment names:

- **The PGI ALE experiment**: *Genetic Basis of Growth Adaptation of Escherichia coli after Deletion of pgi, a Major Metabolic Gene.* [5]
- **The 42C ALE experiment**: *Evolution of Escherichia coli to 42 ÂřC and Subsequent Genetic Engineering Reveals Adaptive Mechanisms and Novel Mutations. Molecular Biology and Evolution.* [21]
- **The *13C* ALE experiment**: *Evolution of E. coli on [U-13C]Glucose Reveals a Negligible Isotopic Influence on Metabolism and Physiology.* [20]
- **The *GLU* ALE experiment**: *Use of Adaptive Laboratory Evolution To Discover Key Mutations Enabling Rapid Growth of Escherichia coli K-12 MG1655 on Glucose Minimal Medium.* [14]

5.4 Evaluation

In this section we describe our approach to formally evaluating the performance of our key mutation automate analysis. The task of finding key mutations is a binary classification problem, where a mutation is or is not a key mutation. To accomplish binary classification on a set of mutations, we have defined a set of features that describe how a mutation should be classified: the enrichment mutation feature and the fixed mutation feature. If a mutation qualifies for either feature, we will classify this mutation as a key mutation.

Table 5.1: The key mutation count is obtained from the ALE experiment published materials and the insignificant mutation count is obtained by finding the difference between the key mutation count and the total mutation count from our variant finding results of an ALE experiment. We use our variant finding total mutation count for consistency since some published ALE experiments didn't include the total set of observed mutations in their supplementary material

ALE Experiment	Total observed mutation count	Published *key mutation* count	Insignificant mutation count
PGI	58	26	32
42C	204	50	154
C13	50	13	37
GLU	281	27	254

In establishing our classification evaluation metrics, we observe that there are many more insignificant mutations than key mutations, or in other words, there exists a class imbalance. We demonstrate this imbalance in Table 5.1. All but the PGI ALE experiment have an obvious class imbalance. Relying on accuracy alone to measure classification performance with a class imbalance can be misleading since one could naively classify all mutations as the majority class of the imbalance and still obtain a high accuracy. We consider avoiding the exclusion of significant mutations from our results as our highest priority; in other words, we consider *recall* as our primary performance metric. In performing with high recall, we can better ensure that all key mutations are returned in our results, therefore providing the best identification of all significant genomic regions in an experimental evolution. Avoiding insignificant mutations in our results is of secondary importance since they can be later excluded by secondary manual investigations performed by the experimentalist. We will use *precision* as the metric to evaluate our approach's ability to avoid insignificant mutations in our results.

PGI ALE Experiment

The PGI ALE experiment and publication focused on the observed adaptive mutations selected for during an experimental evolution due to knocking out the phosphoglucose isomerase PGI gene from the K-12 MG1655 strain of *E. coli*. The PGI gene plays a major role in the central metabolism of *E. coli* and is a good candidate for studying how a strain learns to compensate for a loss of such significant impact. The experiment executed 10 parallel ALEs on replicates of *E. coli* K-12 MG1655 lacking the PGI gene using ALE methodologies for 50 days in minimal media. Clones were taken from each ALE's final flasks and their mutations were defined by first whole-genome sequencing through both Nimblegen hybridization-base tiling arrays and Illumina Solexa technologies, then leveraging the Nimblegen's built-in variant calling capabilities and in-house software. PCR and Sanger sequencing was additionally used to verify mutations identified and the entire sequence of host genomic regions [5]. Our analysis of the PGI ALE experiment samples was executed by processing the same reads generated by the authors using the Breseq 0.23 software pipeline against the *E. coli* K-12 MG1655 reference genome (NCBI accession NC_000913.2).

Table 5.2: PGI ALE experiment key mutation genomic region matching summary between the paper and the ALE Analytics automated enrichment key mutation analysis.

Genomic Region	Paper	ALE Analytics Erichment
rpoA	X	X
rpoS	X	X
udhA (sthA)	X	X
pntA	X	X
pntB	X	X
cpxR	X	X
icd	X	X
rpoB	X	
rpoC	X	
e14 prophage	X	
cyaA	X	
fabZ	X	
trxB, lrp		X
crr		X
pyrE, rph		X

The PGI ALE experiment paper doesn't explicitly list the key mutations as does the other ALE experiment papers involved in our evaluation. We therefore rationalize the paper's key mutations according to the findings presented on each mutation and their host genomic region. The PGI experiment publication identified the perturbation of the following genes as being important for the experiment's observed fitness:

- *rpoS*: mutations observed suppresses stress response by modulating transcription; likely fitness selection in relation to the adaptation to loss of PGI.
- *rpoA*, *rpoB*, *rpoC*, *cpxR* and *cyaA*: mutations observed result in global network-level transcriptional regulation adaptations.
- *sthA*, *pntA*, *pntB*: mutations observed counter the redox imbalance of excess NADPH

production due to PGI knockout.

- *e14 prophage deletion*: target and mutations observed are mechanistically unknown, yet experimental data shown to provide fitness.

The authors speculate that the deletion of the *e14 prophage* is a unique contributor to the fitness of its host strain. This was determined due to not being able to reproduce the fitness of the host strain with any combination of other key mutations in followup ALEs. The *e14 prophage* deletion manifests as unassigned missing coverage in our mutation reports, which we include in our ALE Analytics database but currently do not yet include in reports. The authors do mention the potential importance of the *icd* SNPs coinciding with this deletion. It is speculated that these SNPs may have a fitness benefit in their potential to induce better translation efficiency to *icd*. Due to the high frequency of these *icd* SNPs, the enrichment analysis did include them in its results as key mutations. Though no solid evidence could be found on the *icd* allele's fitness benefit, since the authors speculate that the *icd* mutations are significant and the objective of the ALE Analytics automated analysis is to highlight the mutation that may be a result of the ALE's selective pressures, we consider the paper's *icd* SNPs to be key mutations.

Of the 12 key mutations alleles published in the paper, the enrichment analysis is successful in finding 7. The ALE Analytics enrichment analysis finds multiple *icd* alleles in both ALE 1 and ALE 5. *icd* alleles were reported to be associated with the *e14 prophage* deletion, published as occurring in the ALE 1. The evidence of high *icd* enrichment in ALE 5 motivated an investigation into the possibility of an ALE 5 *e14 prophage* deletion that was in fact found by manually investigating the missing coverage artifacts of the Breseq reports. This investigation serves to demonstrate the importance of automating key mutation identification for ALE experiments.

A minority of the published key mutations which didn't manifest more than once were included in the paper's key mutation set according to their functional association with significantly enriched genes. SNPs uniquely affecting transcription modulation genes *cyaA*, *rpoB*, and *rpoC* were included as key mutations in the paper in addition to the mutations of the more frequently mutated *rpoA*, *rpoS* and *cpxR* genes. These key mutations could be included in future enrichment analysis implementations by additionally considering the mutational enrichment of functional groups rather than only single genes.

The *fabZ* SNP was speculated as important by the authors due to their knowledge of the potential metabolic perturbations caused by the PGI knockout and the idea that the this mutation may reduce its impact. This type of key mutation could be included in future enrichment analysis by including functional data on any type of perturbation introduced into the initial strain of the experimental evolution and identify mutations affecting genes functionally related to the initial perturbation. This approach would therefore also leverage the enrichment of functional groups of genes.

Published mutations in the *rep*, *yfeH*, *fruK*, *rodA*, *bipA* and *ispU* genes only manifest once and weren't discussed by the publication as being key mutations. These mutations were therefore not considered for either the publication's significant mutation set and the ALE Analytics enrichment analysis evaluation.

The ALE Analytics enrichment mutation analysis found 3 additional possible significant genomic regions described in Table 5.3 due to their frequency of mutation. Though these new enriched genomic regions contributed to the lessening of the enrichment analysis' precision metric, they present an opportunity for identifying further significant adaptations not caught by the authors.

Table 5.3: The value 1 used to denote the presence of a mutation describes the approximate frequency in which the mutation was found within the sample population represented in the sample reads [9].

Position	Mutation Type	Sequence Change	Gene	Protein change	PGI A2 F50 I1 R1	PGI A4 F50 I1 R1	PGI A5 F50 I1 R1	PGI A6 F50 I1 R1	PGI A7 F50 I1 R1	PGI A10 F50 I1 R1
931,808	SNP	G→A	trxB, lrp	intergenic (-535/-10)			1			
931,811	SNP	A→C	trxB, lrp	intergenic (-538/-7)		1				
2,534,334	MOB	Δ1 :: IS186 (–) +6 bp :: Δ1	crr	coding (479-484/510 nt)	1					1
2,534,334	MOB	Δ1 :: IS186 (+) +6 bp :: Δ1	crr	coding (479-484/510 nt)		1	1		1	
3,813,824	DEL	Δ1 bp	pyrE, rph	intergenic (-33/+62)				1		
3,813,832	DEL	Δ1 bp	pyrE, rph	intergenic (-41/+54)	1					

No fixed mutations can be established with the PGI ALE experiment data set since all samples are endpoints of different ALEs and therefore do not provide any mutation time-course information for the fixed mutation analysis to work with.

Table 5.4: PGI ALE experiment classification performance.

True Positive	False Positive	False Negative	Recall	Precision
7	3	5	0.583	0.700

42C ALE Experiment

The 42C ALE experiment and publication focused on the observed adaptive mutations selected for during an experimental evolution with a selective pressure of a culture temperature of 42°C. The experiment executed 10 parallel ALEs on replicates of *E. coli* K-12 MG1655 for 45 days in minimal media. Clones were taken from each ALE's final flask, sequenced using the Illumina MiSeq platform and their mutations defined using whole genome re-sequencing with the Breseq 0.22 software pipeline against the *E. coli* K-12 MG1655 reference genome (NCBI accession NC_000913.2) [21].

The authors of the 42C paper considered key mutations as those that perturbed a gene in more than one ALE endpoint. Their key mutation results were clearly annotated in a Table 2 [21]. This ALE experiment experienced two different hypermutator strains, proposed as independently manifesting in ALE 2, where *mutL* was mutated, and ALE 6, where *dnaQ* was mutated. The ALE 2 hypermutator strain went on to contaminate ALE 3 and the hypermutator strain in ALE 6 went on to contaminate ALE 8. Due to this contamination, the authors didn't consider mutations recurring between ALE endpoints derived from the same hypermutator strain as key mutations. The authors recognized unique mutations occurring within the same genomic region within and between hypermutator strain pairs and therefore did not completely disregard the hypermutator samples.

Table 5.5: The 42C ALE experiment key mutation genomic region matching summary between the paper and the ALE Analytics automated key mutation analysis.

Genomic Region	Paper	ALE Analytics Enrichment
secD	X	X
nagC	X	X
nagA	X	X
rne	X	X
hns, tdk	X	X
ydhZ, pykF	X	X
pykF	X	X
yfdI (gtrS)	X	X
ygaH, mprA	X	X
mlaE	X	X
dinQ, arsR	X	X
rph	X	X
ilvL, ilvX	X	X
rpoC	X	X
hfq	X	X
hrpB		X
frmR, yaiO		X
ybfK, kdpE		X
ymfE		X
abgB		X
ynaE, ttcC		X
ydcD		X
dmsD, clcB		X
ydgC		X
araG		X
yeeP, flu		X
yehC		X
yehQ		X
yffP, yffQ		X
yffS		X
ygcB, cysH		X
yhhZ, yrhA		X
tisB, emrD		X
wecC		X
pgi, yjbE		X
yjiT		X
yjjI		X

The ALE Analytics enrichment key mutation analysis finds all 14 of the genomic regions affected by the published key mutations. The ALE Analytics enrichment analysis is successful in finding all key mutations because its implementation is partly based on the key

mutation protocol establish by this paper.

Table 5.6: New enrichment key mutations found by the automated analysis. ALE endpoint pair (2, 3) and (6, 8) are each derived from the same hypermutator strain, therefore explaining the large amount of mutations shared between the endpoints. The value of 1 used to denote the presence of a mutation describes the approximate frequency in which the mutation was found within the sample population represented in the sample reads [9].

Position	Mutation Type	Sequence Change	Gene	Protein change	42C A2 F163 I1 R1	42C A3 F120 I1 R1	42C A6 F164 I1 R1	42C A8 F164 I1 R1
162,973	SNP	G→A	hrpB	C290Y (TGT→TAT)			1	1
379,237	DEL	Δ1 bp	frmR, yaiO	intergenic (-132/+56)	1	1		
720,169	SNP	C→T	ybfK, kdpE	intergenic (+106/+110)			1	1
1,196,962	SNP	A→G	ymfE	S167P (TCC→CCC)			1	1
1,399,868	SNP	A→G	abgB	I471T (ATC→ACC)			1	1
1,432,483	SNP	C→T	ynaE, ttcC	intergenic (-235/+499)			1	1
1,528,093	INS	+A	ydcD	coding (148/483 nt)			1	1
1,663,212	INS	+T	dmsD, clcB	intergenic (+68/-127)			1	1
1,679,956	SNP	T→C	ydgC	P33P (CCA→CCG)			1	1
1,981,785	SNP	C→T	araG	E437K (GAA→AAA)			1	1
2,069,345	SNP	A→G	yeeP, flu	intergenic (+110/-218)			1	1
2,189,454	SNP	C→T	yehC	G72S (GGC→AGC)			1	1
2,208,833	SNP	C→T	yehQ	pseudogene (1712/2001 nt)			1	1
2,561,535	INS	+A	yffP, yffQ	intergenic (+396/-79)			1	1
2,562,547	SNP	G→A	yffS	M1I (ATG→ATA)			1	1
2,885,374	SNP	T→A	ygcB, cysH	intergenic (-133/+226)			1	1
3,580,229	DEL	Δ1,222 bp	[yhhZ], yrhA	IS1-mediated	1	1		
3,851,932	SNP	A→G	tisB, emrD	intergenic (+267/-13)			1	1
3,969,713	SNP	A→C	wecC	Q144P (CAG→CCG)			1	1
4,233,708	INS	+A	pgi, yjbE	intergenic (+278/-221)			1	1
4,570,302	SNP	A→T	yjiS, yjiT	intergenic (+364/-135)			1	
4,571,551	SNP	A→G	yjiT	pseudogene (1115/1503 nt)		1		
4,613,882	SNP	T→G	yjjI	T403P (ACC→CCC)			1	1

The ALE Analytics enrichment mutation analysis finds 22 additional genomic regions affected by more than one mutation. The *mutL* and *dnaA* mutations mentioned by the paper as the cause for ALE 2, 3, 6 and 8 endpoints to become hypermutators were included in the automated analysis' enrichment key mutations. The paper does not include these in their table of key mutations, though do describe their importance; we therefore do

not consider the *mutL* and *dnaA* mutations as false positives.

These 22 additional genomic regions of interest only manifest in the ALEs that have been identified by the authors as being overcome by the same hypermutator strain except for the *yjiT* alleles. *yjiT* was mutated in the ALE 3 and 6 endpoints, hypermutators of separate origins, and therefore make its mutations candidate for classification as key mutations. The exclusion of the *yjiT* mutations in the author's key mutation set may have been an oversight in their manual workflow; the automated enrichment key mutation analysis will reduce the possibility for these errors with its future usage. If our evaluation were to exclude hypermutator mutations from both the published significant mutation set and our results, both the subsets of the published non-hypermutator and automated analysis key mutations would match without any additional key mutations.

The obvious contamination made clear by the amount of hypermutator mutations in contaminated strains lends us intuition on how to automate the identification of contamination among samples. Additionally, one could automate the identification of contamination among samples by recognizing when samples share a large subset of the exact same point mutations.

No fixed mutations can be established with the 42C ALE experiment data set since all samples are endpoints of different ALEs and therefore do not provide any mutation time-course information for the fixed mutation analysis to process.

Table 5.7: The 42C ALE experiment classification performance.

Hypermutators Included	True Positive	False Positive	False Negative	Recall	Precision
Yes	14	22	0	1	0.389
No	14	0	0	1	1

C13 ALE Experiment

The C13 ALE experiment and publication focused on the observed adaptive mutations selected for during an experimental evolution using ^{13}C-glucose as a carbon source for *E. coli* growth. The key mutations found by this experiment were compared to those of [14], which uses ^{12}C-glucose, to investigate if there is any evidence of differing adaptations and therefore additional metabolic stress from using the ^{13}C-glucose isotope. The experiment executed 6 parallel ALEs on replicates of *E. coli* K-12 MG1655 for approximately 1000 generations per ALE. Two clonal samples were taken from each ALE; one sample served as a midpoint clone and the other as the endpoint clone. The clones were sequenced using the Illumina MiSeq platform and their mutations defined using whole genome re-sequencing with the Breseq 0.23 software pipeline against the *E. coli* K-12 MG1655 reference genome (NCBI accession NC_000913.2).

The authors of the C13 published material considered the mutations for genes that were enriched in more than one endpoint to be key mutations. Their key mutation results were clearly stated as those mutations which affected the *pyrE/rph*, *rpoB*, *hns/tdk* and *rhsE* genomic regions. The ALE Analytics enrichment key mutation analysis finds all 4 of the published genomic regions affected by key mutations.

Table 5.8: The value of 1 used to denote the presence of a mutation describes the approximate frequency in which the mutation was found within the sample population represented in the sample reads [9]

Position	Mutation Type	Sequence Change	Gene	Protein change	C13 A6 F58 I1 R1	C13 A6 F133 I1 R1
4,183,563	SNP	C→T	rpoC	P64L (CCG→CTG)	1	
4,184,121	INS	+9 bp	rpoC	coding (749/4224 nt)		1

The ALE Analytics enrichment mutation analysis found 1 additional possible significantly enriched allele described in Table 5.8: *rpoC*. The authors did not include *rpoC* mutations as a key mutation since their approach was limited to high frequency alleles among different ALEs. The ALE Analytics enrichment analysis additionally returns mutations found to affect genomic regions within multiple samples of the same ALE.

The ALE Analytics fixed mutation analysis finds 2 of the 4 published key mutation alleles; those key mutations missed were due to only manifesting in ALE endpoint samples. The ALE Analytics ascending frequency fixed mutation analysis finds the same results. If the C13 ALE experiment had more midpoint samples, the fixing of these mutations may be more evident and would ultimately be captured by the fixed key mutation analysis. Significant mutations should ultimately be all caught by the fixed key mutation analysis with enough samples, though the question that remains is how many samples from an ALE are adequate to provide enough resolution on mutation lineages to capture all fixing mutations.

Table 5.9: The C13 ALE experiment key mutation genomic region matching summary between the paper and the ALE Analytics automated key mutation analyses.

Genomic Region	Paper	ALE Analytics Enrichment	ALE Analytics Fixed
pyrE, rph	X	X	X
rpoB	X	X	X
hns, tdk	X	X	
rhsE	X	X	
rpoC		X	

Table 5.10: The C13 ALE experiment classification performance.

Analysis	True Positive	False Positive	False Negative	Recall	Precision
Enrichment	4	1	0	1	0.8
Fixed	2	0	2	0.5	1

GLU ALE Experiment

The GLU ALE experiment and publication focused on establishing and leveraging novel ALE methods and observing the adaptive mutations selected for using these ALE methods and *E. coli* on glucose minimal media at 37°C. This experiment isolated the selection pressure to the growth rate of the strain by propagating batch cultures to new flasks during their exponential growth phase rather than the stationary growth phase; this avoids the fixing of mutations that grant fitness to attributes other than growth rates. The experiment executed 8 parallel ALEs on replicates of *E. coli* K-12 MG1655, capturing samples from both the final and intermediate ALE flasks for whole genome sequencing with the Illumina MiSeq platform and re-sequencing using the Breseq 0.23 software pipeline against the *E. coli* K-12 MG1655 reference genome (NCBI accession NC_000913.2).

The authors employed two strategies for identifying key mutations. The first was to find genomic regions that were mutated in the endpoint of multiple ALEs. The second was to identify genomic regions within an ALE that experienced the replacement of one mutation with another; both mutations involved were considered key mutations.

Table 5.11: The only difference between the two sets is that the fixed key mutation set does not include the *wecA* alleles.

Position	Mutation Type	Sequence Change	Gene	Protein change	GLU A3 F244 I1 R1	GLU A6 F40 I1 R1	GLU A6 F76 I1 R1	GLU A6 F238 I1 R1	GLU A6 F406 I1 R1	GLU A8 F76 I1 R1	GLU A8 F380 I1 R1	GLU A9 F262 I1 R1	GLU A9 F433 I1 R1	GLU A10 F75 I1 R1	GLU A10 F247 I1 R1	GLU A10 F320 I1 R1	GLU A10 F418 I1 R1
139,326	SNP	C→T	gcd	G634S (GGC→AGC)						1	1						
354,036	SNP	C→T	prpE, codB	intergenic (+220/-110)		1	1	1	1								
1,088,445	SNP	G→A	pgaB	D212D (GAC→GAT)		1	1	1	1								
1,628,622	MOB	IS5 (+) +4 bp	ydfI	coding (313316/1461 nt)								1	1				
1,753,449	DEL	Δ1 bp	ydhZ, pykF	intergenic (-284/-273)												1	1
1,877,853	MOB	Δ1 :: IS186 (+) +6 bp :: Δ1	yeaR	coding (115120/360 nt)		1	1	1	1								
2,222,310	MOB	IS5 (+) +4 bp	pbpG	coding (580583/933 nt)								1	1				
2,531,514	SNP	A→T	cysK, ptsH	intergenic (+112/-272)												1	1
2,626,666	SNP	A→T	yfgF	C99S (TGT→AGT)										1	1	1	1
2,775,999	SNP	A→T	ypjF, ypjA	intergenic (+195/+169)						1	1						
2,984,674	SNP	T→G	yqeG	V269G (GTC→GGC)				1	1								
3,179,196	SNP	G→A	ygiC	G252S (GGC→AGC)										1	1	1	1
3,796,675	DEL	Δ1 bp	waaU	coding (661/1074 nt)		1	1	1	1								
3,966,245	DEL	Δ1 bp	wecA	coding (307/1104 nt)	1												
3,966,923	MOB	IS5 (–) +4 bp	wecA	coding (985988/1104 nt)							1						
4,508,547	SNP	C→A	yjhV, fecE	intergenic (+391/+166)										1	1	1	1

The ALE Analytics enrichment key mutation analysis finds all 8 of the genomic regions affected by the published key mutations. This analysis also finds 15 additional possible significant genomic regions, described in Table 5.11, due to the frequency in which a many mutations reoccurred within an ALE. The ALE Analytics fixed key mutation analysis finds 14 unpublished key mutations. The ascending frequency fixed mutation analysis finds the same results. The subset of unpublished fixed mutations are identical to the subset of unpublished enrichment mutations, with the exception of *wecA*. The *wecA* alleles do manifest in such a way that conforms to the key mutation protocol published for this ALE experiment, though were not included in the paper's key mutation results. Besides the *wecA* alleles, the new key mutations were not published by the authors as being significant since they did not exhibit the criteria of mutating the same genomic region in more than one ALE or replacing a mutation within an ALE. Though the new key mutation genomic regions contributed to the lessening of the enrichment and fixed key mutation analysis' precision metric, they present an opportunity for identifying further significant adaptations not identified by the authors.

Table 5.12: The GLU ALE experiment key mutation genomic region matching summary between the paper and the ALE Analytics automated key mutation analysis.

Genomic Region	Paper	ALE Analytics Enrichment	ALE Analytics Fixed
rph	X	X	X
rpoB	X	X	X
hns, tdk	X	X	X
corA	X	X	X
ygaZ	X	X	X
iap	X	X	
metL	X	X	X
ygeW	X	X	X
gcd		X	X
prpE, codB		X	X
pgaB		X	X
ydfI		X	X
ydhZ, pykF		X	X
yeaR		X	X
pbpG		X	X
cysK, ptsH		X	X
yfgF		X	X
ypjF, ypjA		X	X
yqeG		X	X
ygiC		X	X
waaU		X	X
wecA		X	
yjhV, fecE		X	X

Table 5.13: The GLU ALE experiment classification performance.

Analysis	True Positive	False Positive	False Negative	Recall	Precision
Enrichment	8	15	0	1	0.348
Fixed	7	14	1	0.875	0.333

I would like to acknowledge Dr. Adam Feist, Dr. Ryan LaCroix, Troy Sandberg, Gabriela Guzman, Joon Ho Park, Colton Llyod, Douglas McCloskey and the many others who have contributed their experiment data to enable the work of this thesis.

Chapter 6

Discussion

6.1 Key Mutation Analysis

The implementation of our automated analysis aims to distill the multiple methods published in identifying key mutations, though each of these published methods make no guarantee of including all significant mutations or excluding artifacts such as hitchhiker mutations [14]. A successful key mutation analysis will ultimately leverage multiple factors to judge a mutation's significance in an ALE experiment. Additional factors to those we have implemented in this thesis would be ALE growth rate profiles and gene functional group mutational enrichment.

In executing an evolution, the ALE Machine must track the growth rate of sample for its operations. The compilation of these growth rates contain vital fitness data for the experiment, such as which sample manifested a jump in growth rate during an ALE. Mutations in a population that occur immediately before the growth rate jump and gain dominance once the growth rate stabilizes are likely candidates for key mutations. This fitness data can be integrated into key mutation analysis to serve as an additional dimension

in judging the significance of a mutation [14]. Mutations can also be evaluated according to if their host gene is functionally similar to other mutated genes or is associated with the selection pressure such as experimental conditions or initial perturbations. This list of factors ultimately describe the many additional dimensions that can be further incorporated into automatically evaluating whether a mutation is strongly correlated to the selection pressures of an evolution.

The enrichment key mutation analysis rendered the best classification performance for this thesis; this was likely due to the small amount of samples per ALE with the ALE experiments used in the evaluation. Future ALE experiments may be enabled by ALE Analytics to incorporate more samples per ALE, with each sample being a population rather than clone. In general, more samples would grant more data points in the time-course of an evolution, allowing for more mutation data to describe the evolution. If these samples were populations, our analysis would be able to investigate the population dynamics of evolutions and track not only the consensus mutations but the balance of mutations found in only subsets of the entire population. This higher resolution of samples would enable the fixed key mutation analysis to have a higher probability of finding fixed mutation patterns. The enrichment key mutation analysis may ultimately be identifying mutations that, with more samples, would be identified with fixed key mutation analysis. The fixed key mutation analysis is more clear and intuitive in its intent in describing as to why a mutation may be significant, which is an advantage that the enrichment mutational pattern analysis lacks.

The work of this thesis justifies more population samples per evolution. Before this work, the decision on the sample count to sequence per ALE was based on investigating the genotypes of endpoint samples. Now that we have tools to quickly process a previously infeasible volume of samples, we have the opportunity to investigate new strategies in identifying

key mutations and exploring population dynamics within experimental evolutions.

6.2 Shared Enrichment and Fixed Mutations

A significant opportunity of the consolidation of ALE experiment reporting is that one can easily compare the mutations of multiple ALE experiments and search for mutational trends among all provided ALE experiments. We have leveraged this opportunity and implemented the *shared enrichment* and *shared fixed key mutation* features. These features leverage the nature of the ALE Analytics platform to automate the identification of genomic regions that share key mutations across ALE experiments. In this section, we present the shared enrichment and fixed mutations and elaborate on their significance.

Table 6.1: Shared enrichment and fixed mutation genomic regions among all ALE experiments evaluated.

Genomic Region	Shared Enrichment Mutations				Shared Fixed Mutations	
	PGI	42C	C13	GLU	C13	GLU
rph	X	X	X	X	X	X
hns, tdk		X	X	X		
rpoB			X	X	X	X
rpoC		X	X			
pykF		X		X		

rph was host to enrichment and fixed key mutations in all ALE experiments used in our evaluation. The meta data for the samples hosting these key mutations shows us that each experiment shares a parent strain yet differ by a single feature. Besides the GLU project, these differing features describe the selection pressure on the experiments, therefore each ALE experiment sharing the *rph* enrichment theoretically involves a different selection pressure. We can therefore conclude from the comparison of experimental conditions

through the experiment meta data that the enrichment of *rph* is a general optimization that the parent strain of these experiments can obtain when under any pressure. The *rph* mutations seen in these ALE experiments are in fact thought to help manage a specific defect in the *E. coli* K-12 MG1655 strain in which the starting strain for these ALE experiments derive from [7].

The *hns-tdk* intergenic region was the second most mutated among all enrichment mutations. This region along, with *rpoB* and *rpoC*, has been associated with global transcriptional regulation. Mutating these genes in some manner may have benefited a host's growth rate according to changes in transcription levels [12, 3, 6, 23]. The GLU and C13 ALE experiments sharing the *rpoB* enrichment and fixed mutations were in fact very similar in that the experiments were designed to select for growth rate, where C13 only differed in the isotope of the carbon source provided. These shared mutated key genomic regions therefore confirm the conclusion of the [20], that the C13 carbon source does not significantly affect the host metabolism and therefore enabled the evolutions of the C13 and GLU ALE experiments to follow a significantly similar track.

The *pykF* genomic region is enriched in the 42C and GLU ALE experiments. These experiments have very similar conditions beyond their sample temperatures. Mutations in *pykF* have in fact been associated with enabling an increase in uptake of glucose by reducing or disabling the metabolism of phosphoenolpyruvate to pyruvate [24, 11, 4]. This fitness benefit is likely enabled by the fact that both experiments use M9 glucose minimal media, therefore providing an abundance of glucose to the populations.

6.3 ALE Analytics Platform Feature Overview

The ALE Analytics platform was built to support the analysis needs of its users; many features were therefore implemented to service various perspectives of analysis. This section presents an overview of the current capabilities of the ALE Analytics platform to exemplify how all current analysis features fit together.

On login, users are greeted with the ALE Analytics homepage, known as the *dashboard*, which presents an overview of the mutational database currently available to the platform. This page presents the most frequently mutated genes, the most frequent mutation descriptions and the frequency of mutation types within the mutation database. An example of the dashboard can be seen in Figure 6.1. From the dashboard, users have access to their ALE experiments and all other platform features. Each ALE experiment has a similar home page with the same type of statistics as the dashboard. This page additionally includes alignment statistics for all experiment samples and a mutation needle plot [22] for presenting the spread of mutations across the experiment's reference genome. Users can quickly gain a sense of mutation hot-spots in their experiment according to the mutation needle plot. The experiment home page is exemplified in Figure 6.2. From the ALE experiment home page, users have access to experiment specific applications, such as the experiment's mutation lineage and key mutation analysis reports. Users can also view a report of the meta data associated with each sample of an experiment, shown in Figure 6.3. The experimental condition details, crucial to analysis, are contained within this meta data report.

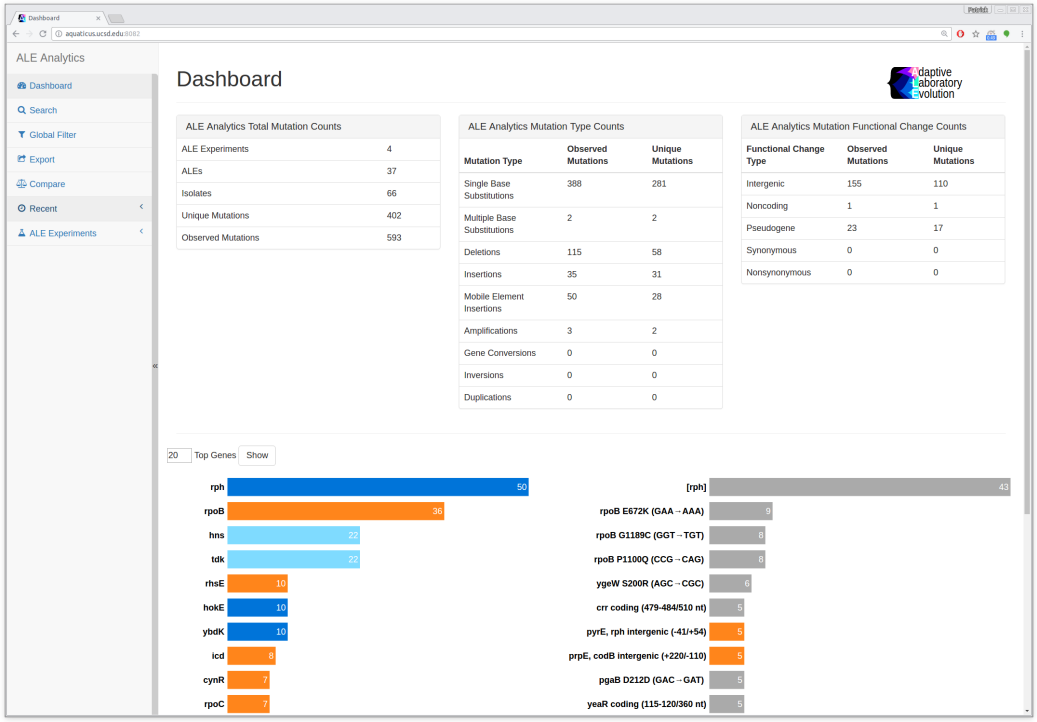


Figure 6.1: A screenshot of the dashboard for the instance of ALE Analytics used to accomplish the analysis contained within this thesis.

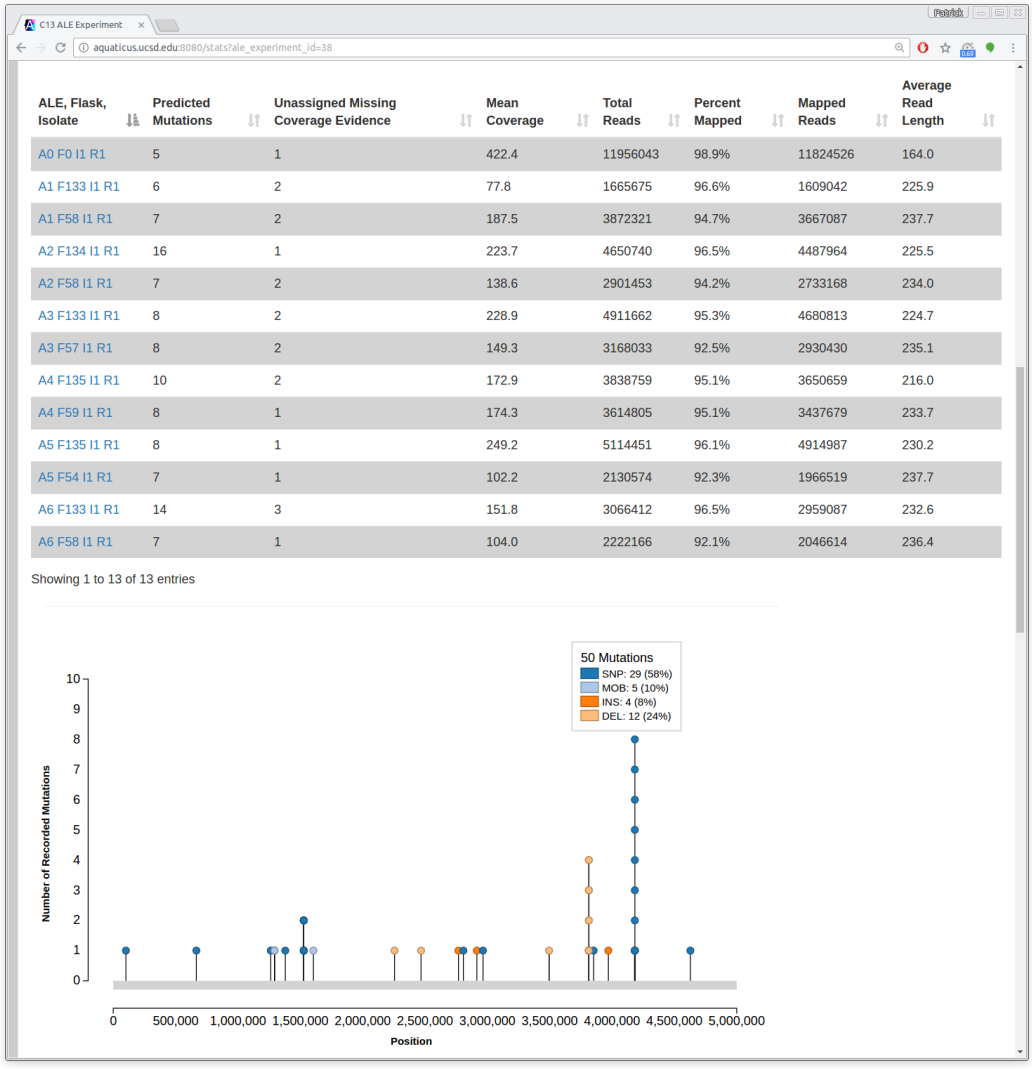
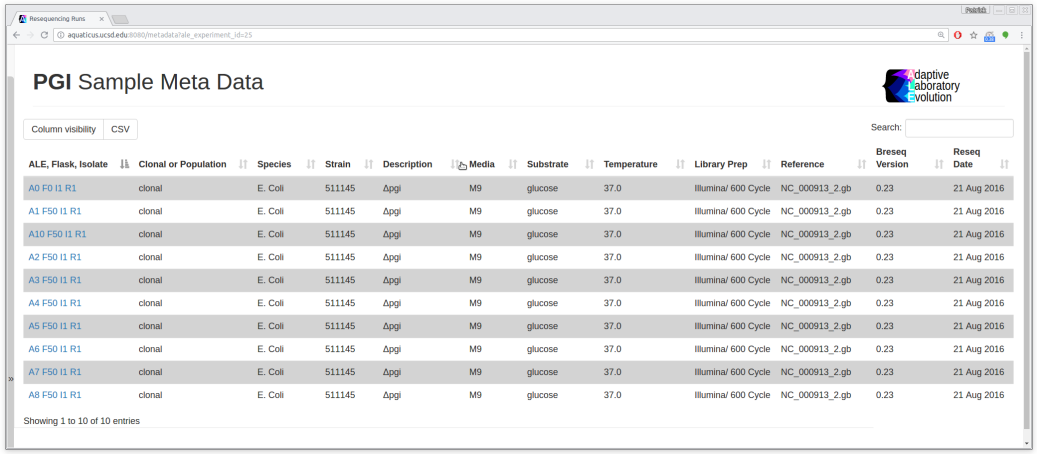


Figure 6.2: A screenshot of the C13 ALE experiment home page for the instance of ALE Analytics used to accomplish the analysis contained within this thesis.



PGI Sample Meta Data

Column visibility | CSV

Search:

ALE, Flask, Isolate	Clonal or Population	Species	Strain	Description	Media	Substrate	Temperature	Library Prep	Reference	Breseq Version	Reseq Date
A0 F0 I1 R1	clonal	E. Coli	511145	Δpgi	M9	glucose	37.0	Illumina/ 600 Cycle	NC_000913_2.gb	0.23	21 Aug 2016
A1 F50 I1 R1	clonal	E. Coli	511145	Δpgi	M9	glucose	37.0	Illumina/ 600 Cycle	NC_000913_2.gb	0.23	21 Aug 2016
A10 F50 I1 R1	clonal	E. Coli	511145	Δpgi	M9	glucose	37.0	Illumina/ 600 Cycle	NC_000913_2.gb	0.23	21 Aug 2016
A2 F50 I1 R1	clonal	E. Coli	511145	Δpgi	M9	glucose	37.0	Illumina/ 600 Cycle	NC_000913_2.gb	0.23	21 Aug 2016
A3 F50 I1 R1	clonal	E. Coli	511145	Δpgi	M9	glucose	37.0	Illumina/ 600 Cycle	NC_000913_2.gb	0.23	21 Aug 2016
A4 F50 I1 R1	clonal	E. Coli	511145	Δpgi	M9	glucose	37.0	Illumina/ 600 Cycle	NC_000913_2.gb	0.23	21 Aug 2016
A5 F50 I1 R1	clonal	E. Coli	511145	Δpgi	M9	glucose	37.0	Illumina/ 600 Cycle	NC_000913_2.gb	0.23	21 Aug 2016
A6 F50 I1 R1	clonal	E. Coli	511145	Δpgi	M9	glucose	37.0	Illumina/ 600 Cycle	NC_000913_2.gb	0.23	21 Aug 2016
A7 F50 I1 R1	clonal	E. Coli	511145	Δpgi	M9	glucose	37.0	Illumina/ 600 Cycle	NC_000913_2.gb	0.23	21 Aug 2016
A8 F50 I1 R1	clonal	E. Coli	511145	Δpgi	M9	glucose	37.0	Illumina/ 600 Cycle	NC_000913_2.gb	0.23	21 Aug 2016

Showing 1 to 10 of 10 entries

Figure 6.3: A screenshot of the PGI ALE experiment's meta data for the instance of ALE Analytics used to accomplish the analysis contained within this thesis.

Users often need to compare ALE experiments to identify shared mutations. ALE Analytics includes a feature named *compare* that will summarize the combination of ALE experiments similar to the experiments home page and build a mutation lineage reporting page from their combined mutations. Along with the mutation lineages, this feature builds the combined enrichment and fixed mutation tables, automating the comparison of the obvious significant mutations among compared ALE experiments.

Mutation filters play a critical role in all ALE experiment analysis. Both experimentalists and automated analysis require functionality that ignore mutations inappropriate for analysis. Mutations are often judged as inappropriate due to being identified as sequencing or alignment artifacts. Mutations can also be filtered on the basis of not containing any information pertaining to an experiment, such as those exhibited by an experiment's starting strain in relation to the reference genome used in alignment. We have defined two levels of filters: *global* and *experiment* levels. Global filters exclude the occurrence of a mutations for all ALE experiment analysis, where experiment filters ignore mutations for specific experiments. The parameters for global mutation filtering are unique mutations and genes.

The parameters for experiment mutation filtering are unique mutations, genes, and observed mutation frequency. The experiment mutation filtering is shown in Figure 6.4.

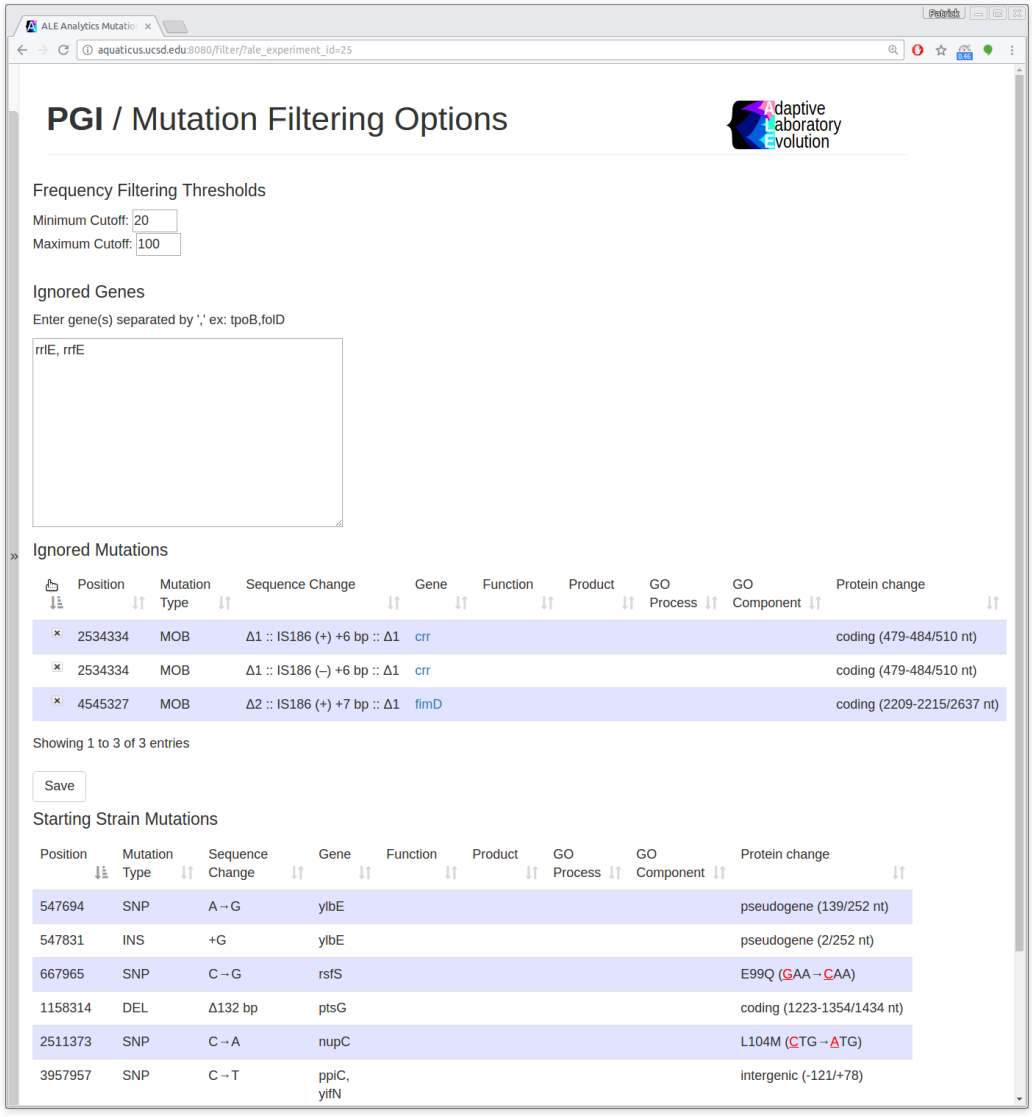


Figure 6.4: A screenshot of the PGI ALE experiment mutation filter page for the instance of ALE Analytics used to accomplish the analysis contained within this thesis.

The compilation of all ALE experiments into one resource enables our ALE Analytics platform to implement a mutation search feature. This feature generates reports of mutations and their host samples according to a set of search parameters. An example of the

search feature is shown in Figure 6.5

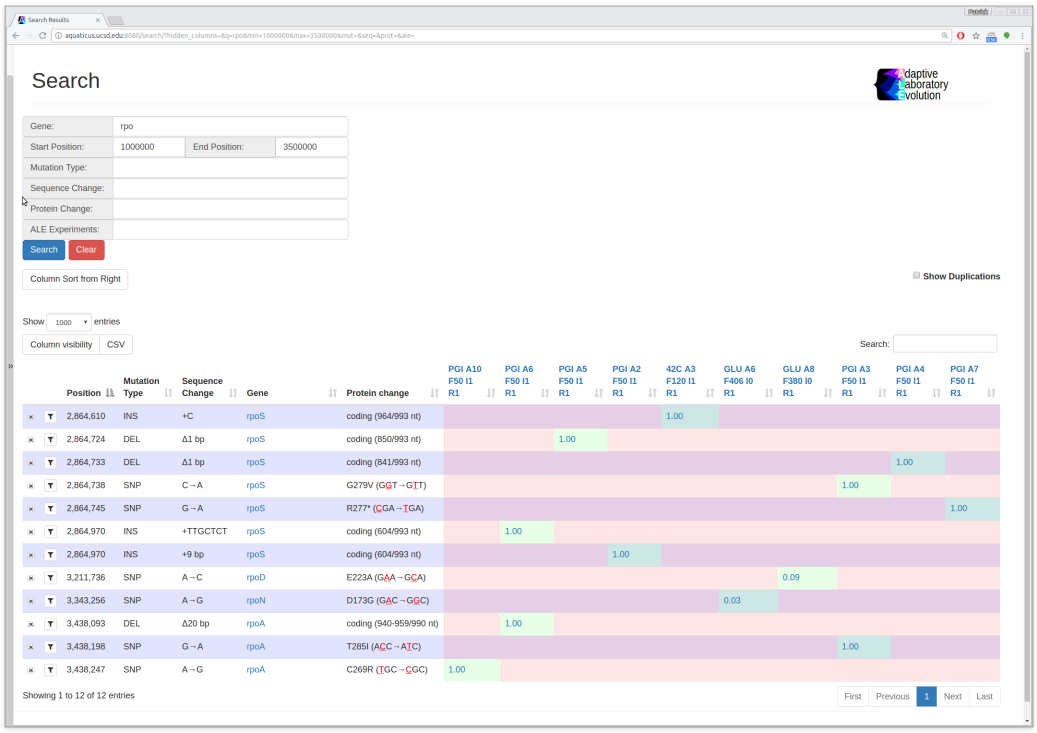


Figure 6.5: The search functionality uses as mutation query parameters the *gene*, *start* and *end positions*, *mutation type*, *sequence change*, *protein change* and ALE experiment of mutations.

Experimentalists invent many different data mining protocols for exploring the data housed within the mutation database. It would not be practical to implement all of these protocols into ALE Analytics, though it is clear that experimentalists should have the ability to easily extract data sets from the mutation database for their own investigations. The ALE experiment export feature was implemented to support this case; it enables users to extract all mutations from one or more ALE experiments. Experimentalists and investigators are then free to implement data mining protocols of their own design on the ALE data. Figure 6.6 presents histograms generated using the export feature and external tools to explore the position and frequency of mutations affecting the genomes of the ALE experiments used in the analysis contained within this thesis.

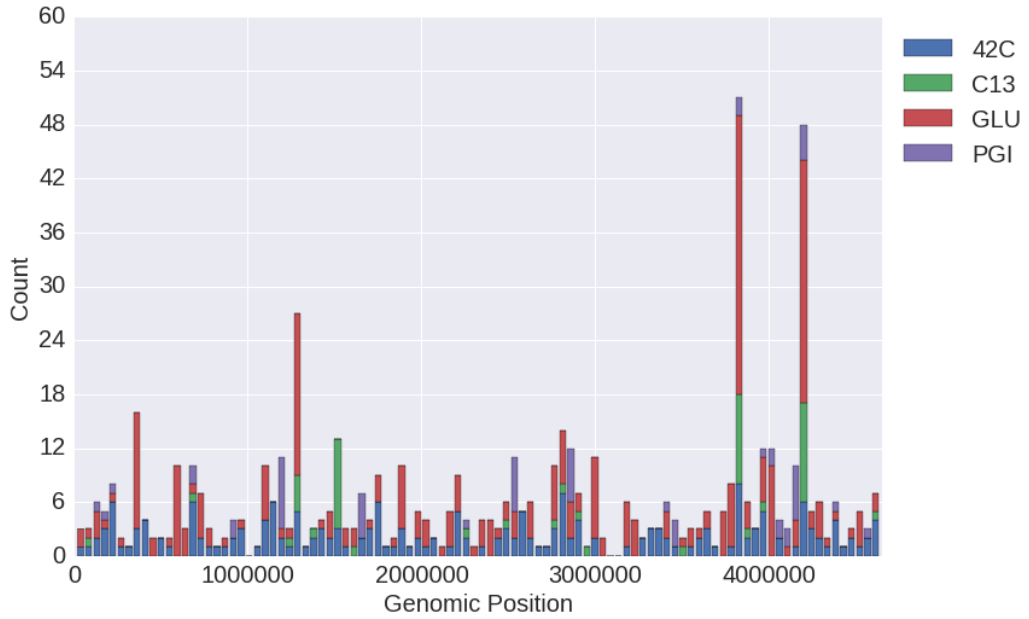


Figure 6.6: Mutation positions are divided into 100 bins and the results of the different ALE experiments involved are stacked to visualize the cumulative mutation frequency within a genomic region.

I would like to acknowledge Dennis Gosting for his work in implementing the current version of the majority of the user features described in this section.

I would like to acknowledge Anand Sastry for his work in implementing external data mining protocols to which I leveraged in generating the histogram figure include in this thesis of all mutations positions for the ALE experiments used in the evaluation of the automated analysis.

6.4 ALE Analytics Platform Deployment Overview

A production version of ALE Analytics has been deployed since August 2015 and has therefore been a live solution for more than a year to the ALE big data to knowledge challenges of the SBRG and the Novo Nordisk Center for Biosustainability of Lyngby, Denmark. Since this deployment, the ALE mutation database and the ALE Analytics platform have seen a dramatic increase in usage and ALE experiment data. Effort was devoted into designing a deployment environment using industry standard technologies and methodologies that would enable the platform's data to be secure and have redundant copies on external secure file servers.

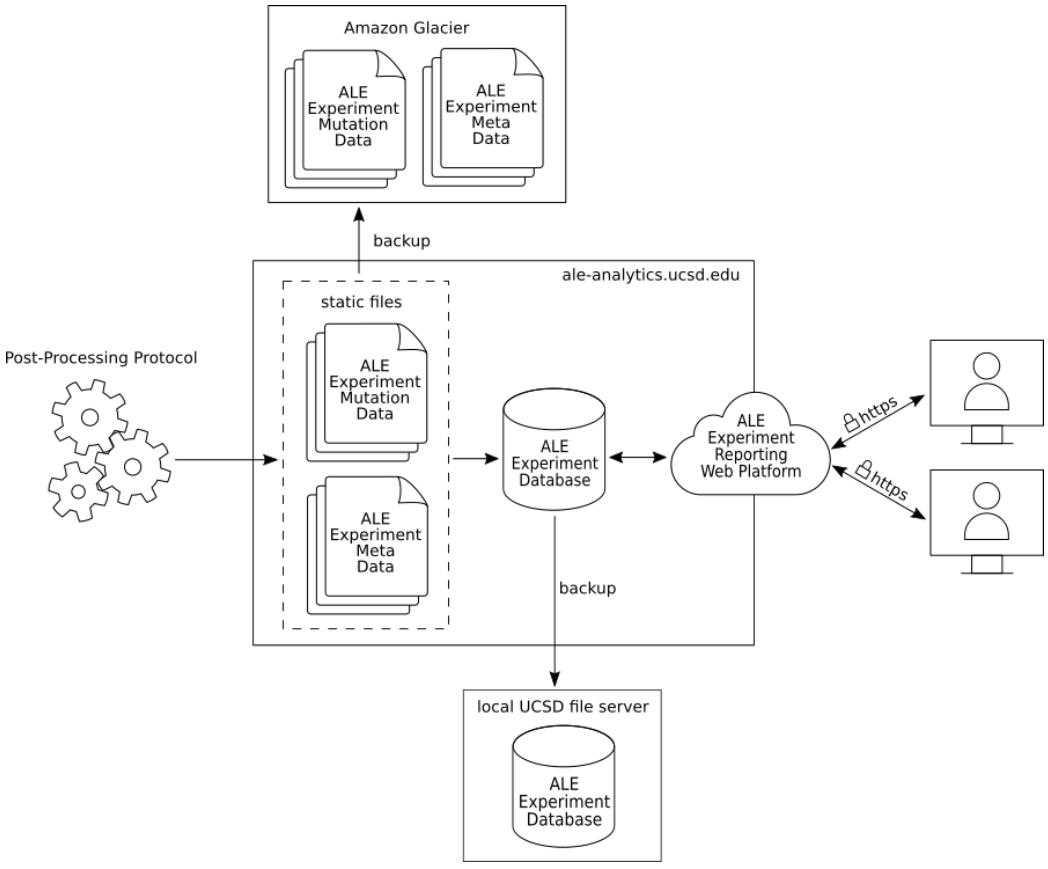


Figure 6.7: An illustration of the deployment environment for the ALE mutation database and ALE Analytics platform that describes important data security and redundancy measures.

The platform makes use of the *Linux*, *Django* [1] and *Nginx* [19] web application technology stack to deploy ALE Analytics from a dedicated server hosted by the San Diego Supercomputer Center. The static files that are used in populating the ALE mutation database and contain mutation reports referenced by ALE Analytics analyses are hosted on the same server and have backups on external servers through the Amazon Glacier file storage service (Amazon Web Services, Seattle, WA). The Amazon Glacier service is used due to the cumulative static file storage footprint of approximately 1.8 terabytes and this service's competitive storage costs. The ALE mutation database has a small enough storage footprint that it can have backups on a file server internal to UCSD and accessible only to SBRG members and IT staff. All reads used in generating the static files are also stored on the same internal file server. User access to the ALE Analytics data is controlled through a user account management system provided by the Django and Nginx technologies and all browsing traffic is encrypted via HTTPS protocol.

Chapter 7

Conclusion

The SBRG's ALE experiment automation has scaled the rate of experiment execution and data generation to the point that data post-processing, consolidation, reporting and common analysis have become a primary bottleneck in interpreting results. The SBRG's ALE methodologies therefore require a big data to knowledge solution that automates these bottlenecks to match the rate of high-throughput ALE experimentation. We have developed a software system that addresses each challenge that defines the big data to knowledge solution. To address quality control and mutation data formatting challenges, we investigate and establish a post-processing software pipeline. To address the challenge of high-throughput ALE experiment data consolidation, data reporting and common analysis, we implement the ALE Analytics software platform. The automated common analysis are evaluated against currently published ALE experiment key mutation results and show that they are precise, maintain high recall and can be expanded upon for more comprehensive predictions. We have additionally developed the ALE Analytics reporting and analysis platform as a web application to address the challenge of accessible experiment reporting. We go beyond these challenges and their solutions and have implemented features that leverage the consolidated

data to find key mutations shared among all ALE experiments. Finally, we have shown how ALE Analytics has implemented beyond the core challenge solving features to culminate in a platform that supports the multitude of services experimentalists currently need to execute their ALE experiment analysis.

Quality control protocols and tools have been investigated and combined into an ALE post-processing protocol. This protocol provides stage-dependent feedback that is crucial for all those involved in the ALE sample preparation and processing, as it informs them of their work's quality and better enables root-causing of quality issues.

Our system's automated common ALE experiment mutation analysis has been shown to be precise and maintain high recall in finding key mutations of published data sets. Of the published ALE experiments, an average recall of 89.6% and an average precision of 71.2% is achieved when excluding hypermutators. The automated key mutation analysis additionally identified key mutations in genes *wecA* and *yjiT* that were not included in the published material yet were aligned with their published key mutation protocols. Our automated key mutation analysis may lead to better result accuracy due to less potential for human error and variation in protocol between experimentalists.

The consolidation of ALE experiment data offers an opportunity for cross-experiment analysis. ALE Analytics has leveraged this opportunity with the implementation of the shared key mutations feature, which generates reports identifying genomic regions affected by key mutations in multiple ALE experiments. The ALE experiments used in evaluating the automated key mutation features manifests five of these shared key mutation genomic regions: *rph*, *hns-tdk*, *rpoB*, *rpoC* and *pykF*. The *rph* genomic region is mutated by key mutations in all provided ALE experiments and is proposed to be an adaptation for a defect that exists with the starting strain of these experiments [7]. The *hns-tdk*, *rpoB* and *rpoC*

genomic regions are each affected by key mutations in at least two ALE experiments and are proposed to be adaptive adjustments to global transcriptional regulation that benefit host growth rates [12, 3, 6, 23]. Key mutations affect the *pykF* gene in two ALE experiments and are speculated to contribute to the hosts growth rate by enabling a larger rate of glucose uptake through the disabling of a phosphenopyruvate metabolic process [24, 11, 4]. Disrupting *pykF* could render a fitness benefit in conjunction with the glucose rich media used in these experiments.

The work of this thesis does not stop at proposing and prototyping a possible big data to knowledge solution, but has in fact been deployed as a tool for ALE experimentalists at the System Biology Research Group and the Novo Nordisk Center for Biosustainability of Lyngby, Denmark, since August 2015. The current deployment leverages an industrial strength technology stack and production environment of Django, Nginx and Linux on a dedicated server hosted by the San Diego Supercomputer Center and strives to ensure security through user accounts and HTTPS encrypted browsing.

With this thesis' work, it is clear that the SBRG's ALE operations can now overcome its consolidation and reporting bottlenecks. This is exemplified by the current count of 42 ALE experiments, 325 ALEs and 14,765 observed mutations currently housed within the ALE experiment and mutation database, where each mutation is represent in an ALE experiment's mutation reports. The SBRG's ALE operations should in fact increase the number of samples per ALE experiment analysis sample set to better enable the automated analysis in predicting key mutations by effectively increasing the resolution of mutations in an evolution. Population samples, which reveal both consensus and population mutations, can additionally be included at higher frequencies in analysis sample sets to enable the exploration of population evolution dynamics. This research has yet to be thoroughly

examined within high-throughput ALE experiments due to the effort necessary in the curation of the magnitude of mutations involved with population samples.

ALE Analytics automated key mutations analysis could be enhanced by including ALE sample growth rate data and by investigating perturbations in the context of functionally related gene groups. ALE sample growth rate data, which describes the growth rate of samples during the progression of an ALE, can be used to automate the identification of mutations correlated with growth rate spikes. Mutations uniquely affecting genomic regions can be considered significant if they perturb a functionally related gene group which hosts additional mutations. Using these new data types and contexts, the automated key mutation analysis can therefore be expanded to consider multiple categories of evidence for significance.

Going forward, the ALE experiment mutation database presents an amazing opportunity for research into mutational trends across all ALE experiments available to the SBRG. Already, data mining protocol have made use of ALE Analytic's experiment export feature and have been used to characterize the general topology of mutations across the *E. coli* K-12 MG1655 genome from multiple ALE experiments. As ALE Analytics continues to integrate new ALE experiments, current and new data mining protocols can scan the ALE experiment mutation database in the hope of identifying previously unseen trends.

Bibliography

[1] Django (version 1.9.0): Web Framework. https://djangoproject.com, 2016.

[2] S. Andrews. Fastqc: a quality control tool for high throughput sequence data, 2010.

[3] Deborah G. Ayers, David T. Auble, and Pieter L. deHaseth. Promoter recognition by escherichia coli rna polymerase. *Journal of Molecular Biology*, 207(4):749 – 756, 1989.

[4] Diana Blank, Luise Wolf, Martin Ackermann, and Olin K. Silander. The predictability of molecular evolution during functional innovation. *Proceedings of the National Academy of Sciences*, 111(8):3044–3049, 2014.

[5] Pep Charusanti, Tom M. Conrad, Eric M. Knight, Karthik Venkataraman, Nicole L. Fong, Bin Xie, Yuan Gao, and Bernhard Palsson. Genetic basis of growth adaptation of escherichia coli after deletion of pgi, a major metabolic gene. *PLoS Genet*, 6(11):1–13, 11 2010.

[6] Kian-Kai Cheng, Baek-Seok Lee, Takeshi Masuda, Takuro Ito, Kazutaka Ikeda, Akiyoshi Hirayama, Lingli Deng, Jiyang Dong, Kazuyuki Shimizu, Tomoyoshi Soga, Masaru Tomita, Bernhard O. Palsson, and Martin Robert. Global metabolic network reorganization by adaptive mutations allows fast growth of escherichia coli on glycerol. *Nature Communications*, 5:3233 EP –, Jan 2014. Article.

[7] Tom M. Conrad, Andrew R. Joyce, M. Kenyon Applebee, Christian L. Barrett, Bin Xie, Yuan Gao, and Bernhard Ø Palsson. Whole-genome resequencing of escherichia coli k-12 mg1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. *Genome Biology*, 10(10):R118, 2009.

[8] Tom M. Conrad, Nathan E. Lewis, and Bernhard Ø Palsson. Microbial laboratory evolution in the era of genome-scale science. *Mol Syst Biol*, 7:509–509, Jul 2011. 21734648[pmid].

[9] D.E. Deatherage and J.E. Barrick. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol. Biol*, pages 165–188, 2014.

[10] Steven R. Head, H. Kiyomi Komori, Sarah A. LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R. Salomon, and Phillip Ordoukhanian. Library construction for next-generation sequencing: Overviews and challenges. *Biotechniques*, 56(2):61–passim, Feb 2014. 24502796[pmid].

[11] Toshihiko Kishimoto, Leo Iijima, Makoto Tatsumi, Naoaki Ono, Ayana Oyake, Tomomi Hashimoto, Moe Matsuo, Masato Okubo, Shingo Suzuki, Kotaro Mori, Akiko Kashiwagi, Chikara Furusawa, Bei-Wen Ying, and Tetsuya Yomo. Transition from positive to neutral in mutation fixation along with continuing rising fitness in thermal adaptive evolution. *PLoS Genet*, 6(10):1–10, 10 2010.

[12] Makoto Kobayashi, Kyosuke Nagata, and Akira Ishihama. Promoter selectivity of escherichia coli rna polymerase: effect of base substitutions in the promoter -35 region on promoter strength. *Nucleic Acids Research*, 18(24):7367–7372, 1990.

[13] Hannon Lab. FASTX Toolkit.

[14] R. A. LaCroix, T. E. Sandberg, E. J. O'Brien, J. Utrilla, A. Ebrahim, G. I. Guzman, R. Szubin, B. O. Palsson, and A. M. Feist. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. *Appl. Environ. Microbiol.*, 81(1):17–30, Jan 2015.

[15] Ryan LaCroix. *Automation, Optimization, and Characterization of Adaptive Laboratory Evolution*. PhD thesis, University of California, San Diego, 2016.

[16] Ronald Margolis, Leslie Derr, Michelle Dunn, Michael Huerta, Jennie Larkin, Jerry Sheehan, Mark Guyer, and Eric D Green. The national institutes of health's big data to knowledge (bd2k) initiative: capitalizing on biomedical big data. *Journal of the American Medical Informatics Association*, 21(6):957–958, 2014.

[17] B Palsson. Adaptive laboratory evolution. *Microbe Magazine*, 2011.

[18] BO. Palsson. *Systems Biology: Constraint-based Reconstruction and Analysis*. Cambridge University Press, 2015.

[19] Will Reese. Nginx: the high-performance web server and reverse proxy. *Linux J.*, 2008(173), September 2008.

[20] T. E. Sandberg, C. P. Long, J. E. Gonzalez, A. M. Feist, M. R. Antoniewicz, and B. O. Palsson. Evolution of E. coli on [U-13C]Glucose Reveals a Negligible Isotopic Influence on Metabolism and Physiology. *PLoS ONE*, 11(3):e0151130, 2016.

[21] T. E. Sandberg, M. Pedersen, R. A. LaCroix, A. Ebrahim, M. Bonde, M. J. Herrgard, B. O. Palsson, M. Sommer, and A. M. Feist. Evolution of Escherichia coli to 42 °C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. *Mol. Biol. Evol.*, 31(10):2647–2662, Oct 2014.

[22] Michael P Schroeder and Nuria Lopez-Bigas. muts-needle-plot: Mutations needle plot v0.8.0, January 2015.

[23] Wenqin Wang, Gene-Wei Li, Chongyi Chen, X. Sunney Xie, and Xiaowei Zhuang. Chromosome organization by a nucleoid-associated protein in live bacteria. *Science*, 333(6048):1445–1449, 2011.

[24] Robert Woods, Dominique Schneider, Cynthia L. Winkworth, Margaret A. Riley, and Richard E. Lenski. Tests of parallel molecular evolution in a long-term experiment with escherichia coli. *Proceedings of the National Academy of Sciences*, 103(24):9107–9112, 2006.