UNIVERSITY OF CALIFORNIA SAN DIEGO

The Diminishing Role of Hybrid Genome Assemblies in Longitudinal and Automated Mutation Detection

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

in

Bioengineering

by

Mia Jade Sales

Committee in charge:

Bernhard O. Palsson, Chair
Vineet Bafna
Gert Cauwenberghs
Adam Feist

2019

The Thesis of Mia Jade Sales is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Chair

University of California San Diego

2019

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Bernhard O. Palsson for his support as the chair of my committee.

I would also like to thank Adam Feist and Jonathan Monk for their support and guidance throughout my research career with the Systems Biology Research Group. It is with their direction that I was able to transition my previous genomics research into implementable and valuable results.

Thank you to Muyao Wu, Patrick Phaneuf, and Troy Sandberg for their support and assistance in navigating the lab and its many resources. Thank you to Richard Szubin and Ying Hefner for providing the crucial Nanopore data.

Chapter 1, in part, is currently being prepared for submission for publication. Mia Jade Sales, George Sakoulas, Richard Szubin, Bernhard Palsson, Jonathan M. Monk. The thesis author was the primary author of this chapter.

VITA

2018	Bachelor of Science in Bioengineering with a specialization in Computational and Systems Biology, University of Illinois at Urbana Champaign
2018-2019	Bioinformatics Technician, University of Illinois at Urbana Champaign
2019	Master of Science in Bioengineering, University of California San Diego

PUBLICATIONS

Sales, Mia J. Complete Genome Sequences of Streptococcus sobrinus SL1 (ATCC 33478 = DSM 20742), NIDR 6715-7 (ATCC 27351), NIDR 6715-15 (ATCC 27352), and NCTC 10919 (ATCC 33402). American Society for Microbiology. 2018.

FIELDS OF STUDY

Major Field: Engineering

Bioengineering
Computational and Systems Biology

ABSTRACT OF THE THESIS

The Diminishing Role of Hybrid Genome Assemblies in Longitudinal and Automated Mutation Detection

by

Mia Jade Sales

Master of Science in Bioengineering

University of California San Diego, 2019

Professor Bernhard O. Palsson, Chair

Hybrid genomes, formed by high accuracy short-reads and long-reads with the ability to span even large repeat regions, allow for analysis of specific mutations that arise in longitudinal isolates of both clinical and lab-evolved samples. Presented here is an analysis of mutations that arise in a clinical isolate of methicillin-sensitive *Staphylococcus aureus*, Tx0117, which displays a high inoculum effect with cefazolin.

The curing of the beta-lactamase gene created strain Tx0117c, which demonstrates an increased resistance to a variety of antimicrobial peptides. Analysis of the acquired mutations using Breseq uncovered five additional genes affected by the curing process, which may contribute to the increased resistance observed. Additionally, this thesis includes a study on automation of copy number variation detection in lab-evolved strains using the ALE system. The implementation of a python function, which takes a map of coverage depth per base position as input, is tested here in its ability to accurately extract amplification events from only short-read data and a reference genome of the parent strain. The results are validated using an automated pipeline which utilizes hybrid genome assemblies of the evolved strains and blastn (version 2.9.0) for identification of local alignments in the genomes. The sample sets analyzed suggest that short-read data alone is sufficient in identifying amplifications by size, depth, and location. Additionally, the new function allows for efficient and accurate analysis of the flanking regions of amplifications and the encoded genes, providing insight into the mechanisms of acquisition of such events.

Introduction

All divergence from an initial life form is a result of mutations. A mutation may be a single-nucleotide variation (SNV), or a larger mutation affecting a sequence of nucleotides. While some areas of the genome may be subject to higher mutation rates, mutations occur at random, and the vast majority are neutral and do not help or hurt an organism. However, some can be beneficial, often leading to fixation in which the mutation spreads throughout the population, and some can be detrimental, leading to elimination of the mutation in the population.

Any population of microorganisms is constantly under some environmental stress, leading to fixation and elimination of randomly occurring mutations. The result is a host of genomes with variability to one another and to any other generation. To determine which mutations are causal mutations, and the result of some specific stress, the genomes of evolving strains are analyzed over many timepoints and considered with their respective growth rates and selective stresses. This process involves first investigating the genetic composition of a single microorganism in a population, requiring the isolation and maintenance of individual cells. Cells are grown in a flask and serially passed to new flasks subject to a new selective stress after a preset time period that allows for the development of adaptive mutations with the ability to produce improved phenotypes. The ALE system utilizes a common naming format of “ExperimentName A.F.I.R.”, where ‘A’ is the ALE number, ‘F’ is the flask number, ‘I’ is the isolate number, and ‘R’ is the technical replicate number. This allows for chronological ordering of evolved strains for more efficient identification of causal mutations. ‘I’ may be either 0 or 1, where 0 represents a population sample and 1 represents a clonal sample. In this study, to ensure the genotypes in a single sample are as uniform as possible, only clonal

samples are used. Due to its ease of cultivation, rapid reproduction, and availability of well-established reference genomes, *Escherichia coli* is the ideal organism for detection of causal mutation events in the ALE system (1).

Accurate and precise detection of acquired mutations is dependent on the quality of the parent and child genomes being compared, and there are three major factors of sequencing that determine the final quality of a genome: coverage, read length, and quality. For a *de novo* genome assembly, with no provided reference for information on genome length or composition, the target sequencing coverage is about 50-100x (2). The goal is to avoid biased coverage, or the sequencing of some regions of the genome with greater frequency than others, such that every region of the genome has the same coverage value. Read length is most significant when the reads encompass repeat regions. In order to unambiguously resolve these regions, there must be some read which spans the entire repeat region and both flanking regions. The third factor is quality, which determines the accuracy of the resulting genome and the quantity of incorporated and discarded reads.

Common methods of DNA sequencing produce two types of data: long-read and short-read. The two datatypes bring different advantages to genome assembly, and when combined to form a hybrid assembly, the analytical capabilities of the resulting genome are much greater. Short-reads have the advantage of a low error rate, a lower cost, and a high read count due to the parallel architecture (3). In these studies, short-read data is produced using Illumina HiSeq sequencing. Short-reads satisfy the need for high quality reads, with Illumina short-reads leading sequencing methods in the proportion of error-free reads (4). However, when repeat regions larger than the length of a read are present in the genome, long-reads must be used for unambiguous resolution.

This is the key benefit of using long-reads in assemblies, and here long-reads are produced by the Oxford Nanopore Technologies MinION DNA sequencer. Aside from its compact design and real-time analytical capabilities, the MinION can produce the longest reads of any commonly used sequencer, as the length of the reads produced is limited only by the length of DNA that can be isolated in library preparation. This is because each read is fed, without digestion, through a nanopore. The nanopore is inserted in an electrically resistant membrane and a voltage is applied across the membrane, such that movement of each nucleotide in the strand through the nanopore produces a characteristic disruption in the electrical signal (5). The downside to this mechanism is high error rates near homopolymer tracts, or areas of the genome with a repeated single nucleotide, as the device cannot reliably differentiate between consecutive nucleotides to determine the length of the tract (6).

The strongest genome assemblies are hybrid assemblies, benefitting from the high accuracy of short-reads combined with the repeat resolution ability of long-reads. After obtaining data with sufficient coverage, read length, and quality, an assembler must be chosen. The assembler used throughout these studies is Unicycler, which utilizes SPAdes to first create the short-read assembly graph. It then selects a set of anchor sequences and bridges the anchors with SPAdes contiguous sequence—or contigs—paths. Unicycler then uses Miniasm to bridge these paths with long-reads, followed by iterative polishing using Racon. Unicycler accepts as inputs short-read data and optional long-read data (7).

Because the read types used in an assembly contribute to the overall genome quality, they also play a factor in the credibility of any detected mutations. A genome deduced from short-reads alone is likely to have high accuracy on a per-

basepair level, but large mutations such as amplifications—or copy-number variations (CNVs)—may not be identified, with the likelihood of detection decreasing as the amplification size and depth increase. A genome deduced from only long-reads can likely resolve these large amplifications but may not be reliable in detecting SNVs due to the high sequencing error rate. As such, hybrid assemblies are crucial to mutation detection, allowing for confidence in all detected mutation sizes.

Breseq is the mutation detection program used in the following studies. It takes as input a high quality, fully assembled reference genome, corresponding to the genome of the parent strain. It also takes genomic reads from the child strain, unassembled. These may be long or short reads, and Breseq's ability to detect mutations using next-generation long-read data has been tested and demonstrated (8). However, due to the price of short-read sequencing and the inclusion of short-read data in the mutation detection workflow currently used by ALE, an efficient method for reliably extracting amplifications using Breseq and short-read data alone would increase the analytical and predictive capabilities of the ALE system.

Chapter 1 introduces a simplistic application of Breseq using only short-reads for both parent and child data, which allows for high confidence in small mutations but lacks information on large mutations like potential amplification events. Presented is a full-genome analysis of the mutations detected in a clinical isolate of methicillin-sensitive *Staphylococcus aureus* strain TX0117 and its Bla-Z cured derivative, TX0117c. Chapter 2 presents a novel program for identifying amplification events in bacterial genomes using only short-read data from the child strain and a high-quality reference genome from the parent strain. The results are validated by sequencing the child strain with both long and short-read technologies and performing a hybrid assembly, which is then

analyzed against the reference in a separate pipeline to identify significant alignments between the two genomes.

References

1. Phaneuf, P. V., Gosting, D., Palsson, B. O. & Feist, A. M. ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. Nucleic Acids Res. **47**, D1164–D1171 (2019).

2. Pightling, A. W., Petronella, N. & Pagotto, F. Choice of reference sequence and assembler for alignment of Listeria monocytogenes short-read sequence data greatly influences rates of error in SNP analyses. PLoS One **9**, e104579 (2014).

3. May, M. The Long and the Short of Sequencing. (2018). at <https://www.biocompare.com/Editorial-Articles/349220-The-Long-and-the-Short-of-Sequencing/>

4. Benefits of SBS Technology | Robust sequencing data quality. at <https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/sbs-benefits.html>

5. Allen, G. & Gillespie, S. R. How it Works. (Tauchnitz, 1895). at <https://nanoporetech.com/how-it-works>

6. O'Donnell, C. R., Wang, H. & Dunbar, W. B. Error analysis of idealized nanopore sequencing. Electrophoresis **34**, 2137–2144 (2013).

7. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13,** e1005595 (2017).

8. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol. Biol.* **1151,** 165–188 (2014).

Chapter 1- Mutation Analysis of MSSA strain TX0117 and BlaZ-cured derivative TX0117c

Beta-lactam therapy has been associated with better outcomes than non-beta-lactam therapy (i.e. vancomycin) in patients with methicillin-susceptible Staphylococcus aureus (MSSA) bacteremia (1–4). Treatment standards recommend therapy either with an isoxazolyl penicillin (eg. nafcillin, oxacillin) or cefazolin (5). Recent data is emerging that shows that patients treated with cefazolin demonstrate similar efficacy but better tolerability compared to those treated with anti-staphylococcal isoxazoyl penicillins (6,7). However, cefazolin treatment failures have been reported due to a cefazolin inoculum effect, which is defined by isolates showing an MIC of >16 mg/L in assays utilizing a bacterial inoculum of 107 CFU/ml compared to the standard inoculum of 105 CFU/mL (8–11). The cefazolin inoculum effect is based on the ability of the beta-lactamase of some MSSA strains to overcome and hydrolyze cefazolin when bacteria are at high inoculum, and has been shown to cause clinical failures in certain deep-seated infections. These isolates may be uncommon, but considerable regional variability is seen in their prevalence (12–15).

In order to examine the effects of different antimicrobial therapies *in vitro* against MSSA with a significant cefazolin inoculum effect, a clinical strain was isolated from a patient with MSSA endocarditis that relapsed after cefazolin therapy (TX0117) (11). This strain was subsequently cured of the beta-lactamase plasmid by heat at 43°C and by novobiocin exposure, yielding TX0117c (16–18). The TX0117 and TX0117c MSSA strain pair have been extensively studied in various *in vitro* models and in *in vivo* rat endocarditis models to better understand the comparative efficacy of different antibiotics against MSSA

exhibiting the beta-lactamase mediated cefazolin inoculum effect and against an isogenic MSSA that has been cured of its beta-lactamase (19,20).

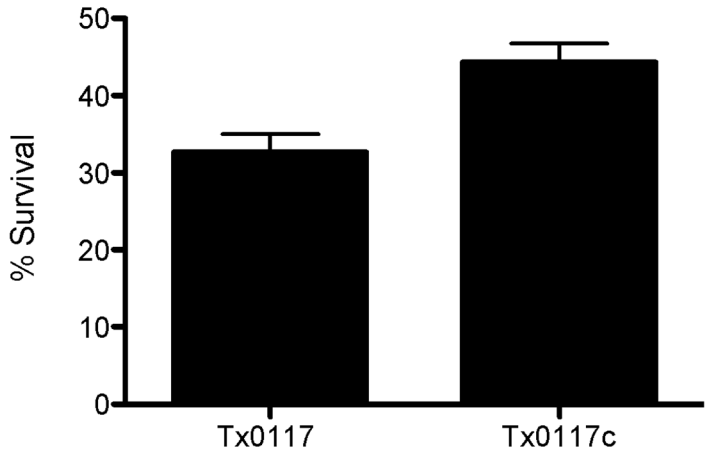


Figure 1.1- Percent survival after 2 hours in LL37 killing assays (32 micromolar) of MSSA TX0117 compared to its beta-lactamase cured derivative TX0117c. (p= 0.006, Mann Whitney-U test)

Our evaluation of TX0117 and TX0117c showed subtle but consistent increased resistance to cationic antimicrobial peptides in strain TX0117c as compared to the TX0117 parent strain, leading us to hypothesize that in addition to curing the strain of beta-lactamase, novobiocin and heat treatment may have additionally co-selected previously uncharacterized mutations in TX0117c. To investigate these mutations, we mapped short-reads from the TX0117c genome to our newly sequenced TX0117 genome using Breseq (version 0.31.0) (21).

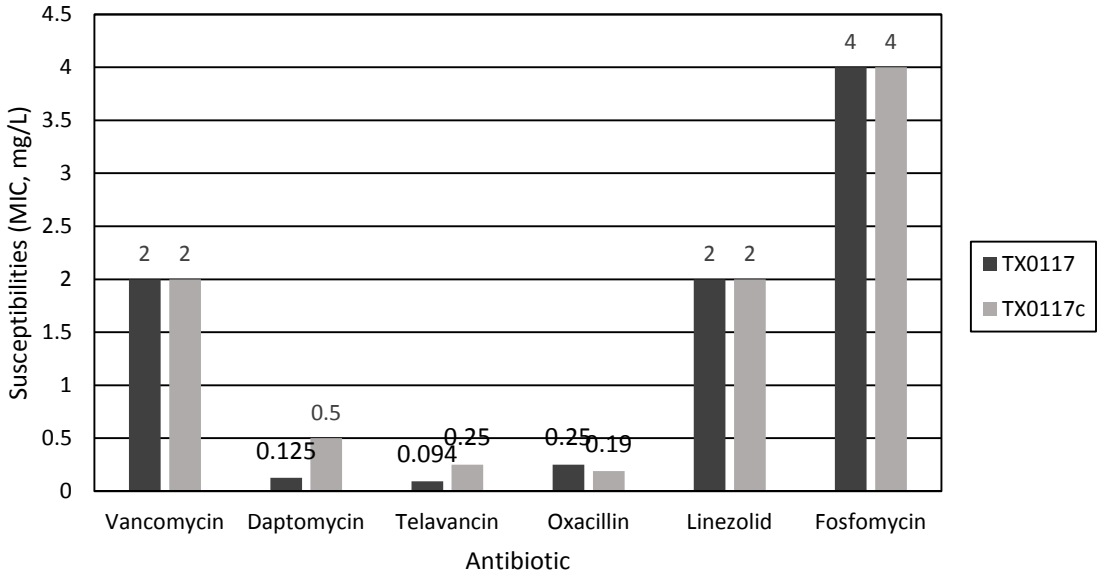


Figure 1.2- Susceptibilities (MIC, mg/L) of different antibiotics against TX0117 and beta-lactamase cured derivative TX0117c determined by Etest.

Genomic DNA was prepared for Illumina sequencing. Illumina libraries were generated using the TruSeq DNA Sample Preparation Kit (Illumina Inc., USA). The libraries were sequenced using the Illumina MiSeq platform with a paired-end protocol and read lengths of 150 nt. The resulting short-reads were assembled with Unicycler (version 0.4.2) (22). The incomplete TX0117 genome consists of 114 contigs and 2.758 Mb. The final assembled genome was annotated using Prokka (version 1.12) (23). The genome has 2,562 annotated coding sequences (CDSs), 16 tRNAs, and 4 rRNAs.

Using the Breseq mutation prediction pipeline, we identified genes altered from TX0117 to the TX0117c strain. Most noteworthy is the curing of BlaZ, as previously documented. The major mechanism of penicillin resistance, involving the hydrolysis of the beta-lactam ring, has been attributed to beta-lactamase, which is encoded by the BlaZ gene (24). Type A Beta lactamases contribute to more efficient inactivation of beta-lactam drugs and therefore correlate to the inoculum effect (25).

TX0117c also displayed a truncation of relA, the GTP pyrophosphokinase involved in the stress response. relA encodes the RELA protein, which most commonly binds NFKB1 to form a NF-kappa-B transcription factor, activated downstream by processes such as inflammation, tumorigenesis, and differentiation (26). Also mutated, via substitution, is the mnaA_1 gene. It encodes a UDP-N-acetylglucosamine 2-epimerase responsible for converting UDP-GlcNAc into UDP-N-acetyl mannosamine, which is then oxidized in the formation of teichoic acids (27). Teichoic acids bind to either peptidoglycans or cytoplasmic membranes and dictate functions from cellular shape to pathogenesis. They have been proven necessary for beta lactam resistance in MRSA (28), and have been shown to control bacteria susceptibility to antimicrobial peptides and cationic antibiotics (29,30). Additional studies will be needed to examine the role of relA and mnaA in S. aureus susceptibility to cationic antimicrobial peptides.

Data availability. This Whole Genome Shotgun project has been deposited in NCBI GenBank under the accession no. VSSN00000000.1, the Illumina short-read data for TX0117 and TX0117c has been deposited in SRA under the accession no. SAMN12622398 and SAMN12622402, respectively.

Table 1.1- Differences in coding regions between TX0117 and TX0117C

gene	description	mutation	annotation
blaZ ←	Beta-lactamase	$(T)_{8\rightarrow7}$	coding (92/846 nt)
relA →	GTP pyrophosphokinase	C→T	Q657* (CAA→TAA) (86% truncation)
TX0117_02559 →	hypothetical protein	T→A	N24K (AAT→AAA)
TX0117_01069 ←	Hyaluronate lyase	T→C	N476S (AAT→AGT)
		T→C	I471V (ATC→GTC)
		T→C	D457G (GAC→GGC)
		T→G	S450R (AGT→CGT)
		T→G	K271T (AAA→ACA)
		A→T	L262M (TTG→ATG)
		C→T	V254I (GTT→ATT)
mnaA_1 →	UDP-N-acetylglucosamine 2-epimerase	C→T	P149S (CCA→TCA)
TX0117_02434 ←	65 kDa membrane protein	A→T	N181K (AAT→AAA)
		C→T	S177N (AGC→AAC)

Chapter 1, in part, is currently being prepared for submission for publication. Mia Jade Sales, George Sakoulas, Richard Szubin, Bernhard Palsson, Jonathan M. Monk. The thesis author was the primary author of this chapter.

Chapter 1 References

1. Chang, F.-Y., Peacock, J. E., Jr, Musher, D. M., Triplett, P., MacDonald, B. B., Mylotte, J. M., O'Donnell, A., Wagener, M. M. & Yu, V. L. Staphylococcus aureus bacteremia: recurrence and the impact of antibiotic treatment in a prospective multicenter study. *Medicine* **82,** 333–339 (2003).

2. Schweizer, M. L., Furuno, J. P., Harris, A. D., Johnson, J. K., Shardell, M. D., McGregor, J. C., Thom, K. A., Cosgrove, S. E., Sakoulas, G. & Perencevich, E. N. Comparative effectiveness of nafcillin or cefazolin versus vancomycin in methicillin-susceptible Staphylococcus aureus bacteremia. *BMC Infect. Dis.* **11,** 279 (2011).

3. Wong, D., Wong, T., Romney, M. & Leung, V. Comparative effectiveness of β-lactam versus vancomycin empiric therapy in patients with methicillin-susceptible Staphylococcus aureus (MSSA) bacteremia. *Ann. Clin. Microbiol. Antimicrob.* **15,** 27 (2016).

4. Stryjewski, M. E., Szczech, L. A., Benjamin, D. K., Jr, Inrig, J. K., Kanafani, Z. A., Engemann, J. J., Chu, V. H., Joyce, M. J., Reller, L. B., Corey, G. R. & Fowler, V. G., Jr. Use of vancomycin or first-generation cephalosporins for the treatment of hemodialysis-dependent patients with methicillin-susceptible Staphylococcus aureus bacteremia. *Clin. Infect. Dis.* **44,** 190–196 (2007).

5. Holland, T. L., Arnold, C. & Fowler, V. G., Jr. Clinical management of Staphylococcus aureus bacteremia: a review. *JAMA* **312,** 1330–1341 (2014).

6. Li, J., Echevarria, K. L., Hughes, D. W., Cadena, J. A., Bowling, J. E. & Lewis, J. S., 2nd. Comparison of cefazolin versus oxacillin for treatment of complicated bacteremia caused by methicillin-susceptible Staphylococcus aureus. *Antimicrob. Agents Chemother.* **58,** 5117–5124 (2014).

7. McDanel, J. S., Roghmann, M.-C., Perencevich, E. N., Ohl, M. E., Goto, M., Livorsi, D. J., Jones, M., Albertson, J. P., Nair, R., O'Shea, A. M. J. & Schweizer, M. L. Comparative Effectiveness of Cefazolin Versus Nafcillin or Oxacillin for Treatment of

Methicillin-Susceptible Staphylococcus aureus Infections Complicated by Bacteremia: A Nationwide Cohort Study. *Clin. Infect. Dis.* **65,** 100–106 (2017).

8. Bryant, R. E. & Alford, R. H. Unsuccessful treatment of staphylococcal endocarditis with cefazolin. *JAMA* **237,** 569–570 (1977).

9. Quinn, E. L., Pohlod, D., Madhavan, T., Burch, K., Fisher, E. & Cox, F. Clinical experiences with cefazolin and other cephalosporins in bacterial endocarditis. *J. Infect. Dis.* **128,** Suppl:S386–9 (1973).

10. Fernández-Guerrero, M. L. & de Górgolas, M. Cefazolin therapy for Staphylococcus aureus bacteremia. *Clin. Infect. Dis.* **41,** 127 (2005).

11. Nannini, E. C., Singh, K. V. & Murray, B. E. Relapse of type A beta-lactamase-producing Staphylococcus aureus native valve endocarditis during cefazolin therapy: revisiting the issue. *Clin. Infect. Dis.* **37,** 1194–1198 (2003).

12. Nannini, E. C., Stryjewski, M. E., Singh, K. V., Bourgogne, A., Rude, T. H., Corey, G. R., Fowler, V. G., Jr & Murray, B. E. Inoculum effect with cefazolin among clinical isolates of methicillin-susceptible Staphylococcus aureus: frequency and possible cause of cefazolin treatment failure. *Antimicrob. Agents Chemother.* **53,** 3437–3441 (2009).

13. Wang, S. K., Gilchrist, A., Loukitcheva, A., Plotkin, B. J., Sigar, I. M., Gross, A. E., O'Donnell, J. N., Pettit, N., Buros, A., O'Driscoll, T., Rhodes, N. J., Bethel, C., Segreti, J., Charnot-Katsikas, A., Singh, K. & Scheetz, M. H. Prevalence of a Cefazolin Inoculum Effect Associated with blaZ Gene Types among Methicillin-Susceptible Staphylococcus aureus Isolates from Four Major Medical Centers in Chicago. *Antimicrob. Agents Chemother.* **62,** (2018).

14. Chong, Y. P., Park, S.-J., Kim, E. S., Bang, K.-M., Kim, M.-N., Kim, S.-H., Lee, S.-O., Choi, S.-H., Jeong, J.-Y., Woo, J. H. & Kim, Y. S. Prevalence of blaZ gene types and the cefazolin inoculum effect among methicillin-susceptible Staphylococcus aureus blood isolates and their association with multilocus sequence types and clinical outcome. *Eur. J. Clin. Microbiol. Infect. Dis.* **34,** 349–355 (2015).

15. Rincón, S., Reyes, J., Carvajal, L. P., Rojas, N., Cortés, F., Panesso, D., Guzmán, M., Zurita, J., Adachi, J. A., Murray, B. E., Nannini, E. C. & Arias, C. A. Cefazolin high-inoculum effect in methicillin-susceptible Staphylococcus aureus from South American hospitals. *J. Antimicrob. Chemother.* **68,** 2773–2778 (2013).

16. Nannini, E. C., Singh, K. V., Arias, C. A. & Murray, B. E. In Vivo Effects of Cefazolin, Daptomycin, and Nafcillin in Experimental Endocarditis with a Methicillin-Susceptible Staphylococcus aureus Strain Showing an Inoculum Effect against Cefazolin. *Antimicrob. Agents Chemother.* **57,** 4276–4281 (2013).

17. McHugh, G. L. & Swartz, M. N. Elimination of plasmids from several bacterial species by novobiocin. *Antimicrob. Agents Chemother.* **12,** 423–426 (1977).

18. Shore, A. C., Brennan, O. M., Ehricht, R., Monecke, S., Schwarz, S., Slickers, P. & Coleman, D. C. Identification and characterization of the multidrug resistance gene cfr in a Panton-Valentine leukocidin-positive sequence type 8 methicillin-resistant Staphylococcus aureus IVa (USA300) isolate. *Antimicrob. Agents Chemother.* **54,** 4978–4984 (2010).

19. Singh, K. V., Tran, T. T., Nannini, E. C., Tam, V. H., Arias, C. A. & Murray, B. E. Efficacy of Ceftaroline against Methicillin-Susceptible Staphylococcus aureus Exhibiting the Cefazolin High-Inoculum Effect in a Rat Model of Endocarditis. *Antimicrob. Agents Chemother.* **61,** (2017).

20. Carvajal, L. P., Santiago, A., Echeverri, A., Rios, R., Rincon, S., Panesso, D., Diaz, L., Miller, W., Sun, Z., Palzkill, T. & Others. 2063. Extracellular Release of β-Lactamase Is Responsible for the Cefazolin Inoculum Effect (CzIE) in Methicillin-Susceptible Staphylococcus aureus. in *Open Forum Infectious Diseases* **5,** S602 (Oxford University Press, 2018).

21. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol. Biol.* **1151,** 165–188 (2014).

22. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13,** e1005595 (2017).

23. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).

24. Olsen, J. E., Christensen, H. & Aarestrup, F. M. Diversity and evolution of blaZ from Staphylococcus aureus and coagulase-negative staphylococci. *J. Antimicrob. Chemother.* **57,** 450–460 (2006).

25. Kernodle, D. S., Classen, D. C., Burke, J. P. & Kaiser, A. B. Failure of cephalosporins to prevent Staphylococcus aureus surgical wound infections. *JAMA* **263,** 961–966 (1990).

26. GeneCards Human Gene Database. RELA Gene - GeneCards | TF65 Protein | TF65 Antibody. at <https://www.genecards.org/cgi-bin/carddisp.pl?gene=RELA>

27. Human Metabolome Database: Showing metabocard for UDP-N-acetyl-D-mannosamine (HMDB0013112). at <http://www.hmdb.ca/metabolites/HMDB0013112>

28. Brown, S., Xia, G., Luhachack, L. G., Campbell, J., Meredith, T. C., Chen, C., Winstel, V., Gekeler, C., Irazoqui, J. E., Peschel, A. & Walker, S. Methicillin resistance in Staphylococcus aureus requires glycosylated wall teichoic acids. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 18909–18914 (2012).

29. Peschel, A., Otto, M., Jack, R. W., Kalbacher, H., Jung, G. & Götz, F. Inactivation of the dlt Operon inStaphylococcus aureus Confers Sensitivity to Defensins, Protegrins, and Other Antimicrobial Peptides. *J. Biol. Chem.* **274,** 8405–8410 (1999).

30. Peschel, A., Vuong, C., Otto, M. & Götz, F. The D-alanine residues of Staphylococcus aureus teichoic acids alter the susceptibility to vancomycin and the activity of autolytic enzymes. *Antimicrob. Agents Chemother.* **44,** 2845–2847 (2000).

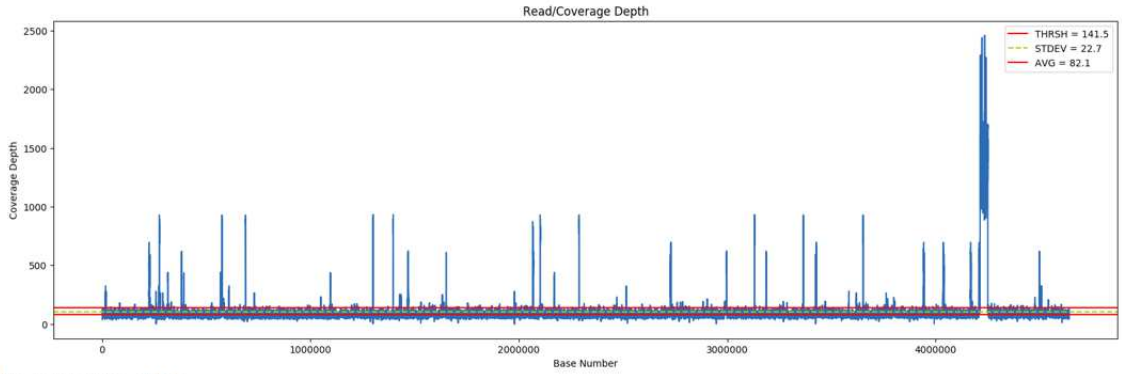
Chapter 2- Amplification Detection in AleDb Samples and Hybrid Data Validation

The Adaptive Laboratory Evolution (ALE) system allows for analysis of causal mutations acquired by a population of microorganisms subject to a controlled environmental condition. As the ALE system accumulates more experimental data and expands its database of genotype-phenotype relationships, automation becomes increasingly important. Current analysis of mutations in ALE is executed via a custom pipeline which utilizes Breseq, and while Breseq mutation prediction can reliably identify SNPs, multiple-base substitutions, and other small variations, there is no dependable amplification detection function (1, 2). The inputted data currently includes a high-quality reference genome of the parent strain and Illumina short-reads from the child strain. Because long-read sequencing is not customary in the ALE system, the purpose of this study is to develop and validate a new amplification function compatible with the ALE-Breseq pipeline that uses only the reference genome and short-read evolved strain data.

2.1: Breseq and NewAmp Amplification Detection

Because Breseq is unable to identify large copy-number variations-- or amplification-- in the child strain, the Systems Biology Research Group developed the OldDups function as a rough solution which produced a coverage map from the data by aligning the child strain short-reads to the reference genome (**Figure 2.1)**.

a) 12-95-1-1 12-95-1-1



b) 22-87-1-1 22-87-1-1

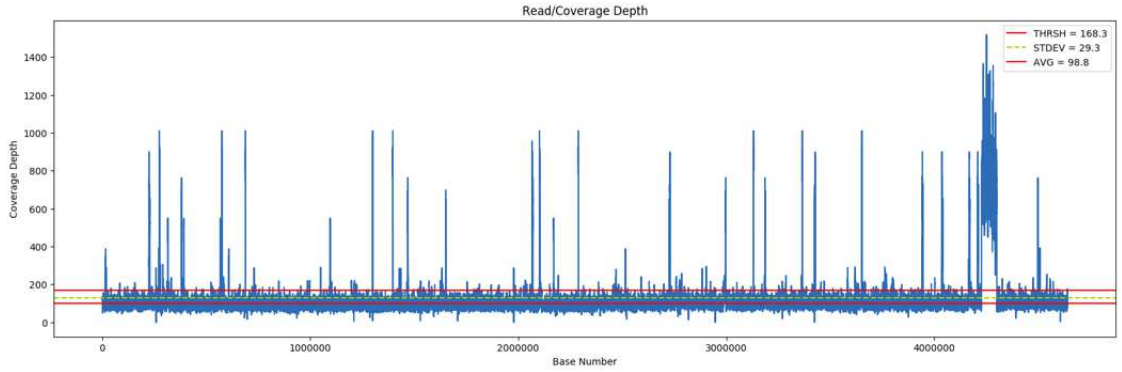


Figure 2.1- Coverage map of SvNS a) HsaPgi 12.95.1.1 and b) BmePgi 22.87.1.1 each containing single major amplification

Many of the coverage maps looked similar to the two example strains shown in **Figure 2.1**, where the large region containing a high coverage depth with respect to the average depth of the genome corresponds to a suspected amplification. In both cases shown, this occurs at about 4.2 Mb (million bases).

The OldAmps function takes as input the coverage map data containing the coverage depth per base number. It also takes a minimum amplification length (*min_amp_length*) specified by the user, the average coverage depth over the genome, and the standard deviation in coverage depth. From this, a threshold is calculated as:

$$\text{threshold} = (2 * \text{avg.cov.}) - \text{st.dev.}$$

OldAmps then reduces this list of data pairs to pairs with coverage depth values greater than the threshold. It organizes the list to group together data pairs in sequence by base number, such that a list of sequences with coverages greater than the threshold remains. The function then checks whether each sequence is greater than some *min_amp_length* and saves the sequences greater than this length threshold.

As a result, OldAmps identifies a sequence in the genome as an amplification if it is greater than the *min_amp_length* and if every base in the sequence has a coverage greater than the threshold. The *min_amp_length* was set to 300 bps by default, resulting in the detection of 59 amplifications in HsaPgi 12.95.1.1 and 82 amplifications in BmePgi 22.87.1.1. However, errors in sequencing and bias coverage may cause interruptions in the coverage maps corresponding to an amplification, and noisy data like that shown in **Figure 2.2** illustrates the need for amplification detection that allows for deviations below the threshold coverage depth.

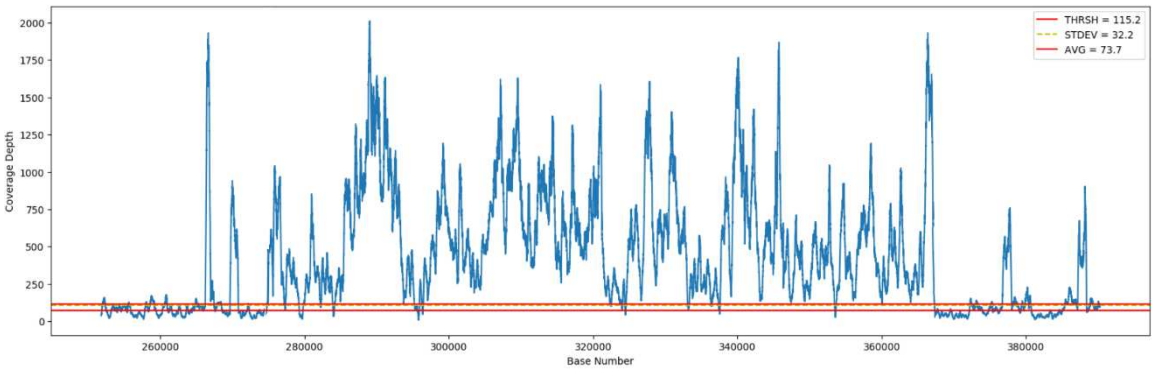


Figure 2.2- FNSCase3 1.8.1.1 coverage map of amplification demonstrating regions of coverage depth below amplification threshold

Additionally, analysis of the encoded genes in the detected amplifications from OldDups shows that many of the regions are smaller than 10 kb and are composed of no detected genes, a single gene, or very few genes-- often insertion elements. These

elements and other regulatory transposon elements occur at high frequency throughout the entire genome. Because the child strain reads that are aligned to the parent reference genome are short, they often contain little sequence information outside of the small repeated elements, and the aligner therefore aligns each instance to every matching sequence on the reference genome. As a result, the coverage depth at each of these locations is very high, though the element may only be present in a single copy at each location. This produces the frequent, narrow spikes present in every coverage map.

The NewAmps function aims to filter out small spikes caused by overalignment of regulatory elements to detect only the complete amplifications. In order to avoid detection of small spikes in the coverage map, NewAmps has a *min_amp_length* of 10 kb by default, mutable as a user-inputted parameter. Rather than considering only bases with coverage greater than the threshold, the new function considers every base and calculates a rolling average of the coverage depths. Beginning at the first base position, the algorithm checks if each base has a coverage greater than the threshold. Once one is found, it sets the current base position as the start of the potential amplification and tracks the average coverage depth, number of total bases, and number of bases greater than the threshold for the growing region. While the average coverage of the potential amplification is greater than the threshold, the region continues to grow. However, since the coverage depths of amplifications is so much greater than the average genome coverage depth, a short segment of an amplified region and a very long segment of an unamplified region may have an average depth greater than the threshold. To prevent these cases from contributing to the detected amplifications, a *quality* parameter is introduced. This parameter is the percentage of bases in the overall amplification which

must have coverage greater than the threshold. By default, this value is set to 0.85, meaning 15% of the bases may be below the threshold value in a detected amplification. To prevent this 15% from accumulating at the end of an amplification, NewAmps includes the final *last* parameter. *last* controls the end measurement of an amplification and dictates the number of consecutive bases at the end of the amplification that must have an average coverage depth greater than the threshold. The second half of this length must also be greater than the threshold coverage depth at every base. NewAmps detected 1 amplification in both cases shown in **Figure 2.1**.

2.2: SvNS and False-Negatives Test Cases

Two experiments are included in the testing and validation: SvNS and false-negatives. Both studies use the bacteria *Escheichia coli* as the target organism. In the SvNS study, genetic engineering was used to replace one of two native genes for glycolytic isomerase – *pgi* or *tpiA*-- with foreign copies from either *Homo sapiens* or *Brucella melitensis.* The strains initially struggled to grow on glucose minimal media, but all eventually evolved to utilize the foreign DNA. This often involves amplification of the foreign gene to achieve the necessary level of enzymatic activity. Here, only results for *pgi* cases are shown, due to the development of amplification following genetic engineering (3).

The false-negatives (FNS) study focuses on identifying promiscuous activity in enzymatic networks – that is, instances of flexibility in catalytic activity of alternative enzymes following the gene knockout of an enzyme that was thought to be essential. Because the newly required functions of these alternative enzymes are often secondary functions, the quantity of enzyme produced needs to be increased to cope with the

applied selective pressure. FNSCase3 follows the laboratory evolution of a gene knockout strain for the enzyme citrate synthase, *gltA*. The knockout identified the gene encoding 2-methylcitrate synthase*, prpC*, as a putative isozyme for *gltA.* Expression analysis showed that *prpC* had been significantly upregulated, suggesting the gene had been copied in the genome during an amplification (4).

2.3: Hybrid Validation and Results

In order to determine the accuracy of the mutations detected using the NewAmps function, long-read data was obtained using the Oxford Nanopore Technologies MinION DNA sequencer. These reads were sequenced using barcodes, so 12 samples could be analyzed in a single run. Each sample was fed through the hybrid validation pipeline, illustrated in **Figure 2.3**.

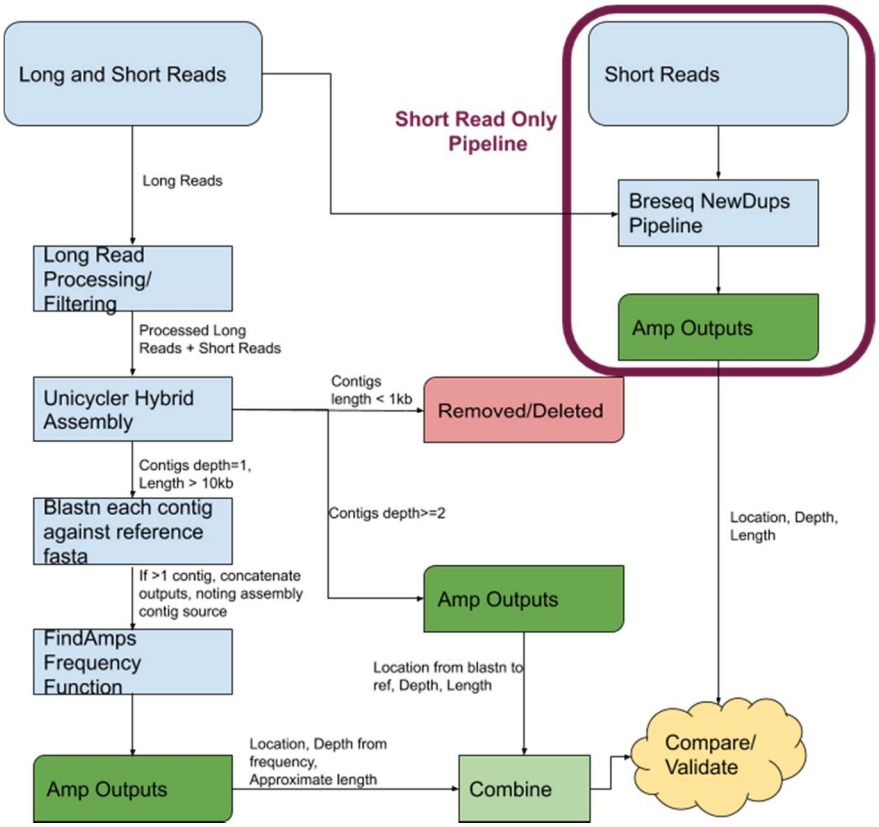


Figure 2.3: Hybrid validation pipeline

The long-reads were first demultiplexed using Guppy, an Oxford Nanopore Technologies data anlysis toolkit. The barcoding and sequencing adapters were then trimmed away using Porechop, and they were filtered using bbmap to discard reads with quality scores lower than 6 and lengths smaller than 1 kb (5). Finally, the data was inputted into the Unicycler assembler, in combination with the short-reads from the Breseq analysis, to generate a hybrid genome for each strain (1,6).

Although the goal was to generate about 8% of the total reads for each sample, the final percentages ranged from 0.68% to 23.29%, as shown in **Table 2.1**. As a result, many of the hybrid genomes remained incomplete, assembled into between 1 and 54 contigs. Unicycler outputs the sequences, lengths, and depths of these contigs, which could be divided into 3 general categories: 1) Contigs greater than *min_amp_length* with

coverage depth greater than 2; 2) Contigs greater than *min_amp_length* with coverage depth of about 1; 3) Contigs smaller than *min_amp_length*.

For all samples in this study, the *min_amp_length wa*s set to 10 kb, as determined by the NewAmps coverage maps, though this value may be adjusted by the user. The first case of contigs, with lengths greater than *min_amp_length* and depth greater than 2, represents large amplifications in the genomes. These are sequences which are not resolved into the remaining contigs, either due to lack of coverage near the ends of the contigs or more likely because the region is so large that a long-read that spans the entire sequence and both flanking regions was not generated. Without this information, it is not clear where the contig appears in the overall genome. However, because this region is aligned to a larger multiplicity of reads compared to the rest of the genome, a higher depth is reported, indicating an amplification.

The second case of contigs ,with depths greater than *min_amp_length* and with depth between 0.8 and 1.2, is less transparent; the depth near 1 seems to imply that no significant increase in coverage occurs in the contig. However, this may be because a long-read that spans the entire amplification *was* generated by the MinION, and the assembler is able to resolve the flanking regions between the amplification and the rest of the genome. To determine whether there are amplifications hidden in these contigs, the contigs are sent through blastn with the parent strain reference genomes to identify local alignments greater than the *min_amp_length* (7). The alignments are added to a frequency array according to the position on the reference. Since amplifications are sequences of the reference genome which are present in tandem as more than one copy on the child strain genome assembly, an amplification can be detected as a region in the reference genome that aligns to multiple regions of the hybrid assembly genome.

If the regions are at least 0.8 * *min_amp_length* apart, they are added to the list of identified amplifications. If no region on the reference maps to multiple regions on the hybrid contig, there is high confidence that no amplification is hidden and resolved in the contig.

The final case includes contigs with length less than *min_amp_length*. While these reads may seem insignificant, and certainly cannot encompass a complete and major amplification, these contigs may contain portions of an amplification which could not be resolved with the rest of the amplification in another one (or more) contigs. This is most likely in cases with poor hybrid genomes which assembled into many contigs. Should this occur, a significant sequence may be discarded if all contigs below *min_amp_length* are removed. To account for this, the depth measured by Unicycler is analyzed for each contig smaller than *min_amp_length* and greater than 0.1 * *min_amp_length*. If a depth is within 20% of the depth of another amplification detected from a major contig, the minor contig is sent through the blastn portion of the pipeline to find where on the reference genome it aligns. If this location is within 0.1 * *min_amp_length* of another detected amplification with matching depth, the amplification sequence is extended in the appropriate direction and the length and location of the amplification is updated. This allowed deviation from the location of the major amplification addresses the possibility of a small but meaningful contig discarded at the start of this case.

The results of the NewAmps short-read amplification detection and the hybrid validation pipeline amplification detection are summarized in **Table 2.1** for each sample tested. Also included in the table is the percentage of sequencing reads corresponding to each sample, the total number of contigs in the Unicycler hybrid assembly, the

number of contigs with depth greater than 2 and length greater than the *min_amp_length* of 10 kb, and the number of amplifications expected from the Breseq analysis, which matched the number of reads detected with the hybrid assembly validation pipeline for every sample. The percent error for each amplification is also shown, and calculated using:

$$\%Error = \frac{AmpLength_{HybridPipeline} - AmpLength_{BreseqNewAmpsDetection}}{AmpLength_{HybridPipeline}}$$

Table 2.1- Amplification measurements of all test and validation cases, where bolded values indicate significant error between Breseq NewAmps predictions and Hybrid Pipeline validation and NA, HP correspond to NewAmps and HybridPipeline, respectively.

Experiment	Strain	Flowcell%	Contigs	D>2,L>10k	#Amps	Amp Lengths			Amp Location				Amp Depth	
						NA	HP	% Error	NA Start	NA End	HP Start	HP End	NA	HP
HsaPgi	**12.95.1.1**	1.76	9	1	1	42757	37371	**0.144**	4208014	4250771	4213397	4250768	14.8	14.69
	16.77.1.1	0.68	50	1	1	17012	17012	0.000	4218435	4235447	4218435	4235447	16.3	15.79
	20.14.1.1	2.09	35	2(+1~7k)	1	35490	35433	0.002	4218567	4254057	4218555	4253988	14	13.5
	21.24.1.1	13.72	2	1	1	36754	35590	0.033	2066177	2102931	2066156	2099752	3.8	3
	22.131.1.1	1.4	2	1	1	38701	38701	0.000	4221463	4260164	4221464	4260165	3.3	3.19
	23.119.1.1	4.82	2	1	2	36748, 25041	35590, 25063	0.033, 0.001	2066185, 4225297	2102933, 4250338	2072268, 4225298	2107858, 4250361	2.6, 3.2	2.03, 3
BmePgi	20.31.1.1	0.8	54	1		17215	16887	0.019	4218540	4235755	4218541	4235428	24.3	15.42
	22.87.1.1	1.13	2			73735	73624	0.002	4227214	4300949	4227215	4296324	7.3	7.04
	2.112.1.1	8.01	1	1	1	14177	14176	0.000	4221370	4235547	4221371	4235547	18.9	6
FNSCase3	**1.8.1.1**	23.29	2	1	1	100747	92843	**0.085**	279900	380647	275641	360143	7.2	7.54

HsaPgi 12.95.1.1, HsaPgi 16.77.1.1, HsaPgi 20.14.1.1, and BmePgi 20.31.1.1 each contain 9 or more contigs in the final Unicycler assembly. In HsaPgi 20.14.1.1, two contigs have depths greater than or equal to 2 and lengths greater than *min_amp_length* = 10 kb. It also has one contig with a length of 7 kb and with depth equal to the other detected amplifications. Upon further analysis of the amplification positions, all three amplifications were joined into one amplification of about 37 kb, producing only 0.2% error as compared to the amplifications detected through the hybrid validation pipeline. The remaining three strains have only one contig in the final assembly with depth greater than or equal to 2, and after further analysis down the pipeline, each contig is evaluated to be the only final amplification.

HsaPgi 21.24.1.1, HsaPgi 22.131.1.1, BmePgi 22.87.1.1, and FNSCase3 1.8.1.1 each have 2 contigs in the final Unicycler assembly. In each case, one contig has a depth of 1, and one contig has a depth greater than or equal to 2.
Each former contig is evaluated to contain no repeated regions and each latter contig was evaluated to be the only final amplification.

HsaPgi 23.119.1.1 contains 2 contigs in the Unicycler assembly. One contig has a depth of 1, while the other has a depth greater than or equal to 2. While there initially appears to be the only amplification present, analysis using blastn reveals one resolved amplification in the contig with depth 1. The positions of these amplifications, found using blastn, indicate that they are not a single mutation, despite the similar depth values. These are the only two amplifications detected for this strain.

In the case of BmePgi 2.112.1.1, the final hybrid assembly contains one single contiguous sequence. In this case, it is known that any amplification that may be present is resolved in the assembly by the long-reads, such that the repeated region is present

the same number of times as its depth value. While this is the strongest Unicycler hybrid assembly, the amplifications are not immediately identifiable. Analysis of local alignments between the reference genome and the hybrid assembly using blastn reveals the regions of the reference genome which align to multiple regions of the hybrid assembly. In this case, a region of about 14 kb on the reference genome, beginning at about 275 kb on the reference, aligns to 6 consecutive regions on the hybrid assembly with approximately the same length. This region is the only detected amplification, and the Breseq NewAmps predicted amplification matches this result with 0.00% error.

2.4 Sample-Specific Parameter Optimization

In every sample case except two, the NewAmps function is able to accurately determine the length, location, and depth of every amplification identified from the hybrid validation pipeline using only the default parameter values. The two cases with significant discrepancy between the detection methods are HsaPgi 12.95.1.1 and FNSCase3 1.8.1.1, which produce 14.4% and 8.5% error, respectively. Every other case shows less than 3.3% error.

HsaPgi 12.95.1.1 and FNSCase3 1.8.1.1 illustrate the analytical power of NewAmps, as a simple adjustment of 1 or 2 of the parameters is sufficient to decrease the error between each pipeline to 0.0% and 0.6%, respectively. In both cases, the *min_amp_length* is left at the default value of 10 kb. The other two parameters, *quality* and *last* are adjusted according to **Table 2.2**, where the bolded values differ from the default.

Table 2.2- Optimal Parameters for HsaPgi 12.95.1.1 and FNSCase3 1.8.1.1

Experiment	Strain	last, quality	Amp Lengths			Amp Location				Amp Depth	
			NA	HP	% Error	NA Start	NA End	HP Start	HP End	NA	HP
HsaPgi	12.95.1.1	0.03, **1**	37376	37371	0.000	4213395	4250771	4213397	4250768	16.3	14.69
FNSCase3	1.8.1.1	**0.3, 0.84**	92305	92843	0.006	274898	367203	275641	360143	7.9	7.54

The NewAmps outputs acquired using the default parameters and the optimal parameters are shown in **Figure 2.4** for HsaPgi 12.95.1.1. In the initial run, a small region preceding the desired region is included in the detected amplification because the rolling average of this region is greater than the threshold value and the proportion of bases greater than the threshold is greater than the *quality* parameter. In order to remove this unwanted region from the amplification, the *quality* is increased from the default value of 0.85 to the optimal value of 1.0, meaning every base position in the detected amplification must be greater than the threshold coverage value. The resulting amplification matches that detected in the hybrid validation pipeline almost exactly in length, location, and depth.

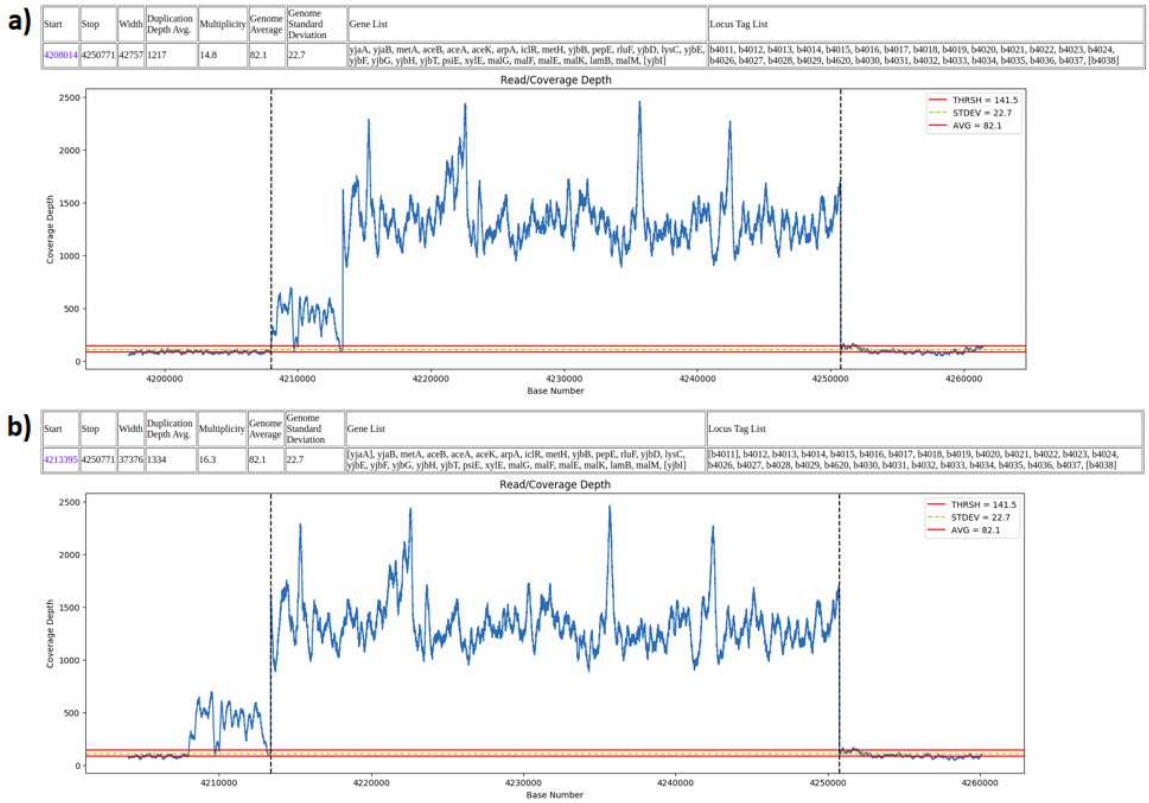


Figure 2.4- HsaPgi 12.95.1.1 coverage map with high quality amplification under a) standard and b) optimal parameters

Figure 2.5 shows the NewAmps output using the default and optimal parameters for FNSCase3 1.8.1.1. This coverage map is the noisiest by far of any sample analyzed. Because it contains frequent instances of coverage drops within the expected amplification, the proportion of bases below the threshold value is likely larger than the *quality* parameter allows. This is especially problematic near the start of the amplification, as indicated by the spikes of coverage depth just before the amplification boundary in **Figure 2.5a**. To allow for increased noise in the detected amplification, the *quality* parameter is decreased slightly from 0.85 to 0.84. Additionally, the detected amplification extends past the visually obvious cutoff. This is due to the size of the amplification. At nearly 100 kb, this is the largest amplification event in the sample set.

The *last* parameter, which controls the end measurement of an amplification, requires that some proportion of the *min_amp_length* have an average coverage greater than the threshold at the end of the detected amplification. It also requires that latter half of this region have every single base position greater than the threshold. The *last* parameter is designed to prevent exactly these cases, but because the amplification is so large compared to the *min_amp_length,* only a very small region, comparatively, is required to measure above the threshold coverage value at the end of the amplification. This is evident in the small spike in coverage depth just below the trailing boundary of the detected amplification. To prevent the low coverage region near the end from being included in the detected amplification, the *last* parameter is increased from 0.03 to 0.3. The resulting NewAmps amplification using optimal parameters contains only 0.6% error as compared to the amplification detected by the hybrid validation pipeline.

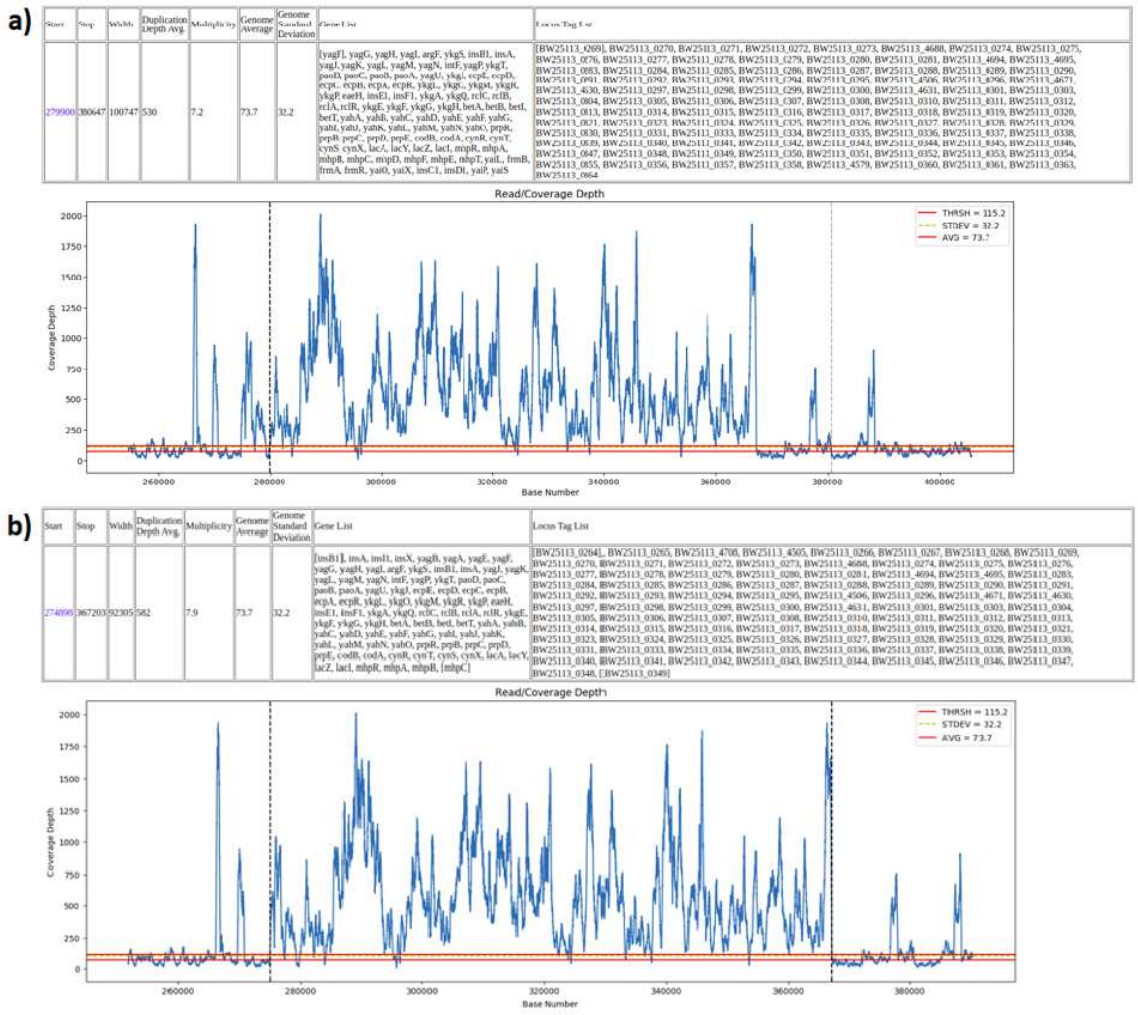


Figure 2.5- FNSCase3 1.8.1.1 coverage map with noisy amplification under a) standard and b) optimal parameters

In general, parameter changes can be anticipated by observing the genome-wide coverage map before amplification detection. If the data appears so noisy that even regions of suspected amplifications appear to dip below the threshold value, indicated by the upper red line, then the *quality* may need to be reduced to avoid breaking apart amplifications. Alternatively, if the data has few fluctuations and regions of suspected amplifications appear to remain above the threshold coverage for the duration of the mutation, the *quality* may be increased to 1 to achieve ideal amplification boundaries.

The *last* parameter can also be anticipated by observing the ends of suspected amplifications. If the ends are particularly noisy and there may be a region of very low coverage before the trailing amplification boundary, *last* may assume a lower value to ensure that the amplification still ends with a coverage depth above the threshold while allowing recent coverage decreases. On the other hand, if there is a region outside and after the suspected amplification that happens to have a high coverage, *last* may be increased to mandate that a larger region at the end of the amplification have every base coverage greater than the threshold value and preventing the region from inclusion.

Because the NewAmps function typically takes under a minute to analyze a full genome, many runs with varying parameter values may be performed on a single sample to gauge optimal parameterization. The testing data suggests that parameter values that are successful for one sample in a project are likely to give strong results for others in the same project.

2.5: Flanking Region Analysis

The high accuracy of identified amplifications using only short-read data from a child strain and a high-quality reference genome from the parent strain allows for greater analytical capabilities in the ALE system and other research focused on automated mutation detection. The precise amplification boundaries produced by NewAmps provides the opportunity to analyze flanking region of amplification events efficiently and reliably.

Start	Stop	Width	Duplication Depth Avg.	Multiplicity	Genome Average	Genome Standard Deviation	Gene List	Locus Tag List
2066177	2102931	36754	726	3.8	193.2	41.2	insH1, yoeG, yoeH, yoeA, insD1, insC1, yeeP, flu, yeeR, yeeS, yeeS, cbeA, cbtA, yeeW, yoeD, yoeF, yeeX, yeeA, sbmC, dacD, sbcB, yeeD, yeeE, plaP, yoeI, yeeY, yeeZ, yoeB, yefM, yefM, hisG, hisD, hisC, hisB, hisH, hisA, hisF, hisI, wzzB, ugd, gnd, [wbbL], insH1	b1994, b4640, b4641, b4582, b1996, b1997, b1999, b2000, b2001, b2002, b2002, b2004, b2005, b2006, b4642, b4538, b2007, b2008, b2009, b2010, b2011, b2012, b2013, b2014, b4678, b2015, b2016, b4539, b2017, b2017, b2019, b2020, b2021, b2022, b2023, b2024, b2025, b2026, b2027, b2028, b2029, [b4571], b2030

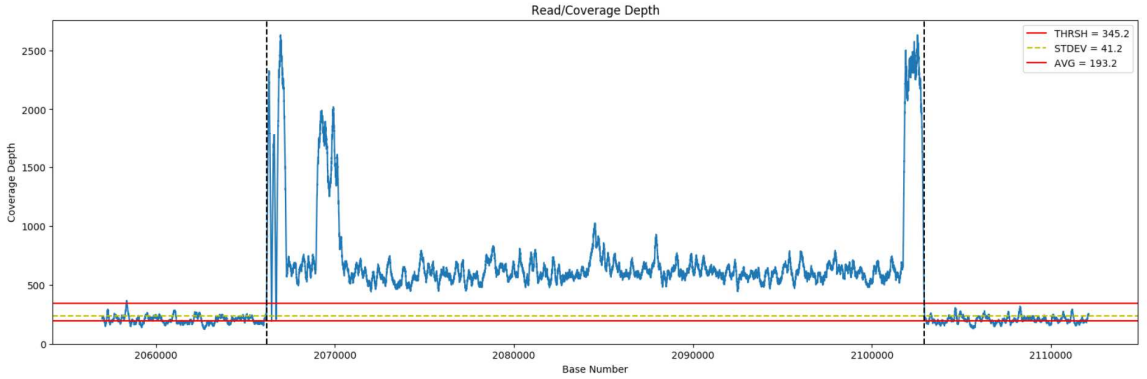


Figure 2.6- Insertion elements in flanking regions of HsaPgi 21.24.1.1

In addition to the amplification location, length, and depth, NewAmps returns information on the genes present in the amplification. The flanking regions, corresponding to genes at the start and end of each gene list, may contain information on the mechanism of amplification formation. When considered with the defined conditions of the ALE experiment, these flanking regions can help to elucidate the factors necessary for formation of an amplification, which may be useful in inducing desired amplifications in the future using only adaptive laboratory evolution.

Analysis of the flanking regions in the HsaPgi 21.24.1.1 amplification shows that the amplification begins and ends with the insertion element insH1. This suggests that the insertion element may play a key role in the mechanism of the acquired amplification. The presence of insertion element insH1 in flanking regions also occurs in HsaPgi 23.119.1.1, which has an amplification of about 37 kb at about 2.07 Mb— the same length and location on the reference genome as HsaPgi 21.24.1.1. This indicates not only a potentially converging genotype used by *E. coli* to upregulate a *Homo sapien*

copy of the *pgi* gene following the knockout of its own, but also a common mechanism of acquiring the genotype (3). Increased understanding of these mechanisms facilitates understanding and identification of causal mutations in the ALE system.

Chapter 2 References

1. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol. Biol.* **1151,** 165–188 (2014).

2. Phaneuf, P. V., Gosting, D., Palsson, B. O. & Feist, A. M. ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. Nucleic Acids Res. **47**, D1164–D1171 (2019).

3. T.E. Sandberg, R. Szubin, P.V. Phaneuf, B.O. Palsson. Synthetic Cross-Phyla Gene Replacement and Evolutionary Assimilation of Major Enzymes in *E. coli*. Submitted to *Nature Ecology & Evolution*.

4. Guzmán, G. I., Utrilla, J., Nurk, S., Brunk, E., Monk, J. M., Ebrahim, A., Palsson, B. O. & Feist, A. M. Model-driven discovery of underground metabolic functions in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. **112**, 929–934 (2015).

5. Bushnell, B. BBMap. (2015). at <https://sourceforge.net/projects/bbmap/>

6. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13,** e1005595 (2017).

7. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. BLAST+: architecture and applications. BMC Bioinformatics **10**, 421 (2009).