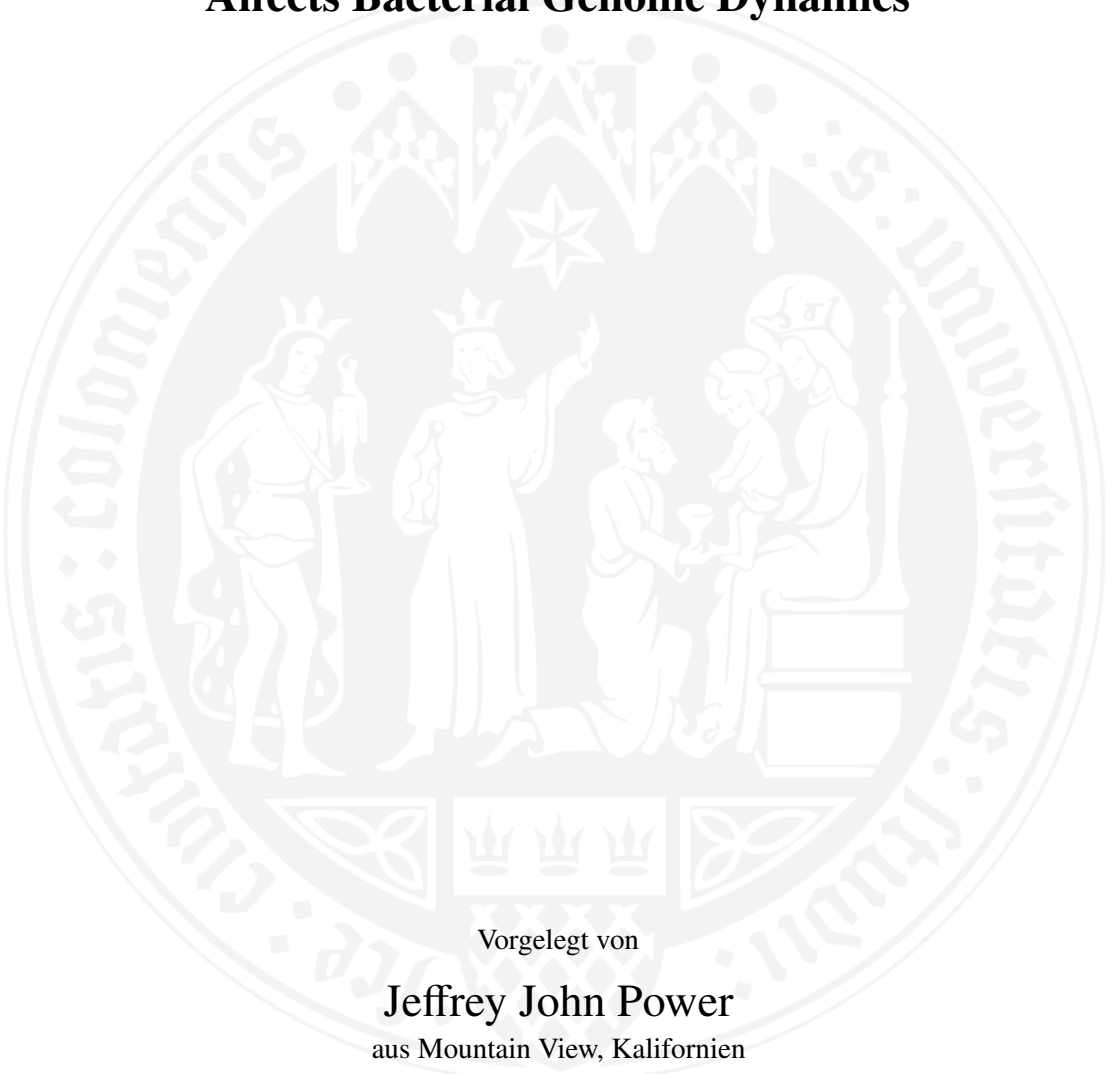
Universität zu Köln
Mathematisch-Naturwissenschaftliche Fakultät
INAUGURAL-DISSERTATION ZUR
ERLANGUNG DES DOKTORGRADES (DR. RER. NAT.)

Horizontal Gene Transfer Between Subspecies Affects Bacterial Genome Dynamics



Vorgelegt von

Jeffrey John Power

aus Mountain View, Kalifornien
Vereinigte Staaten von Amerika

Köln, 2018

Erste Gutachterin:	Prof. Dr. Berenike Maier
Zweiter Gutachter:	Prof. Dr. Tobias Bollenbach
Tag der mündlichen Prüfung:	25.10.2018

Acknowledgements

"Every time I tell my story, I begin with you." – **Broods**

While I may not find the words to completely express the fullness of my gratitude, to all parties involved, I know where to begin. **Berenike**, thank you for letting me grow, not only as a scientist but also learn how to properly behave in the (pseudo)-business world. From better understanding how to design and see a project to the end, to learning how to tell your employer you need sick leave, you have been encouraging and insightful. Thank you for making the transition towards the real, post-graduate-degree world as smooth and harmless as possible – for a now twenty-nine-year-old that still dyes his hair neon colors! While I may not aspire to a career in marketing, I have learned numerous techniques and overall the crucial importance of explaining my research such that my specific audience will be able to go away with my take-home message. Thank you for all of your support.

"After all of this is gone, who would you rather be: The Beatles or The Rolling Stones?" – **Metric**

Tobias, thank you for reading through this beast of a document and concocting questions to make me quiver in my boots. I have benefited from your experimental and statistical knowledge.

"The star maker says, 'it ain't so bad;' the dream maker's going make you mad; the spaceman says, 'Everybody look down. It's all in your mind.'" – **The Killers**

Michael, thank you for our numerous chats about null models and robust statistics. I may not be the one to design the model in the future, but I will certainly be better equipped to spot its short comings and flaws.

"I was up late night ballin', countin' up hundreds by the thousand" – **Vince Staples**

Team Bacillus co-founder **Melih**, thank you. I could not have asked for a better partner in diva cloning crime, and to think it just so happened this way. Thank you for vibrational jam sessions in high fidelity on the Tecan, effortless division of labor, countless descriptions of bacterial strains, friendly company alongside evolving bacteria, the freshest of fresh media, and the uncanny ability to keep they party going (often with the help of those sharp distilled agave juices).

"When gravity seduced me and brought me to this place, I could swear we were together, in an entirely different age." – **Darren Hayes**

Fernanda, thank you for making this collaboration. On the rare occasions I couldn't motivate myself to log into cheops, your enthusiasm for the project reminded me of the good times. You can fit more than one can imagine with one parameter, a physical explanation, and no sleep. I've also thanked Christa for sending a message from the future to make sure that this collaboration came to be.

"Temperature elevated. The air is suffocating. Ego dominating a fear to medicate it. It's not complicated, do what you want." – **The Presets**

Viera, I am in debt to you for all of my knowledge of genome alignment, sequencing pipelines, and variant calling pitfalls. Thank you for being a wizard with fastq files and lighting fast in responding with multiple solutions to my at times substantial road blocks. I will be forever more telling posterity that I learned said trick from Viera. Maybe I'll leave out that you can pull off the same trick at least ten times faster than I can.

"It's like a jungle sometimes, it makes me wonder. . . ." – **Mickey Avalon**

Thorsten, from the mundane to the science, from the morning dawn to the midnight blues, thank you for keeping my experiments running and the laughs.

"Cribbo in the Hills, table full of bills, whatever you want." **– Schoolboy Q**

Gaby, thank you for everything from fully planned and executed competition assays, to meticulous data worksheets, to the best vacation tips.

"Knowledge of the sea-ways, knowledge of how the water flows. Whoever coined the phrase has never had to brave the snow." **– Owen Pallett**

Robert and **Tom**, I want to plot number of times the cow has jumped over the moon in a given month as a function of the length of all the bacterial genomes living in its stomachs during that period. It should be straight forward.... Thank you for not only listening but also trying to help, even when the desires became absurd.

"You don't have to waste your energy, we can just be rockin'." **– Weeknd**

Christina, who would have predicted our friendship with such formal beginnings such as Dear Mrs. Wetzels would have blossomed as it did. Thank you for the tunes, FUNKNROLL, laughs, and wildly fierce looks while keeping the administrative monsters at bay.

"I'm on fire and now I think I'm ready, to bust a move. Check it out I'm rockin' steady." **– Motion City Soundtrack**

Niklas, thank you for reminding me to focus on my strong suits, pipetting (both in the lab and in silico), and helping me to relax after hours.

"Ohoooo, let's live in the moment, come back Sunday morning with that soul to sell. **– Portugal. The Man**

Christof, thank you for letting me rely on your vast biochemical expertise much like one would a calculator. Your relaxed, but still informed, approach to the lab was a lifestyle to learn from and strive for.

"They notice what we're wearing. We notice that they're staring. Hotter than the A-list; next up on our hit-list. Baby, we're gonna blow their minds tonight." **– Britney Spears**

Lena, thank you for putting some spice into the office life. While I am still working on convincing you of the beauty of French fries and a milkshake, you figured me out lickity split and we were dancing till the cows came home. Thank you for paving the way to finish in under five years.

"You know that it's hot, property. That one's hot, hot, hot, hot, property. 'Keep me on top; good to me." **– Jamiroquai**

AG Maier, from the fresh faces to the veterans, thank you for taking a walk on the wild side with me. I'll bet you can count on one hand the number of coworkers you'll have that tap dance.

"The difference between delight and loneliness could be a one way flight, in these San Luis Obispo nights." **– Scissor Sisters**

Thank you **Dr. Christiansen**, for putting a realistic perspective on my ambitions. I didn't always listen (shorts and dress shoes – I am wearing such an outfit while typing this) but am thankful for your continued efforts and guidance. More of it stuck that you think. **Dr. Costanzo**, I have come to accept that your ninja ways cannot be learned but are inherited. Nevertheless, it is a level of perfection that one can strive for; thank you. **Dr. Kat**, thank you for teaching me to appreciate the subtle things in life, like being able to leave a lab without turning in a report and finding the world's best chocolate croissant. To **he who is everywhere and anywhere**, and that we be-aware of. I can't deny that you are an excellent teacher. We should meet up.

"Qui dit etude dit travail. Qui dit argent dit dépenses. Qui dit fatigue dit réveille encore sourd de la veille. Alors on sort pour oublier tous les problèmes. Alors on danse." **– Stromae**

To my tap crew scattered over Bonn, Köln, and Düsseldorf, thank you for keeping me sane. **Cecile**, **Julia**, **Kristina**, and **Sashima**, thank you for making the dance classes more than just a fitness class and answering honestly, if I dared ask if there was a pullback in that step.

"Wenn der Art Direktor komisch guckt und aus dem fetten Deal wird leider nichts. Ich unterhalte mich mit euch. Doch, wir sind nur wegen den Freigetränken hier." **– Von Wegen Lisbeth**

Sven and **Lena**, brothers from another mother, thank you for being kickass neighbors. I'm so happy occasionally asking one another to water the plants or pick up mail turned into such a rewarding friendship.

"There's layers to this – player. Tiramisu. Let my coat tail drag but I ain't tearing my suit." **– Macklemore & Ryan Lewis**

Toby and **Stephi**, it looks like the day has finally come. Thank you for always being just around the corner for a pick me up meal, happy hour special, vacation to the other side of the globe, 2 AM hike, or simply a chat.

"I'll take hold of my enemies; tape their obscenities; kiss them; and leave them like lovers who've gone." **– Darren Hayes**

Ying-Ying one would start to think that the Gods were on our collective side considering how we haven't be able to shake each other after all these years. It would seem like those double bonds are stronger than even the text books imagined. Thank you for hanging in with me for the stories for listening for your "never fainting support helping me" for proof reading this beast of a document and for your love of commas. Keep raging on what's cheaper than free.

"Here I stand, miles apart. Distant land, staring at the sun. You're not here, but we share the same one. One thing's true, just like you, there's only one." **– Mika**

Mom and **Dad**, your thank you would need to be at least the sum of all of those already listed. Thank you for seeing me through this academic journey, letting me wander off to a distant land, being supportive of and interested in (within reason, biophysics isn't everyone's cup of tea) my research, and always just a phone call away. It might not be what you want to hear, but I wouldn't have had the confidence to make it this far without your constant support. Thank you.

"I know it's late but we need another taste. Breakfast can wait." **– His Purple Majesty A –Z**

Abstract

Different bacterial subspecies live in close proximity in soil. Horizontal gene transfer enables them to exchange genetic information. Little is known about the efficiency of gene transfer and whether it occurs randomly across the genome.

In the first part of this thesis, we designed an evolution experiment and analysis algorithm to characterize genome dynamics in the presence of a different subspecies. *Bacillus subtilis* is naturally competent for transformation: it takes up DNA from its environment and integrates segments into its chromosome by recombination. Eight clonal populations of *B. subtilis* 168 (Bsu168) were evolved with genomic DNA from *B. subtilis* W23 (BsuW23) for 21 cycles. To minimize variability in transformation rates as a function of time, we generated a strain with inducible competence, and competence was induced once per cycle. Evolved cultures had more than 100 homologous recombination events, per replicate, and the length of recombined segments was exponentially distributed with a characteristic length of $3500\,\mathrm{bp}^{-1}$. Average recipient genome replacement occurred at a constant rate over the course of the experiment, averaging 0.47% replacement every cycle. In addition to homologous recombined-segments, de novo segments from the BsuW23 auxiliary genome and de novo variants were also detected. The de novo segments from the auxiliary genome had a mean length of 2.2 kbp. Bacteria evolving in the presence of BsuW23 DNA showed five times as many de novo variants in regions upstream of genes and five times as many missense indels, compared to control samples receiving either no DNA or self (Bsu168) DNA. Of those upstream mutations, 75% were inside a recombined segment. This hinted at the possibility of those mutations being compensatory mutations, as upstream and missense mutations are likely to affect gene expression levels. We conclude that the recipient genome was replaced by donor genome at a constant rate and constant segment-length distribution, up to a total of 10% genome replacement. Introduction of de novo variants is likely to affect the levels of gene expression.

The probability of replacing a specific gene was in good agreement with a binomial distribution, suggesting that replacement occurred close to random across the genome. However, we found important deviations from random integration. At both the single-cell level and the population level, we obtained evidence that homologous recombination does not occur stochastically. At the single-cell level, imported segments had a higher average identity, 93.6 %, than the Bsu168 and BsuW23 inter-subspecies identity of 92.4%. Interestingly, recombined segments had one end of integration with a significantly higher identity than expected from simulations using the same length distribution. We found that the increased sequence identity extends to roughly 500 bp. The bias towards higher than average sequence identity is most likely caused by the recombination process. At the population level, we found further evidence that recombination did not occur randomly. On one hand, several genes, such as *leu* and *eps*, were presumptively selected for, as they were replaced in nearly all replicates. On the other hand, recombined donor segments were underrepresented in prophages and mobile-elements genes, most likely because they correspond to auxiliary genes in each subspecies. Essential genes were overrepresented, plausibly because essential genes have a higher average identity compared to nonessential genes. There was a preference to replace genes and operons fully. Two-thirds of all affected genes and operons were fully replaced. Full gene replacement is explained, in part, by the average import length being 1.9 kbp, ~2

genes. The average operon length is also comparable, 3.2 kbp. To summarize, homologous replacement is biased to higher than average sequence-identity. Given the length distribution of recombined segments, the generation of hybrid genes is less likely than the replacement of full genes. Only for a few genes is there evidence of selection.

In the second part of this thesis, we investigated the cost of competence for transformation in the stationary growth phase. We characterized the competition dynamics of strains with various probabilities of entering the K-state, to quantify the effect the K-state has on fitness in *B. subtilis*. Relative fitness was found to decrease with increasing probability of entering the K-state, during both the exponential and stationary growth phases. Using a microfluidic chamber, we were able to characterize generation times in the stationary phase for both K-state and non-K-state cells. We found a strong cost of competence due to growth inhibition, even in the stationary state. These findings emphasized that the stationary phase is dynamic.

We conclude that gene transfer between subspecies of *B. subtilis* is highly efficient, with 10% of the chromosome being replaced in a total of 42 h of competence during the evolution experiment. DNA uptake in *B. subtilis* 168 from donor strain *B. subtilis* W23 occurred at a constant rate and fairly random. Recombined segments had higher identities than the between-subspecies average, pointing to the importance of sequence divergence in recombination. Several alleles were replaced in the majority of the replicates, indicative of a selection advantage. It has been proposed that recombination is beneficial because it can cure the recipient from prophages. In contrast, here we observed that replacement of essential genes was overrepresented, while prophage genes were hardly affected. This suggested that gene transfer between subspecies functions as genome maintenance. The appearance of variants, including indels, in possible regulatory regions calls for studies of the effects of recombination on transcription. We speculate that recombination may be suppressed between subspecies by other factors, such as quorum sensing.

Zusammenfassung

Verschiedene bakterielle Unterarten leben im Erdreich in unmittelbarer Nähe zueinander. Horizontaler Gentransfer ermöglicht es ihnen, genetische Informationen auszutauschen. Über die Effizienz des Gentransfers und darüber, ob er zufällig im Genom auftritt, ist wenig bekannt.

Im ersten Teil dieser Arbeit wurde ein Evolutionsexperiment mit den benötigten Analysealgorithmen entwickelt, um die Genomdynamik in Anwesenheit einer anderen Subspezies zu charakterisieren. *Bacillus subtilis* zeigt natürliche Kompetenz: *B. subtilis* kann DNA aus seiner Umgebung aufnehmen und durch Rekombination Abschnitte in sein Genom integrieren. Acht klonale *B. subtilis* 168 (Bsu168) Populationen wurden für 21 Zyklen mit genomischer DNA aus *B. subtilis* W23 (BsuW23) evolviert. Um die Variabilität der Transformationsrate zu minimieren, wurde ein Stamm mit induzierbarer Kompetenz generiert. Die Kompetenz wurde einmal pro Zyklus induziert. Die evolvierten Kulturen wiesen mehr als 100 homologe Rekombinationsereignisse je Replikat auf. Die Länge der rekombinierten Segmente war exponentiell verteilt mit einer charakteristischen Länge von $3500\,\mathrm{bp}^{-1}$. Der durchschnittliche Austausch des Empfängergenoms trat über den Verlauf des Experiments mit einer konstanten Rate auf, wobei pro Zyklus durchschnittlich 0,47% ausgetauscht wurden. Zusätzlich zu den durch homologe Rekombination eingefügten Segmenten wurden neue Abschnitte aus den nur im Genom von BsuW23 vorkommenden Genen sowie völlig neue Varianten des Ausgangsgenmaterials detektiert. Die neuen Segmente aus spenderspezifischen Genen besaßen eine mittlere Länge von 2,2 kbp. Bakterien, die in Gegenwart von BsuW23 DNA evolviert wurden, zeigten im Vergleich zu Kontrollproben, welchen entweder keine DNA oder eigene (Bsu168) DNA zugegebene wurde, fünfmal so viele neue Varianten in von Genen aus gesehen strangaufwärts gelegenen Regionen sowie fünfmal so viele nicht synonyme Insertionen oder Deletionen. Von den Mutationen in strangaufwärts gelegenen Regionen befanden sich 75% in rekombinierten Abschnitten. Dies legte die Vermutung nahe, dass es sich hierbei um kompensatorische Mutationen handelt, da strangaufwärts gelegene und nicht synonyme Mutationen wahrscheinlich einen Einfluss auf das Genexpressionsniveau haben. Zusammenfassend konnten wir zeigen, dass bis zu 10% des Empfängergenoms durch das Spendergenom ausgetauscht wurde. Die Austauschrate und die Verteilung der Segmentlänge sind dabei konstant. Die Entstehung neuer Varianten hat mit hoher Wahrscheinlichkeit Einfluss auf das Genexpressionsniveau.

Die Wahrscheinlichkeit, dass ein spezifisches Gen ausgetauscht wurde, stimmt mit einer Binomialverteilung überein, was nahelegt, dass der Ort an dem Genmaterial ausgetauscht wird zufällig über das Genom verteilt ist. Dennoch fanden wir sowohl auf Einzelzellniveau als auch auf der Populationsebene wichtige Hinweise auf Abweichungen von einer rein zufälligen Integration. Auf Einzelzellniveau als auch auf der Populationsebene haben wir Hinweise darauf gefunden, dass homologe Rekombination nicht stochastisch auftritt. Auf dem Einzelzellniveau ist die mittlere Übereinstimmung zwischen dem importierten Abschnitt und der Ausganssequenz mit 93,6% höher als die zwischen den beiden Spezies Bsu168 und BsuW23 von 92,4%. Interessanterweise hatten rekombinierte Segmente ein Ende mit einer signifikant höheren Übereinstimmung, als es anhand von Simulationen auf Grundlage derselben Längenverteilung zu erwarten wäre. Die erhöhte Sequenzübereinstimmung erstreckt sich über eine Länge von ungefähr 500 bp. Die Tendenz zu einer erhöhten mittleren Übereinstimmung wird vermutlich durch den

Prozess der Rekombination hervorgerufen. Auch auf der Populationsebene haben wir Hinweise darauf gefunden, dass die Rekombination nicht zufällig auftritt. Einerseits wurde vermutlich auf einige Gene wie *leu* und *eps* selektiert, da diese in nahezu allen Replikaten ersetzt wurden. Andererseits wurden in Prophagengenen und mobilen Genelementen vergleichsweise wenig Spenderabschnitte integriert, vermutlich da es sich bei diesen häufig um Gene handelt, welche ausschliesslich in der Spendersequenz zu finden sind. Essentielle Gene wurden hingegen besonders häufig ausgetauscht, da diese im Vergleich zu nicht essentiellen Genen eine besonders hohe Übereinstimmung zwischen den Subspezies aufweisen. Es konnte eine Präferenz zum vollständigen Austausch von Genen und Operons festgestellt werden. Zwei Drittel aller betroffenen Gene und Operons wurden vollständig ersetzt. Der vollständige Austausch der Gene kann zumindest teilweise durch die durchschnittliche Importlänge von 1,9 kbp erklärt werden, welche ungefähr der Länge von zwei Genen entspricht. Auch die mittlere Operonlänge ist mit 3,2 kbp mit der mittleren Importlänge vergleichbar. Zusammenfassend wurde ein Trend der homologen Rekombination zu höheren Sequenzähnlichkeiten festgestellt. Ausgehend von der Längenverteilung der integrierten Abschnitte ist die Erzeugung von Hybridgenen weniger wahrscheinlich, als der vollständige Austausch. Hinweise auf Selektion konnten für einige wenige Gene beobachtet werden.

Im zweiten Teil der Arbeit wurden die Kosten der Kompetenz in der stationären Wachstumsphase untersucht. Dabei wurde die Konkurrenzdynamik verschiedener Stämme, mit unterschiedlichen Wahrscheinlichkeiten in den K-Zustand zu wechseln, charakterisiert, um den Effekt des K-Zustandes auf die Fitness in *B. subtilis* zu quantifizieren. Es wurde festgestellt, dass sowohl während der exponentiellen als auch der stationären Wachstumsphase die relative Fitness mit steigender Wahrscheinlichkeit sich im K-Zustand zu befinden abnahm. Unter Verwendung einer Mikrofluidikkammer wurden für Zellen in der stationären Phase, die sich jeweils entweder im K-Zustand oder nicht im K-Zustand befanden, die Generationszeiten bestimmt. Wir konnten zeigen, dass Kompetenz sogar im stationären Zustand durch Wachstumsverminderung starke Kosten aufweist. Dies zeigt, dass die stationäre Phase dynamisch ist.

Insgesamt schlussfolgern wir, dass der Gentransfer zwischen *B. subtilis* Subspezies hocheffizient abläuft, da während der 42 stündigen Kompetenz während des Evolutionsexperiments insgesamt 10% des Chromosoms ausgetauscht wurden. Die Aufnahme der DNA des Spenderstammes *B. subtilis* W23 durch den Empfänger *B. subtilis* 168 erfolgte nahezu zufällig und mit einer konstanten Rate. Die rekombinierten Abschnitte hatten eine höhere genetische Übereinstimmung, als die mittlere Übereinstimmung zwischen den Subspezies, was die Bedeutung der Sequenzdivergenz hervorhebt. Der Austausch mehrerer Allele in der Mehrzahl der Replikate weist auf einen selektiven Vorteil hin. Im Allgemeinen wurde angenommen, dass Rekombination von Vorteil sein kann, da es den Empfänger von Prophagen heilen kann. Im Gegensatz dazu zeigen unsere Ergebnisse, dass sich der Austausch genetischen Materials hauptsächlich auf essentielle Gene beschränkt, während Prophagengene kaum betroffen waren. Dies legt die Vermutung nahe, dass Gentransfer zwischen Subspezies hauptsächlich der Erhaltung genetischen Materials dient. Möglicherweise kann die Rekombination zwischen Subspezies auch durch andere Faktoren unterdrückt werden, wie beispielsweise durch die Wahrnehmung der Zelldichte. Das Auftreten von genetischen Variationen, einschließlich Insertionen und Deletionen, in möglichen regulatorischen Regionen zeigt die Notwendigkeit von Studien zum Effekt der Rekombination auf die Transkription.

Contents

Abstract vii

Zusammenfassung ix

1 Introduction 1
1.1 Horizontal Gene Transfer . . . 1
1.1.1 The Role of Recombination in Evolution . . . 1
1.1.2 Bacterial Transformation and DNA Uptake Mechanisms . . . 3
1.1.3 Competence for Transformation in *Bacillus subtilis* . . . 4
1.1.4 Rates of Transformation Depend on Phylogenetic Distance . . . 6
1.1.5 Fitness Effects of Recombination . . . 7
1.2 Experimental Evolution . . . 9
1.2.1 Bacterial Fitness . . . 10
1.2.2 Fitness Landscapes and Epistasis . . . 14
1.2.3 Characterizing Fitness Experimentally . . . 15
1.2.4 Horizontal Gene Transfer in Experimental Evolution . . . 17
1.3 Aims of This Study . . . 19

2 Methods – Evolution Experiment 21
2.1 Experimental Methods . . . 22
2.1.1 Strains, Media, and Growth . . . 22
2.1.2 Whole Genome Sequencing . . . 25
2.1.3 Evolution Experiment Design . . . 25
2.2 Computational and Analysis Methods . . . 26
2.2.1 Sequencing pipelines . . . 26
2.2.2 Orthologous Recombination (CNP) Algorithm . . . 30
2.2.3 SNP Flank Length Bias . . . 39
2.2.4 Orthologous Recombination (SPI) Algorithm . . . 41
2.2.5 De novo Insertions Algorithm (Auxiliary Genes Algorithm) . . . 42
2.2.6 Contaminated or Mislabeled Samples . . . 44

3 Methods – Population Dynamics Experiment 45
3.1 Experimental Methods . . . 45
3.1.1 Strains and Media . . . 45
3.1.2 Population Dynamics – Experiment Design . . . 46
3.2 Computational and Analysis Methods . . . 48
3.2.1 Stationary Phase Dynamics – Selection Coefficients . . . 48
3.2.2 Single Cell Microscopy – Image Analysis . . . 48

4 Results – Evolution Experiment 51
4.1 Orthologous Recombination and De novo Insertions Occur in Multitude . . . 51
4.1.1 Cycles 9, 15, and 21 . . . 52
4.1.2 Time Lapse Replicates, W1, 3, 4, and 5 . . . 57
4.2 CNP Properties: Identity, Composition, and Gene Function . . . 60
4.2.1 CNPs Recombine on Sections of Higher Identity . . . 60

4.2.2 CNP Composition . . . 62
4.2.3 Essential Genes are Overrepresented and Prophages are Underrepresented in CNPs 66
4.3 CNP Recombination Probability . . . 67
4.3.1 Several Putative Hot/Cold Spots Detected in Overall Fairly-Random Gene Replacement . . . 67
4.3.2 CNPs Have Smaller SNP Density on One Side of the Segment . . . 70
4.3.3 Genes are Most Likely Completely Replaced . . . 72
4.4 The Presence of Donor DNA Increases the Number of De novo Variants . . . 73

5 Results – Population Dynamics Experiment 77
5.1 The K-state Confers a Fitness Cost During the Stationary Phase . . . 77

6 Discussion 81
6.1 Evolution Experiment . . . 81
6.1.1 Homologous Recombination Occurs at a Constant Rate with an Exponential Distribution . . . 81
6.1.2 A Constant Recombination Rate Implies Minor Fitness Changes and Small Epistatic Costs . . . 82
6.1.3 Auxiliary Regions are Imported Randomly into the Genome . . . 83
6.1.4 Recombination is Biased towards Higher Identities . . . 84
6.1.5 Recombination Hot Spots Putatively Correspond to Fitness Advantages . . . 85
6.1.6 Essential Genes are Preferentially Replaced in Evolved Strains . . . 86
6.1.7 Genes are More Likely to be Completely Replaced . . . 87
6.1.8 De Novo Variants in Intergenic Regions of CNPs Occur Simultaneously with CNP Integration . . . 88
6.1.9 A Suitable Method for the Detection of Gene Transfer Across *B. subtilis* Subspecies 89
6.2 Population Dynamics Experiment . . . 90
6.2.1 Stochastic Differentiation as a Fitness Trade-Off in Fluctuating Environments . 90

7 Outlook 91

Bibliography 93

Appendix 111

List of Figures 135

List of Tables 137

List of Schemes 139

List of Code Snippets 141

Erklärung 143

Introduction

"We are tied together into a single processing system, made up of many different individuals. You are me, and I am you. Together, we are an amazing, superior"

—Superorganism

1.1 Horizontal Gene Transfer

Life in microbial communities is diverse, with social interactions between strains and species. These social interactions include, cooperation, competition, toxins, quorum sensing, and horizontal gene transfer [1]. Untangling the influence of each interaction is key to understanding how these communities respond to external stresses.

Microbes spend the majority of their time in dense communities with various strains and species [2]–[4]. Phenotypes have evolved which affect other cells, and have both positive and negative effects on their recipients and actors [5]. Cooperation is a phenotype that increases the fitness of another cell. In some microbes, cooperation is mutualistic, such as the production of siderophores to aid in the uptake and metabolism of iron. The majority of bacteria excrete siderophores, which can sequester iron; the iron-siderophore complex can then be recognized by cell receptors on most bacteria [6], [7]. Other microbes have altruistic cooperation such as some cyanobacteria in their division of labor. A fraction of the population differentiates into heterocysts under nitrogen limitation, making them biochemically specialized for nitrogen fixation. This, in turn, causes the heterocysts to lose the ability to reproduce [8].

Bacteria in communities also compete. Competition can take the form of toxins, such as antimicrobial producer *Burkholderia thailandensis* [9]; others take the form of nutrient limitation, such as overproduction of extracellular polysaccharides in *Pseudomonas fluorescens* – which position the bacteria at the air-liquid interface where more oxygen is available [10].

In both cooperative and competitive interactions, bacteria frequently make use of quorum sensing. Quorum sensing is the secretion of signaling molecules to allow neighboring cells to measure population density [2], [11]. Examples of quorum sensing include sporulation and competence in *Bacillus subtilis* [12]–[14] and virulence in *Staphylococcus aureus* [15].

Finally, horizontal gene transfer (HGT) is the transfer of genetic material between individuals that are not linked by inheritance. It is essential to microbial evolution, as it enables genetic information to spread horizontally through diverse communities [16]–[18]. Evidence of HGT has been found not only between bacteria species [19] but also between bacteria and eukaryotes [20]. Here, we explain different mechanisms for HGT and highlight their significance in evolution.

1.1.1 The Role of Recombination in Evolution

In bacteria, there are three mechanisms of HGT: conjugation, transduction, and transformation (Figure 1.1) [21]. Conjugation is the transfer of DNA between two connected cells via nanotubes or membrane fusion [22]–[24]. Transduction is the exchange of genetic information via phages. Bacteriophages can package

bacterial DNA in two ways, specialized or generalized. In specialized transduction, bacterial DNA is intentionally integrated into the bacteriophage by imprecise splicing of its own phage DNA from the bacterial genome. Generalized transduction is the random incorporation of bacterial DNA during cell lysis [25]. Finally, transformation is the uptake of extracellular DNA [26], [27].

The general hypothesis is that HGT is a key agent for adaptation in bacteria [18]. It is thought to speed up adaptation by preventing clonal interference [28], the competition between two beneficial mutations in a population (Figure 1.2). In the absence of horizontal gene transfer, beneficial mutations have to occur on top of previously existing mutations in order to remain in the population. Beneficial mutations can even die out, if another mutation appears that has a greater fitness [29]–[33].

HGT can acquire novel alleles from other strains or species, such as resistance to antibiotics [16] or adaptation to new environmental niches [34]. In *Galdieria sulphuraria*, horizontally transferred genes are thought to make up 5% of the protein-coding genes. Those genes are also involved in important processes such as metabolism, detoxification, and glycerol uptake [20]. The transfer of genetic material

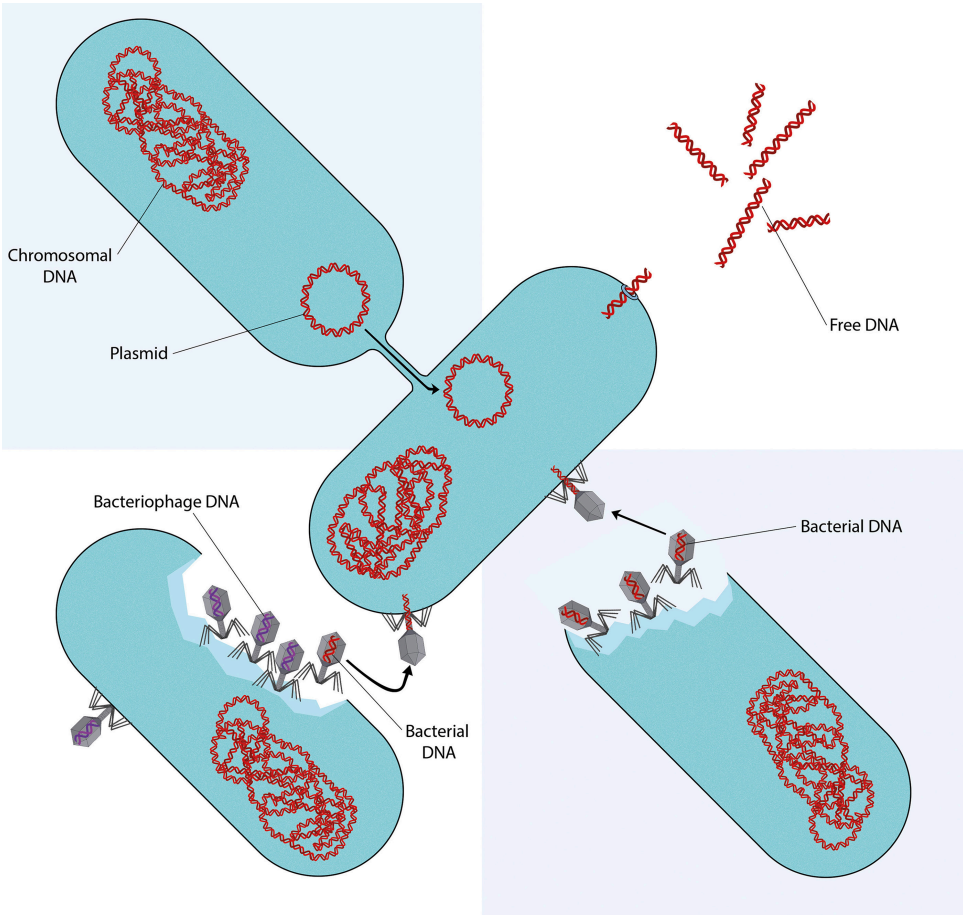


Figure 1.1: Horizontal gene transfer mechanisms in bacteria. (upper left) Conjugation – DNA is transferred through cell to cell contact. (upper right) Transformation – extracellular DNA is taken up from the environment. (lower panels) Transduction – bacterial DNA is loaded into and transferred by phages, both stochastically (lower left) and intentionally (lower right). Image adapted from [27] and under CC-BY license.

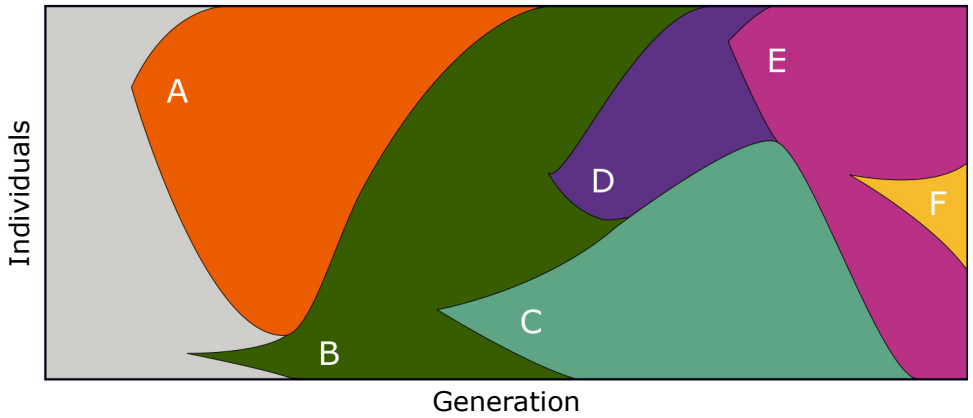


Figure 1.2: Müller diagram showing clonal interference. In a homogeneous population (vertical axis, initially gray) mutations in individuals occur as the population evolves (horizontal axis, generations). Mutations A – F are marked in different colors and grow in height as they outcompete concurrent genotypes. Clonal interference is seen as mutant B outcompetes mutant A, and later as mutant BC is out competed by mutant BDE. Mutant BC is initially fitter than mutant BD, until the latter accumulates the E mutation.

from prophages has also been shown to be beneficial, at times. Cryptic prophages in *Escherichia coli* contributed to antibiotic resistance, withstood osmotic, oxidative and acidic stresses, influenced biofilm formation, and increased growth [35]. Alternatively, it has recently been proposed that HGT expels DNA regions, in particular prophage elements [36], [37]. Theoretically, models have shown that HGT initially helps a population reach higher fitness in a rugged fitness landscape, However, at longer time scales HGT is less effective than the “trapping” of recombining populations on local fitness peaks [38].

In this study, we will focus on one type of horizontal gene transfer, transformation. In bacteria, many organisms can naturally transform by entering a state known as competence.

1.1.2 Bacterial Transformation and DNA Uptake Mechanisms

Competence is the ability to import extracellular DNA. Successful transformation through competence consists of the import of external DNA, and its integration into the host’s chromosome (Figure 1.3). In order to uptake external DNA, cells first have to become competent and build up the necessary machinery for transformation [39]–[41]. Often “competent for transformation" is spoken of, to articulate that cells are in a competent state but are not necessarily undergoing transformation. Competence is found naturally in many bacterial species, both Gram-positive species—e.g., *B. subtilis* and *S. pneumoniae*—and Gram-negative species—e.g., *Neisseria gonorrhoeae* and *Haemophilus influenzae* [26].

In Gram-positives, extracellular DNA binds to the cell surface. In *B. subtilis* ComEA is responsible for DNA binding [42], in *S. pneumoniae* additional proteins seem to be involved [43]. DNA is then transported into the cytoplasm linearly [44]. Uptake is increased in *B. subtilis* by surface endonuclease NucA which introduces double-strand breaks [45] or preceded by single-strand nicks and then double strand breaks in *S. pneumoniae* [46], [47].

In Gram-negatives, DNA uptake begins, for some species such as *Neisseria* [48] and *H. influenzae* [49], with a specific motif (DNA uptake sequence, DUS). These sequence motifs allow the bacteria to favor

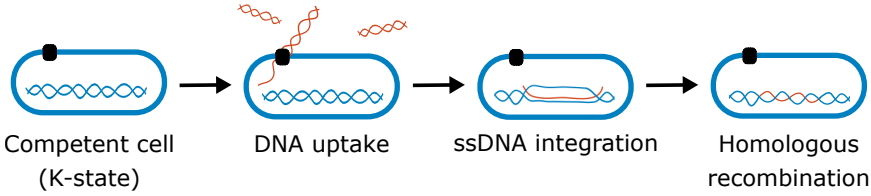


Figure 1.3: Cartoon schematic of competence for transformation. After entering the competent state (K-state), cells import extracellular DNA into the cytoplasm as single-stranded DNA (ssDNA). Once in the cell, single-stranded DNA can integrate into the host's chromosome.

homo-specific DNA uptake [26]. In *N. gonorrhoeae*, DUS binding depends on the major pilin (PilE), which is modified by expression of different minor pilins (e.g., ComP) [50], [51]. Recognition of the DUS triggers uptake across the outer membrane into the periplasmic space, driven by a translocation ratchet where ComE reversibly binds [52]. Finally, DNA is transported through the cytoplasmic membrane via ComA [53].

In both Gram-positives and negatives, only single stranded DNA (ssDNA) reaches the cytoplasm. The ssDNA is bound by single-stranded binding protein SsbB and DNA processing protein DprA [41]. The latter recruits recombinase RecA, which polymerizes on the ssDNA and initiates a homology search along chromosomal DNA. In successful transformation, strand exchange occurs and the foreign DNA is integrated into the host's genome [54]. The other DNA strain is degraded in the extracellular space (Gram-positives) or putatively the periplasmic space (Gram-negatives) [26].

1.1.3 Competence for Transformation in *Bacillus subtilis*

Focusing on the Gram-positive *B. subtilis*, competence genes are expressed in the late exponential phase of growth [55] and regulated by the competence transcription factor ComK (Figure 1.4) [56], [57]. During exponential growth, *comK* expression is inhibited by three repressor proteins: Rok, CodY, and AbrB. Rok is the most important of the three, which is reflected in its name "repressor of *comK*" [58]–[60]. Concurrently, basal levels of ComK are sequestered by MecA, and thereby degraded by the protease ClpCP [61]. During the late exponential growth phase, as the population becomes dense and nutrients scarce, quorum sensing is activated by ComX and Phr peptides [62], [63]. ComX is sensed in the extracellular space by the membrane-spanning protein ComP, which in turn phosphorylates ComA, leading to the regulation of *sfr* [64]. *sfr*, in turn, encodes for the small protein ComS, which competes with ComK for the binding site with the MecA/ClpCP complex. Freed ComK binds to its own promoter as a tetramer [65], creating an auto-catalytic feedback loop [66], [67].

Owing to the auto-catalytic feedback loop, competence is a bistable system in *B. subtilis* [66], [67]. A basal concentration of ComK is maintained just below the threshold needed to activate the auto-catalytic loop. Low copy numbers of *comK* mRNA lead to intrinsic noise in the ComK copy number [57], [68], [69] between cells. Stochastically, some cells will express enough ComK to get over the threshold for activation, and will become competent (enter the K-state). K-state cells escape from competence as ComS levels drop, leaving the MecA/ClpCP complex free to rapidly degrade ComK [70]. Figure 1.5.

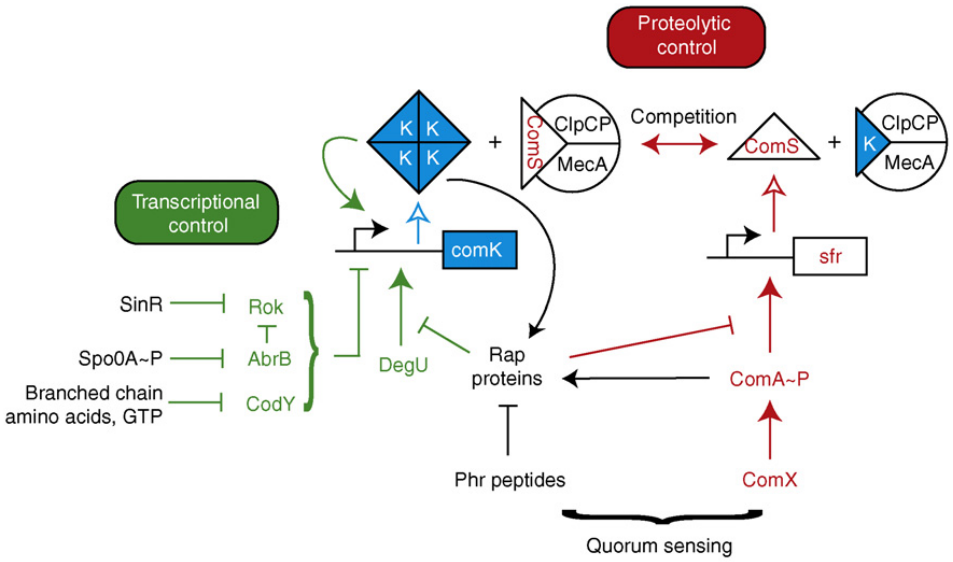


Figure 1.4: Core of the *B. subtilis* competence network. Transcriptional control (green) and proteolytic control (red) of *comK* (blue). Positive and negative regulation are denoted by arrows or t-bars, respectively. Bent arrows represent promoters. Image adapted from [57] and reproduced with permission.

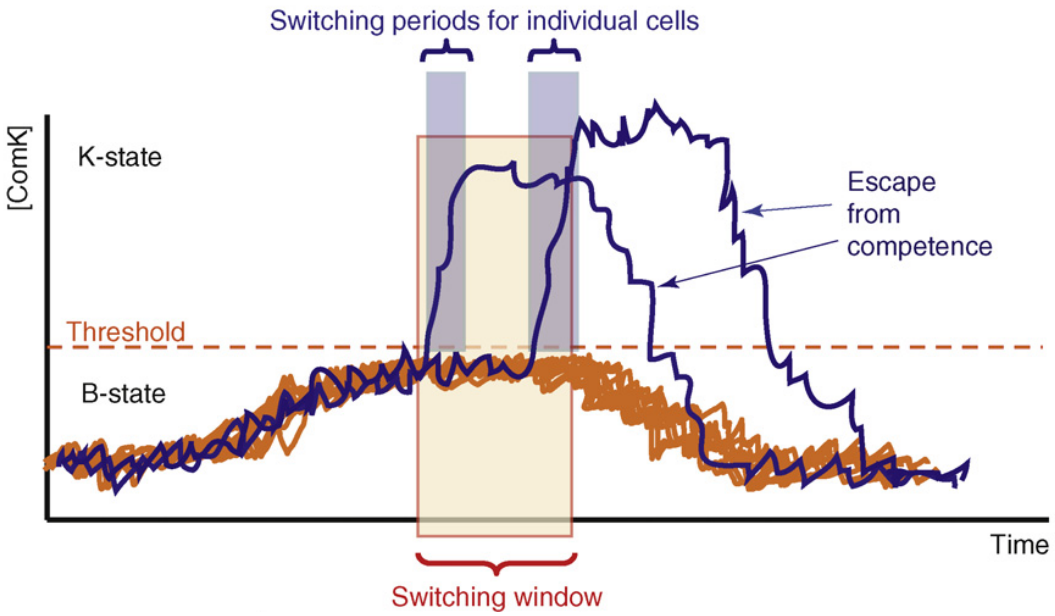


Figure 1.5: ComK levels in *B. subtilis* as a function of time. Cartoon drawing shows how basal ComK levels rise (orange curves), as the end of the exponential phase is reached. Within the switching window (beige, 1.5 h in length), ComK copy number fluctuations, in individual cells, can jump over the threshold needed to activate the auto-catalytic feedback loop (purple curves). Once over the threshold, the feedback loop ensures that those cells immediately enter the K-state. About 2 h later, K-state cells escape from competence, as the MecA/ClpCP complex degrades ComK. After the switching window, the basal level of ComK decreases and thereby the probability of cells (re)entering the K-state, too. Image adapted from [57] and reproduced with permission.

Bistability

Bistability is the phenomenon where cells with identical genotypes express different phenotypes at the same time. Non-uniform expression of genes is not all that uncommon in the bacteria world, and can be seen across many microbe species including *E. coli* (persister cells [71]), *Pseudomonas aeruginosa* (mucoidy and toxicity [72], [73]), and *B. subtilis* (competence [74]). These bistable systems can have increased fitness compared to homogeneous populations, and they allude to a division of labor [75] or bet-hedging strategies to maximize fitness [76].

The bistability of the K-state leads to two phenotypes in wild type strains: K-state and non K-state cells. In a lab wild type strain, 15% of the population will enter the K-state at the late exponential phase [77]. This number can be increased by knocking out *rok* (leading to nearly the entire population entering the K-state) or muted by deleting *comK*, and thereby the competence machinery. Studies playing with ComK and ComS, copy number and basal levels found that the competence system could be tweaked into an excitable, monostable, or bistable system, depending on those four aforementioned variables [78].

Cost and Benefits of the K-state

The cellular machinery for competence is robust, in that it has survived years of evolution. Nevertheless, a bistable system hints at costs and benefits of entering the K-state. K-state cells are competent for transformation – they are capable of taking up extracellular DNA. While there are several hypothesis as to what evolutionary function competence serves (including DNA as a food source [79]), the consensus is that competence evolved to facilitate horizontal gene transfer [80]–[82], and to a more minor degree, repair damaged DNA [83]–[85].

Competence bears the costs of generating machinery for DNA uptake, acquiring deleterious alleles, and reduced fitness in the presence of strong epistasis and persistence [38], [86], [87]. The most significant of these costs is growth arrest; K-state cells are growth arrested for about 2 h after re-inoculation [88], [89]. All of these costs come with the trade-off of acquiring novel functions [80], avoiding antibiotics [87], [90], and faster adaptation due to gene transfer [77].

1.1.4 Rates of Transformation Depend on Phylogenetic Distance

Up until this point, horizontal gene transfer, and in particular competence, has been examined without regard to how likely it is that foreign genes are integrated, at all. Work by Zawadzki *et al.* found that the transformation rate varied with sequence divergence between recipient *B. subtilis* strains and donor *B. mojavensis* [91]. Varying transformation rates, depending on recipient and donor, have also been measured in other species, such as *E. coli* where researchers found the amount of DNA uptake in *E. coli* K12 depended on the donor *E. coli* strain [92].

A more recent study broadened the scope of transformation efficiency between species, by looking at recombination rates of the *rpoB* gene from *B. subtilis* 168 into several *Bacillus* species. The *rpoB* gene is known to confer resistance to rifampicin (rif^R) with a single point mutation. In addition to the point mutation needed to confer resistance, the related *Bacillus* species had additional mismatches, varying from 74 – 624. The authors found that DNA uptake decreased exponentially as a function of sequence

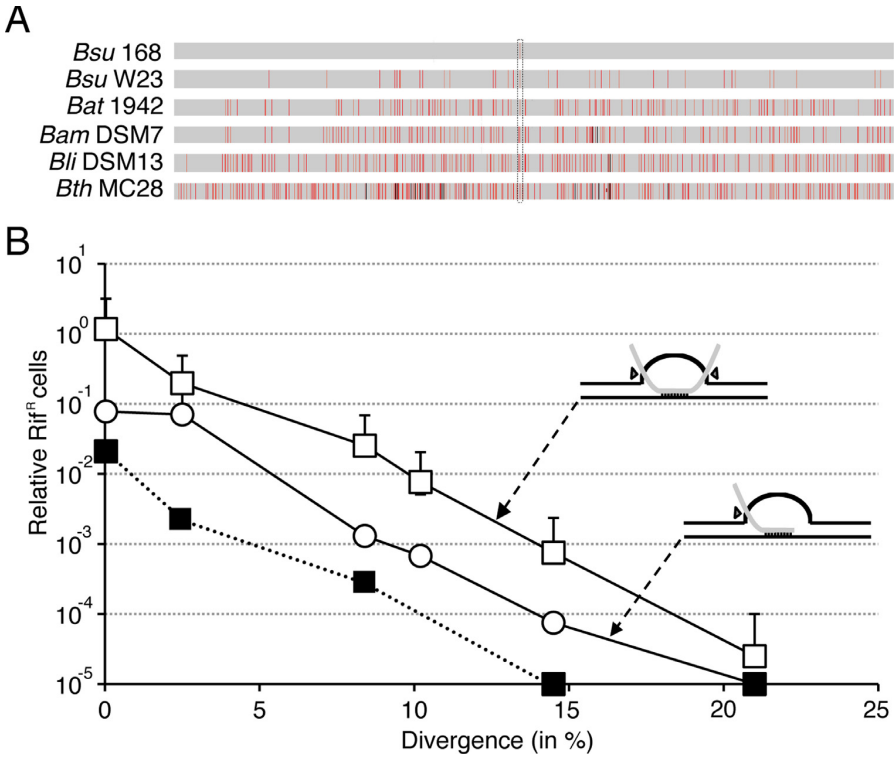


Figure 1.6: DNA uptake as a function of sequence divergence. (A) Similarity between the *rpoB*–rif^R gene from donor *Bacillus* species compared to the recipient (*Bsu*168) non-resistant *rpoB* gene. Red lines denote mismatches and black lines inserts/deletions. (B) Transformation efficiency (number of resistant cells, log scale), as a function of donor sequence divergence. For both recipient strains used (open and filled shapes), an exponential (log-linear) relationship can been seen. Squares and circles differ in how the donor DNA was prepared: (squares) random at both ends, (circles) cut at one end, random at the other. Image adapted from [93] and under CC-BY-NC license.

divergence (Figure 1.6). They also noted that RecA, a key protein in DNA recombination, did not have a preference to recombine on the forward or reverse strain [93].

In numerous species, transformation efficiency decreases exponentially with sequence divergence. Nevertheless, transformation has been shown to have positive effects, in addition to those negative, on fitness. We introduce some of those positive effects in the next section.

1.1.5 Fitness Effects of Recombination

Various studies have been carried out to measure the fitness effects of recombination. A study by Sorek *et al.* passed over the question of how effectively foreign genes are taken up, and instead, investigated the functionality of novel genes when spliced into the *E. coli* genome. They found a fraction of the novel genes, (~0.5% of 250,000) could not be cloned into *E. coli*. When those unclonable genes were put under control of an inducible promoter, growth inhibition was already visible at low induction levels. They concluded, horizontal gene transfer not only has to overcome barriers to transform the foreign DNA but also potential toxicity barriers due to increased gene expression [94].

Additional studies have integrated random segments from foreign species into a host genome and

found, yet, additional barriers to horizontal gene transfer. Knöppel *et al.* study with *Salmonella typhimurium* included integrated segments as large as 5 kbp in length and found that a small fraction of the cloned foreign genes had a negative fitness effect, in agreement with the findings by Sorek *et al.* More importantly, they found the majority of the cloned foreign genes (92% of 98) had no significant fitness effects. They concluded that gene transfer may not immediately confer a fitness benefit, as often assumed, but rather presents itself as a starting point for novel cellular functions [95]. Adding to the complexity, Tuller *et al.* found that codon usage of the donor and recipient strains was a contributing factor to the success of horizontal gene transfer. They found a positive correlation between the number of transferred genes and the similarity of their codon pools [96].

Focusing on competence and its role in evolution, Utnes *et al.* carried out a 175 day evolution experiment with transformation-proficient and deficient *Acinetobacter baylyi* strains. They found it was unclear if there was an overall fitness advantage for transformation. Transformation-proficient cultures showed a fitness advantage in the exponential and early stationary growth phases, but those cultures performed more poorly in the late stationary growth and death phases. The concluded, being competent for transformation does not yield a universal fitness advantage in an evolutionary setting. Transformation, alone, does not make up for the fitness costs of maintaining competence [97].

Baltrus *et al.* performed an experiment over a similar length of time using *Helicobacter pylori*, also looking at the putative benefit of competence in evolving populations. They found that competent populations had higher selection rates than noncompetent populations, when compared to their ancestors. They also found fitness variance was greater in noncompetent populations than competent populations, at early time points. They concluded that greater variance and smaller relative selection rates of the noncompetent population are best explained by clonal interference [98].

These experiments show the wide range of fitness effects recombination can have on transforming populations. Recombination can lead to fitness increases, but often requires recipient strains to adapt their genomes. Other novel genes show no fitness effects or are toxic in the recipient cell. Research remains to be done to form a complete picutre of effects HGT.

Horizontal gene transfer is crucial to microbial evolution. One mechanism of HGT, transformation, is found naturally in many bacterial species, and is hypothesized to aid in the acquisition of beneficial alleles and repair of genomic DNA. In the particular case of *B. subtilis*, competent for transformation, the K-state, is a bistable state, where only ~15% of a wild type population enters the K-state in the stationary phase. The bistability of the K-state alludes that a division of labor or bet-hedging strategy is at play, particularly as K-state cells are growth arrested. Transformation has been seen to have positive, neutral, and negative effects, depending on the similarity of recipient and donor species. Additional experiments have shown that the probability of recombination decreases exponentially with sequence divergence. In the next section of this chapter, we explore all of these aspects of horizontal gene transfer, in the important framework of experimental evolution.

1.2 Experimental Evolution

Experimental evolution is the study of evolutionary dynamics and processes, through controlled laboratory experiments, commonly making use of and manipulating organisms with short generation times and small physical size—i.e., microbes. It allows researchers to track evolutionary processes in real time, addressing questions such as: how species adapt to an environment, what trade-offs are inherent in adaptation – and in particular transformation, how mutation rates and load affect populations, and how to test an evolutionary theory. [99]

In general, there are three types of evolution experiments: single-cell bottlenecks, continuous culture, and serial transfer (Figure 1.7). Single-cell bottlenecks allow for the accumulation of mutations [100]. Continuous culture [101]–[103] and serial transfer [104] allow cultures to adapt to experimental conditions. Experiments do not have to be explicitly one type or the other, and can be mixed and matched according to the experimental question at hand. In addition to a wide range of freedom concerning experimental design, the organism involved in a study can practically be chosen as a matter of convenience; many evolutionary questions apply to a broad range of organisms. This design freedom has resulted in numerous studies being performed on multicellular organisms such as *Daphnia* [105] and *Drosophila* [106], [107] along with countless microbes such as *B. subtilis* [108], *E. coli* [104], [109], and *Saccharomyces cerevisiae* [110]. Microbes are favored organisms for evolution experiments as large populations can be maintained with ease, and samples can be frozen and later reanimated. The length of evolution experiments vary, from one day (for recombination experiments [111]–[113]) to decades (for a serial transfer experiments [106], [107], some of which are still ongoing [114]), depending upon the questions being addressed [115]–[118].

Regardless of the exact question being posed, experimental evolution, combined with fitness characterization and whole genome sequencing (WGS), allows one to better understand the mutational pathways underlying adaptation.

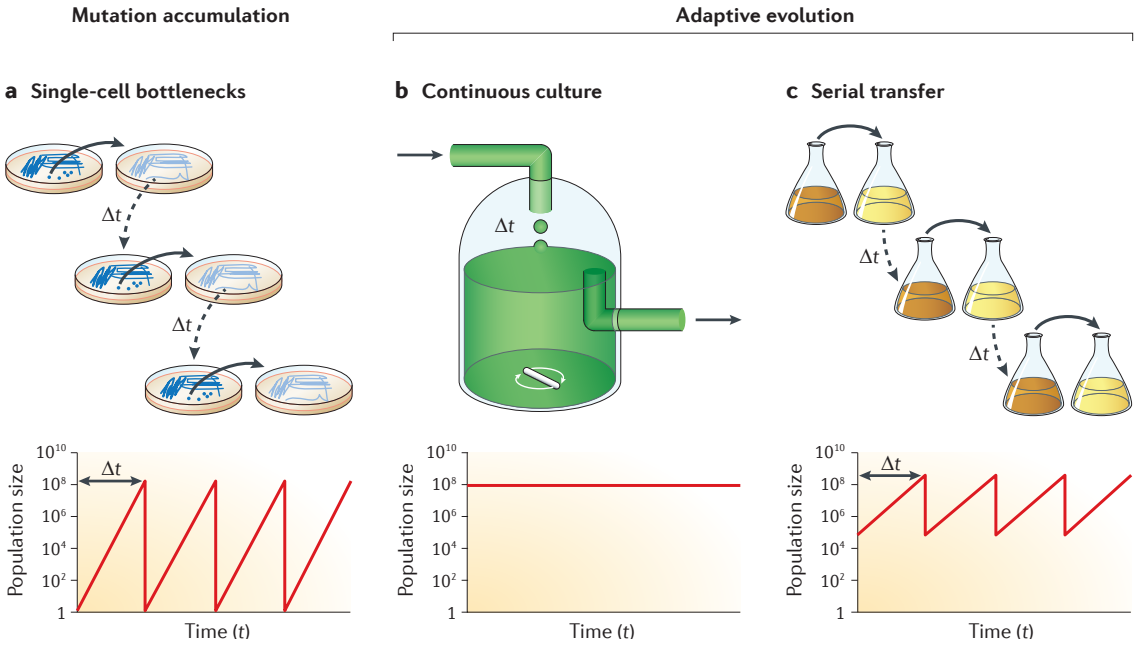


Figure 1.7: Types of evolution experiments. There are three main types of evolution experiments: (a) single-cell bottlenecks, (b) continuous culture, and (c) serial transfer. The top panel shows a graphic representation of each experiment and the bottom panel shows how the population size varies in time, in that particular experimental type. Image adapted from [116] and reproduced with permission.

1.2.1 Bacterial Fitness

The experimental warhorse of quantitative measurements and a corner stone of evolutionary theory is fitness. In evolution, fitness is the ability of organisms to survive and reproduce in their environment [119]–[122]. The following mathematical descriptions for fitness are derived from Nowak and Lässig [123], [124].

Deterministic Fitness

Mathematically, fitness is most easily described beginning with reproduction and large populations. A bacterium growing in a nutrient rich environment will divide to produce two daughter cells. Those daughter cells will, in turn, divide and the population will grow exponentially. For a constant division rate, R, one can write the differential equation

$$\frac{dN}{dt} = RN \tag{1.1}$$

and its solution

$$N(t) = N_0 e^{Rt}, \tag{1.2}$$

where N is the number of cells in the population, N_0 is the initial number of cells, and t is time. The obvious flaw with Equation 1.2 is a bacterial population would grow unstopped and the population size

approaches infinity. One key piece that is missing, is an outcome that catches us all, death[1].

Most simplistically, death can be described in the same way growth was; we assume cells die at a constant rate over the entire experiment. This leads to the following differential equation

$$\frac{dN}{dt} = (R - d)N \tag{1.3}$$

with solution

$$N(t) = N_0 e^{(R-d)t}, \tag{1.4}$$

where d is the death rate. Often, $R - d$ is redefined as r, an *effective* growth rate, also referred to as a Malthusian parameter [125]. Positive Malthusian parameters reflect population growth, and negative values population death.

The other key piece that is missing before we have a functioning description of bacterial growth is a carrying capacity. As resources, or in this case nutrients, are not endless, a maximum carrying capacity, K, must be introduced with the logistic equation

$$\frac{dN}{dt} = rN(1 - N/K) \tag{1.5}$$

and solution

$$N(t) = \frac{KN_0 e^{rt}}{K + N_0(e^{rt} - 1)}. \tag{1.6}$$

When N is much smaller than K, growth is exponential. As N increases, growth slows until $N = K$ where growth stops.

Populations are rarely homogeneous, and often one wants to characterize subpopulations competing for the same space and nutrients. Starting with a population with only two subpopulations, A and B, one can write the following set of differential equations

$$\begin{aligned} \frac{dx}{dt} &= x(a - \phi) \\ \frac{dy}{dt} &= y(b - \phi), \end{aligned} \tag{1.7}$$

where a and b are the growth rates of A and B, x and y are the relative abundance of A and B at time t, and $\phi = ax + by$. Because we force $x + y = 1$, ϕ is the average fitness of the population. Those equations can be combined to give

$$x = \lambda e^{st}/(1 + \lambda e^{st}), \tag{1.8}$$

where $\lambda = x(0)/y(0)$, $s = a - b$, and s is assumed to be constant. This equation is the two subpopulation solution to the replicator equation and s is a selection coefficient (see Equation 1.17 for a formal definition). It has two equilibrium points, $x = 0$ or $x = 1$, reflective of the deterministic nature of two subpopulations with different growth rates. More generally, the replicator equation for many subpopulations takes the

[1] "It is the secret of the world that all things subsist and do not die, but retire a little from sight and afterwards return again." – Ralph Waldo Emerson

form

$$\frac{dx_i}{dt} = x_i\,(r_i - \bar{r}) \tag{1.9}$$

where $\bar{r} = \sum_{j=1}^{n} x_j r_j$.

Stochastic Fitness

More realistically, one can take into account that growth rates fluctuate stochastically over time. These fluctuations are called genetic drift. By adding a noise term, σ, to Equation 1.3, we arrive at

$$\frac{dN}{dt} = rN(t) + \sigma(t). \tag{1.10}$$

Due to the law of large numbers, the noise term is a Gaussian random variable with mean equal to zero [124]. For the simple two subpopulation model, with noise the population fraction of A becomes the Langevin equation

$$\frac{dx}{dt} = \Delta r_{ab} x(1-x) + \sigma_x(t), \tag{1.11}$$

where $\Delta r_{ab} = r_a - r_b$ and

$$\langle \sigma_x(t)\sigma_x(t')\rangle = \frac{x(1-x)}{N}\delta(t-t'). \tag{1.12}$$

Finally, Equation 1.11 is converted into a Fokker-Planck equation to capture the statistics of the evolutionary trajectories [124]. The probability distribution of the genotype composition is

$$\frac{dP(x,t)}{dt} = \frac{1}{2N}\frac{d^2}{dx^2}x(1-x)P(x,t) - \Delta R_{ab}\frac{d}{dx}x(1-x)P(x,t). \tag{1.13}$$

Here, again, $x = 0$ and $x = 1$ are equilibrium points, fixed points of the stochastic dynamics.

Measuring Fitness

When measuring fitness, one generally speaks of two types of fitness: absolute fitness and relative fitness. Absolute fitness, W, is the proportional change in abundance of one genotype over one generation:

$$n(t+1) = Wn(t), \tag{1.14}$$

where $n(t)$ is the abundance of genotype n. Values larger than one correspond to a growth in abundance, whereas negative values indicate a decline [126]. Relative fitness, w, measures changes in genotype *frequency* and is defined as

$$p(t+1) = \frac{w}{\overline{w}}p(t), \tag{1.15}$$

where $p(t) = n(t)/N(t)$ (the frequency of the genotype n in the population), and $\overline{w}$ is the mean relative fitness. Because relative fitness only indicates a change in prevalence of a genotype, only relative values are important. For convenience, the relative fitness of the wild type or reference genome is set to one

[126]. Absolute fitness can be measured from relative fitness as

$$p(t+1) = \frac{n(t+1)}{N(t+1)} = \frac{W}{\overline{W}}p(t), \tag{1.16}$$

where $\overline{W}$ is the mean absolute fitness. This, in turn, implies that relative and absolute fitness are proportional to each other by $w/\overline{w} = W/\overline{W}$.

There are many methods to measure fitness experimentally. The most readily accessible method to measure fitness is *growth rates using optical density*. With this method, the optical density (OD) of a population is monitored throughout the exponential growth phase [104], [127], [128]. Because OD is proportional to number of cells, a linear equation can be fit to the data, in logarithmic space, where the slope is the Malthusian parameter—i.e., growth rate [125]. Because Malthusian parameters are used, experiments need to begin and end within the exponential growth phase, where nutrients are not limited. This measurement of fitness is simple and fast, but it neglects other components of fitness even in simple systems [129].

If two subpopulations are competing in the same culture, fitness can be measured using a *competition assay*. This allows one to focus on relative fitness, which is more important when considering the evolutionary fate of a subpopulation [130]. The individual growth rates for the two subpopulations are measured, usually by placing a fluorescent marker in the reference strain. Growth rates can then be monitored using OD and fluorescence, fluorescence microscopy, or flow cytometry [129], [131], [132]. A selection coefficient, s, is formally defined as

$$s = \frac{d}{dt}ln\left(\frac{p}{1-p}\right) = \frac{dlnN}{dt} - \frac{dlnN}{dt}, \tag{1.17}$$

where p and $(1-p)$ are the frequencies of the two genotypes M, mutant, and N, wild type, in a population. Assuming there are no interactions between the genotypes

$$s = r_m - r_n, \tag{1.18}$$

where r is the respective Malthusian parameters for M and N [123], [133]. Here, again, if Malthusian parameters are used, experiments need to occur within the exponential growth phase. For competitions assays in the stationary growth phase, where one assumes the difference between the effective growth rates is constant, the replicator equation can be used (Equation 1.9).

Depending on the phenotype being examined, *single-cell microscopy* is often the method of choice to measure fitness. Population size can be monitored over time, to determine growth rates, or subpopulations (one marked with a fluorescence marker) can be monitored to determine growth rates and selections coefficients [134], [135]. For bistable phenotypes, lineages can be tracked to determine *cell doubling times* [136]–[138]. In more complex microfluidic chambers, *cell length as a function of time* can be measured [139]–[141]. One such famous microfluidic chamber is the mother machine [142], where multiple long channels, with the width of one bacterium, are filled with single cells. The seed, or "mother" cell continuously divides, pushing daughter cells up and eventually out of the channel.

Up until now we have discussed fitness from the perspective of how organisms reproduce. From an

evolutionary perspective, this is not the only factor determines an organism's ability to survive. A map, linking genotype to fitness, provides additional details on the course of evolution and is explained in the next section.

1.2.2 Fitness Landscapes and Epistasis

Fitness landscapes—i.e., a genotype-fitness map—have crucial effects on the course of evolution including sex, divergence and speciation, genetic robustness, and evolvability [143]–[145]. The most well-known image of a fitness landscape, also known as an adaptive landscape, was introduced by Wright [146], [147]. His fitness landscape was a three-dimensional mountainous landscape, with genotypes forming the floor of the graph and fitness on the vertical axis (Figure 1.8(left)). The landscapes lead to the interpretation of increasing mean fitness as a hill-climbing process. As landscapes often have local maxima, in addition to the global maximum, populations can "get stuck" on a local maximum, because further evolution towards the global maximum would require first going down in fitness.

Wright's fitness landscapes quickly becomes complex and high-dimensional when more than two alleles, forming the floor of the landscape, are taken into account [147]. The concept of fitness landscapes underwent further development as protein mutational pathways were used instead of genotype space [148]. In protein space, one considers a relevant subset of the protein's amino acid or nucleotide sequence, say four amino acids, and calculates the fitness of each permutation of the four amino acids, wild type or mutant. Lines can then be drawn showing the mutations that lead to higher fitness, and the quickest route to the highest fitness (Figure 1.8(right)).

Both Wrightian and empirical landscapes make it evident that fitness it not a linear process from lowest to highest fitness but a complex path. This dependence of mutational effects on genetic background, due to genetic interaction, is known as epistasis [145].

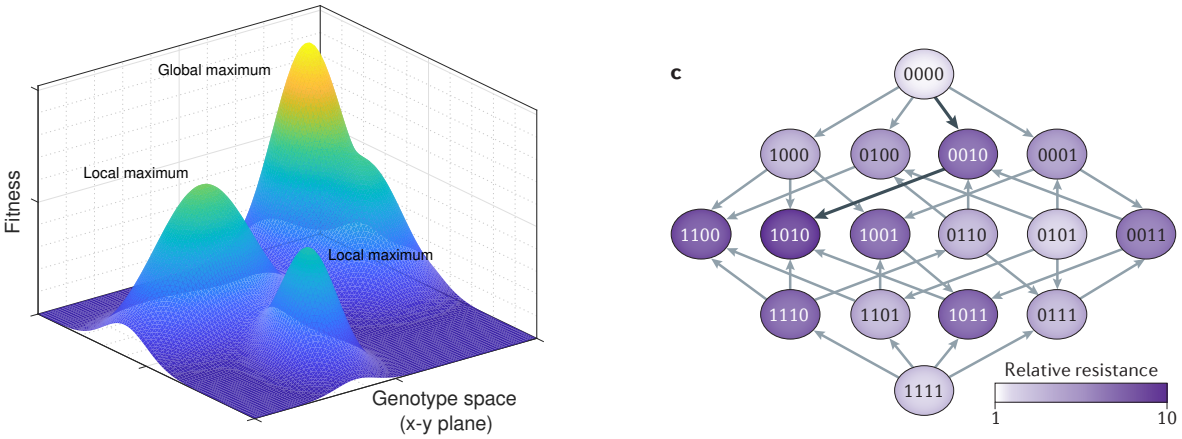


Figure 1.8: Fitness landscapes. (left) Wrightian fitness landscape. The x-y plane is genotype space and the vertical axis fitness. Local maxima "trap" evolving populations, keeping them from reaching the global fitness maximum. (right) Empirical fitness landscape. Each node represents one of 16 genotypes; 0 and 1 for wild type and mutant amino acids, respectively. Fitness (relative resistance conferred by the mutation(s)) are color coded. Gray arrows point towards higher resistance between single mutations. Black arrows show the quickest route to the global maximum. Image (right) adapted from [145] and reproduced with permission.

Epistasis

Epistasis is deviations from independent effects of alleles (genes) on a phenotype. In layman's terms, epistasis is the interaction between genes leading to positive (or negative) interaction at levels higher than expected if the genes were accounted for separately [143], [149]–[151]. Epistasis has been seen in numerous experiments including those on *E. coli* TEM-1 β-lactamase [152], *E. coli* and *S. cerevisiae* metabolic networks [153], and resistance in malaria parasites [154]. Experiments measuring epistasis require exploration and quantification of entire genotypic spaces. This can be quite painstaking, depending on the scope of experiment, as theories have not yet been able to completely or universally predict all epistatic interactions.

The two most generic classes of epistasis are sign and magnitude. Sign epistasis is a mutation that is beneficial on one genotypic background but deleterious on another (Figure 1.9(a,b)). Magnitude epistasis is a mutation that is unconditionally beneficial or deleterious, only the magnitude of that effect is dependent on genotypic background (Figure 1.9(c)) [155]. One can imagine, as genotype space is allowed to become more complex (more alleles), more complex epistasis patterns can be recognized. Of those, the most common in evolutionary studies are diminishing-returns epistasis and all-or-none epistasis. Diminishing-returns epistasis: the combined effect of mutations on fitness is less than expected for the mutations individually. This is often a characteristic of long-term evolution experiments (Figure 1.10) [156]. All-or-none epistasis: as the name implies, a complete set of mutations has to be present to confer a fitness advantage. Any fraction thereof, yields no fitness advantage [157].

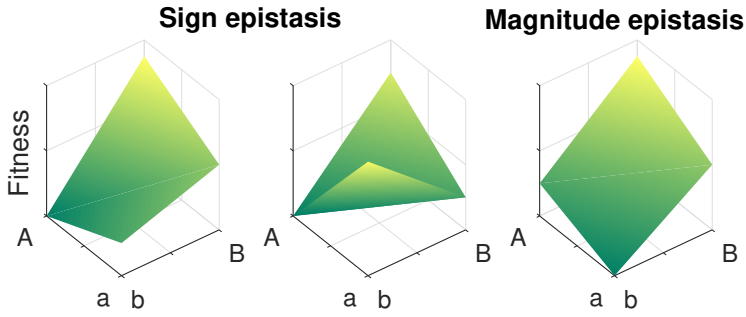


Figure 1.9: Sign and magnitude epistasis. The bottom x-y plane shows genotype space for two alleles (genes) a/A and b/B. Each gene can either be in a native state (a/b) or a mutated state (A/B). In the vertical axis the fitness of each mutation is shown. (A) Sign epistasis: The mutation $a \rightarrow A$ has a negative fitness effect in the "b background" (when gene b/B is not mutated) but positive effect in the B background. (B) Sign epistasis with multiple peaks: The mutation $a \rightarrow A$ still has a negative fitness effect in the b background and positive in the B background, but ab is fitter than either partial mutant (aB or Ab). (C) Magnitude epistasis: The $a \rightarrow A$ mutation is favorable in all backgrounds.

1.2.3 Characterizing Fitness Experimentally

Advances in WGS technology have made the prospect of correlating WGS to fitness enticing. As most evolutionary studies begin with clonal populations, one can link changes in phenotype and response to external stresses directly to genomic mutations. A homogeneous ancestral strain (starting point) also

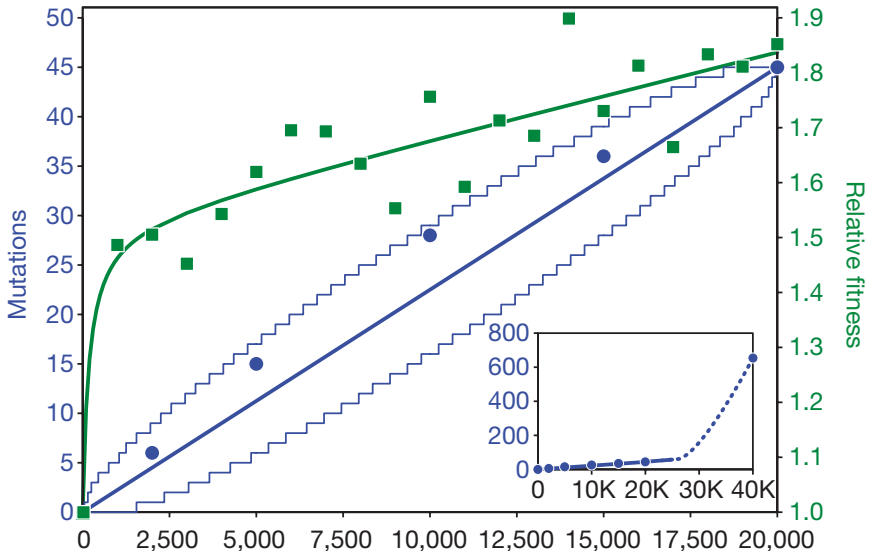


Figure 1.10: Fitness and mutation rates from the Lenski *E. coli* evolution experiment through 40,000 generations. Left axis, blue: (dots) average number of mutations at given time points, (lines) modeled mutation rate assuming it is constant, (steps) 95% confidence interval. Right axis, green: (dots) relative fitness, (line) hyperbolic-linear fit to the data. Inset shows the appearance of a mutator strain, as the mutation rate explodes after ~25,000 generations. Image adapted from [156] and reproduced with permission.

allows for the calculation of mutation rates, genetic diversity, and DNA uptake rates, all of which can be tied back to fitness [115], [116].

One of the longest running, long-term evolution experiments is the *E. coli* adaptation from the Lenski lab. Since 1988, their lab has evolved twelve clonal *E. coli* populations for more than 50,000 generations [158]. Early findings, by generation 2000, already showed that while mean fitness had increased, it increased at a slower rate in later generations [104]. By 10,000 generations, fitness had seemingly plateaued and the twelve replicate populations had diverged from one another, in both fitness and morphology. The initial conclusions from the experiment were based namely on fitness measurements, but decades later, frozen cultures were revived and sequenced. This sequencing brought new insights to "old" data in an ongoing experiment. With WGS, they were able to see that although the rate of increase of relative fitness had dropped, the mutation rate was constant over the 20,000 generations (Figure 1.10). This confirmed that diminishing-returns epistasis was at hand [116].

At 50,000 generations, core genes (defined as single copy orthologous genes shared across 60 *E. coli* strains) were found to have accumulated more non-synonymous mutations than their non-core counterparts [159]. Non-synonymous mutations occurred three times more frequently than synonymous mutations and seventeen times more frequently in the first 500 generations; similar findings were found for point mutations in intergenic regions. At the genomic level, the mutation rate had remained constant (excluding strains that became mutators), and yet still, after 50,000 generations, a high frequency of beneficial non-synonymous mutations was detected. They concluded, even after being in a constant environment for 50,000 generations, most non-synonymous mutations that reached high frequency were beneficial [160].

Looking at the protein functionality of genes having undergone recombination, Bershtein *et al.* replaced

the chromosomal *folA* gene with 35 interspecies orthologs. They found that all replacements caused a reduction in growth rate, despite the orthologous proteins being as stable as the native *E. coli* protein. Orthologous strains were evolved for ~600 generations and sent for sequencing. WGS revealed the accumulation of mutations the Lon protease which caused an increase in abundance of the orthologously replaced protein, and the evolved strains with the highest fitness had accumulated these Lon protease mutations. Their study showed that horizontal gene transfer does necessarily confer fitness benefits immediately. Their work builds on the findings of Sorek *et al.*, who concluded that foreign genes can be toxic in a recipient cell and even recombination is non-toxic, non-self proteins are less functional than their native equivalent [94], [161].

Fitness can be characterized in evolution experiments, even over long time scales. Experiments have been able to correlate WGS results with fitness. Next we will examine how evolution experiments take both fitness and WGS to better understand HGT.

1.2.4 Horizontal Gene Transfer in Experimental Evolution

Having seen the importance of HGT and the ability to correlate WGS with fitness, experimental evolution became an essential method to study those attributes.

One of the first experiments to analyze competence in an evolution experiment was by Engelmoer *et al.* By transforming wild type and competence deficient *S. pneumoniae* strains for 1000 generations, they found that non-competent strains had higher fitness in benign conditions, but the competent populations had fewer mutations and inhibited the emergence of mutators. Under periodic stress, the relative fitness of the non-competent strains was reduced. They concluded that competence was costly, particularly in benign conditions, but helps maintain genome stability [162].

A study using naturally competent *H. pylori* fed cultures DNA from *H. pylori* derivative strains for 28 – 52 days. Donor DNA contained a resistance cassette for chloramphenicol in the *rdxA* locus. Authors found that genomic elements other than the resistance cassette at the *rdxA* locus were imported. The additional genomic elements, secondary elements, had a similar length distribution to colonies evolving on nonselective plates, ranging from 1 - 13,400 bp. Additionally, without selection, 3.3% of the genome was replaced consisting of ~40 individual segments, on average. They drew the conclusion that restriction-modification systems inhibit novel sequences from being integrated [163].

Experiments in *S. pneumoniae* looked specifically into what secondary elements of a donor DNA molecule were imported, independent of a selected antibiotic resistance. They found the amount of secondary elements depended on the concentration of DNA supplied and was independent of the mismatch repair system. A mosaic pattern of the secondary elements, imported independently from the selected antibiotic resistance, was also seen (Figure 1.11). Researchers concluded that unselected recombination events have a fixed per base probability [111].

Transformation of a lab strain of *H. influenzae* with donor DNA from a clinical isolate, without selection, found a similar mosaic recombination pattern. One to three percent of the host chromosome had been replaced, including novel genes and deletions from the donor. There, only 3 – 6 recombined segments were identified, 8.1 kbp in length [164].

Finally, evolution experiments using *E. coli* strains with different carbon sources concluded that the

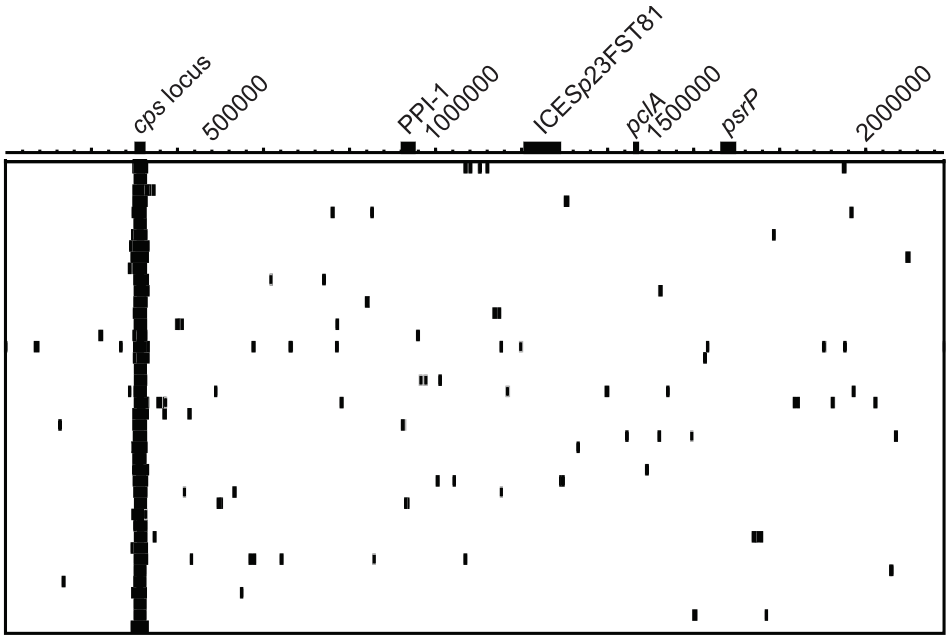


Figure 1.11: Mosaic pattern of recombination events in *S. pneumoniae* 23F-R from *S. pneumoniae* TIGR4Dcps. Transformation occurs in the presence of kanamycin, as the donor DNA had a *cps* allele carrying resistance. Recombination was seen across the genome, outside of the *cps* area of selection. The genome is displayed across the top, denoting key genes in *S. pneumoniae*, and individual replicates are show on the vertical axis. Recombination events are represented as black blocks with lengths proportional to recombination length. Image adapted from [111] and under CC-BY license.

carbon source in question played the central role in how effective recombination was. Recombination proficient strains grown on one carbon source (4-hydroxyphenylacetic acid, HPA), where a gene in the donor was identified to help mediate the metabolism process in the recipient, showed a benefit for recombination. Contrarily, recombination proficient strains grown on a second carbon source (butyric acid), where no metabolism mediating gene was found, showed no benefit from recombination. This was despite the fact that recombining populations went extinct less frequently then a non-recombining counterparts [92].

Evolution experiments have made significant strides in understanding the role of recombination in evolution. Experiments have compatible recombined-DNA-length distributions and consistently seen non-selected (secondary) import recombination events, regardless of homology. Nevertheless, experiments are inconclusive regarding fitness effects and linking them to WGS. Also, in many experiments, temporal resolution is missing.

1.3 Aims of This Study

Evolution experiments have been successful at chipping away at open questions in evolutionary biology. In combination with whole genome sequencing (WGS), they have become a powerful method to understand genotype changes, mutation rates, and adaptation. Significant work has been done examining competence in an evolutionary setting. Previous studies have shown that, under benign and selective conditions, DNA uptake from related subspecies occurs randomly, producing mosaic uptake patterns. The length of imported segments follows a log-linear relationship, where smaller sized segments were favored. By examining the efficiency of uptake, researchers found that efficiency decreased exponentially with increasing divergence.

When the uptake machinery was bypassed and foreign genes were cloned into a host genome, the recipient showed one of two effects: (1) No change in fitness if the novel genes were completely foreign, although some novel genes were toxic to their host. (2) The proteins from the replaced gene were less functional if the genes were homologs, and replaced a native gene. This drop in functionally caused the recipient clones to have a lower fitness. Reduced fitness could be reversed when the orthologous clones were evolved for many generations. WGS revealed that clones had become fitter by increasing the copy number of the less functional proteins, or in some cases, increasing the activity of gene's promoter.

Notwithstanding these contributing results, the role competence for transformation plays in horizontal gene transfer, remains unanswered. Of particular interest is the effect competence has on genome dynamics and interactions between subspecies in mixed communities.

In the first part of this study, we designed and executed evolution experiments with *B. subtilis* 168 (Bsu168), periodically supplying evolving strains with DNA from the related subspecies *B. subtilis* W23 (BsuW23). We characterized the genome dynamics over 21 cycles and examined if gene transfer was limited mechanistically (due to recombination probabilities), or by selection and epistasis. We looked for evidence of gene transfer saturation, and examined any patterns in de novo variants, such as compensatory mutations. Our experimental design uses two novel approaches: parallel evolution and time-resolved WGS.

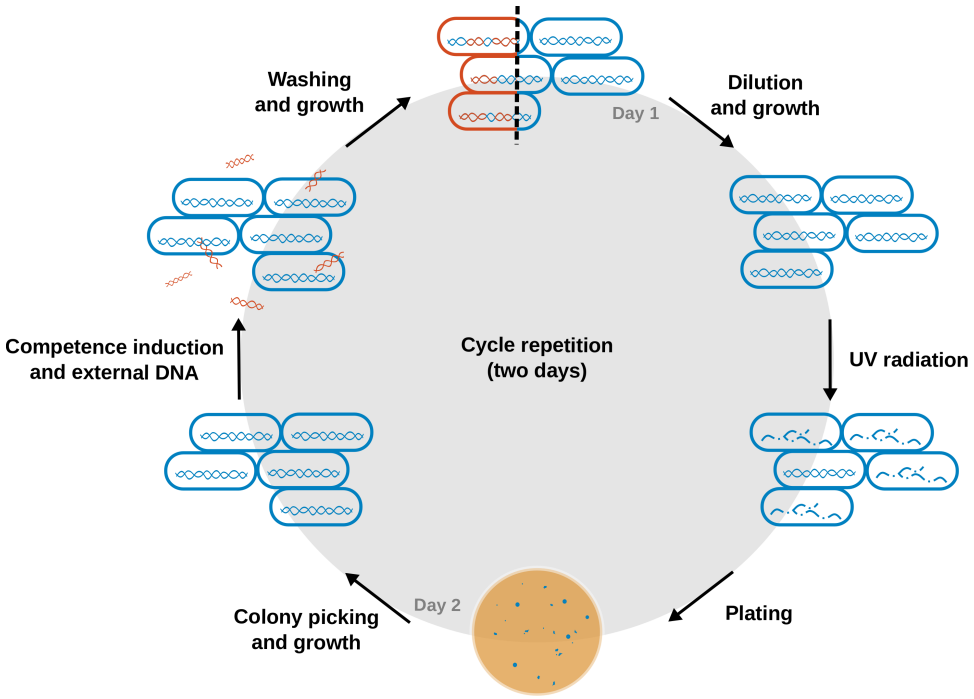
In the second part of this study, we measured the cost of growth arrest during competence to determine its relevance in the stationary phase. By tracking the population dynamics of *B. subtilis* strains (with various probabilities of entering the competent state) in head-to-head competition assays, we were able to quantify the cost of the competence machinery in a benign environment. Using microfluidic chambers, we measured generation times of K-state and non K-state cells in the stationary growth phase. We used the novel approach of creating cell lineages to track competence, growth, and stationary phase dynamics at the single-cell level.

Methods – Evolution Experiment

> *"Working in the lab late one night, his eyes beheld an eerie sight. His monster began to rise from his slab, and suddenly to his surprise – it did the* mash*."*
>
> —Bobby Pickett

In order to characterize genome dynamics in the presence of different subspecies, we designed an experiment and analysis method with parallel evolution and time-resolved WGS. Cells were evolved over 21 two-day cycles. A cycle consisted of six steps: dilution, radiation, plating, colony selection and regrowth, competence induction and addition of extracellular DNA, washing and overnight growth (Scheme 2.1). Samples were frozen every second cycle. All replicates were sequenced at three time points and four replicates were sequenced over all time points. Here we detail our experimental and computational methods.

Scheme 2.1: Evolution experiment design



2.1 Experimental Methods

2.1.1 Strains, Media, and Growth

The bacterial strains used in the evolution experiment are derivatives of BD630 (*B. subtilis* subsp. *subtilis* str. 168, referred to in this text as Bsu168 or recipient) and the related subspecies *B. subtilis* subsp. *spizizenii* str. W23 (referred to as BsuW23 or donor). A description of each strain, including aliases used in this text, is listed in Table 2.1.

Bs166 (Δ*comK*, *comK*-IPTG) was generated by transforming BD3836 (*comK*-IPTG) [66] with genomic DNA from Bs075 (Δ*comK*) [165]. This strain was chosen as the ancestral/recipient strain for the evolution experiment for three reasons: (1) Deleting *comK* removed the problematic of the *comK*-operon acquiring deleterious mutations in the experiment and thereby an immediate fitness advantage – competent cells are growth arrested. (2) Inducible *comK* forced all cells to become competent (not only the ~15% in the wild type) reducing the time the evolution experiment needed to run to obtain results (3) Inducible *comK* decreased the day to day length of the experiment; cells could be induced in the exponential phase. This construct was transformed and characterized by M. Yüksel and G. Schneider.

B. subtilis W23 was chosen for the evolution experiment due to its phylogenetic proximity to *B. subtilis* 168. The two strains shared 3.6 Mbp of their genomic content with an average identity of 92.4 %, making orthologus recombination events not only possible, but also visible via sequencing. In addition to the common genome, both subspecies had auxiliary genomes, 567 kbp in Bsu168 and 386 kbp in BsuW23. The auxiliary content allowed for the detection of de novo insertions. A comparison of the two genomes is summarized in Table 2.2 and Figure 2.1 [166].

Table 2.1: Bacterial strains used in the evolution experiment.

Strain	Alias	Relevant genotype	Source/reference
BD630	Bsu168	*hist leu met*	–
BD3836	–	*hist leu met*, *amyE::*P_{hs}*comK (spc*[a]*)*	[66]
Bs075	–	*hist leu met, comK::kan*[a]	This study, [165]
Bs166	–	*hist leu met*, *amyE::*P_{hs}*comK (spc*[a]*)*, *comK::kan*[a]	This study
2A9	BsuW23	type strain	–

[a] *spc*, *kan* stand for resistance to spectinomycin and kanamycin, respectively

Table 2.2: Key *B. subtilis* 168 and *B. subtilis* W23 statistics

	B. subtilis 168	*B. subtilis* W23
Genome length (kbp)	4.2 Mbp	4.0 Mbp
No. of genes	4421	4116
No. of auxiliary genes	141	157
Length of auxiliary genome	567 kbp	386 kbp
Interspecies identity (ex. auxiliary regions)	92.4%	

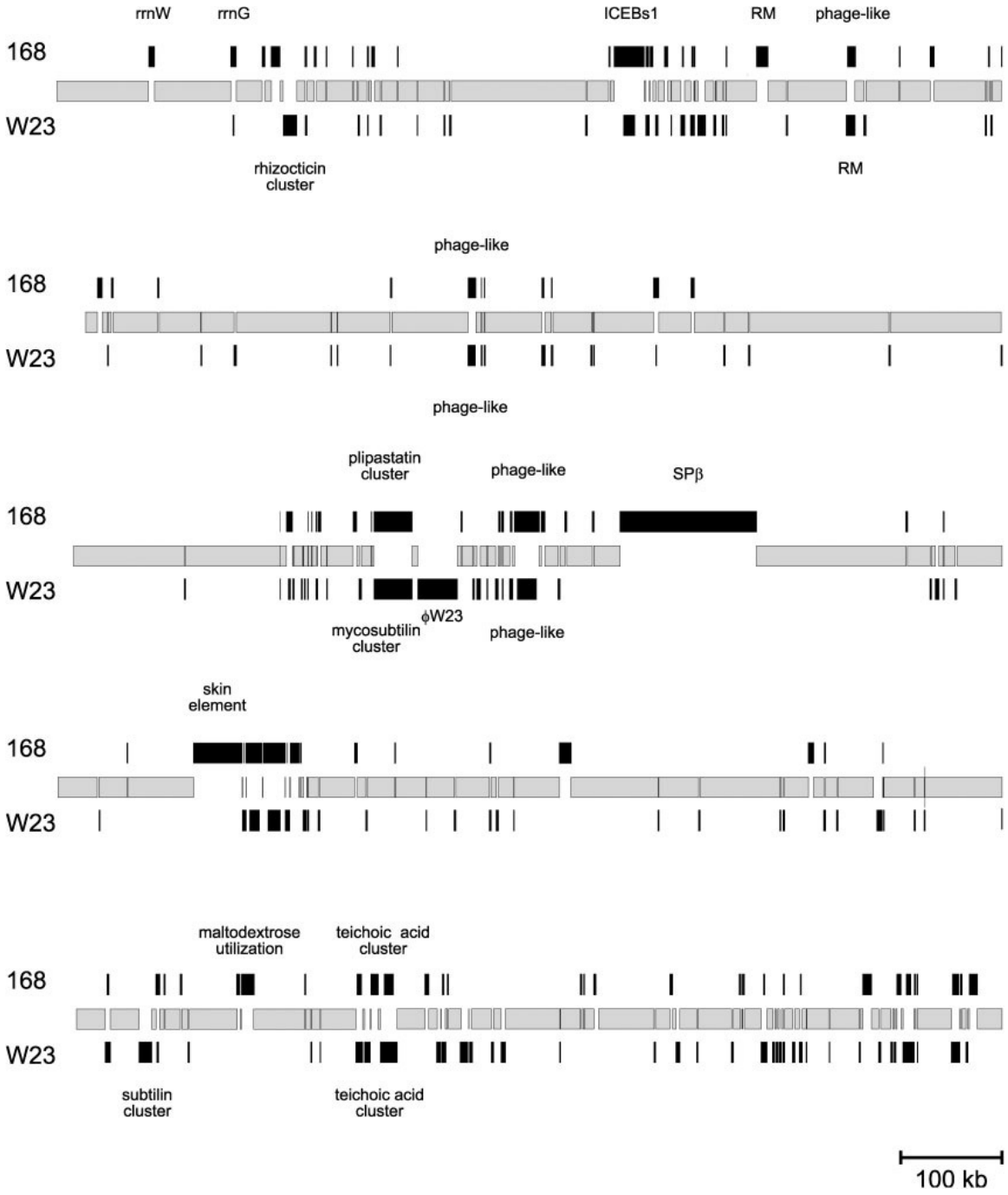


Figure 2.1: Comparison of Bsu168 and BsuW23 genomes. Auxiliary regions are shown in black above (Bsu168) or below (BsuW23) the shared common genome shown in gray. Image adapted from [166] and reproduced with permission.

Media

The composition of all media used in this experiment is listed in Table 2.3. All media were filled with Milli-Q Type 1 water (MQ) to a volume of 1 L, autoclaved, and stored at 4°C, unless otherwise noted. Filtering was done using 0.2 µm pore sterile syringe filters (Whatman).

Table 2.3: Composition of media used in the evolution experiment

Medium	Composition		Manufacturer
LB medium	25 g	LB	1
LB agar (1.5%)	25 g	LB	1
	15 g	Bacto agar	2
Spizizen's salts (10x)	60 g	KH_2PO_4	1
	140 g	K_2HPO_4	1
	20 g	$(NH_4)_2SO_4$	1
	10 g	Trisodium citrate dihydrate	3
		set to pH 7.0	
CM medium	1x	Spizizens salts	–
	0.2 g	Casamino acids	1
	1 g	Yeast extract	1
	0.2 g	$MgCl_2 \cdot 6H_2O$	1
	added after autoclaving, via filtration:		
	5 g	Glucose	1
	0.05 g	Histidine	1
	0.05 g	Leucine	1
	0.05 g	Methionine	1
TAE Buffer	0.8 mM	Tris	3
	0.04 mM	EDTA	1
	0.4 mM	acetic acid in H_2O, pH 8.5	3

[1] Carl Roth, [2] BD, [3] Sigma Aldrich

Growth Conditions

All cultures were grown either in liquid medium at 37°C, 250 rpm in lysogeny broth (LB) or competence medium (CM), or on LB agar plates (1.5%) at 37°C, 250 rpm. Cultures grown overnight in liquid media were grown for 18 – 22 h. Optical density (OD) measurements were performed at 600 nm on an Infinite M200 plate reader (Tecan).

Transformation

Cells competent for transformation were prepared according to the standard protocol developed by Albano, Hahn, and Dubnau. An overnight culture grown in LB medium was diluted in CM to OD = 0.1 and grown at 37°C, 250 rpm. Growth was monitored on an Infinite M200 plate reader (Tecan). Two

hours after the transition from the exponential to the stationary growth phase, a 0.5 mL aliquot of the culture was mixed with 500 ng DNA in a 50 mL centrifuge tube. After shaking for 45 min, 37°C, 1 mL LB was added and the culture was mixed for an additional 60 min. Next, the culture was plated on LB agar plates containing antibiotics and incubated at 37°C, 5% CO_2 overnight. On the next day, several clones were selected and inoculated in LB with antibiotics, overnight. Clones were tested using colony PCR and positive clones were mixed with DMSO (dimethyl sulphoxide, 10% v/v; Sigma Aldrich) and stored at -80°C.

2.1.2 Whole Genome Sequencing

Clonal genomes of ancestral and evolved populations were obtained using next generation sequencing (NGS) methods, in particular Illumina HiSeq. Samples were prepared by growing a frozen culture overnight on an LB Agar plate at 37° C, 5% CO_2. The subsequent day, an individual colony was selected and grown overnight in CM. A 2 mL aliquot of that culture was pelleted at 16.7 xg for 3 min, decanted, and then frozen at -20° C. Additionally, a 1 mL aliquot of the overnight CM culture was mixed with DMSO (10% v/v)and stored for reference at -80° C.

Genomic DNA was isolated from the frozen pellet using the Qiagen Dneasey Blood & Tissue Kit (Hilden, Germany) according to the manufacturer's instructions. A small aliquot of the Isolated DNA was run on a 1% agarose gel with a 1 kb plus DNA Ladder (Thermo Scientific) to check for degradation. The gel was prepared by dissolving 1% agarose (w/v) in 1x TAE buffer, by heating. Once partially cooled, Midori Green Advance DNA stain (Nippon Genetics Europe, Düren, Germany) was added to achieve a final concentration of 4×10^{-5} v/v, and the gel was allowed to harden. Due to dye light sensitivity, gels were covered while hardening and running.

Non-degraded samples were sent to GATC Biotech (Konstanz, Germany) for NGS. Sequencing was performed on an Illumina HiSeq 3000/4000 system with 150 bp paired reads and an average depth of >500.

2.1.3 Evolution Experiment Design

The evolution experiment was composed of continuous repetitions of a two day cycle, consisting of six steps: dilution, radiation, plating, colony selection and regrowth, competence induction and addition of extracellular DNA, washing and overnight growth (Scheme 2.1). Three different extracellular DNA sources were used: no DNA, self DNA (Bsu168) and BsuW23 DNA. Eight replicate ancestral clones of the Δ*comK* strain, with inducible *comK* (Bs166), were used for each DNA source. Initially, all strains were grown overnight in 1 mL LB medium at 37°C, 250 rpm.

Overnight cultures were diluted $1{:}10^3$ in fresh CM medium and grown for 4.5 h (37°C, 250 rpm) in a 24-well microtiter plate (1 mL, final OD ≈ 0.25; Greiner Bio-one). Microtiter plates were covered with rayon film adhesive covers (VWR). Cultures were radiated in the microtiter plate (without cover) in a 600 J cm^{-2} UV-C light chamber (Bio-Link BLX-E crosslinker). A 100 mL aliquot of each radiated culture was diluted $1{:}10^4$, plated onto a 10 cm LB-agar plate, and incubated overnight at 37°C, 5% CO_2. A random colony was picked from each plate using a 200 µL pipette tip, mixed into 1 mL fresh CM, and grown for 2.5 h (OD ≈ 0.25). IPTG [600 µM] was added to each culture along with genomic

Table 2.4: Competence induction conditions for various evolution experiment growth conditions

Condition	[IPTG] for induction	External genomic DNA	UV-C Radiation
Ø	600 μM	none	600 J/cm^2
B	600 μM	0.91 μg/mL Bsu168	600 J/cm^2
W	600 μM	0.87 μg/mL BsuW23	600 J/cm^2

DNA equivalent to two genomes per cell (depending on the culture condition, see Table 2.4). Genome equivalents were calculated assuming 1 bp = 650 Da and a culture density of 10^8 CFU/mL. After growing induced cultures for an additional 2 h (37°C), each culture was washed twice (16,800 xg, 1 mL CM) and grown overnight (37°C, 250 rpm). This completed one cycle and was repeated, starting again with dilution.

Samples were frozen for later analysis every second cycle, starting with cycle 3. For each culture, a 500 μL mixture with DMSO (10% v/v) was stored at -80°C.

2.2 Computational and Analysis Methods

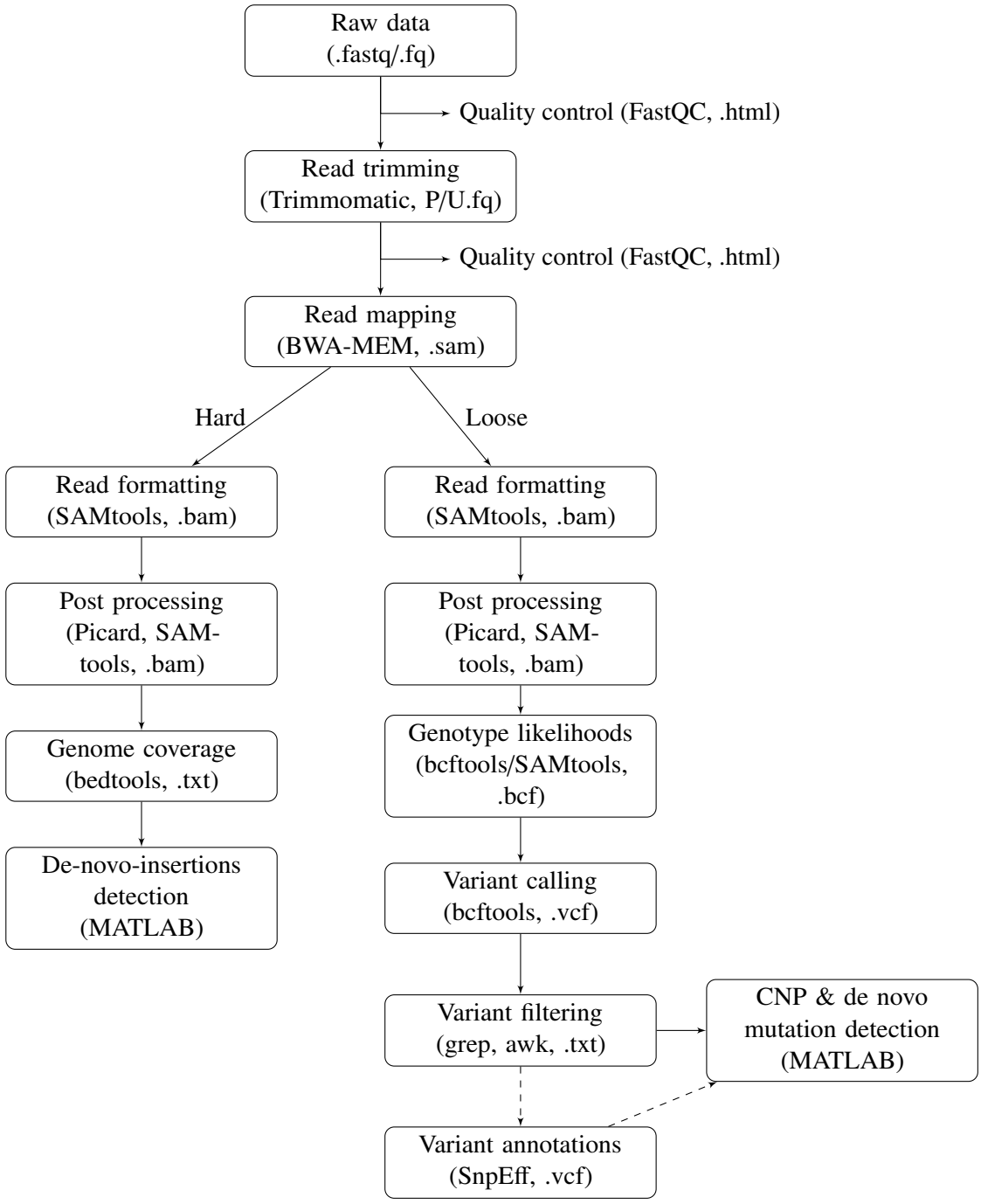
Raw reads from NGS were sent through several analysis pipelines to identify variants and coverage. Those outputs were then sent through one of two analysis algorithms to detect various types of gene transfer. Orthologous recombination was detected using a method adapted from [163] and de novo insertions were detected knowing BsuW23 auxiliary genes and the evolved sample's per base coverage. The sequencing pipelines can be found in Section 2.2.1 and the details of the analysis algorithms, including robustness, can be found in Sections 2.2.2 – 2.2.5.

2.2.1 Sequencing pipelines

A schematic outline of the sequencing pipelines used to detect orthologous recombination (cluster of nucleotide polymorphisms – CNPs), de novo variants, and de novo insertions is shown in Scheme 2.2. The corresponding code can be found in Code Snippets 2.1, 2.2, and 2.3. First, the quality of the raw sequenced genomic DNA files (fastq.gz, fq.gz) were analyzed using FastQC (v0.11.5, [168]). Raw reads were then trimmed using Trimmomatic (v0.36,[169]) to remove adapters and low quality reads. Illumina adapters were removed using the adapter file TruSeq3-PE-2.fa, and the following settings, 2:30:10:4, for seed mismatches, palindrome clip threshold, simple clip threshold, and minimum adapter length, respectively. The minimum quality to retain a base at the beginning or end of a read was set to 3, with *leading* and *trailing*, respectively. The *sliding window* was set to 4, with a minimum average quality of 15. The *minimum length* required for a read to be retained was 36 bp. Trimmed reads were sorted into four files, forward and reverse reads for pairs that survived processing (P) and those where only one partner read survived (U).

Paired reads (P) were, again, checked for quality using FastQC before being passed onto Burrows-Wheeler Aligner (BWA), specifically BWA-MEM (v0.7.12-r1039, [170]). Hard or loose mapping (Code Snip 2.2 or 2.3) was performed using BWA-MEM, depending on the desired analysis routine. Loose

Scheme 2.2: Alignment pipeline for hard and loose filtering. In parenthesis the programs for each step are listed, followed by that step's output format.



mapping was used to determine mutations between the evolved strain and the reference *B. subtilis* 168 genome; reads with a large number of SNPs due to orthologous replacement would still be mapped to the reference genome. A *mismatch penalty* (-B) and *gap open penalty* (-O) of 1 were used. Hard mapping was used to map the evolved strains' reads onto *B. subtilis* W23, in search of de novo insertions from

Code Snippet 2.1: Code snippet of initial steps for sequencing pipeline (hard and loose mapping)

```
%Quality control
./fastqc -o out_folder fwd_reads.fastq.gz
./fastqc -o out_folder rev_reads.fastq.gz
%Trimming
java -Xmx30G -Xms24G -jar trimmomatic-0.36.jar PE -threads 8 -trimlog
    ↪ TrimLog fwd_reads.fastq.gz rev_reads.fastq.gz out_fwd_PAIRED.fq.gz
    ↪ out_fwd_UNpaired.fq.gz out_rev_PAIRED.fq.gz out_rev_UNpaired.fq.gz
    ↪ ILLUMINACLIP:adapters/TruSeq3-PE-2.fa:2:30:10:4 LEADING:3 TRAILING
    ↪ :3 SLIDINGWINDOW:4:15 MINLEN:36
%Quality control
./fastqc -o out_folder fwd_reads_P.fastq.gz
./fastqc -o out_folder rev_reads_P.fastq.gz
```

Code Snippet 2.2: Code snippet for sequencing pipeline with hard mapping (following Code Snippet 2.1)

```
%Mapping
./bwa mem -t 8 -B 100 -O 100 dictionary fwd_reads_P.fq rev_reads_P.fq >
    ↪ output_hard.sam
%Formatting
./samtools view -b input_hard.sam --threads 8 -T dictionary.fasta -o
    ↪ output_hard.bam
%Post processing
java -Xms1g -Xmx3g -jar picard.jar AddOrReplaceReadGroups I=input_hard.
    ↪ bam O=output_hardRG.bam RGID=4 RGLB=lib1 RGPL=illumina RGPU=unit1
    ↪ RGSM=20
./samtools sort input_hardRG.bam --threads 8 --reference dictionary.fasta
    ↪  -o output_hard_sort.bam
./samtools index -b input_hard_sort.bam > output_hard_sort.bam.bai
%Genome coverage
./genomeCoverageBed -d -ibam input_hard_sort.bam -g genome_description.
    ↪ txt > output_coverage_per_bp.txt
```

BsuW23 DNA in the evolved species. Only segments with strong similarity to BsuW23 could be mapped, ensuring that only BsuW23 auxiliary regions would be detected. A *mismatch penalty* (-B) and *gap open penalty* (-O) of 100 were used.

Following both hard and loose mapping, reads were formatted using SAMtools (v1.3.1, [171]) to produce a binary .bam file. The dictionary (reference genome used for mapping) was assembled using BWA to index the reference genome sequence in fasta format, Picard (v2.6.0, [172]) to create a dictionary, and SAMtools to index the fasta file (Code Snippet 2.4). A complete dictionary (with indexed files) only

Code Snippet 2.3: Code snippet for sequencing pipeline with loose mapping (following Code Snippet 2.1)

```
%Mapping
./bwa mem -t 8 -B 1 -O 1 dictionary fwd_reads_P.fq rev_reads_P.fq >
    ↪ output.sam
%Formatting
./samtools view -b input.sam --threads 8 -T dictionary.fasta -o output.
    ↪ bam
%Post processing
java -Xms1g -Xmx3g -jar picard.jar AddOrReplaceReadGroups I=input.bam O=
    ↪ output_RG.bam RGID=4 RGLB=lib1 RGPL=illumina RGPU=unit1 RGSM=20
./samtools sort input_RG.bam --threads 8 --reference dictionary.fasta -o
    ↪ output_sort.bam
./samtools index -b input_sort.bam > output_sort.bam.bai
%Genotype likelihoods
./samtools mpileup -e 10 -t AD -F 0.00001 -h 80 -L 10000 -o 20 -f
    ↪ dictionary.fasta -uv input_sort.bam > output.bcf
%Variant calling
./bcftools call -vc output.bcf > output_bcfcall.vcf
%Variant filtering
grep -v "^#" output_bcfcall.vcf | awk '{split($10,R,":")split(R[1],r,"/")}
    ↪  r[2]==1 && r[1]==1 {print $2,$5}' > output_IndvMutList.txt
awk '{split($10,R,":")split(R[1],r,"/")} r[2]==1 && r[1]==1 {print $0}'
    ↪ input_bcfcall.vcf > output_bcfcall_SNPscleaned.vcf
java -Xms1g -Xmx4g -jar snpEff.jar -noStats -ud 3000 dictionary
    ↪ input_bcfcall_SNPscleaned.vcf > output_snpEff.vcf
```

Code Snippet 2.4: Code snippet of dictionary assembly

```
./bwa index -p dictionary_name dictionary_name.fasta
java -Xms1g -Xmx3g -jar picard.jar CreateSequenceDictionary R=
    ↪ dictionary_name.fasta O=dictionary_name.dict
./samtools faidx dictionary_name.fasta
```

needed to be created once for each reference genome. All dictionary files (.amb, .ann, .bwt, .dict, .fasta, .fasta.fai, .pac, and .sa) needed to have the same file name.

Next, post processing was carried out using Picard and SAMtools. First, Picard's *AddOrReplaceReadGroups* was used to combine all the read groups (sets of reads created by single sequencing runs) into one read group (4, lib1, illumina, unit1, 20, for ID, LB, PL, PU, and SM, respectively). Then, SAMtools was used to sort the reads by genomic position and, finally, index that file for fast random access.

For hard mapping, genome coverage per base was calculated using bedtools (v2.26.0, [173]). The

per base pair depth setting (-d) was used along with a genome description as a tab separated text file, containing chromosome name(s) and length(s). The chromosome name used in the genome description file matched that used as the dictionary name. Genome coverages were analyzed with a MATLAB script to detect de novo insertions.

For loose mapping, genotype likelihoods were calculated from the sorted (and indexed) bam file using SAMtools *mpileup*. The binary call format (bcf) file was generated with a *gap extension sequencing error probability* (-e) of 10, *minimum fraction of gapped reads* (-F) of 0.00001, *coefficient for modeling homopolymer errors* (-h) of 80, *maximum indel calling file depth* (-L) of 10,000, and *gap open sequencing error probability* (-o) of 20. Additionally, the output tags were written in allelic depth (AD) format (-t). Genotype likelihoods were converted to variant calling using BCFtools (v1.5, maintained by SAMtools <http://samtools.github.io/bcftools/>) using the original calling model (-c) and requesting only variant sites in the output file (-v).

Variants were filtered to include only alternate allele homozygous variants (GT = 1/1). Filtered variants were saved in a vcf file, including the complete call records, and an abbreviated version including only the variant position and alternate allele as a text file. The text file was used with various MATLAB scripts to detect CNPs and plot import and mutations events.

To obtain additional information about the mutated genes and their functional affects, the filtered vcf file could additionally be fed into SnpEff (v4.2, [174]). The *B. subtilis* reference database NC_000964 was used to annotate the called variants and predicted their functional affects. The upstream/downstream interval length (-ud) was reduced from the default value (5000 bp) to 3000 bp, because of *B. subtilis*' smaller genome size (compared to eurakyotics) and lack of exons and introns. Enhancer-like elements have been reported to be 1500 bp downstream of the gene's promoter for *B. subtilis* [175] and a study with *E. coli* found promoter activation was not diminished if binding sites were moved more than 1000 bp away from their respective gene [176]. Based on these findings, a safe overestimate of the effective upstream/downstream length was set to 3000 bp. The annonated vcf file would, then, be fed into the MATLAB scripts along with the aforementioned text file.

2.2.2 Orthologous Recombination (CNP) Algorithm

Large recombination events were detected by searching for single nucleotide polymorphisms (SNPs) that matched those found in the donor genome (matched SNPs), but not the host genome. Based upon the proximity of the matched SNP to each other, they were clustered into groups which represented one import event (CNPs, cluster of nucleotide polymorphisms). An overview of how the algorithm works is shown in Figure 2.3 and outlined below; the subsequent subsections detail the algorithm's robustness.

1. The donor genome is aligned to the host genome, and a "master list" of all the SNP differences is created (mℓSNPs). (Indels are ignored.)

2. Experimental samples are aligned loosely to the host genome and the list of SNPs is recorded.

3. The list of SNPs from (2) is compared to the master list from (1). SNPs with a position and nucleotide that do not match the master list are removed.

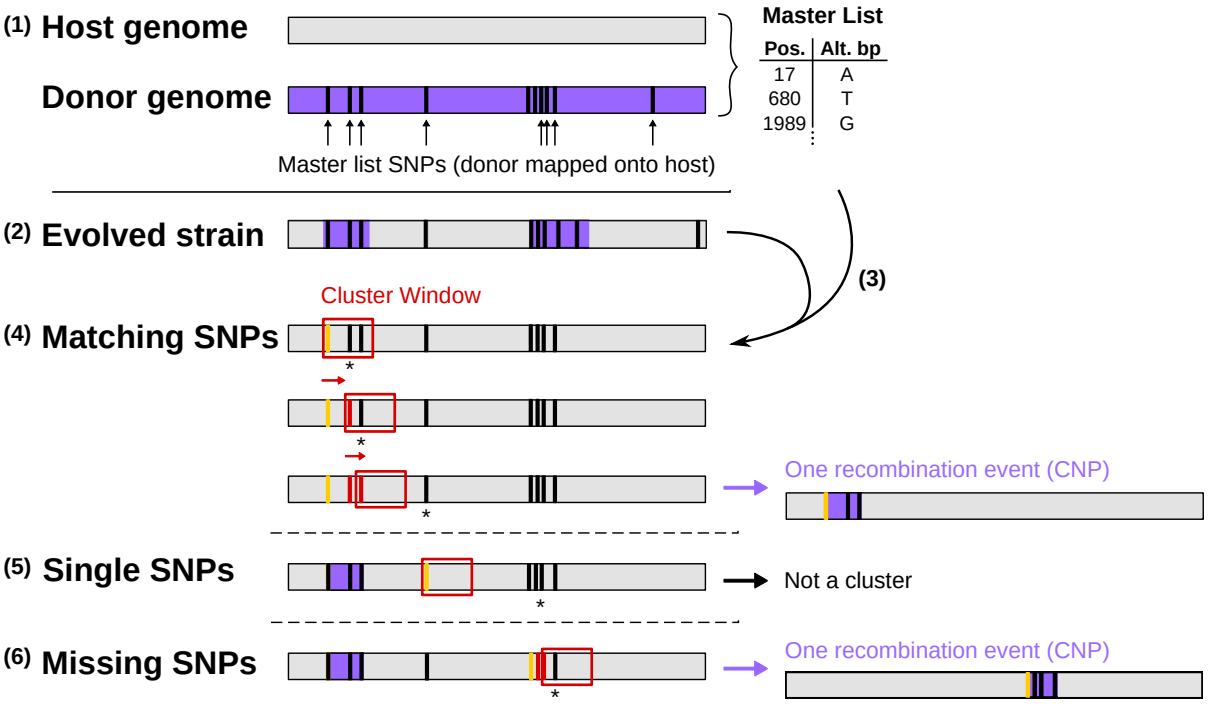
4. Remaining SNPs are combed through to find clusters of SNPs that are within 200 bp of each other (a cluster of nucleotide polymorphisms, CNP).

5. Remaining SNPs with no neighbor within 200 bp are ignored in this algorithm

6. CNPs are allowed to be missing no more than 30% of the excepted SNPs, as per the master list.

For each CNP, a minimum length (from first to last matched-SNP in a CNP), maximum length (from the SNP before the first matched-SNP to the SNP after the last matched-SNP in a CNP), and average length (the mean of the minimum and maximum lengths) are measured.

Scheme 2.3: CNP algorithm to detect orthologous recombination.



(1) Donor genome (purple) is aligned to the host genome (gray) and a "master list" (mℓSNPs) is created. **(2)** Experimental samples are aligned to the host genome and genotype variants are recorded. **(3)** SNPs in the evolved strains that do not match a mℓSNP (position and nucleotide) are removed. **(4)** SNPs are combed through with a 200 bp cluster window (red) starting with the first SNP (gold). If the neighboring SNP (starred) is inside this window, it is added to the cluster (black bar becomes red). The window moves to this new position and the query continues (new starred SNP). If the neighboring SNP is not within 200 bp, the query is closed (one recombination event). A new query opens on the last unsuccessful SNP query (last starred SNP). **(5)** SNPs with no neighbor within 200 bp are not considered CNPs. **(6)** Missing SNPs in a CNP are allowed if the percentage of missing SNPs, when including the query position (starred), is below 30%.

Test Genome

The CNP algorithm was tested using a mock genome of Bsu168 with inserted genomic segments from BsuW23. These two subspecies have a 92.4% average sequence identity, evenly distributed between intra- and intergenic regions. Replacement segments from BsuW23 of various size and position were randomly selected, avoiding only auxiliary regions from Bsu168 or BsuW23 [166]. (Auxiliary genes have no orthologous recombination site in the opposite species' genome.) The list of segments taken from BsuW23, their location in Bsu168, and the final length of the replaced segment (in the mock genome) can be see in Table 2.5. The segments had between 87 and 90% identity.

Using the mock genome outlined above, 10,900,000 mock paired reads were created using wgsim (SAMtools), with an *in silico* read error rate of 0.02 and constant base call quality (Phred score = 30, base read quality in fq file = 9). (Code Snippet 2.5) The reads were aligned using the pipeline outlined in Section 2.2.1.

Code Snippet 2.5: Code snippet to create *in silico* reads

```
./wgsim -1 150 -2 150 -N 10900000 genome.fasta fwd_reads.fq rev_reads.fq
for f in $(ls *_reads.fq)
do; sed -i 's/2.../9.../g' "$f" ; done
```

In the CNP algorithm, only homozygous SNP variants were analyzed. The measured length for each replacement segment is listed in Table 2.5 and shown in Figure 2.2. The detected segment length is smaller than the actual segment size, because replacement segments from BsuW23 were not necessarily bookended by SNPs. The algorithm detected and accurately measured all of the replacement segments down to 60 bp. Below 60 bp the algorithms percent error becomes significant—the location of the SNPs in the replacement segment limit the algorithms ability to accurately detect the replacement segment length [$(90 - 100)$ bp$/100$ bp $= -10\%$ error, $(990 - 1000)$ bp$/1000$ bp $= -1\%$ error]. The algorithm, including the *in silico* 0.02 read error rate, produced no false positives.

Table 2.5: Mock genome segments to test CNP algorithm.

Extracted seg. (BsuW23)			Ident. with	Replacement location (Bsu168)			Recov. seg.
Start pos.	End pos.	Size (bp)	Bsu168 seg.	Start pos.	End pos.	Size (bp)	size (bp)
240,601	265,622	25,021	92%	252,317	277,339	25,022	24,934
224,448	231,129	6681	91%	233,023	239,697	6674	6513
299,999	304,890	4891	94%	314,071	318,962	4891	4828
1,000,000	1,000,719	719	~92%	1,038,844	1,039,563	719	695
1,093,197	1,093,616	419	~92%	1,130,919	1,131,337	418	373
2,069,034	2,069,213	179	~91%	2,080,863	2,081,042	179	160
2,000,535	2,000,603	68	~93%	2,010,534	2,010,602	68	66
2,022,740	2,022,859	119	~87%	2,028,205	2,028,324	119	103
3,052,497	3,052,556	59	~92%	3,269,067	3,269,126	59	30

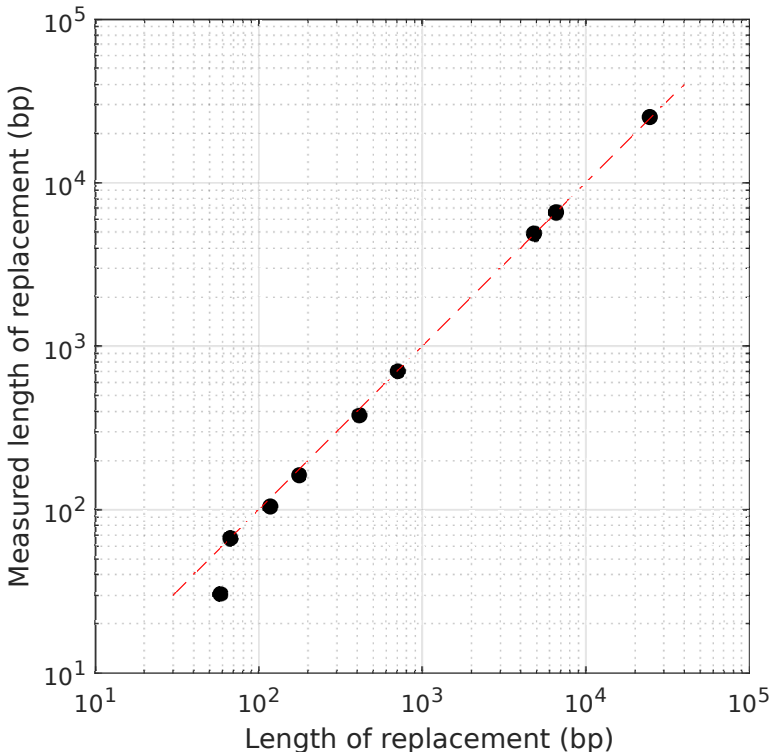


Figure 2.2: Replacement lengths and measured lengths for mock genome Bsubmany2W23 measured using the CNP algorithm. The dashed line (red) denotes a perfect algorithm, replacements are measured as their exact size.

Sensitivity and False Positives

After confirming that the algorithm could correctly identify orthologous recombination of BsuW23 segments in Bsu168, the sensitivity of the algorithm to de novo mutations was analyzed. (Errors in the reads, such as those *in silico* read errors mentioned above, are removed during variant calling.) To rigorously test the algorithm's sensitivity, a list of SNP locations was made by randomly selecting base pair positions in Bsu168, without duplication. That list of random locations was compared to the master list and if a position matched, it was assumed that the base pair change also matched. (This assumption was done to make the computation easier. It results in a three-fold overestimation of the effect of de novo mutations, because the base pair change at these random locations could be one of three de novo base pairs–e.g., A could mutate into T, C, or G. [A to A is ignored because it is not a mutation.])

Up to 4000 random de novo mutations were created *in silico* and run through the CNP algorithm (Figure 2.3). It is first at 4000 random de novo mutations that several false positives are consistently detected in every simulation. These false positives are an artifact of the 200 bp cluster window size (Section 2.2.2) and are further scrutinized by the addition of a "missing SNP" filter (See Interrupted CNPs and Cluster Window Size).

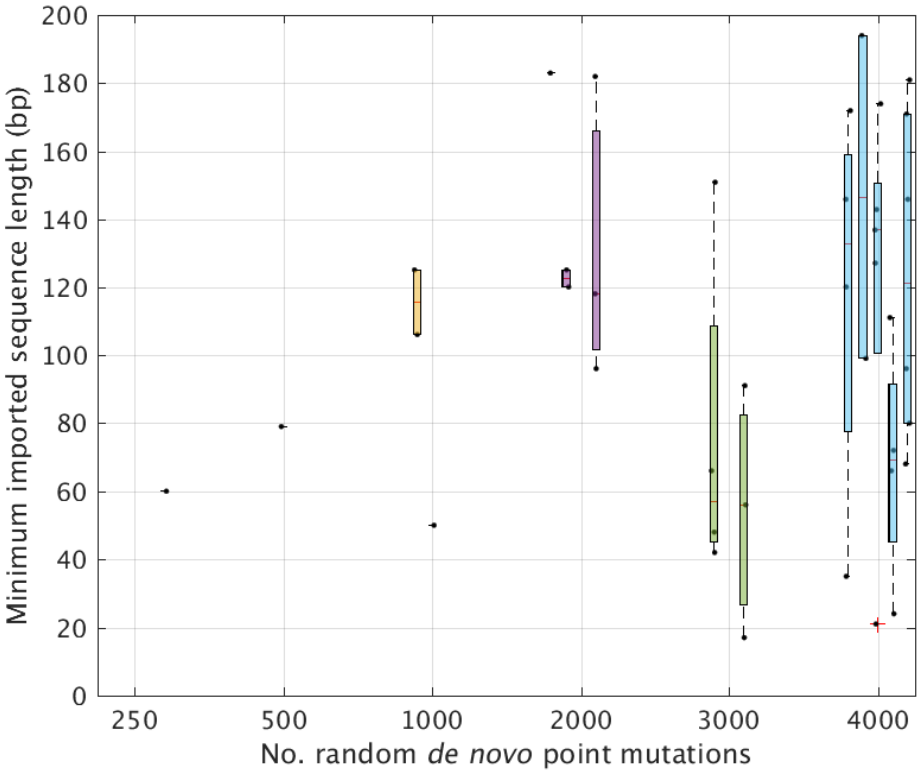


Figure 2.3: Random de novo point mutations *in silico*. A list of 250, 500, 1000 (yellow), 2000 (purple), 3000 (green), 4000 (blue) random de novo mutations were created *in silico* and passed through the CNP algorithm (in 5x replicate). Starting around 4,000 de novo mutations, several < 200 bp false positive CNPs are detected.

mℓSNP Distributions

The distribution of SNPs from the master list (mℓSNPs) were analyzed to determine if the 200 bp cluster window size should be adjusted for these two species. The histogram of the flank lengths (the distance from one mℓSNP to the next, running in ascending position) shows that the majority of the mℓSNPs lie quite close to each other (Figure 2.4). This is more evident when the cumulative probability of the flank lengths is plotted, as in Figure 2.5 (top, blue curve). Looking closely at the asymptotic portion of the curve (Figure 2.5, lower left), one sees that at a length of ~80 bp, there is already a 99% probability that from any given mℓSNP, a neighboring mℓSNP will be found. This value increases to nearly 100% by 200 bp. This argues in favor of a cluster window size of at least 80 bp.

While the distribution of flank lengths describes how probable it is to find a neighboring mℓSNP within x bp, it does not account for the size of the genome and the possibility of large stretches with no mℓSNPs (either due to 100% identity or auxiliary regions). To account for these, the distribution in Figure 2.4, D, was weighted by flank length and normalized by genome size—explicitly, $(D \times \mathit{flank\ length})/\mathit{genome\ length}$. This weighted and normalized distribution, in layman's terms, describes how much of the genome has been accounted for (in bp) when all mℓSNPs with a flank distance of $< x$ are accounted for (Figure 2.4 top, orange curve). Here it is clear, that while a cluster window size of 200 bp accounts for virtually all mℓSNPs, 15% of the genome still remains "untouched", because these regions have extremely low mℓSNPs densities. From 100 – 1000 bp, this distribution asymptotically approaches 87%. After 1000 bp, the cumulative probability jumps stepwise, as large segments of auxiliary regions are incorporated.

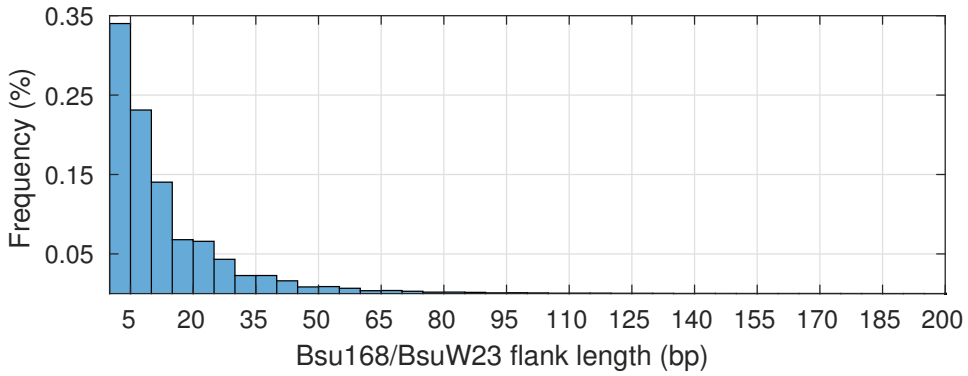


Figure 2.4: BsuW23 flank length distribution (distance between mℓSNPs). Flank lengths are grouped into 5 bp bins.

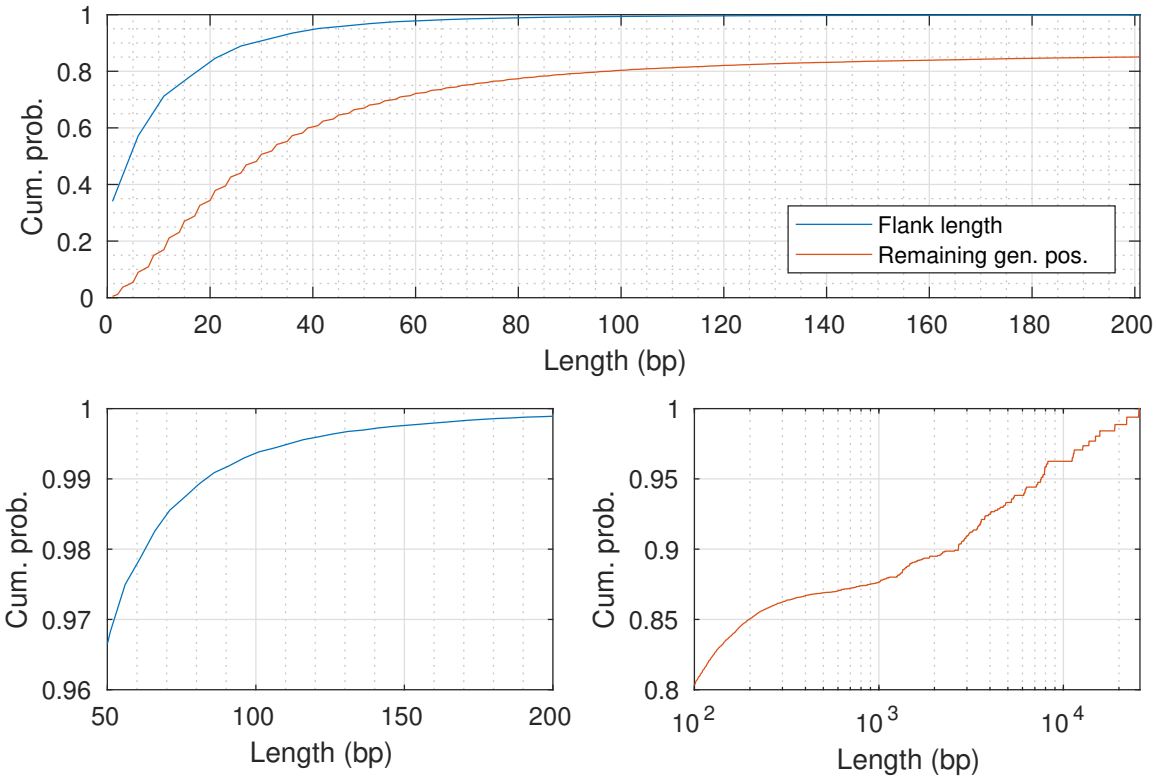


Figure 2.5: Cumulative probability of BsuW23 flank lengths (blue) and weighted flank lengths (orange). Ninety-nine percent of all mℓSNPs have flank lengths less than 80 bp. At the same length of 80 bp, 20% of the genome has yet to be accounted for, due to auxiliary regions. The two bottom plots are zoomed-in sections of the top plot. The x-axis of the lower right plot is logarithmic.

Interrupted CNPs and Cluster Window Size

Analyzing sequencing data from several evolved samples made it evident that a substantial handful of detected CNPs were missing the majority of their mℓSNPs,—i.e., two SNPs were found that matched the master list and were within 200 bp of each other, but the expected mℓSNPs in between were not present. It is likely that these "missing" mℓSNPs were caused by the mismatch repair system [177]. Bubendorfer *et al.* mentions the presence of a low frequency of interrupted CNPs, but do not exclude any from their analysis. To determine if we could, also, overlook the presence of these interrupted CNPs, we created four quality checks to quantize the accuracy of the CNP algorithm on our species and experimental setup.

The first two quality checks focus on SNPs within detected clusters: C_{pos}, the total number of matching SNPs over all clusters and C_{neg}, the total number of mℓSNPs in a cluster which were *not* found in the sample (Figure 2.6(a)). These checks are a direct measure of accuracy of the detected clusters. The remaining two quality checks focus on the non-cluster sections of the genome: nC_{neg}, the total number of matching SNPs not assigned to a cluster and nC_{pos}, the total number of mℓSNPs not detected in the sample and not assigned to a cluster (Figure 2.6(b)). The quality checks were performed on four BsuW23 replicates from cycle 9 (Figure 2.7).

A maximum number of interruptions (mℓSNPs missing from clusters), or in other words, an interruption

Figure 2.6: Graphical description of four quality check measurements. Segments of the genome are labeled as clustered (purple) and non-clustered (grey) sections. Black columns are SNPs present in the sample; white colums with a dashed outline are SNPs not present in the sample but listed in the master list. (a) SNPs contributing to C_{pos} (black) and C_{neg} (white). (b) SNPs contributing to nC_{pos} (white) and nC_{neg} (black).

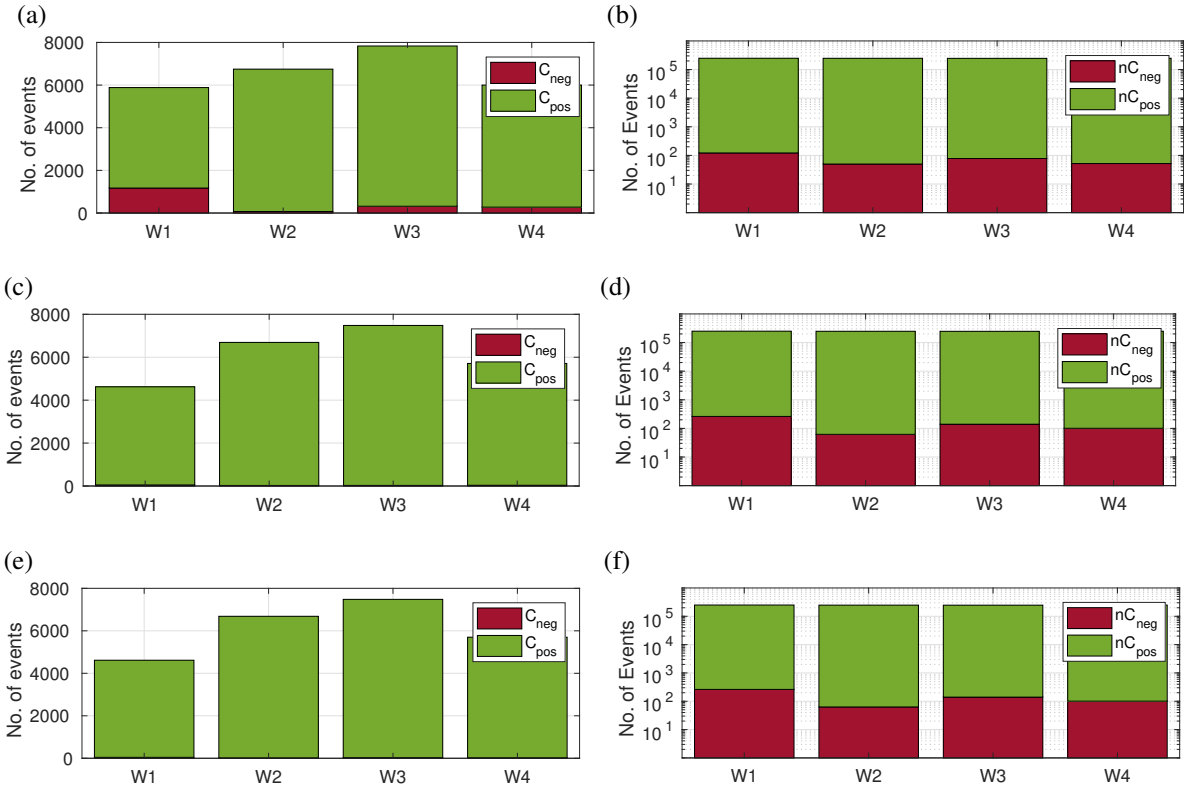


Figure 2.7: Stacked bar plots showing quality checks for various cluster interruption thresholds on cycle 9 samples evolved with BsuW23 DNA (clones 1). (a, c, e) Cluster and (b, d, f) non-cluster quality checks. (a, b) No cluster interruption threshold. (c, d) Cluster interruption threshold of < 30% missing SNPs per cluster. (e, f) Cluster interruption threshold of < 10% missing SNPs per cluster.

threshold, was set to see its effect on the quality check measurements. The upper threshold was defined as *missing mℓSNPs/total expected mℓSNPs* and was calculated dynamically as a cluster was detected and then grew in size (until no matching SNPs were found within 200 bp, or sooner if the threshold was exceeded). Initially, a loose threshold of <30% missing SNPs was applied ($th_{int} = 0.3$), because the cycle 9 BsuW23 replicates showed a gap in the number of interrupted CNPs there (Figure 2.8). The results of $th_{int} = 0.3$ on the quality checks can be seen in Figure 2.7(c, d), along with $th_{int} = 0.1$ in Figure 2.7(e, f).

The implementation of $th_{int} = 0.3$ reduced the number of false positives within clusters, C_{neg}, by an order of magnitude, on average. It dropped the fraction of false positives from 20 to < 1%. Further decreasing th_{int} to 0.1 only reduced the number of false positives by a minimal amount. Looking at the

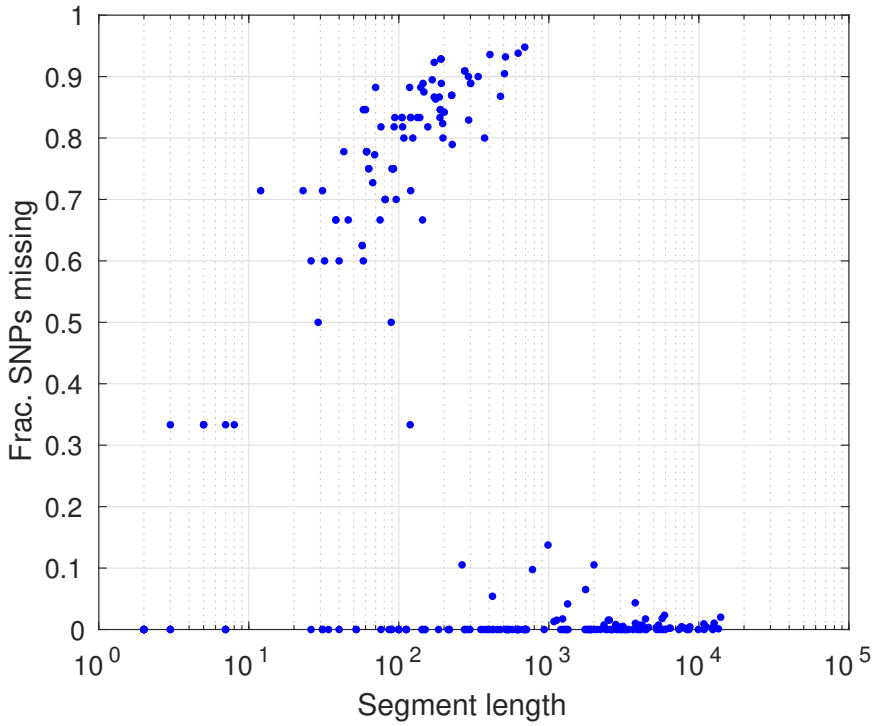


Figure 2.8: Percentage of missing SNPs as a function of cluster length for cycle 9 samples evolved with BsuW23 DNA (clones 1). A gap is visible below 30% missing SNPs.

regions outside of the clusters, $th_{int} = 0.3$ caused the number of false negatives, nC_{neg} to double. This was expected as splitting or eliminating clusters results in a larger number of lone matched SNPs. The fraction of false negatives remained below 0.1% with $th_{int} = 0.3$. Increasing th_{int} further to 0.1 had no effect on the fraction of false negatives.

Because of the sharp decrease in the number of false positives (C_{neg}) and no significant change in the number of false negatives (nC_{neg}), $th_{int} = 0.3$ was incorporated into the CNP algorithm. Further decreasing th_{int} to 0.1 yielded negligible changes in C and nC.

To better understand the influence the cluster window size had on CNPs, the quality checks were also measured keeping $th_{int} = 0.3$ but using three different cluster window sizes: 150, 200, and 300 bp. The results showed varying the cluster window size had a negligible effect on the quality measurements. The percentage of false positives and negatives (C_{neg} and nC_{neg}, respectively) remained below 1% for all three cluster window sizes (Figure 2.9). We chose a cluster window size of 200 bp because of the $> 99.9\%$ probability of finding a second mℓSNP within that distance, and for consistency with previous publications [163].

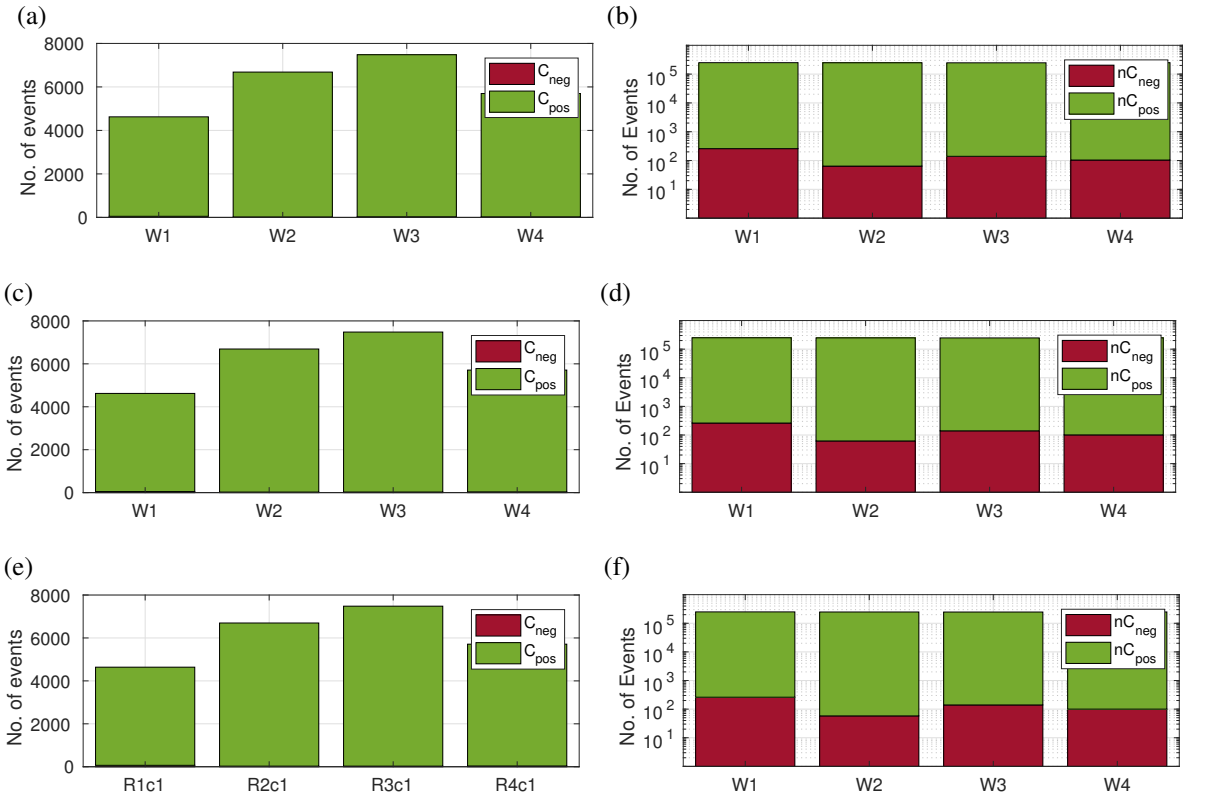


Figure 2.9: Stacked bar plots showing quality checks for various cluster window sizes on cycle 9 samples evolved with BsuW23 DNA. (a, c, e) Cluster and (b, d, f) non-cluster quality checks using th_{int} = 0.3. (a, b) 150 bp, (c, d) 200 bp, and (e, f) 300 bp cluster window size.

2.2.3 SNP Flank Length Bias

Origin of Flank Length Bias

Initially, the flank length of mℓSNPs was measured to determine an appropriate cluster window size for the CNP algorithm (Section 2.2.2). Upon closer inspection, it became clear that there was a bias to multiple-of-three lengths—i.e., 3, 6, 9, etc. (Figure 2.10). To ensure that this module three bias was not an artifact of the alignment pipeline, several mock genomes with known mℓSNP distributions were fed through the process.

To create the mock genomes, one of two distributions were used: (1) An exponential decay distribution, obtained by fitting and exponential function to the BsuW23 flank lengths distribution or (2) An exponential distribution with a bias to multiples of three, obtained by adding a constant multiple of three bias to the distribution from (1). (See Figure 2.11(a)) From each of those distributions, 254,531 values (equivalent to the number of mℓSNPs) were randomly selected and rounded up to the nearest integer. The cumulative sum of those values were taken as point mutation locations, starting with the first mutation occurring at genome position 1. Positions were saved in a text file and those larger than the length of the reference genome were ignored. Next, the Bsu168 fasta file was modified to have a cytosine base at every position listed in the positions file (-mc C) using bedtools (Code Snippet 2.6). Cytosine was chosen as the

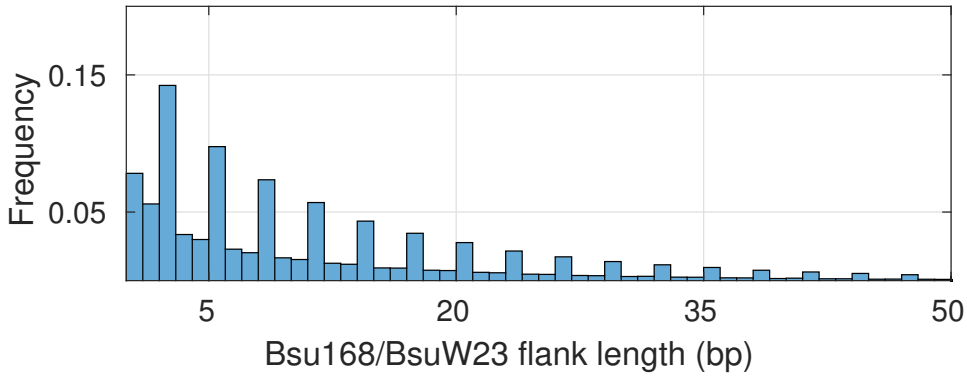


Figure 2.10: Flank length distribution for BsuW23 mℓSNPs with 1 bp bins. A bias to flank lengths of multiples of three is evident.

Code Snippet 2.6: Code Snippet to create a modified fasta file, replacing given positions with the alternate allele C.

```
maskFastaFromBed -fi wildtype.fasta -fo modified.fasta -bed
    ↪ list_of_all_cumulative_sum_positions.txt -mc C
```

replacement base pair for all the positions in the cumulative sum for simplicity. Hard masking, replacing all positions with N (unknown base pair), was not performed because N positions always negatively impact the aligning score when using BWA. Replacing all the cumulative sum positions instead with C, allowed for the possibility that the original base pair would be replaced by itself—i.e., not replaced. No correction was made for those positions. As distribution of A, C, G, T base pairs over the genome was roughly constant, the effective number of mutations present in the mock genomes was reduced by $1/4$. Mock reads were made out of the modified fasta file, now including the C base pair replacements, and run through BWA-MEM, SAMtools, and bcftools, as outlined in Section 2.2.2.

The two distributions, exponential decay and exponential decay with three-bias, were used to test the alignment pipeline. After aligning the mock reads to the Bsu168 reference genome, the flank length distributions were once again measured (Figure 2.11(a)). The input and output for each distribution set were identical in shape. Reads representing a non-bias distribution of SNPs were aligned through the pipeline to yielded a non-bias distribution of flank lengths, and reads with a bias distribution of SNPs yielded a matching bias distribution of flank lengths. This confirmed that the alignment pipeline does not introduce positional SNP artifacts.

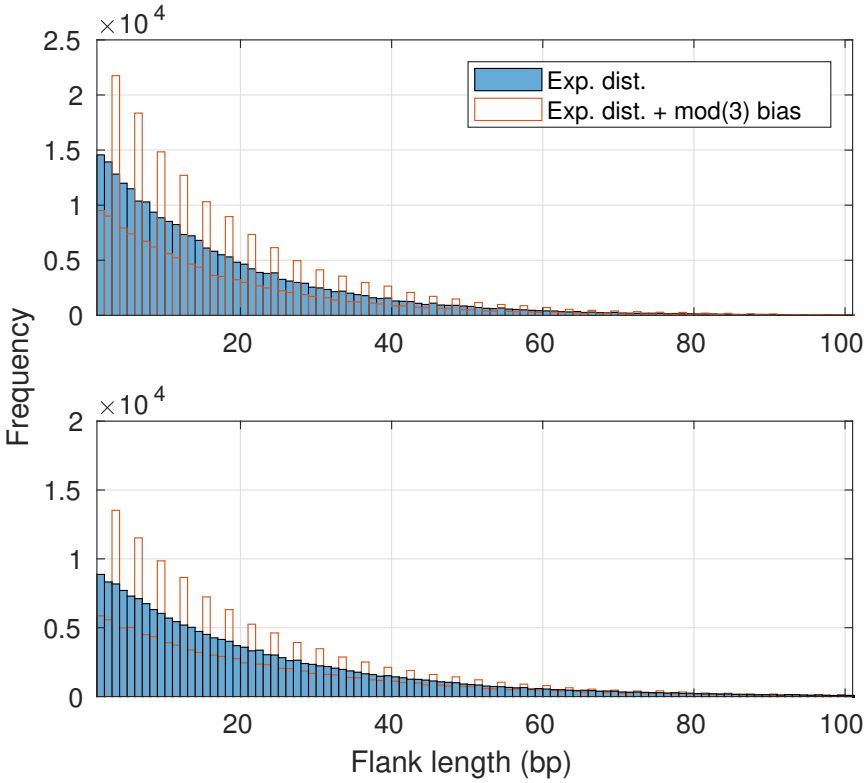


Figure 2.11: (top) Mutation distributions used to create mock genomes to check for artifacts of the alignment pipeline. Exponential distribution with (orange) and without (blue) a bias to multiple of three flank lengths. (bottom) Distributions obtained after aligning the reads from the mock genomes. The legend, here, is identical.

2.2.4 Orthologous Recombination (SPI) Algorithm

Initially, all SNPs that could not be grouped together into CNPs were assumed to be de novo variants. Cycle 9 samples which received no DNA or Bsu168 DNA had on the order of 40 de novo variants per replicate, where as those which received BsuW23 DNA had on the order of 70 mutations (removing those counted in CNPs and indels which matched BsuW23). Our original hypothesis was the addition of external DNA had either no influence on the number of de novo variants or a negative effect, as imported genomic DNA segments replace damaged DNA sections. The two-fold increase in de novo variants, in BsuW23 replicates, hinted that a significant portion of those mutations could actually have been SPIs (single-polymorphism imports, import events that only included only one mℓSNP). SPIs would have had DNA segments, possibly on both sides of the mℓSNP, that were identical in both the host and donor strains.

Of the odd 70 de novo variants found in each of the BsuW23 cycle 9 replicates, several could have been SPIs because the mutation position and alternate base pair matched those in the master list (regardless of the surrounding length of identical sequence). This did not completely account for the overwhelming increase in de novo variants found in the BsuW23 samples. Furthermore, it could not be ruled out that those alleged SPIs were not simply de novo mutations. In cycle 9, replicates which received no DNA or self-DNA, 5 – 15 de novo mutations fell on a master list position with the corresponding base pair – i.e.,

false positives. The small fraction of possible SPIs out of all de novo variants, and the number of SPI false-positives being comparable to the total number of potential SPIs led us to not consider SPIs in our analysis. We assumed all mutations which could not be grouped into CNPs were de novo variants.

2.2.5 De novo Insertions Algorithm (Auxiliary Genes Algorithm)

Auxiliary BsuW23 segments (Section 2.1.1) cannot be found using the CNP algorithm, because they have no orthologous segments in Bsu168. A separate algorithm was designed, to detect the integration of these BsuW23 specific genes in the evolved replicates. Start and end positions of the auxiliary BsuW23 genes were taken from Supplementary Table S1 of [166]. The algorithm works as follows:

1. Experimental samples are aligned to the *donor* genome using hard mapping.
2. The non-zero coverage for the entire genome and the moving-average coverage at each base position for each donor auxiliary-region are calculated.
3. The number of positions within each donor auxiliary-region with a moving-average coverage above the genome average are counted to determine the fraction of gene transfer. Transfers with a percent error greater than 20% are ignored, corresponding to a minimum measured length of 200 bp.

Test Genome

To test the detection limit of the algorithm, a mock genome was created by inserting segments of various BsuW23 auxiliary genes into the Bsu168 genome at their "native" position from the BsuW23 genome. The auxiliary segments taken from BsuW23, along with their import location in the mock genomes, is listed in Table 2.6. Using the mock genome outlined above, mock paired reads were created using wgsim and set to a constant base call quality (Phred score = 30, base read quality in fq file = 9), see Code Snippet 2.5. The reads were, then, aligned to the BsuW23 genome following the hard mapping pipeline outlined in Section 2.2.1.

The gene lengths measured using the novel gene algorithm are listed in Table 2.6 and shown in Figure 2.12. One sees that measured gene length begins to vary significantly from the inserted gene length, starting around 500 bp. We choose to set the upper threshold for percent error at 20%, which corresponds to 200 bp, to allow for the detection of smaller accessory genes. Of the 157 accessory BsuW23 genes, two are less than 200 bp in length and, therefore, not detectable using this algorithm.

Table 2.6: Mock genome segments to test auxiliary genes algorithm.

Extracted seg. (BsuW23)			Replacement loc.	Recov. seg.
Start pos.	End pos.	Size (bp)	start pos. (Bsu168)	size (bp)
1,955,075	1,994,276	39,201	1,955,075	39,177
1,955,075	1,974,675	19,623	1,955,075	19,564
2,527,713	2,532,586	4874	2,527,713	4842
1,955,075	1,959,018	3943	1,955,075	3887
573,231	574,430	1199	573,231	1162
1,955,075	1,955,575	500	1,955,075	458
3,294,677	3,295,000	323	3,294,677	311
1,235,641	1,235,913	273	1,235,641	237
3,886,616	3,886,877	262	3,886,616	225
1,955,075	1,955,300	225	1,955,075	190
1,955,075	1,955,274	199	1,955,075	159
1,955,075	1,955,246	171	1,955,075	131
1,955,075	1,955,224	149	1,955,075	106
1,955,075	1,955,199	124	1,955,075	80
1,955,075	1,955,175	100	1,955,075	39

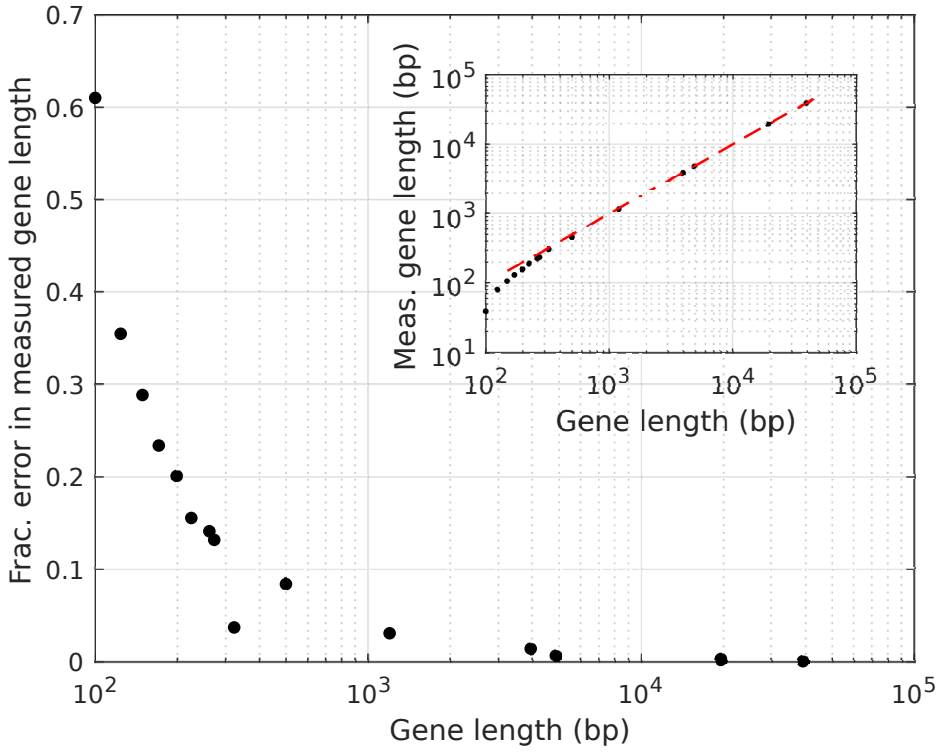


Figure 2.12: Novel gene algorithm test. Percent error of the measured gene length (for BsuW23 auxiliary gene segments in Bsu168) as a function of insert length (logarithmic). (inset) Measured gene length as a function of inserted length (both logarithmic). The dashed line (red) marks $x = y$.

2.2.6 Contaminated or Mislabeled Samples

Due to the extreme detail of the information one receives from WGS and its ability to fingerprint exactly which conditions or samples were used, contaminated or mislabeled samples could easily be removed from the analysis.

Samples which had de novo insertions from species other than that from which they had received DNA throughout the course of the experiment were removed due to mislabeling. The sensitivity and false-positive testing for the CNP algorithm showed that several < 200 bp false positives first appear when at least 4000 SNPs are represent in a sample (see Section 2.2.2). This ruled out the possibility of the Bsu168 recombination events being due to false-positives, because all of our no DNA and Bsu168 DNA replicates had only 40 variants. Furthermore, one-time pipetting errors could be ruled out when a similar number and range of import events were seen in BuW23 samples as in samples receiving Bsu168 DNA or no DNA.

In addition to detecting mislabeled samples, contaminated samples could be detected as they were sequenced at multiple time points, and multiple replicates were sequenced at each time point. The first was detected when, between two consecutive time points, the majority of gene replacements at the later time point were new and those from the prior time point lost. The second was detected when one of eight replicates showed an import or de novo variant pattern that did not match other replicates, and was no longer detected at later time points.

Methods – Population Dynamics Experiment

"Un granito de arroz, una micra, un neutrón, un paramecio, un micro-chip nipón; sin microscopio ya no te veo.... Tan superguay."

—La Casa Azul

3.1 Experimental Methods

3.1.1 Strains and Media

The bacteria used in the population dynamics experiment were derivatives of *B. subtilis* subsp. *subtilis* str. 168 (referred to as Bsu168) and are listed in Table 3.1. Detailed descriptions of how the derived strains were genetically engineered can be found in [165]. All cultures were grown either in liquid medium at 37° C, 250 rpm in lysogeny broth (LB) or competence medium (CM), or on polyacrylamide gels. Cultures grown overnight in liquid media were grown for 18 – 22 h. Optical density (OD) measurements were performed at 600 nm.

Table 3.1: Bacterial strains used in the population dynamics experiment

Strain	Alias	Relevant genotype	Source/reference
BD630	Bsu168	*hist leu met*	–
BD2711	*comK-gfp*	*hist leu met*, P_{comK}-*gfp-cat*[a,b]	[88]
Bs056	Δ*rok*	*hist leu met*, *rok::kan*[a]	This study, [165]
Bs075	Δ*comK*	*hist leu met*, *comK::kan*[a]	This study, [165]
Bs139	wt-*gfp*	*hist leu met*, *amyE::*P_{rrnE}-*gfpmut2*	This study, [165]

[a] *cat*, *spc*, *kan* stand for resistance to chloramphenicol, spectinomycin, and kanamycin, respectively
[b] Inserted by Campbell-like integration

Media compositions used in this experiment are described in Table 3.1.1, or in the case of LB and CM in Table 2.1.1. All media were filled with Milli-Q Type 1 water (MQ) to a volume of 1 L, autoclaved, and stored at 4° C, unless otherwise noted.

Conditioned medium was obtained by diluting an overnight culture of Bsu168 to OD=0.1 in fresh CM and re-growing the culture to the stationary growth phase T_0. T_0 denotes entry into the stationary phase and is defined as the time point at which the OD switches from exponential increase to a nearly constant value. The T_0 culture was then be centrifuged (17,600 xg, 3 min) and filtered (0.2 µm pore sterile filters).

Structured polydimethylsiloxane (PDMS) pads were made by pouring the PDMS mixture, devolatizied, onto negatives of the PDMS structure made out of polyoxymethylene, and heating them overnight at 60°C.

Polyacrylamide gels were made my mixing TEMED and ammonium persulfate to a MQ–acrylamide/bis-acrylamide mixture, in a volume of 10 mL. The solution was quickly pipetted

Table 3.2: Composition of the various media used in the population dynamics experiment

Medium	Composition		Manufacturer
Phosphate buffered solution (PBS)	2	Tablets	3
Polydimethylsiloxane gel (PDMS)	10 parts	PDMS	4
	1 part	Sylgard 184 curing agent (cross linker)	4
		for preparation see text	
Polyacrylamide gel	20% (v/v)	Acrylamide/bis-acrylamide (29:1 ratio)	2
	0.1% (v/v)	Ammonium persulfate	1
	0.1% (v/v)	TEMED	1
		for preparation see text	

[1] Carl Roth, [2] Sigma Aldrich, [3] Thermo Scientific, [4] Dow Corning

between two glass cover slides with 1.5 mm spacers. After waiting several hours to ensure the gel had completely solidified, spacers were removed and the gel (with microscope slides) was soaked in MQ for several minutes. This short soaking phase prevented the gel from tearing when the microscope slides were removed. The gel was cut to the appropriate dimensions and soaked twice in MQ for at least 5 h to ensure diffusion of any remnant acrylamide, out of the gel. The cut gels were stored in fresh MQ until use [178].

3.1.2 Population Dynamics – Experiment Design

Population dynamics were measured collectively and individually. Collectively, two strains were mixed in solution and the dynamics of that mixed population were tracked. The experimental conditions for that part of the experiment are detailed in "Stationary Phase Dynamics, Experiment Design". Individually, cells were grown in a flow chamber for up to 24 h to track individual growth rates and cell linages. "Those experimental conditions are below in Single Cell Microscopy, Experiment Design".

Stationary Phase Dynamics, Experiment Design

Overnight cultures of both strains (competitors) were diluted to OD=0.1 in fresh CM and grown separately to the stationary growth phase T_0. Cells were then diluted 10-fold into conditioned medium and mixed in a 1:1 ratio, where one competitor strain carried a *gfp* reporter. Conditioned medium was made as describing in Section 3.1.1 and was always fresh for each experiment. Cells were incubated for 24 h. Cell suspensions of 10 µL were taken at several time points and mixed with 1 mL PBS. In all competition experiments, one of the strains was labeled fluorescently with *amyE::*P_{rrnE}*-gfpmut2*. The fraction of the fluorescent reporter strain was measured using a BD Canto II flow cytometer (BD Bioscience, Franklin Lakes, USA) equipped with three solid-state lasers at 405, 488, and 561 nm. For detection of GFP fluorescence, a 530/30 filter was used. The photomultiplier voltage for forward scatter (FSC), side scatter (SSC), and GFP was set at 100, 351, and 450, respectively. To exclude particles smaller than *B. subtilis*,

the threshold for FSC was set at 250. At least 10,000 cells/sample were analyzed using BD FACSDiva 6 software (BD Bioscience, Franklin Lakes, USA).

During exponential growth, *B. subtilis* forms chains. FSC and SSC signals of chains observed in the flow cytometer were slightly increased compared to single cells. All events were counted as single cells because flow cytometry measures events and cannot distinguish between single cells and groups of cells forming a chain. Since only non-competent, exponentially growing cells form chains, the ratio between cells with a higher probability of entering the K-state and those with a lower probability of entering the K-state, may be overestimated. This overestimation of the fraction of K-state leads to an underestimation of the fitness cost—i.e., the fitness cost of competence may even be larger.

Single Cell Microscopy, Experiment Design

An overnight culture of BD2711 was diluted to OD=0.1 in fresh CM and grown to T_0. Cells were diluted 20-fold into fresh pre-warmed CM and sandwiched between a glass cover slide and a 1.5 mm-thick polyacrylamide gel (see Section 3.1.1. The glass cover slide, with polyacrylamide gel and diluted sample, was sealed onto the flow chamber using picodent twinsil (Picodent), confining the polyacrylamide gel between the cover slide and the structured PDMS pad (Figure 3.1) [179], [180].

The in-house flow chamber was mounted onto an inverted microscope (Nikon Eclipse TE2000-E). The flow chamber was made of polyoxymethylene with a $6 \times 56 \times 1$ mm channel where a 6×26 mm hole was cut into the center. The hole was sealed using PDMS which formed a structured pad with an array of 0.5 µm tall pillars spaced 1 µm apart from each other. The channel was sealed with the glass cover slide and gel, as described above. Medium was flushed through the chamber at a rate of 10 µL/min during image acquisition and at speeds of up to 500 µL/min to initially fill the chamber. A peristaltic pump was used to control flow rates and batch culture as medium. (the unfiltered batch culture which had been used to create the conditioned medium, kept at 40° C – to obtain 37° C when the culture reached the chamber, constantly stirred).

Differential interference contrast (DIC) images were taken at 10 min intervals for up to 24 h. Images

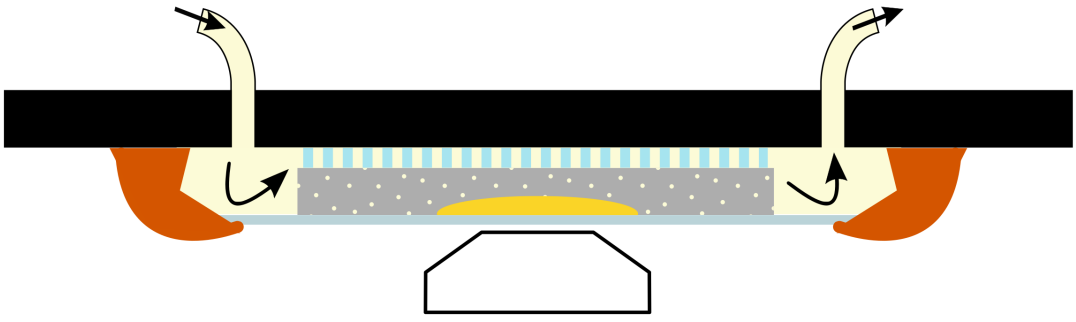


Figure 3.1: The sample (yellow) is sandwiched between a glass cover slide (light gray) and a porous gel (dark gray, here polyacrylamide). The porous gel allows signaling molecules and nutrients from the medium (cream) to diffuse through to the sample, but not bacteria. The porous gel rests on an array of PDMS pillars (blue) to hold the gel fixed on the cover slide. The PDMS pillars are attached to the chamber lid (black) which contains two outlets, allowing for flow injections. The cover slide is sealed to the chamber lid using picodent twinsil silicone (orange).

were taken following the "correlation images" method (Section 3.2.2) [181]. A z-stack height of ±3 μm, vertical step of 250 nm, and standard deviation of 700 nm were used.

3.2 Computational and Analysis Methods

3.2.1 Stationary Phase Dynamics – Selection Coefficients

Collective population dynamics were quantized using selection coefficients. Classic Malthasian parameters could not be calculated because growth was explicitly in the stationary phase. Instead, the replicator equation was used to determine selection coefficients s_{ij} by fitting the fraction in the population as a function of time with Equation 1.8, where $x_i(t)$ is the frequency of type i at time t, and λ is the ratio of initial frequencies of both competitors [123]. It was assumed that the difference $s_{ij} = r_i - r_j$, between the effective growth rates r_i and r_j of the competitors, was constant. (See Section 1.2.1.)

3.2.2 Single Cell Microscopy – Image Analysis

In order to extract growth rates and cell genealogy, z-stack images were converted into a correlation image for each time point, correlation images were filtered to remove background debris, and analyzed using Schnitzcells [138] and an in-house Matlab script.

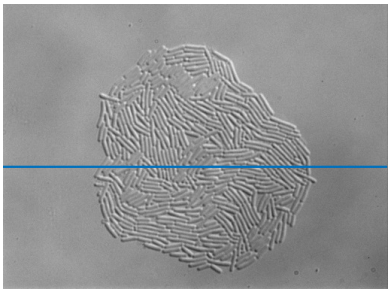
First, correlation images were calculated for a set of z-stack images taken in 250 nm steps over a range of 3 μm above and below the focal plane. This resulted in one final image for each time point, where the cells had a low grayscale intensity and were surrounded by a high grayscale intensity contour (Figure 3.2a,b). The correlation image was defined as the kernel weighted sum of the pixel intensities for each z-stack image, where the kernel was the first derivative of a Gaussian centered around the focal plane ($z_o = 0$) with variance $\sigma = 700\,\text{nm}$ (Equations 3.1 and 3.2).

$$I_c(z=0) = \int I(z')Ker(z-z', z_o = 0, \sigma = 700\,\text{nm})dz' \tag{3.1}$$

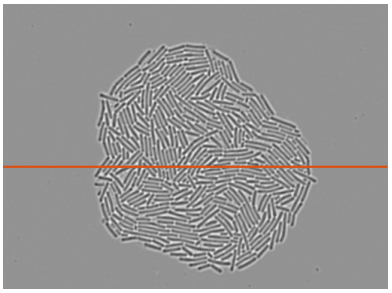
$$Ker(z-z') = \frac{d}{dx}\left(\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(z-z')^2}{2\sigma^2}}\right) = \frac{-(z-z')}{\sigma^3\sqrt{2\pi}}e^{\frac{-(z-z')^2}{2\sigma^2}} \tag{3.2}$$

In addition to uniformly sharpening the contrast around cells, correlation images had the added benefit of creating a constant and uniform background intensity. This is seen in Figure 3.2(d), where the relative grayscale intensities for the same slice of the focal plane (a) and correlation (b) image are shown. In comparison to the DIC intensities (dashed line, blue), the correlated intensities (solid line, orange) have a constant background and clear peaks and valleys for cell boundaries and interiors, respectively.

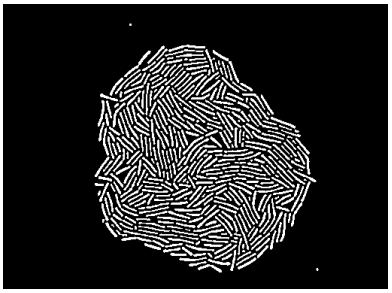
Next, a threshold was applied to all correlation images of a given time lapse using a script developed in conjunction with Stephen Anthony [182]. The script makes use of Otsu's thresholding method [183], in particular, four-thresholding. The interior of the cells corresponded to values below the second threshold. Objects below the second threshold were further filtered by removing noise (objects less than five pixels in size) and cells not fulling contained within the image (Figure 3.2(c)).



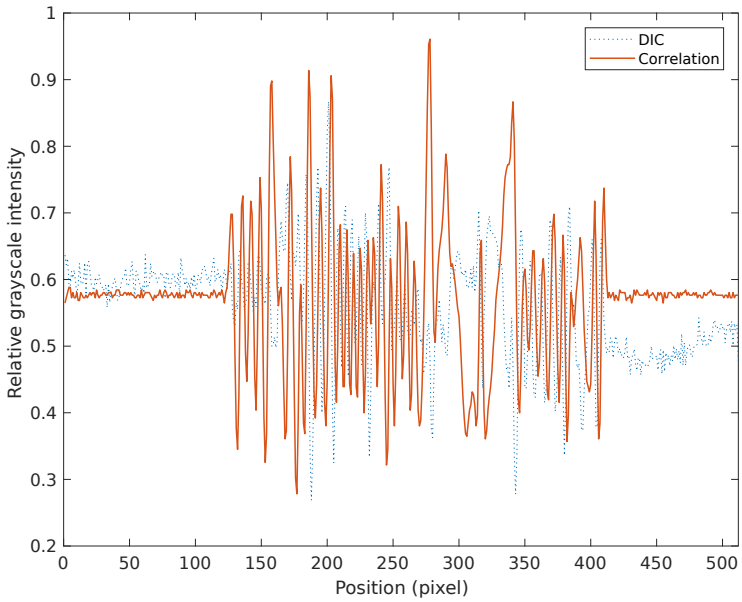
(a) Focal plane DIC image



(b) Correlation image



(c) Post-threshold binary image



(d) Relative intensities of slices shown above in (a) and (b)

Figure 3.2: Image processing of z-stack DIC images. (a) Raw DIC image in the focal plane. (b) Correlation image for the image z-stack corresponding to (a). (c) Binary image of putative cells after Otsu thresholding. (d) Relative intensities of the slices shown in (a) and (b) (dashed line, blue and solid line, orange – respectively).

After applying the threshold and filtering, images were imported into Schnitzcells [138] for tracking, and lineages were analyzed using an in-house Matlab script. Genealogy trees were drawn for each image set and generation time for every dividing cell was calculated, taking note of when a cell had become competent prior to division. Competence was monitored using BD2711. Cells which did not divide during the time frame of the experiment were not included in the mean generation time calculations. As such cells included those with particularly long generation times—i.e., competent cells, the mean generation time of cells that run through a competence period was underestimated.

Results – Evolution Experiment

"Some figures are bigger than others; some figures are bigger than other figure's mothers."

—The Smiths

We aimed to characterize the effect of competence on genome dynamics and interactions between different subspecies of *B. subtilis*. This first chapter focuses on the findings from the evolution experiment with *B. subtilis* 168 (Bsu168) clones which periodically received *B. subtilis* W23 DNA for ~210 generations. In this experiment, the genomic effects of the evolved cultures were measured via whole genome sequencing (WGS), and cataloged by a variety of analysis pipelines and scripts – with the goal of identifying orthologous recombination events, detecting de novo variants, and predicting the effectiveness of interspecies gene transfer.

First, orthologous recombination events are described in full—for the three time points over which all replicates were sequenced and the four replicates sequenced at every second cycle. Next, the recombination events are characterized based on identity, composition, and function. Afterwards, statistical analyses are done to determine if there was any bias in recombination events to specific genes, identity, or partial gene or operon replacement. Finally, de novo variants, occurring in addition to the recombination events, are characterized.

Throughout this chapter, standard deviation was used to report measurement error ($\mu \pm \sigma$), unless otherwise noted.

4.1 Orthologous Recombination and De novo Insertions Occur in Multitude

The evolution experiment consisted of 21 two-day cycles. Each cycle consisted of six steps: dilution, radiation, plating, colony selection and regrowth, competence induction and addition of extracellular DNA, and washing and overnight growth (Section 2.1.3). Replicates received either no DNA, DNA from Bsu168, or DNA from BsuW23. For simplicity, replicates receiving BsuW23 donor DNA are referred to as "BsuW23 replicates", often shortened to W#, when talking about specific replicate #.

WGS was carried out for all eight ancestral clones and the evolved strains from various time points and experimental conditions. All eight replicates evolved with BsuW23 DNA were sequenced at cycles 9, 15, and 21 (cy9, 15, 21). Controls with and without self DNA were sequenced at cy9 (four replicates each) and cy15 (two replicates each). Time course sequencing (every second cycle) was carried out for replicates W1, W3, W4, and W5. Replicate W3 was discarded after cycle 9 and replicate W7 was not considered in that cycle, due to contamination (detected as outlined in Section 2.2.6). Replicate W1 was not sequenced at cycles 17 or 19.

All control samples, receiving either self or no DNA showed no false-positive orthologous recombination events. The total number of SNPs and INDELs for self and no DNA replicates at cycle 9 were on the order of 10^1, while that of BsuW23 samples was 10^4. Further information regarding the de novo mutations in both self and no DNA replicates, along with BsuW23 samples, can be found in Section 4.4.

The CNP (cluster of nucleotide polymorphisms) algorithm outlined in Section 2.2.2 was used to identify orthologous recombination events in the evolved samples. The novel genes algorithm from Section 2.2.5 was used to identify transformed auxiliary BsuW23 genes in the evolved samples receiving BsuW23 DNA.

4.1.1 Cycles 9, 15, and 21

Orthologous recombination events were detected in all BsuW23 replicates for cycles 9, 15, and 21. The mean and median import lengths (CNP lengths) decreased slightly throughout the experiment, along with the variation in those values (Table 4.1).

Table 4.1: Cycles 9, 15, and 21 CNP segment statistics

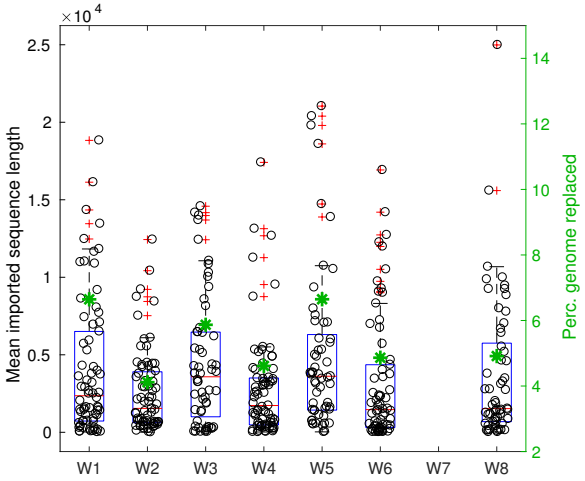
	Mean import length (bp)	Median import length (bp)	Replaced genome	Exponential constant (bp^{-1})
Cycle 9	3700 ± 900	2260 ± 960	5.3 ± 1.0%	3660 ± 170
Cycle 15	3500 ± 630	1960 ± 600	6.9 ± 2.3%	3480 ± 140
Cycle 21	3400 ± 560	1900 ± 390	9.8 ± 2.6%	3500 ± 120

A wide distribution of CNP lengths were found at cycles 9, 15, and 21 (Figure 4.1). Mean and median CNP lengths remained relatively constant, and percent genome replacement varied greatly between replicates at the same cycle – with variance increasing at later cycles. The mean percentage of replaced genome increased in a linear fashion (Table 4.1) and a linear regression calculated a rate of 0.47 % genome replacement, per cycle (Figure 4.2).

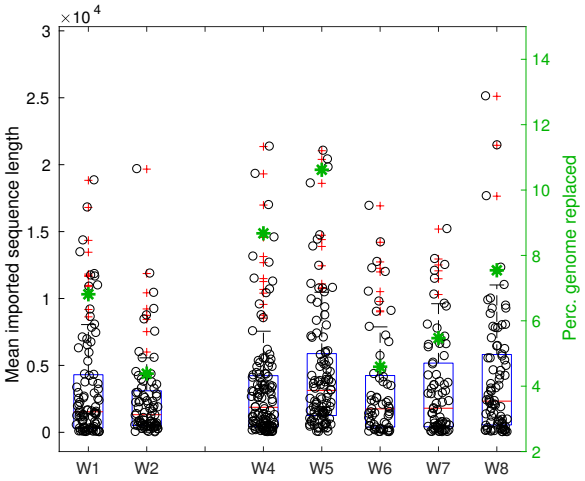
The average mean and median import lengths were more clearly seen in probability distributions (Figure 4.3). Due to the breath of import lengths, the distributions were shown on a log scale. The distributions strongly resemble an exponential decay and have similar decay constants of 3660, 3480, and 3500 bp^{-1} for cycles 9, 15 and 21, respectively (Table 4.1).

Inserts from auxiliary regions of BsuW23 were also detected in the evolved strains but at a much lower frequency than the homologous regions, 5 ± 6 insertions per replicate at cycle 21 (Figure 4.4). The mean insert length grew from 1480 to 1710 to 2240 ± 2900 bp for cycles 9, 15, and 21, respectively. Too few events had occurred by cycle 21 to determine if the length distribution fit an exponential decay function.

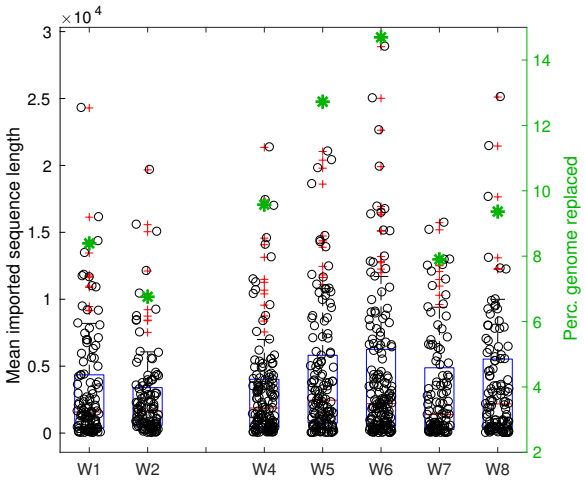
Import events had a mosaic pattern for all replicates over all cycles (Figure 4.5). It was evident that each replicate took a different evolutionary path over the 21 cycles, as the pattern of import events were not identical. There were some similarities between replicates, which are further examined in Section 4.3. (Mosaic import patterns for cycles 9 and 15 can be found in the appendix, Figures A.1 and A.2.)



(a) Cycle 9



(b) Cycle 15



(c) Cycle 21

Figure 4.1: Orthologous recombination for all replicates from cycle 9 (a), 15 (b), and 21 (c). (left axis) Box plot of all CNPs for each sample. Black circles represent the mean length of individual import events. (right axis) Green stars denote the percentage of the ancestral genome that has been replaced by BsuW23 segments.

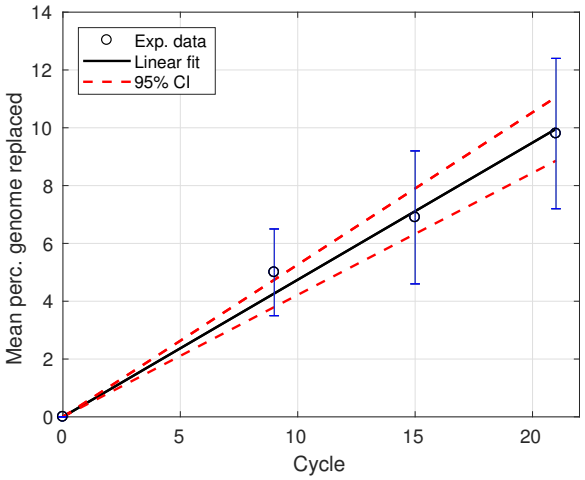
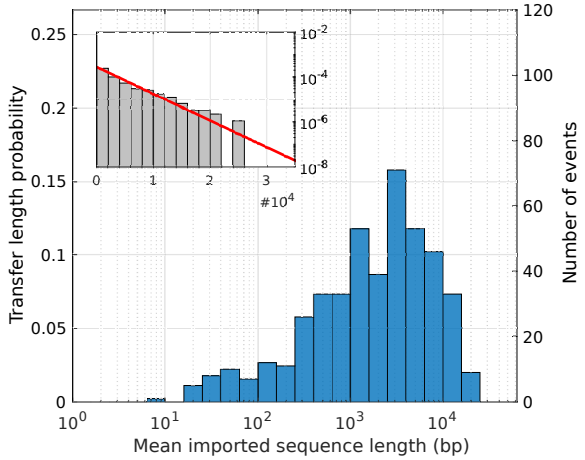
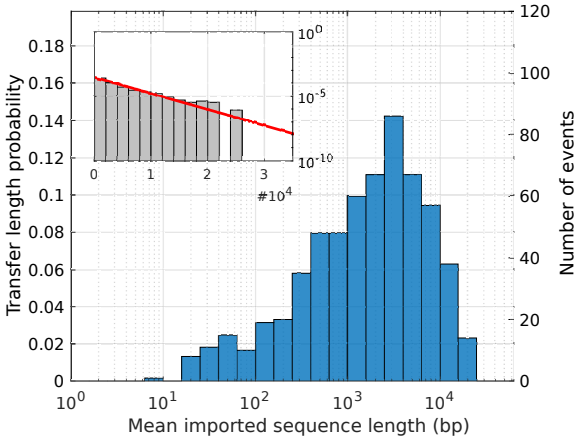


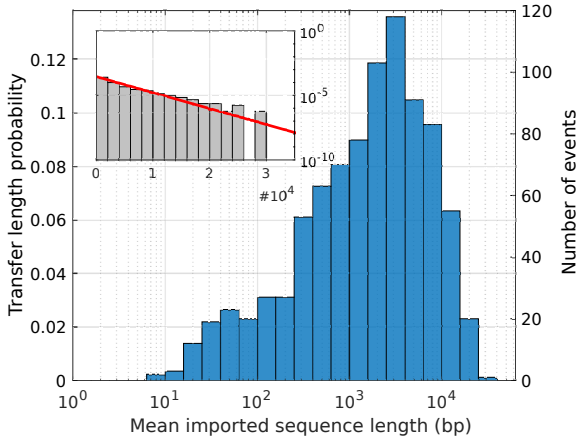
Figure 4.2: Rate of genome replacement. (circles, black) Mean genome replacement for seven BsuW23 replicates. Error bars are the standard deviation. (solid line, black) Best fit linear regression with $y = 0.47x$. (dashed lines, red) 95% confidence interval.



(a) Cycle 9

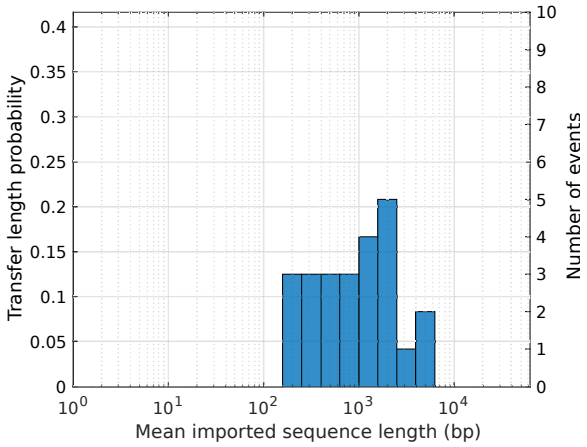


(b) Cycle 15

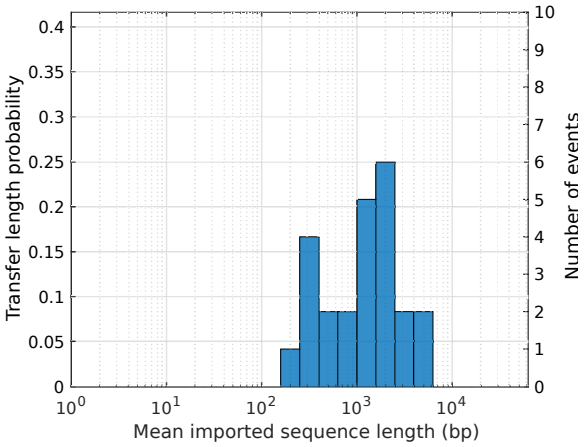


(c) Cycle 21

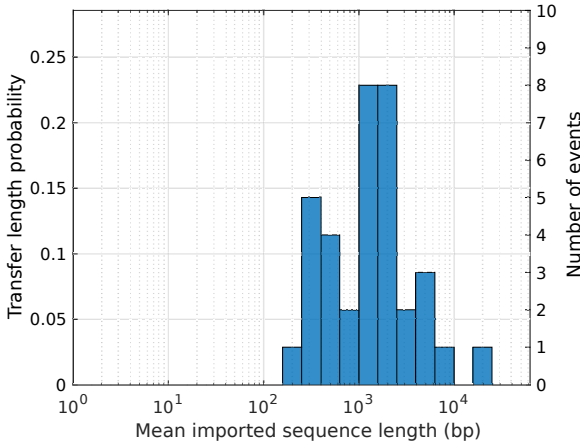
Figure 4.3: Mean CNP import lengths for cycle 9 (a), 15 (b), and 21 (c), plotted logarithmically. The probability of a specific transfer event occurring is given on the left axis, while the total number of events (over all replicates) is given on the right axis. Insets show the same distribution with length on a linear scale and number of events plotted logarithmically. (inset, red line) Exponential decay $y = e^{-\alpha x}$ (red line) with α = 3660, 3480, and 3500 bp^{-1} for cycles 9, 15 and 21, respectively. The probability of a specific transfer event occurring is given on the left axis, while the total number of events (over all replicates) is given on the right axis.



(a) Cycle 9



(b) Cycle 15



(c) Cycle 21

Figure 4.4: Mean lengths of imported segments from auxiliary regions for cycle 9 (a), 15 (b), and 21 (c), plotted logarithmically. The probability of a specific transfer event occurring is given on the left axis, while the total number of events (over all replicates) is given on the right axis.

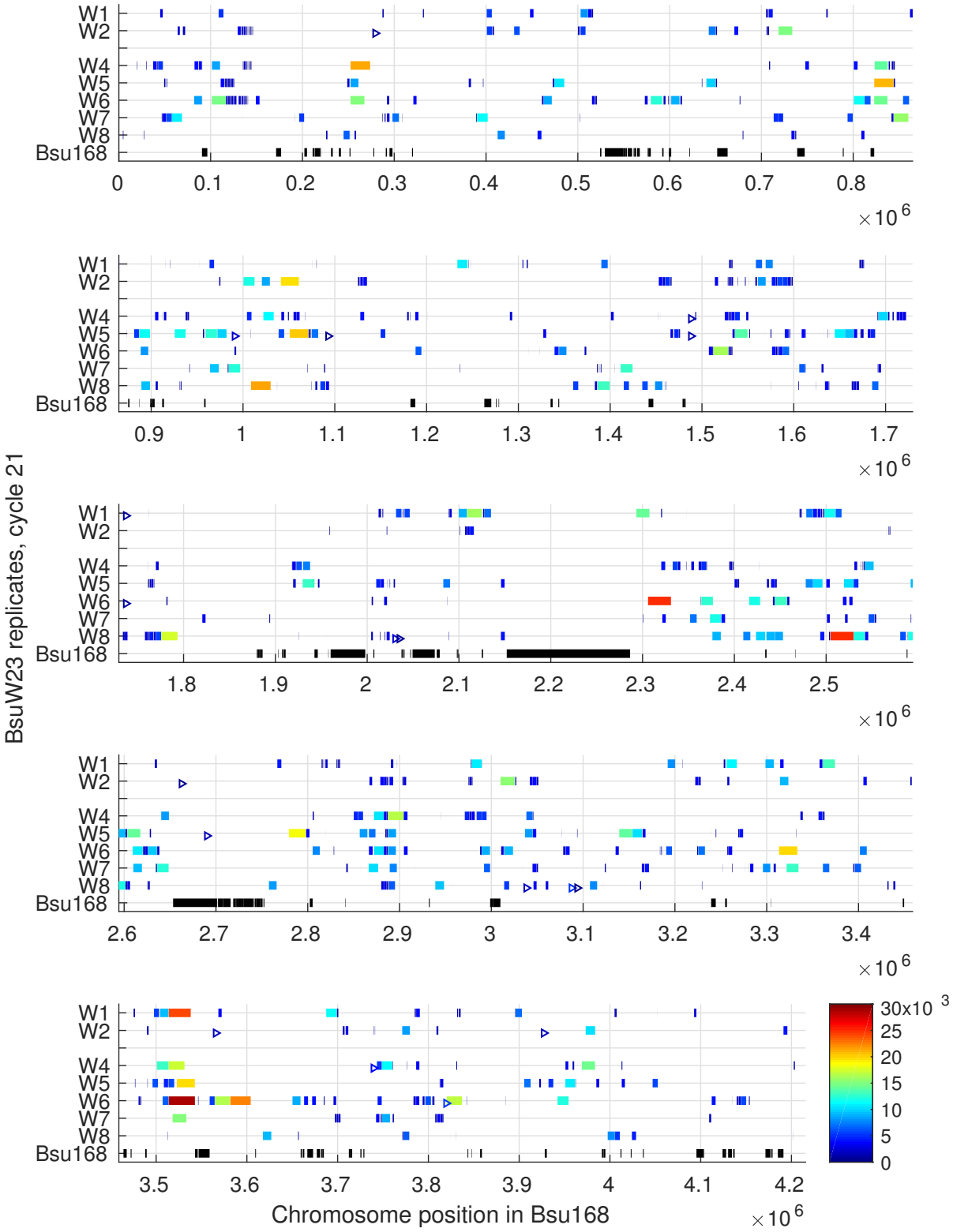
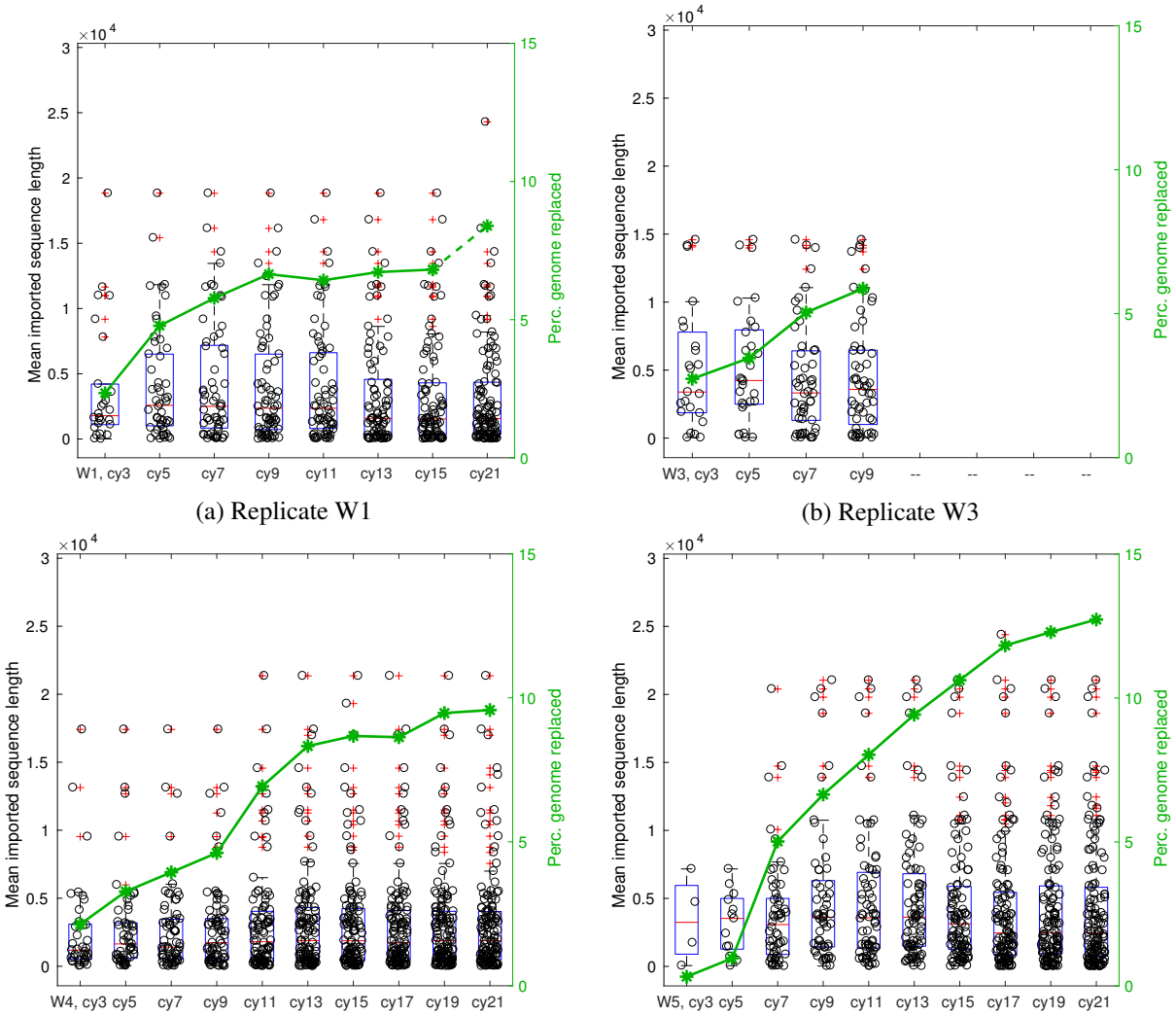


Figure 4.5: Orthologous recombination for various replicates from cycle 21, as a function of chromosome position. The start and end position of an orthologous recombination event (CNP) is denoted using filled boxes. The start of a de novo insertion is marked with an open triangle. All events are color coded to describe the average import length. Sample "Bsu168" denotes Bsu168 auxiliary regions (black).

4.1.2 Time Lapse Replicates, W1, 3, 4, and 5

Replicates W1, 3, 4, and 5 were sequenced at odd cycles, beginning with cycle 3.[1] The mean and median import lengths stayed relatively constant, although different for each replicate, as the number of import events increased (Table A.1). The percentage of replaced genome increased over all cycles, for all four replicates (Figure 4.6). This implied that the genes responsible for competence were still functional.



(a) Replicate W1

(b) Replicate W3

(c) Replicate W4

(d) Replicate W5

Figure 4.6: Orthologous recombination over multiple time points for replicates W1, 3, 4, and 5. (left axis) Box plot of all CNPs for each sample. Black circles represent the mean length of individual import events. (right axis) Green stars denote the percentage of the ancestral genome that has been replaced by BsuW23 segments.

For replicate W5, we see the construction of a mosaic import pattern over time (Figure 4.7, replicates W1, 3, and 4 can be found in the appendix, Figures A.3-A.5). All four replicates had initial import events that grew in length over all cycles ~20% of the time. Such recombinations added onto a preexisting

[1] As explained at the start of Chapter 4, replicate W3 was removed after cycle 9 due to contamination. Replicate W1 was not sequenced at cycles 17 or 19.

segment, thereby changing the length of the original segment from the previous time point—e.g., Figure 4.7, position 0.95×10^6. Occasionally segments were imported next to existing imports—e.g., Figure 4.7, position 1.65×10^6. Most frequently, ~80% of all import events, segments remained the same size. In < 5% of all import events did segments decrease in length.

The majority of import events did not increase in size after recombination. There was little evidence that higher segment identity increased the probability of orthologous recombination in an affected region. More on this and segment identity is discussed in Section 4.3. Between several time points it appears as if acquired CNPs were lost—e.g., Figure 4.7, cycle 7, position 1.21×10^6 (light blue). This was an artifact of the experimental setup, where cultures were frozen after being transformed and growing overnight. The frozen cultures have bacteria which differ amongst themselves in their most recently integrated segments. The clone, that survived the bottleneck and made it to the next cycle of the experiment, was not necessarily the clone that was sent for sequencing.

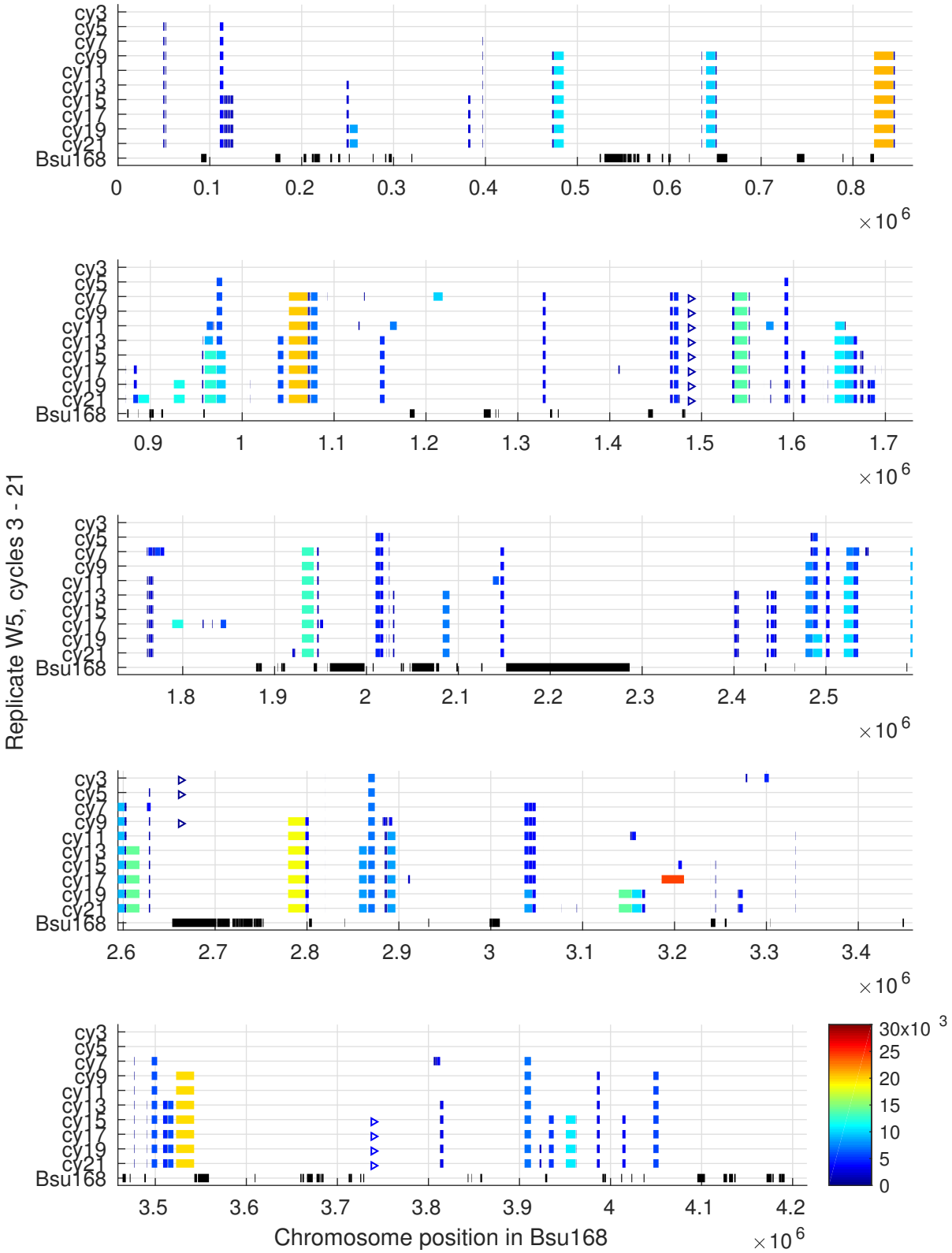


Figure 4.7: Orthologous recombination for replicate W5 from various cycles as a function of chromosome position. The start and end position of an orthologous recombination event (CNP) is denoted using filled boxes. The start of a de novo insertion is marked with an open triangle. All events are color coded to describe the average import length. Sample "Bsu168" denotes Bsu168 auxiliary regions (black).

4.2 CNP Properties: Identity, Composition, and Gene Function

To look for biases in recombination, we calculated the identity of each CNP. We looked at the gene and operon composition of CNPs to see if either of the two were preferentially fully-replaced. Finally, we looked at the individual CNP gene-functions for potential favoritism to replace a particular class of gene.

4.2.1 CNPs Recombine on Sections of Higher Identity

The average sequence identity between donor and recipient for each orthologous recombination event was measured as a function of the integrated segment's length (Figure 4.8). For all cycle 9, 15, and 21 replicates, CNP identity approached the subspecies average-identity (92.4%, between Bsu168 and BsuW23) as segment length increased. Short CNPs had a larger variance in their identity values than long CNPs, as one would expect for evenly distributed mℓSNPs following an exponential-like distribution (Section 2.2.2). The majority of the segments had an identity above the subspecies average-identity, and segment identity approached the subspecies average-identity at larger segment lengths.

The fraction of CNPs with lengths $\leq$ 100 bp increased slightly from 12.7 to 12.9 to 14.5% in cycles 9, 15, and 21, respectively. At the same time, the mean identity of those short CNPs dramatically decreased from 94.0% in cycle 9 to 77.0% in cycle 21 – notably below the subspecies average-identity. Because the identity of short CNPs varied so greatly, they were handled separately from CNPs > 100 bp, for all cycles. The average identity was found to be greater than the subspecies average-identity for all cycles, with $p < 0.001$ using a one sample t-test for the mean. The distributions were assumed to be unimodal and symmetric, as our sample sizes were larger than 50 and the data was not extremely skewed [184]. For CNPs $\leq$ 100 bp, cycles 9 and 15 still showed that the identities were larger than the subspecies average, with $p < 0.001$. Small CNPs at cycle 21 did not have an identity greater than the subspecies average and had 2 – 3 times the variance compared to cycles 9 and 15. (Table 4.2)

The sequence identities calculated here were cross checked against potentially missing mℓSNPs, allowed due to the 30% threshold outlined in Section 2.2.2. While missing SNPs did exist in the detected CNPs (below the aforementioned threshold), they did not change the average CNP identities nor uncertainties for small or long CNPs.

Table 4.2: Mean identities and significance values for cycles 9, 15, and 21, averaged over seven replicates receiving BsuW23 DNA

	Segment length	Mean identity	Significance level
Cycle 9	≤100 bp	94.0 ± 2.5	<0.001
	>100 bp	93.4 ± 0.6	<0.001
Cycle 15	≤100 bp	93.5 ± 1.5	<0.001
	>100 bp	93.9 ± 0.7	<0.001
Cycle 21	≤100 bp	77.0 ± 6.0	–
	>100 bp	93.6 ± 0.5	<0.001

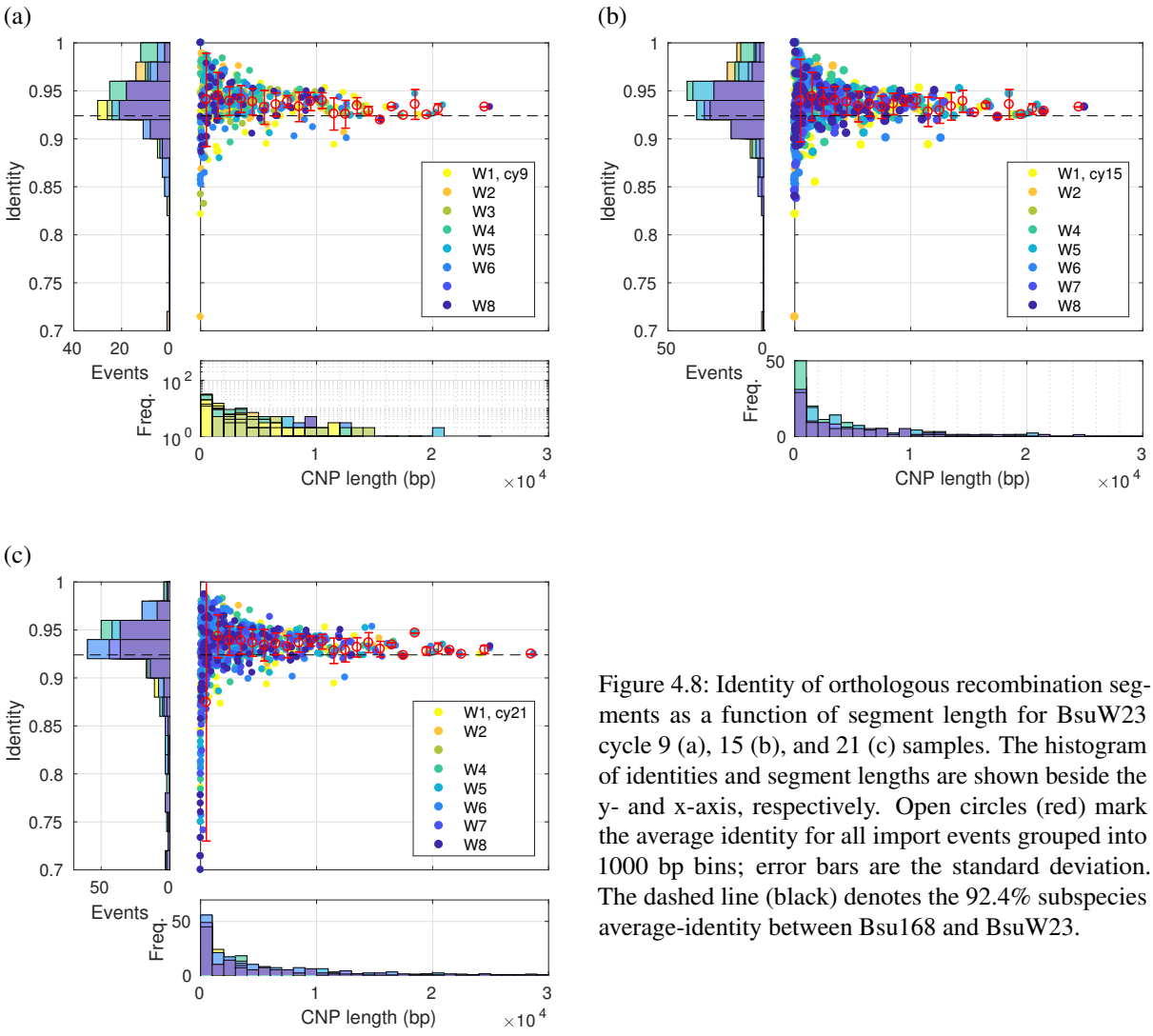


Figure 4.8: Identity of orthologous recombination segments as a function of segment length for BsuW23 cycle 9 (a), 15 (b), and 21 (c) samples. The histogram of identities and segment lengths are shown beside the y- and x-axis, respectively. Open circles (red) mark the average identity for all import events grouped into 1000 bp bins; error bars are the standard deviation. The dashed line (black) denotes the 92.4% subspecies average-identity between Bsu168 and BsuW23.

4.2.2 CNP Composition

No Bias towards Purine or Pyrimidine Enrichment

In addition to identity, the purine/pyrimidine enrichment of each CNP was calculated. We calculated difference between the number of additional purines or pyrimidines in a CNP and normalized that value by CNP length (Figure 4.9). Positive values denoted pyrimidine enrichment and negative values purine enrichment. At cycle 21, replicates had small changes in normalized enrichment values, centered around zero. The average ratio of pyrimidine enriched CNPs to purine enriched CNPs, over all seven replicates, were comparable: 1.02 ± 0.15 , with an average of ~110 non-zero enrichment values per replicate. We did not conclude that there was a bias towards purine or pyrimidine enrichment due to BsuW23 donor DNA.

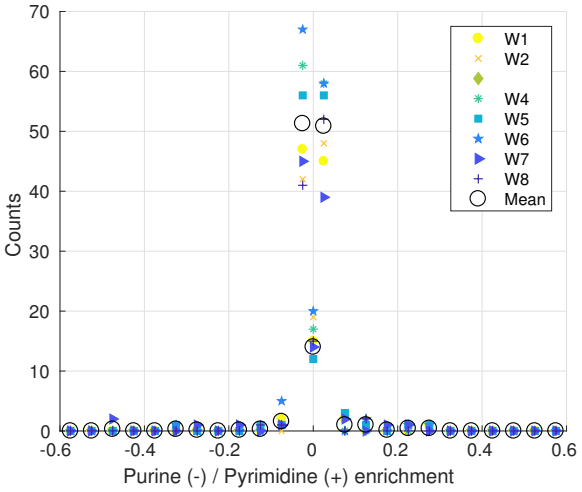


Figure 4.9: Relative purine/pyrimidine enrichment in CNPs from BsuW23 replicates, cycle 21. Histogram counts of the normalized net change in purine or pyrimidine enrichment. The difference in number of pyrimidine and purine changes for each CNP was calculated. The net change then was normalized by CNP length. Negative values denote a normalized net change towards purine enrichment, positive values towards pyrimidine enrichment. Bins were spaced at interval of 0.05, excluding zero, which is plotted separately. (open circles, black) The mean of all seven replicates.

Majority of CNPs Replace Full Genes and Operons

We hypothesized that replacement of partial genes or operons was selected against due to maladaptation of hybrid genes and operons. If hybrid genes and operons conferred a fitness cost, we would expect subsequent DNA imports to tend to complete the gene or operon replacement. Therefore, we investigated to what extent percentage, genes or operons were affected by a CNP replacement.

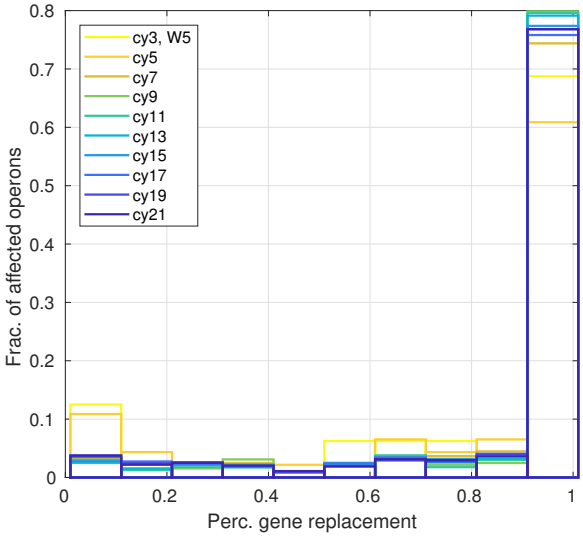
The annotated genome file for Bsu168 was obtained from the National Center for Biotechnology Information (NCBI, Reference Sequence: NC_000964 [185]) and the list of Bsu168 operons was taken from [186]. Percent replacement was calculated for all genes in W5 replicates, all time points. The median gene replacement percentage was 100% and the mean averaged 86 ± 4 %, and the fraction of completely replaced genes hovered between 60 – 80% (Figure 4.10(a,c)).

The majority of the genes that were replaced remained the same percent replaced throughout all 21 cycles. About 20% of CNPs grew in size, thereby replacing more of a given gene, and < 5% of CNPs shrank in size, thereby replacing less of given gene (Figure 4.11(a)). CNPs replaced complete genes in the majority of recombination events. A small fraction of all CNPs grew in length—i.e., built upon

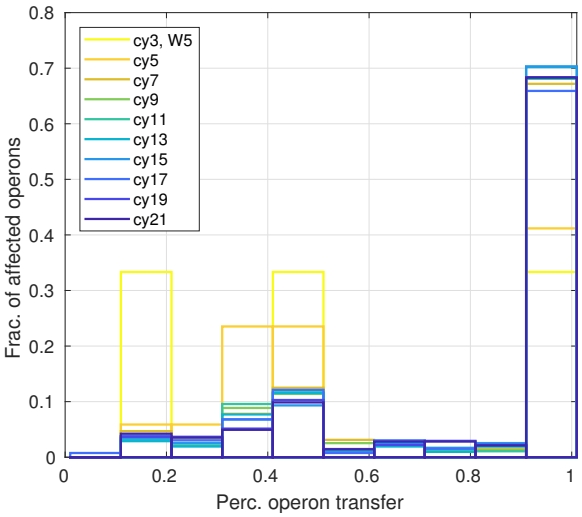
existing CNPs from a previous cycle to yield a larger CNP at the same location.

Results at the operon level were similar to those at the gene level (Figure 4.10(b,d), 4.11). Median percent replacement was 50% until cycle 7, after which the median replacement remained 100% through cycle 21. The mean percent replacement averaged 78 ± 10 %, with that value jumping dramatically from 56% to 81% by cycle 7 and then slowing climbing to 83% by cycle 21. Again, the majority of operons were completely replaced, but there was a second population of operons that were half replaced. Of those operons that were replaced, 19% of them grew and 8% shrank in percent replacement over 21 cycles. In this analysis, operons with only one gene were excluded and all operons were analyzed at the gene level—i.e., if a gene was affected or not, regardless of fully or partially. The latter simplification slightly over estimated the percentage of operon replacement, as the partially replaced genes remained around ~15%, over all cycles.

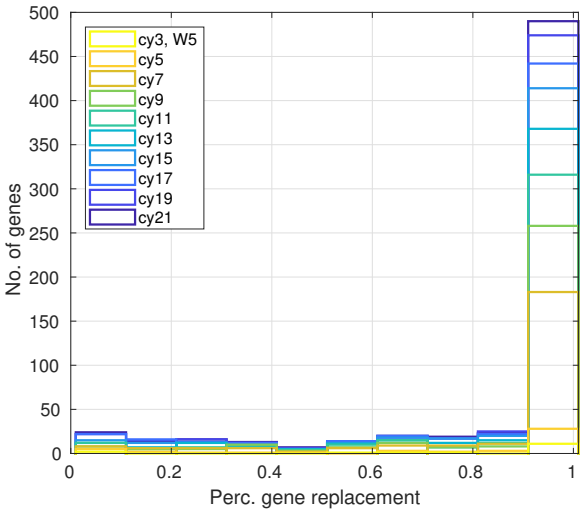
Looking at all replicates from cycle 21, individual genes were fully replaced twice as often as partially (337 ± 115 and 160 ± 22 , respectively). Operons, similarly, were replaced fully (168 ± 53) more often than partially (90 ± 10). The average operon size was 3.2 kbp, comparable to the average CNP length. This could partially explain why we see favoritism towards full operon replacement.



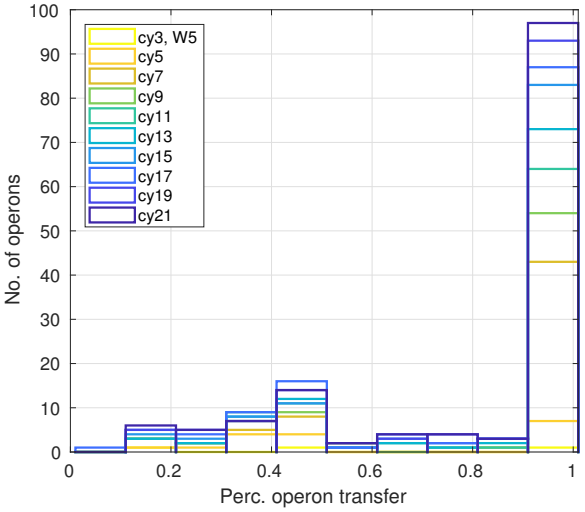
(a) Normalized gene replacement distribution for all cycles, W5



(b) Normalized operon replacement distribution for all cycles, W5

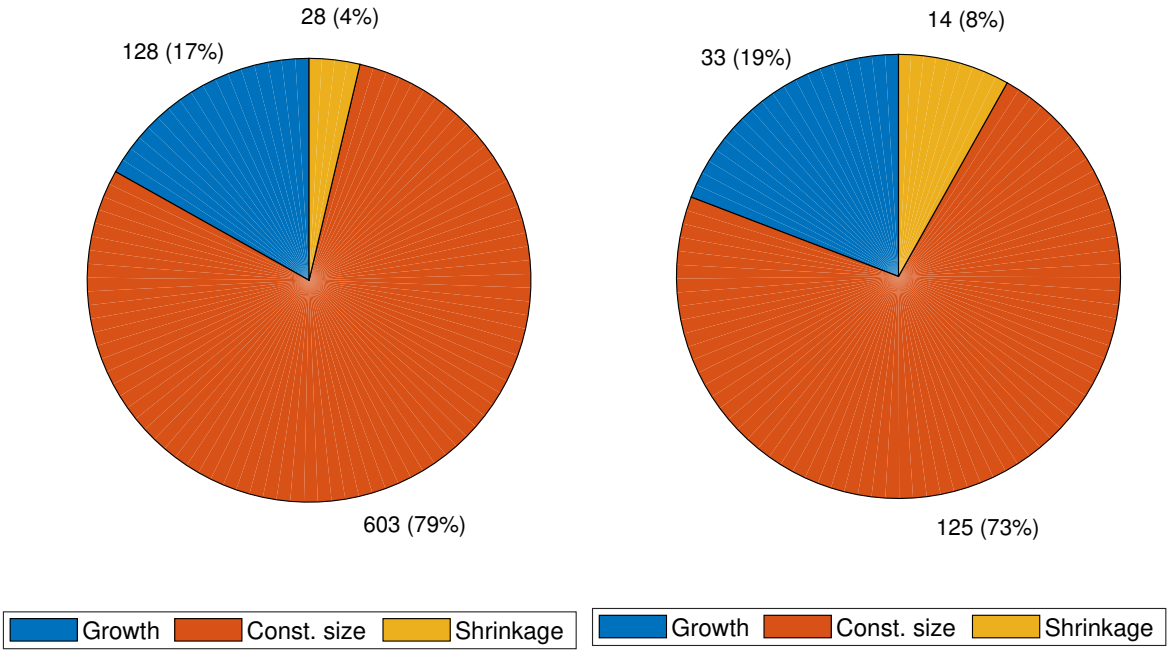


(c) Gene replacement distribution for all cycles, W5



(d) Operon replacement distribution for all cycles, W5

Figure 4.10: Gene and operon replacement distributions for W5, all time points. (a,b) Normalized histograms of gene (a) and operon (b) replacement. (c,d) Raw counts thereof. All operons were characterized at the gene level and single gene operons were omitted.



(a) Dynamics of replaced gene size, 21 cycles, W5

(b) Dynamics of replaced operon size, 21 cycles, W5

Figure 4.11: Change in CNP lengths over time at the gene and operon level, replicate W5. (a,b) Pie charts of gene (a) and operon (b) replacement over 21 cycles. Individual gene/operon replacements either grew (blue), remained the same size (orange), or shrank (yellow). The number of times each event occurred (relative percentage thereof) is listed next to each wedge. All operons were characterized at the gene level and single gene operons were omitted.

4.2.3 Essential Genes are Overrepresented and Prophages are Underrepresented in CNPs

Returning to the gene level, CNPs were analyzed to determine if there was a replacement preference for genes of a particular function. First, the oversweeping categories of essential and non-essential genes were applied to CNPs. Genes were categorized based on the SubtiWiki, University of Gottingen ([187], accessed 1 Mar. 2018) classifications. An affected gene was only counted once, even if multiple functions were indicated in the database.

The fraction of affected essential genes was $8.9 \pm 4.0\,\%$, averaged over all seven replicates from cycles 9, 15, and 21 (Figure A.6). Assuming that replacement was equally likely for all genes, the probability of replacing an essential gene was 3% more than expected. The deviation from the expectation value is significant at the $p < 0.01$ level only for cycle 21, using the one sample t-test for the mean. However, the average sequence identity of essential genes was also higher than the Bsu168/BsuW23 subspecies average, at 95.1%, in agreement with their high replacement probability.

Orthologously replaced genes were further classified using the extensive list of gene categories on the SubtiWiki. The "prophages and mobile genetic elements" class were underrepresented at cycles 9, 15, and 21 (Figure 4.12). This is most likely because the majority of these regions have no homolog in BsuW23—i.e., these are auxiliary regions. (More on recombination probability can be found in Section 4.3.) Gene classifications for cycles 9 and 15 can be found in the appendix, Figure A.7.

The names of all genes affected by CNPs at cycle 21 can be found in the appendix, Figures A.8-A.12.

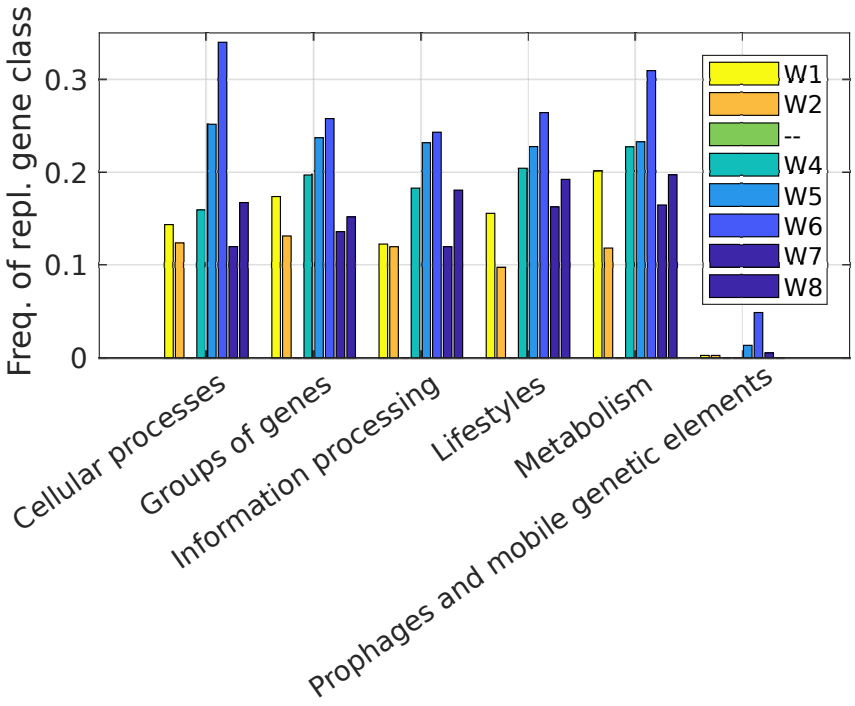


Figure 4.12: CNP gene types, cycle 21. Counts have been normalized by the number of times each gene type occurs in the genome.

4.3 CNP Recombination Probability

Having seen that the average CNP identity was higher than the average subspecies identity between Bsu168 and BsuW23, we further examined CNPs to determine if recombination was happening randomly, or if selection was responsible for the recombination patterns we had seen. Based on previous studies, we expected the segments to have a higher identity, namely because transformation rates are higher for smaller sequence divergences. We looked for evidence of recombination hot spots, calculated the role sequence plays in hypothetical recombination locations, and computed if recombination was biased to replace full genes. The models needed to determine the aforementioned parameters and the simulations necessary to support them were developed in collaboration with Fernanda Pinheiro. Pinheiro drew up the necessary models to compare our null hypotheses to the experimental results and performed the statistical analyses to determine if we could reject said null hypotheses.

4.3.1 Several Putative Hot/Cold Spots Detected in Overall Fairly-Random Gene Replacement

Evolved genomes were analyzed at the gene level to see how often genes were affected, whether fully or partially, by a CNP. We compared the number of times a gene was affected in 0, 1, 2, ..., 7 (all) BsuW23 replicates, for cycles 9, 15, and 21 (normalized to the total number events detected at each cycle), to a binomial distribution to see if gene replacement was random (Figure 4.13). Here, we assume that the probability of replacing a gene is equal for all genes. The rationale behind this assumption was that variation in gene sequence-identity showed only a few outliers (Figure 4.14). For the binomial distribution, the average genome replacement (Table 4.1) was used for the probability of success, p. The experimental data did not significantly differ from the binomial distribution and the null hypothesis was not rejected using the two-sample Kolmogorov-Smirnov test (KS2). We concluded that genes were replaced randomly over the whole genome.

Although our data (Figure 4.13) was consistent with random replacement, the histogram cannot exclude a small number of outliers, potentially selected genes. We looked to see if there were putative recombination hot spots. Again at the level of affected genes, we recorded the number of times a gene was affected across all replicates (Figure 4.14). Several genes were replaced in four or more replicates (Table 4.3), with no obvious relationship between more frequently replaced genes and their identity. *leu* genes were likely hot spots because the recipient strain is a *leu* auxotroph. *eps* genes produce extracellular polysaccharides which are crucial in biofilm formation. Biofilm formation may be advantageous for evolved strains because of their repeated time spent in the stationary phase. Putative recombination-cold-spots, regions with no recombination in any replicate, often corresponded to Bsu168 auxiliary regions. (Putative recombination-hot/cold-spots for cycle 15 can be found in the appendix, Figure A.13.)

To conclude, recombination occurred randomly over the genome. This did not mean that no putative selection was measured. Several genes, in particular *leuABC* and *epsABCDEF*, were replaced in five of the seven replicates, making it likely that they were selected for. Additionally, putative recombination-cold-spots included the auxiliary regions of Bsu168, likely because they have no homolog in BsuW23.

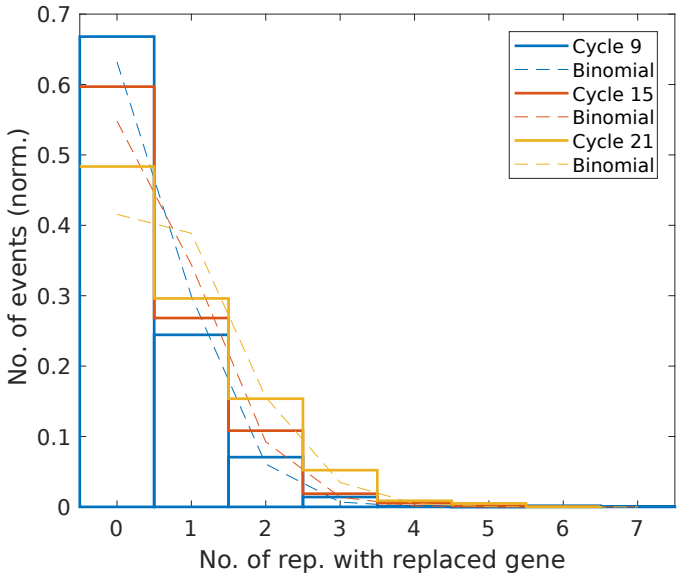


Figure 4.13: Histogram of affected genes did not differ significantly from a binomial distribution. Experimental data (hollow bars) show the number of times a gene was replaced, 0, 1, 2, ..., 7 times, in all replicates by cycle 9 (blue), cycle 15 (orange) and cycle 21 (yellow). Number of events for each bin was normalized to the total number of events over a given cycle. Binomial distributions (dashed lines, same color scheme), using p = average genome replacement.

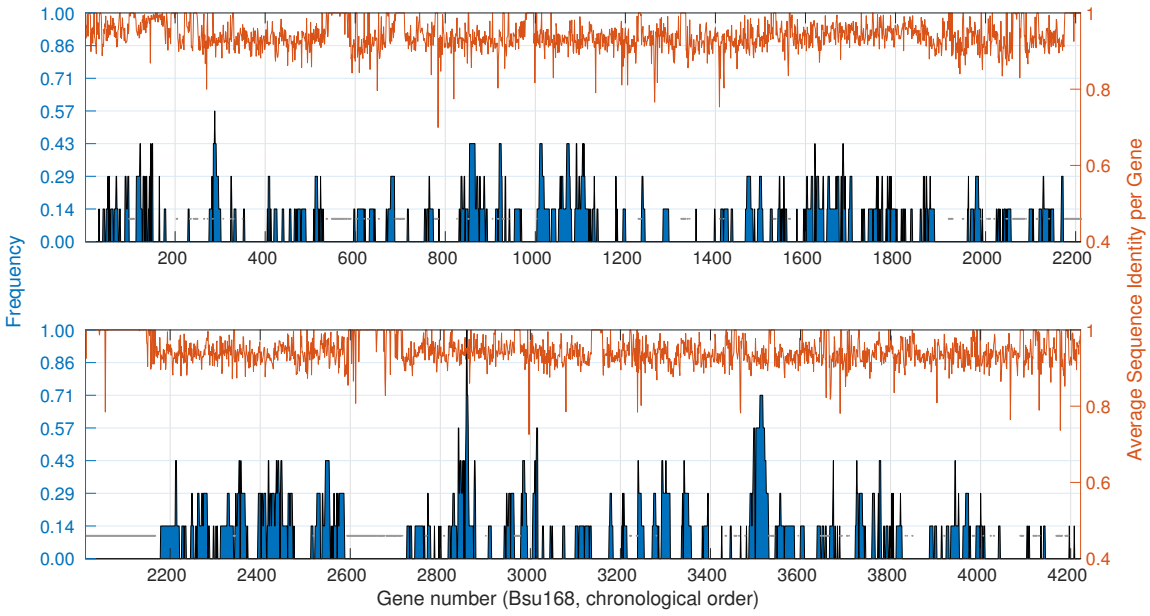


Figure 4.14: Probability of replacement of a specific gene by cycle 21. (blue, left axis) Genes affected by CNPs as a function of gene number and normalized to the number of replicates (seven). (orange, right axis) Average sequence identity per gene. (gray dots) Bsu168 auxiliary regions. (These regions are displayed as having an identity of one.)

Table 4.3: Genes affected in the majority of cycle 21 replicates.

No. replicates with affected gene	7	6	5	4
Genes	*leuB*	*leuC*	*epsABCDEF, ilvC, leuA*	*BSU misc RNA 44, epsGHIJKLMNO, gamA, ganR, ilvH, leuD, padC, pnbA, yszA, ytxGHJ, yveFG, yvfHI*
Gene number	3059	3058	3710 – 3716, 3061, 3060	3040, 3701 – 3709, 288, 3696, 3062, 3057, 3718, 3717, 3041, 3214 – 3216, 3719 – 3720, 3697 – 3698

4.3.2 CNPs Have Smaller SNP Density on One Side of the Segment

In Section 4.2, we reported that the identity of the CNPs had a significantly higher identity than the subspecies average between Bsu168 and BsuW23. The CNPs were further analyzed to determine if there was a higher than average sequence identity at one of the ends of the orthologously replaced segments. This would suggest that high sequence identity is most important at the start of a recombination event.

First, import events were simulated to determine what one would expect if there was no bias. For each of the 868 experimental recombination events, 500 in silico recombination simulations were performed.For each simulation, two starting positions from the master list were chosen at random. Then, the number of mℓSNPs on the forward strand in the first 100 bp was recorded, for the first starting position running in the sense direction and for the second starting position in the antisense direction. We refer to recombination on the forward strand, running in the sense direction, as the "forward direction", and running in the antisense direction as the "reverse direction". In layman's term, one can picture the genome as a number line running from left to right. "Forward" recombination occurred by picking a position and counting the number of mℓSNPs to the right, while "reverse" recombination counted to the left. The SNP density distributions for the forward and reverse directions lay on top of each other. This was expected as the mℓSNPs are evenly distributed across the genome (Section 2.2.2). A KS2 test confirmed that one could not reject the null hypothesis; these two samples came from the same distribution (Figure 4.15).

The same test was performed on the cycle 21 experimental data using the 5' to 3' and 3' to 5' ends on the forward strand as the convention for forward and reverse directions. Again, using a KS2 test the null hypothesis could not be rejected. When the two in silico distributions were compared to the experimental distributions, the experimental distributions were significantly different from both of the in

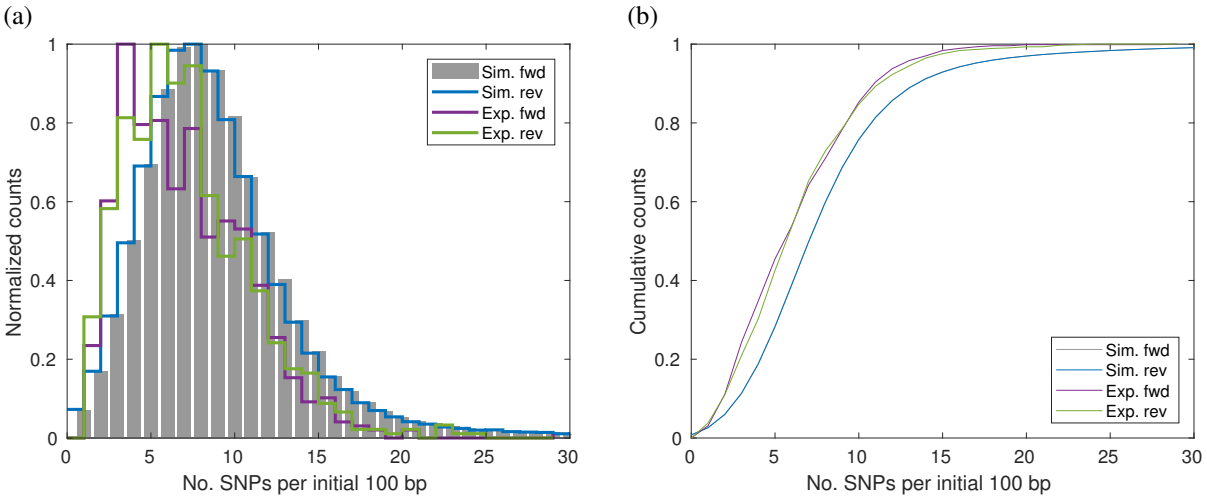


Figure 4.15: SNP density distribtion for the first 100 bp of the in silico and experimental CNPs. (a) Distributions are split into in silico (simulated) CNPs, forward (gray) and reverse (blue) direction, and experimental CNPs, forward (purple) and reverse (green) direction. Forward direction is sitting on the forward DNA strain and reading 5' to 3' direction. Reverse direction is sitting on the forward DNA strain and reading in the 3' to 5' direction. (b) Cumulative sum of distributions in (a), used to determine if the samples come from the same distribution.

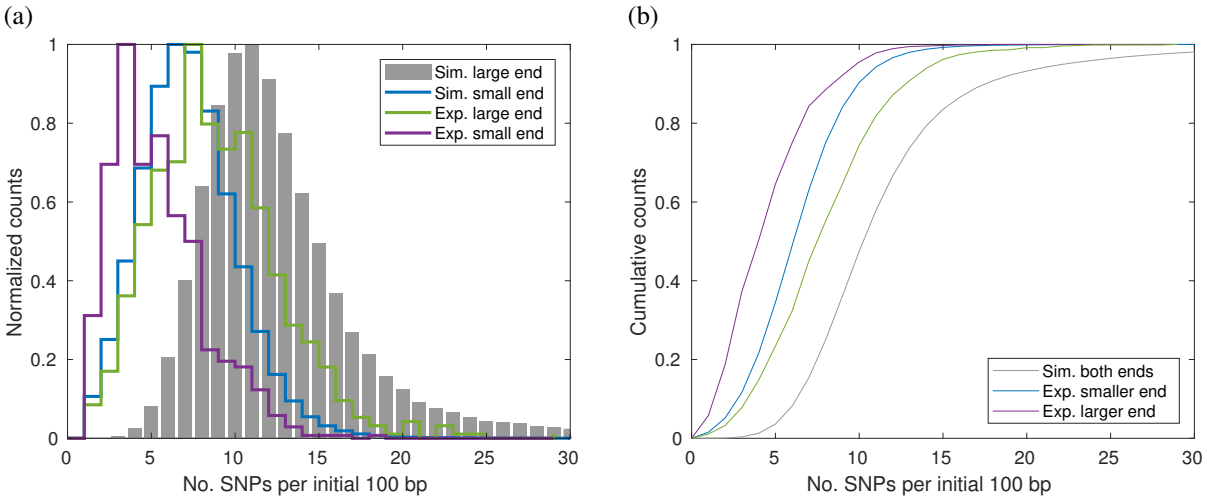


Figure 4.16: First 100 bp SNP density for CNPs and simulations. For each CNP, the ends of each segment (first and last 100 bp) are grouped into smaller (purple) or larger (green), depending on their SNP density with respect to one another. The same is done for the simulated recombinations, smaller (blue) and larger (gray). (a) Normalized distributions. (b) Cumulative sum of the distributions in (a). The null model that the smaller end distribution (blue) came from either simulated distribution (blue or gray) is ruled out at the $p << 0.01$ level.

silico distributions at the $p = 0.01$ level using the KS2 test. This was easily explained as the mean of the in silico and experimental distributions was different, reflective of the higher average identity for the experimental CNPs (Figure 4.15).

Next, SNP density distributions were sorted into smaller and larger end, for both the experimental CNPs and the in silico starting-position pairs. The experimental smaller-number distribution was significantly different from the experimental larger-number and simulated distributions at the $p << 0.01$ level using the KS2 test (Figure 4.16). Additionally, the experimental larger number distribution lay in between the smaller and larger simulated-distributions. In fact, the experimental larger-number distribution was not distinguishable from the non-sorted forward and reverse simulation distributions (Figure 4.15). In layman's terms, one side of each CNP had a higher identity; the lower identity side had a distribution equal to that theoretically predicted for the Bsu168 and BsuW23.

After establishing that one end of each CNP had a significantly lower SNP density, we calculated how far along the CNP one must travel to find the SNP density is no longer significantly different from the expected density (forward and reverse simulations). In other words, how long was the patch of higher identity on the one side of the CNP. First SNP density distributions were calculated for 100 bp segments from the end with the lower SNP density to a length of up to 10,000 bp, for all cycle 21 CNPs. Next, the SNP density distributions of the 100 bp sections (starting at the lower SNP density end) were compared to the distribution of the larger SNP density end, using a KS2 test. The lower SNP density end was used as the reference distribution because we had shown it was indistinguishable from the simulated forward and reverse distributions. The D statistic (the maximum distance or *supremum* between the cumulative density functions was significant at the α=0.05 level for the first 500 bp and from 500 – 2000 bp, significance hovered around the α=0.05 level (Figure 4.17).

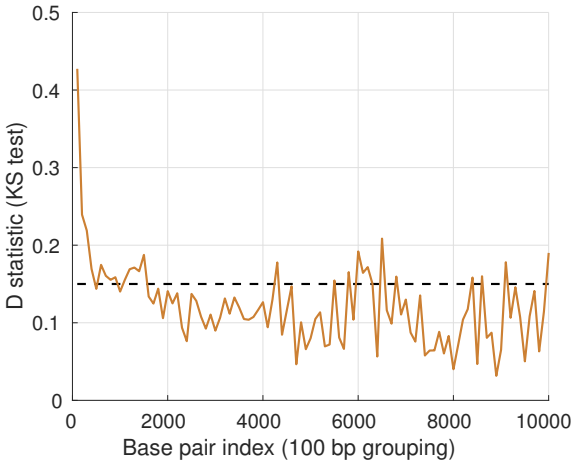


Figure 4.17: D statistic for identity distribution along a 100 bp moving window. Identity distributions were calculated for 100 bp segments in CNPs, starting at the end with the lower SNP density. This moving distribution was compared to the distribution of the larger-SNP-density end. Dashed line (black) marks the $p < 0.05$ significance level. Values below that line are cases where the distributions were not significantly different.

The smaller SNP density distribution was analyzed to see if the smaller end preferentially sat on the 5' or 3' end of each CNP (read off the forward strand). We detected no directional bias for recombination. It was equally likely, for each CNP, that the smaller end sat on the 5' or 3' end of a CNP or the 3' to 5' end (with respect to the forward strand).

Our data showed a higher than average sequence identity at one end of the CNP. The higher sequence identity remained significantly higher for the first 500 bp. It is reasonable to assume that recombination started at this end. We did not measure a bias for the end with a higher identity to sit on the 5' to 3' or 3' to 5' end on the segment, when read off the forward strand. This is consistent with no bias in the direction of recombination with respect to the origin of replication.

4.3.3 Genes are Most Likely Completely Replaced

The in silico CNPs from Section 4.3.2 were reanalyzed to see if genes were more often replaced completely than expected with a length distribution corresponding to Figure 4.3(c). Deviations from the expectations would indicate that replacement of full genes was selected for. We only used in silico CNPs which recombined in the forward direction for this analysis. As described in Section 4.3.2, in silico recombination occurring in the forward and reverse directions could not be distinguished from another, and so taking only the forward direction more closely emulated the CNP detection algorithm.

The number genes a CNP partially covered (0, 1, or 2[2]) in the experimental data was compared to the in silico recombinations. Experimental recombination events more favorably affected one partial gene than two (Figure 4.18(a)). As not all of the experimental CNPs replaced one gene partially as opposed to two, it is unclear how strong this bias, to replace fewer genes partially, is.

The number of fully covered genes did not differ between the experimental and in silico CNPs. This was reflective of the median CNP length being equal to ~2 genes (1900 bp).

[2] Cases with three partially covered genes, due to the possibility of overlapping genes where grouped into the two partially covered genes category. It occurs 572 times (13%) across the genome.

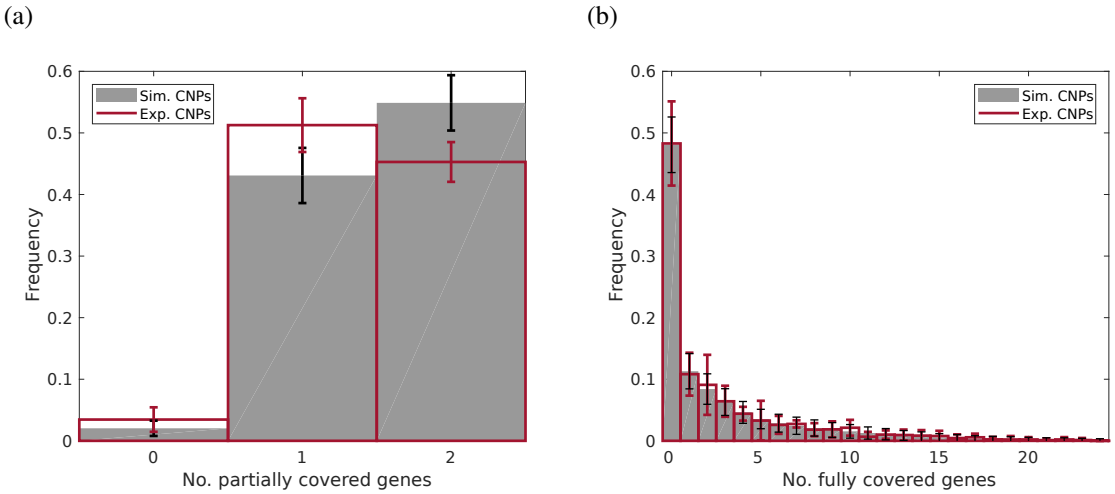


Figure 4.18: Frequency of partially and fully replaced genes from both in silico recombinations (simulations, gray) and experimental CNPs (red) at cycle 21. (a) Number of partially covered genes due to CNP orthologous recombination. (b) Number of genes fully covered by CNP orthologous recombination. Error bars are the standard error of the mean.

4.4 The Presence of Donor DNA Increases the Number of De novo Variants

Finally, we addressed the question whether transformation with DNA from different donors affected the frequency at which de novo variants occurred. We grouped non-CNP variants according to SNP/INDEL type to determine if recombination led to a particular type of SNP/INDEL being overrepresented. Deviations from the SNP/INDEL type distribution could imply that recombination machinery induced the variants. Affected gene types and genome position were analyzed to see if variants had a bias towards either. Such a bias could allude to compensatory mutations in response to recombined segments, either in a particular gene class or upstream of CNPs in promoter regions.

Non-CNP variants were grouped into one of eight SNP/INDEL types: synonymous mutations, non-synonymous mutations, stop codon mutations (including the creation of new stop codons), start codon mutations (again, including the creation of new start codons), missense insertions and deletions, in frame insertions and deletions, upstream mutations (occurring up to 3000 bp in front of a gene, outside of any coding region), and intragenic mutations (occurring outside of any coding region and more than 3000 bp upstream from any gene). The upstream- and intragenic-mutations categories include point mutations, insertions, and deletions. For both no DNA and Bsu168 DNA, the number and distribution of variants were similar. In total, both had replicates with ~35 variants by cycle 9 and ~60 by cycle 15, with greater variance amongst the no DNA replicates. For DNA replicates, 60 ± 20 % of all variants were synonymous or non-synonymous by cycle 15. By comparison, no DNA replicates had 71 ± 3 %. The fraction of upstream mutations remained constant for no DNA replicates at ~13% and doubled for Bsu168 DNA replicates from 7.5% to 15% (Figure 4.19(a,b)).

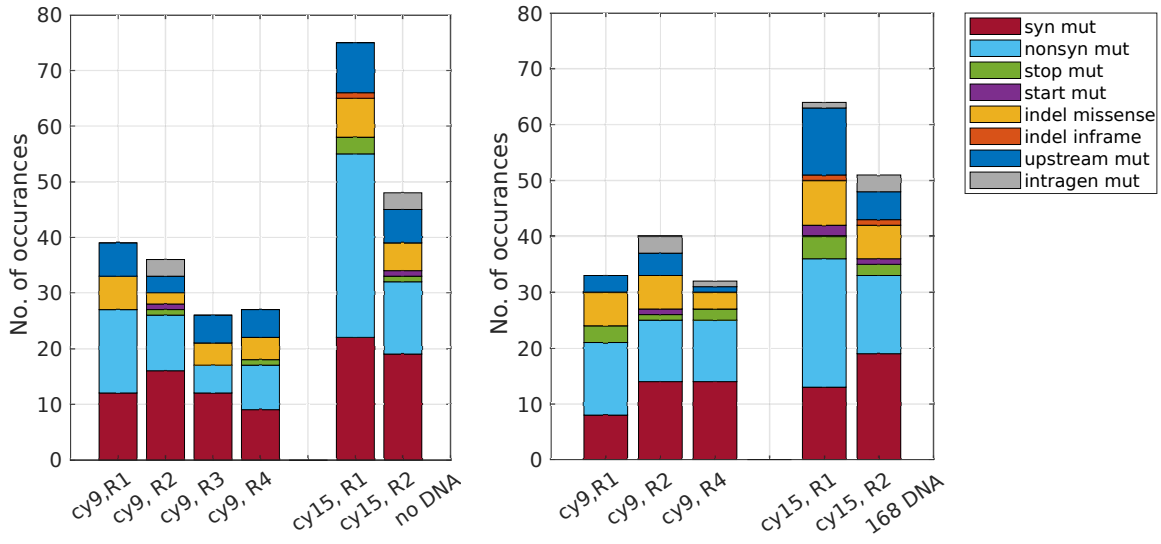
For the BsuW23 replicates, potential de novo variants were first compared to INDELs expected

from mapping BsuW23 onto Bsu168 (a "master list" of INDELs – ignored in the CNP algorithm, see Section 2.2.2) and variants with exact matches were removed.

BsuW23 replicates had twice as many variants as no DNA and Bsu168 DNA replicates by cycle 9 (66 ± 13 variants) and by cycle 21 another doubling with (115 ± 42 variants), however with larger variance (Figure 4.20). In contrast to no DNA and Bsu168 samples, replicates receiving BsuW23 DNA showed a large number of upstream mutations (consistently ~40% over all cycles). These upstream mutations were namely inserts, 57 ± 8 %, and deletions 31 ± 9 %. The number of synonymous and non-synonymous mutations averaged 10% over all cycles. (Variant types for the four time lapse replicates, W1, 3, 4, and 5, can be found in the appendix, Figure A.14.)

The majority of de novo variants were found inside of CNP regions, 58 ± 5 % by cycle 21 (Figure 4.21). In particular, 83 ± 7 % of all the missense and in frame indels, and 75 ± 5 % of all upstream mutations were within a CNP. The appearance of numerous de novo upstream mutations, specifically the large majority of those being within CNPs, hinted that these mutations could have been compensatory mutations.

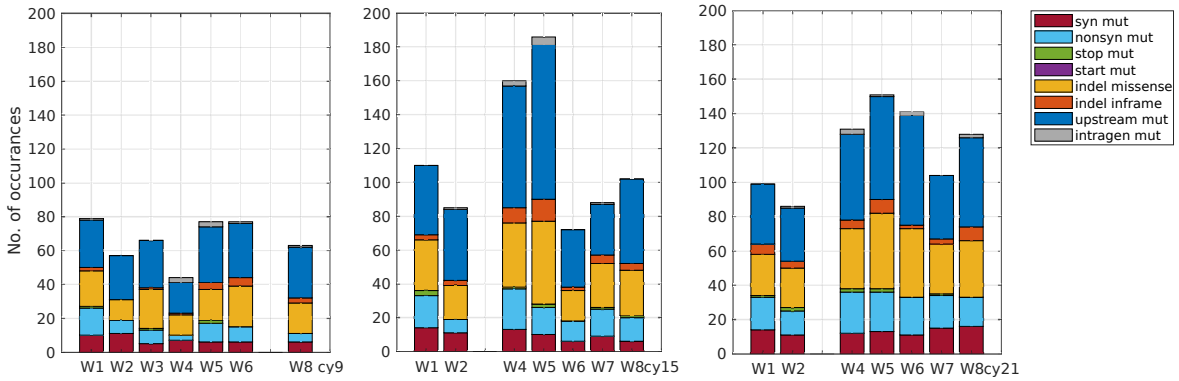
To better determine if the upstream mutations were compensatory mutations, how often de novo variants were found upstream of a CNP replaced gene (whether partially or fully) was calculated (Figure 4.22). Nearly all upstream mutations were found surrounding CNP genes. Of all the affected genes, 80 ± 5 % had no upstream mutation at any time point during the experiment. If an upstream mutation occurred, it occurred at the same time as CNP integration 10 ± 1 % of the time, after CNP integration 10 ± 5 %, and never before CNP integration. Both *same* and *after* events included events where the upstream mutation was later lost. (A representative graphic presentation of CNPs and upstream mutations can be found in the appendix, Figures A.15-A.18 (four parts).) When focusing only on variants



(a) No DNA

(b) Bsu168 DNA

Figure 4.19: De novo mutations for various replicates receiving (a) no DNA or (b) Bsu168 DNA, color coded according to annotation. Upstream mutations are ≤ 3000 bp upstream of a gene's protein coding region. Intragenic mutations are not in a protein coding gene and >3000 bp upstream of a protein coding gene.



(a) Cycle 9 (b) Cycle 15 (c) Cycle 21

Figure 4.20: De novo variants from cycle 9 (a), 15 (b), and 21 (c), BsuW23 replicates. Variants are color coded according to annotation. Upstream mutations are ≤ 3000 bp upstream of a gene's protein coding region. Intragenic mutations are not in a protein coding gene and >3000 bp upstream of a protein coding gene.

which occurred upstream of a CNP during or after CNP integration, we found it was equally likely to find an upstream mutation occur at the same time and CNP integration 52 ± 20 % as after 48 ± 20 %. We could not draw any conclusive evidence of the upstream mutations being compensatory mutations. Transcriptomics could be done using the frozen evolved strains to gain better insight into whether these variants were compensatory.

We concluded that the presence of foreign DNA lead to an increase in the number of de novo variants. The majority of de novo variants were missense indels or upstream mutation and slightly more than half of those variants were found inside of CNPs. It was unclear if these mutations were introduced during recombination or occurred afterwards. Additional studies looking at the transcriptomics of the evolved strains could clarify that open question.

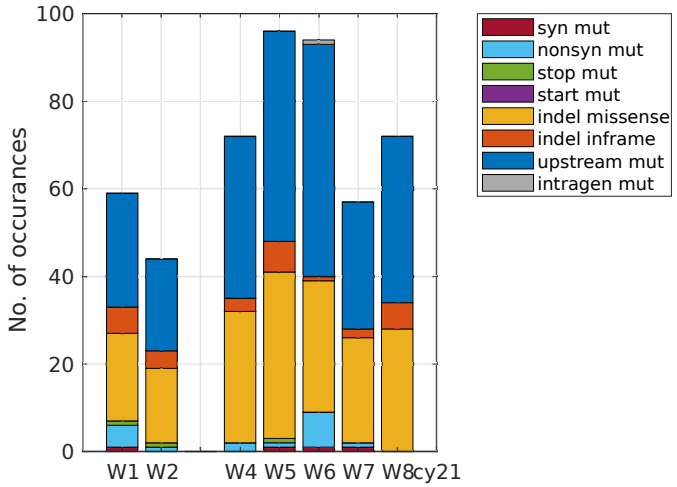


Figure 4.21: De novo variants within CNPs from cycle 21, color coded according to annotation. Upstream mutations are ≤ 3000 bp upstream of a gene's protein coding region. Intragenic mutations are not in a protein coding gene and >3000 bp upstream of a protein coding gene.

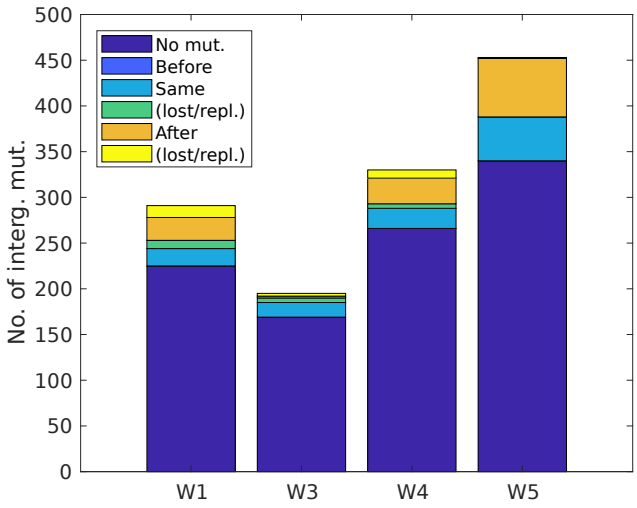


Figure 4.22: Timing of intergenic mutations upstream of CNP affected genes. (left) Timing of all upstream intergenic mutations for replicates W1, 3, 4, and 5. (no mut.) No upstream mutations. (before/same/after) Upstream mutations appeared in cycles before/at the same time as/after CNP segment integration. (lost/repl.) Upstream mutations of the aforementioned groups which disappeared (lost) at a later time point and (possibly) were regained (repl.) afterwards.

Results – Population Dynamics Experiment

"I'll be up in the gym, just working on my fitness. He's my witness."

—Bacteria GAT ATT GTG GCG

Competent *B. subtilis* are growth-arrested [88], and growth arrest is likely the main reason competence development is tightly regulated. Roughly 15% of a lab strain population differentiates into the K-state when entering the stationary phase, but natural isolates show even lower probabilities [80], which vary between isolates. For example, plasmids have been shown to be responsible for competence-development variation in natural isolates. The plasmids either interfered with competence development [188] or encoded for the repressor of competence, *rok* [189]. At the same time, there is no net growth of the population in the stationary phase. It is unclear if all cells are growth-arrested, or the rate of division was comparable to the death rate. In the first case, growth arrest during competence should come at no cost. In the second case, competence would result in a fitness cost. To address these points, population dynamics and flow chamber experiments were constructed, which enabled us to monitor differentiation into the K-state under stationary-state conditions.

5.1 The K-state Confers a Fitness Cost During the Stationary Phase

By immobilizing cells and feeding them a constant flow of conditioned medium (Section 3.1.1), the generation times of K-state and non K-state cells were determined. Although net growth was observed in this experimental setup, bacteria transiently differentiated into the K-state (Figure 5.1(a,b)), as observed previously by Süel *et al.* Figure 5.1(a) shows an example where a K-state cell did not grow or divide while its sibling (from the previous division event) continued to grow and divide. The average generation time of cells that ran through a period of differentiation was considerably increased compared to cells which never differentiated into the K-state (Table 5.1).

Next, the effect of differentiating into the K-state on the relative fitness, while in the stationary phase, was determined. To this end, head-to-head competition experiments between strains with varying differentiation probability were performed (Section 3.1.2). The fraction of wild type (BD630) K-state cells in the early stationary phase was 15%. Differentiation into the K-state was fully inhibited in a *comK* deletion strain (Bs075), whereas nearly 100% of the *rok* deletion strain (Bs056) differentiated into the K-state. As a control, wild type was competed with the reporter strain wild type-*gfp* (Bs139), and the fractions remained close to 0.5 during the time course of the experiment—i.e., the selection coefficient was close to zero.

Table 5.1: Generation times of K-state and non K-state cells averaged over at least 100 cells

	Generation time (min)
K-state cells	250 ± 40
Non K-state cells	116 ± 4

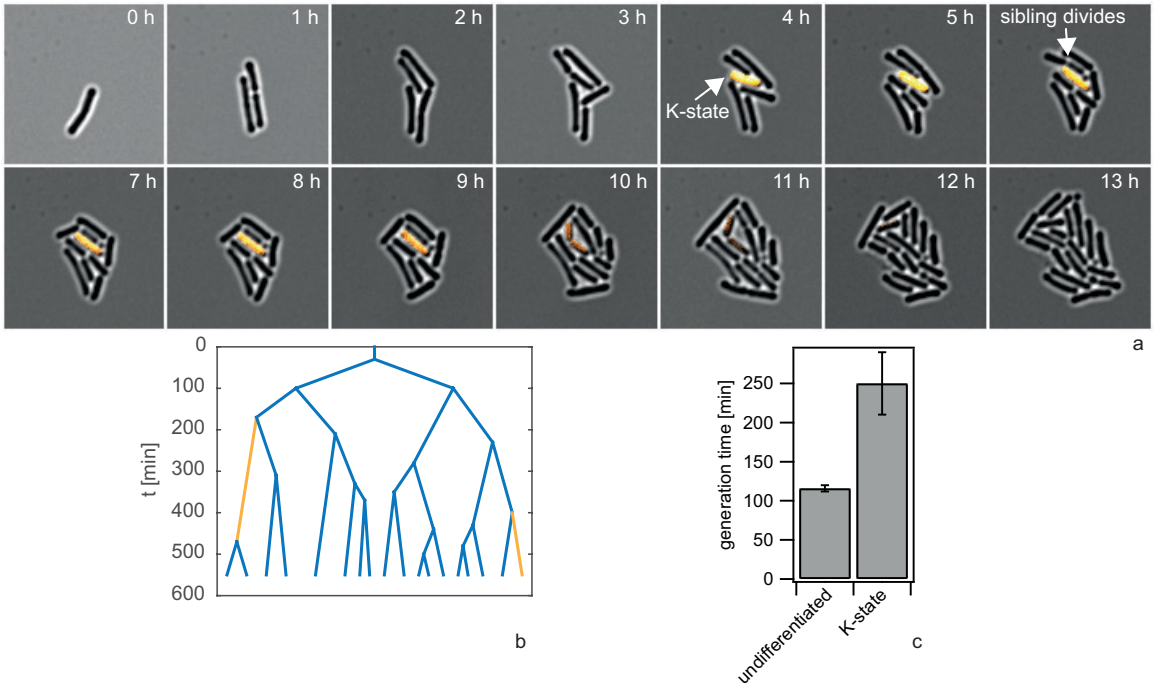


Figure 5.1: (a) Typical time lapse of P_{comK}-*gfp* (BD2711) grown in conditioned stationary state medium (T_0). Bright-field and fluorescence images are merged showing differentiated bacteria in orange. (b) Typical genealogy of cells after T_0, grown in the flow chamber and supplied with conditioned medium. (blue) Non K-state cells. (orange) Cells differentiating into the K-state. (c) Average generation times. Image adapted from [165] and under CC-BY license.

When non-competent Δ*comK* competed with wt-*gfp*, the fraction of wild type cells continuously decreased, and on the contrary, during competition between hyper-competent Δ*rok* and wt-*gfp*, the wild type dominated. The competition dynamics can be seen in Figure 5.2 along with the replicator equation fits used to determine the selection coefficients, Table 5.2.

Table 5.2: Selection coefficients in the stationary phase

Wild type-*gfp* competitor	Selection coefficient (h^{-1})
Wild type	0.006 ± 0.004
Δ*comK*	0.047 ± 0.005
Δ*rok*	−0.052 ± 0.007

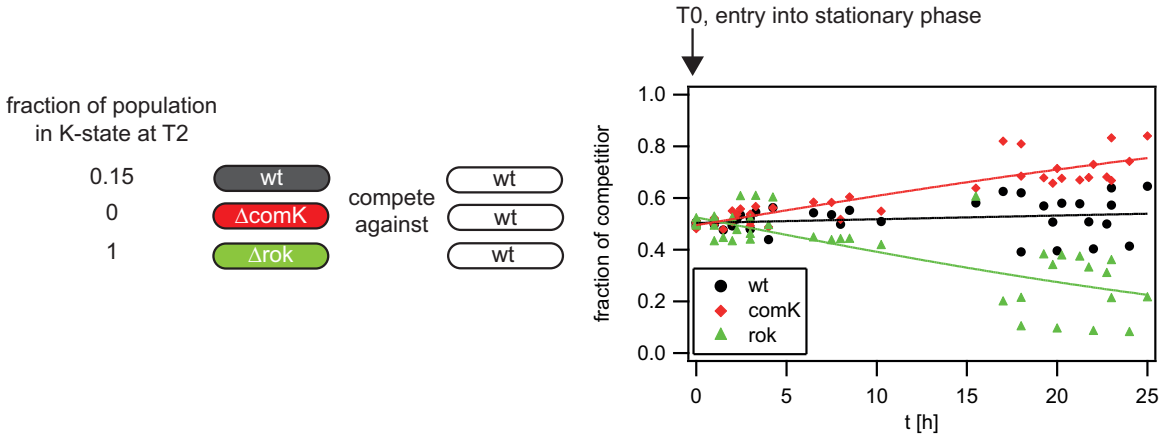


Figure 5.2: Competitors were grown separately to T_0, diluted into fresh competence medium, and mixed in a 1:1 ratio, as outlined in Section 3.1.2. The fraction of wild type (BD630, black circle), hyper-competent Δ*rok* (Bs056, green triangle), and non-competent Δ*comK* (Bs075, red diamond) competing against wild type-*gfp* (Bs139) cells are plotted. Full lines: best fit to replicator equation. Data was obtained from at least three independent experiments for each condition. Image adapted from [165] and under CC-BY license.

These experiments showed that even in the stationary phase where overall growth has plateaued, competence development is associated with a strong cost. This cost is tuned gradually as a function of the probability of differentiating into the K-state. The generation time of K-state cells is more than twice as long as for non K-state cells.

Discussion

"Every[thing] has to pass a series of rigorous tests: How many beats per minute? How many drops? How dope were the drops?... to answer one simple question."

—Is it a Banger?

The bistable K-state in *B. subtilis* has long been hypothesized to be an accelerator of adaptation. Studies have shown that competence for transformation can facilitate HGT, but its effect on genome dynamics and interactions between subspecies in mixed communities is unclear. In this study, we conducted an evolution experiment for 21 cycles, periodically supplying extracellular DNA from *B. subtilis* W23, to measure the beneficial effects of the K-state. We found that recombination occurred fairly randomly with an exponential distribution of lengths. Recombination events were more likely to completely replace genes, and essential genes were more often affected than expected. The largest fraction of de novo variants were in upstream regions, and of those, the majority were within an integration event. This hinted that they may be compensatory mutations.

In the second part of this thesis, K-state costs and dynamics were measured. It has been shown that K-state cells are growth inhibited, which poses the question how competence for transformation has been maintained, from a fitness perspective. Head-to-head competition experiments, between *B. subtilis* strains with differing probabilities of entering the K-state, allowed us to quantify the cost of entering and remaining in the K-state. Microfluidic experiments in the stationary phase allowed us to monitor the dynamics of the stationary phase at the single cell level and determine stationary phase growth rates. We found growth inhibition resulted in a strong cost of entering the K-state, even in the stationary phase.

6.1 Evolution Experiment

Our two-day cycle evolution experiment and the analysis methods adapted from [163] proved highly successful at detecting orthologous recombination and de novo insertions. In a total of 42 h of competence, we were able to see an average genome replacement of about 10%, with an exponential distribution of recombination lengths. Each replicate followed an individual evolutionary path and all showed fairly random gene replacement. There was a bias towards recombined segments with higher identity and, likely, selection for several genes and gene classes. Genes were most likely to be replaced completely, and subspecies donor DNA caused an increase in the number of de novo variants, possibly compensatory mutations.

6.1.1 Homologous Recombination Occurs at a Constant Rate with an Exponential Distribution

DNA uptake over the course of the 21 cycles occurred at a fixed rate with mean and median import lengths remaining constant. Recombination lengths followed an exponential distribution with a well-defined characteristic length, indicating that integration was random.

Our mean segment sizes were significantly smaller than the average 8.5 – 10 kbp measured by Dubnau *et al.* while selecting for tryptophan prototrophs [39], [191], [192]. Our non-selective recombination rate and segment sizes were different from those of similar evolution experiments. One-day experiments with *H. influenzae* found much larger donor segment sizes, 6.9 kbp, but only up to 3.2% of the genome had been replaced [112]. In another one-day experiment with *H. pylori* import lengths varied between 1.3 – 3.9 kbp [113]. Both of these experiments selected for recombination using antibiotics. In longer experiments with *H. pylori without* selection, similar import lengths to our experimental values (1645 bp) were found, with up to 8% genome replacement [163]. Our import lengths and genome replacement rates broadly fell within all of those listed here and highlight how experimental design greatly affects the two values.

In our integrated segments, we did not see a bimodal distribution of import lengths, as was seen in natural transformation experiments with *H. pylori* [163], but rather an exponential distribution with a decay constant of $3500\,\mathrm{bp}^{-1}$. This put our study seemingly at odds with previous studies by Morrison *et al.*, Weinrauch *et al.* (Figure 4.3). In their studies, transformation efficiency increased exponentially with segment length, eventually saturating near 30 kbp. Denaturation tests, using gel electrophoresis, on the genomic DNA supplied in our experiment showed that donor DNA had a tight length distribution around 20 kbp. The width of the distribution was estimated to not be more than several thousand base pairs. This implied that other mechanisms were at play, resulting in integrated segments being much smaller than the donor DNA distribution. We speculate that *B. subtilis* has a preference to import DNA up until a certain length (possibly reflective of our constant mean recombination length), recombination was selective (such that larger imports were broken up [Section 6.1.4]), or extracellular DNA was degraded before uptake. The last two hypotheses were the least likely. Maier *et al.* found DNA of up to 20.5 kbp was taken up at a nearly constant velocity and without considerable pausing in *B. subtilis* [195]; Zafra *et al.* showed that extracellular DNA levels increase during exponential growth, negating the possibility of DNase in the extracellular space [196].

6.1.2 A Constant Recombination Rate Implies Minor Fitness Changes and Small Epistatic Costs

Up to and including cycle 21, orthologous replacement occurred at a constant linear rate of 0.47% of the genome per cycle (Figure 4.2). This constant rate implied that the epistatic barrier to gene replacement was low. This could be due to two primary causes. First, as only a small fraction of the genome had been replaced, sufficient non-replaced regions of the genome were free to facilitate non-epistatic transfer. We would expect this free replacement to level off at future cycles, as a significant portion of the genome was replaced. Second, the fitness cost due to epistasis is nearly constant in orthologous replacement between two subspecies. The cross-species fitness model of Pinheiro and Lässig [197] detects epistasis as fitness valleys when transforming, gene by gene, from one species to another (Figure 6.1). In their model, an evolved species' fitness, F, is measured as function of fraction of the genome which has been transformed q, the proportionality constants for epistatic costs and total number of genes J and g, and a directional selection component d.

$$F = -2gJq(1-q) + d \tag{6.1}$$

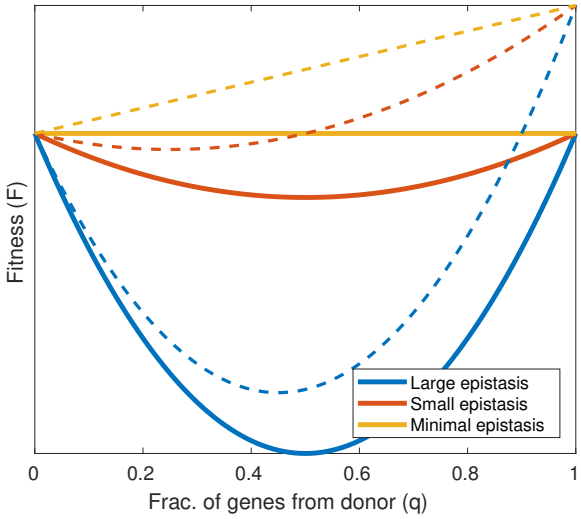


Figure 6.1: Cartoon scheme of cross-species fitness model [197]. The larger the epistasis between two species, the lower the fitness drops when replacing native genes with donor genes, one by one. Solid lines – no directional selection. Dashed lines – constant directional selection.

With no directional selection, $d = 0$, small changes in F are directly proportional to J (solid yellow line, Figure 6.1). We hypothesize that a constant genome replacement rate implies minor changes in fitness, corresponding to a small epistatic cost J.

At cycle 21, there was an emergence of small segments (<100bp) with lower than average identities (Table 4.2). These short, low identity segments could mark the onset of the non-linear regime of DNA import. If the majority of high-identity, low-epistasis, synonymous-mutation-rich segments of the genome have been replaced after ~10% genome exchange, it is plausible that small replacement segments would be favored over larger segments, because they have a lesser impact on gene function and smaller epistatic effect.

6.1.3 Auxiliary Regions are Imported Randomly into the Genome

Auxiliary regions, are sections of a genome that are specific and particular to that species. Auxiliary regions from BsuW23 were found in 6 of 7 replicates at cycle 21. On average, 5 ± 6 imports were found in each replicate with mean lengths of 2200 ± 2900 bp. (The large standard deviation is a result of replicate W6 having more than three times the average number of auxiliary gene imports, 19 imports.) These mean lengths are in agreement with previous *S. pneumoniae* studies were it was found that unselected recombinations without homologous flanking regions (de novo integrations) had a mean length of 2.3 kbp [111].

If we, similarly, assume that recombination of novel DNA segments occurs randomly on the genome and with a constant probability rate, our auxiliary-gene recombination rate-constant would be $4.5 \pm 0.8 \times 10^{-4}\,\mathrm{bp}^{-1}$. This is in excellent agreement with Croucher *et al.* 's $4.40 \times 10^{-4}\,\mathrm{bp}^{-1}$ from *S. pneumoniae* and relative agreement with Mell *et al.* 's $1.37 \times 10^{-4}\,\mathrm{bp}^{-1}$ in *H. influenzae*. Mell *et al.* did not exclude integration events that were selected for in their measurement of the gene recombination rate-constant, leading to a larger number of large segment sizes, and therefore a smaller recombination rate-constant [111], [112].

The same basal recombination rate constant (for de novo insertions) in two different species might hint

that there is a universal recombination rate for foreign DNA. RecA is responsible for matching imported DNA to homologous regions, and a study on RecA pairing by Forget *et al.* found that RecA formed unstable heterologous pairs 22% of the time [41], [198]. This recombination rate-constant would then be reflective of the error rate of RecA, or in other words, the rate at which unstable heterologous pairing leads to successful gene transformation. Additionally, this rate constant could imply that the recombination probability of foreign DNA decreases exponentially with divergence (as shown by [93], [199]) but flattens out at large sequence divergences. Past a certain sequence divergence, ~20%, additional differences between the donor DNA and recipient site could result in marginal changes in binding efficiency, site detection through RecA, or mismatch repair detection.

6.1.4 Recombination is Biased towards Higher Identities

Donor segments that integrated into Bsu168 had a higher average sequence identity than the interspecies average of 92.4%. This was significant at the $p < 0.001$ level, with the exception of segments <100 bp in length at cycle 21. Our results contrast previous findings in *H. influenzae* and *S. pneumoniae* where integrated segments showed no correlation with local identity [111], [112], [164]. Lack of correlation between local identity and recombination is most probably due to the lower divergence between the donor and recipient strains used in those experiments, on the order of 2 - 4% with one experiment reaching 6% divergence. Lower sequence divergence makes it more difficult to detect a correlation between donor sequence identity and recombination rate.

Although Mell *et al.* did not find that higher identity segments were favored during recombination, they did find that integrated segments had a higher than average identity at their bookends. Our work went further to show that the distribution of identities for the first and last 100 bp can be separated, at the 1% significance level using the two-sample Kolmogorov-Smirnov test (KS2) test (Figure 4.16). Furthermore, the distribution of the 100 bp with the lower identity matches the distribution of simulated random import events. It implies that recombination favored a start position with higher sequence identity, regardless of the identity of the end position. This finding differs from Majewski *et al.* 's calculation that *B. subtilis* requires a minimum flanking end on both sides of an import [200]–[202]. Finally, we did not detect a bias for end with the higher sequence identity to sit on the 5' to 3' or 3' to 5' end on the segment, when read off the forward strand. This is consistent with no bias in the direction of recombination with respect to the origin of replication [93].

Using the KS2 test we found that in 100 bp steps, the distribution of local identities for the first 500 bp of an integrated segment differed significantly (at the 5% level) from the distribution for the last 100 bp. The significance level hovered around 5% until about 2000 bp, where the distribution of the 100 bp groups and the last 100 bp could no longer be distinguished from another (Figure 4.17). This result implies that RecA takes advantage of a 500 bp region, at least, of higher local identity to facilitate stable pairing to genomic DNA between divergent subspecies. This higher identity region would affect the stability of pairing, and therefore the transformation rate, up until ~2000 bp, where it no longer plays a significant role. It would be interesting to see if this length range is due to the number of RecA proteins recruited to the ssDNA, in *B. subtilis* possibly due to DprA concentrations, or RecA's inter-segmental transfer pathway (three-dimensional homology search) [198].

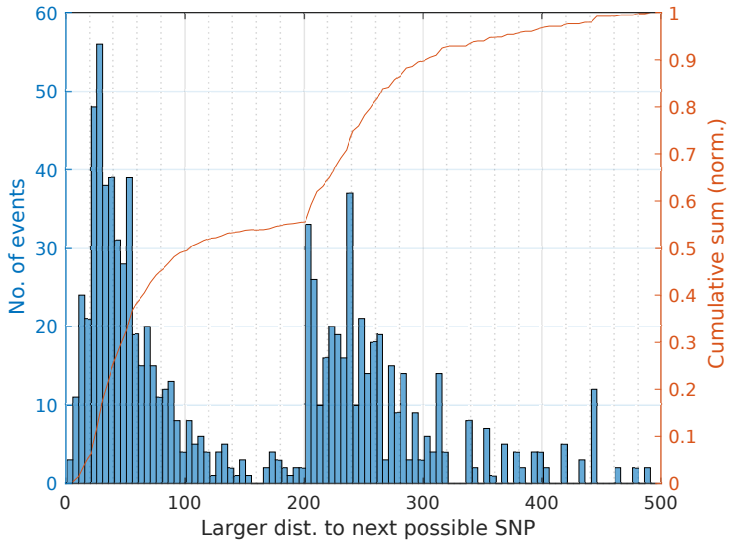


Figure 6.2: Distribution of maximum perfect identity lengths and normalized cumulative distribution. (left, blue) Histogram of the distance from the last detected master list SNP (mℓSNP) to the next possible mℓSNP, the larger of the two for each CNP. (right, orange) The cumulative sum of the distances, normalized. About 25% of the segments had perfect identity lengths less than the RecA-dependent requirement, 41 bp [201].

We did not see evidence that a minimum length of perfect identity is required for RecA-dependent recombination. A study by Majewski *et al.* found that *B. subtilis* required ~41 identical base pairs for RecA to recognize a homologous recombination size, and a previous study in *E. coli* estimated a similar value of 40 – 50 bp [201], [203]. We calculated the distance from the last detected SNP (bookend of a CNP) to the next possible (but not present) mℓSNP (master list SNP); this distance was the maximum length of identical sequence that could be considered part of a recombination event (Figure 6.2). It is clearly visible that recombination occurred with fewer than 41 identical base pairs at the start. (The peak at 200 bp is an artifact of the 200 bp sliding window to detect recombination events.) In this study ~25% of all recombinations had fewer than 41 identical base pairs. Our number is possibly underestimated, as Figure 6.2 shows the larger of the two identical sequence distances.

Our CNP algorithm did not include structural variants (indels or rearrangements), and as such, we did not find evidence of structural variants being more likely to terminate a recombination event. We did find evidence that recombinations significantly favored replacing one gene partially over two, as shown in Section 4.3, and discuss it in Section 6.1.7.

6.1.5 Recombination Hot Spots Putatively Correspond to Fitness Advantages

By cycle 21, 55% of the genome had never been replaced in any of the seven replicates. Our value is in marked contrast to Mell *et al.* where 2/3 of the genome had never been replaced in one transformation cycle *over 88 replicates*. The difference is most likely due to greater sequence divergence between Bsu168 and BsuW23, in comparison to the two *H. influenzae* strains used in their study. At the gene level (simplifying the homologously recombined segments to whether a gene was affected—i.e., replaced either fully or partially, or not at all), a binomial distribution was fit to the number of times a gene

was affected at cycle 9, 15, and 21 (Figure 4.13). Here, we assumed that the likelihood of replacement was equal for all genes. This assumption was reasonable, as the identity of individual genes did not vary greatly. The average rate of genome replacement, over all replicates at a given cycle, was used as the probability p, to calculate the binomial distribution. There were slight deviations between the experimental data and the distribution, at 0 – 2 replicates per affected gene. Nevertheless, a chi-squared test revealed that the null hypothesis, our experimental data came from the given binomial distribution, could not be rejected. Our null model could be corrected for, using the knowledge that genes with a higher identity have a higher probability of recombining. By weighting the original null model by an exponential decay function related to the distribution of genes identities, we would have a clear sign of selection if the weighted distribution matched our experimental data. It is also plausible that more of the genome has to be replaced to see epistatic effects; by cycle 21 only 10% of the genome, on average, had been replaced. Even if selection had been present in the experiment, without epistasis, we would have expected deviations from the binomial distribution. We concluded that little selection and no epistasis, at least at this point in the evolution experiment, were present.

Previous studies have found epistasis through the relative fitness of the evolved or mutated populations. By measuring the fitness and mutation rate of evolved strains over time, they found that fitness increases leveled off while the mutation rate remained constant; this type of negative epistasis is known as diminishing-returns epistasis [204]–[206]. Diminishing-returns epistasis was seen in multiple evolution experiments [116], [207], [208] and was shown by Schoustra *et al.* to be a pattern found not only in unicellular microbes and single genes but also multicellular organisms [209].

We performed fitness tests on our evolved populations but saw no overarching trend towards higher or lower fitness (G. Schneider and J. Power, unpublished data). This reflects our experimental design which does not constantly select for fitter individuals, due to the random bottleneck (created by platting UV-radiated cultures and selecting one colony at random the subsequent day). The lack of selection for fitter individuals explains why we do not see a dropping rate of fitness increase over the course of the experiment, and therefore, cannot confirm the presence of diminishing-returns epistasis. The lack of epistasis by cycle 21 is most likely reflective of the limited number of replicates and recombined segments.

Although no overarching fitness effects were detected, auxiliary regions were clear cold spots for gene transfer, including phage regions in Bsu168. Figure 4.14 strongly suggests recombination hot spots were present, in particular genes from the *leu* and *eps* operons. Our recipient strains are leucine auxotrophs, giving evolved strains which can produce their own leucine a clear advantage. *eps* genes are responsible for exopolysaccharide production and aid in biofilm formation [210]. It is plausible that biofilm formation could be advantageous, as half of the experiment occurs in the stationary phase.

6.1.6 Essential Genes are Preferentially Replaced in Evolved Strains

Essential genes [187] were found to be affected more frequently than non-essential genes, at the $p < 0.01$ level by cycle 21, using a one-sided t-test. Essential genes were affected with an average frequency of 8.9%, 3% more frequently than would be expected if all genes had an equal probability of being affected (Figure A.6). Further investigation found that the essential genes have a higher average identity

than non-essential genes, 95.1% to 93.7%, implying they were replaced by homologous recombination because of their high identity, and not their function. The high level of similarly in essential genes is expected for two subspecies. Subsequent experiments using more divergent species would give more insight into whether the essential genes were being affected solely due to their high similarity.

In general, the essential genes in *B. subtilis* are related to DNA replication, repair, and coiling (*dnaANX*, *gyrAB*, *holB*), ribosomal proteins (*rpl*, *rps*), chaperonins (*groESEL*), and aminotransferases (*gatABC*). Even with a high level of genomic similarity between the two subspecies, it is improbable that essential genes were replaced in Bsu168 because of better functionality. An extensive study by Bershtein *et al.* replaced the *folA* gene in *E. coli* with orthologs from 35 other bacteria. Despite most ortholgous being more catalytically active and at least as stable, growth rates immediately dropped 10 – 90% after recombination. Furthermore, after ~600 serial propagations of the orthologous strains, expression levels of *folA* increased only after mutations accumulated in the *lon* gene, responsible for maintaining protein homeostasis [211] and degrading misfolded or unstable proteins [212], [213]. It was concluded that protein homeostasis imposed an immediate barrier to the functional integration of foreign genes [161]. Similar experiments in *Salmonella typhimurium* drew the same conclusion [95], [214].

Spatial proximity of the Bsu168 essential genes also contributes to them being more frequently affected. The median integration length is ~2 genes (1900 bp). If recombination starts on an essential gene, it will most likely affect at least its neighboring gene because of the median two-gene segment length. Due to the proximity of essential genes on the genome, the neighboring gene is likely an essential gene.

Other than essential/non-essential genes, no bias towards replacement of a particular type of gene was seen. Overrepresentation of purines or pyrimidines was not found in the recombined segments (Figure 4.9).

6.1.7 Genes are More Likely to be Completely Replaced

Recombination events resulting in partial genes are most likely unfavorable due to the possibility of creating frame shifts or start/stop codon mutations. We found genes were completely replaced twice as often as partially, and operons nearly twice as often. This favoritism to completely replace genes can be explained by the mean recombination length, 1.9 kbp, which is about 2 genes. The average operon size is 3.2 kbp, which is also comparable to the mean recombination length. We asked the question if genes were being replaced due to selection.

To determine if complete gene replacement was due to selection, we created a null model where lengths from the experimental length distribution were taken and placed at random mℓSNP positions in the genome (Section 4.3.2). The simulations showed it was more likely that a recombination event would only partially replace one gene (one side of the recombination event ending within a gene) as opposed to two (using a KS2 test). We did not see the number of CNPs partially replacing zero genes—i.e., not partially replacing genes—differ between the in silico and experimental CNPs.

Comparing the mean and standard deviation of the experimental data to in silico recombination events, fewer CNPs partially replaced two genes and more replaced only one gene. Because of this, and our findings that one of the two CNP ends had a significantly lower SNP density (Section 4.3.2), we speculate that the initial binding of foreign DNA to a host's genome favors binding genome sections with low SNP

density and in an intergenic region. Orthologous recombination ends depending on the length of the recombining template, and therefore, ends randomly in an intra or intergenic region.

Our findings are in relative agreement with Dilthey *et al.*, who looked at an association between HGT and distance between the horizontally transferred genes across γ-proteobacteria. They found that spatial clustering was consistent with horizontally co-transferred genes, and probable if the transferred DNA was long enough [215].

While the distribution of completely replaced genes was the same for the experimental data and simulations, experimentally we saw that genes and operons were completely replaced twice as much as partially. There was also moderate selection towards fewer partially replaced genes. Our findings are in agreement with studies that demonstrate horizontally transferred genes are often spatially clustered. We note that the bias towards fully replaced genes could simply be a fitness effect. After transformation, the population grows overnight in liquid media. Cells with negative fitness effects due to partially replaced genes or operons will die; recombinations with fitness benefits will fixate.

6.1.8 De Novo Variants in Intergenic Regions of CNPs Occur Simultaneously with CNP Integration

Ten-fold more de novo variants were found in replicates evolved with BsuW23 DNA than with Bsu168 or no DNA (Figures 4.19 and 4.20). No and self DNA replicates had distributions with about three times as many synonymous and nonsynoymous variants compared to BsuW23 replicates (25% to 70%), and about 1/5 as many missense indels and intergenic mutations (15% to 70%). Of the intergenic mutations in BsuW23 replicates, 57% of them were inserts and 30% were deletions. This contrasts Tenaillon *et al.* 's findings looking at 50,000 generations of Lenski's *E. coli* experiment. There they found the number of nonsynonymous mutations was ~3.4 times greater than synonymous mutations and similarly that intergenic point mutations outnumbered synonymous mutations [160].

Focusing on the de novo variants that occurred within CNPs, 83% of all missense and in frame indels, and 75% of all upstream mutations occurred within CNPs. We speculate that the upstream mutations might be compensatory mutations or occur during recombination.

Assuming the mutations were compensatory and occurred *after* recombination, we calculated the minimum mutation rate needed to see the experimental number of de novo mutations. Each cycle consisted of ~10 generations, split evenly between the first and second day, and samples were sequenced every second cycle. Our wild type strain had a mutation rate of $\mu = 0.5 \times 10^{-9}$ bp^{-1} gen^{-1}, and Bsu168 had a genome size of 4.2×10^{6} bp. This gives a maximum of 4.2×10^{-2} mutations between recombination and clonal selection for sequencing (assuming recombination happened on the first day and sequencing after the second) – two orders of magnitude lower than needed to explain the upstream mutations. A mutator strain, such as Δ*mutSL*, only yields a 40-fold increase in the mutation rate [216] and would not account for the needed difference to claim the upstream mutations occurred post-recombination. Furthermore, no mutator strains were detected in any of our replicates. We concluded that the upstream mutations were not compensatory mutations fixing immediately after recombination.

It was plausible, that the upstream mutations were introduced through recombination. A study by Shee *et al.* found that RecD was responsible for a local increase in mutations following RecBCD repair

around a double strand break in *E. coli* [217]. Nevertheless, our recombination did not accompany double strand breaks and furthermore, RecD has not been shown to participate in bacterial transformation [41]. Research focusing on the polymerases that are called for during D-loop formation (a recombination intermediate), found that pol I and pol IV are both error-prone at D-loops. In particular, pol IV's mutagenic activity may be a direct result of RecA interacting with pol IV or the instability of the RecA mediated D-loop [218].

The addition of BsuW23 donor DNA caused an increase in the number of de novo variants. The majority of those de novo variants were missense indels or intergenic indels. We considered the possibility that those mutations were compensatory mutations or occurred during recombination. We ruled out compensatory mutations, as the minimum mutation rate would have been too high to explain all of the missense indels and intergenic mutations; none of our replicates were mutator strains. It remained possible that these variants were caused by errors in the recombination process. Additional studies would needed to look at how the number of intergenic and missense indels varies with sequence divergence.

6.1.9 A Suitable Method for the Detection of Gene Transfer Across *B. subtilis* Subspecies

The two-day cycle evolution experiment and the analysis methods adapted from [163] proved highly successful at detecting orthologous recombination.

The ability to become competent was not lost or negatively affected during the course of the experiment. Because competence has a large regulatory network [219] and K-state cells are growth arrested, it is an easy target for mutations or recombination. Such variants would render the competence network nonfunctional. To avoid that outcome, we placed competence under the control of an inducible promoter and disabled the native promoter by inserting a resistance cassette. For all seven replicates, DNA uptake continued to increase up through cycle 21. Further, DNA uptake was linear over the course of those 21 cycles, implying competence genes were functional and not negatively affected over the course of the experiment.

There were large variances in the fractions of replaced genomes, even within the same replicate between sequenced time points. Various replicates had recombination in genes that might have affected DNA uptake rates including the competence network *comACENP*, DNA mismatch repair *mutMSL*, and DNA recombination and repair *recAFGNOQUX* (Figures A.8-A.12). We speculate that variations in DNA uptake probability or intracellular targeting could be present.

We used UV radiation once a cycle to induce genomic damage that would result in higher DNA transformation rates, particularly of non-self DNA. Both no DNA and Bsu168 DNA samples showed comparable numbers of de novo variants at cycles 9 and 15. If UV radiation had caused mutations, we would have expected the mutations to accumulate in the no DNA replicates and be repaired every cycle in the Bsu168 replicates. Previous studies on mutation rates and UV radiation in *B. subtilis* found radiation dosages far below those used in this experiment were sufficient to produce mutation rates of up to 10^{-3} mutations/bp [220]. It is unlikely that UV radiation introduced a considerable number of mutations over this experiment.

Finally, the analysis algorithm was robust to the experimental results. Control replicates receiving

either no DNA or Bsu168 DNA showed no orthologous recombination over the course of the experiment. The overall number of SNPs found in individual no DNA and Bsu168 DNA replicates (on the order of tens of SNPs) remained well below sensitivity saturation of 12,000 variants. The 200 bp cluster window and 30% threshold for missing SNPs did not result in a large fraction of CNP-potential SNPs being removed from the CNP algorithm. Missing SNPs amounted to only 0.7 ± 0.2 % of all SNPs linked to a CNP, for seven replicates at cycle 21.

To summarize, the method of inducing competence is useful for maintaining competence in an environment that putatively selects against it. It also had the fortunate byproduct of reducing the overall man-hours needed to conduct the experiment. We speculate that UV radiation was not effective in producing large numbers of mutations nor increasing DNA transformation rates. The analysis algorithm was robust in detecting recombination events and against false positives.

6.2 Population Dynamics Experiment

Studies have shown that K-state cells are growth inhibited, brining into question how competence for transformation has been maintained from a fitness perspective. We quantized those fitness costs in head-to-head competition assays. With single-cell microscopy we found that there was a clear cost of competence, even in the stationary phase. We purpose that the K-state is a fitness trade-off in varying environments and can convey a fitness benefit in the presence of antibiotics by functioning as a persister-like state.

6.2.1 Stochastic Differentiation as a Fitness Trade-Off in Fluctuating Environments

Differentiation into the K-state is associated with growth-arrest [88], and under benign conditions this growth arrest confers a fitness cost, even in the stationary phase. By characterizing the competition dynamics between strains with different probabilities of switching into the K-state, we quantized fitness trade-offs for differentiation into the K-state. Under benign conditions, the relative fitness of the competitors decreased with increasing differentiation probability. In competitions between Δ*rok* and wild type, the Δ*rok* frequency was reduced to 2% after 24 h. This comparison shows that phenotypic heterogeneity strongly reduces the decline of relative fitness caused by differentiation into the K-state.

Considering the pair-wise competitions under benign conditions, the generation of phenotypic heterogeneity by means of the K-state was a useful strategy for exploiting the persister phenotype of the K-state when under stress, while minimizing the cost under benign conditions. This strategy of dealing with fluctuating levels of antibiotics is different for *S. pneumoniae*, which induces competence in response to sub-MIC levels of antibiotics [40], [221] whereas *B. subtilis* does not [165].

Outlook

"And as the harvest moon rose over Cyberland, Elise whispered, 'The only way out is up!' They reared back, springing into a gallop, and leap out of orbit."

—The Only Thing to do is Jump Over the Moon

This study focused on two aspects of competence in evolution: the direct cost and benefits of the K-state and competence machinery, and the putative role of competence in gene acquisition between subspecies.

We successfully measured the physiological cost of generating DNA uptake machinery in a benign environment. The next step would be to use this experimental design to measure the cost of the K-state under selection. There are several point mutations in *B. subtilis* that are known to confer antibiotic resistance (nalidixic acid, novobiocin, and rifampicin) [222]–[224]; by growing wild type bacteria on selective plates, we could clone resistant colonies and harvest their DNA. That DNA could then be used in the stationary phase experiment, along with the corresponding antibiotics, giving K-state cells an advantage when they integrate the resistance gene into their genome. Varying the antibiotic and DNA concentration would yield a deeper understanding as to when the K-state begins to confer an evolutionary advantage.

In addition to K-state costs, we were able to measure the division time of cells in the stationary state. Advances in microfluidic chambers allow for much more extravagant and detailed mining of single cell dynamics. Chambers such as the "mother machines" [142] or those where the height is no larger than a cell diameter could be used to monitor growth in the stationary phase. The open question of how dynamic the stationary phase actually is could be addressed.

The larger portion of this study involved supplying a growing population of *B. subtilis* 168 foreign DNA from *B. subtilis* W23, and analyzing their genomes over numerous cycles. Our experiments proved extremely successful at not only detecting homologous recombination events, but also seeing a large and broad number of events within a short experimental time frame. At the time when this study was published, epistasis had yet to be detected, namely due to a limited number of recombination events. The most obvious first step forward is to continue the evolution experiment to look for signs of epistasis and see when recombination becomes non-random.

Fitness measurements with the existing evolved samples and later experimental time points could be carried out to highlight what recombinations lead to higher fitness. Genes thought to have conveyed a benefit would be cloned directly into the wild type *B. subtilis* 168 strain to confirm if that particular gene lead to the fitness increase. It would also be interesting to look at the transcriptomics of the evolved samples, to see if recombination or putative compensatory mutations lead to different protein expression rates, as seen in previous experiments [161].

A related experiment could be performed to see if mutations upstream of recombination events are compensatory or a side effect of the recombination machinery. Clones conveying antibiotic resistance through a single point mutation—e.g., nalidixic acid, novobiocin, and rifampicin—could be selected

for in various *Bacillius* species. One of those resistant genes would then be transformed into *B. subtilis* 168, and clonal sequencing would reveal how frequently mutations arose as a result of the recombination machinery. One could potentially analyze if the machinery is more error prone at particular bases or motifs.

The experimental design could also be taken further to more divergent species or different external stresses. Using DNA from more divergent species causes the recombination rate to drop, log-linear with divergence, making recombination events more rare and, potentially, beneficial. Using alternate external stress factors,—e.g., temperature—could allow for more efficient recombination from species with larger genomic divergence. One could measure if competence actually speed up the evolutionary process, allowing the recipient strain to grow in the new environment or if the accumulation of point mutations, replication errors, and possibly mutator strains is sufficient to adapt to the same niche.

Our results implied that genes were preferentially replaced completely, and possibly operons, too. An operon with multiple genes could be chosen and clones generated where each had a different percentage of the operon replaced (with DNA from a related species). The fitness of those clones could then be measured to see if operon replacement had strong epistasis: complete operons needed to be replaced, weak epistasis: several elements of the operon needed to be replaced together, or no epistasis: individual genes could be replaced at will with no negative outcome. A similar study could be carried out at the gene level, but extreme care would need to be taken concerning potential reading frame shifts due to inserts and deletions.

Finally, genotype tracking methods have recently been developed [225], [226] that allow one to track an individual bacterium *genomically*, within a population. A random sequence, different for each bacterium, would be integrated into the genome and functions as a bar code. Compared to the time-resolved clonal whole genome sequencing performed in this study, a genotype tracking method allows for higher beneficial-mutation sensitivity.

Bibliography

[1] S. Mitri and K. Richard Foster, "The genotypic view of social interactions in microbial communities," *Annual Review of Genetics*, vol. 47, no. 1, pp. 247–273, 2013. DOI: 10.1146/annurev-genet-111212-133307.

[2] C. D. Nadell, J. B. Xavier, S. A. Levin, and K. R. Foster, "The evolution of quorum sensing in bacterial biofilms," *PLOS Biology*, vol. 6, no. 1, N. A. Moran, Ed., e14, 2008. DOI: 10.1371/journal.pbio.0060014.

[3] R. D. Monds and G. A. O'Toole, "The developmental model of microbial biofilms: Ten years of a paradigm up for review," *Trends in Microbiology*, vol. 17, no. 2, pp. 73–87, 2009. DOI: 10.1016/j.tim.2008.11.001.

[4] L. Hall-Stoodley, J. W. Costerton, and P. Stoodley, "Bacterial biofilms: From the natural environment to infectious diseases," *Nature Reviews. Microbiology*, vol. 2, no. 2, pp. 95–108, 2004. DOI: 10.1038/nrmicro821.

[5] K. R. Foster, "A defense of sociobiology," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 74, pp. 403–418, 2009. DOI: 10.1101/sqb.2009.74.041.

[6] S. A. West and A. Buckling, "Cooperation, virulence and siderophore production in bacterial parasites," *Proceedings of the Royal Society B: Biological Sciences*, vol. 270, no. 1510, pp. 37–44, 2003. DOI: 10.1098/rspb.2002.2209.

[7] M. Miethke and M. A. Marahiel, "Siderophore-based iron acquisition and pathogen control," *Microbiology and Molecular Biology Reviews*, vol. 71, no. 3, pp. 413–451, 2007. DOI: 10.1128/MMBR.00012-07.

[8] J. A. Shapiro and M. Dworkin, Eds., *Bacteria as multicellular organisms*, New York: Oxford University Press, 1997, 466 pp.

[9] B. A. Duerkop, J. Varga, J. R. Chandler, S. B. Peterson, J. P. Herman, M. E. A. Churchill, M. R. Parsek, W. C. Nierman, and E. P. Greenberg, "Quorum-sensing control of antibiotic synthesis in *Burkholderia thailandensis*," *Journal of Bacteriology*, vol. 191, no. 12, pp. 3909–3918, 2009. DOI: 10.1128/JB.00200-09.

[10] P. B. Rainey and K. Rainey, "Evolution of cooperation and conflict in experimental bacterial populations," *Nature*, vol. 425, no. 6953, pp. 72–74, 2003. DOI: 10.1038/nature01906.

[11] L. Keller and M. G. Surette, "Communication in bacteria: An ecological and evolutionary perspective," *Nature Reviews. Microbiology*, vol. 4, no. 4, pp. 249–258, 2006. DOI: 10.1038/nrmicro1383.

[12] A. D. Grossman, "Genetic networks controlling the initiation of sporulation and the development of genetic competence in *Bacillus subtilis*," *Annual Review of Genetics*, vol. 29, pp. 477–508, 1995. DOI: 10.1146/annurev.ge.29.120195.002401.

[13] P. Tortosa, L. Logsdon, B. Kraigher, Y. Itoh, I. Mandic-Mulec, and D. Dubnau, "Specificity and genetic polymorphism of the *Bacillus* competence quorum-sensing system," *Journal of Bacteriology*, vol. 183, no. 2, pp. 451–460, 2001. DOI: 10.1128/JB.183.2.451-460.2001.

[14] A. Oslizlo, P. Stefanic, I. Dogsa, and I. Mandic-Mulec, "Private link between signal and response in *Bacillus subtilis* quorum sensing,"
Proceedings of the National Academy of Sciences, vol. 111, no. 4, pp. 1586–1591, 2014. DOI: 10.1073/pnas.1316283111.

[15] K. Y. Le and M. Otto, "Quorum-sensing regulation in staphylococci – an overview,"
Frontiers in Microbiology, vol. 6, p. 1174, 2015. DOI: 10.3389/fmicb.2015.01174.

[16] O. X. Cordero, H. Wildschutte, B. Kirkup, S. Proehl, L. Ngo, F. Hussain, F. Le Roux, T. Mincer, and M. F. Polz, "Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance," *Science (New York, N.Y.)*, vol. 337, no. 6099, pp. 1228–1231, 2012. DOI: 10.1126/science.1219385.

[17] V. Ojala, S. Mattila, V. Hoikkala, J. K. H. Bamford, and M. Jalasvuori, "Evolutionary rescue of bacteria via horizontal gene transfer under a lethal beta-lactam concentration,"
Journal of Global Antimicrobial Resistance, vol. 2, no. 3, pp. 198–200, 2014.
DOI: 10.1016/j.jgar.2014.02.005.

[18] R. Niehus, S. Mitri, A. G. Fletcher, and K. R. Foster, "Migration and horizontal gene transfer divide microbial genomes into multiple niches," *Nature Communications*, vol. 6, no. 1, 2015.
DOI: 10.1038/ncomms9924.

[19] O. Popa and T. Dagan, "Trends and barriers to lateral gene transfer in prokaryotes,"
Current Opinion in Microbiology, vol. 14, no. 5, pp. 615–623, 2011.
DOI: 10.1016/j.mib.2011.07.027.

[20] G. Schonknecht, W.-H. Chen, C. M. Ternes, G. G. Barbier, R. P. Shrestha, M. Stanke, A. Brautigam, B. J. Baker, J. F. Banfield, R. M. Garavito, K. Carr, C. Wilkerson, S. A. Rensing, D. Gagneul, N. E. Dickenson, C. Oesterhelt, M. J. Lercher, and A. P. M. Weber, "Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote,"
Science, vol. 339, no. 6124, pp. 1207–1210, 2013. DOI: 10.1126/science.1231707.

[21] S. M. Soucy, J. Huang, and J. P. Gogarten, "Horizontal gene transfer: Building the web of life,"
Nature Reviews Genetics, vol. 16, no. 8, pp. 472–482, 2015. DOI: 10.1038/nrg3962.

[22] J. P. J. Hall, M. A. Brockhurst, and E. Harrison, "Sampling the mobile gene pool: Innovation via horizontal gene transfer in bacteria," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1735, p. 20 160 424, 2017. DOI: 10.1098/rstb.2016.0424.

[23] G. P. Dubey and S. Ben-Yehuda, "Intercellular nanotubes mediate bacterial communication,"
Cell, vol. 144, no. 4, pp. 590–600, 2011. DOI: 10.1016/j.cell.2011.01.015.

[24] S. Fulsundar, K. Harms, G. E. Flaten, P. J. Johnsen, B. A. Chopade, and K. M. Nielsen, "Gene transfer potential of outer membrane vesicles of *Acinetobacter baylyi* and effects of stress on vesiculation,"
Applied and Environmental Microbiology, vol. 80, no. 11, M. Kivisaar, Ed., pp. 3469–3483, 2014.
DOI: 10.1128/AEM.04248-13.

[25] J. Chen and R. P. Novick, "Phage-mediated intergeneric transfer of toxin genes,"
Science, vol. 323, no. 5910, pp. 139–141, 2009. DOI: 10.1126/science.1164783.

[26] I. Chen and D. Dubnau, "DNA uptake during bacterial transformation,"
Nature Reviews Microbiology, vol. 2, no. 3, pp. 241–249, 2004. DOI: 10.1038/nrmicro844.

[27] C. J. H. von Wintersdorff, J. Penders, J. M. van Niekerk, N. D. Mills, S. Majumder, L. B. van Alphen, P. H. M. Savelkoul, and P. F. G. Wolffs, "Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer," *Frontiers in Microbiology*, vol. 7, 2016. DOI: 10.3389/fmicb.2016.00173.

[28] H. Muller, "Some genetic aspects of sex," *The American Naturalist*, vol. 66, no. 703, pp. 118–138, 1932.

[29] P. J. Gerrish and R. E. Lenski, "The fate of competing beneficial mutations in an asexual population," *Genetica*, vol. 102-103, no. 1, pp. 127–144, 1998.

[30] M. M. Desai and D. S. Fisher, "Beneficial mutation selection balance and the effect of linkage on positive selection," *Genetics*, vol. 176, no. 3, pp. 1759–1798, 2007. DOI: 10.1534/genetics.106.067678.

[31] I. M. Rouzine, J. Wakeley, and J. M. Coffin, "The solitary wave of asexual evolution," *Proceedings of the National Academy of Sciences*, vol. 100, no. 2, pp. 587–592, 2003. DOI: 10.1073/pnas.242719299.

[32] B. H. Good, I. M. Rouzine, D. J. Balick, O. Hallatschek, and M. M. Desai, "Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 13, pp. 4950–4955, 2012. DOI: 10.1073/pnas.1119910109.

[33] S. Schiffels, G. J. Szöllosi, V. Mustonen, and M. Lässig, "Emergent neutrality in adaptive asexual evolution," *Genetics*, vol. 189, no. 4, pp. 1361–1375, 2011. DOI: 10.1534/genetics.111.132027.

[34] H. Ochman, J. G. Lawrence, and E. A. Groisman, "Lateral gene transfer and the nature of bacterial innovation," *Nature*, vol. 405, no. 6784, pp. 299–304, 2000. DOI: 10.1038/35012500.

[35] X. Wang, Y. Kim, Q. Ma, S. H. Hong, K. Pokusaeva, J. M. Sturino, and T. K. Wood, "Cryptic prophages help bacteria cope with adverse environments," *Nature Communications*, vol. 1, no. 9, p. 147, 2010. DOI: 10.1038/ncomms1146.

[36] N. J. Croucher, R. Mostowy, C. Wymant, P. Turner, S. D. Bentley, and C. Fraser, "Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict," *PLOS Biology*, vol. 14, no. 3, N. H. Barton, Ed., e1002394, 2016. DOI: 10.1371/journal.pbio.1002394.

[37] M. Blokesch, "In and out—contribution of natural transformation to the shuffling of large genomic regions," *Current Opinion in Microbiology*, vol. 38, pp. 22–29, 2017. DOI: 10.1016/j.mib.2017.04.001.

[38] S. Nowak, J. Neidhart, I. G. Szendro, and J. Krug, "Multidimensional epistasis and the transitory advantage of sex," *PLOS Computational Biology*, vol. 10, no. 9, N. Beerenwinkel, Ed., e1003836, 2014. DOI: 10.1371/journal.pcbi.1003836.

[39] D. Dubnau, "DNA uptake in bacteria," *Annual Review of Microbiology*, vol. 53, no. 1, pp. 217–244, 1999. DOI: 10.1146/annurev.micro.53.1.217.

[40] J.-P. Claverys, M. Prudhomme, and B. Martin, "Induction of competence regulons as a general response to stress in gram-positive bacteria,"
Annual Review of Microbiology, vol. 60, no. 1, pp. 451–475, 2006.
DOI: 10.1146/annurev.micro.60.080805.142139.

[41] J.-P. Claverys, B. Martin, and P. Polard, "The genetic transformation machinery: Composition, localization, and mechanism," *FEMS Microbiology Reviews*, vol. 33, no. 3, pp. 643–656, 2009.
DOI: 10.1111/j.1574-6976.2009.00164.x.

[42] G. S. Inamine and D. Dubnau, "ComEA, a *Bacillus subtilis* integral membrane protein required for genetic transformation, is needed for both DNA binding and transport,"
Journal of Bacteriology, vol. 177, no. 11, pp. 3045–3051, 1995.

[43] M. Berge, M. Moscoso, M. Prudhomme, B. Martin, and J.-P. Claverys, "Uptake of transforming DNA in gram-positive bacteria: A view from *Streptococcus pneumoniae*,"
Molecular Microbiology, vol. 45, no. 2, pp. 411–421, 2002.

[44] N. Strauss, "Configuration of transforming deoxyribonucleic acid during entry into *Bacillus subtilis*," *Journal of Bacteriology*, vol. 89, pp. 288–293, 1965.

[45] R. Provvedi, I. Chen, and D. Dubnau, "NucA is required for DNA cleavage during transformation of *Bacillus subtilis*," *Molecular Microbiology*, vol. 40, no. 3, pp. 634–644, 2001.

[46] S. Lacks and B. Greenberg, "Single-strand breakage on binding of DNA to cells in the genetic transformation of diplococcus pneumoniae,"
Journal of Molecular Biology, vol. 101, no. 2, pp. 255–275, 1976.

[47] D. A. Morrison and W. R. Guild, "Breakage prior to entry of donor DNA in pneumococcus transformation," *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis*, vol. 299, no. 4, pp. 545–556, 1973. DOI: 10.1016/0005-2787(73)90226-8.

[48] S. D. Goodman and J. J. Scocca, "Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, no. 18, pp. 6982–6986, 1988.

[49] K. L. Sisco and H. O. Smith, "Sequence-specific DNA uptake in *Haemophilus* transformation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 76, no. 2, pp. 972–976, 1979.

[50] F. E. Aas, M. Wolfgang, S. Frye, S. Dunham, C. Løvold, and M. Koomey, "Competence for natural transformation in *Neisseria gonorrhoeae*: Components of DNA binding and uptake linked to type IV pilus expression," *Molecular Microbiology*, vol. 46, no. 3, pp. 749–760, 2002.

[51] F. E. Aas, C. Løvold, and M. Koomey, "An inhibitor of DNA binding and uptake events dictates the proficiency of genetic transformation in *Neisseria gonorrhoeae*: Mechanism of action and links to type IV pilus expression," *Molecular Microbiology*, vol. 46, no. 5, pp. 1441–1450, 2002.

[52] C. Hepp and B. Maier, "Kinetics of DNA uptake during transformation provide evidence for a translocation ratchet mechanism,"
Proceedings of the National Academy of Sciences, vol. 113, no. 44, pp. 12 467–12 472, 2016.
DOI: 10.1073/pnas.1608110113.

[53] D. Facius, M. Fussenegger, and T. F. Meyer, "Sequential action of factors involved in natural competence for transformation of *Neisseria gonorrhoeae*,"
FEMS microbiology letters, vol. 137, no. 2, pp. 159–164, 1996.

[54] C. Johnston, B. Martin, G. Fichant, P. Polard, and J.-P. Claverys, "Bacterial transformation: Distribution, shared mechanisms and divergent control,"
Nature Reviews Microbiology, vol. 12, no. 3, pp. 181–196, 2014. DOI: 10.1038/nrmicro3199.

[55] D. Dubnau, "Genetic competence in *Bacillus subtilis*,"
Microbiological Reviews, vol. 55, no. 3, pp. 395–424, 1991.

[56] J. Hahn, L. Kong, and D. Dubnau, "The regulation of competence transcription factor synthesis constitutes a critical control point in the regulation of competence in *Bacillus subtilis*,"
Journal of Bacteriology, vol. 176, no. 18, pp. 5753–5761, 1994.
DOI: 10.1128/jb.176.18.5753-5761.1994.

[57] M. Leisner, K. Stingl, E. Frey, and B. Maier, "Stochastic switching to competence,"
Current Opinion in Microbiology, vol. 11, no. 6, pp. 553–559, 2008.
DOI: 10.1016/j.mib.2008.09.020.

[58] T. T. Hoa, P. Tortosa, M. Albano, and D. Dubnau, "Rok (YkuW) regulates genetic competence in *Bacillus subtilis* by directly repressing *comK*,"
Molecular Microbiology, vol. 43, no. 1, pp. 15–26, 2002.

[59] P. Serror and A. L. Sonenshein, "Interaction of CodY, a novel *Bacillus subtilis* DNA-binding protein, with the *dpp* promoter region,"
Molecular Microbiology, vol. 20, no. 4, pp. 843–852, 1996.

[60] L. W. Hamoen, D. Kausche, M. A. Marahiel, D. van Sinderen, G. Venema, and P. Serror, "The *Bacillus subtilis* transition state regulator AbrB binds to the -35 promoter region of *comK*,"
FEMS microbiology letters, vol. 218, no. 2, pp. 299–304, 2003.

[61] P. Prepiak and D. Dubnau, "A peptide signal for adapter protein-mediated degradation by the AAA+ protease ClpCP," *Molecular Cell*, vol. 26, no. 5, pp. 639–647, 2007.
DOI: 10.1016/j.molcel.2007.05.011.

[62] C. Bongiorni, S. Ishikawa, S. Stephenson, N. Ogasawara, and M. Perego, "Synergistic regulation of competence development in *Bacillus subtilis* by two Rap-Phr systems,"
Journal of Bacteriology, vol. 187, no. 13, pp. 4353–4361, 2005.
DOI: 10.1128/JB.187.13.4353-4361.2005.

[63] W. K. Smits, C. Bongiorni, J.-W. Veening, L. W. Hamoen, O. P. Kuipers, and M. Perego, "Temporal separation of distinct differentiation pathways by a dual specificity Rap-Phr system in *Bacillus subtilis*," *Molecular Microbiology*, vol. 65, no. 1, pp. 103–120, 2007.
DOI: 10.1111/j.1365-2958.2007.05776.x.

[64] K. L. Griffith and A. D. Grossman, "A degenerate tripartite DNA-binding site required for activation of ComA-dependent quorum response gene expression in *Bacillus subtilis*,"
Journal of Molecular Biology, vol. 381, no. 2, pp. 261–275, 2008.
DOI: 10.1016/j.jmb.2008.06.035.

[65] L. W. Hamoen, A. F. Van Werkhoven, J. J. Bijlsma, D. Dubnau, and G. Venema, "The competence transcription factor of *Bacillus subtilis* recognizes short A/T-rich sequences arranged in a unique, flexible pattern along the DNA helix,"
Genes & Development, vol. 12, no. 10, pp. 1539–1550, 1998.

[66] H. Maamar and D. Dubnau, "Bistability in the *Bacillus subtilis* K-state (competence) system requires a positive feedback loop: Bistability in *B. subtilis* competence," *Molecular Microbiology*, vol. 56, no. 3, pp. 615–624, 2005. DOI: 10.1111/j.1365-2958.2005.04592.x.

[67] W. K. Smits, C. C. Eschevins, K. A. Susanna, S. Bron, O. P. Kuipers, and L. W. Hamoen, "Stripping *Bacillus*: ComK auto-stimulation is responsible for the bistable response in competence development," *Molecular Microbiology*, vol. 56, no. 3, pp. 604–614, 2005. DOI: 10.1111/j.1365-2958.2005.04488.x.

[68] M. Leisner, K. Stingl, J. O. Rädler, and B. Maier, "Basal expression rate of *comK* sets a 'switching-window' into the K-state of *Bacillus subtilis*: Basal expression rate of *comK* sets a 'switching-window'," *Molecular Microbiology*, vol. 63, no. 6, pp. 1806–1816, 2007. DOI: 10.1111/j.1365-2958.2007.05628.x.

[69] H. Maamar, A. Raj, and D. Dubnau, "Noise in gene expression determines cell fate in *Bacillus subtilis*," *Science*, vol. 317, no. 5837, pp. 526–529, 2007. DOI: 10.1126/science.1140818.

[70] K. Turgay, J. Hahn, J. Burghoorn, and D. Dubnau, "Competence in *Bacillus subtilis* is controlled by regulated proteolysis of a transcription factor," *The EMBO journal*, vol. 17, no. 22, pp. 6730–6738, 1998. DOI: 10.1093/emboj/17.22.6730.

[71] J. E. Ferrell, "Self-perpetuating states in signal transduction: Positive feedback, double-negative feedback and bistability," *Current Opinion in Cell Biology*, vol. 14, no. 2, pp. 140–148, 2002.

[72] J. Guespin-Michel and M. Kaufman, "Positive feedback circuits and adaptive regulations in bacteria," *Acta Biotheoretica*, vol. 49, no. 4, pp. 207–218, 2001.

[73] J. F. Guespin-Michel, G. Bernot, J. P. Comet, A. Mérieau, A. Richard, C. Hulen, and B. Polack, "Epigenesis and dynamic similarity in two regulatory networks in *Pseudomonas aeruginosa*," *Acta Biotheoretica*, vol. 52, no. 4, pp. 379–390, 2004. DOI: 10.1023/B:ACBI.0000046604.18092.a7.

[74] D. Dubnau and R. Losick, "Bistability in bacteria," *Molecular Microbiology*, vol. 61, no. 3, pp. 564–572, 2006. DOI: 10.1111/j.1365-2958.2006.05249.x.

[75] Z. Zhang, D. Claessen, and D. E. Rozen, "Understanding microbial divisions of labor," *Frontiers in Microbiology*, vol. 7, p. 2070, 2016. DOI: 10.3389/fmicb.2016.02070.

[76] J.-W. Veening, W. K. Smits, and O. P. Kuipers, "Bistability, epigenetics, and bet-hedging in bacteria," *Annual Review of Microbiology*, vol. 62, no. 1, pp. 193–210, 2008. DOI: 10.1146/annurev.micro.62.081307.163002.

[77] P. J. Johnsen, D. Dubnau, and B. R. Levin, "Episodic selection and the maintenance of competence and natural transformation in *Bacillus subtilis*," *Genetics*, vol. 181, no. 4, pp. 1521–1533, 2009. DOI: 10.1534/genetics.108.099523.

[78] L. Espinar, M. Dies, T. Cagatay, G. M. Süel, and J. Garcia-Ojalvo, "Circuit-level input integration in bacterial gene regulation," *Proceedings of the National Academy of Sciences*, vol. 110, no. 17, pp. 7091–7096, 2013. DOI: 10.1073/pnas.1216091110.

[79] R. J. Redfield, "Genes for breakfast: The have-your-cake-and-eat-it-too of bacterial transformation," *The Journal of Heredity*, vol. 84, no. 5, pp. 400–404, 1993.

[80] F. M. Cohan, M. S. Roberts, and E. C. King, "The potential for genetic exchange by transformation within a natural population of *Bacillus subtilis*," *Evolution*, vol. 45, no. 6, pp. 1393–1421, 1991.

[81] V. Burrus and M. K. Waldor, "Shaping bacterial genomes with integrative and conjugative elements," *Research in Microbiology*, vol. 155, no. 5, pp. 376–386, 2004. DOI: 10.1016/j.resmic.2004.01.012.

[82] S. Krause, W. Pansegrau, R. Lurz, F. de la Cruz, and E. Lanka, "Enzymology of type IV macromolecule secretion systems: The conjugative transfer regions of plasmids RP4 and r388 and the cag pathogenicity island of *Helicobacter pylori* encode structurally and functionally related nucleoside triphosphate hydrolases," *Journal of Bacteriology*, vol. 182, no. 10, pp. 2761–2770, 2000. DOI: 10.1128/JB.182.10.2761-2770.2000.

[83] R. E. Michod, M. F. Wojciechowski, and M. A. Hoelzer, "DNA repair and the evolution of transformation in the bacterium *Bacillus subtilis*," *Genetics*, vol. 118, no. 1, pp. 31–39, 1988.

[84] M. F. Wojciechowski, M. A. Hoelzer, and R. E. Michod, "DNA repair and the evolution of transformation in *Bacillus subtilis*. II. role of inducible repair," *Genetics*, vol. 121, no. 3, pp. 411–422, 1989.

[85] M. A. Hoelzer and R. E. Michod, "DNA repair and the evolution of transformation in *Bacillus subtilis*. III. sex with damaged DNA," *Genetics*, vol. 128, no. 2, pp. 215–223, 1991.

[86] D. Moradigaravand and J. Engelstädter, "The evolution of natural competence: Disentangling costs and benefits of sex in bacteria," *The American Naturalist*, vol. 182, no. 4, E112–E126, 2013. DOI: 10.1086/671909.

[87] J. Hahn, A. W. Tanner, V. J. Carabetta, I. M. Cristea, and D. Dubnau, "ComGA-RelA interaction and persistence in the *Bacillus subtilis* K-state: ComGA and K-state persistence," *Molecular Microbiology*, vol. 97, no. 3, pp. 454–471, 2015. DOI: 10.1111/mmi.13040.

[88] B.-J. Haijema, J. Hahn, J. Haynes, and D. Dubnau, "A ComGA-dependent checkpoint limits growth during the escape from competence: A ComGA checkpoint limits growth after competence," *Molecular Microbiology*, vol. 40, no. 1, pp. 52–64, 2001. DOI: 10.1046/j.1365-2958.2001.02363.x.

[89] K. Briley Jr, P. Prepiak, M. J. Dias, J. Hahn, and D. Dubnau, "Maf acts downstream of ComGA to arrest cell division in competent cells of *B. subtilis*: Regulation of cell division by maf," *Molecular Microbiology*, vol. 81, no. 1, pp. 23–39, 2011. DOI: 10.1111/j.1365-2958.2011.07695.x.

[90] E. W. Nester and B. A. Stocker, "Biosynthetic latency in early stages of deoxyribonucleic acid transformation in *Bacillus subtilis*," *Journal of Bacteriology*, vol. 86, pp. 785–796, 1963.

[91] P. Zawadzki, M. S. Roberts, and F. M. Cohan, "The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust," *Genetics*, vol. 140, no. 3, pp. 917–932, 1995.

[92] H. Y. Chu, K. Sprouffske, and A. Wagner, "Assessing the benefits of horizontal gene transfer by laboratory evolution and genome sequencing," *BMC Evolutionary Biology*, vol. 18, no. 1, 2018. DOI: 10.1186/s12862-018-1164-7.

[93] B. Carrasco, E. Serrano, H. Sánchez, C. Wyman, and J. C. Alonso, "Chromosomal transformation in *Bacillus subtilis* is a non-polar recombination reaction," *Nucleic Acids Research*, vol. 44, no. 6, pp. 2754–2768, 2016. DOI: 10.1093/nar/gkv1546.

[94] R. Sorek, Y. Zhu, C. J. Creevey, M. P. Francino, P. Bork, and E. M. Rubin, "Genome-wide experimental determination of barriers to horizontal gene transfer," *Science*, vol. 318, no. 5855, pp. 1449–1452, 2007. DOI: 10.1126/science.1147112.

[95] A. Knöppel, P. A. Lind, U. Lustig, J. Näsvall, and D. I. Andersson, "Minor fitness costs in an experimental model of horizontal gene transfer in bacteria," *Molecular Biology and Evolution*, vol. 31, no. 5, pp. 1220–1227, 2014. DOI: 10.1093/molbev/msu076.

[96] T. Tuller, Y. Girshovich, Y. Sella, A. Kreimer, S. Freilich, M. Kupiec, U. Gophna, and E. Ruppin, "Association between translation efficiency and horizontal gene transfer within microbial communities," *Nucleic Acids Research*, vol. 39, no. 11, pp. 4743–4755, 2011. DOI: 10.1093/nar/gkr054.

[97] A. L. G. Utnes, V. Sørum, N. Hülter, R. Primicerio, J. Hegstad, J. Kloos, K. M. Nielsen, and P. J. Johnsen, "Growth phase-specific evolutionary benefits of natural transformation in *Acinetobacter baylyi*," *The ISME Journal*, vol. 9, no. 10, pp. 2221–2231, 2015. DOI: 10.1038/ismej.2015.35.

[98] D. A. Baltrus, K. Guillemin, and P. C. Phillips, "Natural transformation increases the rate of adaptation in the human pathogen *Helicobacter pylori*," *Evolution*, vol. 62, no. 1, pp. 39–49, 2007. DOI: 10.1111/j.1558-5646.2007.00271.x.

[99] Nature. (). Experimental evolution, [Online]. Available: https://www.nature.com/subjects/experimental-evolution (visited on 07/26/2018).

[100] H. J. E. Beaumont, J. Gallie, C. Kost, G. C. Ferguson, and P. B. Rainey, "Experimental evolution of bet hedging," *Nature*, vol. 462, no. 7269, pp. 90–93, 2009. DOI: 10.1038/nature08504.

[101] D. S. Treves, S. Manning, and J. Adams, "Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of *Escherichia coli*," *Molecular Biology and Evolution*, vol. 15, no. 7, pp. 789–797, 1998. DOI: 10.1093/oxfordjournals.molbev.a025984.

[102] R. Maharjan, S. Seeto, L. Notley-McRobb, and T. Ferenci, "Clonal adaptive radiation in a constant environment," *Science (New York, N.Y.)*, vol. 313, no. 5786, pp. 514–517, 2006. DOI: 10.1126/science.1129865.

[103] R. P. Maharjan, T. Ferenci, P. R. Reeves, Y. Li, B. Liu, and L. Wang, "The multiplicity of divergence mechanisms in a single evolving population," *Genome Biology*, vol. 13, no. 6, R41, 2012. DOI: 10.1186/gb-2012-13-6-r41.

[104] R. E. Lenski, M. R. Rose, S. C. Simpson, and S. C. Tadler, "Long-term experimental evolution in *Escherichia coli*. I. adaptation and divergence during 2,000 generations," *The American Naturalist*, vol. 138, no. 6, pp. 1315–1341, 1991. DOI: 10.2307/2462549.

[105] D. Ebert, C. Haag, M. Kirkpatrick, M. Riek, J. W. Hottinger, and V. I. Pajunen, "A selective advantage to immigrant genes in a daphnia metapopulation," *Science (New York, N.Y.)*, vol. 295, no. 5554, pp. 485–488, 2002. DOI: 10.1126/science.1067485.

[106] M. R. Rose, "Artificial selection on a fitness-component in *Drosophila melanogaster*," *Evolution*, vol. 38, no. 3, pp. 516–526, 1984. DOI: 10.1111/j.1558-5646.1984.tb00317.x.

[107] M. R. Rose, H. B. Passananti, and M. Matos, *Methuselah flies: A case study in the evoultion of aging*. New Jersey ; London: World Scientific Pub, 2004, 479 pp., OCLC: ocm57638264.

[108] C. T. Brown, L. K. Fishwick, B. M. Chokshi, M. A. Cuff, J. M. Jackson, T. Oglesby, A. T. Rioux, E. Rodriguez, G. S. Stupp, A. H. Trupp, J. S. Woollcombe-Clarke, T. N. Wright, W. J. Zaragoza, J. C. Drew, E. W. Triplett, and W. L. Nicholson, "Whole-genome sequencing and phenotypic analysis of *Bacillus subtilis* mutants following evolution under conditions of relaxed selection for sporulation," *Applied and Environmental Microbiology*, vol. 77, no. 19, pp. 6867–6877, 2011. DOI: 10.1128/AEM.05272-11.

[109] E. F. Mao, L. Lane, J. Lee, and J. H. Miller, "Proliferation of mutators in a cell population," *Journal of Bacteriology*, vol. 179, no. 2, pp. 417–422, 1997.

[110] M. Lynch, W. Sung, K. Morris, N. Coffey, C. R. Landry, E. B. Dopman, W. J. Dickinson, K. Okamoto, S. Kulkarni, D. L. Hartl, and W. K. Thomas, "A genome-wide view of the spectrum of spontaneous mutations in yeast," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 27, pp. 9272–9277, 2008. DOI: 10.1073/pnas.0803466105.

[111] N. J. Croucher, S. R. Harris, L. Barquist, J. Parkhill, and S. D. Bentley, "A high-resolution view of genome-wide pneumococcal transformation," *PLOS Pathogens*, vol. 8, no. 6, X. Didelot, Ed., e1002745, 2012. DOI: 10.1371/journal.ppat.1002745.

[112] J. C. Mell, J. Y. Lee, M. Firme, S. Sinha, and R. J. Redfield, "Extensive cotransformation of natural variation into chromosomes of naturally competent *Haemophilus influenzae*," *Genes|Genomes|Genetics*, vol. 4, no. 4, pp. 717–731, 2014. DOI: 10.1534/g3.113.009597.

[113] S. Kulick, C. Moccia, X. Didelot, D. Falush, C. Kraft, and S. Suerbaum, "Mosaic DNA imports with interspersions of recipient sequence after natural transformation of *Helicobacter pylori*," *PLOS ONE*, vol. 3, no. 11, N. Ahmed, Ed., e3797, 2008. DOI: 10.1371/journal.pone.0003797.

[114] R. E. Lenski, "Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations," *The ISME Journal*, vol. 11, no. 10, pp. 2181–2194, 2017. DOI: 10.1038/ismej.2017.69.

[115] T. J. Kawecki, R. E. Lenski, D. Ebert, B. Hollis, I. Olivieri, and M. C. Whitlock, "Experimental evolution," *Trends in Ecology & Evolution*, vol. 27, no. 10, pp. 547–560, 2012. DOI: 10.1016/j.tree.2012.06.001.

[116] J. E. Barrick and R. E. Lenski, "Genome dynamics during experimental evolution," *Nature Reviews Genetics*, vol. 14, no. 12, pp. 827–839, 2013. DOI: 10.1038/nrg3564.

[117] T. Giraud, B. Koskella, and A.-L. Laine, "Introduction: Microbial local adaptation: Insights from natural populations, genomics and experimental evolution," *Molecular Ecology*, vol. 26, no. 7, pp. 1703–1710, 2017. DOI: 10.1111/mec.14091.

[118] V. S. Cooper, "Experimental evolution as a high-throughput screen for genetic adaptations," *MSphere*, vol. 3, no. 3, A. C. Gales, Ed., 2018. DOI: 10.1128/mSphere.00121-18.

[119] J. S. (Stuart) Barker, "Defining fitness in natural and domesticated populations," in *Adaptation and Fitness in Animal Populations*, J. van der Werf, H.-U. Graser, R. Frankham, and C. Gondro, Eds., Dordrecht: Springer Netherlands, 2009, pp. 3–14. DOI: 10.1007/978-1-4020-9005-9_1.

[120] J. B. S. Haldane, *The causes of evolution*, ser. Princeton science library. Princeton, N.J: Princeton University Press, 1990, 222 pp.

[121] T. Dobzhansky, *Genetics and the origin of species*, ser. Columbia classics in evolution series. New York: Columbia University Press, 1982, 364 pp.

[122] T. Dobzhansky, "A review of some fundamental concepts and problems of population genetics," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 20, pp. 1–15, 1955. DOI: 10.1101/SQB.1955.020.01.003.

[123] M. A. Nowak, *Evolutionary dynamics: Exploring the equations of life*. Cambridge, Mass: Belknap Press of Harvard University Press, 2006, 363 pp.

[124] M. Lässig, "From biophysics to evolutionary genetics: Statistical aspects of gene regulation," *BMC Bioinformatics*, vol. 8, S7, Suppl 6 2007. DOI: 10.1186/1471-2105-8-S6-S7.

[125] T. R. Malthus, *An Essay on the Principle of Population*. London: J. Johnson, in St. Paul's Church-yard, 1798.

[126] J. F. Crow and M. Kimura, *An introduction to population genetics theory*. Jodhpur; New Jersey: Scientific Publisher (India) : The Blackburn Press, 2010, OCLC: 1027901624.

[127] R. J. Nichols, S. Sen, Y. J. Choo, P. Beltrao, M. Zietek, R. Chaba, S. Lee, K. M. Kazmierczak, K. J. Lee, A. Wong, M. Shales, S. Lovett, M. E. Winkler, N. J. Krogan, A. Typas, and C. A. Gross, "Phenotypic landscape of a bacterial cell," *Cell*, vol. 144, no. 1, pp. 143–156, 2011. DOI: 10.1016/j.cell.2010.11.052.

[128] T. Bollenbach, S. Quan, R. Chait, and R. Kishony, "Nonoptimal microbial response to antibiotics underlies suppressive drug interactions," *Cell*, vol. 139, no. 4, pp. 707–718, 2009. DOI: 10.1016/j.cell.2009.10.025.

[129] F. Vasi, M. Travisano, and R. E. Lenski, "Long-term experimental evolution in *Escherichia coli*. II. changes in life-history traits during adaptation to a seasonal environment," *The American Naturalist*, vol. 144, no. 3, pp. 432–456, 1994. DOI: 10.1086/285685.

[130] M. J. Wiser and R. E. Lenski, "A comparison of methods to measure fitness in *Escherichia coli*," *PLOS One*, vol. 10, no. 5, e0126210, 2015. DOI: 10.1371/journal.pone.0126210.

[131] L. Stannek, R. Egelkamp, K. Gunka, and F. M. Commichau, "Monitoring intraspecies competition in a bacterial cell population by cocultivation of fluorescently labelled strains," *Journal of Visualized Experiments: JoVE*, no. 83, e51196, 2014. DOI: 10.3791/51196.

[132] J. C. Garcia-Betancur, A. Yepes, J. Schneider, and D. Lopez, "Single-cell analysis of *Bacillus subtilis* biofilms using fluorescence microscopy and flow cytometry," *Journal of Visualized Experiments: JoVE*, no. 60, 2012. DOI: 10.3791/3796.

[133] L.-M. Chevin, "On measuring selection in experimental evolution," *Biology Letters*, vol. 7, no. 2, pp. 210–213, 2010. DOI: 10.1098/rsbl.2010.0580.

[134] B. Cerulus, A. M. New, K. Pougach, and K. J. Verstrepen, "Noise and epigenetic inheritance of single-cell division times influence population fitness," *Current biology: CB*, vol. 26, no. 9, pp. 1138–1147, 2016. DOI: 10.1016/j.cub.2016.03.010.

[135] T. Miyashiro and M. Goulian,
"Single cell analysis of gene expression by fluorescence microscopy," in *Methods in Enzymology*, vol. 423, Elsevier, 2007, pp. 458–475. DOI: 10.1016/S0076-6879(07)23022-8.

[136] D. Cottinet, F. Condamine, N. Bremond, A. D. Griffiths, P. B. Rainey, J. A. G. M. de Visser, J. Baudry, and J. Bibette, "Lineage tracking for probing heritable phenotypes at single-cell resolution," *PLOS ONE*, vol. 11, no. 4, S. Rutherford, Ed., e0152395, 2016. DOI: 10.1371/journal.pone.0152395.

[137] M. Ackermann, "A functional perspective on phenotypic heterogeneity in microorganisms," *Nature Reviews Microbiology*, vol. 13, no. 8, pp. 497–508, 2015. DOI: 10.1038/nrmicro3491.

[138] J. W. Young, J. C. W. Locke, A. Altinok, N. Rosenfeld, T. Bacarian, P. S. Swain, E. Mjolsness, and M. B. Elowitz, "Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy," *Nature Protocols*, vol. 7, no. 1, pp. 80–88, 2011. DOI: 10.1038/nprot.2011.432.

[139] G. Ullman, M. Wallden, E. G. Marklund, A. Mahmutovic, I. Razinkov, and J. Elf, "High-throughput gene expression analysis at the level of single proteins using a microfluidic turbidostat and automated cell tracking," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 368, no. 1611, p. 20 120 025, 2013. DOI: 10.1098/rstb.2012.0025.

[140] M. Kaiser, F. Jug, T. Julou, S. Deshpande, T. Pfohl, O. K. Silander, G. Myers, and E. van Nimwegen, "Monitoring single-cell gene regulation under dynamically controllable conditions with integrated microfluidics and software," *Nature Communications*, vol. 9, no. 1, p. 212, 2018. DOI: 10.1038/s41467-017-02505-0.

[141] Ö. Baltekin, A. Boucharin, E. Tano, D. I. Andersson, and J. Elf, "Antibiotic susceptibility testing in less than 30 min using direct single-cell imaging," *Proceedings of the National Academy of Sciences*, vol. 114, no. 34, pp. 9170–9175, 2017. DOI: 10.1073/pnas.1708558114.

[142] P. Wang, L. Robert, J. Pelletier, W. L. Dang, F. Taddei, A. Wright, and S. Jun, "Robust growth of *Escherichia coli*," *Current Biology*, vol. 20, no. 12, pp. 1099–1103, 2010. DOI: 10.1016/j.cub.2010.04.045.

[143] P. C. Phillips, "Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems," *Nature Reviews. Genetics*, vol. 9, no. 11, pp. 855–867, 2008. DOI: 10.1038/nrg2452.

[144] S. Gavrilets, *Fitness landscapes and the origin of species*, ser. Monographs in population biology 41. Princeton, N.J: Princeton University Press, 2004, 476 pp., OCLC: ocm56023319.

[145] J. A. G. de Visser and J. Krug, "Empirical fitness landscapes and the predictability of evolution," *Nature Reviews Genetics*, vol. 15, no. 7, pp. 480–490, 2014. DOI: 10.1038/nrg3744.

[146] S. Wright, "Evolution in mendelian populations," *Genetics*, vol. 16, no. 2, pp. 97–159, 1931.

[147] ——, "The roles of mutation, inbreeding, crossbreeding and selection in evolution," *Proceedings of the Sixth International Congress on Genetics*, vol. 1, no. 6, pp. 356–366, 1932.

[148] J. M. Smith, "Natural selection and the concept of a protein space," *Nature*, vol. 225, no. 5232, pp. 563–564, 1970.

[149] J. A. G. M. de Visser, T. F. Cooper, and S. F. Elena, "The causes of epistasis," *Proceedings. Biological Sciences*, vol. 278, no. 1725, pp. 3617–3624, 2011. DOI: 10.1098/rspb.2011.1537.

[150] J. B. Wolf, E. D. Brodie, and M. J. Wade, "Epistasis and the evolutionary process," in *Epistasis and the Evolutionary Process*, New York: Oxford University Press, 2000.

[151] P. C. Phillips, "The language of gene interaction," *Genetics*, vol. 149, no. 3, pp. 1167–1171, 1998.

[152] J. Franke, A. Klözer, J. A. G. M. de Visser, and J. Krug, "Evolutionary accessibility of mutational pathways," *PLOS Computational Biology*, vol. 7, no. 8, C. O. Wilke, Ed., e1002134, 2011. DOI: 10.1371/journal.pcbi.1002134.

[153] X. He, W. Qian, Z. Wang, Y. Li, and J. Zhang, "Prevalent positive epistasis in *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks," *Nature Genetics*, vol. 42, no. 3, pp. 272–276, 2010. DOI: 10.1038/ng.524.

[154] E. R. Lozovsky, T. Chookajorn, K. M. Brown, M. Imwong, P. J. Shaw, S. Kamchonwongpaisan, D. E. Neafsey, D. M. Weinreich, and D. L. Hartl, "Stepwise acquisition of pyrimethamine resistance in the malaria parasite," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 29, pp. 12 025–12 030, 2009. DOI: 10.1073/pnas.0905922106.

[155] D. M. Weinreich, R. A. Watson, and L. Chao, "Perspective: Sign epistasis and genetic constraint on evolutionary trajectories," *Evolution; International Journal of Organic Evolution*, vol. 59, no. 6, pp. 1165–1174, 2005.

[156] J. E. Barrick, D. S. Yu, S. H. Yoon, H. Jeong, T. K. Oh, D. Schneider, R. E. Lenski, and J. F. Kim, "Genome evolution and adaptation in a long-term experiment with *Escherichia coli*," *Nature*, vol. 461, no. 7268, pp. 1243–1247, 2009. DOI: 10.1038/nature08480.

[157] J. R. Meyer, D. T. Dobias, J. S. Weitz, J. E. Barrick, R. T. Quick, and R. E. Lenski, "Repeatability and contingency in the evolution of a key innovation in phage lambda," *Science (New York, N.Y.)*, vol. 335, no. 6067, pp. 428–432, 2012. DOI: 10.1126/science.1214449.

[158] R. Lenski. (). The e. coli long-term experimental evolution project site, [Online]. Available: http://myxo.css.msu.edu/ecoli (visited on 07/27/2018).

[159] R. Maddamsetti, P. J. Hatcher, A. G. Green, B. L. Williams, D. S. Marks, and R. E. Lenski, "Core genes evolve rapidly in the long-term evolution experiment with *Escherichia coli*," *Genome Biology and Evolution*, vol. 9, no. 4, pp. 1072–1083, 2017. DOI: 10.1093/gbe/evx064.

[160] O. Tenaillon, J. E. Barrick, N. Ribeck, D. E. Deatherage, J. L. Blanchard, A. Dasgupta, G. C. Wu, S. Wielgoss, S. Cruveiller, C. Medigue, D. Schneider, and R. E. Lenski, "Tempo and mode of genome evolution in a 50,000-generation experiment," *Nature*, vol. 536, no. 7615, pp. 165–170, 2016. DOI: 10.1038/nature18959.

[161] S. Bershtein, A. W. R. Serohijos, S. Bhattacharyya, M. Manhart, J.-M. Choi, W. Mu, J. Zhou, and E. I. Shakhnovich, "Protein homeostasis imposes a barrier on functional integration of horizontally transferred genes in bacteria,"
PLOS Genetics, vol. 11, no. 10, M. Achtman, Ed., e1005612, 2015.
DOI: 10.1371/journal.pgen.1005612.

[162] D. J. P. Engelmoer, I. Donaldson, and D. E. Rozen, "Conservative sex and the benefits of transformation in *Streptococcus pneumoniae*,"
PLOS Pathogens, vol. 9, no. 11, D. Dykhuizen, Ed., e1003758, 2013.
DOI: 10.1371/journal.ppat.1003758.

[163] S. Bubendorfer, J. Krebes, I. Yang, E. Hage, T. F. Schulz, C. Bahlawane, X. Didelot, and S. Suerbaum, "Genome-wide analysis of chromosomal import patterns after natural transformation of *Helicobacter pylori*," *Nature Communications*, vol. 7, p. 11 995, 2016.
DOI: 10.1038/ncomms11995.

[164] J. C. Mell, S. Shumilina, I. M. Hall, and R. J. Redfield, "Transformation of natural genetic variation into *Haemophilus influenzae* genomes,"
PLOS Pathogens, vol. 7, no. 7, D. S. Guttman, Ed., e1002151, 2011.
DOI: 10.1371/journal.ppat.1002151.

[165] M. Yüksel, J. J. Power, J. Ribbe, T. Volkmann, and B. Maier, "Fitness trade-offs in competence differentiation of *Bacillus subtilis*," *Frontiers in Microbiology*, vol. 7, 2016.
DOI: 10.3389/fmicb.2016.00888.

[166] D. R. Zeigler, "The genome sequence of *Bacillus subtilis* subsp. *spizizenii W23*: Insights into speciation within the *B. subtilis* complex and into the history of *B. subtilis* genetics,"
Microbiology, vol. 157, no. 7, pp. 2033–2041, 2011. DOI: 10.1099/mic.0.048520-0.

[167] M. Albano, J. Hahn, and D. Dubnau, "Expression of competence genes in *Bacillus subtilis*,"
Journal of Bacteriology, vol. 169, no. 7, pp. 3110–3117, 1987. JSTOR: {PMC}212357.

[168] S. Andrew. (2010). Fastqc: A quality control tool for high throughput sequence data,
[Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[169] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.
DOI: 10.1093/bioinformatics/btu170.

[170] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows-wheeler transform,"
Bioinformatics, vol. 25, no. 14, pp. 1754–1760, 2009.
DOI: 10.1093/bioinformatics/btp324.

[171] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.
DOI: 10.1093/bioinformatics/btp352.

[172] broadinstitute. (2015). Picard,
[Online]. Available: http://broadinstitute.github.io/picard.

[173] A. R. Quinlan and I. M. Hall, "Bedtools: A flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.
DOI: 10.1093/bioinformatics/btq033.

[174] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden, "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w ^{1118} ; iso-2; iso-3," *Fly*, vol. 6, no. 2, pp. 80–92, 2012. DOI: 10.4161/fly.19695.

[175] B. R. Belitsky and A. L. Sonenshein, "An enhancer element located downstream of the major glutamate dehydrogenase gene of *Bacillus subtilis*,"
Proceedings of the National Academy of Sciences, vol. 96, no. 18, pp. 10 290–10 295, 1999.
DOI: 10.1073/pnas.96.18.10290.

[176] L. J. Reitzer and B. Magasanik, "Transcription of *glnA* in *E. coli* is stimulated by activator bound to sites far from the promoter," *Cell*, vol. 45, no. 6, pp. 785–792, 1986.
DOI: 10.1016/0092-8674(86)90553-2.

[177] J. P. Claverys and S. A. Lacks, "Heteroduplex deoxyribonucleic acid base mismatch repair in bacteria," *Microbiological Reviews*, vol. 50, no. 2, pp. 133–165, 1986.

[178] P. Nghe, S. Boulineau, S. Gude, P. Recouvreux, J. S. van Zon, and S. J. Tans, "Microfabricated polyacrylamide devices for the controlled culture of growing cells and developing organisms,"
PLOS ONE, vol. 8, no. 9, A. J. Engler, Ed., e75537, 2013.
DOI: 10.1371/journal.pone.0075537.

[179] A. Ducret, E. Maisonneuve, P. Notareschi, A. Grossi, T. Mignot, and S. Dukan, "A microscope automated fluidic system to study bacterial processes in real time,"
PLOS ONE, vol. 4, no. 9, C. Herman, Ed., e7282, 2009.
DOI: 10.1371/journal.pone.0007282.

[180] S. Boulineau, F. Tostevin, D. J. Kiviet, P. R. ten Wolde, P. Nghe, and S. J. Tans, "Single-cell dynamics reveals sustained growth during diauxic shifts,"
PLOS ONE, vol. 8, no. 4, C. Herman, Ed., e61686, 2013.
DOI: 10.1371/journal.pone.0061686.

[181] T. Julou, T. Mora, L. Guillon, V. Croquette, I. J. Schalk, D. Bensimon, and N. Desprat, "Cell-cell contacts confine public goods diffusion inside *Pseudomonas aeruginosa* clonal microcolonies,"
Proceedings of the National Academy of Sciences, vol. 110, no. 31, pp. 12 577–12 582, 2013.
DOI: 10.1073/pnas.1301428110.

[182] S. M. Anthony, M. Kim, and S. Granick, "Single-particle tracking of janus colloids in close proximity," *Langmuir*, vol. 24, no. 13, pp. 6557–6561, 2008. DOI: 10.1021/la800424t.

[183] N. Otsu, "A threshold selection method from gray-level histograms,"
IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, 1979.
DOI: 10.1109/TSMC.1979.4310076.

[184] R. D. De Veaux, P. F. Velleman, and D. Bock, *Stats: Data and models*.
Boston: Pearson/Addison-Wesley, 2008, OCLC: 783401964.

[185] L. Y. Geer, A. Marchler-Bauer, R. C. Geer, L. Han, J. He, S. He, C. Liu, W. Shi, and S. H. Bryant, "The NCBI BioSystems database,"
Nucleic Acids Research, vol. 38, pp. D492–496, Database issue 2010.
DOI: 10.1093/nar/gkp858.

[186] X. Mao, Q. Ma, C. Zhou, X. Chen, H. Zhang, J. Yang, F. Mao, W. Lai, and Y. Xu, "DOOR 2.0: Presenting operons and their functions through dynamic and integrated views,"
Nucleic Acids Research, vol. 42, pp. D654–D659, D1 2014. DOI: 10.1093/nar/gkt1048.

[187] R. H. Michna, B. Zhu, U. Mäder, and J. Stülke, "Subti wiki 2.0—an integrated database for the model organism *Bacillus subtilis*," *Nucleic Acids Research*, vol. 44, pp. D654–D662, D1 2016. DOI: 10.1093/nar/gkv1006.

[188] M. A. Konkol, K. M. Blair, and D. B. Kearns, "Plasmid-encoded ComI inhibits competence in the ancestral 3610 strain of *Bacillus subtilis*," *Journal of Bacteriology*, vol. 195, no. 18, pp. 4085–4093, 2013. DOI: 10.1128/JB.00696-13.

[189] P. K. Singh, G. Ramachandran, L. Durán-Alcalde, C. Alonso, L. J. Wu, and W. J. J. Meijer, "Inhibition of *Bacillus subtilis* natural competence by a native, conjugative plasmid-encoded *comK* repressor protein: *Bacillus* plasmid encodes competence inhibitor," *Environmental Microbiology*, vol. 14, no. 10, pp. 2812–2825, 2012. DOI: 10.1111/j.1462-2920.2012.02819.x.

[190] G. M. Süel, J. Garcia-Ojalvo, L. M. Liberman, and M. B. Elowitz, "An excitable gene regulatory circuit induces transient cellular differentiation," *Nature*, vol. 440, no. 7083, pp. 545–550, 2006. DOI: 10.1038/nature04588.

[191] D. Dubnau and C. Cirigliano, "Fate of transforming deoxyribonucleic acid after uptake by competent *Bacillus subtilis*: Size and distribution of the integrated donor segments," *Journal of Bacteriology*, vol. 111, no. 2, pp. 488–494, 1972. JSTOR: {PMC}251309.

[192] S. L. Fornili and M. S. Fox, "Electron microscope visualization of the products of bacillus subtilis transformation," *Journal of Molecular Biology*, vol. 113, no. 1, pp. 181–191, 1977.

[193] D. A. Morrison and W. R. Guild, "Activity of deoxyribonucleic acid fragments of defined size in *Bacillus subtilis* transformation," *Journal of Bacteriology*, vol. 112, no. 1, pp. 220–223, 1972.

[194] Y. Weinrauch and D. Dubnau, "Plasmid marker rescue transformation in *Bacillus subtilis*," *Journal of Bacteriology*, vol. 154, no. 3, pp. 1077–1087, 1983.

[195] B. Maier, I. Chen, D. Dubnau, and M. P. Sheetz, "DNA transport into *Bacillus subtilis* requires proton motive force to generate large molecular forces," *Nature Structural & Molecular Biology*, vol. 11, no. 7, pp. 643–649, 2004. DOI: 10.1038/nsmb783.

[196] O. Zafra, M. Lamprecht-Grandío, C. G. de Figueras, and J. E. González-Pastor, "Extracellular DNA release by undomesticated *Bacillus subtilis* is regulated by early competence," *PLOS ONE*, vol. 7, no. 11, T. Msadek, Ed., e48716, 2012. DOI: 10.1371/journal.pone.0048716.

[197] F. Pinheiro, "A fitness model for lateral gene transfer," SFB680 Workshop, Köln, Germany, 2016.

[198] A. L. Forget and S. C. Kowalczykowski, "Single-molecule imaging of DNA pairing by RecA reveals a three-dimensional homology search," *Nature*, vol. 482, no. 7385, pp. 423–427, 2012. DOI: 10.1038/nature10782.

[199] M. S. Roberts and F. M. Cohan, "The effect of DNA sequence divergence on sexual isolation in *Bacillus*," *Genetics*, vol. 134, no. 2, pp. 401–408, 1993.

[200] J. Majewski, P. Zawadzki, P. Pickerill, F. M. Cohan, and C. G. Dowson, "Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation," *Journal of Bacteriology*, vol. 182, no. 4, pp. 1016–1023, 2000. DOI: 10.1128/JB.182.4.1016-1023.2000.

[201] J. Majewski and F. M. Cohan, "The effect of mismatch repair and heteroduplex formation on sexual isolation in *Bacillus*," *Genetics*, vol. 148, no. 1, p. 13, 1998.

[202] J. Majewski and F. M. Cohan, "DNA sequence similarity requirements for interspecific recombination in *Bacillus*," *Genetics*, vol. 153, no. 4, pp. 1525–1533, 1999.

[203] G. R. Smith, "Homologous recombination in prokaryotes: Enzymes and controlling sites," *Genome*, vol. 31, no. 2, pp. 520–527, 1989.

[204] H.-H. Chou, H.-C. Chiu, N. F. Delaney, D. Segre, and C. J. Marx, "Diminishing returns epistasis among beneficial mutations decelerates adaptation," *Science*, vol. 332, no. 6034, pp. 1190–1192, 2011. DOI: 10.1126/science.1203799.

[205] A. I. Khan, D. M. Dinh, D. Schneider, R. E. Lenski, and T. F. Cooper, "Negative epistasis between beneficial mutations in an evolving bacterial population," *Science*, vol. 332, no. 6034, pp. 1193–1196, 2011. DOI: 10.1126/science.1203801.

[206] A. Wong, "Epistasis and the evolution of antimicrobial resistance," *Frontiers in Microbiology*, vol. 8, 2017. DOI: 10.3389/fmicb.2017.00246.

[207] S. F. Elena and R. E. Lenski, "Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation," *Nature Reviews. Genetics*, vol. 4, no. 6, pp. 457–469, 2003. DOI: 10.1038/nrg1088.

[208] A. Couce and O. A. Tenaillon, "The rule of declining adaptability in microbial evolution experiments," *Frontiers in Genetics*, vol. 6, p. 99, 2015. DOI: 10.3389/fgene.2015.00099.

[209] S. Schoustra, S. Hwang, J. Krug, and J. A. G. M. de Visser, "Diminishing-returns epistasis among random beneficial mutations in a multicellular fungus," *Proceedings of the Royal Society B: Biological Sciences*, vol. 283, no. 1837, p. 20 161 376, 2016. DOI: 10.1098/rspb.2016.1376.

[210] R. Gallegos-Monterrosa, E. Mhatre, and Á. T. Kovács, "Specific *Bacillus subtilis* 168 variants form biofilms on nutrient-rich medium," *Microbiology (Reading, England)*, vol. 162, no. 11, pp. 1922–1932, 2016. DOI: 10.1099/mic.0.000371.

[211] A. Mogk, D. Huber, and B. Bukau, "Integrating protein homeostasis strategies in prokaryotes," *Cold Spring Harbor Perspectives in Biology*, vol. 3, no. 4, a004366–a004366, 2011. DOI: 10.1101/cshperspect.a004366.

[212] E. Gur and R. T. Sauer, "Recognition of misfolded proteins by lon, a AAA+ protease," *Genes & Development*, vol. 22, no. 16, pp. 2267–2277, 2008. DOI: 10.1101/gad.1670908.

[213] ——, "Degrons in protein substrates program the speed and operating efficiency of the AAA+ lon proteolytic machine," *Proceedings of the National Academy of Sciences*, vol. 106, no. 44, pp. 18 503–18 508, 2009. DOI: 10.1073/pnas.0910392106.

[214] P. A. Lind, C. Tobin, O. G. Berg, C. G. Kurland, and D. I. Andersson, "Compensatory gene amplification restores fitness after inter-species gene replacements," *Molecular Microbiology*, vol. 75, no. 5, pp. 1078–1089, 2010. DOI: 10.1111/j.1365-2958.2009.07030.x.

[215] A. Dilthey and M. J. Lercher, "Horizontally transferred genes cluster spatially and metabolically," *Biology Direct*, vol. 10, no. 1, 2015. DOI: 10.1186/s13062-015-0102-5.

[216] M. Yüksel, "Role of the competent state in antibiotic persistence and resistance evolution of *Bacillus subtilis*," Doctoral Thesis, Universität zu Köln, 2018, 161 pp.

[217] C. Shee, J. L. Gibson, and S. M. Rosenberg, "Two mechanisms produce mutation hotspots at DNA breaks in *Escherichia coli*," *Cell Reports*, vol. 2, no. 4, pp. 714–721, 2012. DOI: 10.1016/j.celrep.2012.08.033.

[218] R. T. Pomerantz, M. F. Goodman, and M. E. O'Donnell, "DNA polymerases are error-prone at RecA-mediated recombination intermediates," *Cell Cycle*, vol. 12, no. 16, pp. 2558–2563, 2013. DOI: 10.4161/cc.25691.

[219] P. Graumann, Ed., *Bacillus: Cellular and molecular biology*, 2nd ed, Norfolk: Caister Academic Press, 2012, 397 pp.

[220] H. Tanooka, N. Munakata, and S. Kitahara, "Mutation induction with UV- and x-radiations in spores and vegetative cells of *Bacillus subtilis*," *Mutation Research*, vol. 49, no. 2, pp. 179–186, 1978.

[221] J. Slager, M. Kjos, L. Attaiech, and J.-W. Veening, "Antibiotic-induced replication stress triggers bacterial competence by increasing gene dosage near the origin," *Cell*, vol. 157, no. 2, pp. 395–406, 2014. DOI: 10.1016/j.cell.2014.01.068.

[222] N. Munakata, F. Morohoshi, M. Saitou, N. Yamazaki, and K. Hayashi, "Molecular characterization of thirteen *gyrA* mutations conferring nalidixic acid resistance in *Bacillus subtilis*," *MGG Molecular & General Genetics*, vol. 244, no. 1, pp. 97–103, 1994. DOI: 10.1007/BF00280192.

[223] A. Sugino and K. F. Bott, "*Bacillus subtilis* deoxyribonucleic acid gyrase," *Journal of Bacteriology*, vol. 141, no. 3, pp. 1331–1339, 1980.

[224] C. J. Ingham and P. A. Furneaux, "Mutations in the ss subunit of the *Bacillus subtilis* RNA polymerase that confer both rifampicin resistance and hypersensitivity to NusG," *Microbiology (Reading, England)*, vol. 146 Pt 12, pp. 3041–3049, 2000. DOI: 10.1099/00221287-146-12-3041.

[225] S. F. Levy, J. R. Blundell, S. Venkataram, D. A. Petrov, D. S. Fisher, and G. Sherlock, "Quantitative evolutionary dynamics using high-resolution lineage tracking," *Nature*, vol. 519, no. 7542, pp. 181–186, 2015. DOI: 10.1038/nature14279.

[226] S. Venkataram, B. Dunn, Y. Li, A. Agarwala, J. Chang, E. R. Ebel, K. Geiler-Samerotte, L. Hérissant, J. R. Blundell, S. F. Levy, D. S. Fisher, G. Sherlock, and D. A. Petrov, "Development of a comprehensive genotype-to-fitness map of adaptation-driving mutations in yeast," *Cell*, vol. 166, no. 6, 1585–1596.e22, 2016. DOI: 10.1016/j.cell.2016.08.002.

Appendix

Table A.1: CNP segment statistics for replicates W1, 3, 4 and 5 at final cycle

Replicate	Mean length[a] (bp)	Median length[a] (bp)
W1	3800 ± 425	2000 ± 450
W3	4800 ± 475	3600 ± 425
W4	3000 ± 225	1700 ± 225
W5	4250 ± 625	3100 ± 500

[a] Averaged over eight (W1), four (W3), or ten (W4, W5) time points

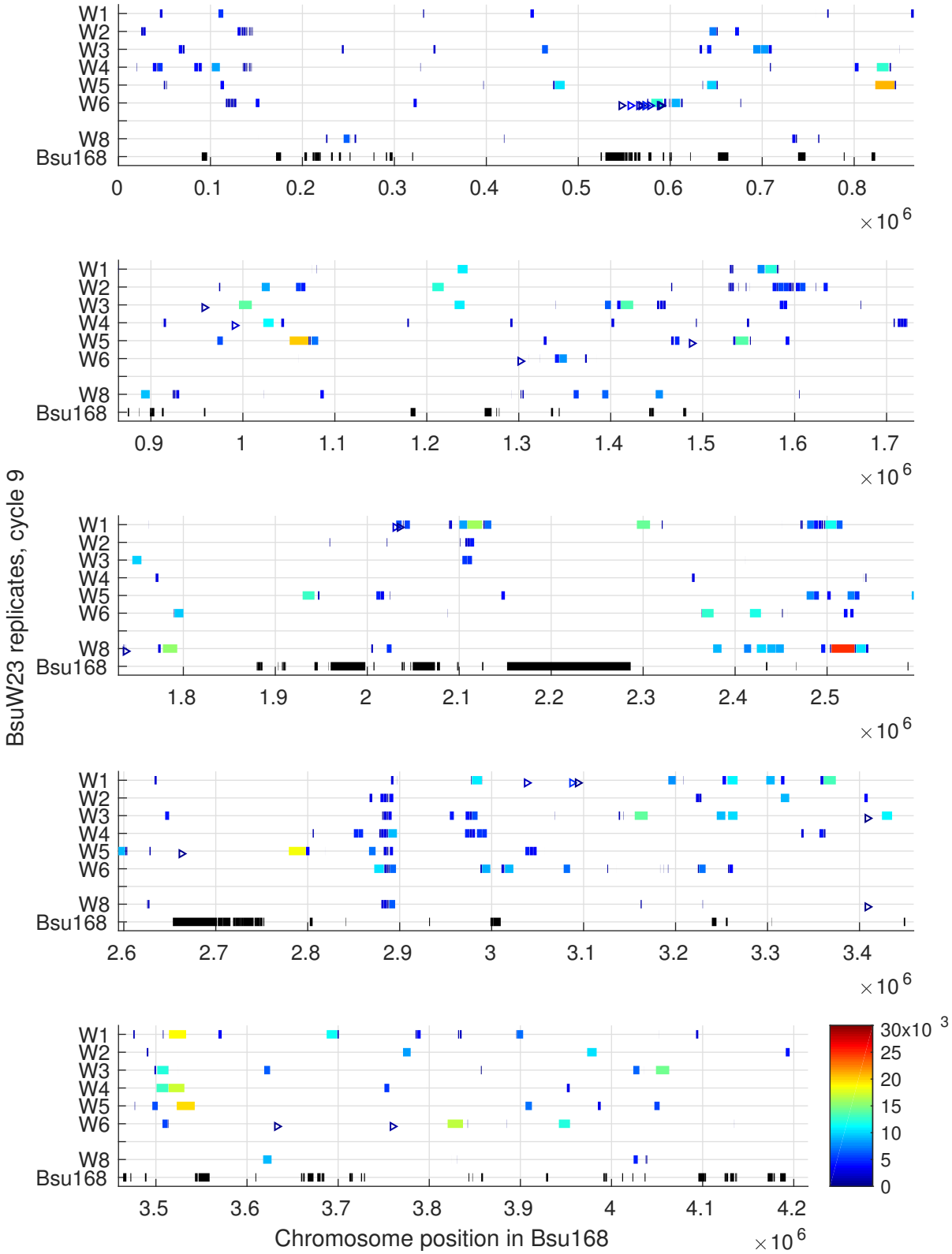


Figure A.1: Orthologous recombination for all BsuW23 replicates from cycle 9, as a function of chromosome position (replicate W7 has been removed). The start and end position of an orthologous recombination event (CNP) is denoted using filled boxes. The start of a horizontal gene transfer event is marked with an open triangle. All events are color coded to describe the average import length. Sample "Bsu168" shows Bsu168 auxiliary regions (black).

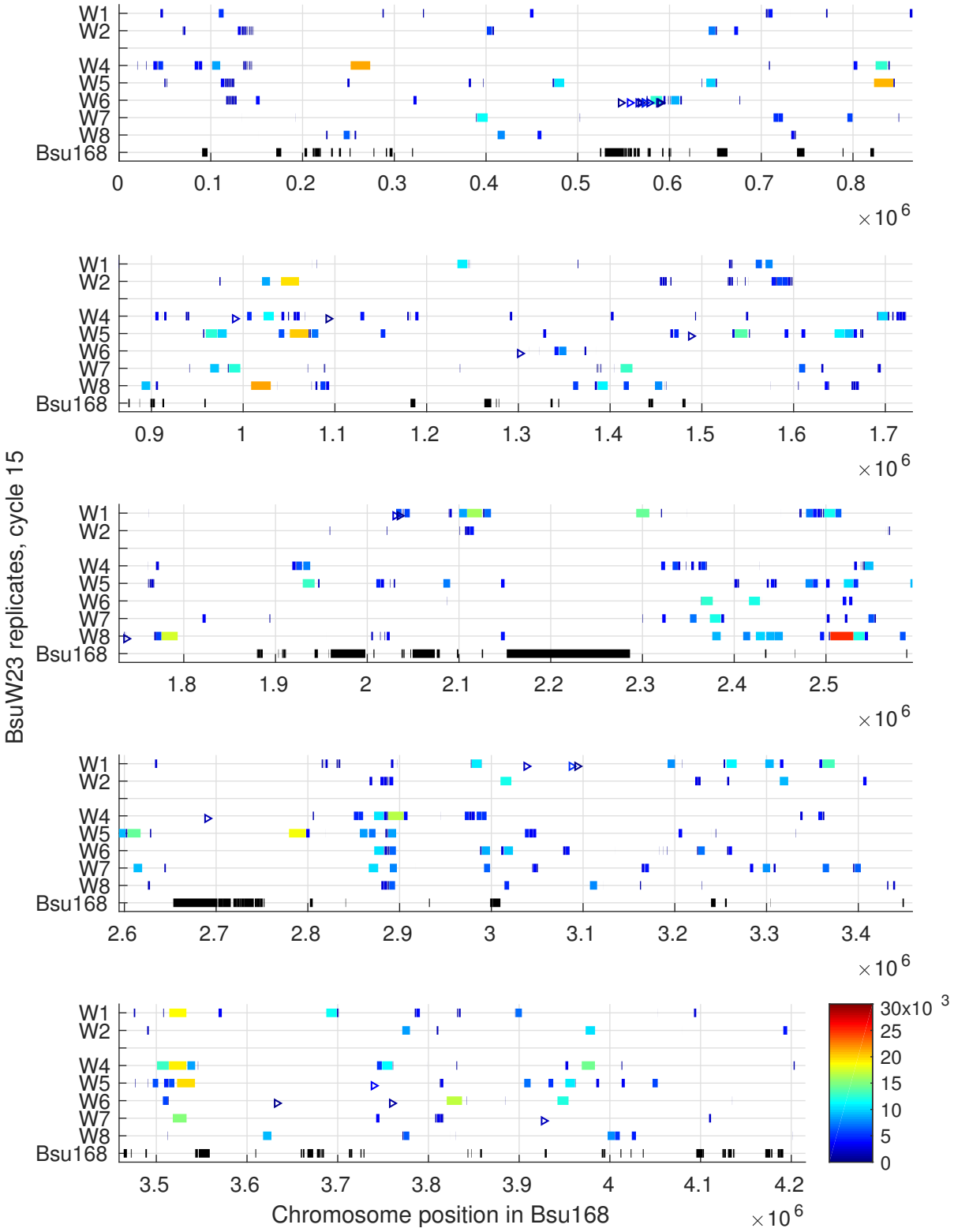


Figure A.2: Orthologous recombination for all BsuW23 replicates from cycle 15, as a function of chromosome position (replicate W3 has been removed). The start and end position of an orthologous recombination event (CNP) is denoted using filled boxes. The start of a horizontal gene transfer event is marked with an open triangle. All events are color coded to describe the average import length. Sample "Bsu168" shows Bsu168 auxiliary regions (black).

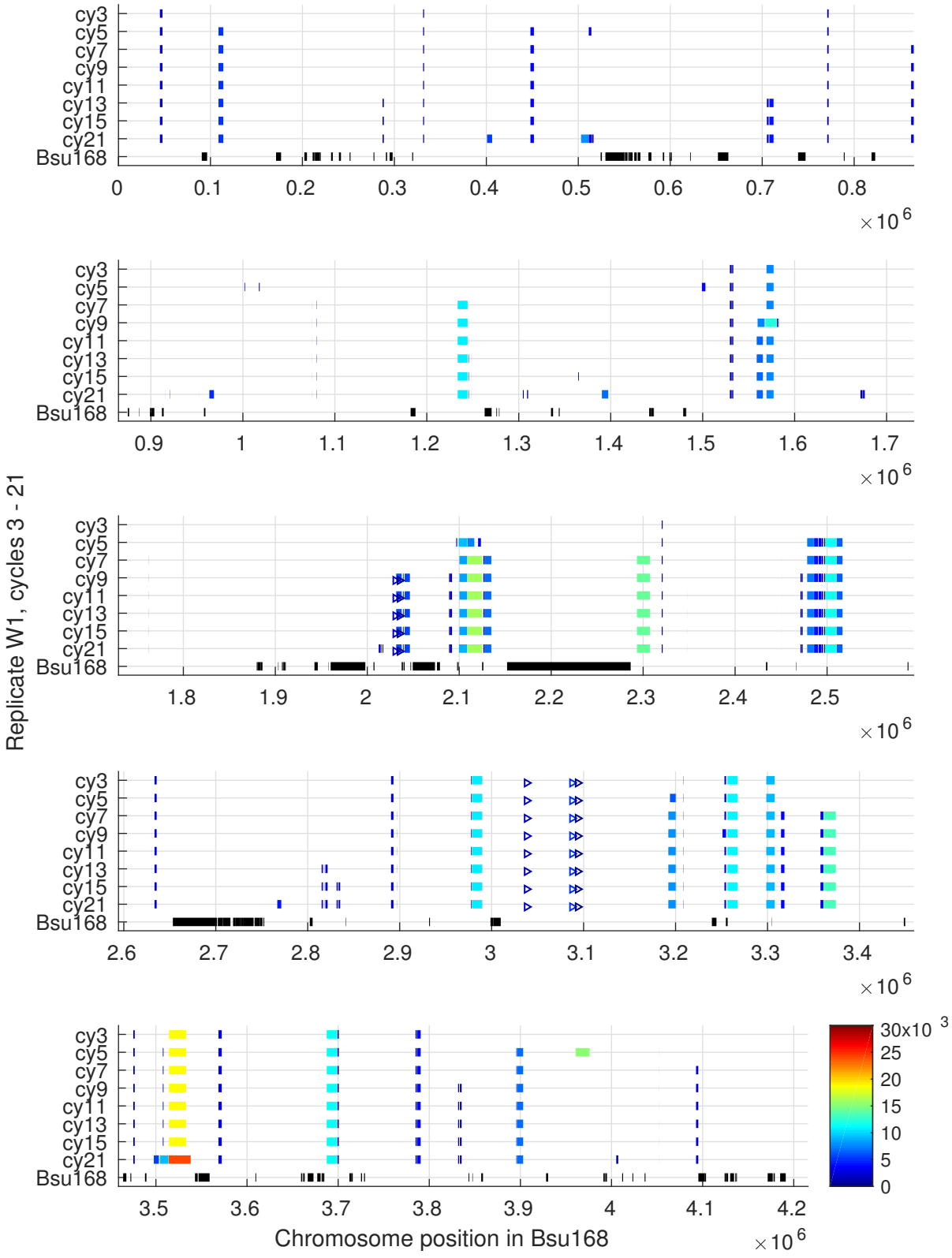


Figure A.3: Orthologous recombination for replicate W1 from various cycles as a function of chromosome position. The start and end position of an orthologous recombination event (CNP) is denoted using filled boxes. The start of a horizontal gene transfer event is marked with an open triangle. All events are color coded to describe the average import length. Sample "Bsu168" shows Bsu168 auxiliary regions (black).

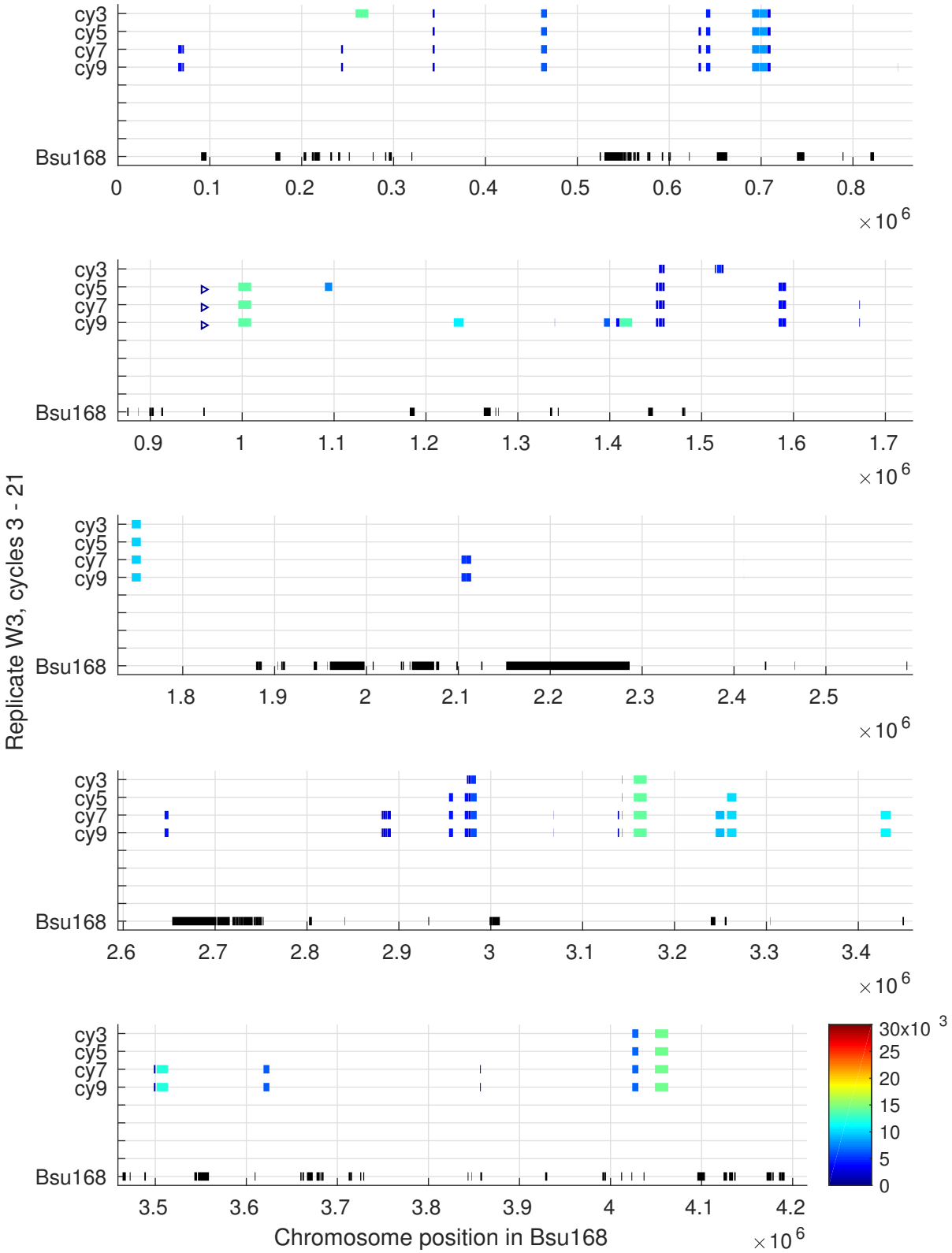


Figure A.4: Orthologous recombination for replicate W3 from various cycles as a function of chromosome position. The start and end position of an orthologous recombination event (CNP) is denoted using filled boxes. The start of a horizontal gene transfer event is marked with an open triangle. All events are color coded to describe the average import length. Sample "Bsu168" shows Bsu168 auxiliary regions (black).

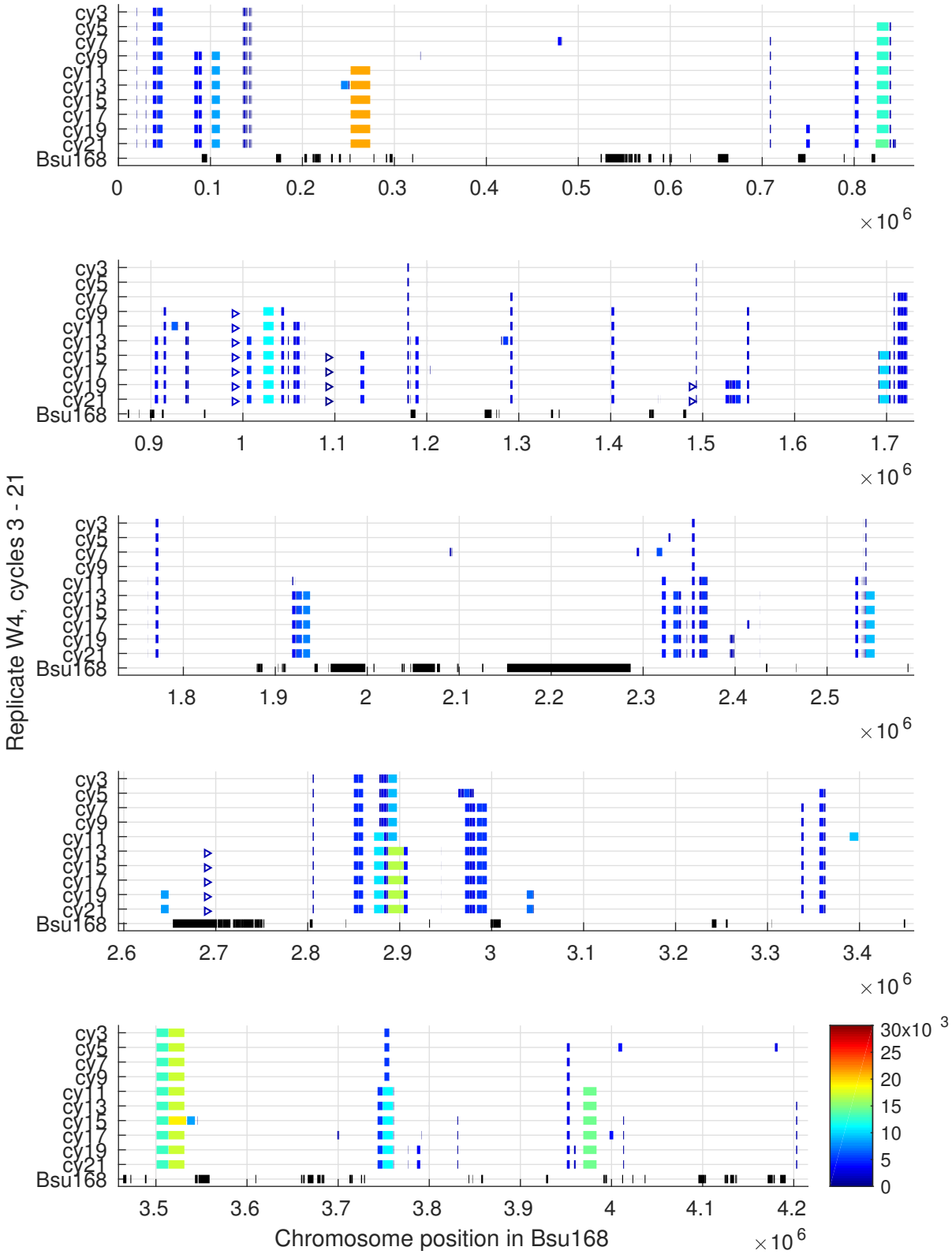


Figure A.5: Orthologous recombination for replicate W4 from various cycles as a function of chromosome position. The start and end position of an orthologous recombination event (CNP) is denoted using filled boxes. The start of a horizontal gene transfer event is marked with an open triangle. All events are color coded to describe the average import length. Sample "Bsu168" shows Bsu168 auxiliary regions (black).

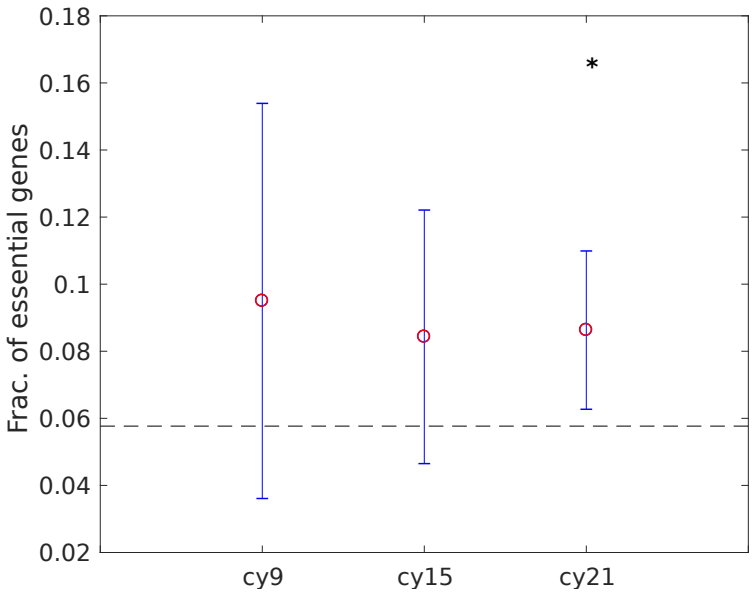


Figure A.6: Fraction of essential genes replaced at cycle 9, 15, and 21, normalized to the total number of genes within a CNP. Circles (red) mark the mean value for seven replicates, and the error bars (blue) the standard deviation. The dashed line (black) is the expected value. $*p < 0.01$

(a) Cycle 9

(b) Cycle 15

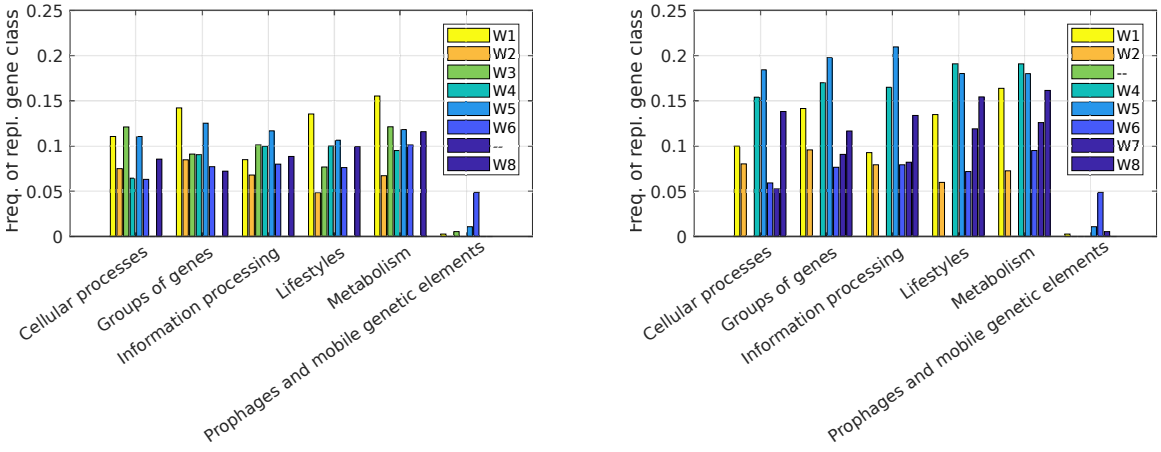


Figure A.7: CNP segment gene types, cycle 9 (a) and 15 (b). Counts have been normalized by the number of time each gene type occurs in the genome.

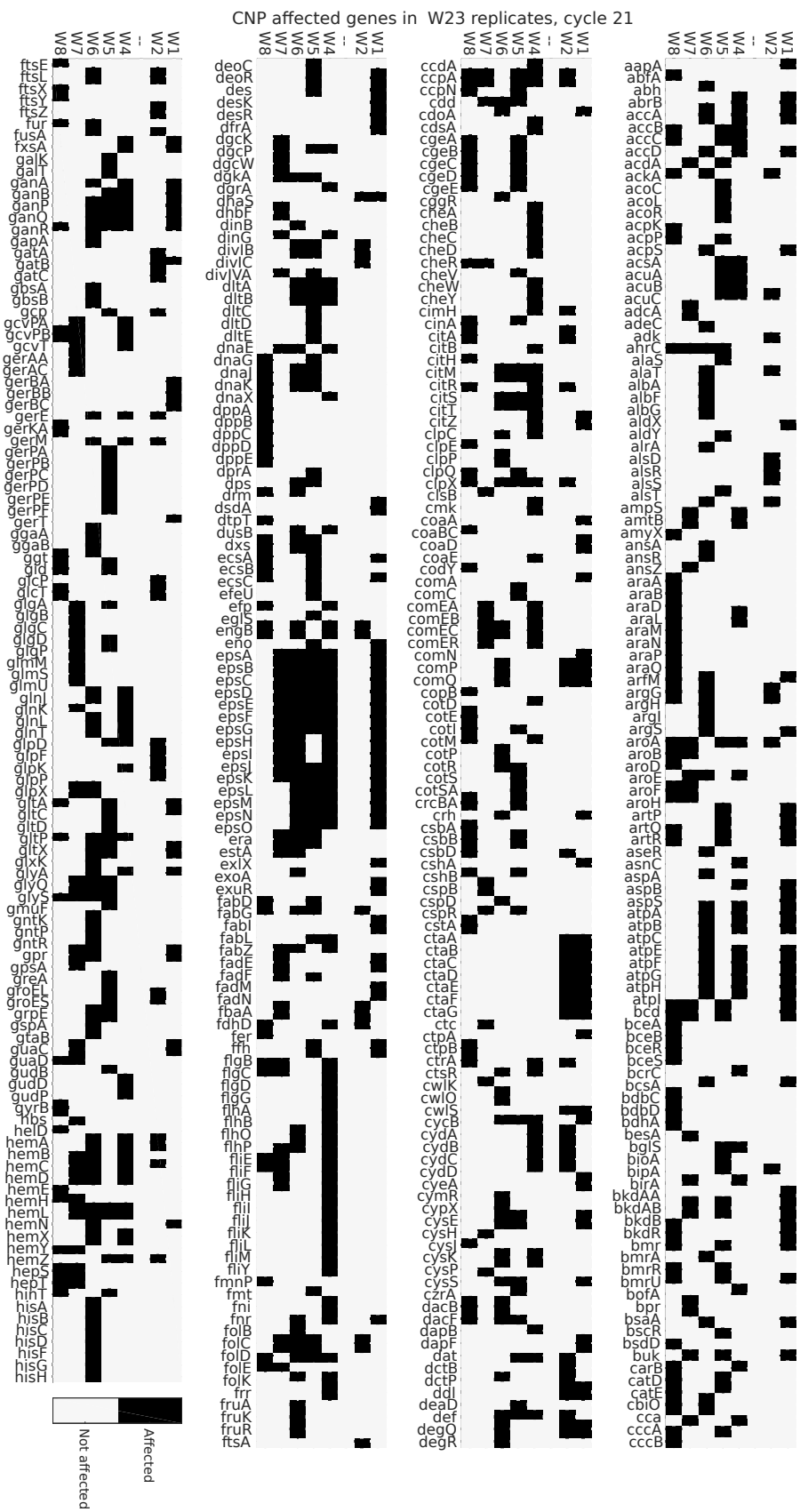


Figure A.8: Genes affected by CNPs, cycle 21 (part A). (black) Genes which were partially or fully affected. (white) Genes which were not replaced. This figure consists of five parts, A – E.

Figure A.9: Genes affected by CNPs, cycle 21 (part B). (black) Genes which were partially or fully affected. (white) Genes which were not replaced. This figure consists of five parts, A – E.

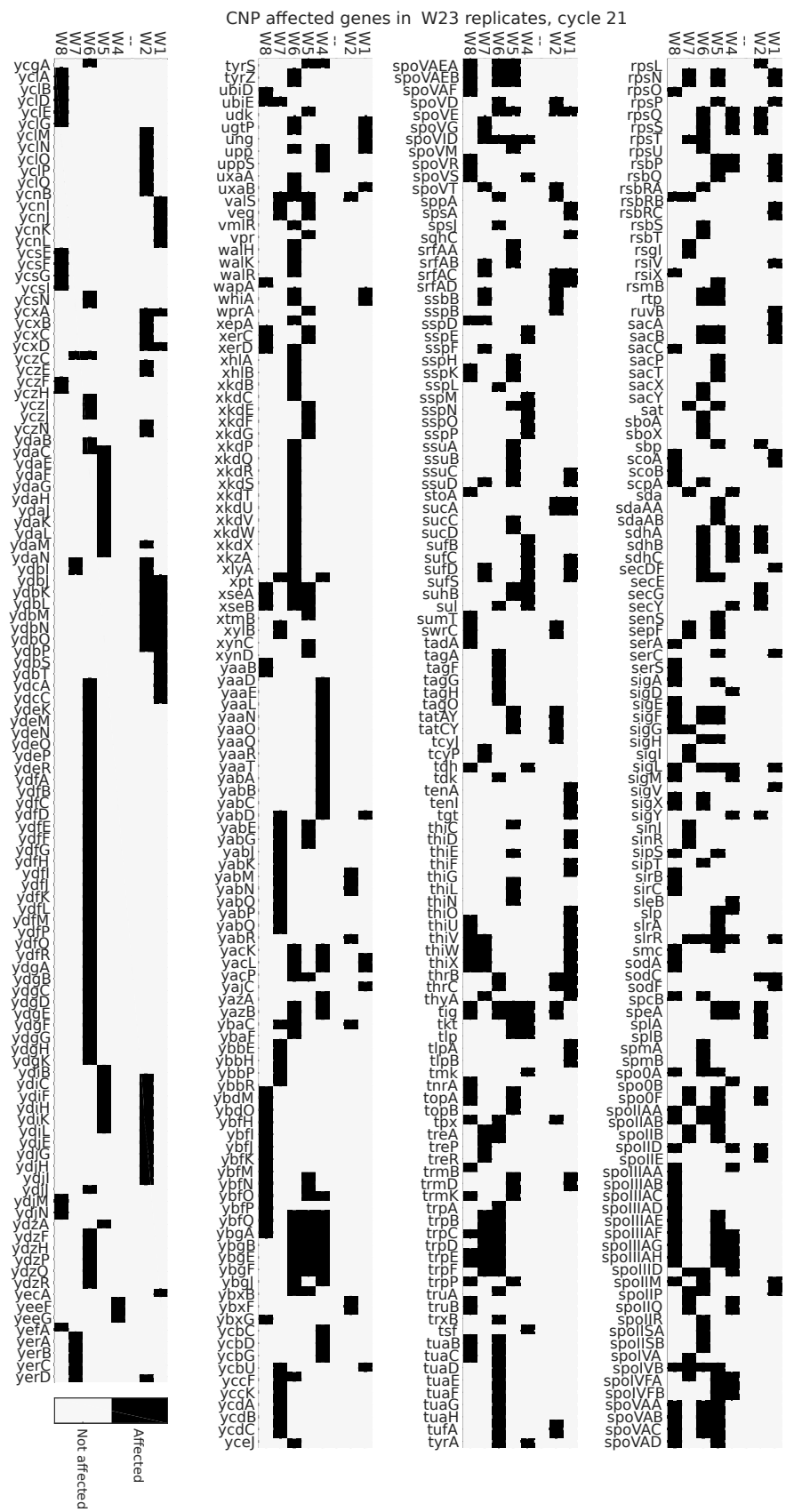


Figure A.10: Genes affected by CNPs, cycle 21 (part C). (black) Genes which were partially or fully affected. (white) Genes which were not replaced. This figure consists of five parts, A – E.

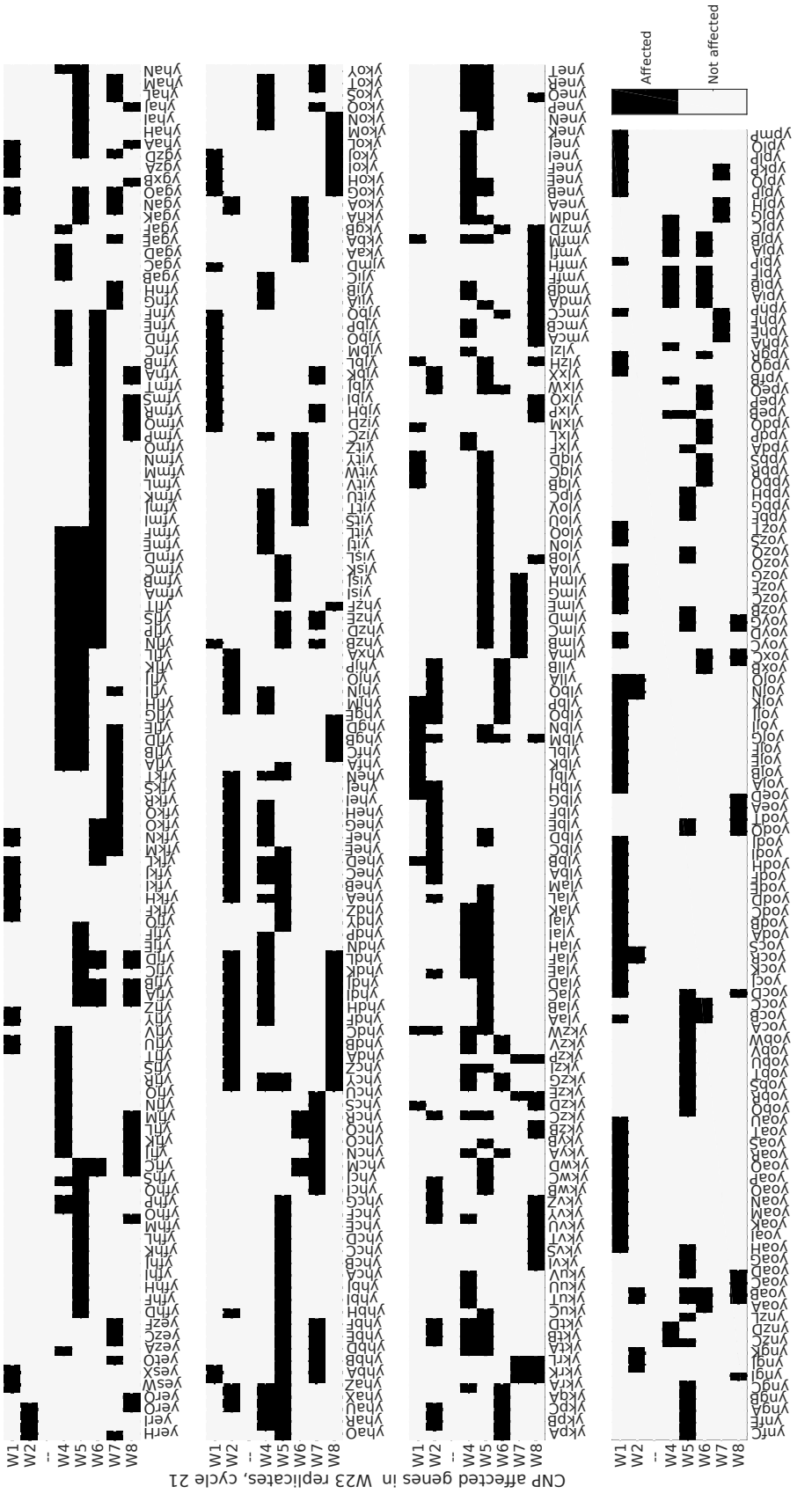


Figure A.11: Genes affected by CNPs, cycle 21 (part D). (black) Genes which were partially or fully affected. (white) Genes which were not replaced. This figure consists of five parts, A – E.

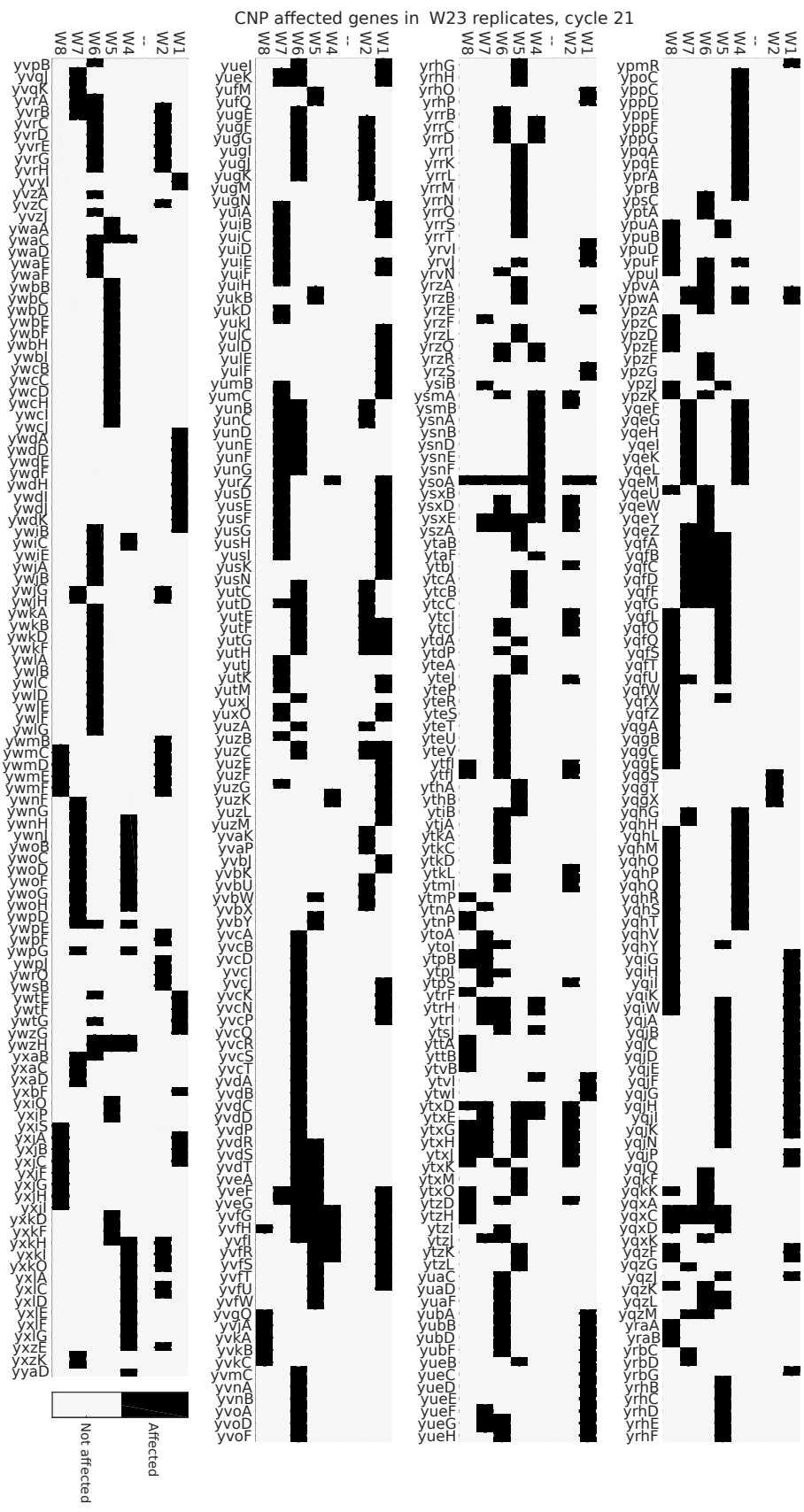


Figure A.12: Genes affected by CNPs, cycle 21 (part E). (black) Genes which were partially or fully affected. (white) Genes which were not replaced. This figure consists of five parts, A – E.

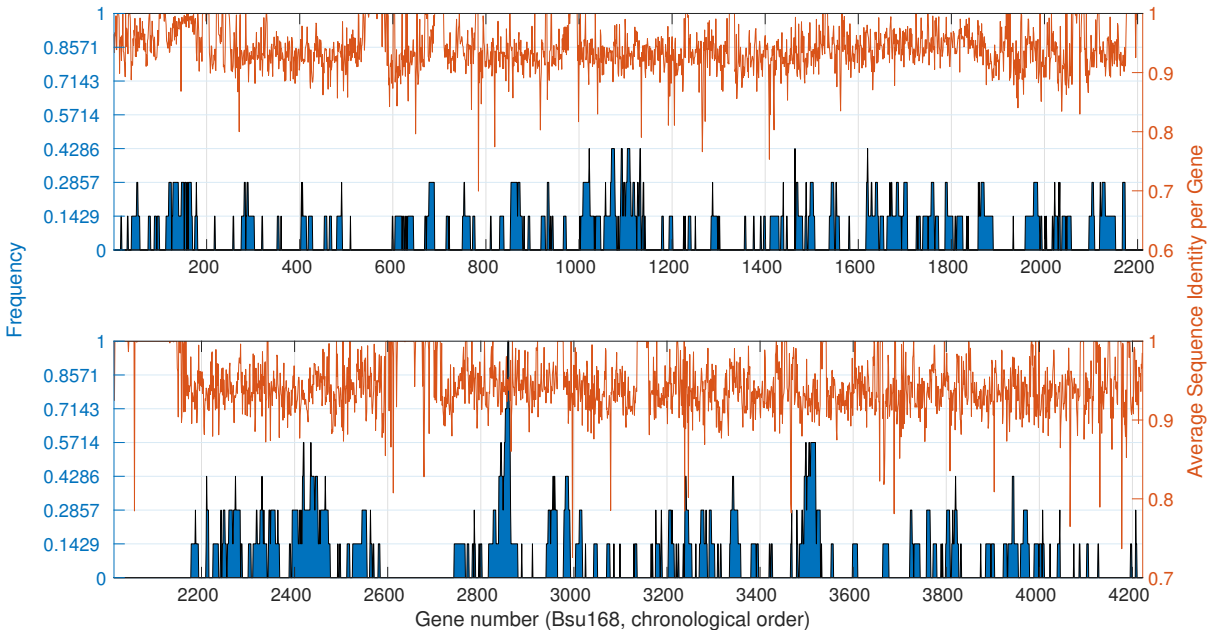
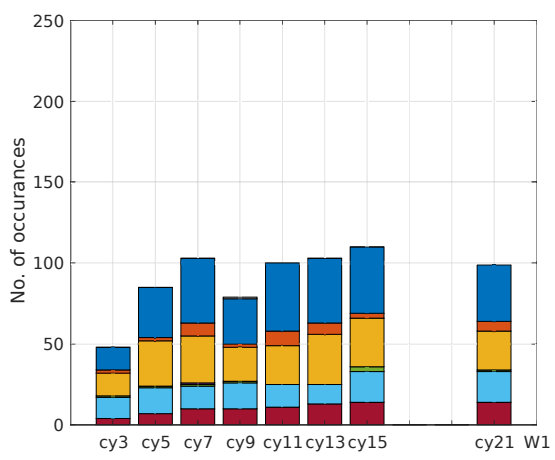
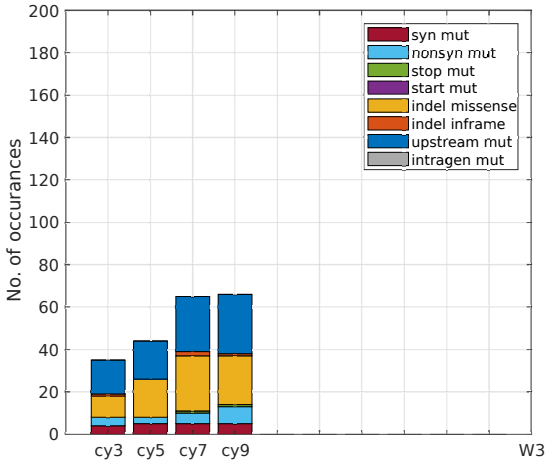


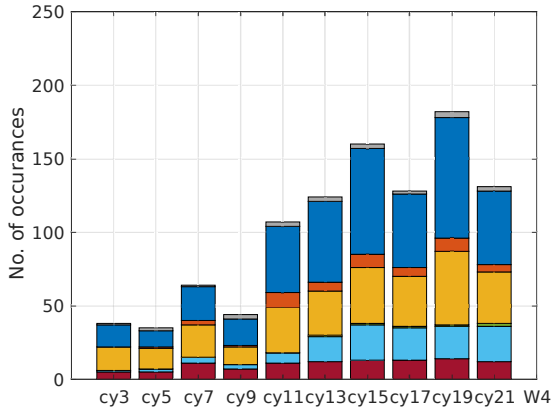
Figure A.13: Probability of replacement of a specific gene by cycle 15. (blue, left axis) Genes affected by a CNPs as a function of gene number and normalized to the number of replicates (seven). (orange, right axis) Gene identity. (gray dots) Bsu168 auxiliary genes. (These genes are displayed as having an identity of one.)



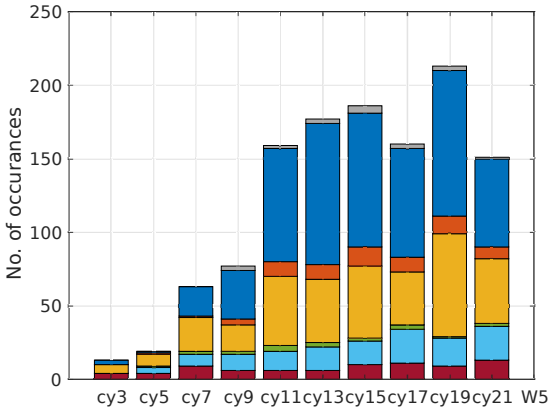
(a) Replicate W1



(b) Replicate W3



(c) Replicate W4



(d) Replicate W5

Figure A.14: De novo variants for replicates W1, 3, 4, and 5, color coded according to annotation. Upstream mutations are ≤ 3000 bp upstream of a gene's protein coding region. Intragenic mutations are not in a protein coding gene and >3000 bp upstream of a protein coding gene.

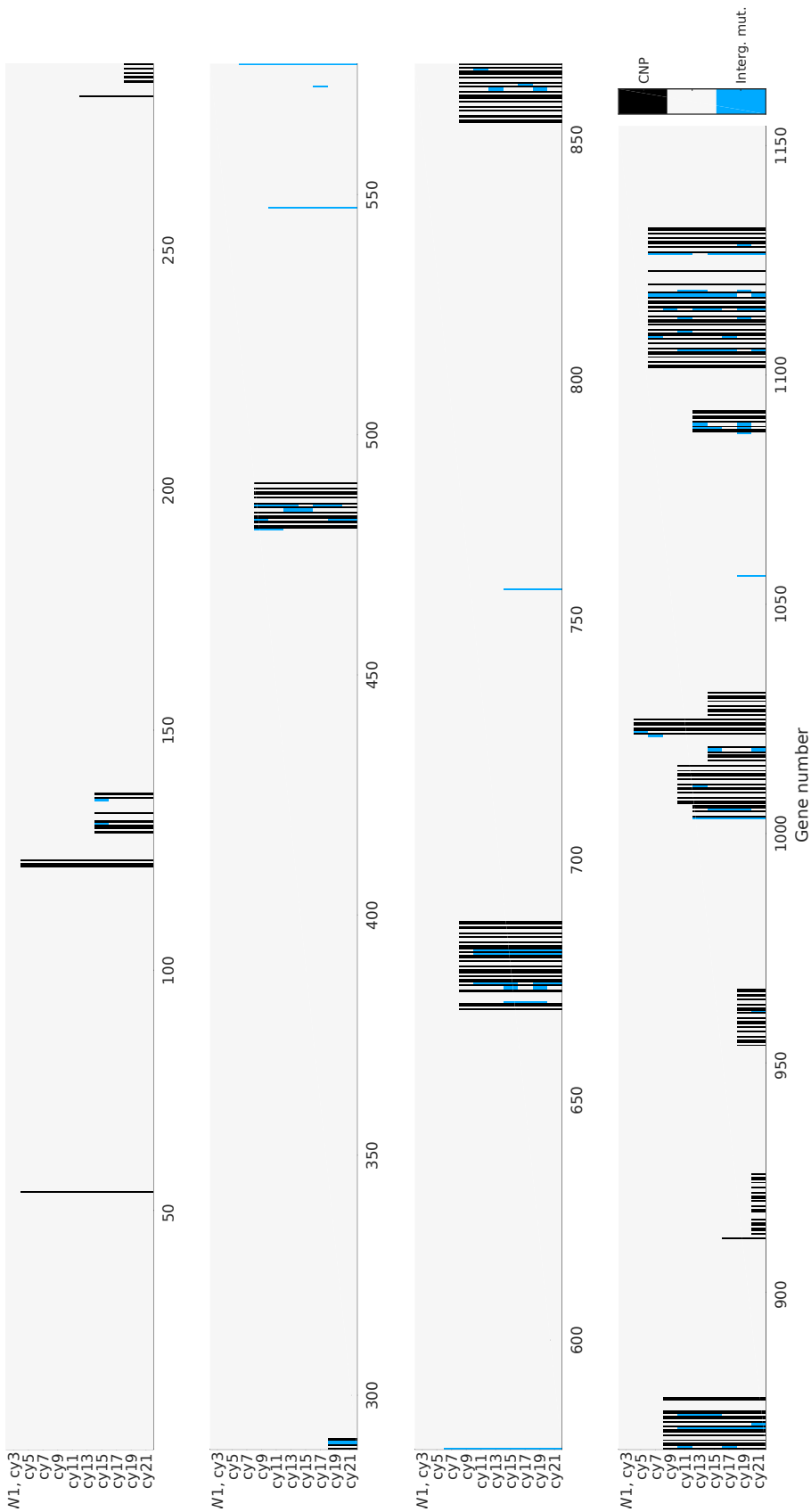


Figure A.15: CNP associated genes and upstream de novo mutations, replicate W1 (part A). (black) Genes which were partially or fully replaced with CNPs. (white) Genes which were not replaced. (blue) Intragenic regions with de novo mutations. This figure consists of four parts, A - D.

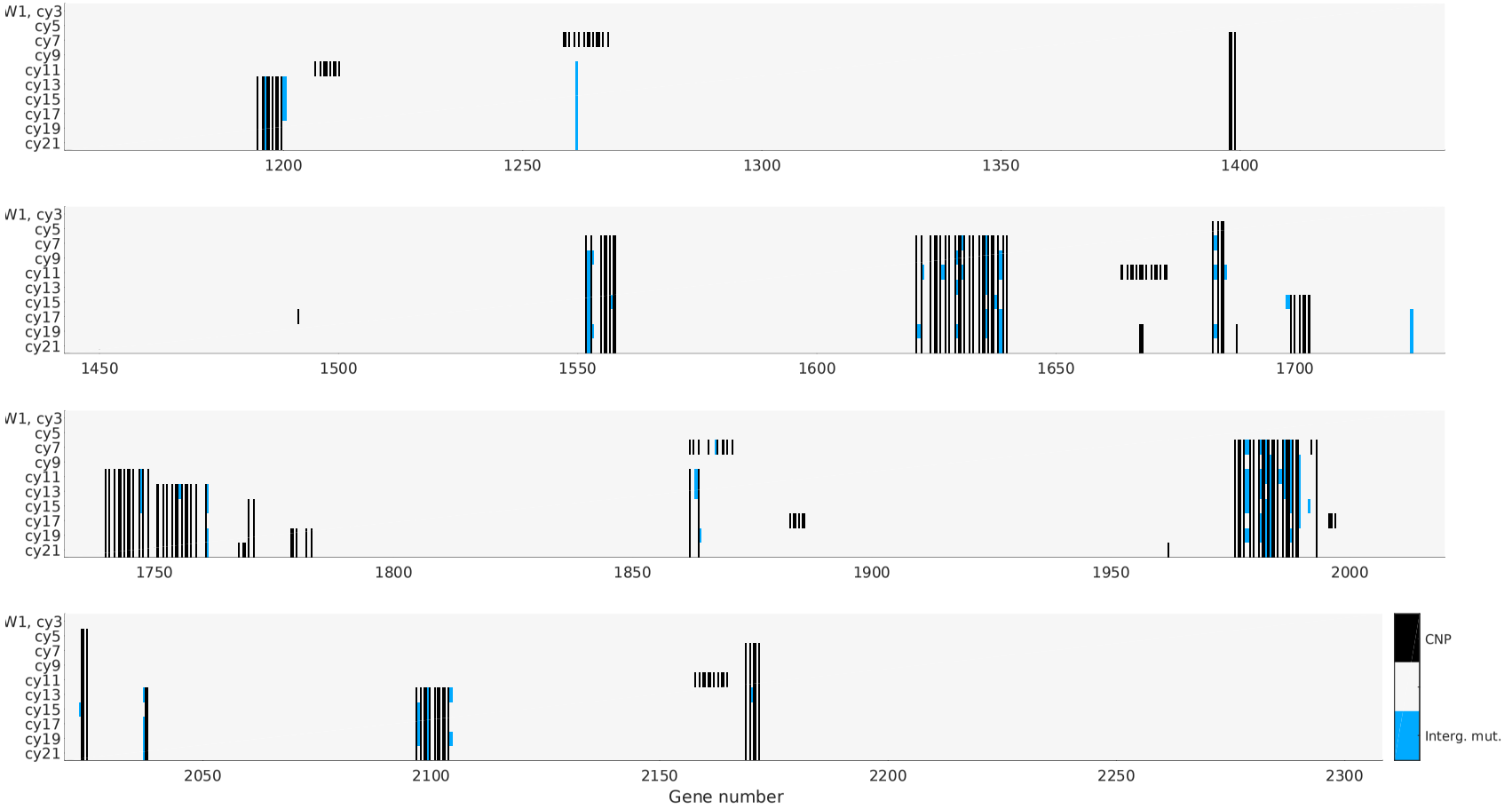


Figure A.16: CNP associated genes and upstream de novo mutations, replicate W1 (part B). (black) Genes which were partially or fully replaced with CNPs. (white) Genes which were not replaced. (blue) Intragenic regions with de novo mutations. This figure consists of four parts, A - D.

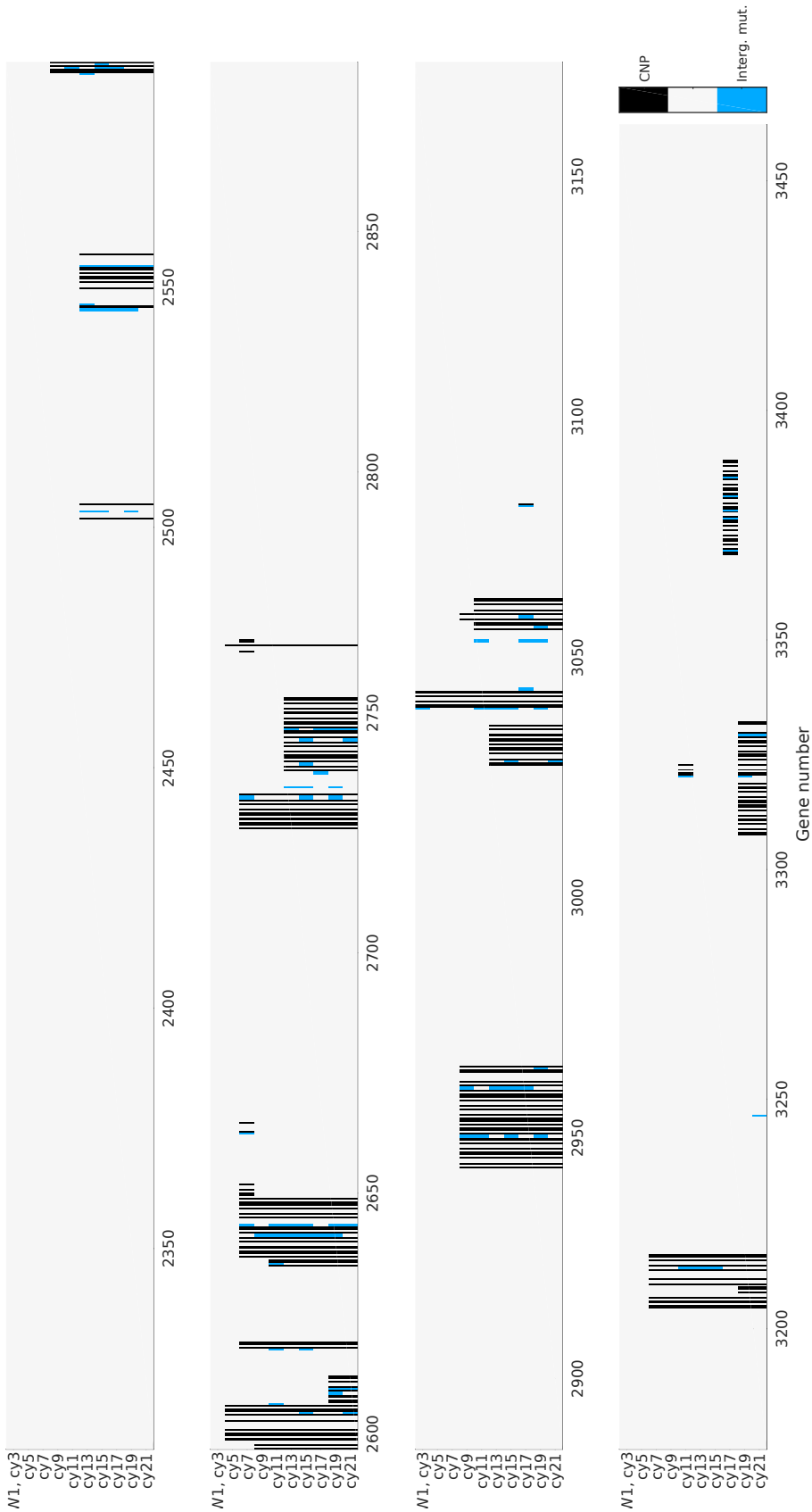


Figure A.17: CNP associated genes and upstream de novo mutations, replicate W1 (part C). (black) Genes which were partially or fully replaced with CNPs. (white) Genes which were not replaced. (blue) Intragenic regions with de novo mutations. This figure consists of four parts, A - D.

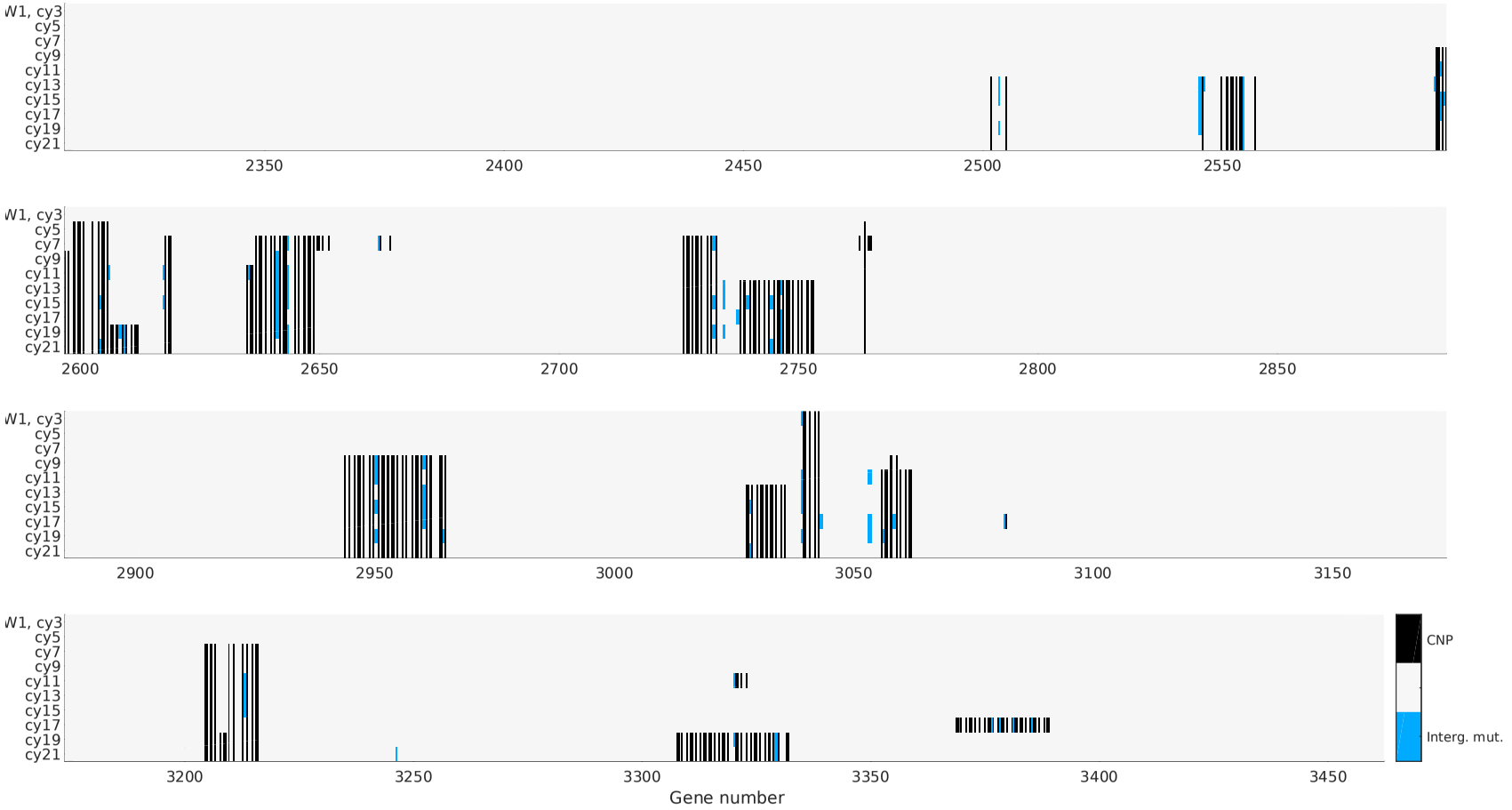


Figure A.18: CNP associated genes and upstream de novo mutations, replicate W1 (part D). (black) Genes which were partially or fully replaced with CNPs. (white) Genes which were not replaced. (blue) Intragenic regions with de novo mutations. This figure consists of four parts, A - D.

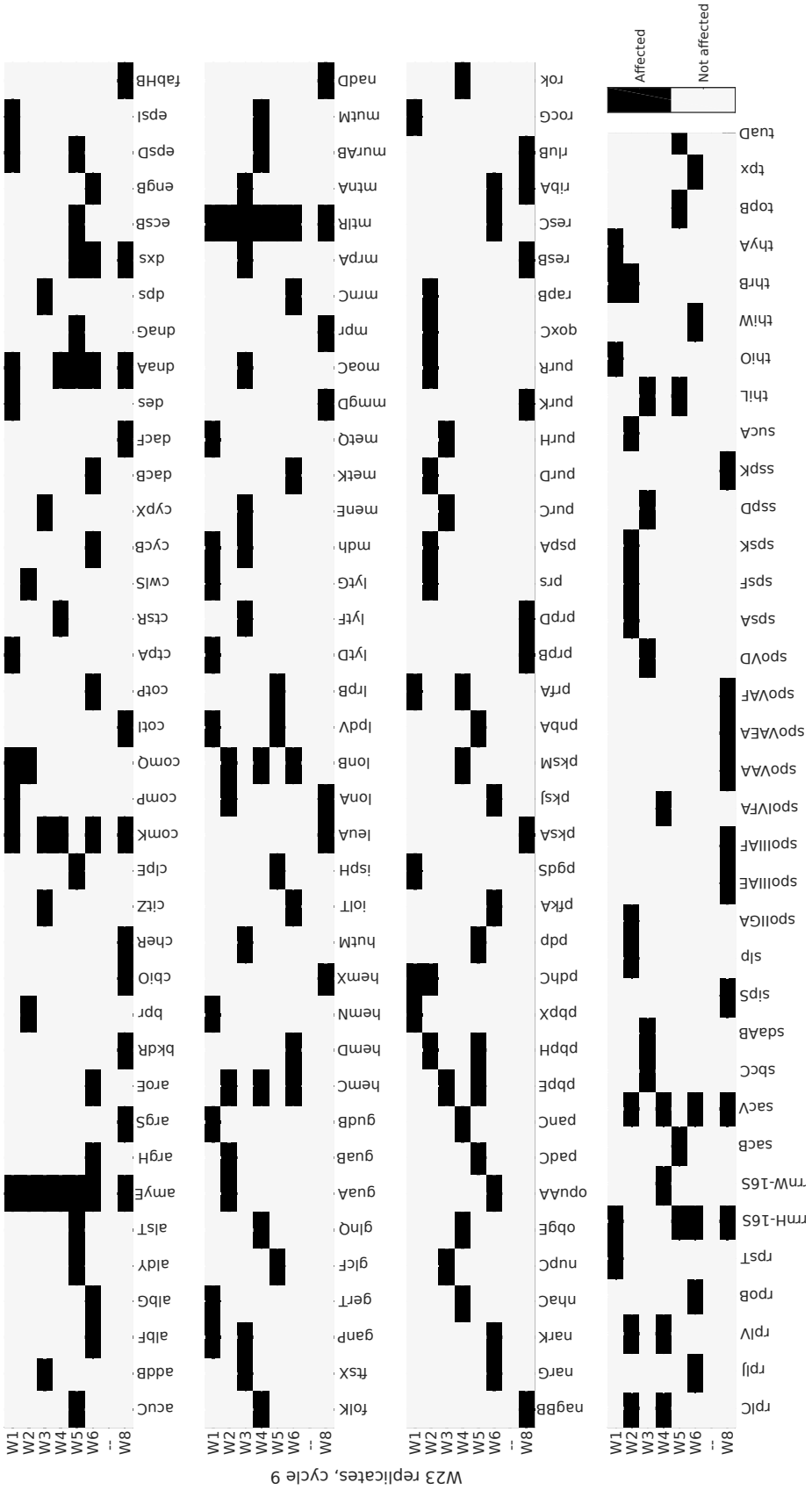


Figure A.19: Genes affected by de novo variants, cycle 9, all BsuW23 DNA replicates (part A). For a given replicate, affected genes are marked in black, non affected in white. This figure consists of two parts, A and B.

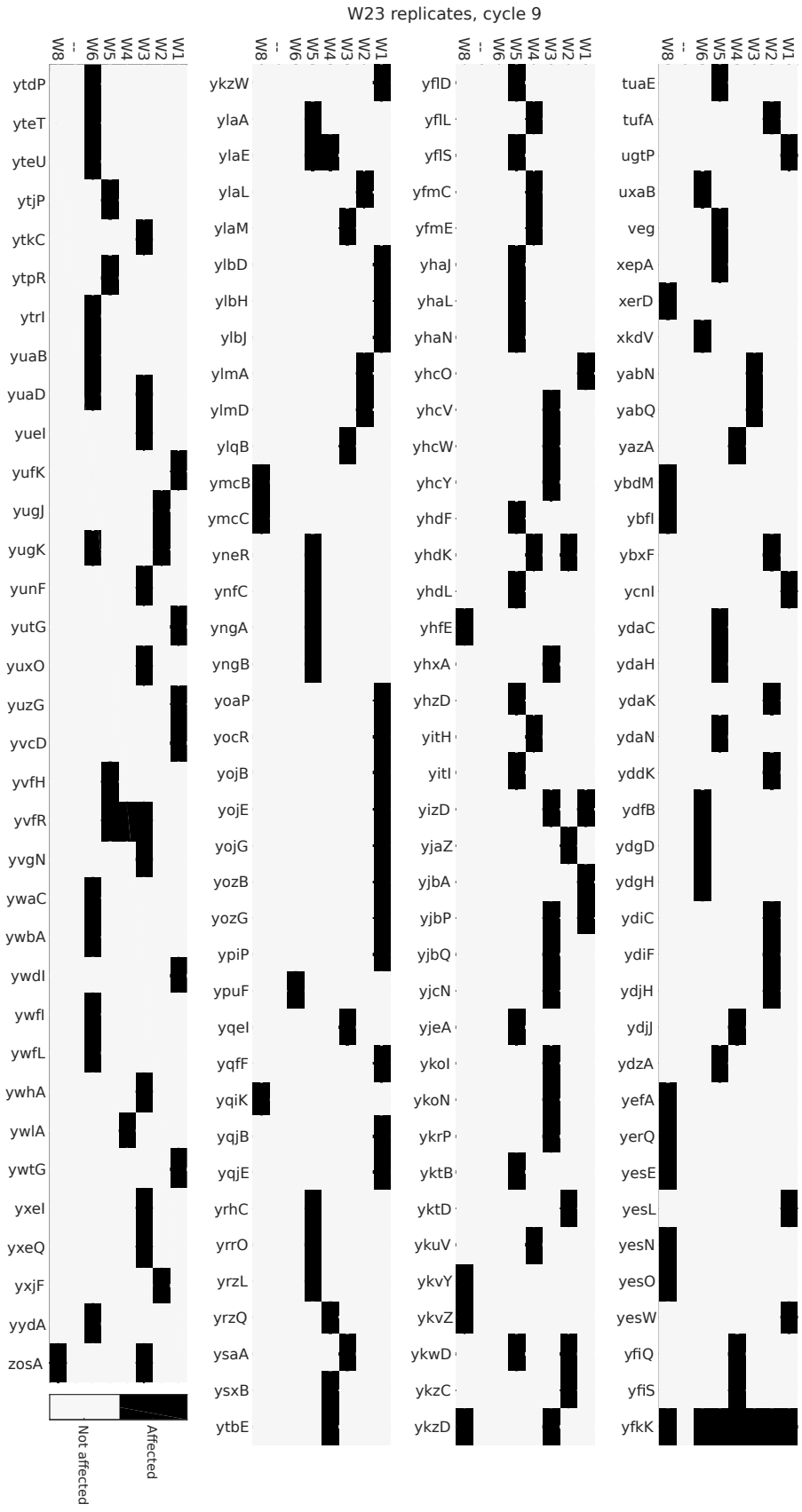


Figure A.20: Genes affected by de novo variants, cycle 9, all BsuW23 DNA replicates (part B). For a given replicate, affected genes are marked in black, non affected in white. This figure consists of two parts, A and B.

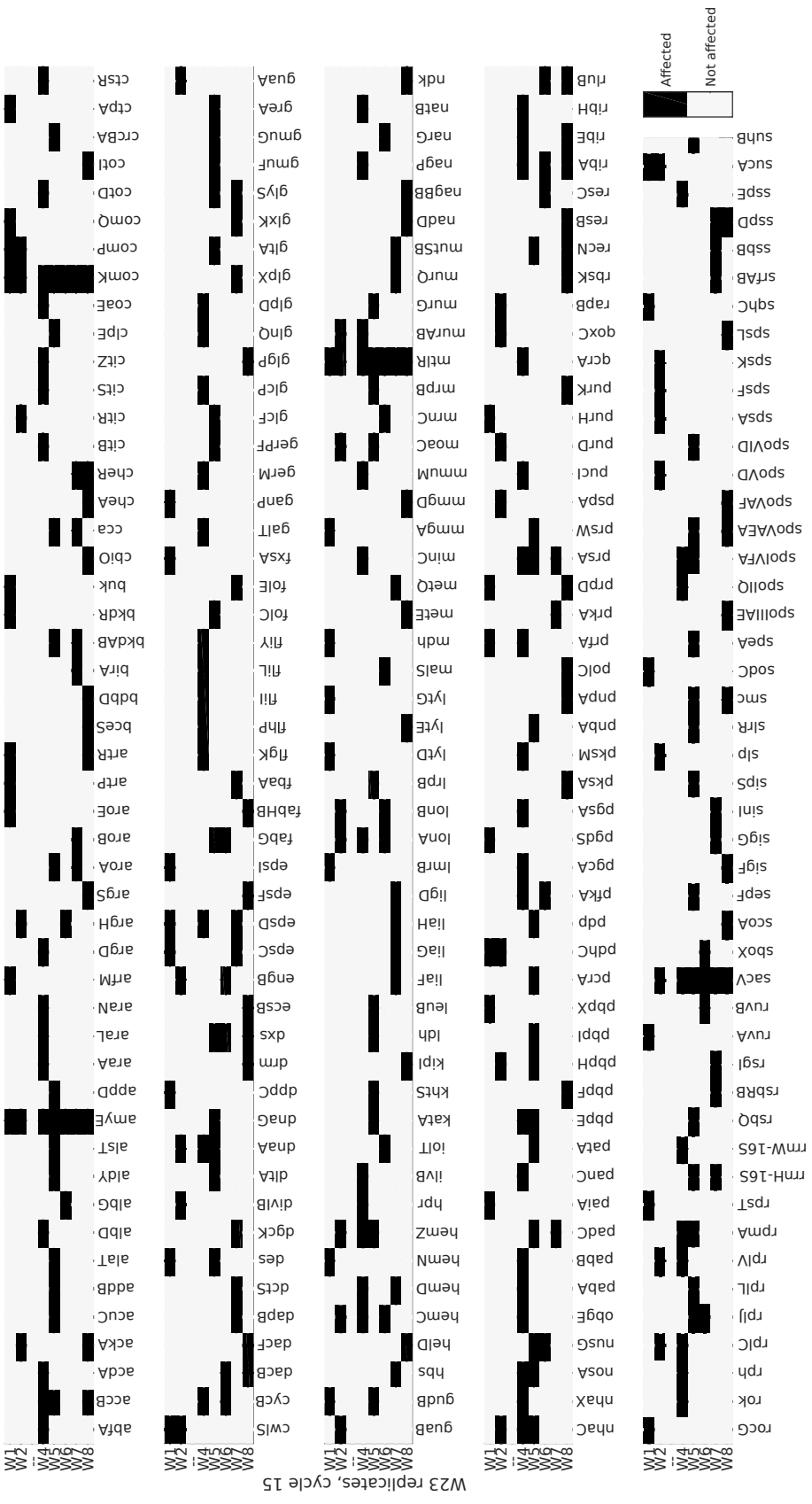


Figure A.21: Genes affected by de novo variants, cycle 15, all BsuW23 DNA replicates (part A). For a given replicate, affected genes are marked in black, non affected in white. This figure consists of two parts, A and B.

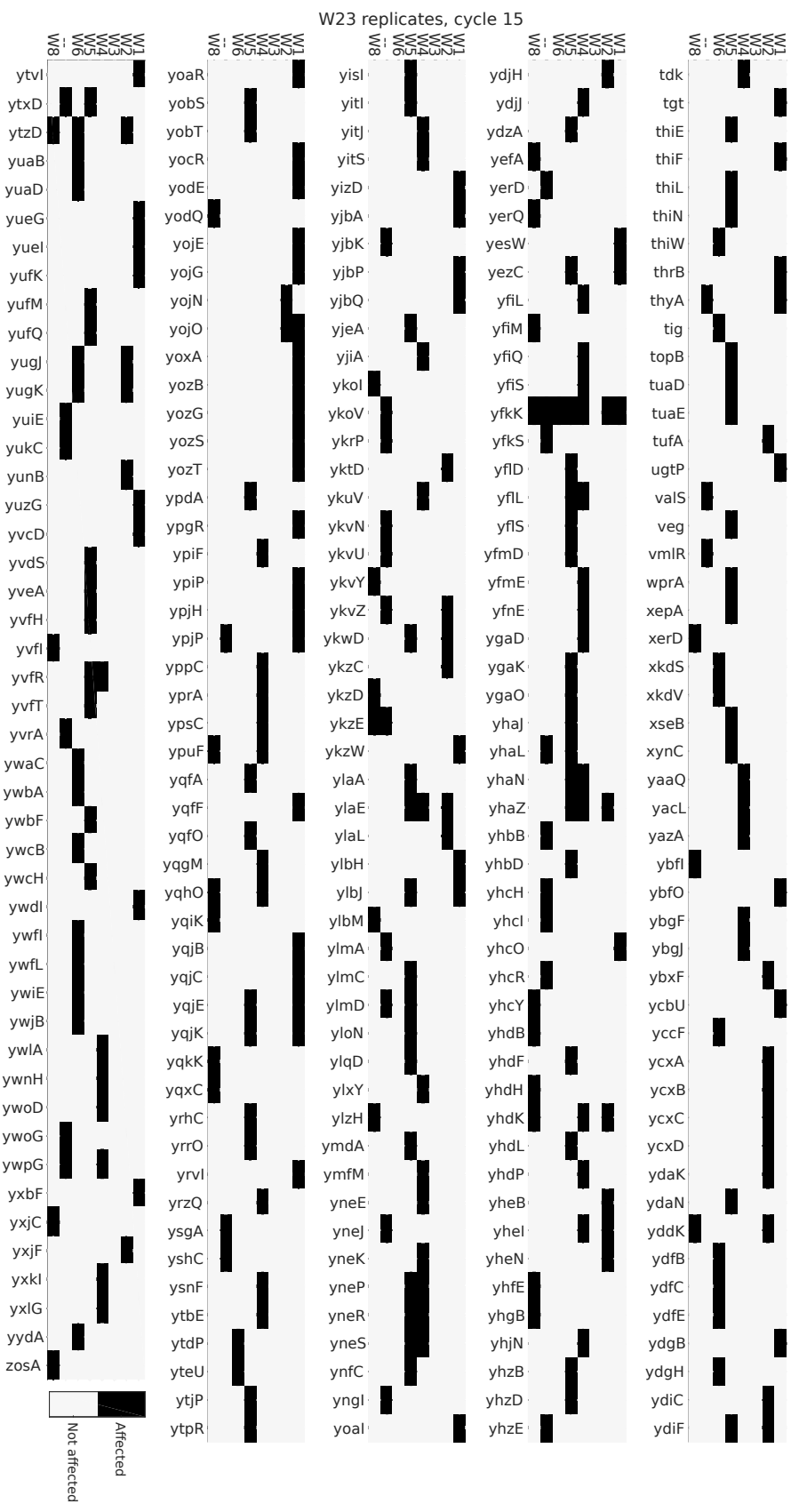


Figure A.22: Genes affected by de novo variants, cycle 15, all BsuW23 DNA replicates (part B). For a given replicate, affected genes are marked in black, non affected in white. This figure consists of two parts, A and B.

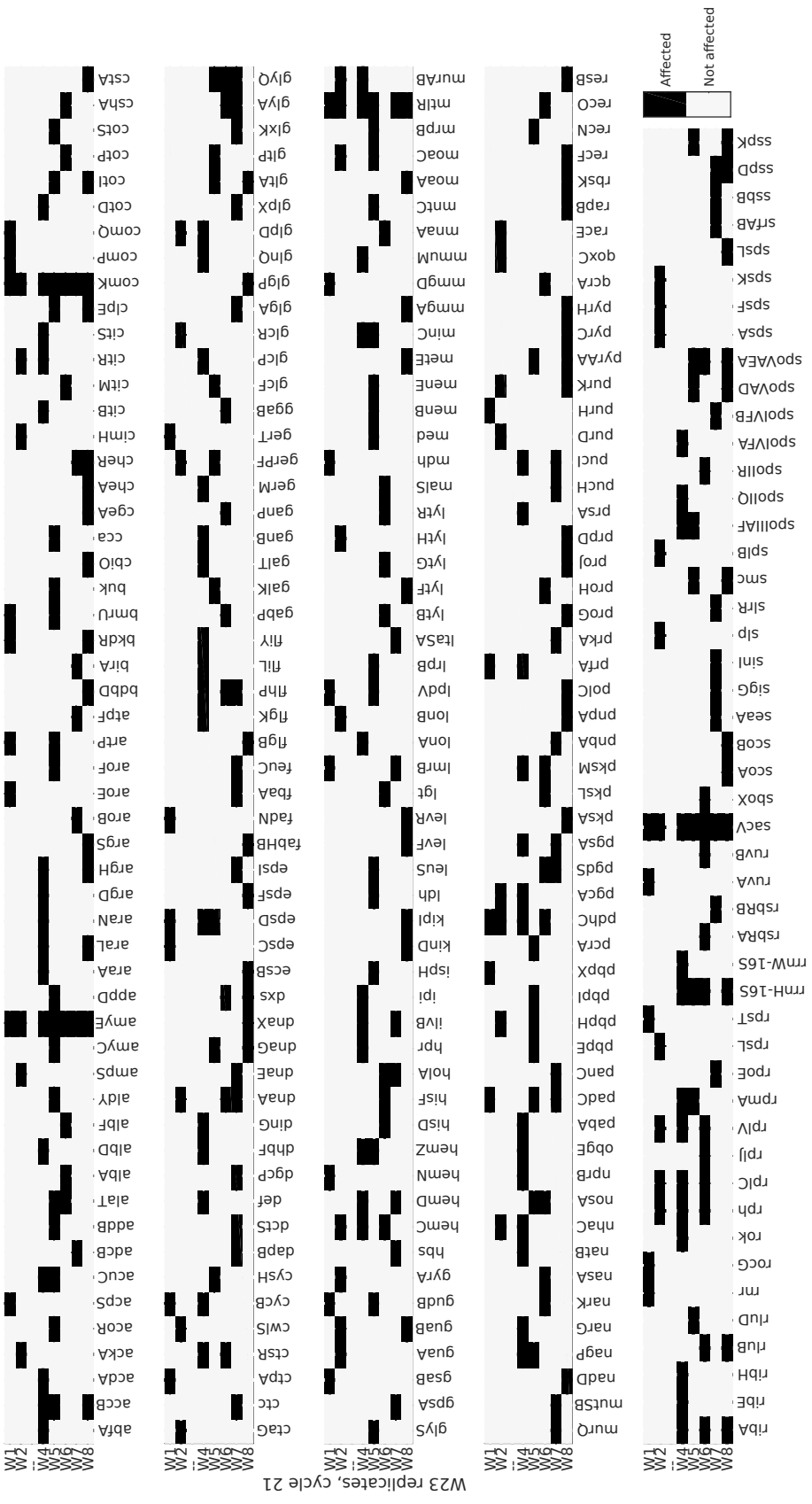


Figure A.23: Genes affected by de novo variants, cycle 21, all BsuW23 DNA replicates (part A). For a given replicate, affected genes are marked in black, non affected in white. This figure consists of two parts, A and B.

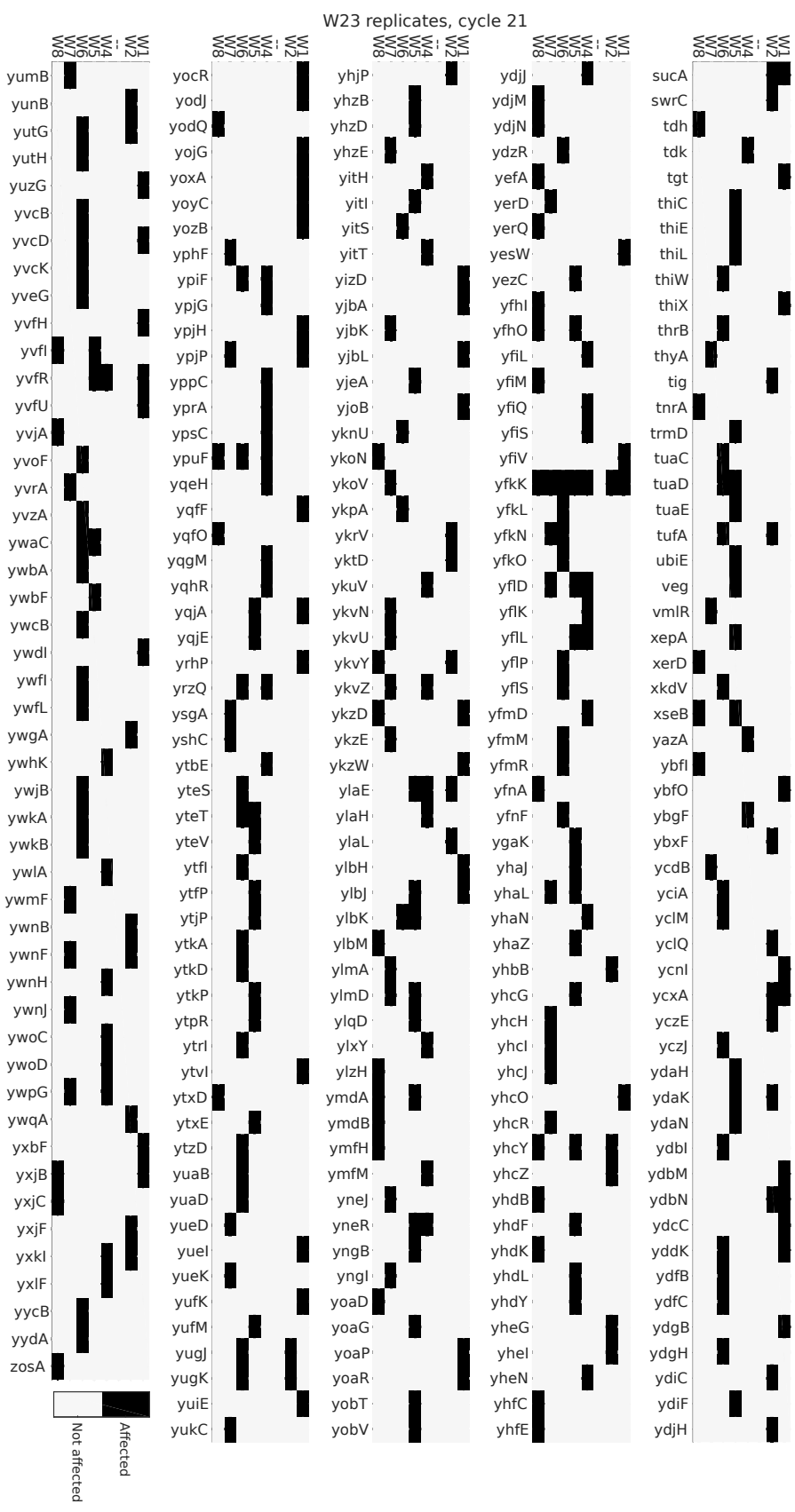


Figure A.24: Genes affected by de novo variants, cycle 21, all BsuW23 DNA replicates (part B). For a given replicate, affected genes are marked in black, non affected in white. This figure consists of two parts, A and B.

List of Figures

1.1 Horizontal gene transfer mechanisms in bacteria . . . 2
1.2 Clonal interference Müller diagram . . . 3
1.3 Cartoon schematic of competence for transformation . . . 4
1.4 Core of the *Bacillus subtilis* competence network . . . 5
1.5 ComK levels in *Bacillus subtilis* . . . 5
1.6 DNA uptake as a function of sequence divergence . . . 7
1.7 Types of evolution experiments . . . 10
1.8 Fitness landscapes . . . 14
1.9 Sign and magnitude epistasis . . . 15
1.10 Fitness and mutation rates in a long-term evolution experiment . . . 16
1.11 Mosaic pattern of recombination events . . . 18

2.1 Bsu168 and BsuW23 genome comparison . . . 23
2.2 Mock genome recovered and replacement gene segments . . . 33
2.3 Mock genome replacement and measured lengths. . . . 34
2.4 BsuW23 flank length distribution . . . 35
2.5 BsuW23 flank length cumulative probabilities . . . 36
2.6 Graphical description of four quality check measurements . . . 37
2.7 Quality checks for various cluster interruption thresholds . . . 37
2.8 Percentage of missing SNPs as a function of cluster length . . . 38
2.9 Quality checks for various cluster window sizes . . . 39
2.10 Multiple of three bias in BsuW23 flank lengths . . . 40
2.11 Mutation distributions for alignment pipeline mock genomes . . . 41
2.12 Novel gene algorithm test . . . 43

3.1 Flow chamber layout . . . 47
3.2 Image processing of z-stack DIC images . . . 49

4.1 Cycle 9, 15, and 21 orthologous recombination events of BsuW23 samples . . . 53
4.2 Rate of genome replacement . . . 53
4.3 Cycle 9, 15, and 21 CNP mean import lengths histogram, BsuW23 . . . 54
4.4 Cycle 9, 15, and 21 mean lengths of imported segments from auxiliary regions . . . 55
4.5 Cycle 21 orthologous recombination and de novo insertions from BsuW23 . . . 56
4.6 Time lapse orthologous recombination events of replicates W1, 3, 4, and 5 . . . 57
4.7 Replicate W5 orthologous recombination and de novo insertions from BsuW23 . . . 59
4.8 Identity of orthologous recombination segments as a function of segment length for BsuW23 cycle 9, 15, and 21 samples . . . 61
4.9 Purine and pyrimidine enrichment in CNPs from BsuW23 replicates . . . 62
4.10 Gene and operon replacement distributions . . . 64
4.11 Change in CNP lengths over time . . . 65
4.12 Cycle 21 CNP gene types . . . 66
4.13 Genes are randomly affected across the entire genome . . . 68
4.14 Probability of replacement for a specific gene . . . 68
4.15 SNP density distributions of CNP first 100 bp . . . 70
4.16 First 100 bp SNP density . . . 71

4.17 D statistic for identity distributions along moving window 72
4.18 Frequency of partially and fully replaced genes 73
4.19 Annotated de novo variants for replicates receiving no or Bsu168 DNA 74
4.20 Annotated de novo variants for cycles 9, 15, and 21 75
4.21 Annotated de novo variants within CNPs, cycle 21 75
4.22 Temporal occurrence of intergenic mutations upstream of CNP affected genes 76

5.1 Differentiation into the K-state is associated with growth arrest 78
5.2 Determination of selection coefficients in the stationary state 79

6.1 Cross species fitness model 83
6.2 Distribution of maximum perfect identity lengths 85

A.1 Cycle 9 orthologous recombination events and horizontal gene transfer from BsuW23 . 112
A.2 Cycle 15 orthologous recombination events and horizontal gene transfer from BsuW23 113
A.3 Replicate W1 orthologous recombination events and horizontal gene transfer from BsuW23 114
A.4 Replicate W3 orthologous recombination events and horizontal gene transfer from BsuW23 115
A.5 Replicate W4 orthologous recombination events and horizontal gene transfer from BsuW23 116
A.6 Normalized fractions of essential genes 117
A.7 CNP segment gene types, cycles 9 and 5 117
A.8 Genes affected by CNPs, cycle 21 (part A) 118
A.9 Genes affected by CNPs, cycle 21 (part B) 119
A.10 Genes affected by CNPs, cycle 21 (part C) 120
A.11 Genes affected by CNPs, cycle 21 (part D) 121
A.12 Genes affected by CNPs, cycle 21 (part E) 122
A.13 Probability of replacement for a specific gene, cycle 15 123
A.14 Annotated de novo variants for replicates W1, 3, 4, and 5, all cycles 124
A.15 CNP associated genes and intragenic de novo mutations, replicate W1 (part A) 125
A.16 CNP associated genes and intragenic de novo mutations, replicate W1 (part B) 126
A.17 CNP associated genes and intragenic de novo mutations, replicate W1 (part C) 127
A.18 CNP associated genes and intragenic de novo mutations, replicate W1 (part D) 128
A.19 Genes affected by de novo variants, cycle 9, all BsuW23 replicates (part A) 129
A.20 Genes affected by de novo variants, cycle 9, all BsuW23 replicates (part B) 130
A.21 Genes affected by de novo variants, cycle 15, all BsuW23 replicates (part A) 131
A.22 Genes affected by de novo variants, cycle 15, all BsuW23 replicates (part B) 132
A.23 Genes affected by de novo variants, cycle 21, all BsuW23 replicates (part A) 133
A.24 Genes affected by de novo variants, cycle 21, all BsuW23 replicates (part B) 134

List of Tables

2.1 Evolution experiment bacterial strains 22
2.2 Key *B. subtilis* 168 and *B. subtilis* W23 statistics 22
2.3 Evolution experiment media 24
2.4 Experimental competence induction conditions 26
2.5 Mock genome segments to test CNP algorithm 33
2.6 Mock genome segments to test auxiliary genes algorithm 43

3.1 Population dynamics experiment bacterial strains 45
3.2 Population dynamics media 46

4.1 Cycles 9, 15, and 21 CNP segment statistics 52
4.2 Mean identities and significance values for cycles 9, 15, and 21 60
4.3 Genes affected in the majority of cycle 21 replicates 69

5.1 Generation times of K-state and non K-state cells 77
5.2 Selection coefficients in the stationary phase 78

A.1 CNP segment statistics for replicates W1, 3, 4 and 5 at final cycle 111

List of Schemes

2.1 Evolution experiment design 21
2.2 Alignment pipeline 27
2.3 CNP algorithm to detect orthologous recombination 31

List of Code Snippets

2.1 Initial sequencing pipeline steps 28
2.2 Hard mapping sequencing pipeline 28
2.3 Loose mapping sequencing pipeline 29
2.4 Dictionary assembly 29
2.5 In silico read production 32
2.6 Fasta file modification 40

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Frau Prof. Dr. Berenike Maier betreut worden.

Köln, den 11. November 2018

Jeffrey John Power

Teilpublikationen

M. Yüksel, J. J. Power, J. Ribbe, *et al.*, "Fitness trade-offs in competence differentiation of *Bacillus subtilis*," *Frontiers in Microbiology*, vol. 7, Jun. 7, 2016. DOI: 10.3389/fmicb.2016.00888