UNIVERSITY OF CALIFORNIA,
IRVINE

Computational and Quantitative Approaches to Challenges in Genetics, Ecology, and Evolution

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical, Computational and Systems Biology

by

Roy Nan Zhao

Dissertation Committee:
Professor John S. Lowengrub, Chair
Professor Peter A. Bowler
Professor Xiaohui Xie

2023

DEDICATION

To my family, friends, mentors,
and everyone who has been part of the journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am immensely grateful to my mentors and thesis committee, Professors John S. Lowengrub, Peter A. Bowler, and Xiaohui Xie, for their guidance, support, and encouragement during my PhD studies.

I would also like to thank all of my collaborators, friends, and family for helpful discussions, camaraderie, and support.

Portions of Chapters 2 and 3, published by Springer Nature [22] and Oxford University Press [157], are included with permission under Creative Commons CC BY.

I thank all of my co-authors and collaborators, including Tamas Lukacsovich, Rebecca Gaut, J.J. Emerson, Mahul Chakraborty, Xinwen Zhang, Nicholas VanKuren, Shannon Kalsow, Katie Pearson, and all others.

This work was made possible by support from the National Science Foundation Graduate Research Fellowship (DGE-1321846), National Institutes of Health (T32-EB009418-07), and the University of California, Irvine.

VITA

Roy Nan Zhao

EDUCATION

Ph.D. in Mathematical, Computational and Systems Biology **2023**
University of California, Irvine *Irvine, CA*

B.A. in Biological Sciences **2015**
University of Chicago *Chicago, IL*

RESEARCH EXPERIENCE

Graduate Student Researcher **2016–2023**
University of California, Irvine *Irvine, CA*

TEACHING EXPERIENCE

Teaching Assistant **2020–2023**
University of California, Irvine *Irvine, CA*

JOURNAL ARTICLES

R. Zhao. Automated segmentation and transcription of digitized herbarium specimens using computer vision and machine learning techniques. 2023 [manuscript in preparation].

R. Zhao, T. Lukacsovich, R. Gaut, and J. J. Emerson. FREQ-Seq^2: a method for precise high-throughput combinatorial quantification of allele frequencies. *G3: Genes|Genomes|Genetics*, 13(10):jkad162, 2023.

M. Chakraborty, N. W. VanKuren, R. Zhao, X. Zhang, S. Kalsow, and J. J. Emerson. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nature Genetics*, 50(1):20–25, 2018.

ABSTRACT OF THE DISSERTATION

Computational and Quantitative Approaches to Challenges in
Genetics, Ecology, and Evolution

By

Roy Nan Zhao

Doctor of Philosophy in Mathematical, Computational and Systems Biology

University of California, Irvine, 2023

Professor John S. Lowengrub, Chair

The proliferation of and continual improvement in sequencing technologies over the last three decades have vastly increased the rate of biological data generation, while concurrently reducing cost per base sequenced. However, in many ways, statistical and computational strategies for extracting meaningful biological insights from this unprecedented wealth of data have not kept pace with these molecular advances. For instance, biological features such as genomic structural variants (SVs) can remain challenging to capture and evaluate with the most common types of sequencing, despite their implication in complex traits including adaptive phenotypes and disease predisposition. To address these unmet needs in this rapidly changing scientific landscape, we developed novel algorithms and applied cutting-edge computational and quantitative techniques to a diverse set of scientific questions.

First, we developed a method called FREQ-Seq^2, along with associated algorithms and software, to enable the rapid and targetable study of allele frequencies, while addressing the challenges inherent to traditional techniques, such as throughput and scalability. We applied our approach to studying evolutionary genetics in *E. coli*, demonstrating its accuracy and precision in characterizing population dynamics and genotype distributions. We also contributed a valuable resource to the scientific community by presenting a new reference-quality genome assembly for the model organism *Drosophila melanogaster* based on long-read sequencing, along with a thorough investigation of hidden genetic variation and evolutionary insights that we uncovered. For example, we discovered previously unidentified SVs potentially linked to complex traits, such as a duplication associated with nicotine resistance, as well as novel understanding about the cosmopolitan distribution of SVs in *Drosophila*. Additionally, we leveraged advancements in machine learning to further explore SVs, training a deep convolutional neural network-based model to identify SVs with high precision using widely available short-read *Drosophila* datasets. Finally, we demonstrated the utility of applying computer vision and machine learning techniques to ongoing initiatives to digitize herbaria catalogs, by developing algorithms and a pipeline to extract phenological and ecological data from digitized specimen images, facilitating rapid and accurate labeling and annotation as well as expanding ease of end-user access to these resources. Taken as a whole, this thesis provides a comprehensive study of computational and quantitative approaches and their application to wide-ranging challenges in genetics, ecology, and evolutionary biology.

CHAPTER 1

Introduction

The proliferation of high-throughput technologies in recent years has reshaped the landscape of biological research by enabling the generation of vast amounts of data, prompting a broad-based integration of quantitative and computational approaches to take advantage of these advancements. This change has opened new avenues for scientific exploration which were previously infeasible with traditional methods. This is certainly the case for the fields of genetics, ecology, and evolutionary biology, where data-driven methodologies have facilitated research on complex systems at population scales [76, 108]. However, despite these advancements, many fundamental challenges persist, in addition to new challenges introduced by this immense inflow of information. In this dissertation, we tackle a number of the challenges faced in effectively using these data, and we draw upon both computational and mathematical techniques in our analysis and interpretation of biological data in order to ultimately tackle biological questions in these fields.

Over the past few decades, the advent of next-generation sequencing technologies has generated incredible amounts of sequencing data of various types. Within the last decade in particular, the cost of Illumina sequencing has dramatically decreased, with a corresponding increase in its accessibility and range of applications [108]. Naturally, this has resulted in vast amounts of data being generated at an unprecedented scale and speed. Indeed, in many cases, the ability to generate large amounts of data has greatly outpaced researchers' ability to extract all the meaningful insights stored within, demanding novel solutions for effectively interfacing with these data.

The work discussed in the following chapters sits at the intersection of genetics, ecology, and evolutionary biology with mathematical and computational biology, representing a nexus of disciplines that offers a holistic perspective on the complexity of biological systems. Accordingly, we have employed a variety of computational and quantitative methodologies, in many cases integrating them with more traditional biological research techniques to more effectively study complex, and often classic, research questions. Specifically, we have focused on uncovering and characterizing genetic variation, understanding its evolutionary and ecological implications, and utilizing both 21st century genomic data and the digitization of rich historical data sources to tackle contemporary problems in the biological sciences.

These methodologies range from the development of algorithms for quantifying allele frequencies in mixed population samples, to the assembly of reference-quality genomes for uncovering hidden structural variation, the development of new methods for extracting useful information from existing data, application of machine learning techniques for genotyping structural variation, and the use of computer vision and natural language processing for extracting valuable phenological information from preserved organisms. Furthermore, we use these approaches to learn about ecological and evolutionary dynamics, utilizing sophisticated computational techniques to extract valuable phenological data from digitized herbaria, which are a treasure trove of historical and geographical biodiversity data. Work that addresses these challenges is important not only for the continued advancement of knowledge in these fields but also for the broader implications they hold for conservation, health, and our understanding of life's diversity and resilience.

1.1 RESEARCH OBJECTIVES

Despite significant progress in biotechnology and our understanding of biodiversity, several critical challenges remain at the forefront of genetics, ecology, and evolution. One such challenge lies in accurately quantifying allele frequencies on a massive scale, particularly at arbitrary polymorphic sites in the genome, in mixed population samples, with reasonable labor, and with reasonable cost [75, 77]. These difficulties have often resulted in reducing sample size or limiting the range of genetic loci that are investigated as a practical necessity, sometimes defining a fixed set of polymorphisms to target. Given the sheer number of genetic variants present in natural populations, these constraints represent substantial limitations, especially on population-scale studies.

Another major challenge lies in accurately characterizing genetic variation within and between populations, a task that is crucial for a myriad of applications, from understanding evolutionary dynamics to predicting disease susceptibility. Furthermore, while single-nucleotide variations have been extensively studied, structural variations, a class of much larger alterations in the genome which affect a significantly greater proportion of DNA sequence, are often overlooked due to the much greater difficulty in locating and studying them [5, 149]. Nevertheless, they represent a substantial contribution to genetic diversity, and likely phenotypic diversity [24, 121]. Considering the open questions around the so-called missing heritability problem, where even large-scale genome-wide association studies have been persistently unable to explain the vast majority of variation in heritable traits [95], paying attention to structural variation is a critical part of closing the gaps in our understanding of functional genomics.

In recent years, the digitization of herbarium specimens has opened new avenues of research in phenology, the study of periodic plant and animal life cycle events and their influences by seasonal and interannual variations in climate. These preserved plant specimens, typically accompanied by detailed information recorded during their collection, provide invaluable snapshots of biodiversity across time and space [104]. Thus, there is simultaneously an extraordinary opportunity and challenge in harnessing the wealth of historical and geographical information related to biodiversity stored in herbaria worldwide. Such resources are continually becoming more crucial for understanding and addressing the challenges we face in the modern world, in the context of major unresolved problems such as climate change. However, extracting meaningful data from these specimens on a large scale is a difficult task necessitating the development of new computational techniques.

The research presented herein tackles a range of diverse yet interconnected problems. In the following chapters, we will discuss several ways in which we address the challenges described above. The research topics include the efficient quantification of allele frequencies, characterization of genetic variation, uncovering and genotyping structural variation, and unlocking the immense amount of ecological and biodiversity data in digitized herbaria. Underlying each of these topics is an interdisciplinary approach which leverages computational and quantitative methods, including machine learning, computer vision, statistical modeling, and bioinformatics, to formulate novel strategies and reveal insights towards fundamental questions in genetics, ecology, and evolution.

1.2 OVERVIEW OF CONTRIBUTIONS

Chapter 2 addresses a persistent challenge in quantitative and population genetics: the precise and efficient quantification of allele frequencies in mixed population samples. We address this through the development of a novel method, FREQ-Seq^2, which leverages high-throughput next-generation sequencing technologies along with a double-barcoding library preparation strategy to achieve high accuracy, precision, and scalability at low costs. Through the application of this method to real-world experiments involving competitive experimental evolution, we demonstrate its versatility and potential for a wide range of biological contexts, from tracking candidate genes in association studies to estimating distributions of fitness effects.

Chapter 3 delves into the realm of structural variation, a relatively unexplored, yet critically important, facet of genetic diversity. A number of long-standing constraints have limited the ability to accurately and comprehensively elucidate the spectrum of structural variation. At the same time, it has suspected that a large amount of structural variation remains hidden from detection, and that these variants contribute to the extensive phenotypic diversity observed. By assembling a reference-quality genome of *Drosophila melanogaster* and creating a comprehensive map of structural variants, this chapter uncovers a wealth of hidden genetic variation with potential implications for phenotypic variation. This work underscores the importance of high-quality reference genomes in deciphering the genetic underpinnings of complex traits and sheds novel insights into the prevalence, distribution, and evolutionary implications of structural variation.

Building on the findings of Chapter 3, Chapter 4 tackles the challenge of detecting structural variants from short-read sequencing data—a task that has proven difficult due to the inherent limitations of short-read technologies and the complexity of structural variants. By developing and training a machine learning model, this chapter demonstrates the potential of sophisticated computational techniques to improve the accuracy of structural variant detection and contribute to our understanding of genetic diversity.

Finally, Chapter 5 bridges the gap between genetics, ecology, and evolution by developing a computational pipeline for the automated transcription and segmentation of herbarium specimens. By harnessing recent advances in computer vision, optical character recognition, natural language processing, and machine learning, this chapter provides a powerful tool for extracting valuable phenological data from digitized herbaria on a large scale, thereby unlocking the research potential of these invaluable collections.

CHAPTER 2

Efficient characterization of genetic variation in mixed population samples

2.1 INTRODUCTION

The accurate determination of allele frequencies is crucially important across a wide range of problems in genetics, such as developing population genetic models, making inferences from genome-wide association studies, determining genetic risk for diseases, as well as other scientific and medical applications. Furthermore, understanding how allele frequencies change over time in populations is central to ascertaining their evolutionary dynamics. Currently available sequencing technologies provide vast amounts of data describing genetic variation in a fast and cost-effective manner [76, 108]. Targeting specific alleles with methods that leverage these technologies can produce a wealth of data at modest cost with substantial sample sizes for the particular genomic regions of interest, which are comparably infeasible using traditional whole-genome sequencing or experimental assays [75, 77, 148]. Methods for accurately and efficiently quantifying allele frequencies are valuable in a wide variety of biological contexts, such as in tracking candidate genes identified in an association study, constructing and validating population genetic models, and estimating distributions of fitness effects, among other topics [92].

To tackle this challenge, we developed a precise, efficient, and economical method, named FREQ-Seq^2, for quantifying the relative frequencies of different alleles at loci of interest in mixed population samples. The original FREQ-Seq method amplifies loci of interest from mixed population samples using short customizable oligonucleotides and plasmid-based barcoded bridging primers [26]. The amplification products consist of fragments containing the DNA sequence for a query region of the genome along with a barcoded adapter sequence, where each barcode can be assigned to a specific sample, as well as Illumina sequencing adapters at each end. The resulting libraries can be sequenced to determine allele frequencies in the locus of interest without requiring additional library preparation.

A principal limitation of this approach is that every sample within a library requires its own unique barcode, and thus the construction and maintenance of a barcoded adapter plasmid, in order to generate the required bridging primer. Due to the linear scaling of library preparation labor and complexity with the number of samples in an experiment, this can quickly become infeasible for experiments requiring large numbers of samples. For example, the number of samples in data from longitudinal population studies or highly replicated experiments can easily number in the hundreds and thousands [136].

Through the application of paired barcode sequences, we exponentially increased the throughput of the FREQ-Seq concept from 48 to 2,304 samples. By applying double-barcoding, we considerably expand the available throughput and scalability. In particular, our use of two independent barcodes to uniquely label a sample produces a substantial increase in the method's scalability by allowing the number of samples to scale quadratically, rather than linearly, with the complexity of library preparation. At the same time, our method preserves the original advantages of FREQ-Seq, including the ease and flexibility of creating custom libraries for specific experiments.

We introduce a new plasmid library for preparing sequencing libraries that exponentially increases the number of possible unique labels, with minimal impact on complexity and cost. FREQ-Seq^2 can be targeted to specific genomic regions of interest, which are amplified using universal barcoded adapters to generate Illumina sequencing libraries. A FREQ-Seq^2 library consists of DNA segments spanning the locus of interest, along with two adapter sequences that are tagged with a unique pair of barcode sequences. With 48 unique sequences available for each of the two barcodes, the range of barcoded adapter fragment libraries consists of 2,304 (48^2) unique combinations that can be used to label and identify samples within a single library.

We validated the performance of our NGS-based approach with a highly multiplexed set of control samples, testing several unique combinations of barcodes on a control dataset with known target allele frequencies and quantify the accuracy, precision, efficiency, and throughput that the method achieves. Additionally, we demonstrate the real-world performance of FREQ-Seq^2 on a series of competitive evolution experiments, competing two strains of *E. coli* that differ in an inactivating single-nucleotide polymorphism (SNP) on the L-arabinose isomerase (*araA*) gene over 2,000 generations [39, 84, 135]. Then, we use FREQ-Seq^2 to label experimental samples, genotype the samples over the course of the competition assays, and analyze the data to determine change in allele frequencies and fitness over time. We compare the results of our FREQ-Seq^2 analysis to estimates obtained from manually counting colonies. Finally, we discuss the combined results of these experiments as well as the utility of FREQ-Seq^2 for tackling questions in population, evolutionary, and quantitative genetics.

Our analyses demonstrate that FREQ-Seq^2 is flexible, inexpensive, and produces large amounts of data with low error, low noise, and desirable statistical properties. The versatile analysis of highly multiplexed libraries makes it well-suited for profiling mixed populations, among other applications requiring detailed resolution of genotype distributions. Our implementation of FREQ-Seq^2 includes an available kit with two sets of 48 plasmids containing barcoded adapter fragments as well as fast and efficient open-source software for analyzing sequenced libraries, which enables detection and removal of errors that are undetectable when using a single barcode as well as other conventional methods for allele frequency quantification. In summary, FREQ-Seq^2 is a powerful method for quantifying allele frequencies within and between populations that is accurate, precise, flexible, high-throughput, and economical.

2.2 METHODS

2.2.1 Constructing a barcoded adapter plasmid library

To enable the double-barcoding in FREQ-Seq^2, we constructed an adapter library for storing the universal barcoded adapters. The FREQ-Seq^2 adapter library is stored in a plasmid vector, similar to that of the original FREQ-Seq method [26]. The library utilizes the Thermo Fisher Scientific TOPO TA PCR cloning vector for the plasmid. The 48 double-stranded adapters were generated by 48 parallel overlapping PCR reactions on the annealed template of the partially complementary single-stranded oligonucleotides following the experimental arrangement shown in Fig. 2.1, using the forward and reverse amplifying primers AAGCAGAAGACGGCATACG and GTAAGCAGTGGGTTCTCTAG, respectively, analogous to the primers ABC1 and ABC2 from the original FREQ-Seq.

Amplification of the adapter fragments for cloning was carried out with 25 PCR cycles with 15 second elongation periods. The QIAGEN Taq DNA polymerase was used in order to provide overhanging A residues required for TOPO TA cloning. The resulting 87 bp double-stranded oligonucleotides were cloned into the TOPO TA vector following the manufacturer's recommended protocol for the TOPO TA cloning kit. Half of the reaction mixture (3 μL) from each reaction was transformed into competent *E. coli* DH5α cells provided by the kit. Plasmid DNA was prepared from single white-colored colonies chosen based on blue-white selection and were confirmed via sequence by vector-specific M13f Illumina sequencing primers.

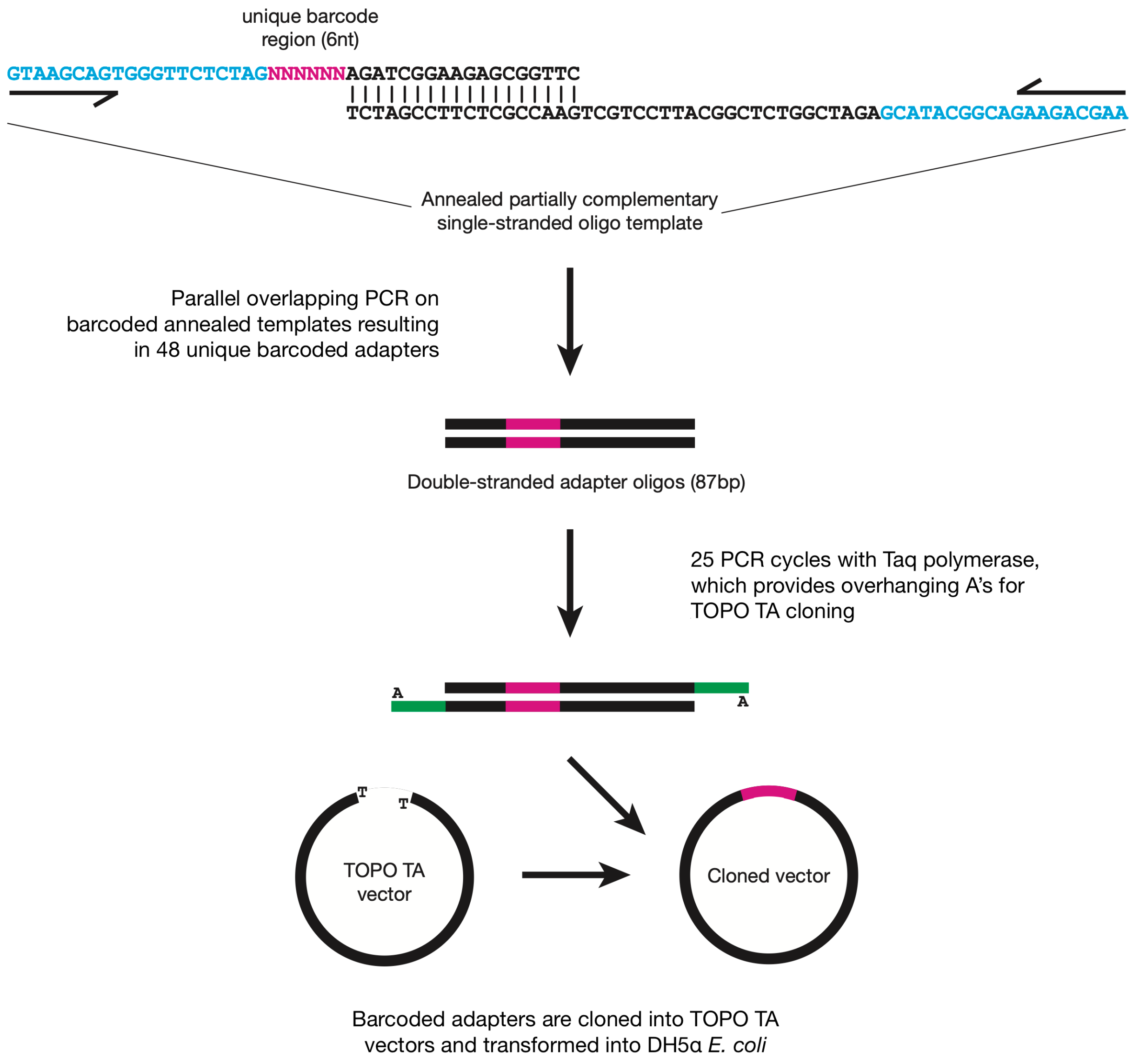

unique barcode region (6nt)
GTAAGCAGTGGGTTCTCTAGNNNNNNAGATCGGAAGAGCGGTTC
TCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTCTGGCTAGAGCATACGGCAGAAGACGAA
Annealed partially complementary single-stranded oligo template
Parallel overlapping PCR on barcoded annealed templates resulting in 48 unique barcoded adapters
Double-stranded adapter oligos (87bp)
25 PCR cycles with Taq polymerase, which provides overhanging A's for TOPO TA cloning
A
A
T
T
TOPO TA vector
Cloned vector
Barcoded adapters are cloned into TOPO TA vectors and transformed into DH5α *E. coli*

Figure 2.1: Protocol for generating the FREQ-Seq^2 adapter library. Partially complementary single-stranded oligonucleotides containing the barcodes are annealed together, extended, and PCR amplified with primers corresponding to the regions in blue. Next, they are amplified with Taq polymerase to add overhanging adenosines, for cloning into the TOPO TA vector. After cloning into the plasmids, the vectors are transformed into competent DH5α *E. coli* bacteria and plated, and plasmid DNA is extracted from the transformed bacteria.

2.2.2 Generating the FREQ-Seq^2 sequencing library

With the adapter library available, the barcoded Illumina bridging primers for paired-end sequencing can be PCR amplified from the plasmids of the adapter library, regardless of the orientation of the adapter fragment in the plasmid vector (the TOPO TA cloning of the insert is not orientation-dependent), using the same small forward and reverse amplifying primers that were used for the parallel overlapping PCR to generate the adapter fragments. These amplified adapters were gel-purified on 2% agarose gel and then used in conjunction with their corresponding original FREQ-Seq barcoded adapters for double-barcoded labeling of the fragment mixtures.

Amplification of the specific region of interest is performed using the following primers, complementary to ABC2 from the original FREQ-Seq and the FREQ-Seq^2 reverse amplifying primer described above: GTAAAACGACGGCCAGT plus a 20-nucleotide locus-specific forward primer, and CTAGAGAACCCACTGCTTAC plus a 20-nucleotide locus-specific reverse primer. The PCR reaction was carried out using the Thermo Fisher Phusion DNA polymerase, and the resulting PCR products were gel-purified on 2% agarose gel to remove unincorporated primers and diluted 100-fold for the second stage of PCR.

For barcoding of the amplified mixtures, the diluted templates were PCR amplified with Phusion DNA polymerase using 10X molar equivalent of primers ABC1 from the original FREQ-Seq and the FREQ-Seq^2 forward amplifying primer (0.1-0.2 μM) against the original FREQ-Seq and the FREQ-Seq^2 purified adapters (10-20 ng). The pooled barcoded amplification products, consisting of a proportional mixture of the sequences from different samples, constitute an Illumina-compatible library for paired-end sequencing. A final purification step (e.g., using a gel or Pippen) may be performed at this stage if desired to remove residual adapters and primers.

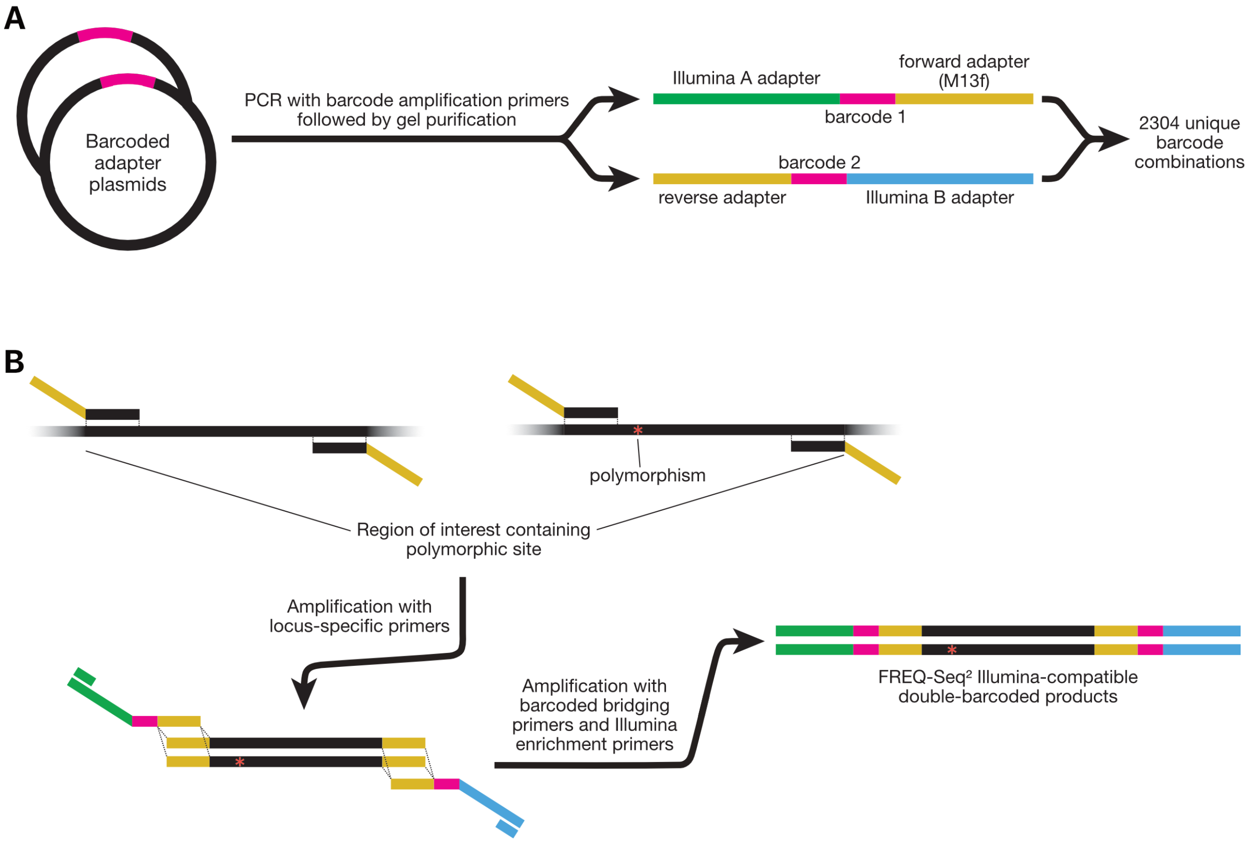

A
Barcoded adapter plasmids
PCR with barcode amplification primers followed by gel purification
Illumina A adapter
forward adapter (M13f)
barcode 1
barcode 2
reverse adapter
Illumina B adapter
2304 unique barcode combinations
B
polymorphism
Region of interest containing polymorphic site
Amplification with locus-specific primers
Amplification with barcoded bridging primers and Illumina enrichment primers
FREQ-Seq² Illumina-compatible double-barcoded products

Figure 2.2: (A) The Illumina-compatible FREQ-Seq^2 barcoded bridging primers for paired-end sequencing can be amplified from the adapter plasmids using the same amplification primers used to generate the adapter fragments. These adapters can be used in conjunction with their corresponding FREQ-Seq barcoded adapters for double-barcoded labeling of fragment mixtures. (B) To generate a FREQ-Seq^2 sequencing library, amplification is first performed using locus-specific primers to produce a pool of fragments in a region of interest. These fragments contain adapters on each end that are complementary to the barcoded bridging primers, enabling double-barcoded labeling. Amplification is then performed using the barcoded bridging primers and enrichment primers, resulting in Illumina-compatible double-barcoded products.

2.2.3 Estimating fitness of evolved strains

To examine the application of FREQ-Seq^2 in a real-world evolutionary biology application, we performed competition assays in which an evolved strain of *E. coli* was competed against an ancestral strain to estimate the adaptive trajectory of the evolved strain's relative fitness. At several time points over 2,000 generations, the evolved strain was competed against the ancestral strain, and their relative frequencies were measured and used to establish a fitness trajectory for the evolved line.

The evolved strains had previously been serially propagated for 2,000 generations in Davis minimal broth supplemented with 25 mg/L of glucose (DM25) at 42.2 °C and periodically stored as frozen glycerol stocks at -80 °C [18, 135]. These strains originated from a clone of *E. coli* B strain REL1206, which was isolated from the *E. coli* long-term evolution experiment (LTEE) and possesses an *Ara*$^-$ neutral marker [84]. REL1206 had been evolved for 2,000 generations at 37 °C in the LTEE and so was adapted to the DM25 medium. The ancestral strain used for the competitions, REL1207, is equivalent to REL1206 aside from possessing a single-nucleotide *Ara*$^+$ mutation.

For each generation, a sample of the evolved strain and of the ancestral strain were each collected on a sterile loop from frozen glycerol stock, inoculated into 10 mL of Luria-Bertani (LB) broth, and incubated at 37 °C overnight in a shaking water bath. For each strain, the culture was diluted 100-fold in phosphate-buffered saline, and 0.1 mL was transferred into 9.9 mL of DM25 and incubated at 37 °C for 24 hours. Then, 0.1 mL of each culture was transferred into 9.9 mL of DM25 and incubated at 42.2 °C for 24 hours. From their respective incubated cultures, an aliquot of the evolved strain (*Ara*$^-$) along with an aliquot of the ancestral strain (*Ara*$^+$) were transferred to a 1.5 mL centrifuge tube, and the tube was vortexed. For the colony counting samples, 0.025 mL and 0.225 mL were transferred of the evolved and ancestral lines, respectively, over six replicates. This protocol was repeated for the FREQ-Seq^2 samples, with 0.005 mL and 0.245 mL transferred of the evolved and ancestral lines, respectively, over eight replicates to optimize the utilization of a 96-well plate. The ratio of the strains in the centrifuge tubes represents the initial (prior to competition) frequencies.

We used small initial proportions of *Ara*$^-$ in order to increase the resolution and decrease the measurement error in the downstream fitness calculations, since these strains have substantially different fitness from the ancestral strain due to the adaptive environment under which the *Ara*$^-$ strain was previously propagated. In a competition assay, as the gap in the relative fitness between competing strains increases, the measurement error increases when the counts of the lower-fitness ancestor (in the denominator in the fitness calculation) become increasingly small and difficult to quantify [146]. The precision and sensitivity of FREQ-Seq^2 enabled the use of a very small initial frequency (2%) of the *Ara*$^-$ strain. A target 10% *Ara*$^-$ initial frequency was used for colony counting, as the 2% initial frequency for visual measurement was not feasible due to insufficient visual signal for the pre-competition counts.

The pre-competition mixtures were created by transferring 0.1 mL from each centrifuge tube to a culture tube containing 9.9 mL of DM25. For the colony counting samples, 0.1 mL of a 100-fold dilution from each pre-competition mixture was plated on tetrazolium arabinose (TA) agar to obtain measurements of the initial frequencies. The culture tubes were incubated at 42.2 °C for 24 hours to compete the strains. For the colony counting samples, 0.05 mL of a 10,000-fold dilution from each post-competition mixture was plated on TA agar. When plated on TA agar, *Ara*$^-$ and *Ara*$^+$ colonies appear red and white, respectively. A visual measurement of the distribution of the evolved strain versus the ancestral strain was taken by counting the plated colonies. For the FREQ-Seq^2 samples, genomic DNA from each pre-competition and post-competition mixture was extracted using the Promega Wizard Genomic DNA Purification Kit. The FREQ-Seq^2 sequencing library was prepared as described above with the locus-specific forward primer containing a 20-nucleotide flanking sequence upstream of the allele of interest and a unique combination of barcoded adapters for each sample. Following library preparation, the samples were paired-end sequenced on an Illumina HiSeq 2500 system.

2.2.4 Obtaining allele frequencies from barcoded reads

Sequencing data from a FREQ-Seq^2 library can be directly processed by our open-source software tool, *fsdm*, from the raw FASTQ files. The sequencing reads are analyzed to compare each read to the library's barcode, adapter, and allele sequences in order to identify which samples the reads belong to. These sequences can be specified by the user and are provided to the program in a FASTA file.

Reads are demultiplexed first by matching the sequence of segments at the beginning and end of each read to the barcode pairs used in the library preparation utilizing a hash table optimized for this application. The barcode pair with which each read is labeled is identified, filtering out reads that are not valid FREQ-Seq^2 reads if they lack a valid barcode combination according to the predetermined sequence information. Then, the adapter sequences and the regions flanking the query allele are extracted from the reads and their sequences are compared against the user-specified sequences. Each read is either verified as a match, up to a user-specified threshold of mismatches, or it is filtered out as an invalid read. Last, reads with an allele matching one of the possible genotypes are recorded, and reads containing unrecognized sequence for the allele are filtered out.

After filtering out reads with unmatching barcode, adapter, flanking, and target allele sequences, the counts for each allele are quantified. The relative frequencies of each allele within a given sample are obtained by dividing the read count for each allele by the total number of valid reads matching the sample's barcode pair. The software reports the computed frequency for all 2,304 combinations of FREQ-Seq^2 barcodes. In the results presented here, no mismatches were allowed in the barcodes, and a maximum edit distance of four was allowed across the adapter and flanking sequences for a read pair.

2.2.5 Demultiplexing and read rescue algorithms

Barcode sequences are identified by comparing the corresponding regions within each read pair to the set of possible barcode combinations using a fast hash table lookup. Barcode comparisons are performed for exact sequence matches as well as an optionally specified single-nucleotide mismatch threshold based on Hamming distance. In the case of allowed mismatches in barcodes, reads are only assigned to a particular barcode combination if the mismatching sequence is not ambiguous, that is, the sequence is not within the same Hamming distance to two or more possible barcodes [58].

Mismatches in the adapter and flanking regions of each read are determined using the Damerau-Levenshtein distance, an edit distance metric which accounts for substitutions, insertions, deletions, and adjacent transpositions [36, 86]. The Damerau-Levenshtein distance is computed between each of the specified adapter and allele flanking sequences and the corresponding portions of the read pair. Reads that exceed the specified edit distance threshold in the adapter and flanking sequences are discarded.

For reads that uniquely match a barcode pair and match within the edit distance threshold for the adapters and flanking sequences but fail to match to a recognized allele, a rescue algorithm is employed to find and genotype reads which contain shifted sequences due to a small insertion or deletion. The reference flanking sequence to the left of the allele (in $5'$ to $3'$ orientation) is aligned to the corresponding region in the read using a Needleman-Wunsch optimal global sequence alignment [103]. If the alignment contains an overhang, indicating the presence of a small indel in the read, the shift is corrected by reindexing the read according to the length of the overhang.

The allele position of the read is once again queried and recorded if it matches a recognized allele sequence.

2.2.6 Calculating relative fitness

The relative fitness of the evolved *Ara⁻* strain is calculated as

$$w_E = 1 + s, \tag{2.1}$$

where s is the selection coefficient, expressing the degree to which the evolved strain is fitter than the ancestral strain with relative fitness $w_A = 1$ [53, 138]. The selection coefficient describes the magnitude of the change in the strains' relative abundance per generation, as the fold-change in the count of the fitter strain is $1 + s$ times the fold-change of the less fit strain after each generation, and thus reflects the rate at which the strains' frequencies change. Over t generations,

$$\frac{E_t}{A_t} = \frac{E_0}{A_0}(1+s)^t,$$

$$t \log(1+s) = \log \frac{E_t/A_t}{E_0/A_0}, \tag{2.2}$$

where E and A represent the absolute frequency of the evolved and ancestral strains, and subscripts 0 and t represent the initial time point and the time at which fitness is estimated. As $\log \frac{E}{A} = \log \frac{f_E}{f_A}$, where f is the relative frequency,

$$\log(1+s) = \frac{1}{t} \log \frac{f_{E,t}/f_{A,t}}{f_{E,0}/f_{A,0}}, \tag{2.3}$$

or in the continuous-time case,

$$\begin{aligned}\log(1+s) &= \frac{d}{dt}\left(\log\frac{E_t}{A_t}\frac{A_0}{E_0}\right) = \frac{d}{dt}\left(\log\frac{E_t/E_0}{A_t/A_0}\right) \\ &= \frac{d}{dt}\left(\log\frac{E}{A}\right) = \frac{d}{dt}\left(\log\frac{f_E}{f_A}\right)\end{aligned} \tag{2.4}$$

For bacteria which reproduce by binary fission, the number of generations t can be calculated as $\log_2 \frac{N_t}{N_0} = \log_2 100$ based on a 100-fold dilution from stationary phase at the start of each competition assay.

2.2.7 Power law fitness model

To verify that the difference in the magnitude of *Ara*⁻ allele frequencies derived from colony counting compared to the FREQ-Seq^2 data is a result of the higher initial frequencies in the colony count samples, we compared the *Ara*⁻ frequencies measured after competition at each time point to those under a power law model of fitness increase under asexual adaptation in a constant environment.

Previous research on data obtained from the *E. coli* LTEE has demonstrated that the trajectory of relative fitness increase is well-described by an offset power law relating mean fitness as a function of time in generations. The power law fitted to a subset of the data from a set of LTEE populations accurately predicts later measurements [83, 147]. Power-law dynamics are consistent with the impact of clonal interference and diminishing-returns epistasis on the evolutionary dynamics of the evolved strain's fitness. This power law relationship can be expressed in the form of

$$\overline{w} = (at+1)^b, \tag{2.5}$$

where $\overline{w}$ represents mean relative fitness and t represents time in generations, with two model parameters a and b representing the magnitude of clonal interference and diminishing-returns epistasis, respectively [147].

We fitted the above model to the fitness trajectory for the FREQ-Seq^2 samples with non-linear least squares regression using the Levenberg–Marquardt algorithm. This power law model allows us to estimate the expected post-competition frequencies for the colony count samples, conditioned on their initial frequencies and with a model derived from an independent dataset obtained via an independent method with separate initial conditions.

We used the fitted model to obtain model predictions of mean fitness for the *Ara*$^-$ strain at each time point. Using these fitness predictions and the measured initial frequencies for the colony counts, we solved for the expected post-competition frequencies in accordance with Eq. (2.3) at each time point:

$$\widehat{\overline{w}}_E = \left(\frac{\dfrac{\overline{f}_{E,t}}{1-\overline{f}_{E,t}}}{\dfrac{\overline{f}_{E,0}}{1-\overline{f}_{E,0}}} \right)^{1/t}, \tag{2.6}$$

where $\widehat{\overline{w}}_E$ represents the model predictions of mean relative fitness of the evolved strain.

To examine the evolutionary dynamics of the evolved strain's genetic background, we can look at degree to which the scale of its distribution of fitness effects changes over time. As the emergence of a beneficial mutation can be treated as a Poisson process, mutations conferring a fitness advantage are exponentially distributed

$$s \sim \mathrm{Exp}(\lambda), \tag{2.7}$$

such that the mean fitness advantage $\mathrm{E}[s] = 1/\lambda$, where λ is the distribution's rate parameter [147]. Due to diminishing-returns epistasis, the phenomenon by which the fitness effect of a beneficial mutation is attenuated when it occurs on a fitter genetic background [151], the mean of the distribution of positive fitness effects decreases with each successive fixation of a beneficial mutation during the course of evolution. For n fixed beneficial mutations,

$$\lambda_{n+1} = \lambda_n(1 + g\, s_{n+1}), \tag{2.8}$$

where the factor g represents the effect of diminishing-returns epistasis, which has been shown to relate to parameter b of the power law by $g = 1/(2b)$ [147].

2.2.8 Statistical analyses

Competitions between the *Ara*$^+$ and *Ara*$^-$ strains were performed independently, and were measured using both the FREQ-Seq^2 and colony counting methods. We used an analysis of variance (ANOVA) to test for a correlation between the error in estimated allele frequency and barcode. We additionally used a two-way ANOVA to check for the existence of interactions between the method of determining frequency and the replicates at each time point. A significance level of 0.05 was used. Confidence intervals were estimated using a nonparametric empirical CDF-based method, which does not assume that the data follows a particular distribution, and the standard error of the mean.

2.3 RESULTS

2.3.1 Accuracy and precision

To test the accuracy of FREQ-Seq^2, we generated libraries from control samples with known relative DNA concentrations and compared the allele frequency estimates obtained with FREQ-Seq^2 to the target values for each sample. Our test dataset is comprised of 96 control samples consisting of combinations of eight separate barcodes for the first adapter with each of twelve barcodes for the second adapter. Additionally, we tested four different frequencies of *Ara*$^+$ set at

0.1, 0.45, 0.55, and 0.9. The estimated allele frequencies for all the control samples compared to their target values are shown in Fig. 2.3.

Variance in FREQ-Seq^2 allele frequency estimates is small and tightly clustered near the target frequency for a broad range of values. The average error in allele frequencies estimated using FREQ-Seq^2 in the control samples was 1.47%, with a standard deviation of 0.73%. Note that this estimate of error accounts for not just the variance in the method itself, but also external sources of error, such as sequencing error, contamination, and pipetting error introduced in creating the test samples. Error statistics for the control samples are summarized in Table 2.1. To investigate whether the method exhibits biases, we examined the distribution of errors and looked for the existence of correlated errors, as deficiencies in these metrics can indicate systematic bias in PCR amplification or sequencing [4, 117]. The error was not correlated with the barcode sequences at either of the possible positions and is close to normally distributed.

Error	**Percent**
Average error	1.47%
Minimum error	0.02%
Maximum error	2.91%
Standard deviation	0.73%
95% confidence interval	(0.12%, 2.72%)

Table 2.1: Error in control sample allele frequency estimates.

2.3.2 Applying to real evolutionary biology samples

To evaluate the performance of FREQ-Seq^2 with real biological samples, we used the method to obtain allele frequency estimates over evolutionary time for a competition experiment between two strains of *E. coli* that differ at a SNP in the *araA* gene. These estimates were then used to compute the fitness trajectory of this experiment. The *araA* gene encodes the L-arabinose isomerase protein, and is part of the L-arabinose operon. One of the strains we use (*Ara*$^-$) possesses an inactivating SNP in the gene [28], and is routinely used as a neutral visible marker in exper-

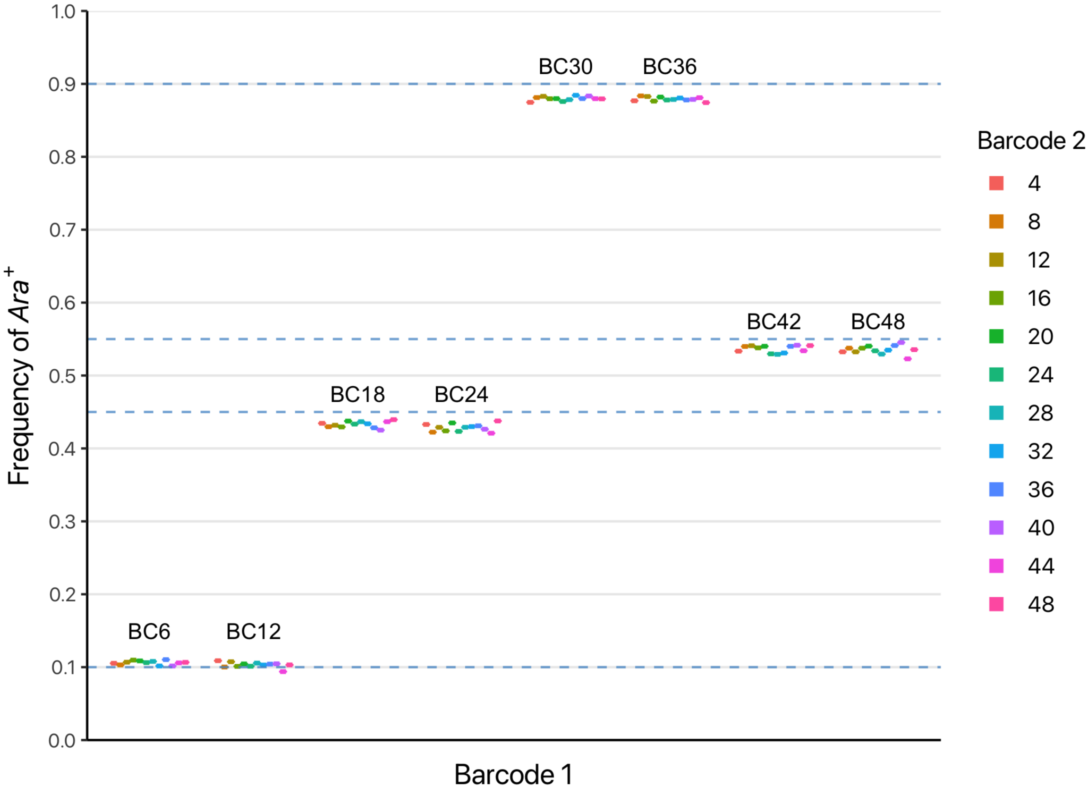


Figure 2.3: Estimated *Ara*$^+$ allele frequencies using FREQ-Seq^2 for 96 independent loading controls with unique barcode combinations. Dashed blue lines represent the four target allele frequencies of *Ara*$^+$ that were used to benchmark the controls.

imental evolution studies [84]. Two independent competition assays were performed, in which several independent aliquots of the *Ara*$^{+}$ strain and evolving *Ara*$^{-}$ strain were taken and amplified together at each of eleven evolutionary time points spaced over the course of 2,000 generations.

We used FREQ-Seq^{2} to determine the allele frequency for both of the strains at each time point, and then estimated relative fitness based on the allele frequency estimates using the method described by Lenski *et al.* [84]. The frequency and fitness trajectories for the competitions are shown in Fig. 2.4a and 2.4b, respectively. The *Ara*$^{-}$ allele frequency and relative fitness both increase steadily over the 2,000-generation experiment and on average exhibit near-monotonic upward trajectories. Notably, with a comparatively small number of samples and generations, the characteristics of the FREQ-Seq^{2} frequency and fitness trajectories in our *E. coli* competition assay resemble those of the extensive *E. coli* long-term evolution experiment [85, 147].

The observed variation in fitness trajectories among the different samples at each time point is not necessarily surprising. First, noise inherent in the various steps of a competition assay produces some degree of variation between samples. Second, stochasticity in the traversal of rugged evolutionary fitness landscapes naturally causes rises and dips in fitness on the path towards an optimum [120]. This principle regarding evolutionary trajectories with respect to fitness landscapes, including individual sample variation in frequency and fitness at each time point measured in our experiment, has been observed in a wide range of experiments [30, 59, 120]. Additionally, the mean fitness measured in our samples exhibits an initial increase within the first few hundred generations, followed by an eventual deceleration in the fitness increase over time, which is consistent with theoretical expectations as well as the results of long-term studies in experimental evolution [39, 49, 106].

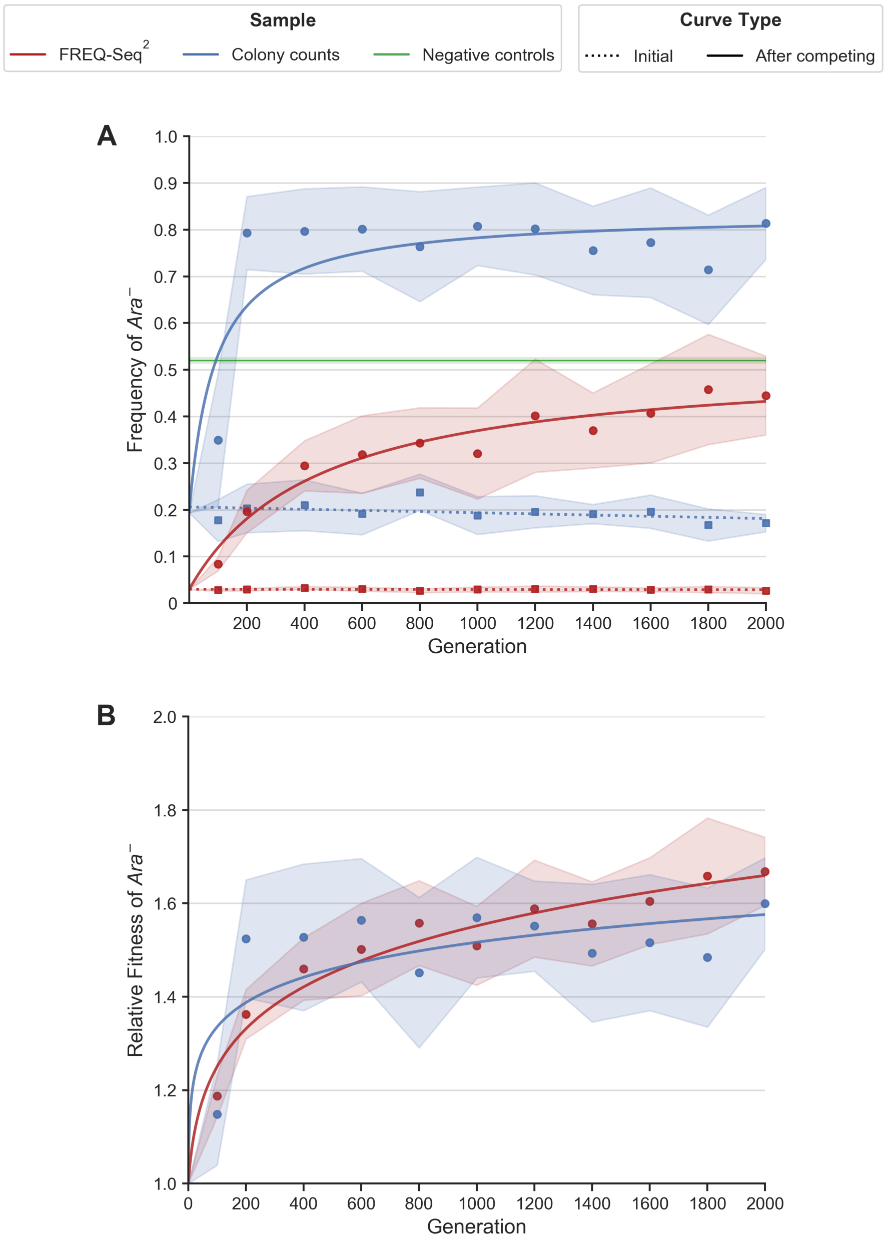


Figure 2.4: FREQ-Seq^2 allele frequency and fitness trajectories over time for the evolved Ara^- strain. The Ara^- strain was competed with the ancestral Ara^+ strain, and their frequencies were measured at several time points over 2,000 generations. (A) Ara^- allele frequency and (B) relative fitness across eleven generations of the competition assay measured using both FREQ-Seq^2 and manual colony counting. The blue and red dots represent the mean allele frequency or relative fitness at each time point. In (A), the dotted lines correspond to the initial Ara^- frequency before the strains were conducted and the solid lines correspond to the Ara^- frequency after competing. The line and curves show the fit of a linear, hyperbolic, and power law model to the initial frequencies, post-competition frequencies, and fitnesses, respectively. Note that the higher magnitude of the Ara^- frequencies for colony counting are due to the higher initial frequencies. The green line is the mean allele frequency measured using FREQ-Seq^2 for sixteen independent target 50/50 negative controls. The shaded regions represent 95% confidence intervals based on the standard error of the mean.

2.3.3 Comparing to manual quantification methods

We compared the estimates of allele frequencies and fitness determined using FREQ-Seq^2 to those computed by manual colony count measurements. Plating and competitions were performed at the same eleven time points that were used for the sequenced data. Colony counts of each allele were obtained at each generation for all samples and replicates. The mean allele frequency trajectories of *Ara*$^-$ determined by the FREQ-Seq^2 and colony counting methods are shown alongside each other in Fig. 2.4a.

Compared to the estimates of allele frequency and fitness determined by colony counts, the FREQ-Seq^2 data produced more stable measurements for both frequency and fitness, as well as trajectories that more closely match predictions from theory for a population adapting to a fixed environment over time [34, 49]. This was particularly true for relative fitness, where the estimates derived from manual counting exhibited much less stable measurements over time along with substantially higher variance (Fig. 2.4). The FREQ-Seq^2 fitness measurements produce a trajectory that exhibits the gradual reduction in the average rate of fitness increase over time characteristic of classic adaptive walks, following an initial increase before generation 400 [59, 106]. Additionally, the mean fitness trajectory is consistent with the expected presence of clonal interference and diminishing-returns epistasis, with the latter's estimated effect in line with estimates from the LTEE [147].

The substantially higher magnitude of the post-competition allele frequencies for the colony counts versus FREQ-Seq^2 is a predictable consequence of the initial *Ara*$^-$frequency in each of the experiments (Fig. 2.4a). We confirmed this is the case by considering that the adaptive dynamics of fitness in clonal populations is consistent with a power law relationship of mean fitness as a function of time in generations (Eq. 2.5) [147]. Fitting this power law to the fitness trajectory derived from the FREQ-Seq^2 data, computing the relative fitness predicted at each time point using the fitted model, and then solving for the expected post-competition allele frequencies in the colony count samples given their initial frequencies shows that the higher magnitude in the observed post-competition frequencies tracks with expectations.

The greater variance and increased jaggedness in the colony count-based allele frequency and fitness estimates illustrate a major practical benefit of FREQ-Seq^2's accuracy and precision, particularly with smaller numbers of samples and degrees of replication. Though this may be mitigated to a degree with larger datasets and increased replication, such changes entail additional costs and labor, or may not be readily available depending on the difficulty in obtaining and preparing samples. As a quality control measure, a negative control sample targeting a 50/50 distribution of *Ara*$^+$ and *Ara*$^-$ was included with each group of FREQ-Seq^2 samples from our competition assay during library preparation, with sixteen independent negative controls in total. The frequency measurements for the initial frequency at each time point in combination with the negative controls demonstrate that substantial variations between different samples or time points are unlikely to be an artifact of the FREQ-Seq^2 method itself. Both the negative controls and initial frequencies are extremely consistent, fall within a very narrow range of variation, and closely track the target aliquot ratio across all the time points. No statistical interaction was observed between the different replicates at each time point and the method used to measure allele frequency.

2.3.4 Coverage, noise, and resolution

We used the control samples to evaluate the random variation of our method. We compared our frequency observations in these controls to the frequencies among barcode combinations that were not introduced into the experiment. These combinations represent a class of false positives against which we measure the intended barcode combinations. The false positive barcode combinations are divided into two categories. The first is for combinations matching two possible combinations of barcodes that actually exist in the library. The second represents the case where either barcode matches a barcode actually in the library but not both. Notably, the single spurious match category is an exaggerated estimate of the degree of barcode hopping given that contaminating a real category would require matching both barcodes. Nevertheless, we present these results as a conservative upper bound demonstrating how rare errors are. Thus, counts in the first category correspond to an upper bound for the risk of misidentifying a particular sample based on a specious barcode pair, derived from fragments of one or more samples that were erroneously barcoded at any point prior to sequencing [134].

When the frequency of these spurious barcodes approaches that of the lower coverage control samples among the expected barcode combinations, the risk of undetected error increases for allele frequency estimates in these lower coverage samples. Comparison of the three categories (two spurious barcode combinations and the true barcode combinations expected from control groups) shows that the potential for contamination via barcode misassignment is quite low, with the distribution of the single spurious barcode category not overlapping that of the true category. The two spurious distributions share a substantial degree of overlap, and neither class of errors represents an appreciable risk of confounding (Fig. 2.6a).

The coverage for the 96 FREQ-Seq^2 barcode pairs in our control and experimental samples (a total of 192 unique combinations) are visualized in Fig. 2.5. Different samples in our library obtained a range of coverage levels, though we did not observe any particular barcode being associated with unusually low or high efficiency. Additionally, the lowest-coverage sample in our library, which was sequenced using a small fraction of a single lane, still produced a read count in the thousands, with the largest sample size reaching well over 10,000. The vast majority of FREQ-Seq^2 reads are uniquely identified as one of the true combinations in the sequencing library. The coverage of the expected barcode combinations was substantially higher than that of any spurious combinations when comparing the across all 2,304 possible barcode pairings. In fact, the coverage of erroneous barcode combinations only approaches within an order of magnitude of the coverage of properly barcoded reads at the very bottom of the coverage distribution (Fig. 2.6a).

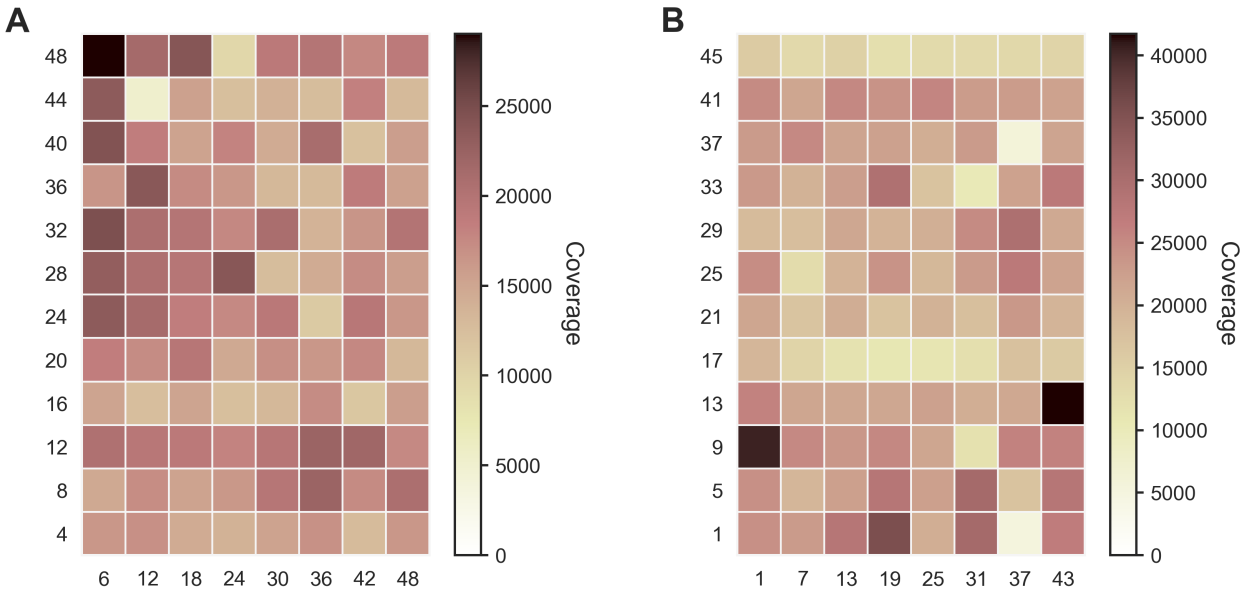


Figure 2.5: Sequencing read coverage measured for the FREQ-Seq^2 barcode combinations used in the control and experimental samples. Different sets of 96 distinct barcode pairs were used to label the loading controls and experimental evolution samples, which are clearly identifiable by coverage from the background noise. The labels on the x-axis and y-axis show the first and second barcodes used to label each of the 96 sample barcode pairs in each heatmap for (A) control samples and (B) experimental evolution samples. Coverage for barcodes outside the barcode combinations used for sample labeling represent spurious signal from noise in the method or errors during preparation and sequencing.

2.3.5 Throughput, efficiency, and scalability

To evaluate the throughput and scalability of the FREQ-Seq^2 method, we first looked at the distribution of reads that contain a matching barcode pair but do not contain a proper allele. This statistic examines the accuracy of the barcoding protocol itself, and therefore, the likelihood of correctly identifying a particular sample based on FREQ-Seq^2 reads. Fig. 2.6b shows the coverage ratio of reads containing a proper allele to those with a mismatched allele for the set of reads with a matching barcode pair. This represents a desirable result, as the number of erroneous reads is far lower than the number of reads with a proper allele for all 96 barcode combinations. Additionally, the distribution does not show a correlation between barcode and error rate.

Next, we examined the frequency distribution of sequencing reads generated from our libraries. Specifically, we investigated the rank-frequency distribution of reads that contain both a valid combination of FREQ-Seq^2 barcodes and a matching target allele sequence, representing the true positives. This statistic is useful for evaluating the method's effective throughput relative to total coverage, as it looks specifically at the reads that are usable for downstream analyses. To gauge the representation of usable reads compared to erroneous reads, we computed the worst-case coverage ratio of the true positive samples to the highest-coverage erroneous barcode combination comprised of two individually valid barcodes. Mirroring the results in Fig. 2.5, the rank-frequency distribution indicates that the method produces comparable and substantial coverage for the vast majority of samples in a library, and thus will scale well by simply increasing the library size until a desired sample size is reached.

Finally, to evaluate the overall efficiency of the FREQ-Seq^2 pipeline, we examined the overall rate of useful reads generated from our barcoded library. We first extracted the set of all reads from the raw sequencing output that were associated with the FREQ-Seq^2 library sequences. This broader set of FREQ-Seq^2-associated reads was defined as any read pair that contained two individually recognizable barcodes, regardless of whether or not the particular combination was valid, as well as adapter and flanking sequences which each matched the respective reference sequence within an edit distance of four. The rate of usable reads was calculated as the proportion of reads which contained a valid barcode combination and allele and passed the quality control thresholds for matching adapter and flanking sequences, out of the total number of reads derived from the FREQ-Seq^2 library. After filtering and demultiplexing the reads as described in the Methods, the proportion of useful reads in our control samples was over 91%.

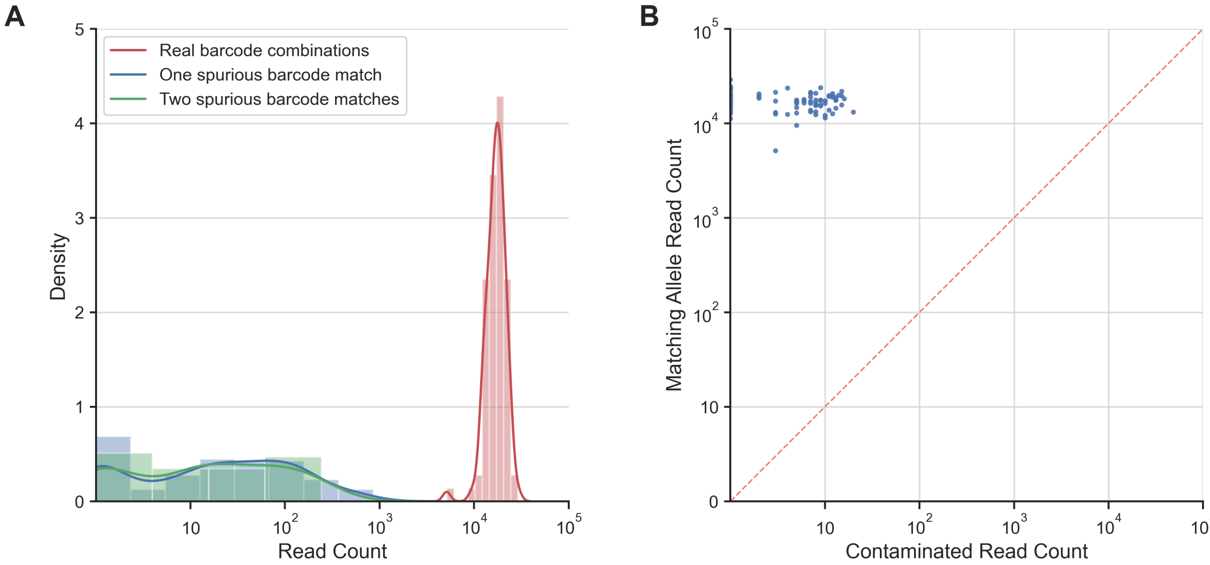


Figure 2.6: (A) Histograms comparing the coverage of properly barcoded reads to that of reads with either one or two improper barcodes for 96 unique control sample barcode combinations. The distributions of one and two spurious barcode matches represent the relative risk of misbarcoding in a FREQ-Seq^2 library. (B) Coverage of reads containing a valid genotype (y-axis) versus the coverage of contaminated reads containing an unrecognized allele (x-axis) among properly barcoded control sample reads for each of the 96 barcode combinations. The dashed red line is a one-to-one scaled diagonal between the axes.

2.4 DISCUSSION

Traditional quantification of allele frequencies by counting colonies is laborious and time-consuming due to the nature of the methods and the sheer number of individual measurements required [68, 101, 111]. Our results show that FREQ-Seq^2 is an effective method for bypassing these problems, while simultaneously improving throughput, repeatability, and cost efficiency. Our method significantly improves upon the scalability of its predecessor, enabling highly multiplexed sample combinations to be analyzed in a single sequencing library, while retaining the original benefits such as simple library preparation and precise quantification.

The barcode redundancy in FREQ-Seq^2 ensures a high degree of accuracy and minimizes the false positive rate for detecting a given allele. In the hundreds of samples comprising our present results, the great majority of datapoints produced by the method consists of true positives, that is, reads that contain two correct barcodes as well as one of the expected alleles of the target gene. This represents the desired signal, as this indicates that a read corresponds uniquely to one of the barcode combinations with which the library was prepared.

One way to evaluate the efficiency of a method like FREQ-Seq^2 is to compare the level of each true positive signal to that of the single highest-coverage erroneous group of reads, in which each of the reads' two individual barcodes are present in the library but are not expected in that particular combination. This provides a useful worst-case noise component as a basis for evaluating the impact of barcoding and sequencing errors on the method, because it judges accuracy with respect to the most highly represented class of erroneous reads which actually presents a risk of confounding the analysis of a particular sample [134]. The closer the coverage of this error signal is to that of proper reads which uniquely identify a real sample based on a matching barcode pair, the less confidence one has that a particular sample has been accurately measured. Our data demonstrates that FREQ-Seq^2 performs exceptionally well in this respect.

In evaluating the resolution of FREQ-Seq^2, it is also useful to note that this noise component is in fact a conservative estimate of the overall error in the dataset. This is due to the fact that many samples exhibit a far lower degree of error than the worst-case, which is based on the coverage for the highest observed erroneous sample that poses a Type I or Type II error risk to any one of our 96 true samples. Indeed, most of the barcode pairs in our samples do not have any barcode in common with this group of spurious reads.

This particular metric does not have any overlap with the various types of obviously erroneous reads, for example, reads that do not contain two individually valid barcodes and aberrant reads due to a sequencing, PCR, or ligation error or some other library preparation issue. For these more forgiving classes of errors, the redundancy inherent in FREQ-Seq^2 allows for unambiguous identification and filtering of erroneous reads. The data show that these error components, despite collectively comprising the most diverse class of non-useful reads, are by and large so low in frequency as to be negligible compared to true positives (Fig. 2.6a). Additionally, they can be clearly identified and distinguished from a valid FREQ-Seq^2 read (i.e., reads with a barcode pair corresponding to a known combination in the library), so they can be easily and reliably filtered out from a dataset.

Since the FREQ-Seq^2 adapter library enables 48^2 distinct barcode combinations, one can run a very large number of combinations on a single lane of a modern sequencer, providing the latitude and throughput to discard noisier read groups if desired without being constrained by the number of unique identifiers that can be assigned to different samples. Alternatively, replicates of the same libraries can be differentially barcoded to increase and balance sample sizes. In applications where high sensitivity is required, natural random variation in the coverage among samples labeled with particular barcode combinations can be mitigated using such strategies [98, 123], as the variation in sample size for different barcode combinations within a given run of the sequencer is in principle random. This is a particularly important characteristic in an NGS-based method, as undetected barcoding and amplification biases can confound inferences based on coverage and degrade library performance and consistency [7, 35].

Indeed, the large sample sizes obtained from this method are another major advantage, one which will only increase with improvements in the read counts and base-pair accuracy of sequencing technologies. Because FREQ-Seq^2 libraries are prepared such that every read ideally contains two independent barcodes that uniquely identify a sample, in addition to known adapter sequences and an allele at the target locus, every read from the raw output of a sequencer is a potentially usable sample. The efficiency of the method is limited only by the precision of the library preparation and sequencing process itself. In real data, some reads must be discarded due to errors and noise, such as in cases where one or more barcodes do not match or where no target allele is present, and here the large sample sizes combined with the barcode redundancy of FREQ-Seq^2 are advantageous.

Out of the 96 barcode combinations in our control samples, the most efficient sample had an effective sample size of over 29,000, which was produced from a small fraction of a single lane on a run-of-the-mill short-read sequencer. Additionally, the lowest coverage sample still had a sample size in the thousands. This indicates that one could further scale the library to contain many more samples than the 96 we included and achieve a higher sample size for each combination than would be possible using traditional quantification methods, without an increase in cost or sequencing resource usage [142, 145, 148]. At the current levels of sequencing throughput and cost, outstanding quantities of high-precision measurements can be achieved for relatively modest sums [108].

Our results show that, compared to a manual approach to estimating allele frequencies by counting colonies, the FREQ-Seq^2 method produces much more stable trajectories, while successfully reproducing a qualitatively similar trend consistent with both theory and empirical data for clonal populations evolving towards a fitness peak [55]. The fitness trajectories are likewise qualitatively similar, and we observe similar final values between the two methods. In our evolution experiments, the FREQ-Seq^2 data exhibits a markedly smoother trajectory for both frequency and fitness across several time points over 2,000 generations. Combined with the small magnitude and uncorrelated nature of its error, FREQ-Seq^2 provides a substantial reduction in error and increase in precision compared to manual counting colonies. This is not surprising, as the method eliminates unpredictable sources of human and experimental error [68] while at the same time massively boosting sample sizes.

The allele frequencies at particular loci of interest in a given population can have major effects on the accuracy and outcome of biological inferences, which can go undiscovered if the frequencies are not precisely quantified. For example, it has been shown that the minor allele frequency of a candidate SNP in a genome-wide association study can have a large impact on the likelihood of obtaining a false positive result [133]. Additionally, inaccuracies in the determination of allele frequencies in a sample can substantially confound the results and analysis of studies into gene regulatory architecture, population and evolutionary genetic inference, *cis/trans*-variation, and allele-specific expression, among other major topics of active research [119, 127, 156]. Our error analysis illustrates how numerous false positives and false negatives can go undetected without adequate redundancy and sample size, often at rates surpassing common thresholds for statistical significance in large datasets [47].

FREQ-Seq^2 represents a versatile tool for supplementing and validating results and inferences in applications such as high-throughput genetic experiments, long-term evolution studies, genome-wide association studies, allele-specific expression studies, as well as other applications across population, evolutionary, and quantitative genetics.

CHAPTER 3

Uncovering hidden structural variation in functional elements of *Drosophila*

3.1 INTRODUCTION

Mutations underlying phenotypic variation remain elusive in trait-mapping studies [115] despite the exponential accumulation of genomic data, suggesting that many causal variants are invisible to current genotyping approaches [44, 95, 99, 149]. In fact, mutations like duplications, deletions, and transpositions are systematically under-represented by standard methods, even as a consensus emerges that such structural variants (SVs) are important factors in the genetics of complex traits [5, 44, 45]. Addressing this problem requires compiling an accurate and complete catalog of the genomic features that are relevant to phenotypic variation, a goal most readily achieved by comparing nearly complete high-quality genomes [5]. Although the development of high-throughput short-read sequencing led to a steep drop in cost and a commensurate increase in the pace of sequencing [1], it also led to a focus on single-nucleotide changes and small indels [51, 149]. Paradoxically, this has also resulted in deterioration of the contiguity and completeness of new genome assemblies, due primarily to read-length limitations [6].

Mutations that add, subtract, rearrange, or otherwise refashion genome structure often affect phenotypes, although the fragmented nature of most contemporary assemblies obscures them. To discover such mutations, we assembled the first new reference-quality genome of *Drosophila melanogaster* since its initial sequencing. Here we present the reference-quality assembly of a second *D. melanogaster* strain called A4 and introduce a comprehensive map of SVs, which identifies a large amount of hidden variation exceeding that due to SNPs and small indels, and which includes strong candidates to explain complex traits. By comparing this new genome to the existing *D. melanogaster* assembly, we created a structural variant map of unprecedented resolution and identified extensive genetic variation that has remained hidden until now. Many of these variants constitute candidates underlying phenotypic variation, including tandem duplications and a transposable element insertion that amplifies the expression of detoxification-related genes associated with nicotine resistance. The abundance of important genetic variation that still evades discovery highlights how crucial high-quality reference genomes are to deciphering phenotypes.

3.2 METHODS

3.2.1 DNA sequencing and genome assembly

A4 DNA was extracted from females and used in SMRTbell library preparation as described previously [20], yielding 18.7 Gb of sequence. We then followed the method described previously [20] to assemble the A4 genome. We assembled a draft genome using PBcR-MHAP [12] in wgs 8.3rc1 and PacBio reads (NG50 = 13.9 Mb, 147 Mb in total; NG50 is the contig length such that 50% of an assumed assembly size is contained within contigs of this length or longer) and then generated a hybrid assembly with DBG2OLC [152] using the longest 30x PacBio reads and 75x paired-end Illumina reads from ref. [63] (assuming a genome size of 130 Mb; NG50 = 4.23 Mb, 129 Mb in total). We merged the two assemblies using quickmerge v0.1 with default settings, except hco = 5, c = 1.5, and l = 2 Mb. The merge yielded an assembly (NG50 = 21.3 Mb, 130 Mb in total) that was both smaller than expected [63] and smaller than the PacBio-only assembly. Therefore, we added contigs that were unique to the PacBio assembly to the hybrid assembly using quickmerge as described above but with I = 5 Mb. Finally, we generated the final assembly by running finisherSC [80] with default settings, polishing the assembly twice with quiver (SMRT Analysis v2.3), and with Pilon v1.3 [141] (using A4 reads from ref. [63]). This yielded a final assembly of 144 Mb with N50 = 22.3 Mb.

A4 embryos less than 12 hours old were collected on Petri dishes containing apple juice and agar, dechorionated using 50% bleach, rinsed with water, and stored at −80 °C. DNA was extracted from frozen embryos using the Animal Tissue DNA Isolation kit (Bionano Genomics). Bionano Irys optical data were generated and assembled with IrysSolve 2.1 at Bionano Genomics. We then merged the Bionano assembly with the final assembly contigs using IrysSolve, retaining Bionano assembly features when the two assemblies disagreed.

The scaffold for the A4 assembly was prepared with the software mscaffolder (see URLs) using the release 6 *D. melanogaster* genome (r6.09) assembly [62] as the reference. Prior to scaffolding, TEs and repeats in both assemblies were masked using default settings for RepeatMasker (v4.0.6). The repeat-masked A4 assembly was aligned to the repeat-masked major chromosome arms (X, 2L, 2R, 3L, 3R, and 4) of the *D. melanogaster* ISO1 assembly using MUMmer [78]. Alignments were further filtered using the delta-filter utility with the -m option, and the contigs were assigned to specific chromosome arms on the basis of the mutually best alignment. Contigs showing less than 40% of the total alignment for any chromosome arms could not be assigned a chromosomal location and therefore were not scaffolded. The mapped contigs were ordered on the basis of the starting coordinate of their alignment that did not overlap with the preceding reference chromosome–contig alignment. Finally, the mapped contigs were joined with 100 Ns, a convention representing assembly gaps. The unscaffolded sequences were named with a 'U' prefix.

We used BUSCO (v1.22) [122] to evaluate the completeness and accuracy of the A4 and ISO1 release 6 assemblies. ISO1 contains five BUSCOs (BUSCOaEOG75R3J9, BUSCOaEOG7SJRJ9, BUSCOaEOG7SJRK2, BUSCOaEOG7WMR0H, and BUSCOaEOG71S8ZH) that are missing from the A4 assembly. To validate the absence of these five BUSCOs in the A4 assembly, the full-length sequences of the ISO1 genes (*Ftz-f1*, *CG7627*, *Raw*, *Maf1*, and *Cv-c*) were downloaded from FlyBase [42] and queried against the A4 assembly with MUMmer. MUMmer found all five 'missing' BUSCOs in the A4 assembly in single copies. The BUSCO counts for A4 were adjusted accordingly.

3.2.2 Detecting structural variants via whole-genome alignment

We aligned the ISO1 and A4 assemblies using MUMmer (mummer -mumreference -l 20 -b) and then clustered maximal exact matches (MEMs) between the two mgaps (mgaps -C -s 200 -f 0.12 -l 100). The -l parameter in mgaps was set to 100 to detect duplicates that were 100 bp or longer. We used a pipeline called svmu (structural variants from MUMmer) to automate CNV detection from overlapping mgaps clusters. When reference sequence regions in two separate alignment clusters overlapped, the overlapping segment of the reference sequence regions was inferred to be duplicated in the query sequence. This approach can also identify (i) a duplicated sequence that is present in both the genomes but has diverged owing to the presence of repeats or indels and (ii) CNVs containing TE sequences. We filtered the latter using RepeatMasker (v4.0.6). We identified false-positive duplication calls by aligning the putatively duplicated reference sequences back to the ISO1 and A4 genomes using nucmer (nucmer –maxmatch –g 200) and then counting the copy number using checkCNV, which is also included in the svmu pipeline. svmu was run with the default parameters; checkCNV was run with c = 500 (max copy number 500), qco = 10,000 (10 kb of insertion or deletion allowed within a copy), and rco = 0.2 (unaligned length of up to 20% of the sequence length between reference and query copies allowed). CNVs occurring within 2 kb of each other were designated as complex events and combined (bedtools merge –d 2000)

[114] for the purpose of counting the total number of CNVs present in the genome. However, the total sequence affected by CNVs was counted before merging. Functional annotation of CNVs was based on gene annotation of ISO1 release 6.

Insertions (>100 bp) in the A4 genome appear as alignment gaps between two adjacent syntenic blocks when ISO1 is aligned to A4 (and vice versa). We aligned the A4 sequence to the ISO1 sequence using nucmer (default parameters) and then identified adjacent syntenic blocks with gaps >100 bp in length between them in the A4 assembly but <10% the gap length in the ISO1 assembly. Indel detection was carried out with the svmu utility findInDel. A deletion was inferred for a specific gene (e.g., *Cyp6a17*) when an ortholog of the gene was present in the closely related species *Drosophila simulans*.

We identified inversions in the A4 genome by aligning it to the ISO1 genome using nucmer (-mumreference) and then processing the outputted delta file using findInDel. A4 regions that ran in the reverse direction with respect to the ISO1 sequence were recorded as inversions. TEs were removed from this list using RepeatMasker annotations for ISO1.

SNPs and small (<100-bp) indels in the A4 assembly were identified using the show-snps utility from MUMmer. We aligned A4 scaffolds to ISO1 scaffolds using nucmer (-mumreference) and then filtered repeats using delta-filter in conjunction with the –r and –q options. SNPs and small indels were called from the filtered data using show-snps with –Clr options.

3.2.3 Genotyping CNVs, indels, and inversions using Illumina reads

Three common, complementary strategies are typically used to discover CNVs with paired-end Illumina reads: read depth, read-pair mapping orientation, and split-read mapping. We identified duplications (100 bp to 25 kb long) in the A4 genome using 70x paired-end reads [74] with CNVnator [3] for the read depth approach, pecnv [116] for the read-pair orientation approach, and Pindel [153] for the split-read mapping approach. We mapped reads to ISO1 release 6 using bwa-mem for CNVnator and Pindel and bwa-aln for pecnv [87]. We required at least three supporting read pairs for pecnv calls and used a bin size of 100 for CNVnator because of the data's high coverage. Furthermore, we used CNVnator and Pindel to identify large (>100-bp) indels and Pindel to identify inversions. We manually compared these short-read-based calls to our alignment-based CNV calls for all of chromosome arm 2L. TE insertion coordinates for A4 were obtained from DSPR. We manually compared our TE insertion calls and those from ref. [32] for all of chromosome arm 2L.

3.2.4 Validation of duplicates and indels

Dot plots between A4 and ISO1 for all SV loci on chromosome arm 2L were manually inspected to confirm the accuracy of the MUMmer-based genotyping. All manually inspected loci corresponded to the automated genotype calls. To quantify the effect of assembly errors in A4 on SV calls, we required that unassembled, corrected long reads from A4 agree with the A4 assembly in the region spanning the entire mutation. To do this, we mapped the PBcR-MHAP-corrected long reads to the A4 assembly using blasr v1.3.1.142244 (-bestn 1 –sam) and identified all of the reads that spanned the mutation-containing region with anchors in the flanking sequence of at least 250 bp on each side. For our stringent validation criteria, we required at least two fully spanning reads to overlap each SV. These fully spanning reads were required to have at least 99.5% alignment coverage and less than a ratio of 0.005 of gaps to read length. For our standard validation criteria, we permitted validation under the following relaxed criteria: (i) overlap-spanning reads (at least two on each side) that otherwise fit the stringent criteria above and (ii) fully spanning reads with at least 97.5% alignment coverage and less than a ratio of 0.025 of gaps to read length.

Half of our sequencing data were present in reads that were 17,885 bp or longer, which was enough to achieve more than 60-fold coverage across the entirety of the euchromatin and more than 10-fold coverage of the genome in reads that were 30 kb or longer. Such long reads contained unique sequences flanking each side of the mutation, as well as the mutation breakpoints and the mutation itself, making this a powerful approach to validating SV calls.

We assayed for the presence and absence of *Cyp28d1* and *p24-2* copies using PCR. We extracted DNA from 25 flies from each strain using the Magattract HMW DNA kit (Qiagen), and we used Phusion (New England Biolabs) for PCRs that had an amplification time of 15 s for the *Cyp28d1* reactions and 30 s for the *p24-2* reactions.

3.2.5 Temperature-preference assay

We created a linear temperature gradient on a solid aluminum bar (total dimensions: 24 inches by 4 inches by 4 inches) by placing 4 inches of one end of the bar inside a reservoir containing ice water (0 °C) and 4 inches of the other end inside a reservoir containing warm water (35 °C). This left ~40 cm of aluminum bar exposed between the baths. Temperatures along the bar were measured by 11 temperature sensors (Tmp36 analog temperature sensors from Adafruit) that were evenly spaced at 4-cm intervals and sealed into holes drilled into the bar after being secured with thermal epoxy (OMEGABOND 101 Two-Part Epoxy). The probes were connected to three four-channel 16-bit analog-to-digital converters (ADS1115 from Adafruit), which were in turn calibrated and monitored by a Raspberry Pi 3 single-board computer. Automated temperatures were recorded every second during the experiment to verify the stability of the gradient. The temperature measurements at the end of the experiment were used in assigning temperatures to individual flies. The temperature gradient on the aluminum bar ranged from 9 °C to 30 °C (Fig. 3.2b). We compared the preference of A4 flies, which lack the *Cyp6a17* gene, to that of w^{1118} flies (BDSC stock 5905), which have an intact copy of *Cyp6a17* [69]. We collected groups of 100 1- to 3-day-old flies of mixed sex and kept them at 25 °C for 24 hours. Before the assay, flies were immobilized with light anesthesia and placed between a thin aluminum sheet cut into the shape of the aluminum bar surface and an acrylic lid possessing a partition to create two lanes for the flies to behave without interacting with each other. Quinine sulfate was applied to the roof and walls of each channel in the lid so that the flies would avoid these surfaces and be constantly in contact with the aluminum surface. Flies were allowed to recover on the aluminum sheet in a 25 °C incubator for 40 minutes after being anesthetized. The aluminum sheet was then placed on top of the aluminum bar and left for 40 minutes in the dark. A photo was taken to record the positions of the flies on the block after 40 minutes. We recorded fly positions and interpolated their temperatures using linear regression based on temperature-probe readings.

We replicated the temperature preference assay experiment six times. Three replicates were conducted with A4 flies in lane 1 and w^{1118} flies in lane 2, and three replicates were conducted with the lane assignments reversed. We performed a nonparametric Wilcoxon rank-sum test, which does not assume a particular distribution for the data, on each of these six replicates to test for a difference in temperature preference between the two strains. These six individual tests produced P values of 2.12×10^{-10}, 6.76×10^{-10}, 1.89×10^{-6}, 9.21×10^{-14}, 1.96×10^{-6}, and 1.25×10^{-24}. To obtain a combined P value, we performed a meta-analysis using Fisher's method, which gave a very low meta P value ($P \ll 10^{-16}$).

3.2.6 RNAi strain construction and screening

Strain 60100 (Vienna *Drosophila* Resource Center) contains two *attP* sites at 2L: 22,019,296 (near tiptop; VIE260B) and 2L: 9,437,482 (VIE260B-2). Activation of RNAi constructs inserted into VIE260B results in ectopic activation of tiptop and phenotypes independent of the RNAi target [56]. PCR screening showed that KK109179 contained insertions at both sites and likely caused the lethal phenotype observed in [23]. We removed the insertion at VIE260B following the crossing scheme outlined by [56] and kept two of the resulting lines with insertions only at VIE260B-2.

We generated a new *p24-2* RNAi line as previously described [41]. We designed the RNAi construct CG33105_RNAi using the E-RNAi server [61]. CG33105_RNAi was the only possible construct >50 bp in length with 100% of the possible 19-mers uniquely matching *p24-2*. CG33105_RNAi was cloned into pKC26 and then injected into flies from strain 60100 at 250 ng/μl. We isolated transformants using Bloomington *Drosophila* Stock Center (BDSC) balancer stock 9325 to ensure that the RNAi construct was inserted only at VIE260B-2 using PCR54. NV-CG33105-2 and NV-CG33105-6 are derived from different transformants, but carry the same CG33105_RNAi construct. We drove RNAi expression using lines that constitutively expressed GAL4 under the control of the *Act5C* or *αTub84B* promoter (BDSC lines 4414 and 5138, respectively). Five males and five virgin driver females were allowed to cross for 9 days at 25 °C and a 12-hour light/12-hour dark cycle; they were then removed from the vials. F1 progeny flies were counted 19 days after crossing. The proportion of wild-type (RNAi-active) F1 flies was compared to the proportion of wild-type F1 flies from control crosses between 60,100 males and the driver strains. We confirmed presence of the *p24-2* duplicate in each of these lines using PCR and Sanger sequencing.

3.2.7 Expression analysis

Genome-wide gene expression differences between A3 and A4 larvae were analyzed as described previously [96]. Sequences of the genes from A3 larvae were obtained from an A3 genome assembly constructed with publicly available A3 Illumina paired-end reads. To compare the expression levels of *Cyp28d1*, *CG7742*, and *Ugt86Dh* gene copies, we aligned publicly available 100-bp RNA-seq reads [96] to A4 mRNA sequences using Bowtie2 [81] (with –score-min L,0,0 to ensure that only perfectly aligned unique (i.e., copy-specific) reads were kept for FPKM calculations). We adjusted transcript length by subtracting the length of regions to which no SNP-covering read aligned because only reads overlapping the SNPs could be included in FPKM calculations. For example, *Cyp28d1* gene copies are distinguishable by 15 SNPs. When regions that cannot be spanned by perfectly aligned unique reads are removed from the effective transcript length, 310 bp is subtracted from the total 1,509-bp transcript length, leaving an effective transcript length of

1,199 bp. Similarly, for *Ugt86Dh* and *CG7742*, transcript lengths of 1,065 bp and 755 bp were used to calculate FPKM values, respectively. No such adjustments were made for the single-copy genes not segregating for duplications. The total number of reads aligned to the genomes was calculated based on alignment of the single-end RNA-seq reads aligned to the A4 and A3 genomes using TopHat [137].

3.2.8 Detecting evidence of recent natural selection

To assess the likelihood of natural selection at loci containing newly discovered copy number variants, we scanned the genome for evidence of a recent selective sweep. In a selective sweep, a mutation conferring a substantial fitness advantage quickly increases in frequency within a population. This process produces characteristic signatures in the site frequency spectrum, the distribution of allele frequencies at polymorphic sites in the population [48, 125]. For a recent selective sweep, these signatures may be detectable as differences in the site frequency spectrum relative to the expectations under neutral evolution.

For n genotypes in a population sample, the site frequency spectrum can be represented as a vector of the number of polymorphic sites with derived allele frequency i, $\vec{p_i} = [p_1, p_2, ..., p_{n-1}]$. In conditions of neutral evolution, the site frequency spectrum is typically biased towards alleles with lower frequency due to the prevalence of rare mutations. Under a selective sweep, alleles linked to the beneficial mutation increase in frequency concurrently with the beneficial mutation itself due to the genetic hitchhiking of alleles at loci linked to the allele under selection [125]. As a result, there is a reduction in the genetic variation around the locus containing the beneficial mutation. This results in a corresponding change in the shape of the local site frequency spectrum towards an overabundance of high-frequency derived alleles [48].

In the event of a selective sweep, an allele in a nearby region to the mutation under selection can escape the hitchhiking effect by recombining onto a selected haplotype. The probability of a lineage escaping the sweep can be approximated by

$$P_e = 1 - e^{-\alpha d}, \tag{3.1}$$

with the strength of the sweep $\alpha = r \ln(2N)/s$, where N is the effective population size, s is the selection coefficient, r is the recombination rate, and d is the genomic distance between the polymorphic site and the site of the selective sweep [43, 105]. The composite likelihood under assumption of a selective sweep is calculated as the product of the probabilities of the observed derived allele frequency at each polymorphic site across all recombination scenarios, maximizing for α [105]. The likelihood ratio can then be calculated using the composite likelihood of a selective sweep (alternative hypothesis) and the composite likelihood of neutral evolution (null hypothesis) based on the derived allele frequency at each site and the overall site frequency spectrum.

We polarized the site frequency spectrum using a multiple sequence alignment of an outgroup, consisting of reference genomes for other species in the *D. melanogaster* species subgroup, to obtain the unfolded site frequency spectrum. This is an important step in accurately estimating the composite likelihoods, as using the folded site frequency spectrum in such an analysis would result in the minor allele frequency at each polymorphic site being treated as the derived allele. The multiple alignment included *D. simulans*, *D. yakuba*, and *D. erecta*, representing all the species complexes within the *D. melanogaster* subgroup. Invariant sites that differed from the inferred ancestral state were included in the analysis, improving power and robustness to demographic factors such as bottlenecks [64].

To test the significance of the results, we compared the values of the composite likelihood ratio across the segregating sites to those calculated on the result of 100 coalescent simulations of neutral evolution. To account for ascertainment bias, we conditioned our statistical inference on neutral simulations with parameters matching the number of segregating sites observed in the data as well as measured recombination frequencies in *Drosophila* at the relevant locations in the genome to obtain an unbiased confidence interval. Estimates of the effective population size, neutral mutation rate, and recombination rate were taken from previous publications [50, 71]. 95% confidence intervals were computed using the largest composite likelihood ratio values from each neutral simulation to minimize the possibility of Type I errors. We used SweepFinder2 to compute the composite likelihood ratio over a 250 bp grid and the ms coalescent simulator to compute the neutral simulations [40, 66].

3.2.9 Estimating frequencies of duplicate alleles

The frequencies of duplicate alleles across the global *Drosophila* population panel were estimated using next-generation sequencing data. The Illumina reads used to determine the presence of a *Cyp6a17* deletion or *Cyp28d1* and *Ugt86Dh* duplication were obtained from Africa (Cameroon, Ethiopia, Kenya, Rwanda, Gabon, Guinea, South Africa, Zambia, and Zimbabwe [74, 79]), Europe (France and the Netherlands [11, 57]), and North America (North Carolina, Georgia, and New York [11, 57, 93]). Reads were mapped against the release 6 ISO1 reference genome using bwa-mem to produce SAM alignments [87, 88].

Duplications were genotyped at these loci by decomposing the alignment signatures of paired-end reads to extract the set of divergently mapped orphan reads and analyzing these signatures to call duplication variants. Reads in the SAM alignment that were assigned a discordant orientation were first extracted to obtain the pairs of reads that were not properly aligned with respect to each other onto ISO1 (e.g., the forward read maps to a location on the reference upstream of the reverse read due to the presence of a duplication with an intervening spacer sequence). We took the intersection of this set of reads with the complement of all read pairs in the alignment which mapped to the same chromosome on opposite strands with a concordant orientation (well-mapped read pairs in an expected configuration with respect to the reference). We then plotted the coverage of these reads versus reference genome coordinate to produce a genome track of coverage density for each strand of the paired-end reads.

Duplication variants were called for strains that showed distinct peaks and high signal-to-noise ratio in their coverage density for divergent orphan read pairs at breakpoints proximal to genes that were found to be duplicated in A4. Strains which had low genomic coverage (less than 10 Mb over the chromosome containing the potential duplication) or which were inferred to be identical by descent to other strains over the region containing the duplication, using data from the *Drosophila* Genome Nexus [79] for estimates of homozygous coverage and identity by descent, were excluded from the variant calling analysis. Populations were excluded from the analysis if they contained fewer than ten samples.

3.3 RESULTS

3.3.1 A4 strain assembly

The A4 strain is a part of the *Drosophila* Synthetic Population Resource (DSPR) [74], a resource for mapping phenotypically relevant variants. We assembled the new A4 genome using high-coverage (147x) long reads through single-molecule real-time sequencing of DNA extracted from females, following an approach that has been shown to yield complete and contiguous assemblies [20]. The A4 assembly is more contiguous than release 6 of the ISO1 strain [62], which is arguably the best metazoan whole-genome sequence assembly, with 50% of the genome contained in contiguous sequences (contigs) 22.3 Mb in length or longer. As compared to the ISO1 assembly, the A4 assembly comprises far fewer sequences (161 scaffolds versus 1,857 non-Y-chromosome scaffolds [42]) while maintaining comparable completeness [122]. The two genomes are collinear across all major chromosome arms, making large-scale misassembly unlikely (Fig. 3.1a). An optical map of the A4 genome also supported the accuracy of the assembly.

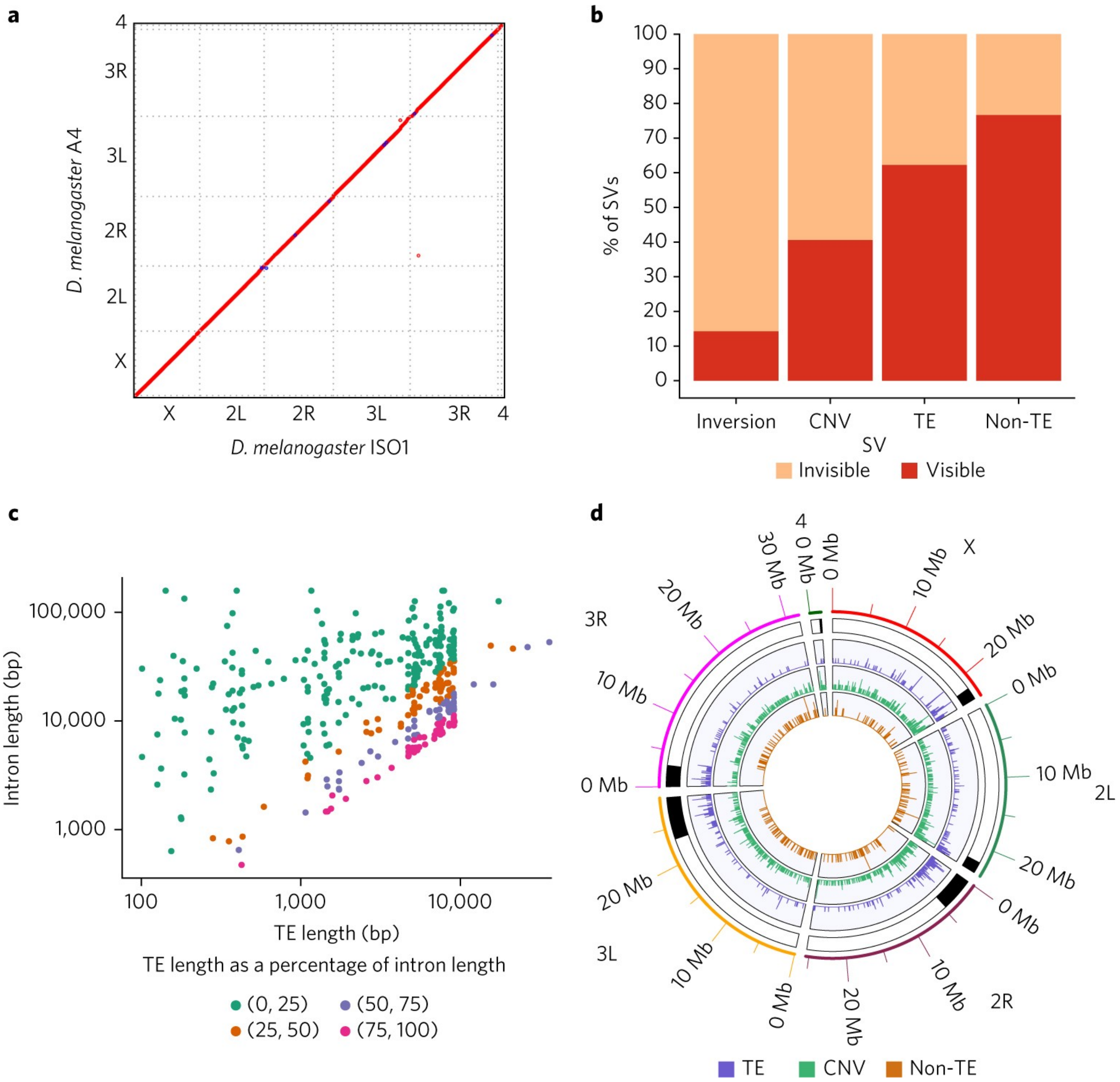


Figure 3.1: (A) Dot plot between the *D. melanogaster* reference (ISO1) and A4 assemblies. The A4 assembly is as contiguous as the ISO1 assembly (scaffold N50 = 25.4 Mb versus 25.2 Mb). Repeats and TEs were masked to highlight the correspondence of the two genomes. (B) The proportions of large (>100 bp) SVs in the A4 chromosome 2L assembly relative to the ISO1 2L assembly that were identified (visible) or missed (invisible) by short-read methods. (C) Relationship between the lengths of TEs in ISO1 (median 5.1 kb) and the lengths of the introns into which they are inserted. Nearly equal intron and TE lengths indicate that many introns comprise mainly TEs. (D) Distribution of SVs (>100 bp) across chromosome arms in the A4 genome. Track 1 shows pericentric heterochromatin (black). Tracks 2–4 show TEs, duplicate CNVs (relative to ISO1), and non-TE indels >100 bp in length, respectively. CNVs and TEs are present in higher densities in heterochromatin as compared to euchromatin, whereas non-TE indels are less numerous in heterochromatin.

Mutation type (>100 bp)	**Number of mutations in A4 euchromatin**
Insertion (TE)	768
Deletion (TE)	718
Insertion (non-TE)	223
Deletion (non-TE)	181
Copy-number variant	390
Inversion	27

Table 3.1: Number of different types of structural variants uncovered by aligning the A4 and ISO1 genomes.

3.3.2 Structural variants identified relative to ISO1

We identified putative SVs by classifying regions of disagreement in a genome-wide pairwise alignment of the A4 and ISO1 assemblies as indels, copy number variants (CNVs), or inversions (Table 3.1). Reads spanning SVs showed that genotyping error was rare (<2.5%). However, because extremely long repeats are common in heterochromatin and require specialized approaches for assembly and validation [72], we focused on euchromatin. We discovered 1,890 large (>100-bp) indels, which affected more than 7 Mb. In contrast, mutations <100 bp in length affected only 1.4 Mb (indels, 722 kb; SNPs, 687 kb). Among large indels, 79% (1,486/1,890) were transposable element (TE) insertions. A previously published catalog of TE insertions in A4 based on 70x short-read coverage [32] failed to find 38% of the TE insertions in A4 reported here (Fig. 3.1b). These insertions, which are invisible to short-read approaches, often occur (in 34% of instances) when a TE is inserted near another TE, resulting in complex, non-uniquely mapping reads that are difficult to interpret. One such insertion was found in the A4 allele of the *MRP* gene (encoding multidrug-resistance-like protein 1), which is a candidate gene for resistance to the chemotherapy drug carboplatin [73].

We found that many TE insertions affected introns (395/718 in ISO1, 435/768 in A4), often greatly lengthening them (Fig. 3.1c). Additionally, TEs inserted into exons can be spliced out, effectively becoming new introns. We saw evidence of this in cDNA from ISO1 [126] and in RNA-seq reads in A4 that showed exon junctions flanking TE insertions, which represents a genome-wide view of

TE-derived introns segregating in a population. TE insertions within introns are associated with decreased transcription [33], possibly caused by a phenomenon called intron delay, which slows transcription in long introns [132]. TE insertions can affect phenotype directly [90], perhaps by modulating or disrupting the expression of important genes. Because most TEs are rare in *D. melanogaster* [113], they are poorly tagged by common variants, complicating genome-wide association study (GWAS) approaches for mapping traits; this mirrors similar complications in human GWAS [89].

Non-TE insertions represented 20% of ISO1 and 23% of A4 insertions, and they accounted for 170 kb of sequence variation (Fig. 3.1d and Table 3.1). Although these mutations were much smaller than TEs (median 213 bp versus 4.7 kb), they often affected genes, and 23% even escaped detection by short reads (Fig. 3.1b). For example, among both hidden and visible deletions, there were 18 genes that were present in ISO1 and partially or completely absent in A4, including *Cyp6a17* (Fig. 3.2a). Knockout of *Cyp6a17* in a previous study increased cold preference [69]. Indeed, A4 flies preferred colder temperatures than flies from a strain carrying an intact copy of *Cyp6a17* (Fig. 3.2b). Furthermore, this mutation was more common than expected for a deleterious allele (Fig. 3.2c), suggesting that it has a role in regulating how flies respond to temperature in the wild. One deletion missed by short-read genotyping removed the second exon of *Mur18B* (and 41 amino acids of the encoded chitin-binding protein that confers resistance to high-temperature stress [94]), likely rendering the A4 *Mur18B* allele defective.

We discovered 27 inversions, ranging from 100 bp to 21 kb in length, that affected 60 kb of sequence, only 4 of which were detected by paired-end methods (Fig. 3.1b). These inversions often (in 21/27 instances) affected regions harboring genes, including a 21-kb region that spanned five genes encoding gustatory receptors: *Gr22a*, *Gr22b*, *Gr22c*, *Gr22d*, and *Gr22e*. Although such clusters of related sequences may obscure the read-mapping information used to detect inversions, we could not find genomic features that might explain why the other inversions were missed. The A4 optical map identified a putative inversion occupying 300 kb of the proximal end

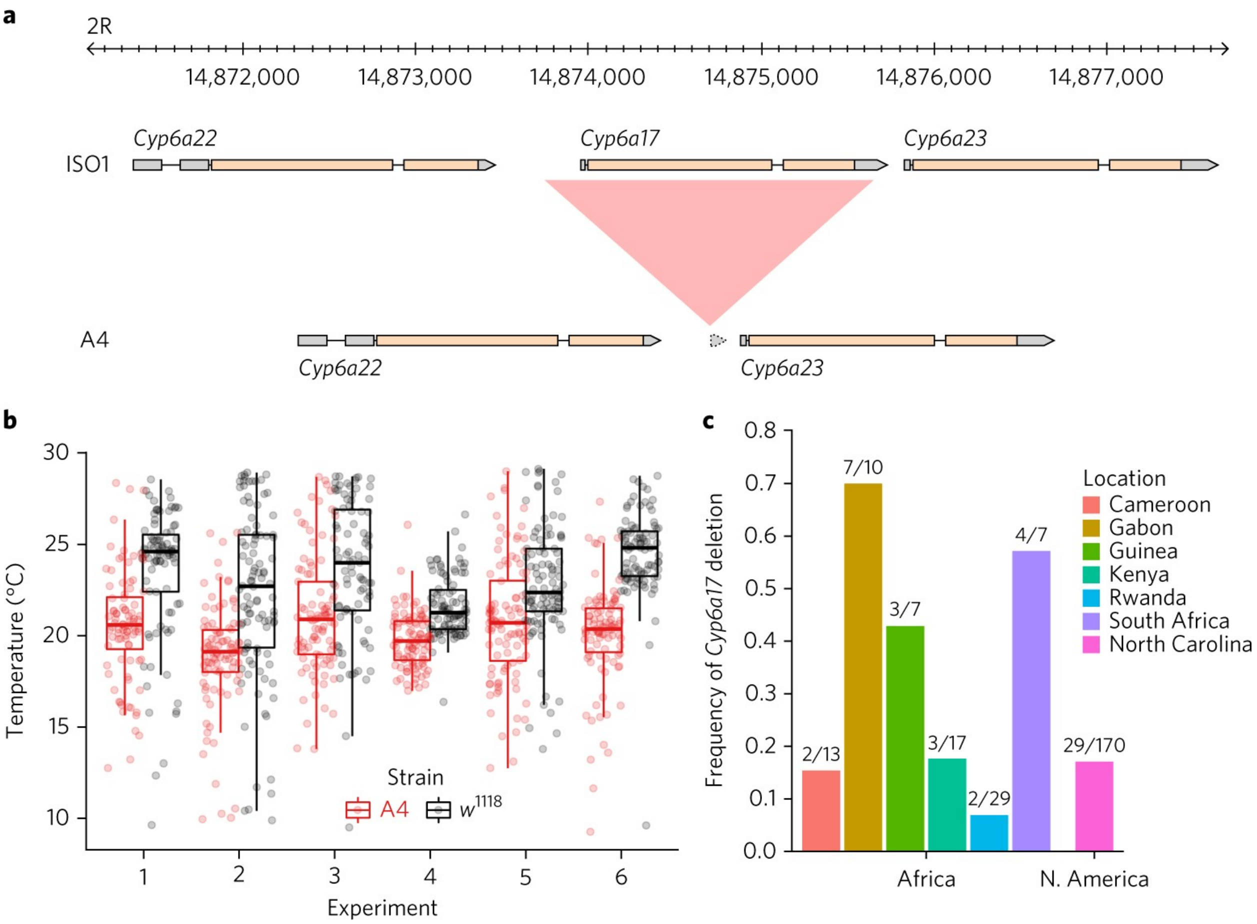


Figure 3.2: (a) *Cyp6a17* is deleted in the A4 genome relative to the ISO1 genome. Alignment between annotated ISO1 and A4 assemblies on chromosome arm 2R shows a large ISO1 region (red) missing in A4. Gene models are shown (gray indicates noncoding sequences, and yellow indicates coding sequences). (b) Temperature preference of strains A4 (Δ*Cyp6a17*) and w^{1118} (intact *Cyp6a17*). Preference was measured by recording the position of 100 flies along a linear 8-30 °C temperature gradient after an adjustment period. Each dot represents the position of a fly along the gradient. Each experiment number is an independent pairwise trial. A4 flies occupy colder regions of the gradient than w^{1118} flies (Fisher's method on Wilcoxson rank-sum tests, meta P value « 10^{-16}). Upper and lower hinges of the box plots represent 25% and 75% quantiles, respectively; the upper whisker indicates the largest observation less than or equal to the upper hinge + 1.5 times the interquartile range (IQR); the lower whisker indicates the smallest observation greater than or equal to the lower hinge – 1.5 times the IQR; and the middle horizontal bar indicates the median, 50% quantile. (c) Frequency of the *Cyp6a17* deletion in African (DPGP2) and North American (DGRP) populations.

of the X-chromosome scaffold that was not resolved by the A4 assembly. Failure to resolve this inversion is not unexpected because assembly methods tuned for euchromatin perform poorly in heterochromatic regions [72].

We discovered 390 CNVs (209 in A4 and 181 in ISO1) that affected ~600 kb (Fig. 3.1d). Although some CNVs were missed by paired-end methods owing to spacer sequences between copies that were longer than the library fragments (Fig. 3.3a,d), most (~90%) of the CNVs were missed because they occurred in complex tandem repeats. Unlike indels, most CNVs (64%) affected exons. Additionally, short-read CNV genotyping methods missed 13 of 34 protein-coding genes that were duplicated in A4. In total, only ~40% of CNVs were discoverable with high-specificity split-read and read-orientation methods (Fig. 3.1b). Consistent with previous observations [65], coverage-based methods were extremely nonspecific and were therefore excluded from analysis. We next compared published gene expression data from larvae of A4 to expression data for a DSPR strain called A3 [96] and identified 17 A4 duplicate genes that are single copy in ISO1 with increased expression, including genes previously identified as candidates for cold adaptation, olfactory response, and toxin resistance, among others (Fig. 3.3a,d). Notably, eight of these CNVs were invisible to short-read methods.

3.3.3 Functional elements in hidden structural variants

A longstanding concern in trait-mapping studies is failure to genotype candidate mutations [44]. Because A4 is a parental line of the DSPR trait-mapping panel, we could confront this problem directly. Among the eight duplicate genes with increased expression in A4 that escaped detection, *Cyp28d1* and *Ugt86Dh* fell under quantitative trait loci (QTLs) for resistance to nicotine, a plant defense toxin [54, 96]. One QTL (Q1) contains two genes, *Cyp28d1* and *Cyp28d2*, that encode cytochrome P450 enzymes, both of which were upregulated. The other candidate region that showed a major effect contains the *Ugt86D* gene cluster, which includes several differentially regulated genes, including *Ugt86Dh* (Fig. 3.3d,e). Candidate mutations like these are of obvious interest to researchers trying to dissect complex traits, and yet they were not visible in the initial study [96].

In the A4 assembly, Q1 contains a 3,755-bp tandem duplication in which the duplicated regions are separated by a 1.5-kb spacer, resulting in two copies of *Cyp28d1* (Fig. 3.3a). We compared paralog-specific expression levels of the *Cyp28d1* copies in A4 to expression of the single copy in A3. In the absence of nicotine, the proximal and distal copies in A4 exhibited ~41-fold and ~6.3-fold higher expression, respectively, than the single copy in A3 (Fig. 3.3b). The intervening spacer sequence proved to be the 5′ end of *Accord*, a long terminal repeat (LTR) retrotransposon (Fig. 3.3a). Insertion of *Accord* upstream of another gene called *Cyp6g1* has been linked to upregulation of the encoded cytochrome P450 enzyme [27], suggesting that the retrotransposon may be responsible for the upregulated expression rather than the tandem duplication of the *Cyp28d* gene. The second nicotine-resistance QTL contains several *Ugt* genes, including *Ugt86Dh*, which have previously been implicated in increased resistance to the pesticide DDT [110]. Of note, we found that *Ugt86Dh* was duplicated in A4 and several other strains (Fig. 3.3d,f); this mutation escaped detection by conventional paired-end short read methods. Although several *Ugt* genes in the Q4 QTL showed higher expression in nicotine-resistant A4 larvae than in sensitive A3 larvae [96] (Fig. 3.3e), candidate variants that explain these differences have yet to be identified.

3.3.4 Prevalence of hidden variants in global populations

Because nicotine analogs are widely used pesticides, we predict that resistance-conferring mutations are common, mirroring observations for DDT. Indeed, we found that four duplicate alleles spanning *Cyp28d1* and *Cyp28d2* segregated at intermediate to high frequencies in multiple populations (Figs. 3.3c and 3.4) in a 25-kb region where we expected duplicate heterozygosity to be less than 0.1. Similarly, the single duplicate allele of *Ugt86Dh* segregated at high or intermediate frequency in nearly all the populations we examined (Fig. 3.3f). Additionally, the site frequency

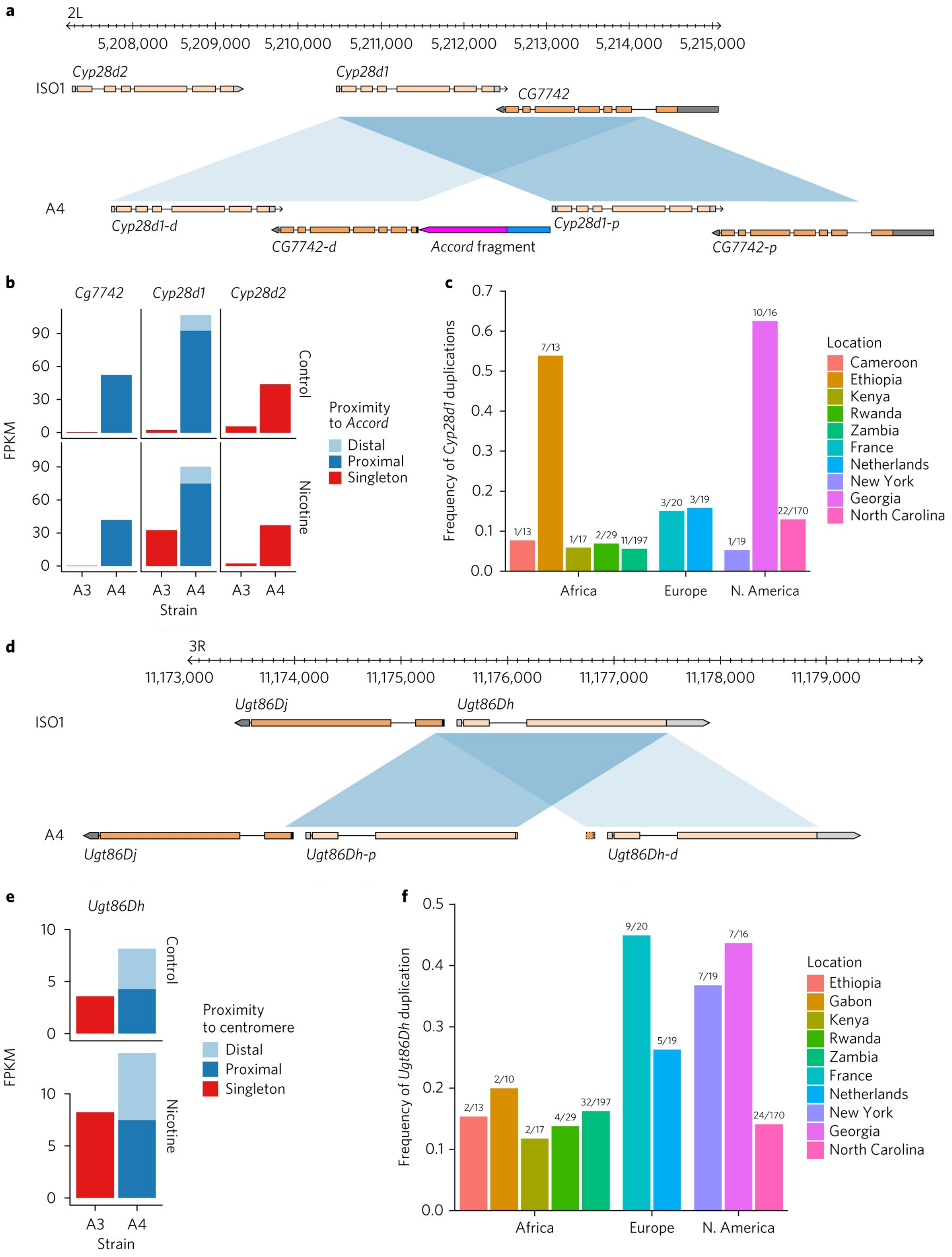


Figure 3.3: (a, d) Duplication of *Cyp28d1* and *CG7742* (with a 1.5-kb *Accord* fragment between proximal *Cyp28d1* and distal *CG7742*) and tandem duplication of *Ugt86Dh* in A4. (b, e) Paralog-specific expression (fragments per kb of transcript per million mapped reads) of candidate QTL genes in A4 and A3 with and without nicotine in the food. *CG7742* and *Cyp28d1* copies nearer to *Accord* are transcribed at higher levels, and the copies of *Ugt86Dh* are expressed at similar levels. (c, f) Combined frequency of *Cyp28d1* duplicate alleles in African, European, and North American populations, and frequency of *Ugt86Dh* duplicate in African and North American populations. Four duplication alleles were identified at the *Cyp28d* locus.

spectrum of SNP variation surrounding both *Cyp28d1* and *Ugt86Dh* are consistent with recent bouts of positive natural selection (Fig. 3.5), suggesting recent adaptation to nicotinoids.

Although we focus on genetic variation in A4 relative to ISO1, there is no biologically meaningful sense in which any individual of a species is a more appropriate reference than another. Yet, despite the prevalence of heritable phenotypic variation, functional work often describes results derived from individuals with diverse genotypes as applying to an entire species [100]. Approaches like RNA interference (RNAi) or gene editing with CRISPR require precise sequence information about their targets and can be easily misled by hidden structural variation. One study on the origin of new genes in *D. melanogaster* argues that new genes rapidly become essential, and the authors even report a new gene called *p24-2* that is so young that it is present in only *D. melanogaster* [23]. Experiments targeting *p24-2* using RNAi constructs suggested that, although new, *p24-2* is essential. However, *p24-2* was absent in eight of the ten strains we examined, including A4 and Oregon-R, which calls into question its essential nature in *D. melanogaster*. Because the original construct actually targeted both *p24-2* and its essential paralog *eca* [9, 118], we tested two other constructs targeting *p24-2*, neither of which resulted in any reduction in viability, thus bolstering the suggestion that *p24-2* is not essential.

3.4 DISCUSSION

The ubiquity of hidden variation in genome structure is merely an indication of the extent of the underlying genetic variation governing phenotypes. Together with careful phenotypic measurements, a new generation of high-quality genomes will identify previously invisible heritable phenotypic variation. Our results show that popular genotyping approaches miss a plethora of SVs, including ones that affect gene expression and organismal phenotype, suggesting that previous estimates of the contribution of SVs to regulatory and phenotypic variation are misleading [52, 129]. A substantial proportion of these newly discovered SVs, including large indels, trans-

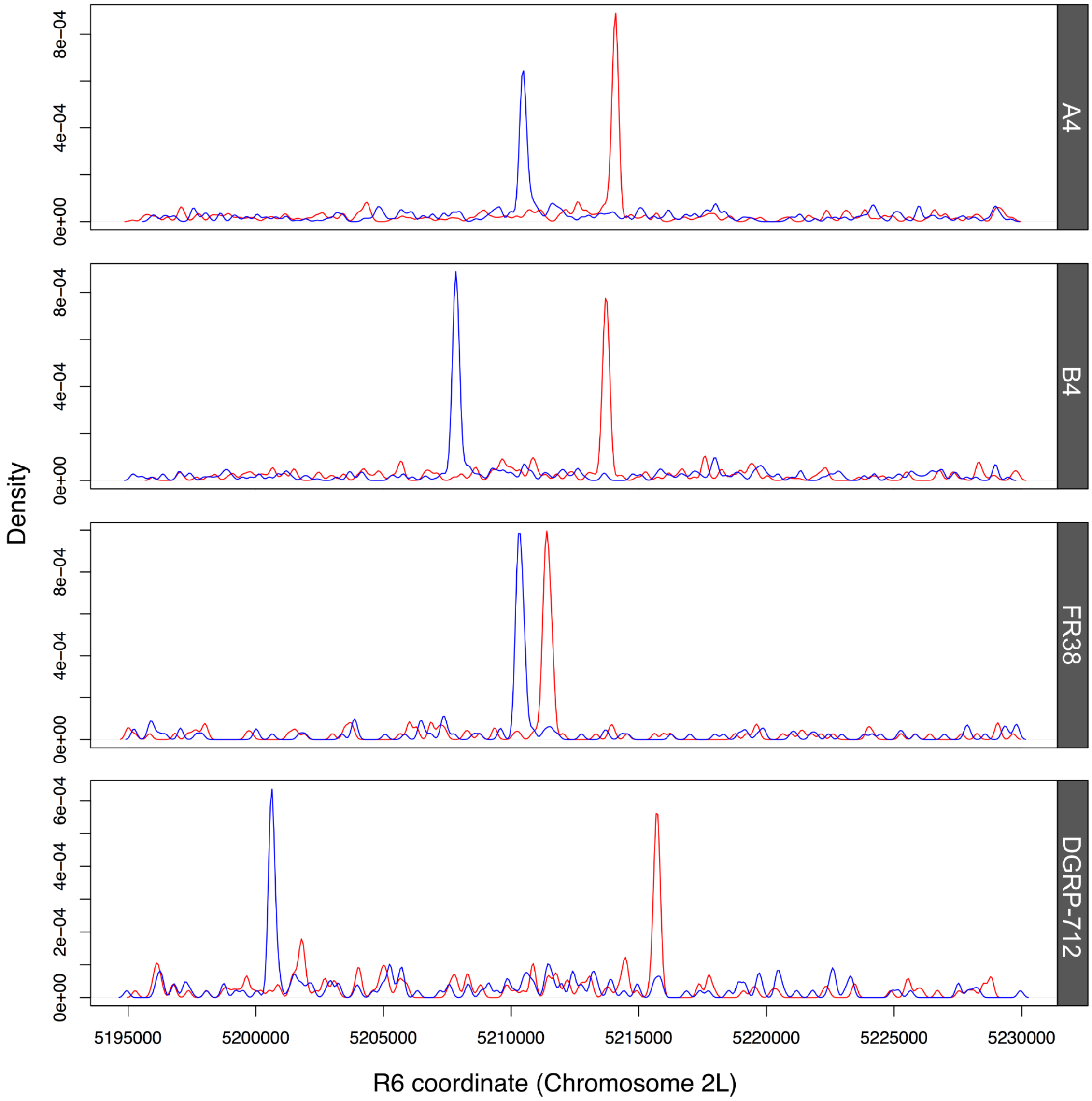


Figure 3.4: Duplications observed at the *Cyp28d* locus based on divergently mapped orphan read coverage density. The tracks show four separate alleles for duplications that were observed to be segregating in various populations around the world. The strains observed here, from top to bottom, are from Zimbabwe, California, France, and North Carolina. The red tracks correspond to reads aligning to the forward strand and the blue tracks correspond to reads aligning to the reverse strand.

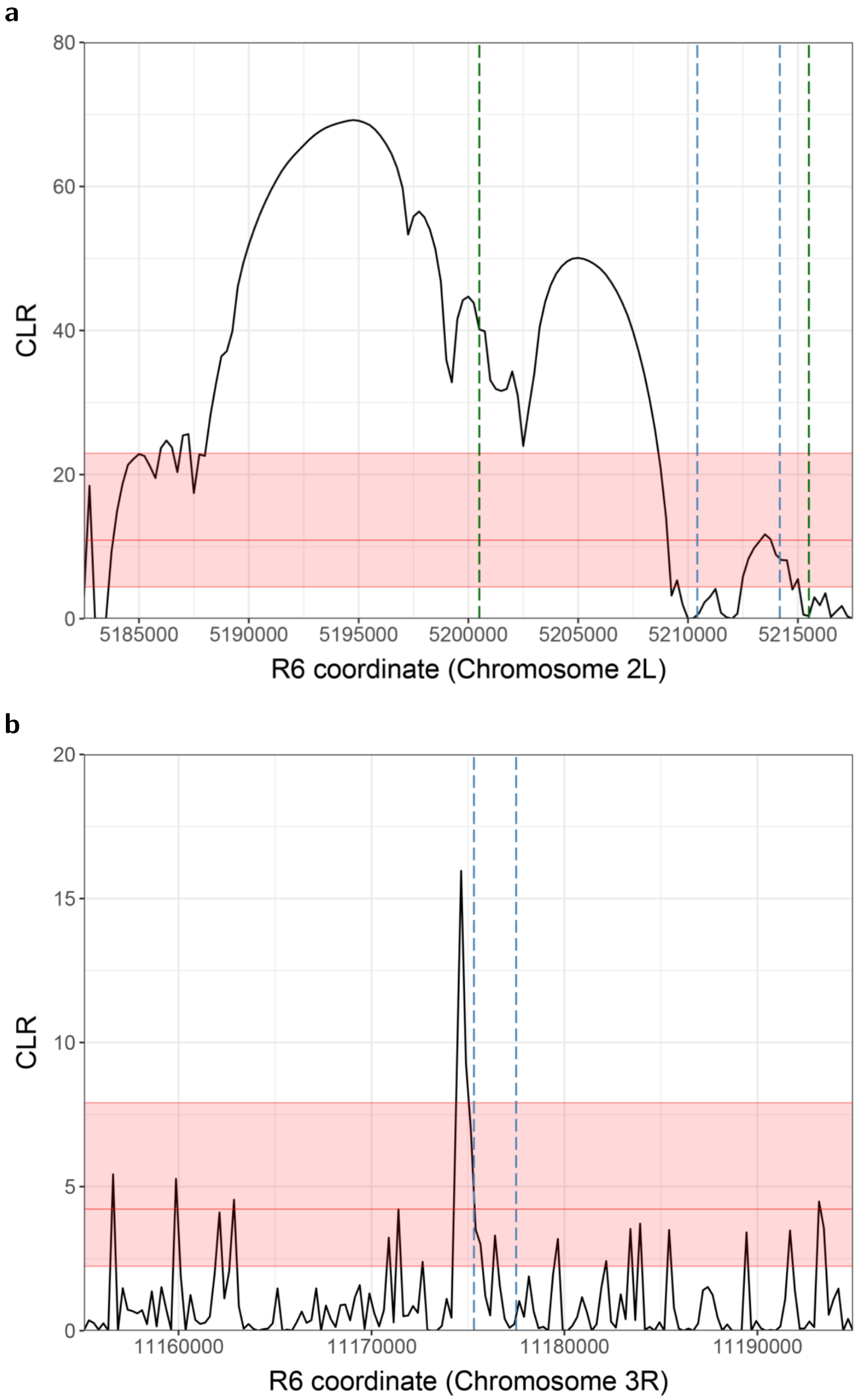


Figure 3.5: Distribution of the composite likelihood ratio statistic for the SNP site frequency spectrum in the genomic region containing (a) *Cyp28d1* in a French population and (b) *Ugt86Dh* in sub-Saharan African populations. The red shaded region represents the empirical 95% confidence interval for the maximum CLR values based on 100 neutral simulations, conditioned on the number of segregating sites and recombination rates in these regions. Green lines indicate the span of all duplication alleles observed in the global population panel, and blue lines indicate the span of the duplication discovered in A4.

posable elements, inversions, and copy number variants, are not detectable through conventional short-read sequencing methods, highlighting the value of high-quality, long-read sequencing in uncovering the full range of genetic variation.

These findings also highlight the potential pitfalls of relying on a single reference genome in functional genomics studies. The presence of hidden SVs can lead to confounded or misleading results, particularly in techniques which require precise sequence information to accurately reflect phenotypic outcomes. Additionally, we found that purportedly essential genes in *D. melanogaster* are actually absent in the A4 and other strains, cautioning against the generalization of findings from a reference strain to the entire species. At the same time, we observe numerous previously unidentified SVs in genes associated with adaptive phenotypes. Identified CNVs affect many exons and represent a substantial amount of hidden genetic variation, which may impact downstream protein coding and thus phenotypic variation through a number of mechanisms such as modulating gene expression levels or encoding multiple paralogs. Some of these variants, which had escaped detection due to having a single copy in ISO1, occur in genes related to traits such as resistance to toxins and temperature preference and display signatures of recent positive selection.

The A4 strain represents just one of the many possible reference genomes for *D. melanogaster*. The extensive hidden variation we observe segregating in *D. melanogaster* occurs in a species that likely harbors fewer complex structural features than humans or livestock, as well as crop species like wheat and maize. Consequently, we suggest that the true medical and agricultural impact of structural variation is likely to be much greater than the already considerable estimates made without recourse to multiple reference-grade assemblies [65]. This underscores the importance of comparing multiple high-quality reference genomes within a species in order to decipher the genetic architecture underlying phenotypic diversity and its evolutionary implications.

CHAPTER 4

Genotyping structural variation using machine learning

4.1 INTRODUCTION

Structural variants (SVs) are genomic variants that include duplications, insertions, deletions, transpositions, and inversion of DNA sequences. Balanced structural variations, such as transpositions and inversions, do not result in a net change in DNA copy number or dosage. In contrast, genomic imbalances can occur with duplications and deletions and are known as copy number variants (CNVs). Unlike small variants, such as single nucleotide polymorphisms (SNPs) and short insertion and deletion variants, SVs that arose from distinct mechanisms can span from less than 50 bp up to several megabases in size. It is estimated that genomic differences between humans due to SVs are many-fold higher than those arising from SNPs [24, 107, 130]. Large scale characterization of deeply-sequenced human genomes estimate that humans carry, on average, 2.9 rare SVs that affect 4.2 genes, and have revealed rare megabase-scale SVs in 2-4% of individuals [2, 29]. Furthermore, SVs contribute to deleterious rare alleles, and certain classes of SVs may tend to be more deleterious than others [21].

Some SVs can cause overt genetic disease. For example, Charcot Marie Tooth disease, the first autosomal dominant disease identified in humans, is caused by duplication that results in three copies of the *PMP22* gene and a gene dosage effect [91]. Meanwhile, having a single copy only of the gene results in phenotypic haploinsufficiency that causes hereditary neuropathy with liability to pressure palsies. Acquisition of either CNVs that affect proto-oncogenes, or translocations, inversions, and deletions that lead to fusion genes [82], later in life can lead to the development of cancers and affect response to therapy.

Many SVs do not directly cause disease but may contribute to disease susceptibility or normal phenotypic diversity within populations or between species by interfering with coding sequences, altering gene dosage, or affecting long-range control of gene expression [112, 143]. Numerous studies of human genomes suggest that SVs help to shape heritable differences in gene expression [24, 129, 130] and complex traits. In the genetic model organism, *Drosophila*, SVs have been shown to be overrepresented in candidate genes associated with quantitative trait loci (QTLs) [21]. From an evolutionary perspective, transpositions may play a role in divergence and postzygotic isolation. One manifestation is inviability as seen in interspecies crosses of Arabidopsis thaliana [13] or the development of hybrid sterility as seen in crosses between *Drosophila melanogaster* and *D. simulans* [97]. SVs have also been directly linked to phenotypically important natural variation in *Drosophila*, such as in the case of the *Cyp6g1* locus, a DDT-resistance gene under positive selection [45].

However, despite substantial interest in their evolutionary origins, mutational mechanisms, topological effects, and phenotypic consequences, these variants have been difficult to detect, identify and study compared to smaller variants such as SNPs. One of the main difficulties has a dearth of reference databases for SVs [29] that are comparable to those for SNPs. This problem has persisted despite the continuous growth and abundance of high-quality short-read genomic data. Although there have been many efforts in developing algorithms to detect SVs based on techniques such as split-read mapping and coverage analysis [116, 153], technical and computational challenges persist in the mapping of complex rearrangements or those that span repetitive elements. For example, SV identification from short paired-end sequencing data is limited by a limited recall, high false discovery rate, and missed SVs (Fig. 4.1). Additionally, existing short-read detection methods are strongly biased towards detection of small SVs (often less than 100 bp) and struggle to call larger or more complex variants. This underlying complexity has made identifying causative genetic variants for most traits a steep challenge for which the scientific community has had limited success through conventional SNP loci association [93, 99]. As a result, there is a large discrepancy between the known heritability of most traits and the fraction of that heritability that can be explained by known causative genetic variants [95].

Compared to short-read sequences that are typically less than 300 bp and mostly around 50-150 bp, long-read sequencing by PacBio or Oxford Nanopore generate significantly longer sequences with reading lengths ranging from 10 kb to 900 kb. This results in substantially larger contigs with a greater degree of overlap, enabling more reliable high-contiguity de novo assembly. Long-read mapping approaches can improve mapping by spanning problematic regions, especially those with stretches of repetitive or heterozygous elements, which can elude the heuristic inferences relied upon by short-read methods. At the current time, however, long-read sequencing technologies have a higher error rate, are more expensive, and require distinct expertise in library preparation and sequencing, resulting in a lower abundance of long-read genomic data, as well as numerous caveats in analysis and SV identification. Given the prevalence of high-accuracy short-read genomic data, a current-unmet need has been the development of algorithms to accurately predict SVs from short-read data, particularly for false negatives and hidden variation.

Recent advances in machine learning, specifically deep convolutional neural networks (CNNs), have resulted in generational leaps in a number of pattern recognition tasks. Through the convolution steps, these models continuously learn to recognize specific, predictive patterns, tuning them progressively throughout the training. As a result, provided a sufficiently large and representative training dataset, CNNs can make more accurate predictions on complex data than less sophisticated models, such as random forests or support logistic regression [10].

High-coverage long reads obtained by single-molecule real-time sequencing have been used to assemble reference-quality genomes of 14 strains of *D. melanogaster* and identify novel, functionally important SVs [21]. Thirteen of these strains are founder strains of the *Drosophila* Synthetic Population Resource (DSPR), and the other strain is Oregon-R, a common wild-type strain used in projects such as *Drosophila* modENCODE. DSPR represents an advanced intercross panel of roughly 1600 recombinant inbred strains designed to facilitate genetic analysis of complex traits by addressing the shortcomings of genome-wide association studies and QTL mapping panels derived from only two parents.

Here, we leverage the long-read sequencing dataset that we previously generated to develop and train a machine learning model with the goal of improving SV calling in short-read datasets. We show that our model is able to learn the important combinations of sequence and alignment metrics and achieve a mean accuracy of over 95%, outperforming other machine learning algorithms such as random forest and statistical models like logistic regression.

4.2 METHODS

4.2.1 Preparing a set of structural variant candidates

SV candidates were compiled based on reference-quality assemblies of 14 *Drosophila* melanogaster strains (13 founder strains from the *Drosophila* Synthetic Population Resource and the Oregon-R strain) using PacBio long-read sequencing [21, 74, 100]. Initial quality control was performed on each of the source data files for SV candidates in the fourteen strains to validate each entries/row in the raw output of SV calls produced by svmu [21]. Additionally, data was cleaned by removing

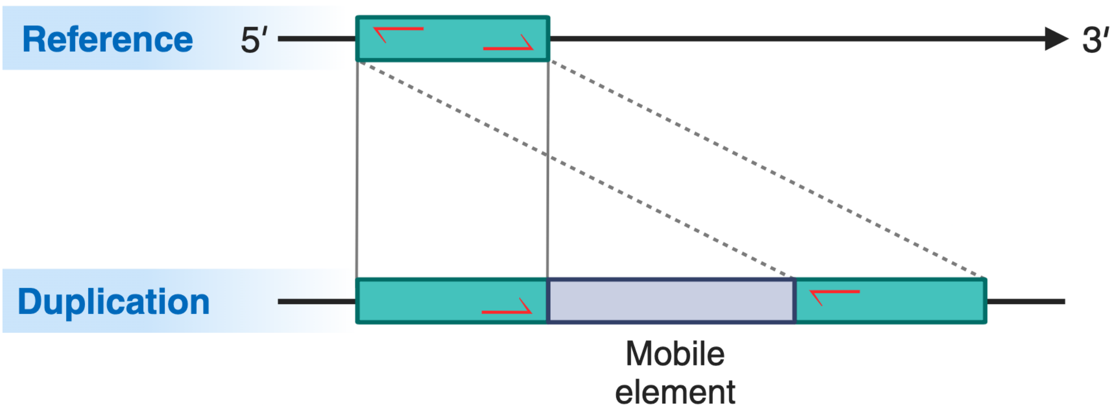


Figure 4.1: Short-read sequencing has inherent shortcomings in detecting structural variants, including balanced variants as well as duplications and copy number variations. This gene schematic illustrates a frequent duplication phenomenon, in which duplicated paralogs are separated by a spacer DNA element between them. When the size of the spacer element is larger than the insert size of the read pair, short-read-based techniques such as split-read mapping are insufficient to reliably detect such variants.

all errant or non-conforming entries that do not match or are missing data fields in the standard svmu output format. Biologically nonsensical entries were filtered out, such as those recording a negative, zero, or infinity-valued coverage for either the ISO1 reference or one of the fourteen assemblies.

Then, we selected the calls for all SV candidates that could be identified by an ISO1 coordinate range, such that the range is referenceable for any short-read alignment to the reference genome. The length of each SV candidate was computed based on the ISO1 start and end coordinate breakpoints. All rows that contained zero or negative-length calls were removed. For each of the fifteen assemblies, duplicated SV calls were filtered out, and the SV types for each call were collapsed into the four main types (deletion, insertion, CNV, or inversion); for instance, calls labeled as a 'CNV-INS' were folded into the larger CNV category. We filtered out redundant calls that were multiply labeled across different SV types. Finally, for each assembly, we sorted the calls by the reference-genome chromosome and then by reference-genome start coordinate. These final post-QC SV candidates were re-collated by merging overlapping and contiguous ranges for the same SV type within each chromosome for each assembly.

4.2.2 Generating features from short-read sequence data

Data for the neural network was produced by generating a collection of features from Illumina short-read sequence data using a custom pipeline. Raw short reads for each of the fourteen strains were aligned to release 6 of the ISO1 *Drosophila* melanogaster reference genome assembly [62] using BWA-MEM version 0.7.17 [87]. The aligned reads were processed with samtools version 1.11 to mark PCR duplicates [88].

Features were generated by extracting, decomposing, and transforming the data in the short-read alignment fields and optional tags generated by BWA-MEM. The features for each read were summed and collated into per-base values according to the coordinates that they mapped to on the ISO1 reference genome, accounting for alignment gaps and clipping at the ends of reads.

To normalize the scale of features across samples, collated values were scaled based on either coverage depth or aligned length (in the case of features such as per-read edit distance and alignment score). This data was generated for each of the candidate SV ranges identified from the fourteen long-read assemblies. The set of twenty different features is listed in Table 4.2.

4.2.3 Neural network architecture

We implemented a multi-layer deep convolutional neural network using Keras 2.2.4 as the frontend and Tensorflow 1.15.0 as the backend [25], implemented in Python 3.7.4, and trained on 4 Nvidia Tesla T4 GPUs over up to 50 epochs at a batch size of 1024.

Taking the featured input matrices across aligned sequence windows as described above, we feed this input layer into a series of convolutional layers to further encode our features before generating predictions. In the first set of convolutions in the network, we scan 128 filters of size 16 over four consecutive layers. These are followed by four layers of 256 filters of size 12 and four additional layers of 256 filters of size 8, where each subsequent layer operates on the output of the previous layer. After the first convolution in each group of four, we perform pooling by max values within a window to reduce the output size and thus the number of parameters in the model. After each convolution operation, we normalized weight parameters using batch normalization, before applying a rectified linear (ReLU) activation to the output.

As a result of these stacked convolutions, constituting the deep learning capacity of the network, our initial input of features across a window of 1000 bp is then mapped into a series of high-level feature patterns that better capture the information content relating to each type of SV. The output of the convolutional layers are then flattened and connected to a single hidden layer of 128 fully connected neurons and a final fully connected layer of 4 neurons (with a 50% dropout rate) activated by the softmax function, which outputs the probability of an input sequence being representative of a given SV class. These layers were regularized using an elastic net penalty of L1 = 0.01 and L2 = 0.1.

4.2.4 Partitioning datasets into training, validation, and test sets

To assess the performance of our neural network models, we used two different strategies to partition the entire set of features from all 15 assemblies. For the "leave-one-out" approach, we selectively excluded one assembly and combined the other 14 before applying the processing and normalization steps described above. Here, we randomly select 20% of the combined data from 14 assemblies as the validation set and the remainder as the training set for fitting the models. The left out assembly is used as the test set, for each of the 15 assemblies. For the assembly-blind k-fold approach, we took the entire set of combined feature data from all 15 assemblies, and randomly selected 10% for testing, 10% for validation, and used the remaining 80% for training.

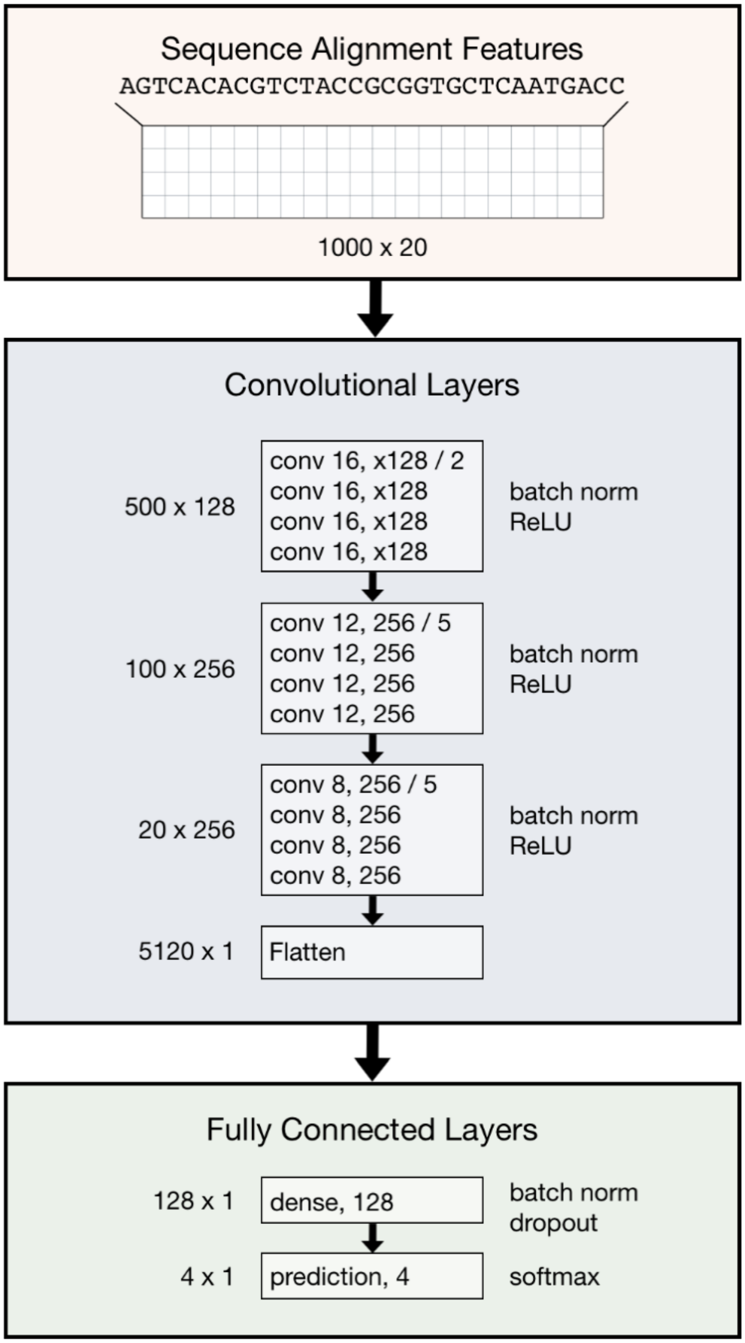


Figure 4.2: Schematic flowchart of the deep CNN architecture for SV classification. First, the sequence alignments are divided into 1000 bp windows and encoded at each position as a combination of 20 features. Subsequent convolutional layers scan across these positions, followed by batch normalization, rectified linear unit (ReLU) nonlinear transformation, and periodic pooling of adjacent positions by taking the maximum value. These final extracted features are fed into a hidden layer of fully connected neurons, which itself connects to a final layer of four neurons, each representing the probability of one of four SV classes.

4.2.5 Comparison of neural network to conventional classifiers

To evaluate the performance of our neural network model against other classification models, we used conventional linear classifiers and other statistical models to infer the presence of structural variants on the same short-read sequences as used for the neural network. We compared the various methods across several metrics, including overall accuracy, false-positive rate, false-negative rate, precision, recall, and receiver operating characteristic.

We trained a logistic regression, support vector machine, and random forest model using the scikit-learn machine learning package with the same featured data that was used to train the neural network. To reshape the 3D matrices into a format compatible with scikit-learn functions, the window (n=1000) and feature (n=20) axes were collapsed into a flat vector (n=20,000). The generalized linear models with L1 and L2 regularization (elastic net) were optimized using a grid search for alpha and L1 ratio respectively.

4.3 RESULTS

4.3.1 Characterization of SVs in reference long-read assemblies

Although SVs are increasingly recognized to be important in regulating gene expression to influence the diversity of complex traits in health and disease, they remain understudied as the identification and characterization of SVs have remained a technical and computational challenge. The sequenced DSPR founder strains are particularly well-suited for the present investigation. This resource represents a large panel of recombinant inbred lines derived from a small set of *D. melanogaster* founder strains. Previous studies using DSPR have produced a wealth of genetic data, including phenotypic observations as well as candidates from mapping of quantitative trait loci [74]. Additionally, Oregon-R from the modENCODE project is similarly useful as a long-read reference assembly, with an extensive dataset of RNA-seq and ChIP-seq data available [100].

Preliminary model testing and tuning on simulated and real short-read data showed that the general type and structure of the model architecture can be trained to detect variable-sized structural variation based on windows of short-read information at the base-pair level. Various models trained and validated on this data performed very well but had suboptimal performance on noisier sequences from biological samples. Using the features generated from real short reads based on A4-ISO1 long-read alignment showed possible issues with model generalization or overfitting. To tackle these challenges, we characterized the SVs present in each of the fourteen long-read assemblies (Table 4.1). This information motivates an updated framework to generate a large amount of genomes with much more biologically representative artificial SVs that can be used to tune a generalizable neural network architecture. Their statistical parameters will inform how we simulate data for model optimization.

For each strain, we analyzed the distribution of SV lengths and distances across the five major chromosome arms (2L, 2R, 3L, 3R, and X) in addition to chromosome 4. For all four classes of SVs under study, the shape of their respective size and distance distributions approximate an inverse gamma distribution. This is reflected in a peak of values flanked by a long right tail and a very small density of values close to zero. The general shape of the distribution is consistent with prior analyses into the statistical properties of structural variants in various species [159].

4.3.2 Feature engineering of short-read alignment data

While the mechanisms of gene regulation are well-described by epigenomic and transcriptomic machinery that recognizes sequence-specific binding sites, the principles governing structural variation are far more nebulous. As structural variants encompass a wide range of disparate genomic events, and can vary from as little as 50 bp to several megabases, attempting to learn the

Strain	Deletions	Insertions	CNVs	Inversions
A1	2469	2282	333	179
A2	2926	2541	352	215
A3	2559	2187	270	178
A4	2597	2240	376	210
A5	2729	2301	304	174
A6	2602	2489	266	134
A7	3157	2545	483	188
AB8	2428	2186	330	229
B1	1911	1691	260	188
B2	2654	2610	349	168
B3	2656	2342	355	153
B4	2728	2440	392	133
B6	1628	1412	203	195
Oregon-R	2660	2229	343	205

Table 4.1: Number of uniquely identifiable SVs with respect to ISO1 in the DSPR founder and Oregon-R strains.

biological rules behind SVs from sequence alone is likely to be challenging. Furthermore, though these sequence-only models have proven very successful for predictions of samples similar to their training data, they do not generalize well to datasets representing foreign cellular contexts [102].

Thus, we incorporated several features that are technical in nature derived from various metrics in the alignment files that should be agnostic to biological backgrounds after normalization. In order to build quantitative predictive models using these short read alignment data, the genomic sequences and associated alignment metrics were featurized by enumerating the scores of 20 characteristic features at each genomic position (Table 4.2). These features were chosen based on their relevance in patterns of mapping paired-end reads to a reference genome when one or more read sequences span a structural variant breakpoint. For example, in previous analyses of short-read signals of SV presence given a long-read prior, we successfully leveraged the coverage density of discordant-orientation orphan reads to detect duplication breakpoints at loci that were initially identified through de novo assembly of the A4 DSPR founder strain from PacBio long reads. We were able to genotype multiple duplication alleles of Cytochrome P450 gene, *Cyp28d1*, relative to the ISO1 reference using short-read sequences in a global panel of *D. melanogaster*. These alleles, including the allele variant identified in the A4 strain, were previously hidden variants invisible to existing short-read detection methods [22].

Naturally, the various types of data contained within a sequence alignment differ significantly in their numerical magnitude and dynamic range, as well as the shape of their distribution. This is prevalent for certain signals, such as the rate of well-mapped read pairs, which are heavily skewed, hard bounded within a range of values (such as 0 to 1), or have discontinuous density functions. Other features, such as template length, contain outlier values which can span orders of magnitude above the mean value. However, as a reference alignment signal for inferring the presence of SVs, such outliers are useful and therefore must be accounted for rather than simply removing them from the dataset. Our model combines these data in combination with other signals potentially contained in short-read data, including staple heuristics in existing short-read SV detection methods. For instance, signatures of split read mapping and coverage are the basis of many existing short-read structural variant detection methods, such as Pindel and pecnv [116, 153]. Our method incorporates and extends upon this information by including whole-genome distributional data on secondary and supplementary alignments, while not making a priori assumptions about the exact signatures being searched for. Instead, we allow the model to construct novel and potentially unintuitive nonlinear associations between combinations of short-read features and SV classifications.

Another consideration when selecting our features and processing these data were their suitability to predicting our structural variants of interest. Indels and copy number variants can prove difficult to map accurately due to being composed of repetitive sequences and often spanning a region of the genome much larger than a paired-end insert. The problem is exacerbated in analyses of short-read alignments, as the cumulative signal of SV signatures is a small fraction of the total alignment information reported by a read mapper.

This returns a very high total mapping rate from the FASTQ, because it will assign lower quality secondary alignments, attempt poorly mapped mate rescue, and assign mates in a pair to different chromosomes or with very large template lengths if the mapping algorithm finds a better score than for a close-by-the-mate low-quality mapping. This is why signals like read clipping densities are high-importance features, as well as the average alignment score, etc. A richer and stronger signal for SV inference can be extracted by considering combinations of various paired-end read mapping rates in relationship to one another, e.g., the density of unpaired mates and orphan reads and correlated relationships of PE mapping signatures extracted from bit flags.

Thus, in order to achieve proper convergence of model parameters, an appropriate transformation must be applied to the features which adequately accounts for these differences in scale to avoid instability in gradient updates during model training. We found that the default methods of regularizing feature scale, namely standardization and min-max normalization perform suboptimally on the short-read alignment features, due to their susceptibility to heavy influence by outliers. Superior performance was achieved using quantile normalization targeting a normal distribution, which is robust to outliers and equalizes the scale of different features while maintaining relative rank within features.

When training a deep-learning model for SV applications, it is important to incorporate spatial information from genomic data into the features that are the model input. This is a challenge with short reads due to the limited amount of genomic information contained within each read pair. To overcome this problem, we used a sliding window approach in which the model inference at each position in the genome is centered in a 1000 bp window. Each input to the model was represented as a two-dimensional 1000x20 tensor, with the twenty columns corresponding to the feature value at each position in the genome. Spatial information around SV breakpoints was also considered by applying a flanking region to the first and last window at the edges of each SV range. This approach has the additional advantage of naturally representing the total genomic coverage of each type of SV, as the number of windows generated for each class is based on both the frequency of the SV and its size distribution, instead of just the count.

Type of information	Number of features	Details
Coverage	1	Based on CIGAR string "M" count
CIGAR string	5	Read density of match, insertion, deletion, soft clipping, and hard clipping labels
Paired-end alignment signatures	6	Read mapped in proper pair; Mate unmapped; Read reverse complemented; Mate reverse complemented; First in pair; Second in pair
Non-primary alignment	2	Secondary and supplementary or split read mappings (bit flag and XA/SA tags)
Mapping quality	1	MAPQ score in alignment
Chimeric and split reads	1	Mates map to different chromosomes (RNEXT field is not "=")
Template length	1	Inferred read pair insert size
Edit Distance	1	Based on NM tag
Alignment score	2	BWA alignment (AS) and suboptimal (XS) alignment scores

Table 4.2: The 20 features generated from short-read sequence data used to train the neural network model.

4.3.3 Constructing a CNN for structural variant detection

In order to learn the sequence biology as well as technical metrics associated with each class of SV, we implemented a deep convolutional neural network (CNN). CNNs have been successfully deployed in a wide variety of machine learning applications, from computer vision and natural language processing to analyzing sequence-based multi-omics datasets. One particularly powerful characteristic of CNNs is the ability to extract high-level representations of the biological sequences over the course of many convolutions, effectively converting a one-dimensional sequence into a series of complex patterns, albeit at the expense of biological interpretability compared to linear models. When applied to understanding regulatory biology, convolution operations on the initial sequences are analogous to scoring patterns, such as position weight matrices (PWMs) for transcription factor motif binding, and subsequent combinations of them thereof [159].

By employing deep neural networks, we are able to integrate both of types of features and recognize unique combinatorial patterns predictive of SVs. In the initial convolution step, our model learns the weights of 128 unique patterns scoring the 20 features we chose over a sliding window of 16 bp, which are then pooled by taking the maximum value of adjacent pairs of positions. This convolution is repeated without the pooling three times to further convolve our features into abstract but higher-level representations of the data. We repeat this process of four convolutions two more times, but with 256 kernel filters of sizes 12 and 8 respectively and with an initial pooling distance of size 5, which is then flattened into a 1D vector. Prior to activation with a rectified linear unit (ReLU) transformation, each layer is normalized with batch normalization. The flattened output from the convolutional modules are taken as the input to a single hidden layer of fully connected neurons, which are also subjected to batch normalization, elastic net regularization, and neuron dropout. Finally, this layer is connected to the prediction layer consisting of four neurons, each representing the probability of one of the four SV classes.

Our architecture design was guided by observations in recent literature that deeper networks with more convolutional layers can perform just as well, if not better than, models featuring many dense layers of fully connected neurons. This type of architecture not only has the advantage of more sophisticated feature extraction but also significantly reduces the number of parameters to be trained, with benefits to both overfitting concerns and computational runtime. We optimized our parameters during training by comparing prediction accuracy and minimizing the loss function (categorical cross-entropy), and tuning the model accordingly to improve predictions (Fig. 4.3).

4.3.4 Identifying different classes of SVs in *Drosophila* short-read data

To assess the predictive performance of our models, we performed 8-fold cross validation where we randomly held out 5% of the overall combined dataset across all 15 assemblies as the test set. We reserved another 5% as the validation set for training and early stopping, and the remaining 90% was used for model training. As neural networks are inherently stochastic in nature, our models were evaluated over 10 independent iterations, achieving an average accuracy of over 95% in predicting the correct SV class on the test set (Fig. 4.4a). Additionally, we evaluated the area under the receiver operating characteristic (auROC) curve to assess the trade-off between model sensitivity and specificity by plotting the true positive rate against the false positive rate (Fig. 4.4b). Our models achieve an impressive mean overall auROC of over 95% for deletions, insertions, copy number variations, and inversions respectively (in comparison, with no model, random estimates would yield on average an auROC of 0.25 given four classes). Notably, the models were consistently better at predicting deletions and insertions compared to copy number variation and inversions, which was the least accurate of the SV classes. This can be attributed in part to the lengths, as the former two are much longer and thus covered by numerous windows compared to the latter. Given this initially unbalanced nature of the different SV classes, we also calculated the area under the precision-recall curve (auPRC). Here, our models achieve a mean overall auPRC of over 95% for deletions, insertions, copy number variations, and inversions respectively, indicating

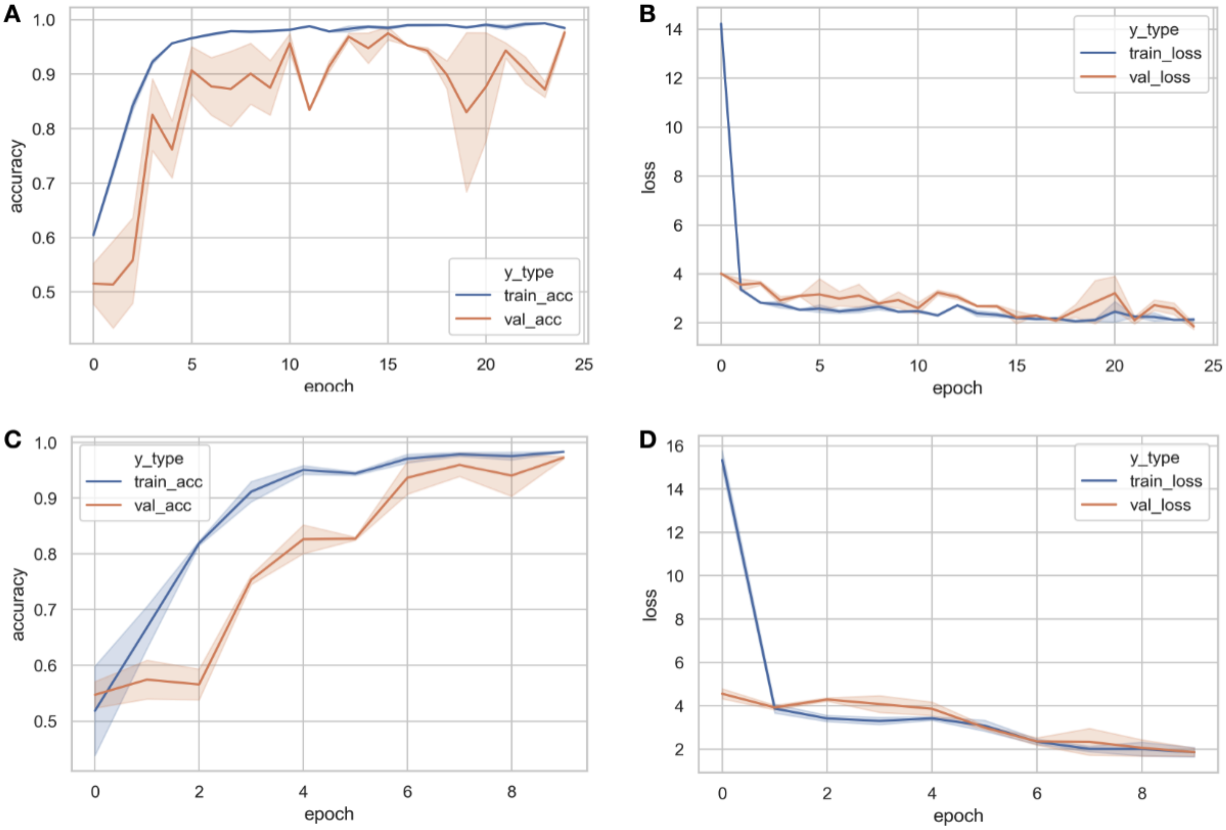


Figure 4.3: (A) Line plots showing the mean trajectory of training and validation accuracy over training batches for the 8-fold cross-validated models (shaded area represents IQR range) (n=10). (B) Line plots showing mean trajectory of training and validation loss (categorical cross-entropy) for the 8-fold cross-validated models. (C) Line plots showing the mean trajectory of training and validation accuracy for the leave-one-assembly-out models (n=15). (D) Line plots showing mean trajectory of training and validation loss (categorical cross-entropy) for the leave-one-assembly-out models (n=15).

that the balancing step in during preprocessing and our network architecture are able to account for this class imbalance.

To further evaluate our models, we performed cross-validation by leaving one assembly out as the test set, before splitting the combined data from the other 14 into training (80%) and validation (20%) sets. This is a much more challenging scenario, as we are leaving out an entire strain, so any strain-specific effects are not seen during training. Strong performance using this approach is highly desirable and informative because it indicates our models are generalizable to truly orthogonal datasets unseen by the models in the training data. Across all fifteen left-out-assemblies, the models achieved a mean accuracy of 37% (Fig. 4.4c). In terms of mean auROC, the values were 76% overall and 52%, 68%, 64%, and 66% for deletions, insertions, copy number variations, and inversions respectively. As with k-fold cross-validation, the model appears to have the most difficulty recognizing inversions but also struggles with insertions compared to deletions. Similarly, auPRC values were 39% overall and 61%, 31%, 42%, and 34% for each SV class respectively. As compared to the equivalent metrics for the 8-fold cross-validation, the models are less performant, which is expected and can be attributed to the more challenging problem of predicting on unseen data.

We achieve high performance with our models in overall prediction metrics as well as for each of the individual classes. Given the diversity of information in the fifteen training assemblies and the ability of the architecture to generalize to unseen datasets, these models can provide tremendous value in identifying putative structural variants in short-read datasets lacking an accompanying long-read dataset.

4.3.5 CNN detects variation hidden from other models

We benchmarked our models against a number of existing approaches to determine if our deep learning strategy could discover structural variation hidden from these current algorithms and

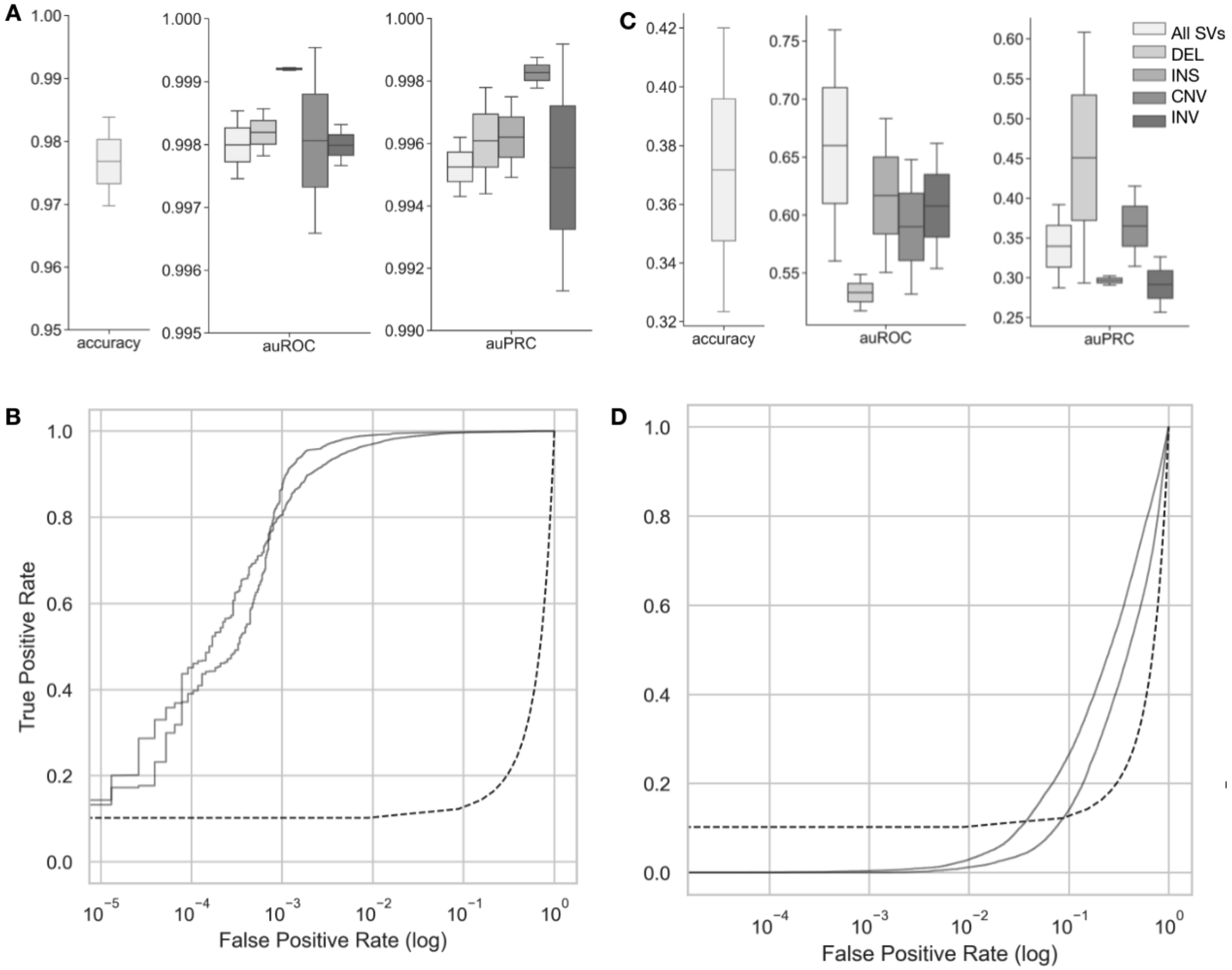


Figure 4.4: (A) Box plots showing distribution of accuracy and weighted accuracy of 8-fold cross-validated models, where a random 5% of the dataset is held out as the test set, another 5% is used for validation, and the remaining 90% is used for training (n=10). (B) Plots of the receiver operating characteristic (ROC) curves for overall predictions for 10-fold cross-validated models (n=10). (C) Box plots showing distribution of accuracy and weighted accuracy for leave-one-assembly-out cross-validation, where each assembly is held out and the remaining 14 are partitioned with 10% for validation and 90% for training (n=15). (D) Plots of the ROC curves for leave-one-assembly-out cross-validation (n=15).

tools. Using the same preprocessing methods to featurize the data as with our convolutional neural network, we fitted these data to a random forest (RF) classifier and multinomial logistic regression. Evaluated under the same metrics as above, we find that these algorithms underperform our CNN by substantial margins.

In terms of classification accuracy, both the random forest and logistic regression models performed worse, at 77% and 58%, respectively. Similarly, when considering sensitivity versus specificity, the mean area under the receiver operating characteristic measures at 50% for RF and 48% for logistic regression. Finally, the area under the precision-recall curve tells a similar story, at 25% for RF and 21% for logistic regression.

Unsurprisingly, given the complexity and non-linearity of the features, logistic regression performed the worst, but all the methods suffered at classifying any of the SV types except inversions, with poor auROC and auPRC for the other three SV classes. Though the random forest, and to a lesser degree logistic regression, was able to predict SV classes with some degree of success, both methods suffer from scalability and memory issues compared to neural networks, which are also very sensitive to kernel optimization and generalize poorly to overlapping classes. Overall, our models are able to achieve a high level of accuracy on the problem of predicting SVs from short-read data, and they significantly outperform the less sophisticated models.

4.4 DISCUSSION

The detection and characterization of structural variants have long been a challenging endeavor in genomics. SVs are known to contribute significantly to genetic diversity and disease susceptibility, but their detection has been limited by the technical and computational constraints of traditional short-read sequencing technologies. This study aimed to overcome such limitations by developing a deep learning-based approach to detect SVs from short-read sequencing data. This approach, leveraging convolutional neural networks, was trained and validated on a high-quality, long-read sequencing dataset of 14 *Drosophila melanogaster* strains.

Our CNN-based model demonstrated high predictive performance, achieving an average accuracy of over 95% in predicting the correct SV class on the test set. The model also achieved an impressive mean overall auROC of over 95% for deletions, insertions, copy number variations, and inversions, respectively. The model demonstrated robust performance across different SV classes and was able to account for the class imbalance inherent to the SV detection problem.

The success of our approach is rooted in the ability of CNNs to extract high-level representations of biological sequences over the course of many convolutions, converting a one-dimensional sequence into a series of complex patterns. The model was designed with multiple convolutional layers to capture the complex patterns associated with SVs. The use of a sliding window approach also enabled the model to capture spatial information around SV breakpoints, which is critical for accurate SV detection.

While our model demonstrated robust performance in detecting SVs, it is important to note that it was trained and validated on a single species (*D. melanogaster*). Future work should investigate the generalizability of our approach to other species. Additionally, while our model excelled in detecting SVs from short-read sequencing data, long-read sequencing technologies continue to advance and are becoming increasingly accessible. Future studies should also explore the integration of long-read sequencing data into our deep learning framework to further enhance SV detection.

Our study also compared the performance of our CNN-based model to traditional linear classifiers and existing short-read SV detection software. We found that our model significantly outperformed these methods, demonstrating the power of deep learning for SV detection. However, it is important to note that these traditional methods are not without their merits. For instance, linear models, while less sophisticated, offer greater interpretability compared to deep learning models.

Similarly, existing short-read SV detection software have been widely used and validated in numerous studies. Therefore, while our study presents a new and powerful tool for SV detection, these traditional methods continue to have a place in the genomics toolkit.

In conclusion, we have developed a novel, CNN-based approach for detecting SVs from short-read sequencing data. This approach leverages the power of deep learning to capture the complex patterns associated with SVs, achieving high predictive performance. Our study contributes to the ongoing efforts to improve SV detection, offering a new tool that can aid in the understanding of genetic diversity and disease susceptibility. Future work should continue to refine and expand this deep learning framework, with the ultimate goal of enabling comprehensive and accurate SV detection across different species and sequencing technologies.

CHAPTER 5

Extracting phenological data from digitized herbarium specimens

5.1 INTRODUCTION

Herbaria, collections of preserved plant specimens stored at institutions around the world, serve as indispensable repositories of biodiversity data for the study of evolution, ecology, and plant phenology. These catalogs provide a rich source of historical, geographical, and phenological data. Herbarium specimens, which are generally accompanied by detailed labels containing information about the circumstances of their collection, provide a snapshot in time of a species at a particular location [104, 131]. The volume of these collections, combined with their extensive historical and geographical reach, makes herbaria a valuable resource for a wide range of biological and ecological research over time and space [38]. In the field of phenology, the study of periodic plant and animal life cycle events and how these are influenced by seasonal and interannual variations in climate, herbarium specimens play a crucial role. The phenological data that can be gleaned from herbarium specimens, such as flowering times and fruiting periods, provide a unique opportunity to study changes in these traits across temporal geographical gradients [37].

The digitization of herbaria has opened up a new frontier in herbarium-based research, facilitating the extraction of information from herbarium specimens on an unprecedented scale [104]. The process of digitizing herbarium specimens has been performed for a broad range of herbarium collections throughout the state of California and around the world. For example, the Consortium of California Herbaria (CCH) is an organization of dozens of herbaria, including the UC Irvine Herbarium, that collectively house the world's largest collection of California flora and contain millions of herbarium specimens [31, 139]. This process involves the creation of high-resolution digital images of herbarium specimens, which can be further processed and analyzed to extract valuable data.

These digitization efforts have the potential to revolutionize phenological research and other scientific fields by providing researchers with access to vast amounts of digitized specimen data. The existence of digitized herbarium collections has spurred an array of computational approaches aimed at extracting data from digitized specimens. These approaches range from automated species identification [19], to the extraction of phenological data [37], and the georeferencing of collection localities [144]. While significant progress has been made in some areas, the automated transcription and segmentation of herbarium specimen labels remain challenging tasks [60].

Image segmentation and transcription of text from these digitized specimens are key steps in the digitization process. Image segmentation refers to the partitioning of an image into separate components to extract the regions of the image that contain information of interest [128]. In the context of herbarium specimens, this involves identifying the boundaries of different components such as the plant material, the label, and annotations. The accurate segmentation of digitized specimen images is an essential step in obtaining useful information from each of the components at scale. After image segmentation, the text on the specimen label must be transcribed and then further parsed into standardized data fields, a process known as text segmentation. These steps are necessary for the extraction of specific pieces of information from the label, such as the scientific name of the species, the location and habitat where the specimen was collected, the date of collection, and the collector's name.

The primary challenge in this context is the variability of the herbarium specimen labels. The format, layout, font or handwriting, language, and level of detail can vary widely between different herbaria, between different collectors, and even between different specimens collected by the same person [60]. Additionally, accurately segmenting the different pieces of information in the labels into their correct data field is a complex task requiring domain knowledge. These challenges present substantial obstacles to using optical character recognition software for label transcription. As a result, label transcription is mostly performed by humans and often requires a significant amount of time and labor. Despite the essential role that label information plays in herbarium-based research, there are currently no widespread automated solutions for the transcription and segmentation of specimen labels.

In this chapter, we address this gap by developing a pipeline for the accurate automated transcription and segmentation of herbarium specimen labels. Our pipeline, which builds upon recent advances in computer vision, optical character recognition, natural language processing, and machine learning, is capable of handling the variability and complexity of herbarium specimen labels. It presents a significant step forward in the field of herbarium digitization, with the potential to greatly facilitate the extraction of data from digitized specimens and thereby further unlock the research potential of herbarium collections worldwide.

The digitization of herbaria presents a wealth of research opportunities, yet the transcription and segmentation of specimen labels remain significant challenges. Our pipeline addresses these challenges, providing a powerful tool for the extraction of data from digitized herbarium specimens. By facilitating the automated transcription and segmentation of specimen labels, our pipeline has the potential to greatly enhance the usability of digitized herbarium specimens and thereby contribute to a wide range of scientific research.

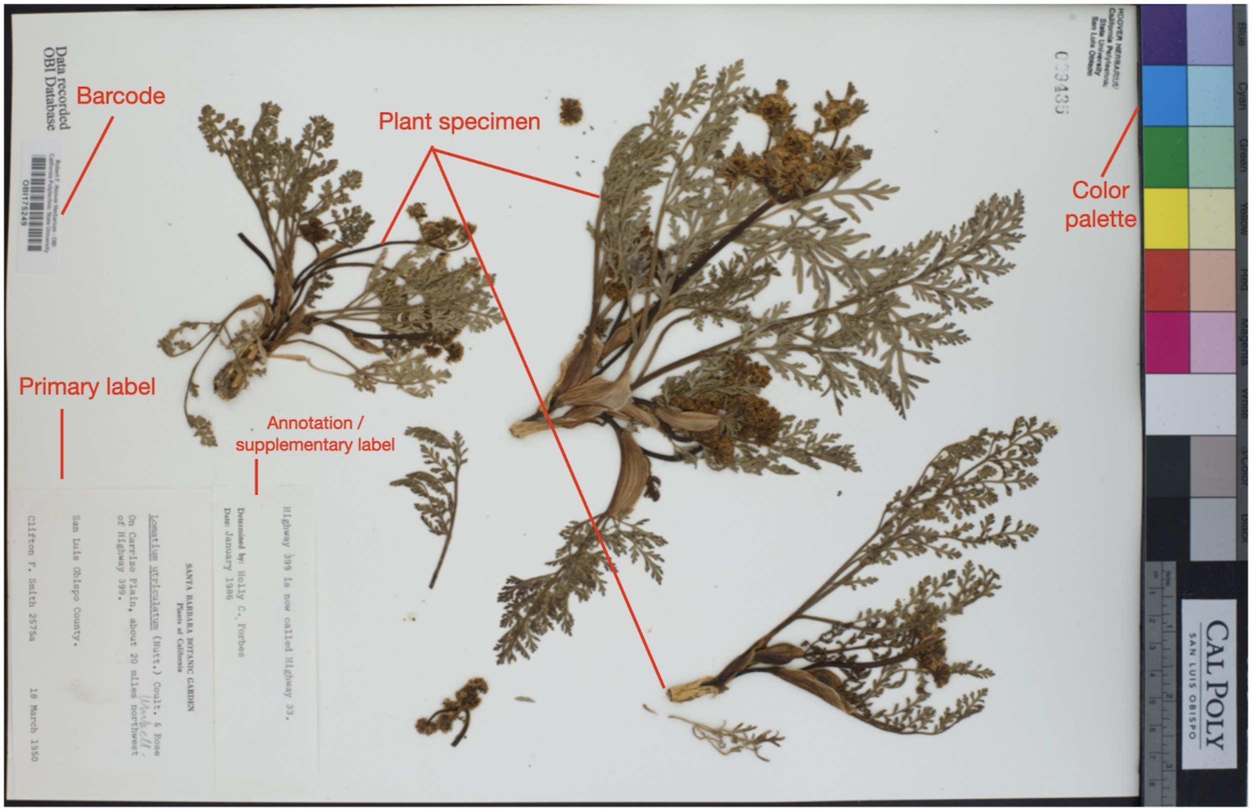


Figure 5.1: Example of an herbarium specimen, highlighting several important elements which are targets for segmenting and extracting for downstream analyses.

5.2 METHODS

5.2.1 Cleaning digitized specimen images

The procedure for cleaning the images as preprocessing steps includes adaptive steps, so specimen images are first cropped to remove the common header, which contains the reference color palette and ruler as well as the outer margins of the page, before establishing the calibration. The images are then preprocessed with a background-cleaning algorithm to remove shadows, noise, and extraneous background colors from the photograph. For this process, each specimen is split into three 8-bit images (i.e., the individual color channels). Each channel then undergoes a series of noise reduction and morphological operations, consisting of median blurring, dilation, and closing operations. Then, the luminosity of each channel is min-max normalized to the full 8-bit color range before the channels are recombined to produce a cleaned three-channel image. The contrast of the cleaned image is boosted to increase the separation between faint edges and the background.

5.2.2 Edge detection and enhancement

Edge detection represents a major challenge in the processing of herbarium specimen images. An effective edge detection methodology is necessary to accurately segment the image. However, specimens from different time periods and collections vary significantly in their condition, the number of segmentable elements, and the contrast between important elements and the background or other extraneous visual elements on the sheet. As various image transformations entail their own tradeoffs and information losses, we employ an ensemble of four different edge detection algorithms on the cleaned image in order to improve the discrimination of faint edges, such as the boundary between specimen labels and the sheet background, isolate the plant specimen itself from the rest of the elements on the page, and minimize the impact of extraneous signals.

The first algorithm, which is used to detect the location of the plant specimen, is based on lightness rescaling using a contrast-enhancing sigmoid transformation [15]. The use of sigmoid remapping functions preserves the perceived lightness contrast for color mapping tasks. The image is transformed according to Eq. (5.1), where the value represents the pixel intensity on a scale from 0 to 1:

$$\frac{1}{1 + e(\mathrm{Gain} * (\mathrm{Cutoff} - 1))}. \tag{5.1}$$

To increase the contrast between the plant specimen and the background while attenuating unwanted/confounding elements, we used a cutoff of 0.5 with a high gain of 25. The image was then converted to grayscale and binarized by thresholding it using the triangle method [155]. The transformed and binarized image was contour mapped. The specimen contours were heuristically selected by filtering for contours above a plausible contour-area threshold and comprised of at least 1000 unique points. The second algorithm uses dynamic range compression, followed by adaptive thresholding to binarize the image, in addition to contour stacking.

The third algorithm is based on the Canny edge detector, which is used to produce an initial binarized outline of the edges [17]. The output of the Canny edge detection then undergoes a series of sequential morphological dilations followed by morphological closing, which closes small gaps within edges and increases the distinction between background noise and actual edges. Small background artifacts from the transformations are removed from the image by contour detection the image and masking out objects below a certain contour area.

The edges detected through each of these algorithms are added together to obtain a combined map of detected edges. To close some small gaps in the detected edges due to a faint or obscured edge in the original image, we applied a Hough line transform on the detected edges and overlaid these lines back onto the edges to obtain a final outline of edges in the image. As shown in Fig. 2, the edges detected with each of these algorithms differ in their scope and resolution. When combined, the resulting ensemble provides a significantly more comprehensive map of edges by complementing missed elements or broken edges in each other's output, which substantially improves downstream image segmentation performance. The contour detection, morphological operations, Canny edge detection, Hough line transform, and thresholding were performed using functions from OpenCV 4.5.5, and the sigmoid transform was performed using scikit-image 0.19.1 [14, 140].

5.2.3 Visual segmentation of specimen images

Following the edge detection and enhancement, the next step involves the identification of recognizable shapes in the cleaned images. For the purpose of downstream processing of specimen labels, rectangular regions are of particular interest, as these regions could represent the outlines

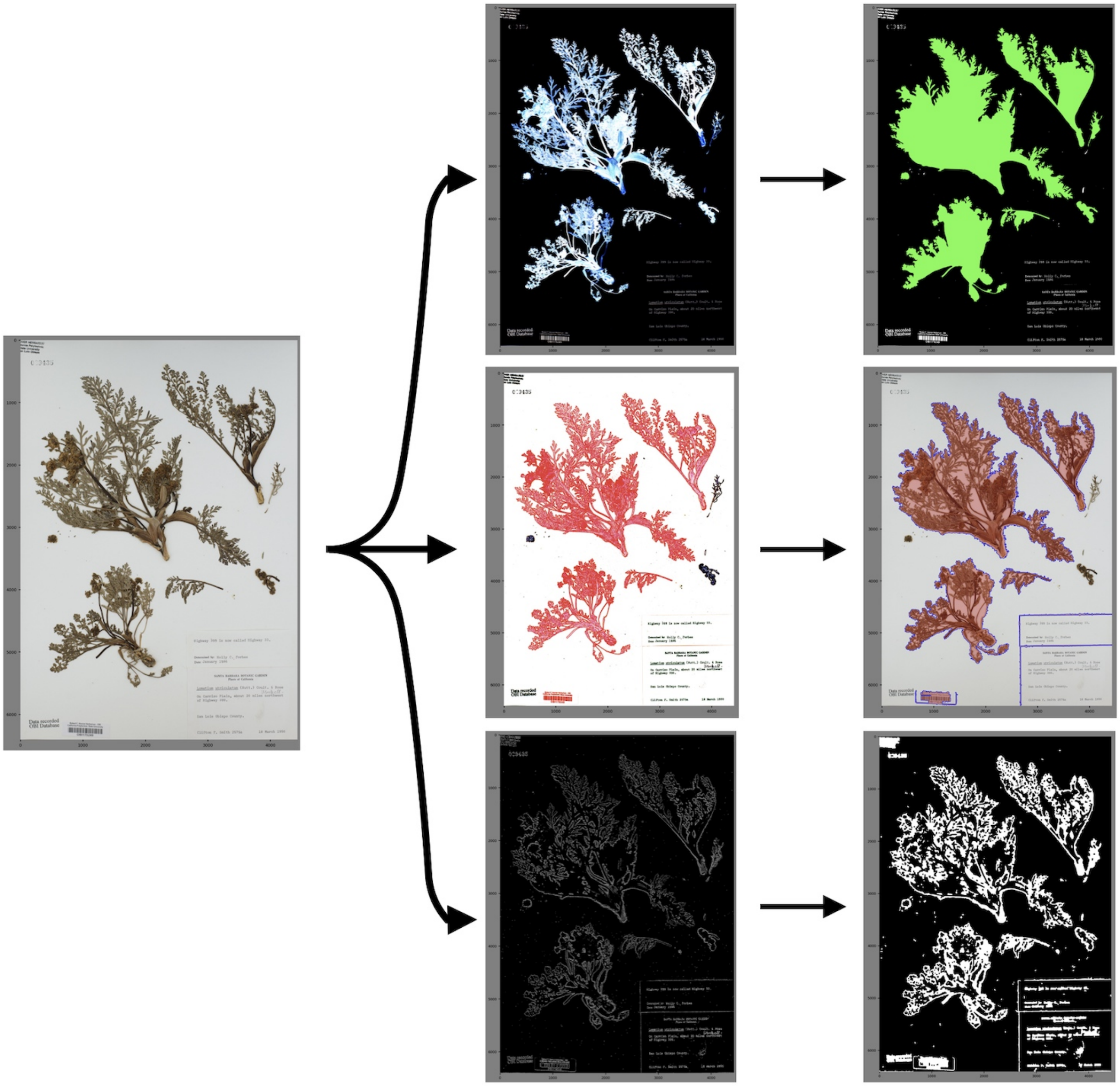


Figure 5.2: Intermediate representations of the three edge detection algorithms used for the computer vision-based segmentation of specimens. Red, green, and blue overlays have been used to illustrate the boundary detection and masking operations on different regions of the specimen at various stages of image processing.

of labels, annotations, collector notes, or other objects of interest on the herbarium specimen. We utilized the OpenCV library to detect these rectangular regions using a technique based on contour detection and approximation [14].

First, the combined edge map from the previous step was thresholded to create a binary image. The outlines in the binary image were found using contour detection. For each contour, a minimum bounding rectangle was computed. To filter out non-rectangular contours, only contour shapes that exceeded a certain threshold in terms of area and aspect ratio were retained. The remaining contours were assumed to be the outlines of other regions of interest.

Next, we generated prompts for the Fast Segment Anything Model (FastSAM) [158]. SAM is a model trained to segment objects in an image based on provided prompts. In our case, the prompts were the centroids of the detected contours in our previously generated edge map. The centroid of a contour was computed as the arithmetic mean of the coordinates of its four vertices. These centroids served as target points in the image to guide the SAM model in segmenting objects. The SAM model was run on the cleaned specimen images for each of the prompts. Following these prompted segmentations, we used the automatic mask generator to extract any additional segments detected by SAM.

Once the segments were obtained, a manual labeling of 200 hundred image segments was conducted. Each segment was assigned to a class that corresponded to the various objects that we were interested in extracting from a herbarium specimen image. To automate this task, a convolutional neural network was trained to classify the segments. Given the limited number of images, we decided to use transfer learning to leverage the knowledge gained from pre-trained models, with a 50/50 train/test split.

We used the Keras library and the VGG16 model, which is pre-trained on the ImageNet dataset [25, 124]. The last few layers of the pre-trained model were fine-tuned to adapt to our specific task. The images were resized to match the input size of the VGG16 model (224 x 224 pixels), and the pixel values were normalized. The model was trained using a categorical cross-entropy loss function and an Adam optimizer with a learning rate of 0.0001.

For segmented regions classified as a label or other text-containing rectangle, a final check was performed using the Efficient and Accurate Scene Text (EAST) detector [160]. EAST is a performant detector for identifying regions in an arbitrary image that contain text. The Intersection over Union (IoU) of the combined area of any text bounding boxes detected by EAST and the area of the detected rectangle was computed as a sanity check to ensure that rectangles classified as containing text have an IoU of at least 0.3.

The herbarium specimens used in the training and testing of our pipeline are from various collections in the Consortium of California Herbaria. Our dataset for evaluating performance comprised 200 herbarium specimens from 18 different institutions in the California Phenology Network [16]. This extensive collection provides a diverse and representative sample of specimen labels for training and validating our pipeline. As part of the California Phenology Thematic Collections Network (CAP-TCN) project, these specimens have been digitized as high-resolution photographs and were retrieved through the Consortium of California Herbaria web portal (CCH2) [109, 154].

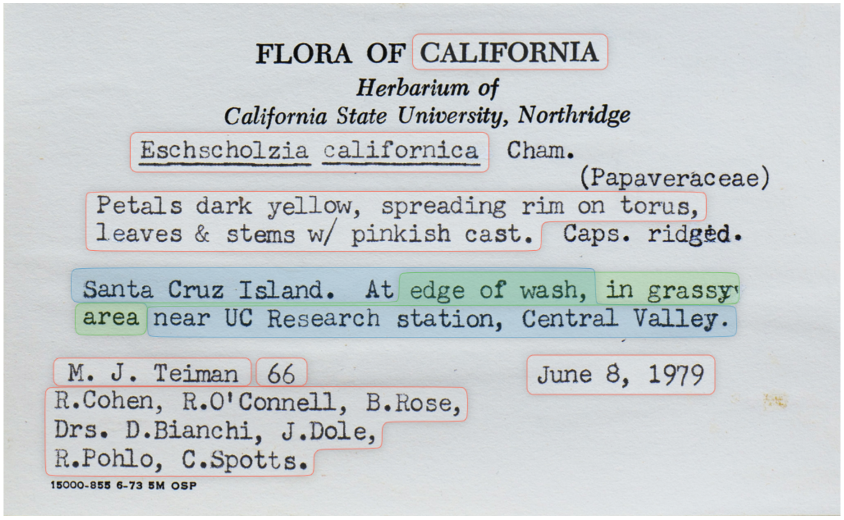
5.2.4 Detecting and extracting text from specimen labels

The labels on herbarium specimens contain the majority of the collected data on each specimen's record. Specimens in the CCH2 collection generally possess a common set of data fields in their labels, including information such as the specimen's scientific name, date of collection, collector name, locality, and habitat. However, due to the varied sources of the individual specimens across the consortium's collection as well as the extensive range of specimen ages and dates, the particular structure, and presentation of this information can vary significantly between specimens, presenting challenges such as handwritten labels, missing fields, and fading and physical degradation of the label. Due to the inherent complexity and inconsistency in the automated conversion of handwriting to text, specimens with handwritten labels were excluded.

First, the label boxes on each specimen isolated during the segmentation process are converted from an image to digitized text using the Tesseract 5.1.0 optical character recognition (OCR) software with a custom-trained long short-term memory model. This model was trained on the manually transcribed labels of 100 herbarium specimens from the CCH2 collection. To preseve the original layout of the label as much as possible, Tesseract was run with the ‘–psm 12’ and ‘-c preserve_interword_spaces=1’ options, directing the algorithm to treat the input as unstructured text and preserving horizontal whitespace. A dictionary containing the scientific names in the CCH2 collection and a catalog of locations across the United States was used to facilitate the accurate transcription and detection of important proper nouns and named entities in the OCR output.

A specimen label often contains multiple layouts of text within a single label, such as alternating numbers of columns on different rows. As a result, using the direct string output from the OCR will often result in lines of text from different semantic groupings being improperly combined and intermixed. To maintain the separation of these different regions of the label, the raw output from the OCR was manually parsed and concatenated by block to obtain a set of initial paragraphs, instead of using the OCR string output.

To split the paragraphs into sentences, the paragraphs were tokenized and split along sentence start and end tokens, using the SpaCy v3.5.3 library’s language model as the base for the tokenizer. We then segmented identifiable entities, such as scientific names, states and counties, dates, elevations, named locations, and geographical coordinates were extracted from the processed text using a combination of named entity recognition and regular expression pattern matching. To extract substrings for remaining fields such as the locality and habitat, the processed text was input to the GPT-3.5 large language model, using a structured multi-shot system prompt with the temperature parameter set to 0.5, at or below which the output was effectively deterministic in our testing.



FLORA OF CALIFORNIA
Herbarium of
California State University, Northridge
Eschscholzia californica Cham.
(Papaveraceae)
Petals dark yellow, spreading rim on torus,
leaves & stems w/ pinkish cast. Caps. ridged.

Santa Cruz Island. At edge of wash, in grassy
area near UC Research station, Central Valley.

M. J. Teiman 66 June 8, 1979
R.Cohen, R.O'Connell, B.Rose,
Drs. D.Bianchi, J.Dole,
R.Pohlo, C.Spotts.
15000-855 6-73 5M OSP

Figure 5.3: Example of a specimen label with various data fields highlighted, including the state, scientific name, collectors, number, date, habitat, locality, and notes. The green and blue outlines for the habitat and locality information illustrate the difficulty and complexity involved in transcribing many specimen labels, as these data, in addition to other types of information such as collector notes, are often overlapping or interspersed and are not easily identified or isolated from other data fields.

5.3 RESULTS

5.3.1 Specimen segmentation performance

We evaluated the performance of our computer vision and machine learning-based segmentation protocol by measuring its ability to identify, extract, and correctly categorize several important elements of an herbarium specimen. We tested the ability of our algorithms to segment four separate regions of interest in an specimen image, which are very likely to contain valuable information: the plant specimen itself, the primary label, the barcode, and annotations or other collector notes. The segmentation performance for each of these four elements is shown in Table 5.1.

The segmentation rate column in Table 5.1 measures the rate at which our segmentation algorithm correctly isolated a given type of element from the rest of the herbarium specimen sheet.

An element was only scored as a correct segmentation if the algorithm (1) correctly identified the element and (2) extracted a mask of the element within the specified IoU threshold discussed earlier. However, the accuracy column refers to the final accuracy over the entire pipeline, including both the segmentation and classification stages. When measuring our method's performance on the test dataset, since an element is only input into the neural network classifier if it was first successfully segmented, the accuracy value in Table 5.1 for a given type is element is always less than or equal to the segmentation rate. On the other hand, the precision column measures the positive predictive value of the neural network classifier itself. Thus, it is possible for the precision metrics to have higher values than that of the segmentation rate and net accuracy value.

Element	Segmentation Rate	Accuracy	Precision
Plant specimen	92%	90%	97.8%
Primary label	94%	91%	96.8%
Barcode	92%	92%	100%
Annotation	95.5%	89.5%	93.7%

Table 5.1: Performance of the segmentation algorithms and pipeline on the 100 test specimens. None of the specimens used for evaluating performance were included in any training step. The segmentation rate refers to the proportion of elements that were cleanly segmented from the rest of the specimen sheet, that is, we correctly detected its boundaries with an intersection over union (IoU) metric of over 90%, such that the element could be masked from the rest of the image for downstream analysis. The accuracy column refers to the overall rate at which we correctly segmented the element and correctly classified it into its type. The precision column refers to the positive predictive value of the classification among correctly segmented elements.

We achieve good performance on the segmentation of all four elements, and importantly, a very high accuracy for the extraction of the primary label. This is one of the most difficult elements to reliably segment across a wide range of specimens from the raw digitized specimens, as there is a large and unpredictable range of contrasts and anomalies which can reduce separation from the background. It is worth noting that although there are some misclassifications of primary labels and annotations labels, reducing accuracy as calculated by the metric reported in Table 5.1, the misclassification of one of these two element types is essentially always being assigned as the other. Since annotations often contain textual information of interest for biological analyses, such as collector's notes about the plant organism, such misclassifications often have no practical negative effect, as both primary labels and annotations are processed downstream with the label transcription and segmentation algorithms to extract their semantic content. An example of a segmented specimen, with masks overlaid on each of the elements of interest, is shown in Fig. 5.4.

5.3.2 Label transcription and segmentation

For segmenting the text from the specimen labels into separate categories, we followed the structure of the occurrence data fields in the Symbiota biodiversity data management system, which is the system used by many herberia collections around the world, including the CCH2 portal. We evaluated the label segmentation performance of our pipeline by measuring the rate at which we accurately assigned text transcribed from the specimen label to the appropriate field. We measured the segmentation accuracy for scientific name, collector, collection date, locality, and habitat fields, which together comprise a set of core information related to a specimen. For example, these fields often contain valuable ecological and phenological information such as flowering date, the precise location and habitat where a specimen was observed, and other species occurring alongside the collected specimen. The segmentation accuracy for these fields is shown in Table 5.2.

Data Field	**Text Segmentation Accuracy**
Scientific name	97%
Collector	89.6%
Date (Verbatim)	92%
Date (Month/Year)	94%
Locality	95.9%
Habitat	78.5%

Table 5.2: Performance of the text transcription and segmentation pipeline on the test specimens, measured as the rate of the accurate transcription, extraction, and assignment of information from the specimen label to the appropriate data field. The accuracy for each data field was calculated out of the total number of specimens whose labels contained the field. The five data fields benchmarked represent essential information about a specimen and the circumstances of its collection, and correspond to data fields in biodiversity data management portals.

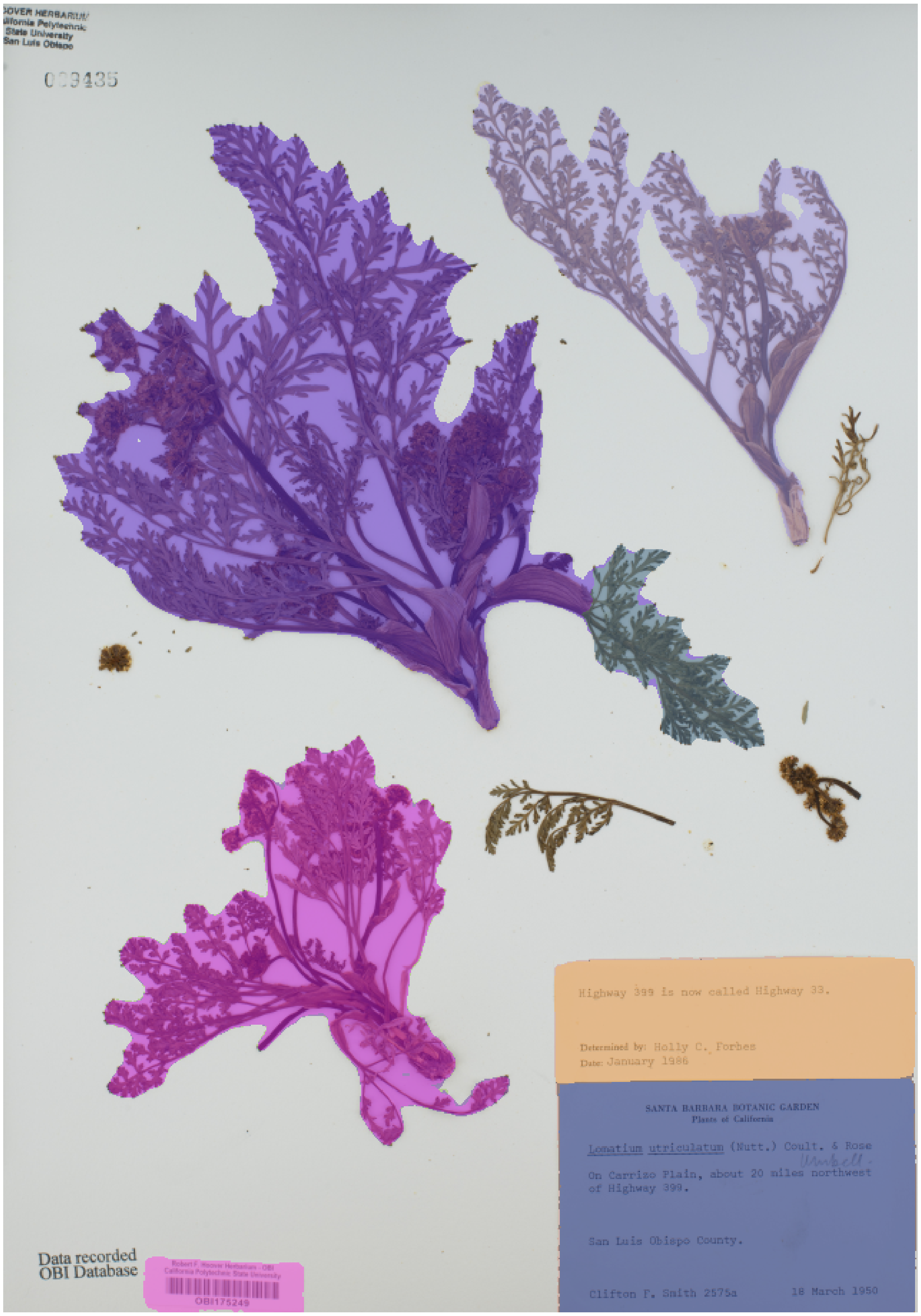

OVER HERBARIUM
lifornia Polytechnic
State University
San Luis Obispo
039435
Highway 399 is now called Highway 33.
Determined by: Holly C. Forbes
Date: January 1986
SANTA BARBARA BOTANIC GARDEN
Plants of California
Lomatium utriculatum (Nutt.) Coult. & Rose
Umbell.
On Carrizo Plain, about 20 miles northwest of Highway 399.
San Luis Obispo County.
Clifton F. Smith 2575a
18 March 1950
Data recorded
OBI Database
Robert F. Hoover Herbarium - OBI
California Polytechnic State University
OBI175249

Figure 5.4: The output of the segmentation algorithm with masks overlaid on top of the segmented elements of interest. The label, barcode, herbarium stamp, and preserved plant are each isolated with clean boundaries and can be extracted for downstream processing. The segmentation of the label enables more structured and accurate downstream optical character recognition and natural language processing, and the segmentation of the plant specimen enables automated annotation of flowering status.

5.4 DISCUSSION

The digitization of herbarium specimens at institutions around the world over the last decade has presented an extraordinary opportunity to leverage this rich resource at an unprecedented scale using 21st century computational techniques. One of the primary areas of need is an automated process for accurate transcription and segmentation of specimen labels. The specimen labels, while inherently containing much critical information about the specimen, have remained an obstacle for automated workflows and continue to rely on substantial human time and labor to transcribe.

Consequently, addressing this long-standing challenge was a central priority in the work described in this chapter. We developed a framework for converting raw digitized specimens, i.e., an image of a specimen sheet, into a segmented and cataloged set of information for downstream analyses. In particular, we developed a series of algorithms and a pipeline for segmenting regions of interest in a specimen image, extracting specimen labels, and transcribing the labels and organizing their information into the appropriate data fields.

Automated approaches to utilizing digitized herbaria, including machine learning approaches, are typically focused on species identification and phenological classification following segmentation of the physical plant specimen from specimen images [70]. However, herbarium label transcription can be a deceptively difficult task for computers to perform. There are numerous challenges inherent to this task, such as the amount and density of information contained on the labels, the repetition and overlap of information across different fields, the variable layout of the information, and the inconsistency in the appearance, location, presence, and terminology of information across different specimens and over time.

Many of these challenges have been noted in earlier work on utilizing digitized herbarium specimens. Even when OCR tools or commercial services are used to facilitate the transcription process, intervention by a human with domain knowledge is often still required to accurately identify and separate the various fields and appropriately catalog them [8]. For example, an earlier digitization effort found that transcriptions provided by a company from a commercial contract still required cleaning or adjustments for 12-60% of occurrences depending on the data field [46]. Additionally, many older specimens have handwritten labels, which is an ongoing challenge in the domain of OCR software.

Due to the aforementioned difficulties in extracting data from herbarium specimen labels, there have been relatively few tools for automating this task, especially compared to other uses of digitized herbaria such as automated species identification. The work in this chapter expands upon the scope of previous efforts towards automated label transcription. While our approach and semi-automated workflows, such as the SALIX method [8], have certain features in common, such as OCR and named-entity recognition techniques, our approach is a fully automated pipeline which seeks to produce segmented data fields of the label information from a digitized specimen image without the need for human input at intervening steps.

Although species identification and extraction of visual phenological features are not the focus of the present study, we hope that our versatile preprocessing algorithms and high-resolution segmentation algorithms, which aim to robustly isolate relevant portions of digitized specimens and preserve fine detail in features such as flower and leaf boundaries, can facilitate such efforts.

Systemic literature reviews of computer vision and machine learning methods applied to digitized herbarium specimens have noted that most studies did not apply much or any segmentation to the specimens before applying downstream classification algorithms [67, 150]. Relatedly, the specimen images used tend to have plain backgrounds with a single mature leaf, though we realistically expect to deal with specimens with multiple leaves as well as variable backgrounds and lighting conditions across an herbarium collection [150]. While deep-learning-based models can outperform traditional image processing and machine learning methods for species identification and phenological annotation tasks, they can also be more susceptible to noise and biases caused by the numerous elements on a specimen image which are extraneous to the task at hand, such as barcodes, labels, and color palettes [37].

It has also been noted that studies tackling species identification overall have achieved greater success than those for detecting phenological features, in part because of the difficulties in object detection and segmentation tasks with herbarium specimens [67]. As species identification tasks do not necessarily require fine-scale analysis of specific features of the plant specimen, which are essential for phenological studies, it is comparatively easier to compile a ground truth dataset associating digitized herbarium specimen images in their entirety with their corresponding plant species [67].

As robust and versatile image segmentation was a primary methodological focus of this work, since reliably segmenting the label itself is a nontrivial computer vision task and often more difficult than segmenting the plant itself from a specimen sheet, we hope that some of our techniques will be able to complement existing tools for working with digitized herbarium specimens, alleviating issues related to visual differences due to variations in the digitization imaging process, the presence of extraneous elements, inadequate segmentation of different parts of the specimen, and other confounding factors.

CHAPTER 6

Conclusion

In the preceding chapters, we have discussed the development and application of novel algorithms to a diverse set of scientific questions to address the unmet need for extracting meaningful biological features from large-scale genomic and other datasets. The research undertaken in these chapters demonstrates the power of computational and quantitative approaches to tackle complex problems in genetics, ecology, and evolution, as well as the opportunity and need for ongoing development in this area. It also highlights the importance of high-quality reference genomes, machine learning, and digitization efforts in advancing these fields.

The accurate determination of allele frequencies is essential for various genetic applications, including population genetics, genome-wide association studies, and disease risk assessment. Sequencing technologies have made it possible to obtain vast genetic variation data efficiently. To address the limitation of barcoding in traditional methods, we developed new method called FREQ-Seq^2, allowing the simultaneous analysis of allele frequencies in a significantly larger number of samples. By using paired barcode sequences, this approach exponentially increases the throughput of the FREQ-Seq concept, enabling scalability. FREQ-Seq^2 is designed to target specific genomic regions, making it versatile and cost-effective. It has been validated for accuracy and precision and is particularly useful for studying population dynamics, evolutionary genetics, and genotype distributions in mixed populations. The method includes a kit with barcoded adapter fragments and open-source software for data analysis, making it a powerful, flexible, high-throughput, and economical tool for allele frequency quantification.

Despite the rapid accumulation of genomic data, many causal mutations linked to phenotypic variations remain unidentified, primarily due to the limitations of current genotyping approaches which often overlook structural variants (SVs), genomic rearrangements such as inversions, translocations, duplications, and deletions. Recognizing the importance of SVs in understanding complex traits, our team produced a new reference-quality genome assembly for the *Drosophila melanogaster* strain A4. This assembly revealed an abundance of previously undetected genetic variations, including those potentially responsible for complex traits, such as a transposable element linked to nicotine resistance. This study underscores the significance of high-quality reference genomes in unearthing crucial genetic variations that contribute to phenotypes.

SVs have a significant impact on phenotypes, and can cause or influence the course of various diseases. Nevertheless, despite their significance, they have been challenging to detect compared to smaller genetic variations, such as single nucleotide polymorphisms. This is due to technological constraints and the lack of robust reference databases for SVs. However, with the advent of long-read sequencing technologies, like PacBio and Oxford Nanopore, researchers can generate more comprehensive genomic data, although these techniques have their own set of challenges. We leveraged a previously assembled high-quality genome dataset of 14 strains of Drosophila melanogaster to develop a machine learning model aimed at enhancing SV identification in short-read datasets. By employing deep convolutional neural networks, the model demonstrated a high accuracy level of 92%, surpassing other models and algorithms. Using this model, the team could identify SVs with high precision from vast published Drosophila short-read datasets.

Together, our work utilizes a combination of computational approaches to improve the detection of SVs and demonstrates their application in established sample datasets. Future approaches in the development of computational methods for SV detection in both bulk and single-cell sequencing techniques would provide improved tools for the identification and functional interpretation of SVs from large-scale sequencing datasets.

In the final chapter of this thesis, we extend our work to demonstrate the utility of applying machine learning, computer vision, and natural language processing techniques to historical and ongoing digitization of herbaria catalogs, as a powerful tool for extracting valuable phenological data from these invaluable collections. Herbaria, worldwide collections of preserved plant specimens, serve as invaluable resources for biodiversity research, offering insights into evolution, ecology, and plant phenology. These collections offer a temporal and geographical snapshot of plant species. The ongoing digitization of herbaria is revolutionizing research by converting these physical specimens into digital formats for easier access and analysis. However, extracting data from these digital images, especially transcribing and segmenting the varied and complex specimen labels, remains challenging. Addressing this, we introduce a new pipeline that employs computer vision, optical character recognition, natural language processing, and machine learning to automatically transcribe and segment herbarium labels. Using specimens from the Consortium of California Herbaria, the pipeline's utility was demonstrated by replicating a phenological analysis, emphasizing its potential in transforming how digitized herbarium data is used for various research endeavors.

Bibliography

[1] The human genome at ten. *Nature*, 464(7289):649–650, 2010.

[2] H. J. Abel, D. E. Larson, A. A. Regier, C. Chiang, I. Das, K. L. Kanchi, R. M. Layer, B. M. Neale, W. J. Salerno, C. Reeves, S. Buyske, NHGRI Centers for Common Disease Genomics, T. C. Matise, D. M. Muzny, M. C. Zody, E. S. Lander, S. K. Dutcher, N. O. Stitziel, and I. M. Hall. Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, 583(7814):83–89, 2020.

[3] A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 21(6):974–984, 2011.

[4] S. G. Acinas, R. Sarma-Rupavtarm, V. Klepac-Ceraj, and M. F. Polz. PCR-induced sequence artifacts and bias: Insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and Environmental Microbiology*, 71(12):8966–8969, 2005.

[5] C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363–376, 2011.

[6] C. Alkan, S. Sajjadian, and E. E. Eichler. Limitations of next-generation genome sequence assembly. *Nature Methods*, 8(1):61–65, 2010.

[7] S. Alon, F. Vigneault, S. Eminaga, D. C. Christodoulou, J. G. Seidman, G. M. Church, and E. Eisenberg. Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Research*, 21(9):1506–1511, 2011.

[8] A. Barber, D. Lafferty, and L. R. Landrum. The SALIX method: A semi-automated workflow for herbarium specimen digitization. *TAXON*, 62(3):581–590, 2013.

[9] S. Bartoszewski, S. Luschnig, I. Desjeux, J. Grosshans, and C. Nüsslein-Volhard. *Drosophila* p24 homologues *eclair* and *baiser* are necessary for the activity of the maternally expressed Tkv receptor during early embryogenesis. *Mechanisms of Development*, 121(10):1259–1273, 2004.

[10] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[11] C. M. Bergman and P. R. Haddrill. Strain-specific and pooled genome sequences for populations of *Drosophila melanogaster* from three continents. *F1000Research*, 4:31, 2015.

[12] K. Berlin, S. Koren, C.-S. Chin, J. P. Drake, J. M. Landolin, and A. M. Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33(6):623–630, 2015.

[13] D. Bikard, D. Patel, C. L. Metté, V. Giorgi, C. Camilleri, M. J. Bennett, and O. Loudet. Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science*, 323(5914):623–626, 2009.

[14] G. Bradski. The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 2000.

[15] G. J. Braun. Image lightness rescaling using sigmoidal contrast enhancement functions. *Journal of Electronic Imaging*, 8(4):380, 1999.

[16] California Phenology Network. California Phenology Collections Network, 2023. https://www.capturingcaliforniasflowers.org.

[17] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.

[18] B. C. Carlton and B. J. Brown. *Manual of methods for general bacteriology*, chapter Gene mutation, page 222–242. American Society for Microbiology, Washington, D.C., 1981.

[19] J. Carranza-Rojas and E. Mata-Montero. On the significance of leaf sides in automatic leaf-based plant species identification. In *2016 IEEE 36th Central American and Panama Convention*. IEEE, 2016.

[20] M. Chakraborty, J. G. Baldwin-Brown, A. D. Long, and J. J. Emerson. Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, 44(19):e147, 2016.

[21] M. Chakraborty, J. J. Emerson, S. J. Macdonald, and A. D. Long. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nature Communications*, 10(1), 2019.

[22] M. Chakraborty, N. W. VanKuren, R. Zhao, X. Zhang, S. Kalsow, and J. J. Emerson. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nature Genetics*, 50(1):20–25, 2018. doi: 10.1038/s41588-017-0010-y.

[23] S. Chen, Y. E. Zhang, and M. Long. New genes in *Drosophila* quickly become essential. *Science*, 330(6011):1682–1685, 2010.

[24] C. Chiang, A. J. Scott, J. R. Davis, E. K. Tsang, X. Li, Y. Kim, T. Hadzic, F. N. Damani, L. Ganel, S. B. Montgomery, A. Battle, D. F. Conrad, and I. M. Hall. The impact of structural variation on human gene expression. *Nature Genetics*, 49(5):692–699, 2017.

[25] F. Chollet et al. Keras, 2015.

[26] L. M. Chubiz, M.-C. Lee, N. F. Delaney, and C. J. Marx. FREQ-Seq: a rapid, cost-effective, sequencing-based method to determine allele frequencies directly from mixed populations. *PLoS One*, 7(10):479–59, 2012.

[27] H. Chung, M. R. Bogwitz, C. McCart, A. Andrianopoulos, R. H. ffrench Constant, P. Batterham, and P. J. Daborn. *Cis*-regulatory elements in the *Accord* retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene *Cyp6g1*. *Genetics*, 175(3):1071–1077, 2007.

[28] P. P. Cleary and E. Englesberg. Transcriptional control in the L-arabinose operon of *Escherichia coli* B/r. *Journal of Bacteriology*, 118(1):121–128, 1974.

[29] R. L. Collins, H. Brand, K. J. Karczewski, X. Zhao, J. Alföldi, L. C. Francioli, A. V. Khera, C. Lowther, L. D. Gauthier, H. Wang, N. A. Watts, M. Solomonson, A. O'Donnell-Luria, A. Baumann, R. Munshi, M. Walker, C. W. Whelan, Y. Huang, T. Brookings, T. Sharpe, ..., J. I. Rotter, C. Nusbaum, A. Philippakis, E. Lander, S. Gabriel, B. M. Neale, S. Kathiresan, M. J. Daly, E. Banks, D. G. MacArthur, and M. E. Talkowski. A structural variation reference for medical and population genetics. *Nature*, 581(7809):444–451, 2020.

[30] S. Collins, J. de Meaux, and C. Acquisti. Adaptive walks toward a moving optimum. *Genetics*, 176(2):1089–1099, 2007.

[31] Consortium of California Herbaria. CCH: About, 2023. https://ucjeps.berkeley.edu/consortium/about.html.

[32] J. M. Cridland, S. J. Macdonald, A. D. Long, and K. R. Thornton. Abundance and distribution of transposable elements in two *Drosophila* QTL mapping resources. *Molecular Biology and Evolution*, 30(10):2311–2327, 2013.

[33] J. M. Cridland, K. R. Thornton, and A. D. Long. Gene expression variation in *Drosophila melanogaster* due to rare transposable element insertion alleles of large effect. *Genetics*, 199(1):85–93, 2014.

[34] J. F. Crow. Perspective: Here's to Fisher, additive genetic variance, and the fundamental theorem of natural selection. *Evolution*, 56(7):1313–16, 2002.

[35] J. Dabney and M. Meyer. Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques*, 52(2), 2012.

[36] F. J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.

[37] B. H. Daru, D. S. Park, R. B. Primack, C. G. Willis, D. S. Barrington, T. J. S. Whitfeld, T. G. Seidler, P. W. Sweeney, D. R. Foster, A. M. Ellison, and C. C. Davis. Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist*, 217(2):939–955, 2017.

[38] C. C. Davis. The herbarium of the future. *Trends in Ecology & Evolution*, 38(5):412–423, 2023.

[39] J. A. G. de Visser and R. E. Lenski. Long-term experimental evolution in *Escherichia coli*. XI. Rejection of non-transitive interactions as cause of declining rate of adaptation. *BMC Evolutionary Biology*, 2:19, 2002.

[40] M. DeGiorgio, C. D. Huber, M. J. Hubisz, I. Hellmann, and R. Nielsen. Sweepfinder2: increased sensitivity, robustness and flexibility. *Bioinformatics*, 32(12):1895–1897, 2016.

[41] G. Dietzl, D. Chen, F. Schnorrer, K.-C. Su, Y. Barinova, M. Fellner, B. Gasser, K. Kinsey, S. Oppel, S. Scheiblauer, A. Couto, V. Marra, K. Keleman, and B. J. Dickson. A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. *Nature*, 448(7150):151–156, 2007.

[42] G. dos Santos, A. J. Schroeder, J. L. Goodman, V. B. Strelets, M. A. Crosby, J. Thurmond, D. B. Emmert, and W. M. G. and. FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research*, 43(D1):D690–D697, 2014.

[43] R. Durrett and J. Schweinsberg. Approximating selective sweeps. *Theoretical Population Biology*, 66(2):129–138, 2004.

[44] E. E. Eichler, J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450, 2010.

[45] J. J. Emerson, M. Cardoso-Moreira, J. O. Borevitz, and M. Long. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science*, 320(5883):1629–1631, 2008.

[46] H. Engledow. Herbarium specimen label interpretation and transcription: First steps used to clean digitized data. *Biodiversity Information Science and Standards*, 6, 2022.

[47] J. Fadista, A. K. Manning, J. C. Florez, and L. Groop. The (in)famous GWAS *P*-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, 24:1202–1205, 2016.

[48] J. C. Fay and C.-I. Wu. Hitchhiking under positive darwinian selection. *Genetics*, 155(3):1405–1413, 2000.

[49] R. A. Fisher. The genetical theory of natural selection, 1930.

[50] A.-S. Fiston-Lavier, N. D. Singh, M. Lipatov, and D. A. Petrov. *Drosophila melanogaster* recombination rate calculator. *Gene*, 463(1-2):18–20, 2010.

[51] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251, 2009.

[52] E. R. Gamazon, D. L. Nicolae, and N. J. Cox. A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. *PLoS Genetics*, 7(2):e1001292, 2011.

[53] J. H. Gillespie. *Population Genetics: A Concise Guide*. Johns Hopkins University Press, 2nd edition, 2004.

[54] J. I. Glendinning. How do herbivorous insects cope with noxious secondary plant compounds in their diet? *Entomologia Experimentalis et Applicata*, 104(1):15–25, 2002.

[55] I. Gordo and P. R. A. Campos. Evolution of clonal populations approaching a fitness peak. *Biology Letters*, 9:1–4, 2012.

[56] E. W. Green, G. Fedele, F. Giorgini, and C. P. Kyriacou. A *Drosophila* RNAi collection is subject to dominant phenotypic effects. *Nature Methods*, 11(3):222–223, 2014.

[57] J. K. Grenier, J. R. Arguello, M. C. Moreira, S. Gottipati, J. Mohammed, S. R. Hackett, R. Boughton, A. J. Greenberg, and A. G. Clark. Global diversity lines–a five-continent reference panel of sequenced *Drosophila melanogaster* strains. *G3: Genes|Genomes|Genetics*, 5(4):593–603, 2015.

[58] R. W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160, 1950.

[59] J. P. Heredia, B. Trubenová, D. Sudholt, and T. Paixão. Selection limits to adaptive walks on correlated landscapes. *Genetics*, 205(2):803–825, 2017.

[60] A. Hill, R. Guralnick, A. Smith, A. Sallans, R. Gillespie, M. Denslow, J. Gross, Z. Murrell, T. Conyers, P. Oboyski, J. Ball, A. Thomer, R. Prys-Jones, J. de la Torre, P. Kociolek, and L. Fortson. The notes from nature tool for unlocking biodiversity records from museum records through citizen science. *ZooKeys*, 209:219–233, 2012.

[61] T. Horn and M. Boutros. E-RNAi: a web application for the multi-species design of RNAi reagents—2010 update. *Nucleic Acids Research*, 38(suppl_2):W332–W339, 2010.

[62] R. A. Hoskins, J. W. Carlson, K. H. Wan, S. Park, I. Mendez, S. E. Galle, B. W. Booth, B. D. Pfeiffer, R. A. George, R. Svirskas, M. Krzywinski, J. Schein, M. C. Accardo, E. Damia, G. Messina, M. Méndez-Lago, B. de Pablos, O. V. Demakova, E. N. Andreyeva, L. V. Boldyreva, M. Marra, A. B. Carvalho, P. Dimitri, A. Villasante, I. F. Zhimulev, G. M. Rubin, G. H. Karpen, and S. E. Celniker. The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Research*, 25(3):445–458, 2015.

[63] R. A. Hoskins, C. D. Smith, J. W. Carlson, A. B. Carvalho, A. Halpern, J. S. Kaminker, C. Kennedy, C. J. Mungall, B. A. Sullivan, G. G. Sutton, J. C. Yasuhara, B. T. Wakimoto, E. W. Myers, S. E. Celniker, G. M. Rubin, and G. H. Karpen. Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. *Genome Biology*, 3(12):research0085.1–0085.16, 2002.

[64] C. D. Huber, M. DeGiorgio, I. Hellmann, and R. Nielsen. Detecting recent selective sweeps while controlling for mutation rate and background selection. *Molecular Ecology*, 25(1):142–156, 2015.

[65] J. Huddleston and E. E. Eichler. An incomplete understanding of human genetic variation. *Genetics*, 202(4):1251–1254, 2016.

[66] R. R. Hudson. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[67] B. R. Hussein, O. A. Malik, W.-H. Ong, and J. W. F. Slik. Applications of computer vision and machine learning techniques for digitized herbarium specimens: A systematic literature review. *Ecological Informatics*, 69:101641, 2022.

[68] B. Jarvis. *Statistical Aspects of the Microbiological Examination of Foods*, chapter Errors associated with colony count procedures, pages 119–140. Academic Press, 3rd edition, 2016.

[69] J. Kang, J. Kim, and K.-W. Choi. Novel cytochrome P450, *cyp6a17*, is required for temperature preference behavior in *Drosophila*. *PLoS One*, 6(12):e29800, 2011.

[70] N. Katal, M. Rzanny, P. Mäder, and J. Wäldchen. Deep learning in plant phenological research: A systematic literature review. *Frontiers in Plant Science*, 13, 2022.

[71] P. D. Keightley, R. W. Ness, D. L. Halligan, and P. R. Haddrill. Estimation of the Spontaneous Mutation Rate per Nucleotide Site in a Drosophila melanogaster Full-Sib Family. *Genetics*, 196(1):313–320, 2014.

[72] D. E. Khost, D. G. Eickbush, and A. M. Larracuente. Single molecule long read sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. 2016.

[73] E. G. King, G. Kislukhin, K. N. Walters, and A. D. Long. Using *Drosophila melanogaster* to identify chemotherapy toxicity genes. *Genetics*, 198(1):31–43, 2014.

[74] E. G. King, C. M. Merkes, C. L. McNeil, S. R. Hoofer, S. Sen, K. W. Broman, A. D. Long, and S. J. Macdonald. Genetic dissection of a model complex trait using the *Drosophila* Synthetic Population Resource. *Genome Research*, 22(8):1558–1566, 2012.

[75] G. Kirov, I. Nikolov, L. Georgieva, V. Moskvina, M. J. Owen, and M. C. O'Donovan. Pooled DNA genotyping on Affymetrix SNP genotyping arrays. *BMC Genomics*, 7(1):27, 2006.

[76] D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. Mardis. The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1):27–38, 2013.

[77] S. W. Kong, I.-H. Lee, X. Liu, J. N. Hirschhorn, and K. D. Mandl. Measuring coverage and accuracy of whole exome sequencing in clinical context. *Genetics in Medicine*, 20(12):1617–1626, 2018.

[78] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. Versatile and open software for comparing large genomes. *Genome Biology*, 5(2):R12, 2004.

[79] J. B. Lack, C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, K. A. Stevens, C. H. Langley, and J. E. Pool. The *Drosophila* Genome Nexus: A population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics*, 199(4):1229–1241, 2015.

[80] K.-K. Lam, K. LaButti, A. Khalak, and D. Tse. FinisherSC: a repeat-aware tool for upgrading *de novo* assembly using long reads. *Bioinformatics*, 31(19):3207–3209, 2015.

[81] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[82] N. S. Latysheva and M. M. Babu. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Research*, 44(10):4487–4503, 2016.

[83] R. E. Lenski. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *The ISME Journal*, 11(10):2181–2194, 2017.

[84] 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*, 138(6):1315–41, 1991.

[85] R. E. Lenski and M. Travisano. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences*, 91(15):6808–14, 1994.

[86] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[87] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[88] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. D. and. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[89] K. E. Lohmueller, T. Sparsø, Q. Li, E. Andersson, T. Korneliussen, A. Albrechtsen, K. Banasik, N. Grarup, I. Hallgrimsdottir, K. K. T. O. Kilpeläinen, N. T. Krarup, T. H. Pers, G. Sanchez, Y. Hu, M. DeGiorgio, T. Jørgensen, A. Sandbæk, T. Lauritzen, S. Brunak, K. Kristiansen, Y. Li, T. Hansen, J. Wang, R. Nielsen, and O. Pedersen. Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. *The American Journal of Human Genetics*, 93(6):1072–1086, 2013.

[90] A. D. Long, R. F. Lyman, A. H. Morgan, C. H. Langley, and T. F. C. Mackay. Both naturally occurring insertions of transposable elements and intermediate frequency polymorphisms at the achaete-scute complex are associated with variation in bristle number in *Drosophila melanogaster*. *Genetics*, 154(3):1255–1269, 2000.

[91] J. R. Lupski, R. M. de Oca-Luna, S. Slaugenhaupt, L. Pentao, V. Guzzetta, B. J. Trask, O. Saucedo-Cardenas, D. F. Barker, J. M. Killian, C. A. Garcia, A. Chakravarti, and P. I. Patel. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*, 66(2):219–232, 1991.

[92] M. Lynch, D. Bost, S. Wilson, T. Maruki, and S. Harrison. Population-genetic inference from pooled-sequencing data. *Genome Biology and Evolution*, 6(5):1210–18, 2014.

[93] T. F. C. Mackay, S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. H. Anholt, M. Barrón, ..., L. Nazareth, I. Newsham, L. Perales, L.-L. Pu, C. Qu, M. Ràmia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L. Turlapati, K. C. Worley, Y.-Q. Wu, A. Yamamoto, Y. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman, and R. A. Gibbs. The *Drosophila melanogaster* Genetic Reference Panel. *Nature*, 482(7384):173–178, 2012.

[94] H. A. MacMillan, J. M. Knee, A. B. Dennis, H. Udaka, K. E. Marshall, T. J. S. Merritt, and B. J. Sinclair. Cold acclimation wholly reorganizes the *Drosophila melanogaster* transcriptome and metabolome. *Scientific Reports*, 6(1), 2016.

[95] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[96] T. N. Marriage, E. G. King, A. D. Long, and S. J. Macdonald. Fine-mapping nicotine resistance loci in *Drosophila* using a multiparent advanced generation inter-cross population. *Genetics*, 198(1):45–57, 2014.

[97] J. P. Masly, C. D. Jones, M. A. F. Noor, J. Locke, and H. A. Orr. Gene transposition as a cause of hybrid sterility in *Drosophila*. *Science*, 313(5792):1448–1450, 2006.

[98] O. V. Matveeva, Y. D. Nechipurenko, E. Riabenko, C. Ragan, N. N. Nazipova, A. Y. Ogurtsov, and S. A. Shabalina. Optimization of signal-to-noise ratio for efficient microarray probe design. *Bioinformatics*, 32(17):i552–i558, 2016.

[99] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

[100] modENCODE Consortium. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*, 330(6012):1787–1797, 2010.

[101] B. Monsion, H. Duborjal, and S. Blanc. Quantitative single-letter sequencing: a method for simultaneously monitoring numerous known allelic variants in single DNA samples. *BMC Genomics*, 9(1):85, 2008.

[102] S. Nair, D. S. Kim, J. Perricone, and A. Kundaje. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics*, 35(14):i108–i116, 2019.

[103] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[104] G. Nelson and S. Ellis. The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1763):20170391, 2018.

[105] R. Nielsen, S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. Genomic scans for selective sweeps using SNP data. *Genome Research*, 15(11):1566–1575, 2005.

[106] H. A. Orr. Fitness and its role in evolutionary genetics. *Nature Reviews Genetics*, 10(8):531–39, 2009.

[107] A. W. Pang, J. R. MacDonald, D. Pinto, J. Wei, M. A. Rafiq, D. F. Conrad, H. Park, M. E. Hurles, C. Lee, J. C. Venter, E. F. Kirkness, S. Levy, L. Feuk, and S. W. Scherer. Towards a comprehensive structural variation map of an individual human genome. *Genome Biology*, 11(5):R52, 2010.

[108] S. T. Park and J. Kim. Trends in next-generation sequencing and a new era for whole genome sequencing. *International Neurourology Journal*, 20(Suppl 2):S76–83, 2016.

[109] K. D. Pearson, G. Nelson, M. F. J. Aronson, P. Bonnet, L. Brenskelle, C. C. Davis, E. G. Denny, E. R. Ellwood, H. Goëau, J. M. Heberling, A. Joly, T. Lorieul, S. J. Mazer, E. K. Meineke, B. J. Stucky, P. Sweeney, A. E. White, and P. S. Soltis. Machine learning using digitized herbarium specimens to advance phenological research. *BioScience*, 70(7):610–620, 2020.

[110] J. H. F. Pedra, L. M. McIntyre, M. E. Scharf, and B. R. Pittendrigh. Genome-wide transcription profile of field- and laboratory-selected dichlorodiphenyltrichloroethane (DDT)-resistant *Drosophila*. *Proceedings of the National Academy of Sciences*, 101(18):7034–7039, 2004.

[111] J. T. Peeler, J. E. Leslie, J. W. Danielson, and J. W. Messer. Replicate counting errors by analysts and bacterial colony counters. *Journal of Food Protection*, 45(3):238–240, 1982.

[112] G. H. Perry, F. Yang, T. Marques-Bonet, C. Murphy, T. Fitzgerald, A. S. Lee, C. Hyland, A. C. Stone, M. E. Hurles, C. Tyler-Smith, E. E. Eichler, N. P. Carter, C. Lee, and R. Redon. Copy number variation and evolution in humans and chimpanzees. *Genome Research*, 18(11):1698–1710, 2008.

[113] D. A. Petrov, A.-S. Fiston-Lavier, M. Lipatov, K. Lenkov, and J. Gonzalez. Population genomics of transposable elements in *Drosophila melanogaster*. *Molecular Biology and Evolution*, 28(5):1633–1644, 2010.

[114] A. R. Quinlan. BEDTools: The swiss-army tool for genome feature analysis. *Current Protocols in Bioinformatics*, 47(1), 2014.

[115] M. V. Rockman. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution*, 66(1):1–17, 2011.

[116] R. L. Rogers, J. M. Cridland, L. Shao, T. T. Hu, P. Andolfatto, and K. R. Thornton. Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*. *Molecular Biology and Evolution*, 31(7):1750–1766, 2014.

[117] M. G. Ross, C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, and D. B. Jaffe. Characterizing and measuring bias in sequence data. *Genome Biology*, 14(5):R51, 2013.

[118] S. Saleem, C. C. Schwedes, L. L. Ellis, S. T. Grady, R. L. Adams, N. Johnson, J. R. Whittington, and G. E. Carney. *Drosophila melanogaster* p24 trafficking proteins have vital roles in development and reproduction. *Mechanisms of Development*, 129(5-8):177–191, 2012.

[119] J. S. Sanjak, A. D. Long, and K. R. Thornton. A model of compound heterozygous, loss-of-function alleles is broadly consistent with observations from complex-disease GWAS datasets. *PLoS Genetics*, 13(1):1–30, 2017.

[120] S. E. Schoustra, T. Bataillon, D. R. Gifford, and R. Kassen. The properties of adaptive walks in evolving populations of fungus. *PLoS Biology*, 7(11):e1000250, 2009.

[121] A. J. Scott, C. Chiang, and I. M. Hall. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Research*, 31(12):2249â€“2257, 2021.

[122] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, 2015.

[123] A. T. Simonsen, M. C. Hansen, E. Kjeldsen, P. L. Møller, J. J. Hindkjær, P. Hokland, and A. Aggerholm. Systematic evaluation of signal-to-noise ratio in variant detection from single cell genome multiple displacement amplification and exome sequencing. *BMC Genomics*, 19(1):681, 2018.

[124] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[125] J. M. Smith and J. Haigh. The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(1):23–35, 1974.

[126] M. Stapleton, G. Liao, P. Brokstein, L. Hong, P. Carninci, T. Shiraki, Y. Hayashizaki, M. Champe, J. Pacleb, K. Wan, C. Yu, J. Carlson, R. George, S. Celniker, and G. M. Rubin. The *Drosophila* Gene Collection: Identification of putative full-length cDNAs for 70% of *D. melanogaster* genes. *Genome Research*, 12(8):1294–1300, 2002.

[127] K. A. Steige, B. Laenen, J. Reimegård, D. G. Scofield, and T. Slotte. Genomic analysis reveals major determinants of *cis*-regulatory variation in *Capsella grandiflora*. *Proceedings of the National Academy of Sciences*, 114(5):1087–92, 2017.

[128] G. Stockman and L. G. Shapiro. *Computer Vision*. Prentice Hall PTR, USA, 1st edition, 2001.

[129] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavaré, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, 2007.

[130] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, ..., J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, and J. O. Korbel. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015.

[131] J. R. Sullivan and M. Nazaire. Specimen collection and preparation for a changing flora. *Rhodora*, 123(993), 2022.

[132] I. A. Swinburne and P. A. Silver. Intron delays and transcriptional timing during development. *Developmental Cell*, 14(3):324–330, 2008.

[133] M. E. Tabangin, J. G. Woo, and L. J. Martin. The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proceedings*, 3(Suppl 7):41, 2009.

[134] Y. Tanabe and T. Ishida. Quantification of the accuracy limits of image registration using peak signal-to-noise ratio. *Radiological Physics and Technology*, 10(1):91–94, 2017.

[135] O. Tenaillon, A. Rodriguez-Verdugo, R. L. Gaut, P. McDonald, A. F. Bennett, A. D. Long, and B. S. Gaut. The molecular diversity of adaptive convergence. *Science*, 335(6067):457–461, 2012.

[136] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[137] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–578, 2012.

[138] M. Travisano and R. E. Lenski. Long-term experimental evolution in *Escherichia coli*. IV. Targets of selection and the specificity of adaptation. *Genetics*, 143(1):15–26, 1996.

[139] UCI Herbarium. The University of California, Irvine Herbarium – IRVC, 2023. https://arboretum.bio.uci.edu/plant-exhibits/herbarium/.

[140] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[141] B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, and A. M. Earl. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9(11):e112963, 2014.

[142] J. Wasson. Allele quantification and DNA pooling methods. *Methods in Molecular Biology*, 373:63–74, 2007.

[143] J. Weischenfeldt, O. Symmons, F. Spitz, and J. O. Korbel. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, 14(2):125–138, 2013.

[144] J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais. Darwin Core: An evolving community-developed biodiversity data standard. *PLoS One*, 7(1):e29715, 2012.

[145] S. Wilkening, K. Hemminki, R. Kumar Thirumaran, J. Lorenzo Bermejo, S. Bonn, A. Försti, and R. Kumar. Determination of allele frequency in pooled DNA: comparison of three PCR-based methods. *BioTechniques*, 39(6):853–858, 2005.

[146] M. J. Wiser and R. E. Lenski. A comparison of methods to measure fitness in *Escherichia coli*. *PLoS One*, 10(5):1–11, 2015.

[147] M. J. Wiser, N. Ribeck, and R. E. Lenski. Long-term dynamics of adaptation in asexual populations. *Science*, 342(6164):1364–1367, 2013.

[148] R. Woods, D. Schneider, C. L. Winkworth, M. A. Riley, and R. E. Lenski. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 103(24):9107–12, 2006.

[149] N. R. Wray, J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard, and P. M. Visscher. Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, 14(7):507–515, 2013.

[150] J. Wäldchen and P. Mäder. Plant species identification using computer vision techniques: A systematic literature review. *Archives of Computational Methods in Engineering*, 25(2):507–543, 2017.

[151] A. Wünsche, D. M. Dinh, R. S. Satterwhite, C. D. Arenas, D. M. Stoebel, and T. F. Cooper. Diminishing-returns epistasis decreases adaptability along an evolutionary trajectory. *Nature Ecology and Evolution*, 1(4):0061, 2017.

[152] C. Ye, C. M. Hill, S. Wu, J. Ruan, and Z. Ma. DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific Reports*, 6(1), 2016.

[153] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, 2009.

[154] J. M. Yost, K. D. Pearson, J. Alexander, E. Gilbert, L. A. Hains, T. Barry, R. Bencie, P. Bowler, B. Carter, R. E. Crowe, E. Dean, J. Der, A. Fisher, K. Fisher, L. Flores-Renteria, C. M. Guilliams, C. Hatfield, L. Hendrickson, T. Huggins, L. Janeway, C. Lay, A. Litt, S. Markos, S. J. Mazer, D. McCamish, L. McDade, M. Mesler, B. Mishler, M. Nazaire, J. Rebman, L. Rosengreen, P. W. Rundel, D. Potter, A. Sanders, K. C. Seltmann, M. G. Simpson, G. A. Wahlert, K. Waselkov, K. Williams, and P. S. Wilson. The California phenology collections network: using digital images to investigate phenological change in a biodiversity hotspot. *Madroño*, 66(4):130, 2020.

[155] G. W. Zack, W. E. Rogers, and S. A. Latt. Automatic measurement of sister chromatid exchange frequency. *Journal of Histochemistry & Cytochemistry*, 25(7):741–753, 1977. PMID: 70454.

[156] X. Zhang and J. J. Emerson. Inferring compensatory evolution of *cis*- and *trans*-regulatory variation. *Trends in Genetics*, 35(1):1–3, 2019.

[157] R. Zhao, T. Lukacsovich, R. Gaut, and J. J. Emerson. FREQ-Seq^2: a method for precise high-throughput combinatorial quantification of allele frequencies. *G3: Genes|Genomes|Genetics*, 13(10):jkad162, 2023. doi: 10.1093/g3journal/jkad162.

[158] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang. Fast Segment Anything, 2023.

[159] J. Zhou and O. G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 2015.

[160] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. EAST: An efficient and accurate scene text detector. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.