UNIVERSITY OF CALIFORNIA, SAN DIEGO

The Genetic Basis of Hypoxia Tolerance

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics & Systems Biology

by

Nitin Udpa

Committee in charge:

Professor Vineet Bafna, Chair
Professor Kelly A. Frazer, Co-Chair
Professor Trey G. Ideker
Professor Pavel A. Pevzner
Professor Glenn P. Tesler
Professor Dan Zhou

2013

The dissertation of Nitin Udpa is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2013

DEDICATION

To my parents, for their many sacrifices over the years.

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Table of Contents v

List of Figures viii

List of Tables x

Acknowledgements xi

Vita xiv

Abstract of the Dissertation xv

Chapter 1 Introduction 1
1.1 What is hypoxia? 1
1.2 How do we identify the genetic basis of a phenotype? . 2
1.3 Specific goals of the dissertation 5

Chapter 2 RAPID detection of gene-gene interactions underlying quantitative traits 6
2.1 Abstract 6
2.2 Introduction 7
2.3 Materials and Methods 10
2.3.1 Simulated dataset 10
2.3.2 Uncoupling the mixture of Gaussians 11
2.4 Results 12
2.4.1 Determination of optimal EM parameters 12
2.4.2 The impact of quantitative phenotypes on the randomized algorithms 15
2.4.3 Application to Drosophila dataset 17
2.5 Discussion 19
2.6 Acknowledgments 21

Chapter 3 Exploring tests of selection in case-control data 22
3.1 Abstract 22
3.2 Introduction 24
3.3 Materials and Methods 27
3.3.1 Statistics for detecting selection 28
3.3.2 Forward simulation 29

3.3.3 Preprocessing . . . 32
3.3.4 Power and significance computations . . . 34
3.4 Results . . . 36
3.4.1 Power versus time under different measures of θ 36
3.4.2 Window size . . . 41
3.5 Discussion . . . 45
3.6 Acknowledgments . . . 47

Chapter 4 The role of the Notch pathway in hypoxia-tolerant *Drosophila melanogaster* . . . 48
4.1 Abstract . . . 48
4.2 Introduction . . . 49
4.3 Results . . . 50
4.3.1 Hypoxia-selected regions and genetic profiles in the adapted genome . . . 50
4.3.2 Candidate genes underlying hypoxia tolerance in AF . . . 53
4.3.3 Notch activation is critical for hypoxia tolerance in *Drosophila melanogaster* . . . 53
4.3.4 Spatial-temporal activation of Notch and its downstream genes in hypoxia tolerance . . . 55
4.4 Discussion . . . 57
4.5 Materials and Methods . . . 62
4.5.1 *Drosophila* stocks and culture . . . 62
4.5.2 Whole-genome resequencing . . . 63
4.5.3 Data analysis . . . 64
4.5.4 Test for selection in pooled data . . . 65
4.5.5 DAPT treatment . . . 65
4.5.6 Hypoxia tolerance and vulnerability tests . . . 66
4.6 Acknowledgments . . . 66

Chapter 5 Evolutionarily conserved hypoxia tolerance genes in Ethiopian highlanders . . . 68
5.1 Abstract . . . 68
5.2 Introduction . . . 69
5.3 Results . . . 70
5.4 Discussion . . . 79
5.5 Materials and Methods . . . 81
5.5.1 Ethics statement . . . 81
5.5.2 Sample description . . . 82
5.5.3 DNA extraction, library construction and sequencing . . . 82

5.5.4 Read alignment, score recalibration and variant calling . . . 82
5.5.5 Variant filters . . . 83
5.5.6 Control populations . . . 84
5.5.7 Identifying regions under positive selection . . . 84
5.5.8 Tests of selection . . . 85
5.5.9 Population simulations and power estimation . . . 86
5.5.10 Power of different tests under varying model parameters . . . 87
5.5.11 Region filtration . . . 90
5.5.12 Power of whole genome sequencing . . . 92
5.5.13 Drosophila stocks and test of hypoxia tolerance 93
5.6 Acknowledgments . . . 94

Chapter 6 The genetic basis of chronic mountain sickness (Monge's disease) in Andean highlanders . . . 98
6.1 Abstract . . . 98
6.2 Introduction . . . 99
6.3 Materials and Methods . . . 100
6.3.1 Subjects and clinical characterization . . . 100
6.3.2 DNA extraction, library construction and sequencing . . . 100
6.3.3 Read alignment, score recalibration and variant calling . . . 101
6.3.4 Tests of selection . . . 102
6.3.5 Human fibroblast cell culture, hypoxia treatment and real-time qPCR assay . . . 102
6.3.6 Fly lines and culture . . . 103
6.3.7 In-vivo hypoxia tolerance test . . . 103
6.4 Results . . . 104
6.4.1 Lowlander control populations . . . 104
6.4.2 Variant filters . . . 104
6.4.3 Identifying regions under positive selection . . . 106
6.4.4 Region filtration . . . 107
6.4.5 In vitro and in vivo validation of candidate genes 110
6.5 Discussion . . . 114
6.6 Acknowledgments . . . 115

Bibliography . . . 117

LIST OF FIGURES

Figure 2.1: The impact of the EM confidence threshold on the ability to capture an interacting pair of SNPs 13
Figure 2.2: The expectation of correctly classified points given a 0.65 threshold 14
Figure 2.3: The impact of standard deviation on RAPID's performance . 16
Figure 2.4: The performance of RAPID and SixPAC on simulated datsets 17
Figure 2.5: Manhattan plot showing significance of single-locus interactions with the hypoxia phenotype 19

Figure 3.1: Schematic of simulator 31
Figure 3.2: Power of different statistics as a function of time since bottleneck 33
Figure 3.3: Tests of the distributions of the S_{f} and S_{W} statistics in simulated and *Drosophila* (C1 & C2) control versus control data 35
Figure 3.4: SNP frequency spectrum analysis 38
Figure 3.5: Power versus time after bottleneck 40
Figure 3.6: ROC Curve evaluating the impact of threshold on true positive and false positive rates 41
Figure 3.7: Time to optimal power as a function of selection pressure . . 42
Figure 3.8: The impact of simulator parameters on the power of the S_{f} statistic 43
Figure 3.9: Contour plot describing determination of window size 44

Figure 4.1: Genomic regions under selection for hypoxia tolerance 52
Figure 4.2: Enrichment of fixed SNPs and indels in an extended Notch pathway 54
Figure 4.3: Hypoxia tolerance in Notch mutants and γ-secretase inhibitor-treated flies 56
Figure 4.4: Increased hypoxia tolerance with Eaat-GAL4-driven NICD overexpression in fly glial cells 58
Figure 4.5: Increased hypoxia tolerance in flies with P{GawB}17A-Gal4-driven NICD overexpression in specific glial cells 59
Figure 4.6: Notch activation-conferred hypoxia tolerance requires E(spl)mα 60

Figure 5.1: Computational analysis work flow 71
Figure 5.2: ADMIXTURE analysis with 6 clusters on the Ethiopian highlanders, along with the 1000 Genomes populations 72
Figure 5.3: Power as function of time for neutrality tests used in this study (S_{π}, F_{st}, S_{f}, and PBS) 73
Figure 5.4: The impact of sampling $n = 14$ haplotypes (the sample size of our Amhara population) from a population on estimated allele frequencies 74

Figure 5.5: PBS statistic across chromosome 19 in the Oromos population compared to both Luhya (LWK) and European (CEU) populations 76
Figure 5.6: Evidence supporting EDNRB as a gene candidate 77
Figure 5.7: RNAi-mediated knockdown of orthologs of candidate genes enhanced hypoxia tolerance in Drosophila 79
Figure 5.8: Impact of whole genome sequencing on selection signals . . . 95
Figure 5.9: The impact of sequenced sample size on power, using S_π as an exemplar test 96
Figure 5.10: Illustration of a selective bottleneck in one of two diverged populations, leading to a loss of genetic diversity 97

Figure 6.1: Computational analysis workflow 101
Figure 6.2: ADMIXTURE analysis (4 clusters) applied to the 1000 Genomes populations jointly with our Andean highlander subjects . . . 105
Figure 6.3: Profile of the only two candidate regions that are significant in non-CMS vs. CMS and non-CMS vs. MXL tests 111
Figure 6.4: Hypoxia response of top candidate genes in non-CMS and CMS cells 113
Figure 6.5: Down-regulation of human SENP1 and ANP32D ortholog in Drosophila enhances survival under hypoxic condition 114

LIST OF TABLES

Table 2.1: List of single loci associated with the hypoxia phenotype, sorted by P-value 20
Table 2.2: List of pairs of loci jointly associated with survival rates under hypoxic conditions, sorted by P-value 21

Table 3.1: Notation and definition of the eight statistics tested in this chapter 29

Table 5.1: List of significant genomic regions found in Amhara and Oromos populations 78

Table 6.1: List of the top 11 regions prioritized in non-CMS relative to all lowlander controls 110

ACKNOWLEDGEMENTS

Although I will get most of the credit for the work in this thesis, this would have been nothing without the help of many others. I would first like to thank my advisor, Dr. Vineet Bafna. Like any good advisor, he has taught me how to be competent in my field, encompassing various aspects of population genetics, biology, and computer science. However, unlike most advisors, his more important contribution to me has been to teach me life skills that will be useful outside my field as well. I feel I have learned the most from him over the past six years in clarity. Specifically, I believe I have made substantial progress in approaching and formulating problems clearly, communicating ideas with both colleagues and non-scientists clearly, and understanding existing algorithms clearly. I would also like to thank the rest of my thesis committee (Dr. Kelly Frazer, Dr. Pavel Pevzner, Dr. Glenn Tesler, Dr. Trey Ideker, and Dr. Dan Zhou) for their thoughtful comments and guidance throughout this journey.

As a computational person in a field defined by biology, I have been fortunate to work with many brilliant scientists who have helped guide and validate the projects I have worked on. While all of my co-authors (see below) have contributed to this work, in particular, I would like to thank (soon-to-be Dr.) Roy Ronen for virtually daily discussions over the direction in which our analyses should proceed. In addition, none of this work would have been possible without the vision of our collaborators from the Department of Pediatrics (especially Dr. Gabriel G. Haddad and Dr. Dan Zhou) — the design and cleanliness of the datasets made our work very easy.

While the research aspect was essential, the friendships I have built up over the years made me realize that virtually everyone had the same sets of problems with paper rejections, deadlines, and underprepared lab meetings. The members of the Bafna, Pevzner, and Bandeira labs (such as Anand Patel, Stefano Bonissone, Christine Lo, Son Pham, and Sangtae Kim) were instrumental in providing a good balance between academic and non-academic mindsets. Students in the Bioinformatics and Systems Biology Graduate Program (such as Joshua Lerman, Josue Perez Santiago, and Vipul Bhargava) and others (such as Balaji Sriram,

Arun Manohar, and Lauren Sadler) have mentored me and kept me grounded. In particular, I would like to acknowledge the other three members of my class – Harish Nagarajan, Nisha Rajagopal, and Roger Chang – for all the shared memories. Despite our small number, your talent and hard work has caused our year to be one of the most successful in the history of the bioinformatics program.

Finally, I would not even be in this position without the sacrifices made by my family over the years. For instance, several years ago, my father quit his PhD, since a grad student's salary was insufficient to feed a demanding baby. My mother went to work at 5AM every day so she could transport me to my extracurricular activities in the afternoon. My sister has taught me that unconditional love and happiness lead to fulfillment. I have been fortunate enough to have wonderful uncles, aunts, grandparents, and even great-grandparents that have fostered my intellectual curiosity over the years.

Chapter 2 is unpublished work. The dissertation author was the primary author of this paper responsible for the research.

Chapter 3 is in full adapted from *Udpa N, Zhou D, Haddad GG, Bafna V. Tests of selection in pooled case-control data: an empirical study. Front Genet. (2011);2(83).* The dissertation author was the primary author of this paper responsible for the research.

Chapter 4 is in full adapted from *Zhou D, Udpa N*, Gersten M*, Visk DW*, Bashir A, Xue J, Frazer KA, Posakony JW, Subramaniam S, Bafna V, Haddad GG. Experimental selection of hypoxia-tolerant Drosophila melanogaster. Proc Natl Acad Sci U S A. (2011);108(6):2349-2354.* The dissertation author was a second author of this manuscript, and was responsible for the "coarse-grained" computational analysis.

Chapter 5 is in full material submitted for publication from *Udpa N*, Ronen R*, Zhou D*, Liang J, Stobdan T, Appenzeller O, Yin Y, Du Y, Guo L, Cao*

R, Wang Y, Jin X, Huang C, Jia W, Cao D, Guo G, Claydon VE, Hainsworth R, Gamboa JL, Zibenigus M, Zenebe G, Xue J, Liu S, Frazer KA, Li Y, Bafna V, Haddad GG. Whole Genome Sequencing of Ethiopian Highlanders Reveals Evolutionarily Conserved Hypoxia Tolerance Genes. The dissertation author was a joint first author of this paper, responsible for much of the computational analysis.

Chapter 6 is in full material submitted for publication from *Zhou D*, Udpa N*, Ronen R*, Stobdan T, Liang J, Appenzeller O, Zhao HW, Yin Y, Du Y, Guo L, Cao R, Wang Y, Jin X, Huang C, Jia W, Cao D, Guo G, Gamboa JL, Villafuerte F, Callacondo D, Xue J, Liu S, Frazer KA, Li Y, Bafna V*, Haddad GG*. Genetic mechanisms underlie chronic mountain sickness (Monge's disease) in Andean highlanders.* The dissertation author was a joint first author of this paper, responsible for much of the computational analysis.

VITA

2007	B.S. in Biomedical Engineering, University of Texas at Austin
2013	Ph. D. in Bioinformatics & Systems Biology, University of California, San Diego

PUBLICATIONS

Udpa N*, Ronen R*, Zhou D*, Liang J, Stobdan T, Appenzeller O, Yin Y, Du Y, Guo L, Cao R, Wang Y, Jin X, Huang C, Jia W, Cao D, Guo G, Claydon VE, Hainsworth R, Gamboa JL, Zibenigus M, Zenebe G, Xue J, Liu S, Frazer KA, Li Y, Bafna V, Haddad GG. Whole Genome Sequencing of Ethiopian Highlanders Reveals Evolutionarily Conserved Hypoxia Tolerance Genes. *Submitted* *Equal contribution

Zhou D*, **Udpa N***, Ronen R*, Stobdan T, Liang J, Appenzeller O, Zhao HW, Yin Y, Du Y, Guo L, Cao R, Wang Y, Jin X, Huang C, Jia W, Cao D, Guo G, Gamboa JL, Villafuerte F, Callacondo D, Xue J, Liu S, Frazer KA, Li Y, Bafna V*, Haddad GG*. Genetic mechanisms underlie chronic mountain sickness (Monge's disease) in Andean highlanders. *Submitted* *Equal contribution

Ronen R, **Udpa N**, Halperin E, Bafna V. Learning natural selection from the site frequency spectrum. *Submitted*

Dixon-Salazar TJ, He H, Nguyen MB, **Udpa N**, et al. Mutations in ADPRHL2, encoding ADP2-ribosylhydrolase like 2, lead to a novel, potentially treatable, lethal epilepsy syndrome. *Submitted*.

Dixon-Salazar TJ, Silhavy JL, **Udpa N**, Schroth J, Bielas S, Schaffer AE, Olvera J, Bafna V, Zaki MS, Abdel-Salam GH, Mansour LA, Selim L, Abdel-Hadi S, Marzouki N, Ben-Omran T, Al-Saana NA, Sonmez FM, Celep F, Azam M, Hill KJ, Collazo A, Fenstermaker AG, Novarino G, Akizu N, Garimella KV, Sougnez C, Russ C, Gabriel SB, Gleeson JG. Exome sequencing can improve diagnosis and alter patient management. Sci Trans Med. (2012);4(138):p.138ra78.

Udpa N, Zhou D, Haddad GG, Bafna V. Tests of selection in pooled case-control data: an empirical study. Front Genet. (2011);2(83).

Zhou D, **Udpa N***, Gersten M*, Visk DW*, Bashir A, Xue J, Frazer KA, Posakony JW, Subramaniam S, Bafna V, Haddad GG. Experimental selection of hypoxia-tolerant Drosophila melanogaster. Proc Natl Acad Sci U S A. (2011);108(6):2349-2354. *Equal contribution

ABSTRACT OF THE DISSERTATION

The Genetic Basis of Hypoxia Tolerance

by

Nitin Udpa

Doctor of Philosophy in Bioinformatics & Systems Biology

University of California, San Diego, 2013

Professor Vineet Bafna, Chair
Professor Kelly A. Frazer, Co-Chair

Research into hypoxia (or low oxygen levels) has been a hot topic for a number of decades, because many harmful diseases, such as heart attacks and cancer, create much of their damage by inducing hypoxia. It has been suspected for years that the ability of a cell (or an organism) to cope with a hypoxic environment is, at least in part, influenced by genetic factors. However, for financial reasons, virtually all studies that have attempted to find these factors have been constrained to a subset of variant sites (targeted genes, exons, or genotyping arrays). As the costs of sequencing drop, though, whole-genome sequencing will become increasingly used. The primary goal of this dissertation is to build computational tools that use the power of whole-genome sequencing to identify genetic variants that

can confer tolerance to hypoxia.

Even though the basic computational problem is one of correlation, the experimental design plays a huge role in determining the best way to measure this correlation. First, we discuss ways to identify correlated sites in a typical association study. While the single-locus case is trivial to solve, extending this to multiple loci is intractable using a naive approach. We discuss existing randomized algorithms that solve this problem quickly, and extend these algorithms to handle quantitative phenotypes. We then apply one of these approaches to identify interacting sites correlating with survival rate under acute hypoxia. We then focus on a different problem — detecting natural selection. As the signatures of natural selection are dependent on several parameters, such as selection pressure and time under selection, which are largely unknown, we compare the performance of a number of tests over a wide range of parameters and identify optimal regimes for each of them. We then select a statistic appropriate for strong, laboratory selection and use it to identify elements of the Notch repression mechanism in flies that have adapted to extreme hypoxia (4% O_2). Finally, we apply a number of these statistics to two different populations of humans adapting to mild hypoxia, identifying novel and distinct mechanisms in both cases.

Chapter 1

Introduction

1.1 What is hypoxia?

The genomic era has advanced at a staggeringly fast pace. Just one decade ago, the scientific community had "finished" sequencing the first human genome, at the cost of approximately 3 billion dollars [1]. Since that time, technological improvements have increased the throughput of sequencers while decreasing the cost tremendously, far surpassing the rate predicted by Moore's law [2]. This has resulted in a number of hypotheses (directly or indirectly related to human health) which can be feasibly tested in a "high-throughput" manner.

For instance, hypoxia (or low oxygen supply to part or all of an organism) is an important trait of several diseases. There are two major causes of hypoxia — malfunctions of part of the body and changes to the environment. For the former, cardiovascular diseases such as heart attack or stroke act by reducing blood (and thus, oxygen) flow to the heart or brain, respectively. Cancer cells proliferate so rapidly that the oxygen level per cell is much lower, necessitating mechanisms such as angiogenesis to cope with the demand. Respiratory diseases like emphysema lead to less oxygen entering the bloodstream and thus, hypoxic stress throughout the organism. Another form of hypoxia occurs not due to any inherent malfunction in the organism, but rather due to environmental conditions. For instance, in high altitudes, the overall atmospheric pressure is reduced, lowering the concentration of oxygen in the air.

The body responds to hypoxia based on the severity and duration of the stress. The body's natural response is essentially one of panic. It initially tries to increase blood flow to the hypoxic region. For more acute cases, this is often unable to restore oxygen levels (for instance, due to a blockage). As a result, the body initiates apoptotic pathways in affected cells. The resulting tissue necrosis is the major damage associated with the cardiovascular diseases mentioned above. In chronic cases (such as high altitude dwellers), increasing the blood flow helps in the short term, but the perpetual nature of the stress leads to a condition known as *chronic mountain sickness* (CMS), or Monge's disease [3]. The root cause of the symptoms is increased blood viscosity, in an attempt to increase the capacity of oxygen per unit volume of blood. This results in symptoms such as headache, dizziness, and fatigue. The only known treatments for this are descending to lower environments and phlebotomies (blood-letting).

Interestingly, many individuals who live at high altitudes have adapted well to the environment and are relatively symptom-free. It is suspected that this adaptation is at least partly heritable, although it was only in recent years that researchers attempted to identify the genetic basis of this adaptation.

1.2 How do we identify the genetic basis of a phenotype?

In order to identify the genetic basis of phenotypes such as these, we first need a good representation of the data inherent in a population. For the following few paragraphs, the specific numbers represent data for human beings (*Homo sapiens*), but the same principles apply for all species. The heritable genetic material (DNA) in an individual is organized as two sets of 23 chromosomes per individual, where each chromosome is composed of between 40 and 250 million base pairs. The reason that it is labeled as a base *pair* is that DNA is naturally double-stranded. For a first order approximation (i.e. ignoring events such as DNA methylations), we can treat each base (per strand) as having one of four possible values: adenine (A), cytosine (C), guanine (G), and thymine (T). From the point

of view of information content, the two strands are redundant — an A on one strand almost always parallels a T on the other, while a C on one strand almost always parallels a G on the other. We can thus treat each copy of the data as though it were single-stranded, and represent an individual's DNA as 23 pairs of "words", each composed of the same set of 4 distinct letters. One copy of the 23 chromosomes is known as a *haplotype* of the individual, while both sets are referred to as a *genotype*.

We can line up the haplotypes of all individuals in a population and compare them to an established, "reference" sequence, which represents an exemplar DNA sequence of a human being. Much ($> 95\%$) of the sequence is identical to the reference across the entire population. Any differences in a haplotype compared to the reference occur as a result of a *mutation*, which adds, deletes, rearranges, or changes one or more consecutive bases in an individual. As the population shares common ancestry, ancestral mutations are often shared among current haplotypes. Practically, the sequencing data we get is from a jumble of both copies of an individual. While high sequencing depths are sufficient to detect most mutations, it is nontrivial to determine the *phase* of a number of mutations — in other words, which mutations belongs to which copy. As a result, we work with genotypes instead of haplotypes.

For a genetic study such as this, the question we are interested in can be distilled down to identifying why a subset of individuals (say, the well-adapted mountain dwellers) is different from the others. Since the underlying question is one of differences, we can safely ignore the sites that are invariant in our populations and identical to the reference. The result is what is usually referred to as a *SNP matrix*. This matrix (represented as S below) has dimensions corresponding to m individuals and n variant sites (in other words, mutation events across the population), and has elements defined as follows:

$$S_{i,j} = \text{the number of occurrences of mutation } j \text{ in individual } i$$

As humans are diploid, this means that the possible values of $S_{i,j}$ can be 0, 1, or 2. A column j in this matrix represents the values of the j^{th} SNP across the population. A row i in this matrix represents the set of variations belonging

to the i^{th} individual and thus, the genotype of that individual.

The basic approach of virtually any analysis in the realm of population genetics is one of *association* — that is, to look for an overrepresentation of one allele in mal-adapted individuals (*cases*) relative to well-adapted individuals (*controls*). In other words, it can be formulated as follows:

Input:

-A SNP matrix S (as defined above)

-A length m phenotype vector y, where:

$$y_i = \begin{cases} 1 \text{ if row } i \text{ corresponds to a case individual in } S \\ 0 \text{ if row } i \text{ corresponds to a control individual in } S \end{cases}$$

Output:

A single SNP j with the most skew in cases relative to controls

Objective Function:

argmax_j $(\chi^2_{j,y})$, calculated from the 3×2 contingency table formed by j and y

For the objective function, we note that there are several tests that can be utilized to capture these skews (for instance, Fisher's exact test), and the choice of the χ^2 statistic is arbitrary. This objective can be trivially determined in time linear to the size of S, and is commonly implemented in association studies (known as *GWAS*). However, there are a few assumptions that this formulation makes. For instance, this assumes that the "causal" loci correlate individually with the phenotype. In addition, there is an implicit assumption that the same "causal" variant is shared among the cases. Finally, the approach requires thousands of cases and controls in order to achieve the requisite P-value significance levels for a whole genome scan.

In many studies, these assumptions either do not hold or are not optimal for the dataset, and new methods must be devised. The work of this dissertation helps address this. Interestingly, this shows the importance of the underlying assumptions — each chapter attempts to solve the same biological question (the genetic basis of hypoxia tolerance).

1.3 Specific goals of the dissertation

Chapter 2 attempts to identify associations missed by a single-locus sweep. For instance, it is possible that two distal loci interact to create a phenotype. Consider an situation in which two loci j_1 and j_2 jointly determine a binary phenotype, in a "lock-and-key" mechanism. In other words, if both j_1 and j_2 both show the reference allele or both show the derived allele, the individual is considered as control. However, if either j_1 or j_2 (but not both) show the derived allele, the individual is considered as case. In this situation, single-locus tests may fail, as the determinant of the phenotype is the interaction between j_1 and j_2.

As a result, we must correlate pairs of SNPs to the phenotype. A brute force solution along the same lines as the one above is quadratic in the number of SNPs and, for a genome-wide scan with $O(10^6)$ SNPs, is thus infeasible. As a result, randomized algorithms such as RAPID [4] and SixPAC [5] have been developed to approximately solve this problem in subquadratic time. However, both of these tools require a binary phenotype to work on. For a quantitative phenotype, such as survival rate under various hypoxic stress, this may not easily present itself. Chapter 2 attempts to convert quantitative phenotypes into binary classes for these tools to use, and then explores their ability to capture a variety of interactions.

The rest of this dissertation looks at changes in the formulation due to experimental design. Specifically, putting a population of animals under generations of stress (such as hypoxia) leads to a genetic adaptation (natural selection) to cope with this stress. Genetically, this leads to systematic changes in a region which provide more information than simple, single-locus correlations. Chapter 3 describes an empirical framework to evaluate various tests that capture regions under natural selection. This is then used to characterize the regimes in which each test does best. Chapter 4 uses the test deemed most appropriate from the previous chapter to identify regions under known, strong selection (caused by 4% hypoxia) in flies. Chapters 5 and 6 then use a battery of tests, plus some custom prioritization pipelines, to identify regions under unknown selection in humans.

Chapter 2

RAPID detection of gene-gene interactions underlying quantitative traits

2.1 Abstract

In complex phenotypes, the trait under question may not be explained adequately by single locus effects — multiple loci might interact to confer the phenotype. For instance, two proteins may interact in a lock-and-key mechanism, and a mutation in one protein can be offset by a complementary mutation in the other. In this situation, a specific genotype value for either individual SNP does not matter as much as the relationship between the two genotypes. The problem with detecting these interactions is primarily computational – a naive way takes $O(nm^2)$, where n is the number of individuals and m is the number of variants. This is infeasible for a typical human GWAS, where $n \approx 10^3$ and $m \approx 10^6$. Randomized algorithms (RAPID [4] and SixPAC [5]) have been built which improve upon this significantly. They take advantage of the fact that we can represent genotypes as a series of binary high-dimensional vectors (where each dimension represents the allelic value of an individual). Given this framework, correlated loci are proximal in this vector space, and the problem reduces to finding such adjacent pairs efficiently.

They both use a variant of a geometric approach (locality-sensitive hashing, or LSH [6]) to solve this without explicitly checking every pair of loci.

Here we extend this approach to handle quantitative phenotypes, as opposed to the standard "case" and "control" delineations. The correlation between quantitative phenotypes and multilocus genotypes is usually performed using ANOVA statistics that do not lend themselves to geometric/discrete approaches. To get around this, we follow a two-stage strategy: first, we describe an algorithm to partition the continuous phenotypes into two sub-populations under the assumption that the phenotypes are a mixture of Gaussians. Second, we use the LSH frameworks implemented in RAPID and SixPAC to identify epistatic loci within each sub-population. We use simulated datasets to determine the best way to partition these points given different levels of mixture between the two phenotypes. An application of this algorithm to a number of isogenic Drosophila lines tested for hypoxia tolerance show the speed and power of our approach.

2.2 Introduction

The simple approach described in Chapter 1 for correlating genotype and phenotype may not work for many reasons. For instance, let us consider a phenotype caused by an interaction of two proteins. If a mutation occurs in a locus in one protein that destroys the interaction, a compensatory mutation can occur in the other to reestablish this. If only one of the two is mutated, the phenotype is not present. The underlying computational problem is also one of correlation — one way to frame this is that within the case (and/or control) population, two SNPs show a much higher level of correlation with each other than expected. In other words, it can be formulated as follows:

Input:
-A SNP matrix S (simplified as binary haplotypes)

-A length m phenotype vector y, where, where:

$$y_i = \begin{cases} 1 \text{ if row } i \text{ corresponds to a case individual in } S \\ 0 \text{ if row } i \text{ corresponds to a control individual in } S \end{cases}$$

Output:

A pair of SNPs j_1, j_2 with the most skew in cases relative to controls

Objective Function:

argmax_{j_1,j_2} $(\chi^2_{j_1,j_2,y})$, calculated from the $3 \times 3 \times 2$ contingency table formed by j_1, j_2, and y

Similarly to before, several objectives can similarly be used. A brute force approach (calculating the χ^2 tables for all pairs of SNPs) can also be implemented for this problem. For a typical genome-wide study, however, there can be on the order of 10^3 individuals in both cases and controls and 10^6 loci. To exhaustively search through all pairs of loci for interacting pairs can take $O(nm^2) = O(10^{15})$ computations, which may not be feasible. Randomized algorithms exist which can identify such a pair in much shorter time. Almost all of these follow a two-step procedure: quickly filtering the full set of SNPs into a smaller candidate subset of SNPs which can be interacting, followed by performing brute force calculations on this subset. The ones described below all identify candidate interacting pairs as being correlated in one population only (say, the cases), then calculate the appropriate two-population statistic on the resulting set. For exposition's sake, the algorithms below are described using haplotype data, although genotype analogs are available.

For instance, members of our lab developed a tool called RAPID [4]. RAPID works by transforming all SNPs into points on an n-dimensional hypersphere (essentially, by converting each dimension into its z-score). Through this conversion, the problem becomes one of finding nearby points in this geometric space quickly. For the purposes of exposition, let us say that SNPs j_1 and j_2 are "interacting", and thus, adjacent. RAPID attempts to find this pair by using the concept of locality sensitive hashing (LSH) [6]. LSH works on the principle that, when projected onto random vectors, nearby points should have similar values (or in this

situation, that the cosine between a random vector and j_1 should be similar to the cosine between the random vector and j_2). We can capture this proximity through binning these projections into a small number of discrete bins, as with a high probability, the projection of both j_1 and j_2 with a random vector will fall into the same bin. Of course, with just one random vector, many unrelated SNPs will also fall into this bin. As a result, RAPID projects onto multiple random vectors. The bin size and the number of random vectors needed to distinguish an interacting pair from a random pair can be optimized based on the expected "distance" between interacting SNPs on the hypersphere. On average, the number of calculations needed is $O(nm^{1.07} \cdot \ln(1/\epsilon))$, a significant savings over the naive approach.

A similar approach was implemented by Prabhu et al. [5] in a tool called SixPAC, but with some simplifying assumptions that sped up searches. It works under the assumption that the phenotype is rare and caused by the interaction between the minor allele of two SNPs (in other words, having both minor alleles leads to the phenotype, while having any other combination does not lead to the phenotype). This problem thus becomes finding instances where the minor alleles of two sites occur quite often together in the case population, which is an instance of the "light bulb" problem [7]. The approach starts by transforming each column (SNP) in a SNP matrix such that the label "1" refers not to the derived allele, but to the minor allele. The solution implemented is to simply take random subsets of the case individuals (rows), and to look for all SNPs where the subset contains only minor alleles. A logical AND operation can be used to quickly find these SNPs. If more than one SNP fits this criterion, then they are identical at these sites, and thus, candidates for being interacting. This approach can be considered as a special case of RAPID, with the analog of the "random" vector being a vector parallel to one of the axes (representing one dimension, or one case individual).

Both of these approaches are significantly faster than a brute force approach. However, the underlying problem behind both of these is to identify correlated sites that lead to qualitative classes of individuals. With many phenotypes (such as tolerance to hypoxia), we may have a wide spectrum of responses (for

instance, survival rates of different populations may vary from 0% to 100%). The focus of this chapter is to utilize the above methods on quantitative data. In other words, the problem can be defined as follows:

Input:
-A SNP matrix S
-A length m phenotype vector y, where $y_i \in \mathbb{R}$
Output:
A pair of SNPs j_1, j_2 with the most skew in cases relative to controls
Objective Function:
argmax_{j_1,j_2} $(P_{j_1,j_2,y})$, calculated from an ANOVA statistic by j_1, j_2, and y

Once again, other approaches (e.g. regression) can be used for the objective function. Our goal is to convert the quantitative phenotypes to qualitative statuses. We do this by treating the phenotypes as coming from a mixture of Gaussians. If we can run a preprocessing step to identify the parameters associated with the Gaussians (the means (μ_i) and standard deviations (σ_i) for all classes i), we can determine class memberships for each individual and run the LSH algorithms as before.

2.3 Materials and Methods

2.3.1 Simulated dataset

In order to test the performance of our approach, we built a simulated dataset, composed of $n = 10000$ sites and $m = 2000$ individuals. Out of these individuals, 1000 come from a "case" population, while the remainder come from a "control" population. Once again, we work with haplotype data for simplicity's sake, and each value in the SNP matrix is chosen completely at random (50% probability of either 0 or 1). However, one pair of SNPs is chosen to be completely identical in the case population only. The phenotype vector is chosen such that the case population has mean $\mu = 0$, while the control population has mean $\mu = 1$.

The standard deviations of both populations are identical, and vary from $\sigma = 0.1$ to $\sigma = 4$, in increments of 0.1. Thus, on the lower end of the standard deviation spectrum, the two classes are perfectly separable using a threshold of approximately 1.5, while at the higher end, the two classes are largely intermixed. We note that many aspects of this setup do not generalize to any dataset — for instance, many phenotypes will be present at much less than 50% frequency in a sampled population. This study is meant to be a preliminary step towards solving this, discussing, for instance, minimum requirements in order to detect an association. We simulated 20,000 trials per standard deviation in order to understand the best approach for our preprocessing step. In the interest of time, when applying the interacting site algorithms (RAPID and SixPAC), we increment the σ values by 0.3 and run only 100 trials per standard deviation.

2.3.2 Uncoupling the mixture of Gaussians

We use a classic approach known as expectation maximization (EM) to uncouple the mixture of Gaussians. Firstly, we assume that the phenotype data comes from two (one-dimensional) Gaussians. We provide an initial guess for the means of these Gaussians at the lower and upper terciles, with equivalent standard deviations and mixing proportions. From there, this is a two-stage process. First, for each point, we determine the likelihood that it comes from each of the two distributions given their parameters, and normalize these probabilities such that they sum to 1. Second, given these probabilities for each point, we re-estimate the underlying parameters of the Gaussians (means, standard deviations, and mixing proportions). We then iteratively repeat these steps with the new parameters, until convergence. We utilized the R package mclust [8] to calculate these parameters.

2.4 Results

2.4.1 Determination of optimal EM parameters

We first attempted to evaluate and optimize how well the EM approach performed on classifying the mixture, using the χ^2 statistic as an objective function. The output of the EM algorithm in this setup can be boiled down to a probability matrix, P, where $P_{i,j}$ represents the probability that individual i belongs to class j. We need to partition each individual into the discrete classes in order to run any of the above algorithms, and a logical approach would be to take the majority class for each individual. However, it is possible that the more marginal points (where the maximum probability is close to 50%) contain too many misclassified points, and thus, hurt the χ^2 statistic. We can thus fix a confidence threshold between 0.5 and 1, and toss out all individuals that have confidence levels below this threshold. If the threshold is too high, the gain in correctly classified points is more than offset by a drop in sample size, and a higher P-value results.

Trials identifying a "sweet spot" for thresholding are described in Figure 2.1. As mentioned above, for a given standard deviation (column), we run 20,000 trials, each of which tests all confidence thresholds from 0.5 to 1 (in intervals of 0.01). In each trial, we focused only on the interacting pair of loci. For these two sites, we constructed a 2×2 contingency table from all individuals deemed to be at a high enough confidence level. We then calculated a P-value (defined by a χ^2 statistic). As the figure shows, we generally have very good power (fraction of sites with P-value < 0.05), even with a high standard deviation. Including all individuals in their most likely class (i.e. setting the threshold at 0.5) is suboptimal relative to setting the threshold higher (approximately 0.6-0.65, for instance). As expected, though, the more intermixed the two classes are, the more misclassified points there are. As a result, we need to increase the filtration threshold as a function of standard deviation. Since we do not know the standard deviation *a priori*, we pick a threshold at the higher end of the spectrum (0.65) for future tests. Given this threshold, the expectation of correctly and incorrectly classified points is shown in Figure 2.2.

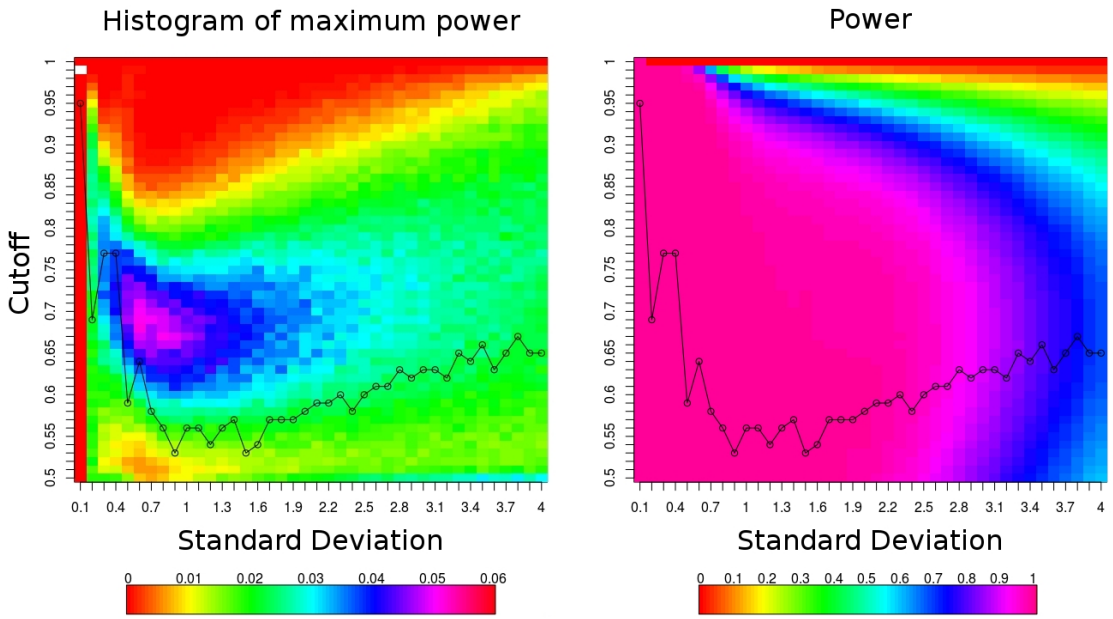


Figure 2.1: **The impact of the EM confidence threshold on the ability to capture an interacting pair of SNPs.** For a given standard deviation (columns in both plots), we attempted to see how important the point classification threshold (rows in both plots) matters in identifying an interacting pair of SNPs. In both plots, the black line represents the threshold that returns the minimum average P-value. Left: Each column of the underlying heatmap represents a histogram of the threshold which returns the lowest P-value. Right: The power (as defined by the fraction of points with P-value < 0.05) is shown for each standard deviation/threshold pair. As can be seen, the power remains very high in this scenario, even as the standard deviation increases. At low standard deviations (< 1), the classes are separable enough that virtually all thresholds will work. At higher standard deviations, though, we need to filter out more and more borderline cases. Also of note, regarding the P-value, the mode low threshold is generally higher than the mean low threshold. This is because, when the EM algorithm does not do a good job at reconstructing the original distributions, setting the thresholds too high eliminates many true candidates as well, leading to a high P-value (data not shown).

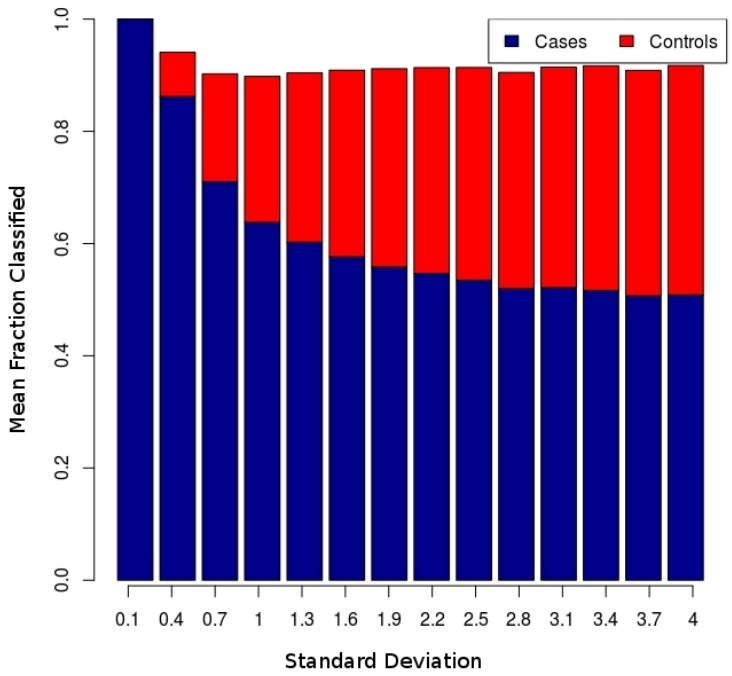


Figure 2.2: **The expectation of correctly classified points given a 0.65 threshold.** As expected, the higher the standard deviation, the lower the amount of correctly classified individuals. However, we still have reasonable signal which can be detected (50% of the cases vs. 40% of the controls even in trials with standard deviation 4).

Given this threshold, we now have an input SNP matrix (with qualitative phenotypes) for each of the approaches mentioned above.

2.4.2 The impact of quantitative phenotypes on the randomized algorithms

The major negative consequence of any approach converting quantitative values into qualitative classes is that we can misclassify points, weakening any signal that may be present. To explain this, let us take a pair of sites in which we have perfect correlation between two sites in our case dataset and perfect anti-correlation in our control dataset. In this setup, the amount of misclassified individuals directly relates to the correlation between the two sites.

RAPID quantifies this using what the authors refer to as θ, the distance between two correlated points once projected onto a unit hypersphere. Two perfectly correlated points have a θ value of 0, while two random points have a θ value of $\sqrt{2}$. As the fraction of misclassified points increases, θ increases as well, approaching the value of a random pair of points. In order to accommodate the rise in θ, we need to perform more random projections to separate truly interacting pairs from false positives. If this distance is too high ($\theta > 0.6$), then the amount of calculations required exceeds the amount needed in a brute force operation. Figure 2.3 shows that this θ distance corresponds to a standard deviation of approximately 0.5. Although SixPAC does not model θ explicitly as an input parameter, it has generally similar behavior.

As a standard deviation of 0.5 is relatively small, we choose a different approach: to set the parameters (θ for RAPID, significance value for SixPAC) as constant across all our simulations. We selected roughly comparable values for both statistics — the default Bonferroni-corrected P-value of 0.05 for SixPAC, and a θ of 0.2 for RAPID. This results in a speedup relative to a brute force approach under all instances of the input dataset. However, this increases the false negatives for all datasets generated with a higher than expected standard deviation. As can be seen in Figure 2.4, both perform similarly well. RAPID takes approximately three to four times as long (mean time of 466 seconds vs. 136 for SixPAC). Part

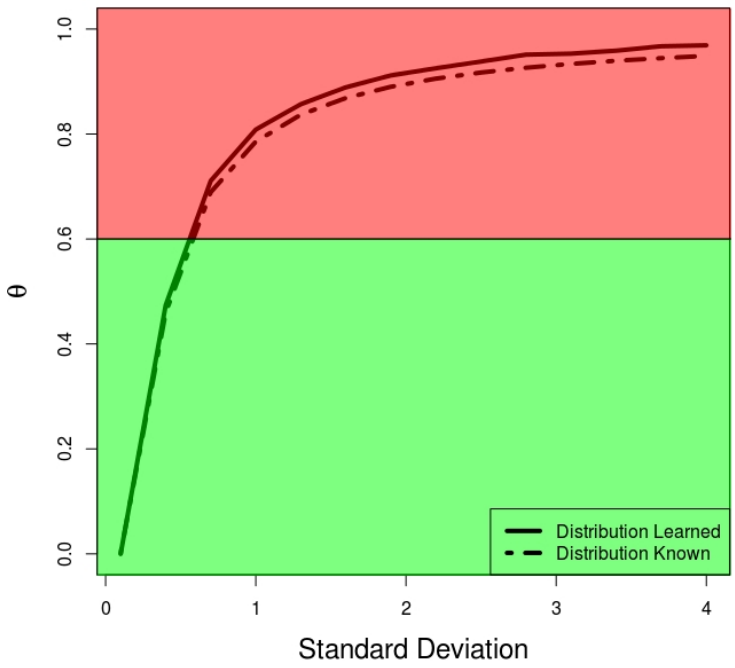


Figure 2.3: **The impact of standard deviation on RAPID's performance.** Firstly, we note that using the EM approach to define classes (solid line) returns results that are relatively similar to having prior knowledge of the classes themselves. Searching for θ values above 0.6 using RAPID yields more calculations than a brute force approach. With our approach, θ increases to greater than 0.6 when the standard deviation of both classes is greater than 0.5.

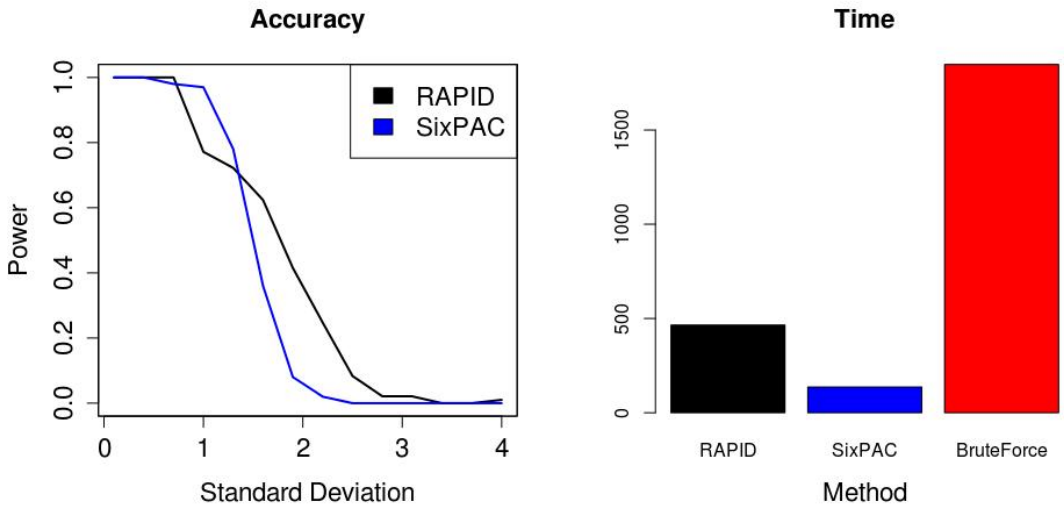


Figure 2.4: **The performance of RAPID and SixPAC on simulated datasets.** As expected, both tools perform similarly well in the regimes where the classes are easily separatble ($\theta < 0.5$). Above this, the power starts to decay, as the estimated distances between the two sites exceeds the θ (or significance) parameter. Both are significantly faster than a brute force method, although the assumptions that SixPAC makes allows it to be approximately 3× faster than RAPID.

of this is because RAPID tests both cases and controls (as opposed to just cases), while another part is because RAPID tests skews in both directions (abundance of "1-1/0-0" and "1-0/0-1" pairs). However, we note that SixPAC has the potential to be much faster than represented here due to their implementation of multi-threading.

2.4.3 Application to Drosophila dataset

We now apply our algorithm to a Drosophila dataset. This dataset is composed of 136 isogenic lines from the Drosophila Genetic Reference Panel (DGRP) [9], each of which is put independently in 5% oxygen conditions. Each line has been sequenced, and approximately 2.6 million variant sites have been discovered across all lines. The ability of the fly to adapt is measured using eclosion rate, the percentage of pupae that emerge as adult flies, and is thus a real number between 0% and 100%. We note that this sample size is much smaller than a typical

GWAS study (which numbers in the thousands), and thus, we do not expect to find variants at a Bonferroni-corrected P-value (approximately 10^{-8}). For single locus tests, we thus use the protocol described in the DGRP paper [9], which identifies all sites present at a much lower significance ($\leq 10^{-5}$) as being worthy of further investigation.

Before we look for multilocus interactions, we first attempt to identify possible (single) loci that directly correlate to the phenotype. This could yield a simpler mechanism for linking genotype and phenotype. In addition, as described in Brinza et al. [4], it is possible for one or more of these singly-associated loci to be involved in multilocus interactions. We determine our test statistic using linear regression of the genotype to the quantitative phenotype, as implemented by Plink [10]. In order to correct for outliers caused by low variant frequencies, we permute the phenotype labels either until we get a P-value that we can confidently determine is non-zero, or until we have permuted the data 10,000 times. The result is shown in Figure 2.5. As expected, no site even comes close to the Bonferroni P-value. The list of sites less than the threshold are described in Table 2.1. The top hit is located in a gene called *Rala*, part of the Ras family. This pathway has been linked to many stress responses, including hypoxia [11], and is involved with regulation of many important cellular pathways, such as the Notch pathway (see Chapter 4 for the link of Notch to hypoxia).

For multilocus interactions, we choose to run SixPAC over RAPID due to its speed, particularly after multi-threading. Our GMM approach separates the set of individuals into 39 "cases" which survive well under hypoxia, 92 "controls" which do not survive well under hypoxia, and 5 individuals which are inconclusive, and thus, tossed out. A Bonferroni-corrected P-value is even more infeasible when dealing with pairs of loci — the significance threshold drops to 10^{-14}. We thus arbitrarily set a cut-off of 10^{-7} to identify potentially interacting pairs. The resulting pairs of SNPs are shown in Table 2.2. As the threshold is low enough for many of the P-values to occur by chance, we expect many false positives. However, there are some interesting candidate pairs. For instance, there is a correlated pair of SNPs in the coding regions of *whd* and *CG17219*. *whd* is a gene that has

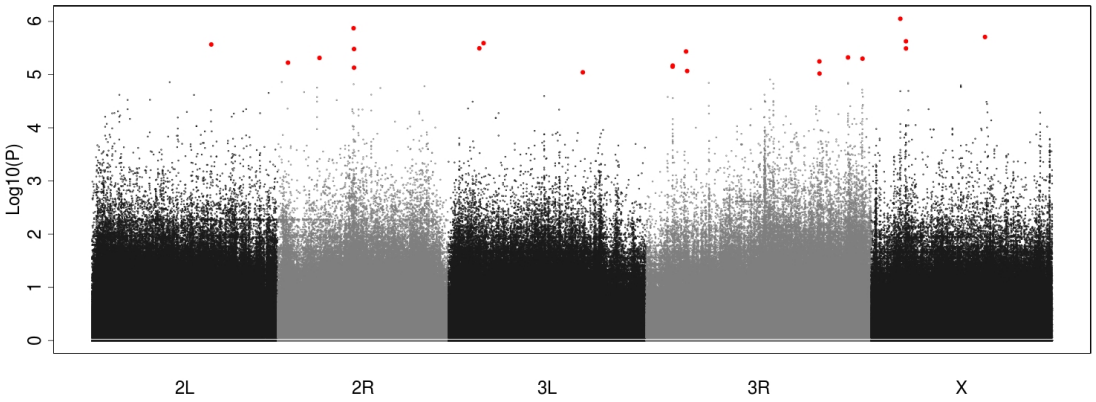


Figure 2.5: **Manhattan plot showing significance of single-locus interactions with the hypoxia phenotype.** The 21 points in red have significance value less than 10^{-5}, and are shown in Table 2.1. Although some points are co-located, they are generally spread out across all five chromosome arms.

been implicated in oxidative stress tolerance [12]. In contrast, not much is known about *CG17219*, but one of the few manuscripts linked to the gene at the time of analysis shows that *CG17219* is significantly upregulated ($\sim$6$\times$) during blood cell proliferation in flies [13]. Thus, it is conceivable that these two genes can be involved in a fly's response to a hypoxic environment.

2.5 Discussion

The major emphasis of this paper is exploring how RAPID and SixPAC deal with noise in phenotype values. This noise manifests itself as misclassified individuals, which results in weaker correlations. As expected, datasets where the case and control populations are easily separable are trivial for both of these tools. However, even if there is a mild level of mixing (particularly for RAPID), we can no longer theoretically guarantee that we will find the correlation quickly (as much as any randomized algorithm can guarantee events). That said, these approaches do hash more distal, but correlated points together quite often, as evidenced in Figure 2.4.

This is a preliminary step towards applying fast approaches that identify

Table 2.1: **List of single loci associated with the hypoxia phenotype, sorted by P-value**

Chr.	Position	P-value	Nearest Gene
X	3605554	8.90E-007	*Rala* (intron)
2R	9420056	1.34E-006	*Vmat* (intron)
X	14101925	1.96E-006	*l(1)G0007* (intron), *l(1)G0469* (intron)
X	4302488	2.36E-006	*CG32773* (intron)
3L	4384686	2.56E-006	*slow* (5' UTR)
2L	14789315	2.71E-006	*CG18636* (5k upstream)
3L	3861646	3.20E-006	*Awh* (3' UTR)
X	4305555	3.22E-006	*CG32773* (intron)
2R	9476269	3.30E-006	*CG6191* (intron)
3R	4945662	3.67E-006	*pum* (intron)
3R	25021192	4.75E-006	*CG14509* (500bp upstream)
2R	5194479	4.85E-006	*Myd88* (intron)
3R	26817207	5.01E-006	*5-HT7* (intron)
3R	21471993	5.65E-006	*jigr1* (intron)
2R	1289386	5.96E-006	*laccase2* (12k upstream)
3R	3294726	6.77E-006	*tRNA:V3b:84Dc* (3k downstream)
3R	3278714	7.08E-006	*tRNA:V3b:84Da* (1k downstream)
2R	9476357	7.41E-006	*CG6191* (intron)
3R	5079269	8.55E-006	*CR16735* (pseudogene)
3L	16685688	9.04E-006	*Lasp* (intron), *CG43954* (intron)
3R	21488562	9.55E-006	*RASSF8* (syn. coding)

multilocus interactions on quantitative phenotypes. While it does show some promise on simulated datasets, it is important to note that real-world datasets do not generally conform to the simulation parameters tested here. For instance, in many disease scenarios, the dataset is composed of far fewer cases than controls. There is also no guarantee that a quantitative dataset is composed of exactly two classes, that both of them are Gaussians, that both Gaussians have equal variance, and that the interaction model is as straightforward as presented (see Marchini et al. [14] for a more comprehensive list of two-locus interactions. In some cases, we can account for this. For instance, if we know *a priori* that the classes come from a mixture of exponentially distributed variables, we can adjust the EM approach accordingly. In addition, we can use the output of the EM algorithm to guide our parameter settings. As an example, let us consider a situation where our sample size is large enough and we have many more controls than cases. In this setting,

Table 2.2: **List of pairs of loci jointly associated with survival rates under hypoxic conditions, sorted by P-value.** "Syn" refers to synonymous SNPs, while "up" and "down" are upstream and downstream of the target gene, respectively.

SNP 1	Gene 1	SNP 2	Gene 2	P
2L:10329047	CG5375 (syn)	2L:15284676	CG15263 (500bp up)	3e-9
2L:10329050	CG5375 (syn)	2L:15284676	CG15263 (500bp up)	3e-9
2L:10329045	CG5375 (L426R)	2L:15284676	CG15263 (500bp up)	3e-9
3R:9650579	Mst87F (2kb down)	3R:14061566	repo (300bp up)	1e-8
3L:10090664	dpr6 (intron)	3R:26958719	CR43458 (ncRNA)	3e-8
3L:10090684	dpr6 (intron)	3R:26958719	CR43458 (ncRNA)	3e-8
2L:17267450	beat-IIIc (7k up)	X:15795050	shi (intron)	5e-8
2L:3029533	CG17219 (syn)	2R:6358197	whd (syn)	8e-8
2L:11690459	CG4988 (5k up)	2L:17159035	beat-IIIa (intron)	9e-8

we can afford to be more selective with our control population (i.e. set a higher confidence threshold) than with our case population.

2.6 Acknowledgments

We would like to thank Glenn Tesler and Matthew Schultz for constructive feedback on the RAPID code, as well as Dan Zhou and Gabriel G. Haddad for providing the Drosophila phenotype data used in this chapter.

Chapter 2 is unpublished work. The dissertation author was the primary author of this paper responsible for the research.

Chapter 3

Exploring tests of selection in case-control data

3.1 Abstract

Because our sample size from the previous section was too small to identify linked sites with any level of significance, we now focus on a different, but related problem, that does not require as many individuals to be sequenced. This problem is one of detecting natural selection. In this approach, a population is put under a stress (such as hypoxia) for a number of generations. This poses a different set of challenges from a GWAS-style approach. In the previous framework, we do not know if each fly line adapts using a common or distinct mechanism (i.e. a set of rare variations or a few common mutations). Under natural selection, the underlying selective pressure leads to the proliferation of beneficial mutations over time (see below for more information). In other words, the beneficial mutations should rise in frequency, and we are necessarily looking for common mutations within a population.

In addition to natural selection, for smaller organisms with faster breeding cycles (such as flies), artificial (or laboratory) selection can be used to create sub-populations with different phenotypic traits. Genetic tests can be employed to identify the causal markers for the phenotypes, as a precursor to engineering strains

with a combination of traits. Traditional approaches involve analyzing crosses of inbred strains to test for co-segregation with genetic markers. In this chapter, we take advantage of cheaper next generation sequencing techniques to identify genetic signatures of adaptation to the selection constraints. While we develop tests that can be applicable to individual sequences, it is important to note that such data is often suboptimal (power-wise) for a population of flies. There are two reasons for this. First, in order to generate sufficient sample to sequence a fly, the most common approach is to interbreed siblings for a number of generations, eventually creating a completely homozygous (isogenic) line. This process is time-consuming. Even if the process is trivial, though, pooled population-level sequencing gives much more insight per unit coverage on processes like selection than individual data (see below for reasons why). The main focus of this chapter is thus on pooled genomic data.

There are two major issues with testing for either natural or laboratory selection. Firstly, not every phenotype is linked to a selection pressure. For instance, a variant influencing eye color does not confer any obvious selective advantages, leading to a situation similar to the previous chapter. Even if a trait is linked to a selection pressure (particularly with the weaker ones often found in nature), several sites may segregate similarly, due to genetic drift. As a result, the standard GWAS-style correlations may not be sufficient to distinguish beneficial loci, and more directed tests must be implemented. Chapters 3-6 attempt to address this.

Specifically, in this chapter, we explore a series of statistical tests for selection using pooled case (under selection) and control populations. The tests generally capture skews in the scaled frequency spectrum of alleles in a region, which are indicative of a selective sweep. Extensive simulations are used to show that these approaches work well for a wide range of population divergence times and strong selective pressures. Control vs. control simulations are used to determine an empirical False Positive Rate, and regions under selection are determined using a 1% FPR level. We show that pooling does not have a significant impact on statistical power. The tests are also robust to reasonable variations in several different parameters, including window size, base-calling error rate, and sequencing

coverage. In Chapter 4, we show the vast potential of one of these methods for selection for hypoxia.

Overall, we outline a strategy for finding regions under selection using pooled sequences, then devise optimal tests for that strategy. The approaches show promise for detecting selection, even several generations after fixation of the beneficial allele has occurred.

3.2 Introduction

Laboratory selection methods have been used for centuries to selectively breed organisms for desired phenotypes. The organisms are bred under directional selective pressure to create a stable, adapted population with the desired phenotype. For smaller organisms with faster breeding cycles, this approach can be used to create many populations with different phenotypic traits. Genetic tests can be employed to identify the causal markers for the phenotypes, as a precursor to engineering strains with a combination of traits.

For sexually reproducing organisms, the typical approach entails generating and crossing pure-bred strains from sub-populations with different phenotype levels. Second (and higher) generation crosses can be used to identify markers that co-segregate with the phenotype. The approach is effective especially with a sparse array of genetic markers. However, the generation of crosses is often labor intensive, and the linked regions are large, requiring additional genetic mapping effort to identify the causal variation.

In recent years, deep sequencing technologies have been increasingly available, making whole genome sequencing feasible for small organisms. Even so, given the low quantities of DNA in smaller organisms, it may not be feasible to individually sequence each organism. Even if feasible, cost constraints often lead to a sacrifice in sample size or sequencing coverage, leading to a loss in statistical power. We consider the following experimental approach to identifying the genetic basis of an adapted phenotype: (a) Separate a neutrally-evolving population into two sub-populations, and breed the two sub-populations with (*case*) and without

(*control*) directional selective pressure; (b) sequence large pools of individuals from the two sub-populations; and, (c) identify regions that show a genetic signature of selection relative to the control sub-population. While step (a) is common to any forward genetics approach, steps (b) and (c) do not require labor-intensive crosses. Pooling allows for sequencing to be done in a cost-effective manner. We show below that, under certain regimes, the signal has higher resolution, reducing the additional effort needed to identify candidate genes.

The genetic signature for laboratory selection is similar to that of natural selection. Consider a trait such as hypoxia. The population is bred in an increasingly hypoxic environment, forcing it to gradually adapt. A genetic variant that helps the individual survive will eventually go to fixation. Neighboring SNPs (in LD) also approach fixation, leading to loss of genetic diversity, or a selective sweep, in a region. When multiple loci contribute independently to the adaptation, recombination events bring advantageous alleles together, and the adapted population shows multiple unlinked regions under selection. Various tests of neutrality capture the loss of heterozygosity (as in Tajima [15] or Fay and Wu [16]), exact haplotype frequencies (as in Fu [17]), and other departures from neutral evolution as a test for selection. For recent selection events, the region is characterized by extended haplotypes in high LD with a core set of alleles (see Sabeti et al. [18]). However, all of these approaches are designed to be used with individual haplotypes in the population.

One common approach for testing for causal variations involve analyzing an aggregate of individuals [19, 20, 21, 22]. The feature that many of these algorithms look for is a rise in frequency of a collection of rare alleles in the case population. However, one of the implicit assumptions of these approaches is that the entire population is fully mixing. Under the setup described above, however, the exact opposite situation arises — the populations are completely isolated. While it may be possible to adapt the above statistics after accounting for this substructure, we focus on a class of statistics that do not have any such assumptions.

A formulation for identifying selection is not as clean as it is for association, since the demographics and the strength of selection play a large role in the

underlying signature. We can attempt to formulate the problem as follows:

Input:
-A SNP matrix S, combined from a case and control population (where each variant is linked to its location on the chromosome)
-A length m phenotype vector y, where:

$$y_i = \begin{cases} 1 \text{ if row } i \text{ corresponds to an individual in the case population in } S \\ 0 \text{ if row } i \text{ corresponds to an individual in the control population in } S \end{cases}$$

Output:
A contiguous subset of columns (SNPs) $j_1, ..., j_2$ with the most signs of selection in the case population relative to controls
Objective Function:
Discussed in this chapter

In this chapter, we investigate various tests of selection based not on the departure from neutrality in the case population, but rather, on direct comparisons of allele frequency spectra in the two populations. Many of the tests can be applied to both individual and pooled genomic data, where we only have allele frequencies at each location. By analyzing the scaled allele frequency spectra of case and control populations, we explain how the power of the proposed statistics depends critically on the selection pressure and the time since the selection pressure was implemented. We then identify the relative strengths and weaknesses of each of these tests over a large range of times and pressures. We also investigate the power of the proposed statistics on a number of parameters, including selection pressure, mutation rate, and recombination rate, but also technology-dependent ones like depth of sequencing and base-calling error rate.

We select one approach (the S_f statistic) as being robust to many of the tested parameters (and particularly applicable to strong selection), and apply it to an existing experimental population of *Drosophila melanogaster* that have been adapted to severe hypoxia (Chapter 4). We identify a clear signature of selection that is significant on a genome-wide scale. Our results suggest that in many

experimental populations of interest, direct tests of selection provide an effective alternative to cross-based analyses in identifying the genetic determinant of a phenotype.

3.3 Materials and Methods

Consider a mutation that confers a selective advantage for a specific phenotype (like hypoxia tolerance), and assume it lies in a region with scaled mutation rate $\theta = 4N\mu$. It is important to note that the θ variable defined here (and in subsequent chapters) is completely unrelated to the θ variable used in the RAPID approach from Chapter 2. Here, μ is the mutation rate per base per generation, and N is the effective population size (see, for example, Durrett [23]). Under directional selective pressure for the phenotype, the mutation is driven to fixation. Neighboring (linked) mutations are co-inherited and also go towards fixation, leading to an overall loss of diversity, captured by a lower value of θ. Tests of neutrality often compare two different estimates of θ on the same population that behave differently under departure from neutrality. A significant difference in the two measures is indicative of non-neutral evolution, and possibly, selection. However, the mutation rate μ can vary throughout the genome and might confound estimates of the scaled mutation rate, even with normalization. In addition, population-specific effects (such as a founder population composed of siblings) may lead to false positives.

Instead, consider a case-control scenario in which identical populations are split, and one population is subject to directional selection. It is generally reasonable to assume that the mutation rate μ is identical in the two populations in any specific region. For any measure of θ, the log ratio statistic

$$S(1,2) = \log \frac{\theta_2}{\theta_1} = \log \frac{4\mu N_2}{4\mu N_1} = \log \frac{N_2}{N_1} \tag{3.1}$$

computes the ratio of effective population sizes. A high value of the statistic implies that $N_1 \ll N_2$, or that the region is under a selective sweep in sub-population 1. While we could work with difference estimates, the ratio has a direct interpretation as the relative decrease in population size.

3.3.1 Statistics for detecting selection

The LR-statistic depends upon estimates of θ. Many estimates have been derived and will behave very differently under different regimes of selection. Consider a population sample of size n and assume that an outgroup is known making it possible to distinguish the derived allele. Let ξ_i denote the fraction of sites with exactly i derived alleles. A classical result due to Fu states that, under a neutral model, $\mathrm{E}(\xi_i) = \theta/i$ [24]. Let us define the scaled frequency spectrum as:

$$\hat{\theta}_i = i\xi_i$$

Under neutral evolution, for any i, $\hat{\theta}_i$ is an unbiased estimator of θ. Likewise, for any linear combination:

$$\mathrm{Exp}\left(\frac{1}{\sum_i \omega_i}\sum_i \omega_i\hat{\theta}_i\right) = \theta \tag{3.2}$$

Achaz [25] shows that many of the classical measures of θ are variants of Equation 3.2 with appropriate weight functions ω_i. For instance, we have

$$\begin{array}{lcll}
\theta_{\mathrm{W}} & = & \frac{1}{a_n}\sum_i \frac{1}{i}\hat{\theta}_i & (\omega_i = \frac{1}{i}) \quad [26] \\
\theta_{\pi} & = & \frac{2}{n\cdot(n-1)}\sum_i (n-i)\hat{\theta}_i & (\omega_i = n-i) \quad [15] \\
\theta_{\mathrm{f}} & = & \frac{1}{n}\sum_i \hat{\theta}_i & (\omega_i = 1) \quad [27] \\
\theta_{\mathrm{H}} & = & \frac{2}{n\cdot(n-1)}\sum_i i\hat{\theta}_i & (\omega_i = i) \quad [16]
\end{array}$$

All estimates toss out fixed, derived mutations which are likely to have occurred between the outgroup and the most recent common ancestor of the individuals in the pool. Note that each of these estimators can be derived from the allele frequency spectrum, and therefore, pooled data. A change in $\hat{\theta}_i$ between the case and control population is indicative of selection. We label applications of the θ estimates to the log ratio statistic (Equation 3.1) as S_{W}, S_{π}, S_{f}, and S_{H}, respectively.

Another set of approaches would be based on measuring differences in relative SNP frequencies in the two populations. For instance, Hudson's F_{st} is defined as $1 - \pi_{\mathrm{within}}/\pi_{\mathrm{between}}$ [28]. As our hypothesis is directional (we have defined "case" and "control" populations), we can replace the π_{within} term from this equation with just the heterozygosity from the case population, creating a "directional"

F_{st}. A final approach is based on the principle that selection would lead to a much longer ancestral branch length in the cases, and thus, a significant increase in fixed SNPs in the population. We can construct a 2×2 contingency table composed of counts of fixed and polymorphic sites, in the case and control populations. A one-sided Fisher exact test can be used to test against the null hypothesis of no correlation between the variables. Finally, we can also use one sample tests on the case population, such as Tajima's D [15]. Table 3.1 summarizes all of the statistics used.

Table 3.1: **Notation and definition of the eight statistics tested in this chapter**

List of statistics	
Test	Formula
Tajima's D (one-sample)	$\frac{\theta_\pi - \theta_{\mathrm{W}}}{\sqrt{e_1 \cdot S + e_2 \cdot S \cdot (S-1)}}$
Directional F_{st}	$1 - \frac{\pi_{\mathrm{within,case}}}{\pi_{\mathrm{between}}}$
Hudson's F_{st}	$1 - \frac{\pi_{\mathrm{within}}}{\pi_{\mathrm{between}}}$
Excess of fixed case SNPs	Fisher test
$S_{\mathrm{f}}(1,2)$	$\log \frac{\theta_{\mathrm{L},2}}{\theta_{\mathrm{L},1}}$
$S_{\mathrm{W}}(1,2)$	$\log \frac{\theta_{\mathrm{W},2}}{\theta_{\mathrm{W},1}}$
$S_{\pi}(1,2)$	$\log \frac{\theta_{\pi,2}}{\theta_{\pi,1}}$
$S_{\mathrm{H}}(1,2)$	$\log \frac{\theta_{\mathrm{H},2}}{\theta_{\mathrm{H},1}}$

3.3.2 Forward simulation

We built a simulator that captured aspects of the pooled, diploid *Drosophila* populations described below. There are several different parameters for this simulator.

In a typical laboratory setting, a small group of individuals is used as a founder population. The simulator first has to generate this founder population before the application of selective pressure. In other words, this is a forward simulator that can be thought of as having three stages: 1) generation of a diverse founder population, 2) institution of a population bottleneck (to create the founder individuals) and introduction of the beneficial mutation, and 3) population expansion, followed by application of selection pressure (see Figure 3.1 for an overview). Within each stage, a Wright-Fisher model with defined population size and time (in generations) is used. Parameters that are characteristic of the species, such as the per-base mutation and recombination rates, are set to be identical across all stages. With the *Drosophila* studies and most of the simulations, the per-base mutation rate ($8.5 \cdot 10^{-4}$/bp) was taken from Watterson's estimator applied to the C1 population, while the per-base recombination rate was taken from Fiston-Lavier et al. [29] to be $1.892 \cdot 10^{-8}$/bp.

Since we do not know the genotypes or the relative differences between the individuals that spawned the biological populations, we simulate differences by generating a known "reference" haplotype and setting 2000 individuals to have two copies of this haplotype. As per coalescent theory, a neutrally evolving population of size 2000 will have a shared common ancestor after approximately 4000 generations [30]. There is some variance in this number, so to ensure that this criterion is met, we ran the simulator for 14000 generations. The specific choice of 14000 was somewhat arbitrary — it is long enough away from 4000 to ensure that the two populations would share a common ancestor, but short enough to make thousands of trials feasible. As a result, this process generates a population based on a known reference, but each individual is sufficiently different from the reference as well as other individuals in the population.

After this, a 54-haplotype founder population is taken from the pool of genotypes. The beneficial mutation is introduced into exactly one of the founders, and the population is then immediately separated into two sets of 2000 individuals derived from these founders. At this stage, the selective pressure is applied to one of the sets, and the populations are allowed to evolve independently at

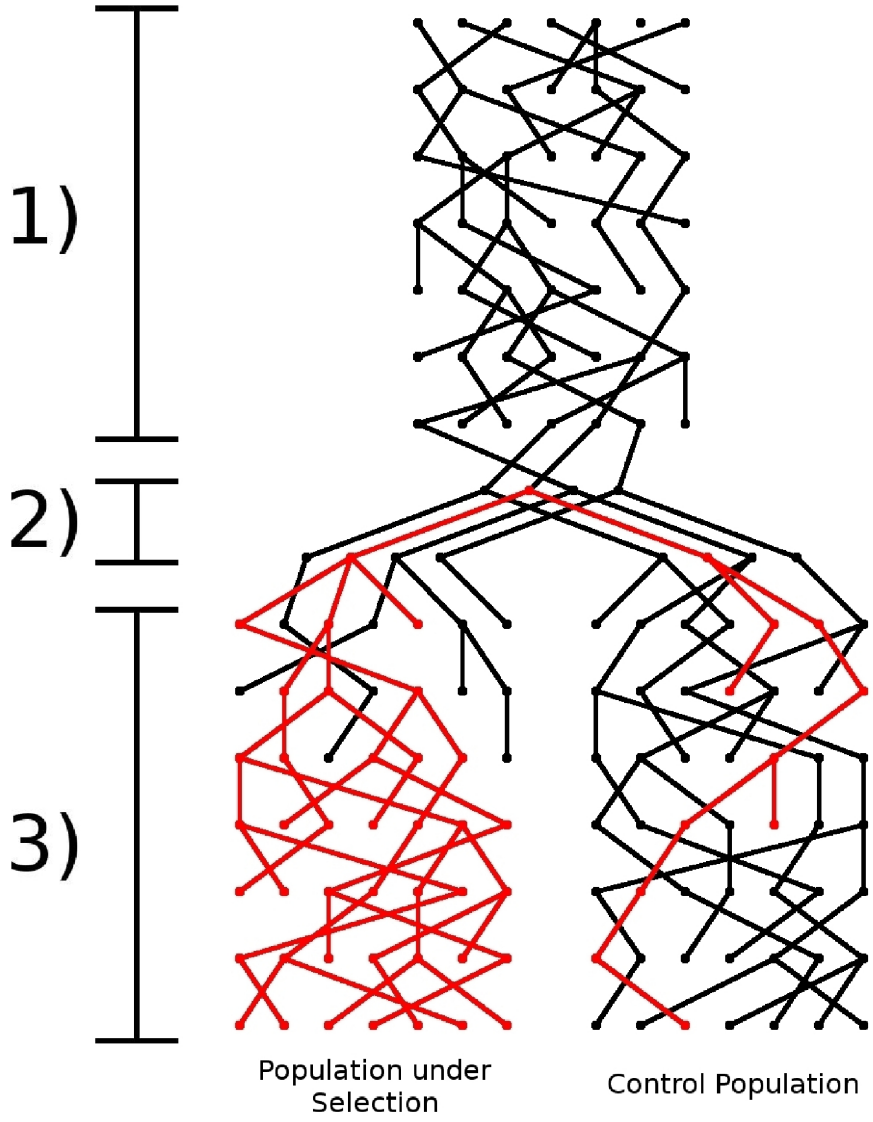


Figure 3.1: **Schematic of simulator.** There are three stages to this simulator: 1) Generation of founder population, 2) Shrinking of population to "founders", introduction of the beneficial mutation (represented by the red lineage), and separation into subpopulations, and 3) Application of selection pressure in one subpopulation and expansion.

constant size, usually for 200 generations. As far as measuring signal, we need to capture regions large enough to accurately estimate the true tree topology, yet small enough such that the signal is not masked by recombinations. Fixed window sizes (generally 50kbp) are used for θ calculations, and the beneficial mutation is located exactly in the middle of these windows. The beneficial mutation is defined by two parameters: selection coefficient, s; and degree of dominance, h. The relative fitness of homozygous wild-type individuals is 1, heterozygous individuals is $1 + hs$, and homozygous mutant individuals is $1 + s$. s is generally variable, but h is fixed to be 0.5 in all trials. Parameters involved in the sequencing stage include sequencing sample size (generally 200 individuals), sequencing coverage (generally 70×), and base-calling error rates (generally 1%). The default values are intended to be representative of typical experimental conditions, and as such, are derived from the hypoxia dataset (see Methods). Each trial is repeated 500 times to get a more complete picture of the behavior of the statistics.

3.3.3 Preprocessing

Assuming even a 1% sequencing error rate, it is unlikely that we would be able to distinguish low frequency variants from sequencing errors at reasonable levels of coverage. Since under the null model, the SNP frequencies would follow a similar distribution, we tossed out all SNPs with a minor allele frequency of less than 10% from consideration. In addition, this would lead to a potential gain of power, as many of the *de novo* mutations would initially have frequencies less than 10% (see Figure 3.2B). At least 10× read depth at a site was required to have enough reasonable confidence in its frequency. In addition, when the selective pressures are high, it is common to see no SNPs with minor allelic frequencies $\geq 10\%$, leading to θ estimates of 0. In order to take the log ratio, pseudocounts of 0.1 were added to both numerator and denominator. We note that this is not a problem with individual genotype data.

With the application to *Drosophila* genotypes, it is possible that the phenotype impacts the mutation rate. For instance, it has been shown that temperature impacts mutation rate in *Drosophila* [31]. Particularly in the regime well after fixa-

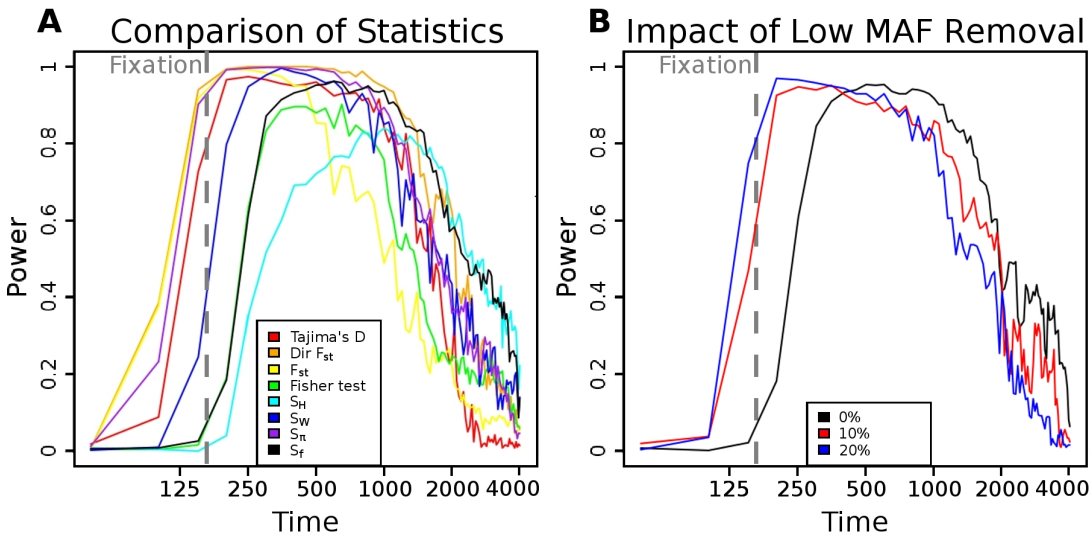


Figure 3.2: **Power of different statistics as a function of time since bottleneck.** A) Power versus time plots for multiple statistics given a fixed selection pressure (s = 0.08). Statistics that weigh low frequency SNPs more, such as S_{W}, have a higher peak power, but decay faster. On the other hand, statistics that weigh high frequency SNPs more, such as S_{H}, have a lower peak power, but retain power for a longer timespan. S_{f} (black, equal weights) and directional F_{st} (orange) both provide a reasonable compromise between peak power and duration. B) The influence of removing low minor allelic frequency SNPs on the power of the S_{f} statistic. The black line represents the power of the S_{f} statistic at s = 0.08. Through tossing out low frequency SNPs (MAF $<$ 10%, red curve and $<$ 20%, blue curve), we have a buffer from moderate error rates. As we weigh intermediate frequency SNP counts more, we have a slight gain of peak power, and this power is gained prior to fixation (as the extremely high frequency SNPs are tossed out). However, the counts are lower and thus, more susceptible to noise, so we lose power over longer time spans. The red curve (10%) seems to provide a reasonable compromise between time and duration of high power.

tion of the beneficial allele, we may see frequency spectrum differentials (and thus, false positives) only because of differences in *de novo* mutation counts. Under the assumption that the regions under selection are relatively small compared to the whole genome, if we take the genome as a whole, the effective population sizes of both populations should be similar. Adding an additive corrective factor, equivalent to $\log \frac{\theta_1}{\theta_2} = \log \frac{\mu_1}{\mu_2}$, where the θ values are computed for the entire genome, thus cancels the effect of mutation rate differences. In practice, this term was negligible for H1/C1, H2/C1, A1/N1, and A2/N1 calculations (see the following section for definitions).

3.3.4 Power and significance computations

The Type I error can be obtained by getting bounds on the tail probabilities of the distribution of the statistic in a scenario of no-selection. However, the distribution is not well understood, and is not normal. A quantile-quantile plot (Figure 3.3) suggests a strong deviation from the normal distribution for both statistics. If the underlying distribution is normal, the quantiles of the observed data would be linearly related to the quantiles of the standard normal distribution. As can be seen, this is not the case — the plots indicate fatter tails than expected. The Lilliefors test for normality [32] quantifies the probabilities that any of these distributions are normal as $P < 2.2 \cdot 10^{-16}$. Therefore, we use empirical cut-offs for the statistics. To accurately determine Type I error, we use control vs control tests. In the simulations, we can replace the case population with a population evolving separately, but under identical conditions to the controls (for instance, under no selective pressure). In this situation, any deviation that appears to be significant is a false positive caused by genetic drift. To determine significance, we use a threshold corresponding to a 1% false positive rate from 500 control versus control simulations (in other words, the significance threshold was set at the fifth highest control vs control statistic value). For the *Drosophila* applications, we do not know the false positive rate, since there may be an unknown selective pressure acting on what we consider to be "controls". In addition, some of the parameters we have assumed for the simulations (for instance, uniform coverage) may not

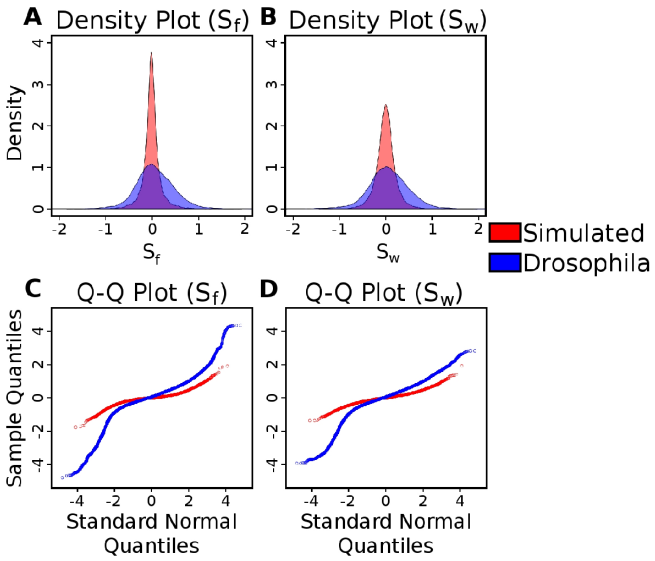


Figure 3.3: **Tests of the distributions of the S_f and S_W statistics in simulated and *Drosophila* (C1 & C2) control versus control data.** A) Density plots of simulated (red) and *Drosophila* (blue) control S_f data (blue). B) Density plots of simulated (red) and *Drosophila* (blue) control S_W data. As can be seen in both cases, with C1 and C2, the variance of the statistic is much higher than in the simulated control data. C) Quantile-quantile plots of simulated (red) and *Drosophila* (blue) control S_f data, compared to a standard normal distribution. D) Quantile-quantile plots of simulated (red) and *Drosophila* (blue) control S_W data, compared to a standard normal distribution. In these plots, a straight line indicates that the data comes from a normal distribution. In all cases, the plots indicate a non-normal distribution (quantified by the Lilliefors test for normality as having P-values less than $2.2 \cdot 10^{-16}$).

hold throughout the genome. As a result, we utilize a biological replicate of the controls (such as C2 above) to determine a 1% false discovery rate (FDR) instead. To calculate this FDR, we calculate the fractions provided by the complementary CDFs of our statistics, applied across the genome for both control versus control and case versus control studies. We set the threshold of significance where the ratio of these fractions is roughly 1%. For instance, for the summed frequency statistic, a 1% FDR occurs when the threshold is set at 4 in both hypoxia populations mentioned above.

3.4 Results

3.4.1 Power versus time under different measures of θ

To test the power of different statistics, we simulated case and control populations under fixed, high selection pressure (s = 0.08), and sampled $n = 400$ individuals from each (See Methods, and Figure 3.1). For each test, power was determined as the fraction of cases with a test statistic more significant than a 1% FDR level cut-off determined in control versus control simulations. As Figure 3.2A shows, the power for all methods is high shortly after fixation of the beneficial mutation (at roughly $t = 150$ generations). However, if the populations are sampled at times prior to fixation ($t < 150$), or subsequent to it ($t > 400$), the different tests behave very differently. For example, F_{st} has relatively high power prior to fixation, but it decays subsequently, while S_{H} shows the opposite behavior. S_{f} shows at least 50% power over a wide range of generations ($250 - 3000$). Interestingly, by removing sites with low minor allele frequencies, there is a shift in the power plot to earlier generations (Figure 3.2B).

To understand the reason for these trends, we plotted the mean scaled allele frequencies ($\hat{\theta}_i$) for the case and control populations at $t = 100, 200, 2000$ generations after bottleneck (Figure 3.4). As mentioned earlier, Achaz [25] showed that θ measures can be interpreted as linear combinations of these frequencies. A test of selection would have the most power when the chosen weights maximize the difference between the case and control populations. Under selective pressure, the lineage carrying the beneficial mutation expands rapidly, and we see an increase in intermediate to high frequency alleles in regions under selection (Figure 3.4A). However, the frequency spectra are generally too close to distinguish at this stage, though, and most tests do not have high power. At fixation (Figure 3.4B), we see an almost complete loss of intermediate frequency SNPs. At this stage, the populations are well separable, and nearly any weighting can distinguish the two populations. Thus, most tests do well. However, as θ_π and θ_{W} weigh high frequency alleles lower than lower frequencies, S_π and S_{W} show higher power than S_{f} (equal weights) and S_{H} (higher weight for high frequency SNPs). Subsequent to fixation,

with $t \in [500, 2000]$ (Figure 3.4C), the high frequency alleles drift to fixation. However, we start seeing a fair number of *de novo* mutations at low frequencies. As a result, the best signal is obtained by methods that weigh low frequency alleles lower than high frequency ($S_{\mathrm{f}}, S_{\mathrm{H}}$).

Tajima's D starts to lose power rapidly in the post-fixation regime. In addition, with this (or any other one sample test), we lose the benefit of a case-control setup, and we could lose power if there are any founder-specific or region-specific anomalies, which the simulator does not capture. The directional F_{st} generally performed well over a large time interval. Recall that the diversity in the case population is θ_π which weighs low frequency alleles higher than high frequency alleles. It is not surprising that its power rises quickly, as the signal comes from differences between individual SNP frequencies between two populations. We only need to see a loss of diversity within a population (which occurs prior to fixation). As the intermediate frequency SNPs start to rise (around 1500 generations), power is maintained longer than S_π due to the large number of fixed SNPs in cases that increase the diversity between the case and control populations. However, in order to accurately determine π_{between}, we need a reasonably high subset of SNPs to be sampled in both populations. In circumstances where this may not be feasible (see below), this approach will be underpowered.

Pooling reduces the cost of sequencing, but loses exact information on haplotype frequencies. To test the corresponding loss of power due to lack of haplotype information, we computed the power of the S_{f} statistic on the underlying haplotypes and compared it to the corresponding power of sampling at $70\times$ coverage, removing low frequency SNPs, and adding in 1% base-calling errors. The results (Figure 3.2B) suggest that even though we lose the ability to tell the exact frequencies, pooling does not have much impact on power. Assuming no sampling biases, the coverage is high enough such that we can accurately estimate the summed frequencies in a window. By removing SNPs with a minor allelic frequency of less than 10%, we gain peak power (since most *de novo* mutations in the cases get filtered out, leading to higher signal), but lose power as time goes on (as the SNP counts in both cases and controls become much lower, and thus, are more

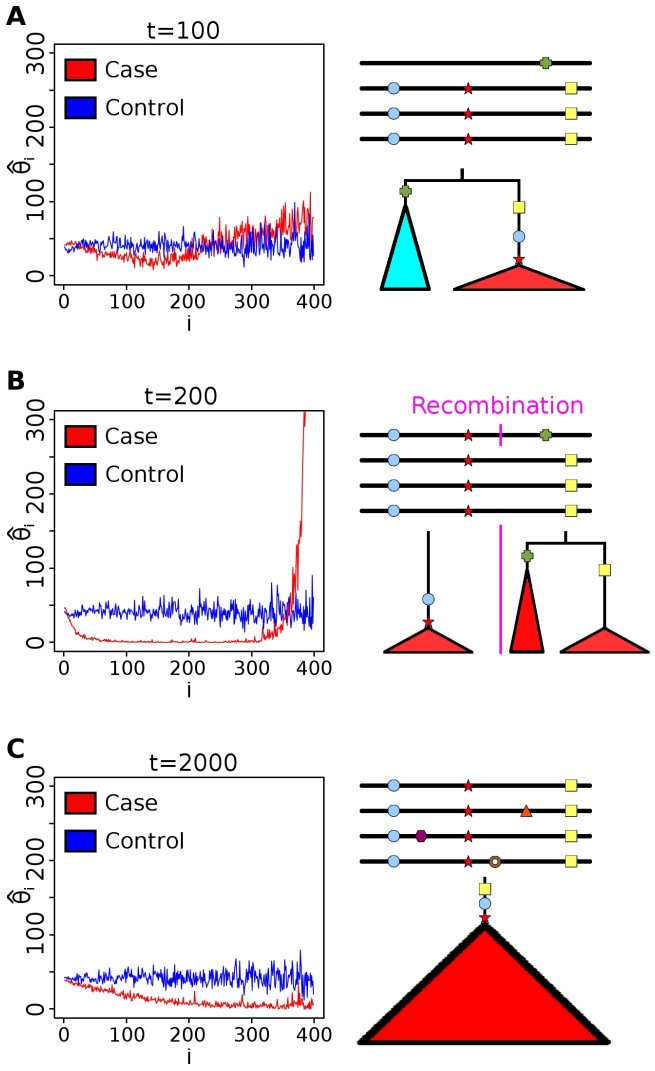


Figure 3.4: **SNP frequency spectrum analysis.** Scaled allele frequencies $\hat{\theta}_i = i\xi_i$ of simulated cases and control populations at different times after selection ($s = 0.08$). Here, ξ_i is the number of SNPs present in exactly i haplotypes. The ideal statistic is one that weighs alleles to maximize the separation between the blue and red line. A) At $t = 100$, individuals with the beneficial mutation start to dominate the population (leading to a rise of high-frequency hitchhiking SNPs), but there is still a substantial fraction of individuals without the beneficial mutation. B) At $t = 200$, many alleles that hitchhike with the beneficial mutation are near fixation, and the bulk of the mutations are very high frequency. Thus, measures that gives less weight to high frequency alleles show high power. C) At $t = 2000$, drift and recombination events decrease high frequency alleles, while *de novo* mutations increase lower frequency. Thus, measures that weigh lower frequency alleles more rapidly lose power even with a clear signal between case and control.

susceptible to noise). An additional benefit of filtering out low frequency SNPs is to dilute the impact of base-calling errors — due to the preprocessing, reasonable base-calling error rates barely influence the power.

Our results suggest that a variety of tests that use a case-control setup do well over different regimes of selection. We work with relatively high selection coefficients ($s = 0.08$) in these simulations, as it is often possible in model organism settings. By contrast, in naturally evolving populations, the selective pressure is usually lower and other tests might be better. We examine the impact of other parameters on power using the S_{f} and F_{st} statistics as exemplars.

Selection coefficient

According to coalescent theory, the time to the most recent common ancestor (MRCA) for a neutrally growing population with N haplotypes is $2N$ generations [30]. Under the assumption of a single beneficial mutation with viability $1 + s$, the number of copies of the mutant allele will increase exponentially until fixation. Thus time to MRCA of a sample containing mutant alleles is similar to a coalescent under an exponentially growing populations, and scales as $O(\frac{\ln 2Ns}{s})$ [33]. The power of the test for selection will increase up to this point, as the separation between the two populations increases. Once the majority of the case populations have reached fixation (i.e. their MRCA is after the introduction of the beneficial mutation), however, de novo mutations subsequently reduce the power of the test. Some intuition can be provided in Figure 3.4. At the time that the beneficial mutation is close to the MRCA, the tree has a long main branch as lineages not carrying the mutation are less likely to survive. All of the mutations on the same linkage block as the beneficial mutation fall on the main branch of this tree and are consequently near fixation. Further, the branch lengths of the lineages descending from the main branch are reduced. Thus, the allele frequencies are dominated by sites having very low and very high frequencies, and we see relatively small numbers of sites with intermediate frequencies (Figure 3.4B). Thus, any statistic that scores the difference between the observed and expected (under neutral selection) will be at peak power near this time. However, different statistics measure this

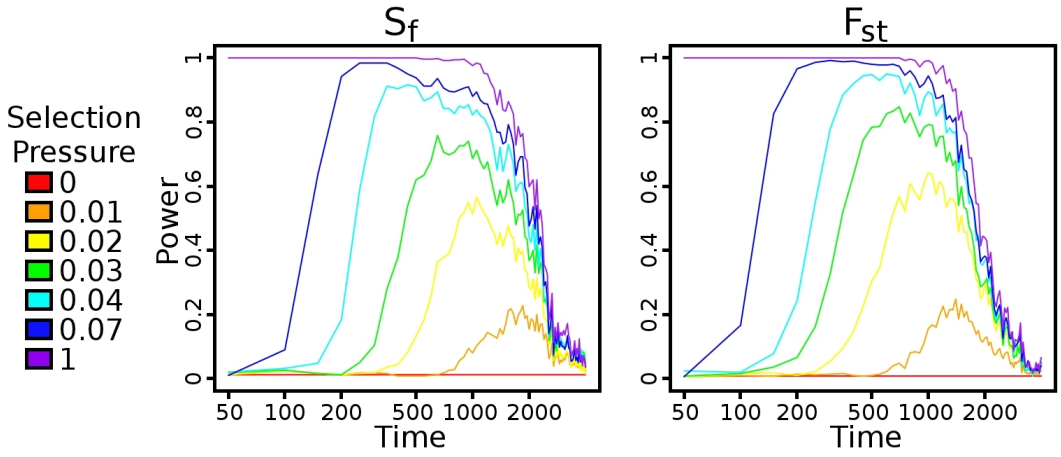


Figure 3.5: **Power versus time after bottleneck.** Power is calculated using a 1% false positive rate, as described in Section 2.5. Once a haplotype gets fixed in the population, the power increases rapidly for both statistics, although with slightly more power using F_{st} than using S_{f}. As time continues, though, the signal decays faster in F_{st} than in S_{f}, largely because the controls vary more sharply as time progresses.

skew in different ways, and reach peak power at slightly different times. With further passage of time, more lineages come out, and the main branch becomes shorter, and the scaled frequency spectrum starts to match the neutral spectrum. Consequently, the power of all statistics reduces (Figure 3.4C).

We tested how power of each selection pressure was impacted by the number of generations that the populations were allowed to diverge after the bottleneck (Figures 3.5 and 3.6). The time to maximum power with the S_{f} statistic indeed scales according to $\frac{\ln 2Ns}{s}$ (Figure 3.7), but reasonable power is achieved at a large number of generations. In all cases, the power drops rapidly at around N (2000) generations, disappearing at $2N$ generations. Additional trials of the impacts of coverage, base-calling error rate, founder population size, and sequencing sample size on S_{f} are in Figure 3.8. Given our population setup, around 20× pooled coverage is sufficient to achieve peak power. As mentioned earlier, the preprocessing provided tolerance to reasonable base-calling error rate. The population sample size seems to have minimal impact on power. The bottleneck has a fairly large

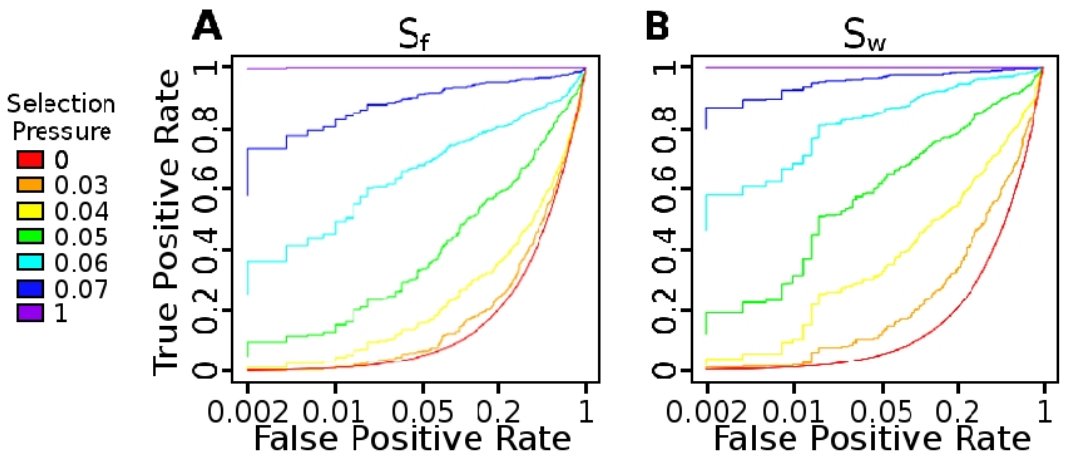


Figure 3.6: **ROC Curve evaluating the impact of threshold on true positive and false positive rates.** Using F_{st} leads to better performance than using S_{f} in this regime. For instance, at s = 0.05, with a 5% false positive rate, we can detect 33.8% of the true positives with S_{f} and 61.4% with F_{st}.

impact, in large part because the beneficial mutation is introduced in exactly one haplotype. Thus, a large bottleneck size corresponds to a small initial frequency as well. Over 200 generations and relatively small selection pressure, the beneficial allele has not had sufficient time to fix in the population.

3.4.2 Window size

For model organism studies, we intend on determining regions under selection using a statistic on a sliding window of fixed size w. The choice of w may be important in determining the signature of selection. The overall per-window recombination rate ($\rho' = \rho \cdot w$) and mutation rate ($\mu' = \mu \cdot w$) increase linearly with increasing window size, directly influencing the power of the statistic.

Figure 3.9 plots, as contours, the power of S_{f} and F_{st} as a function of ρ' and μ'. In general, the power increases with an increase in the mutation rate, as the selective sweep becomes easier to distinguish from genetic drift with a larger number of linked mutations. However, a very high mutation rate could create too many *de novo* mutations that mask the selection signature, leading to a reduction in power. In a similar fashion, the power increases with decreasing recombination

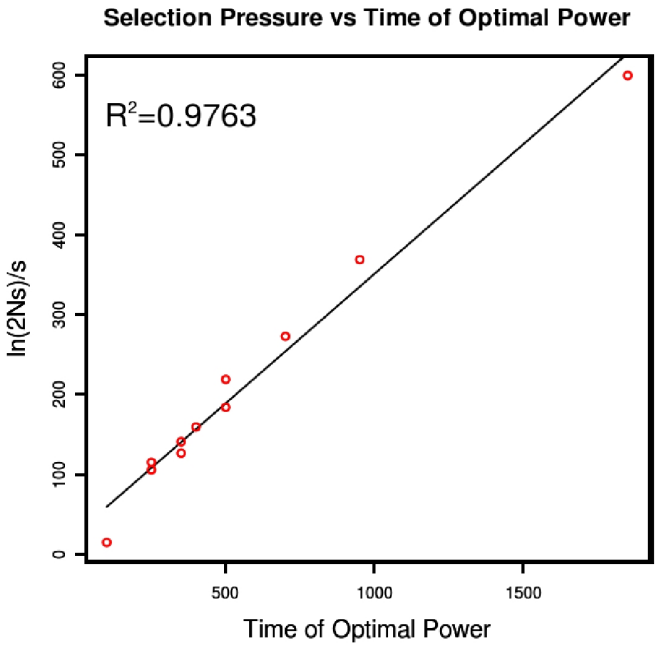


Figure 3.7: **Time to optimal power as a function of selection pressure.** Campbell [33] has shown that, under strong selection, the expected time to fixation can be approximated by (ln(2Ns))/s. This plot shows that, for a selection pressure s, the time that yields optimal power is linearly related to this quantity.

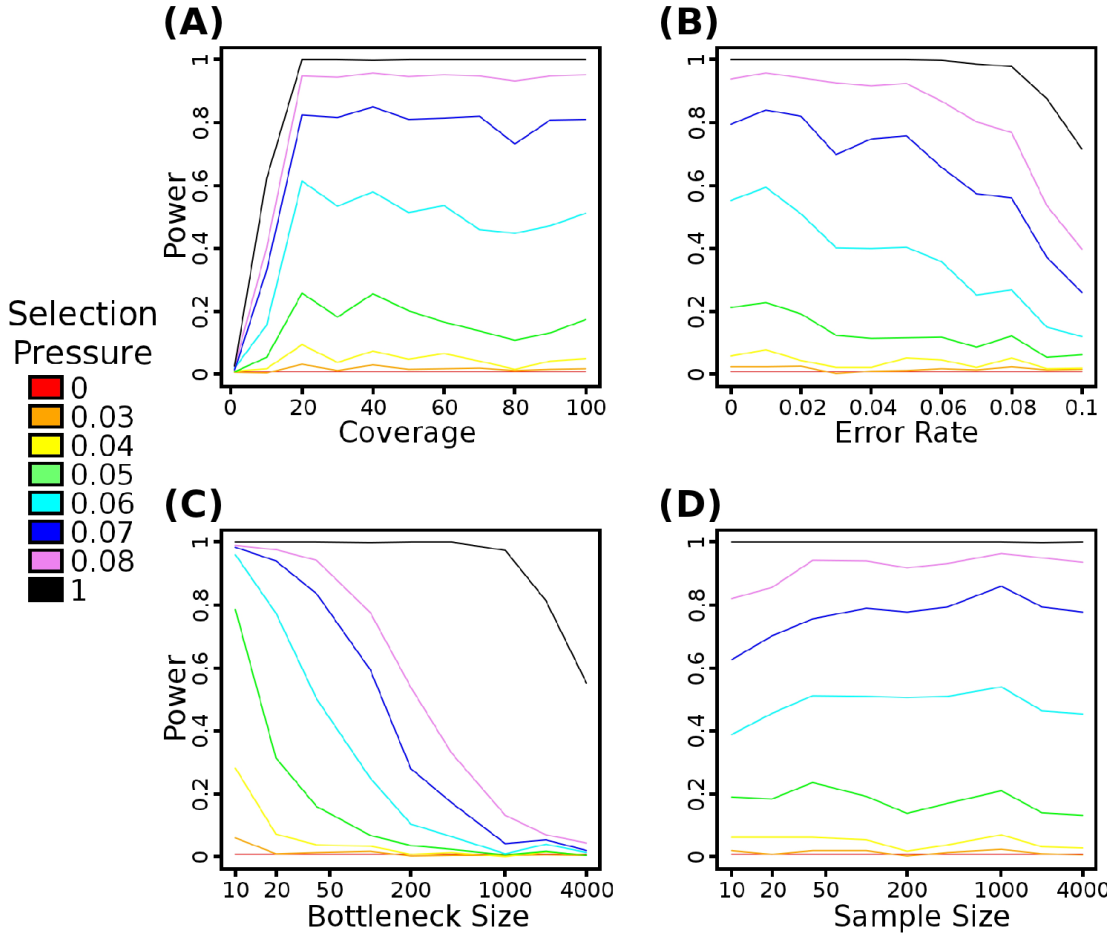


Figure 3.8: **The impact of simulator parameters on the power of the S_f statistic.** All trials were performed at 200 generations after the application of selection pressure and with all other simulator parameters as default, as per the Methods. A 1% false positive rate is also used to identify significance, as described in the Methods. A) The impact of average read depth on power. As can be seen, approximately $20 - 30\times$ coverage per population is sufficient to achieve maximal power. B) The impact of base-calling error rates on power. Due to the removal of SNPs with less than 10% MAF, the statistic is robust to reasonable levels of error. C) The impact of founder population size on power. The beneficial allele is present in exactly one haplotype in each setting, so this also represents the impact of the beneficial allele's initial frequency on power. As expected, a low initial frequency (and high initial haplotype diversity) leads to a loss of power. D) The impact of sequencing sample size on power. This parameter does not have a major impact on power — even 10 individuals are sufficient to determine signatures of selection in the population.

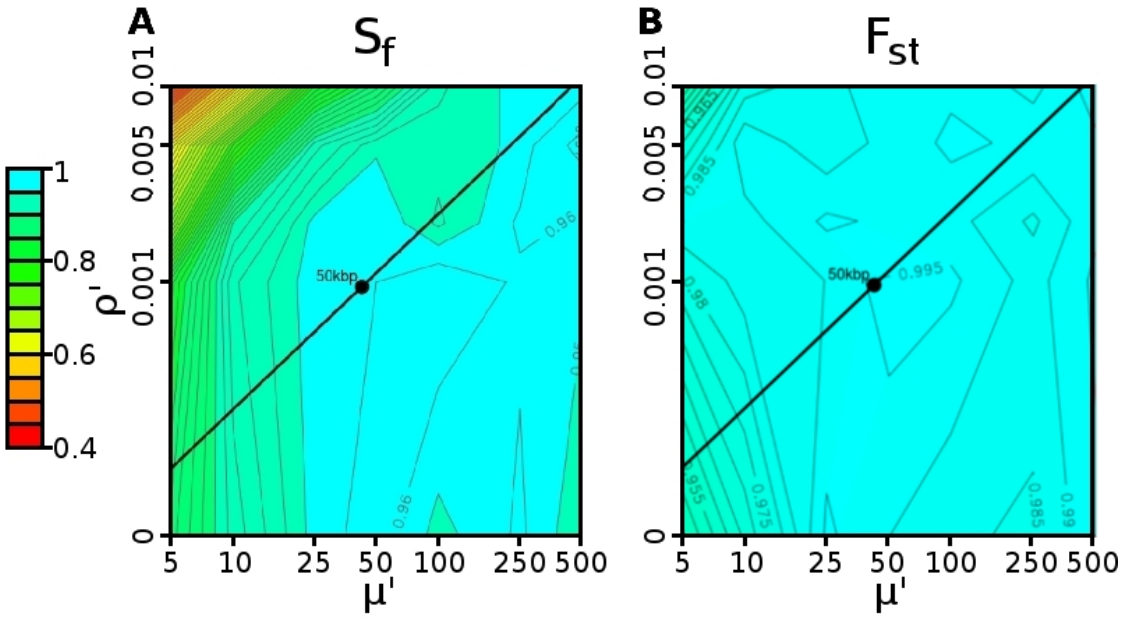


Figure 3.9: **Contour plot describing determination of window size.** The power of the (left) S_{f} and (right) F_{st} statistics for various combinations of recombination rate ($\rho' = \rho \cdot w$) and mutation rate ($\mu' = \mu \cdot w$). In both cases, the selection pressure was 0.08 (strong), and thresholds were defined by a 1% false positive rate, as described in Section 2.5. The black line represents the range of possible values caused by varying the window size in *Drosophila*. The black circle represents the 50kbp window size ultimately chosen, which shows high ($\geq 95\%$) power for both S_{f} and F_{st}.

rates, as recombination events increase the genetic diversity in the region.

The mutation and recombination rates both depend on factors such as sequence complexity, but we can estimate the average per base rates. For *Drosophila*, the per-base mutation rate ($8.5{\cdot}10^{-4}$/bp) was taken from Watterson's estimator applied to the C1 population (defined in Materials and Methods), while the per-base recombination rate was taken from Fiston-Lavier et al. [29] to be $1.892 \cdot 10^{-8}$/bp. If we treat the ratio of mutation to recombination rate as constant ($\mu'/\rho' = \mu/\rho \approx 44926$), varying the window size yields the black lines in these diagrams. Both statistics are robust to a wide range of window sizes. For $w = 50$ kbp (shown by the black circle in Figure 3.9), we have more than 95% power. As a result, we use 50 kbp windows in all trials.

3.5 Discussion

Our test shows great promise in identifying signatures of selection. The results suggest an economical, yet effective, approach for utilizing the capabilities of whole-genome sequencing to identify genetic determinants of phenotypes. The method requires that the phenotypes be used as a basis for laboratory selection and that the genomic data of the selected population be tested for signatures of selection. However, the test does not require high levels of sequencing to have full power. $20-30\times$ pooled read depth per population was sufficient in our simulations; for typical model organisms with small genomes, this is relatively inexpensive.

In addition, the test provides higher resolution than genetic crosses and does not require a second level of sequencing to identify causal variants. For instance, let us take the cross-based protocol described in Leips et al. [34]. In this protocol, F1 offspring between two Drosophila strains are backcrossed to the parental line without the beneficial mutation. The resulting offspring are interbred for four generations, and then 98 recombinant inbred lines are created (which takes 25 more generations). In the paper, the QTL region sizes range from roughly 100kbp to 3Mbp. Our approach does not require the labor associated with constructing the cross and maintaining the RI lines, and we can easily reduce the LD block

size further by two main mechanisms: increasing the number of generations that the populations are allowed to mix and increasing the initial genetic diversity (for instance, by increasing the number of parental strains).

At the same time, the test may not be universally applicable. Not all phenotypes present a strong selection pressure, and weak selection results in a weak signature. However, our empirical data in the scaled allele frequency spectrum provides a strong theoretical foundation for when a specific test might be effective, and the difference between the scaled allele frequencies of regions under selection versus controls suggest alternative tests that will be explored in the future. Furthermore, with dropping sequencing costs, it may be feasible to test the population at many time points during evolution, and develop tests of selection that also look at trends in the scaled allele frequency spectrum over time. Recently, tests have been developed for testing recent selection events, using long range haplotypes and other signals [18]. We plan to develop analogs for pooled data to improve the power of the statistic for recent selection events. For older selection events, we plan on improving the statistics by using tests that depend upon an excess of coding and functional variants.

The test would work best given the sequences of a second control population (such as C2 above), so that a more accurate null distribution can be determined. In the absence of this, we can estimate the null distribution either via simulations or by assuming that the bulk of the genome is not under selection and using genomic control. Also, we determined that 50kbp windows were appropriate by determining that an average window (as determined by per-base estimates of mutation and recombination rate) of this size has high power (see Figure 3.9). Not all regions of the genome are equivalent, though. For instance, if we assume that a recombination hotspot has the same properties as a typical region of the genome, we might overestimate the homogeneity of the region encompassed by the selective sweep. If we have an accurate recombination map, it would be beneficial to improve this by adjusting the window size (and significance threshold) based on location. In the absence of this, an alternate approach would entail running these tests over multiple window sizes. In this scenario, we would determine significance based on

the presumed window size.

The statistic can potentially be applied to naturally occurring (or industrial) strains of organisms that have been evolving independently under different selective constraints, and even human populations. However, the lack of proper controls in naturally evolving sub-populations, will possibly require additional tests to associate the genetic signatures with the appropriate phenotypes. In this case, factors such as the time of divergence and isolation of sub-populations will also need to be considered. Development of these ideas will provide us with new tools for associating genotypes and phenotypes.

3.6 Acknowledgments

This work was supported by the National Science Foundation (NSF-III-0810905 and NSF-CCF-1115206); the National Human Genome Research Institute (5R01HG004962); the Eunice Kennedy Shriver National Institute of Child Health and Human Development (5P01HD032573); the National Institute of Neurological Disorders and Stroke (5R01NS037756); and an American Heart Association Award (0835188N).

Chapter 3 is in full adapted from *Udpa N, Zhou D, Haddad GG, Bafna V. Tests of selection in pooled case-control data: an empirical study. Front Genet. (2011);2(83).* The dissertation author was the primary author of this paper responsible for the research.

Chapter 4

The role of the Notch pathway in hypoxia-tolerant *Drosophila melanogaster*

4.1 Abstract

After determining the relative strengths and weaknesses of a number of tests, we are now ready to apply the appropriate tests to a number of datasets. Firstly, through long-term laboratory selection (over 200 generations), we have generated *Drosophila melanogaster* populations that tolerate severe, normally lethal, levels of hypoxia. Because of initial experiments suspecting genetic mechanisms underlying this adaptation, we compared the genomes of the hypoxia-selected flies with those of controls using deep resequencing. Since we know the levels of hypoxia are typically lethal, this implies that, at least for the top hits, the selection pressure is very strong. By applying the appropriate computing and analytical methods we identified a number of DNA regions under selection, mostly on the X chromosome. Several of the hypoxia-selected regions contained genes encoding or regulating the Notch pathway. In addition, previous expression profiling revealed an activation of the Notch pathway in the hypoxia-selected flies. We confirmed the contribution of Notch activation to hypoxia tolerance using a spe-

cific γ-secretase inhibitor, N-[N-(3,5-Difluorophenacetyl)-L-alanyl]-S-phenylglycine t-butyl ester (DAPT), which significantly reduced adult survival and life span in the hypoxia-selected flies. We also demonstrated that flies with loss-of-function Notch mutations or RNAi-mediated Notch knockdown had a significant reduction in hypoxia tolerance, but those with a gain-of-function had a dramatic opposite effect. Using the UAS-Gal4 system, we also showed that specific overexpression of the Notch intracellular domain in glial cells was critical for conferring hypoxia tolerance. Unique analytical tools and genetic and bioinformatic strategies allowed us to discover that Notch activation plays a major role in this hypoxia tolerance in *Drosophila melanogaster*.

4.2 Introduction

As mentioned previously, oxygen homoeostasis is essential for development, growth, and integrity of cells, tissues, and organisms. Limited oxygen supply to cells and tissues (hypoxia) has a wide range of physiologic and potentially pathologic consequences, ranging from ischemic/hypoxic heart disease, stroke, and pulmonary hypertension to a number of obstetrical/perinatal complications, to high-altitude illnesses, to organ transplantation, and finally to intratumor hypoxia and cancer progression. Despite the clinical importance and societal disease impact of such a wide range of disorders, the molecular underpinnings of susceptibility or tolerance of cells or tissues to lack of O_2 are not well understood. Many studies have investigated the mechanisms that lead to injury when cells are deprived of O_2, but to potentially treat or prevent the consequences of hypoxia necessitates also the understanding of the inherent tissue mechanisms that are critical for tolerance and survival. To do so, we use a long-term laboratory selection strategy that unmasks mechanisms that play an important role in hypoxia tolerance in a genetic model, *Drosophila melanogaster* [35, 36]. In this attempt, starting with 27 isofemale D. melanogaster strains, and applying decreasing levels of O_2 over more than 200 generations, we generated *Drosophila* populations that tolerate severe levels of hypoxia, which are lethal to the original parental lines. These hypoxia-

adapted flies (AF) pass the tolerance trait from generation to generation and the trait persists even in the absence of hypoxic stress (i.e., after several generations in a normoxic environment), suggesting that a genetic rather than a physiological mechanism underlies adaptation. We have discovered that a) genetic mechanisms allowed D. melanogaster to adapt and tolerate extremely low O_2 environments, b) there were genomic intervals that were selected for, mostly on the X chromosome, that occurred during long-term hypoxia, and c) the Notch pathway played an important role in hypoxia tolerance.

4.3 Results

4.3.1 Hypoxia-selected regions and genetic profiles in the adapted genome

To determine whether there are DNA signatures of hypoxia selection, we sequenced two control (C1 and C2) and two AF (H1 and H2) populations that had been under hypoxia selection for 180 generations in separate environmental chambers at $> 60\times$ coverage, using the Illumina GA II sequencer. We aligned between 120 million and 200 million 54-bp paired-end reads per population to the D. melanogaster reference sequence. Because individual genotypes and the number of individuals sampled at any given base and standard linkage disequilibrium (LD) information could not be determined from the pooled sequence data, standard tests of selection could not be used [15, 17]. Consequently, we used an approach defined in the previous chapter to determine the hypoxia-selected regions and genetic profiles in the AF genome.

This was a coarse-grained approach to compare SNP distributions in both control and AF pools. In this approach, we used the S_{f}(C1, H1) statistic to determine potential regions under selection using the sequence data generated from the pooled populations. This represents the log ratio of control and AF scaled mutation rates and provides a comparison of the effective population sizes. We identified S_{f}(C1, H1) $/gg0$ as indicative of deviation from neutrality and consis-

tent with a purifying selection in hypoxia (reduction in effective population size). We used S_{f}(C1, C2) as an empirical control for false discovery rate (FDR) computation. We also investigated the concordance between the AF populations from different chambers by comparing the independent estimates S_{f}(C1, H1) and S_{f}(C1, H2). Using a S_{f}(C1, H1) cutoff of 4, corresponding to an FDR of 1%, and overlapping 50-kbp windows, we observed remarkable concordance between regions under selection in H1 and H2. A total of 1,509,436 bp comprising 24 distinct hypoxia-selected regions and containing a total of 188 genes were under selection in both H1 and H2. Twenty of these regions (>80%) were located within a 10-Mb interval on chromosome X and the remaining 4 were located within a 1.2-Mb interval on chromosome 3R (Figure 4.1 A-D). These results demonstrate that two populations of flies, independently selected for hypoxia tolerance (i.e., in different environmental chambers) had the same intervals in the genome undergo a high degree of fixation (hypoxia-selected regions) and suggest that the genes required for adaptation to severe hypoxic conditions are localized rather than distributed across the genome. The latter observation is reinforced by the distribution of C1 vs. H1 or H2 fixed SNPs across the genome: Whereas all three populations have a median value of 28 fixed SNPs (range 0–363) per 50-kb interval across genome, there is a large difference between C1 and H1 and H2 in the hypoxia-selected regions where the AF populations have threefold higher fixed SNPs compared with the control flies ($\approx$74; Poisson P $= 2.57 \cdot 10^{-13}$). Furthermore, conservation between H1 and H2 of fixed SNPs was also higher in the hypoxia-selected regions compared with those in the non-hypoxia-selected regions (93% vs. 78%; hypergeometric P $= 3.12 \cdot 10^{-43}$). Consistent with a hypoxic stress-mediated population bottleneck leading to an overall loss of diversity in the AF populations, we observed a much higher genetic similarity between H1 and H2 (66% of the fixed SNPs are common), compared with a concordance of 35-37% between AF and control lines, despite the identical ancestry (Figure 4.1 E and F).

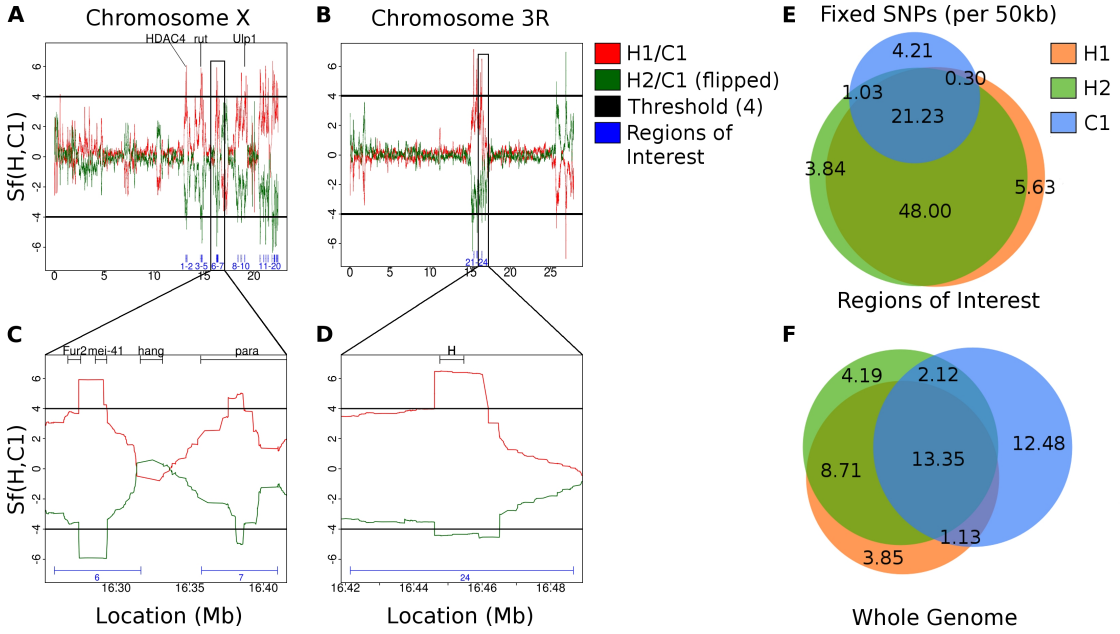


Figure 4.1: **Genomic regions under selection for hypoxia tolerance.** The S_{f} statistic was computed on overlapping 50-kbp windows, comparing two hypoxia samples (H1 and H2) to a normoxia one (C1). Regions with S_{f} values >4 are shown, corresponding to 1% FDR (based on comparing two normoxia samples). The results for chromosomes (A) X and (B) 3R show near perfect concordance between H1 vs. C1 (red) and H2 vs. C2 (green, inverted). Regions deemed significant for both hypoxic samples are shown as blue intervals. (C) Regions 6 and 7, including genes Fur2, mei-41, hang, and para, and (D) region 24, containing Hairless, are shown in greater detail. Note that the value of the statistic is plotted at the median of the region and decays as it goes toward the edge of the selected region. (E and F) Venn diagrams, describing the frequency of fixed SNPs (per 50 kbp) in (E) just the regions under selection and (F) the entire genome, were tabulated. Orange represents H1, green represents H2, and blue represents C1.

4.3.2 Candidate genes underlying hypoxia tolerance in AF

There are 188 genes located in the hypoxia-selected regions. To identify the set of potentially causal genes in hypoxia tolerance, we started initially with 28 genes because these were previously implicated in either hypoxia or similar phenotypes (such as oxidative stress and aging). We filtered the remaining 160 genes for evidence of a selective sweep using three complementary tests: i) the McDonald-Kreitman test between H1 and C1 for evidence of adaptive evolution based on correlation between fixed (i.e., >90% frequency) and nonsynonymous mutations in a population ($P \leq 0.05$) [37], ii) ≥1.5-fold transcriptomic change under hypoxia [35], and iii) sorting intolerant from tolerant (SIFT) evaluation of the impact of fixed, nonsynonymous mutations on the functions of proteins encoded by the genes in these regions ($P \leq 0.05$) [38]. A total of 68 genes from the 160 genes were identified and many of these were also observed using the fine-granularity assessment. These include 12 genes that interact with, activate, or are targets of the Notch pathway. For example, there are two members of the Notch repressor complex, i.e., Hairless and HDAC4, that are located in the hypoxia-selected regions in AF. HDAC4 contains a SNP fixed in both AF populations with significant impact on function (SIFT P value = 0.05) but virtually absent in both control populations (A1004S in isoform B).

Because a) there were a number of genes in the Notch pathway or related to Notch signaling that were selected for in AF in contrast to the control flies and b) our previous expression profiling studies demonstrated that the Notch pathway is activated in the AF flies in comparison with control flies [35], we focused our investigation next on the role of the Notch pathway in hypoxia tolerance.

4.3.3 Notch activation is critical for hypoxia tolerance in *Drosophila melanogaster*

To dissect the contribution of Notch activation to hypoxia tolerance, we used both genetic tools and pharmacologic agents. We first examined the role of Notch in hypoxia tolerance using three homozygous-viable Notch mutants: N[Ax-tsl],

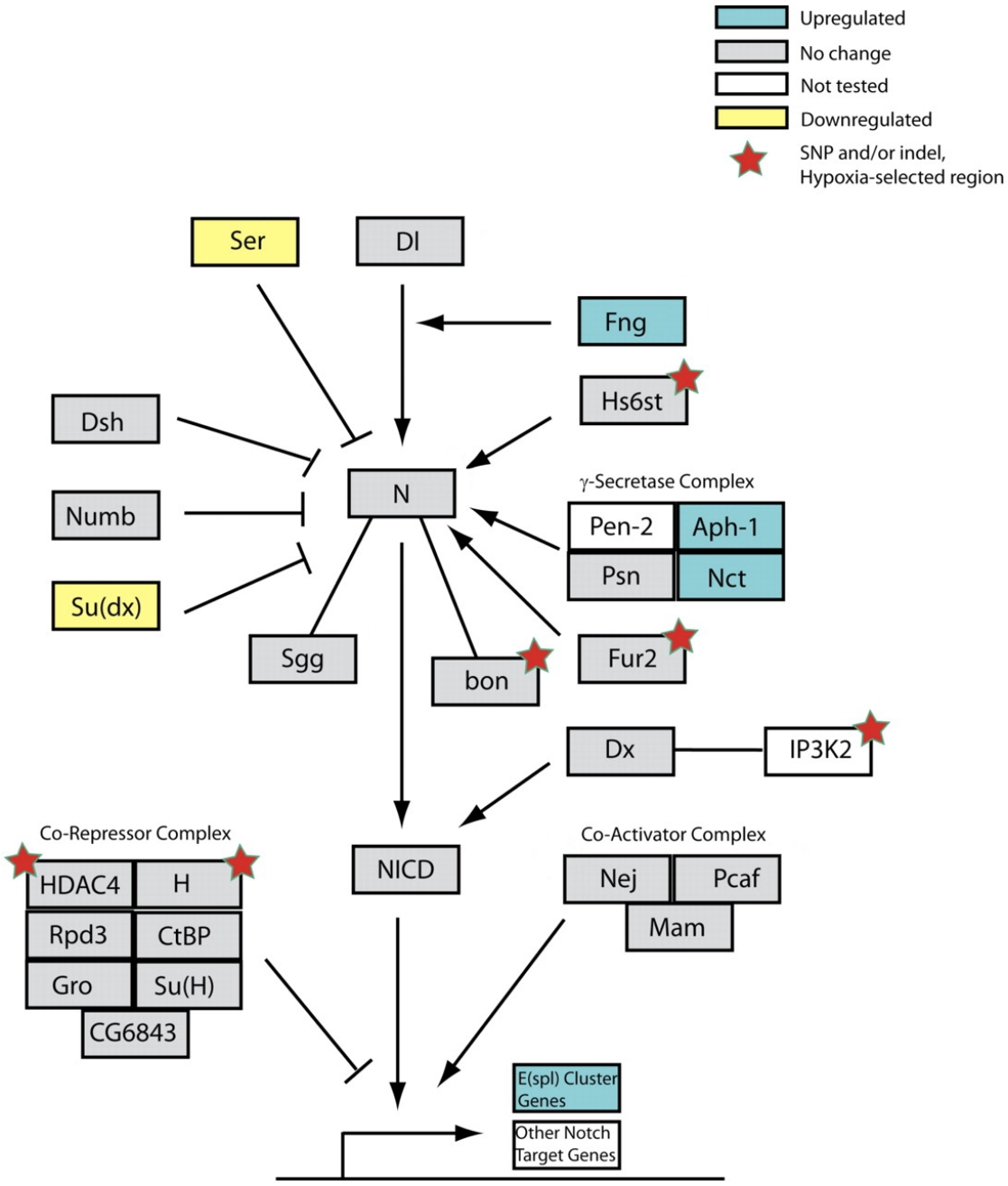


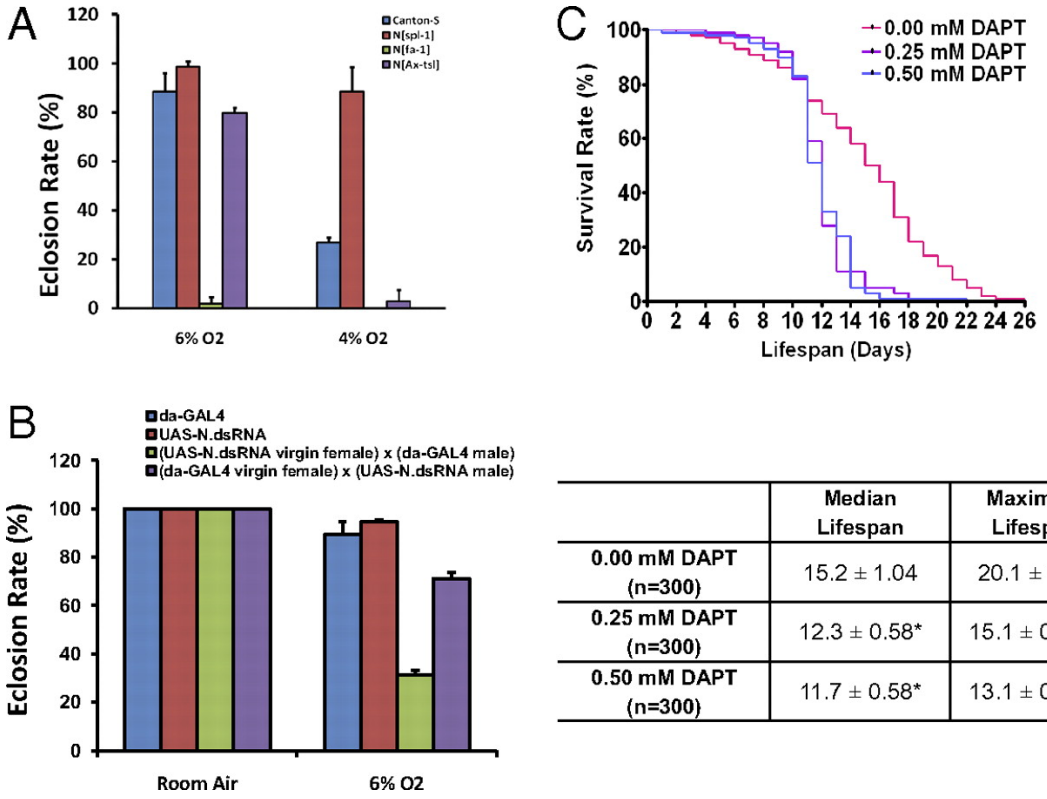
Figure 4.2: **Enrichment of fixed SNPs and indels in an extended Notch pathway.** The Notch pathway was adapted from KEGG [39] by adding Notch interactors from the literature to create an expanded Notch signaling pathway. Genes differentially expressed in larva (expression levels from Zhou et al.) [35] are cyan (up-regulated) or yellow (down-regulated), genes showing no change in expression are gray, and untested genes are white. Genes located in a hypoxia-selected region are indicated with red stars.

N[fa-1], and N[spl-1] [40, 41, 42]. The N[fa-1] mutation is caused by an insertion of a transposable element (opus) in the second intron of the Notch gene, and N[Ax-tsl] is generated by ethyl methanesulfonate-induced mutagenesis with both N[fa-1] and N[Ax-tsl] having loss-of-function mutations. Unlike N[fa-1] and N[Ax-tsl], N[spl-1] is a Notch gain-of-function allele carrying a point mutation in EGF repeat 14 of the Notch protein that replaces the Isoleucine578 with a Threonine [43]. We found that N[fa-1] and N[Ax-tsl] were hypersensitive to hypoxia and had a lowered survival rate, even in much milder hypoxic conditions (i.e., 6% O_2) (Figure 4.3A). In contrast, N[spl-1] exhibited remarkable hypoxia tolerance and survived 4% O_2, much like the AF flies. We next used an RNAi strategy and determined that flies with a knockdown of Notch had a hypoxia-sensitive phenotype (Figure 4.3B). These results clearly indicate that Notch function is critical for survival under hypoxia.

Because the up-regulation of genes encoding aph-1 and nct subunits of γ-secretase suggested that activation of Notch signaling in the AF flies might involve γ-secretase, we used a specific γ-secretase inhibitor, N-[N-(3,5-Difluorophenacetyl)-L-alanyl]-S-phenylglycine t-butyl ester (DAPT) [44] and examined their life span in severe hypoxia in the AF population. We found that DAPT treatment indeed reduced significantly both median and maximum life span in the AF flies (Figure 4.3C) but had no significant effect on control flies.

4.3.4 Spatial-temporal activation of Notch and its downstream genes in hypoxia tolerance

The UAS/GAL4 system was used to determine the critical spatial-temporal activation of the Notch pathway in hypoxia tolerance [45, 46]. Several available GAL4 lines were crossed with a UAS-Notch intracellular domain (UAS-NICD) transgenic stock to generate progeny that had specific Notch activation in specific cells/tissues or during specific developmental stages. The results showed that specific expression of NICD in the neurons and/or glial cells conferred hypoxia survival in the progeny. For example, the progeny derived from crosses in which NICD was up-regulated in a specific subset of glial cells showed a remarkable



	Median Lifespan	Maximum Lifespan
0.00 mM DAPT (n=300)	15.2 ± 1.04	20.1 ± 1.25
0.25 mM DAPT (n=300)	12.3 ± 0.58*	15.1 ± 0.55**
0.50 mM DAPT (n=300)	11.7 ± 0.58*	13.1 ± 0.72**

Figure 4.3: **Hypoxia tolerance in Notch mutants and γ-secretase inhibitor-treated flies.** (A) N[spl-1] (a gain-of-function allele) has an increased survival rate in hypoxic conditions, while N[fa-1] and N[Ax-tsl] (loss-of-function mutants) show reduced survival rates even in mild hypoxic condition (*P $<$ 0.01, compared with Canton-S control). (B) RNAi-mediated Notch knockdown induces increased sensitivity to hypoxia in flies. Flies carrying a UAS-N.dsRNA transgene on the X chromosome were used to determine the function of Notch in hypoxia tolerance. Two crosses were used to generate flies that had N.dsRNA expression either in all progeny (cross A) or only in the female progeny (cross B). Cross A progeny showed an increased sensitivity to hypoxia relative to cross B (*P $<$ 0.01). Both male and female progeny were included in scoring. (C) γ-secretase activation plays an important role in hypoxia tolerance in hypoxia-selected flies. Five-day-old adult hypoxia-selected flies were collected and treated with 0.25 or 0.50 mM DAPT and their life span was determined under 1.5% O_2. Median life span was the time when 50% of death occurred in the sample, and maximum life span was the time when 90% of sample were dead. Compared with the vehicle-treated controls, flies treated with DAPT showed a significant reduction of both median and maximum life span (*P and **P $<$ 0.01, Student's t test).

increase in hypoxia tolerance and survival. As shown in Figure 4.3.4 and Figure 4.3.4, the Eaat-GAL4 and P{GawB}17A-Gal4-driven NICD overexpression in glial cells significantly enhanced both eclosion rate and adult life span (after eclosion) in hypoxia.

To further explore the mechanisms underlying Notch-mediated hypoxia tolerance, we tested the role of a Notch downstream gene, E(spl)mα, that is located at the E(spl) genomic region and that is significantly up-regulated in the hypoxia-selected flies [35]. We generated first a homozygous fly to use in a subsequent cross to provide a Notch overexpression background [i.e., (Eaat1-GAL4/Eaat1-GAL4; UAS-NICD/UAS-NICD) stock, EN line]. This EN line was then crossed to homozygous flies that carry UAS-E(spl)mαRNAi. The progeny of this cross carried Eaat1-GAL4, UAS-NICD, and UAS-E(spl)mαRNAi that allowed us to knock down the target E(spl)mα gene on the Notch overexpression background and determine if Notch activation requires E(spl)mα to regulate hypoxia tolerance. As shown in Figure 4.3.4, Notch activation, which conferred hypoxia tolerance without the RNAi for E(spl)mα, was totally abolished with E(spl)mα knockdown, demonstrating a critical role of E(spl)mα in regulating Notch-induced hypoxia tolerance.

4.4 Discussion

D. melanogaster has been used as a powerful genetic model for about a century. Because many genes and pathways are evolutionarily conserved between *D. melanogaster* and humans, *Drosophila* has become one of the most effective tools for dissecting the genetic mechanisms of human diseases, including developmental and neurological disorders, cancer, cardiovascular disease, and metabolic and storage diseases (for selected reviews see [47, 48, 49, 50]. The current study has taken advantage of this Drosophila model and used population genetics, deep-sequencing, and molecular strategies to identify potential causative gene(s) underlying hypoxia tolerance. The use of experimental selection methods followed by deep sequencing of pooled individuals is a unique experimental technique in its own right. In a recent publication [51], a similar evolutionary experiment was performed, but

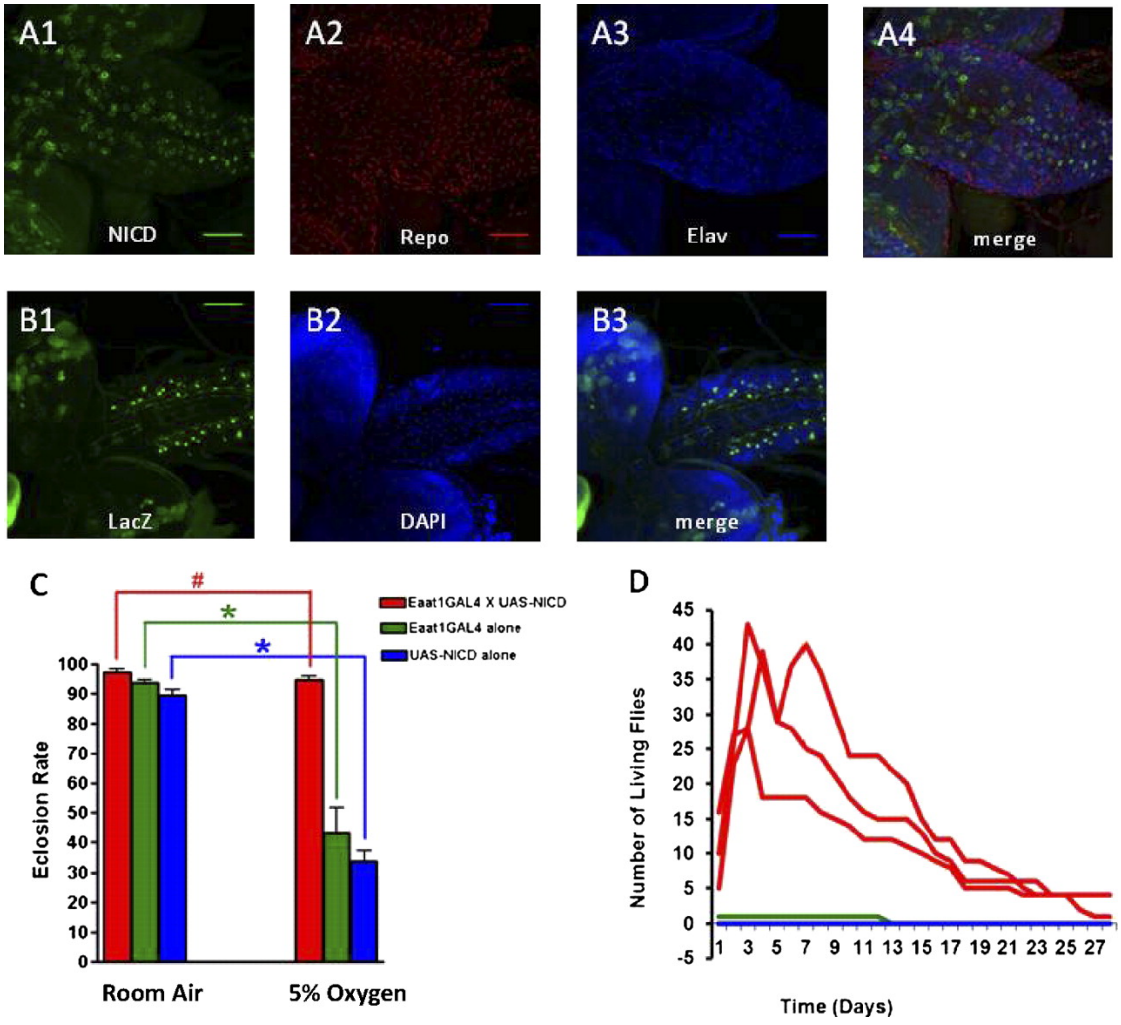


Figure 4.4: **Increased hypoxia tolerance in flies with Eaat-GAL4-driven NICD overexpression in glial cells.** (A) NICD overexpression (green) in the third instar larval brain was determined in the progeny of the (Eaat1-GAL4) x (UAS-NICD) cross. Staining with Repo (red) and Elav (blue) reveals that NICD overexpression is mostly in glia (Repo+). (B) NICD overexpression-induced activation of Notch in glia was determined by crossing (Eaat1-GAL4/Eaat1-GAL4; UAS-NICD/UAS-NICD) to Su(H)-LacZ reporter. LacZ (green) was detected in the cells in the same locations as in A, demonstrating that the Eaat1-GAL4-driven NICD overexpression is functionally active. (C) Glial-specific overexpression of NICD enhances the eclosion rate in hypoxia. #P value = 0.205; *P value <0.001 (unpaired Student's t test, n = 6). (D) Glial-specific overexpression of NICD enhances adult survival in hypoxia. Each day, the survival statistics changed by adding the number of newly eclosed experimental and control flies (which were separated from the pupae) and subtracting the number of new adult dead flies.

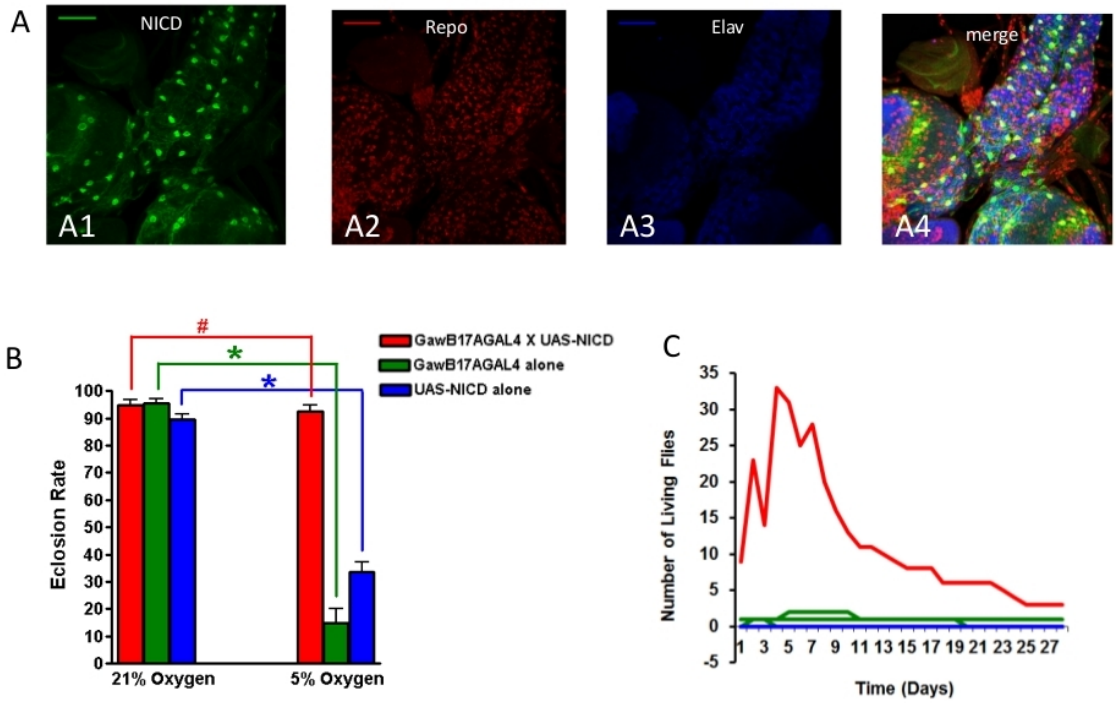


Figure 4.5: **Increased hypoxia tolerance in flies with P{GawB}17A-Gal4-driven NICD overexpression in specific glial cells.** (A) NICD is over-expressed in the third instar larvae brains (A1). The majority of these NICD over-expressing cells express Repo (A2), while a few express Elav (A3), and a couple express neither (A4). (B) The eclosion rate of GawB17A>NICD compared to parental lines is significantly higher. (C) An analysis of the adult flies who survive hypoxic eclosure and are maintained in hypoxia clearly displays the advantage of this line when compared to parental lines. Images taken on Olympus confocal FV1000 at 400X, with Z-projections of planes respectively. Scale bar=50 um. For eclosion rate, means ± SEM are displayed; *p<0.0001, #p=0.456 as determined by unpaired t-test with 6 vials total, with 3 vials each in two independent experiments.

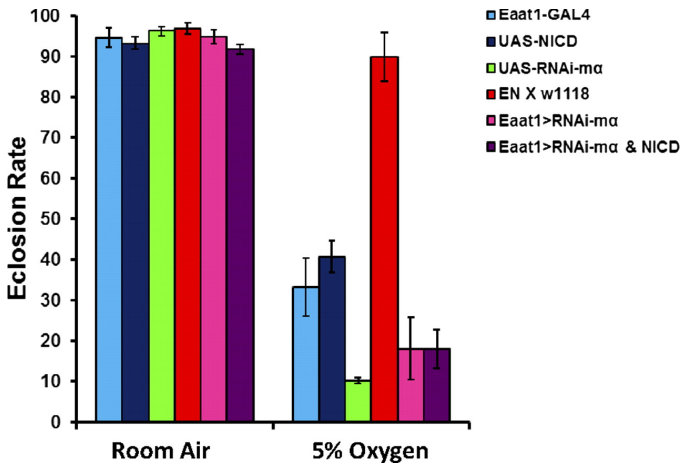


Figure 4.6: **Notch activation-conferred hypoxia tolerance requires E(spl)mα.** The EN-line flies, which are homozygous for both the Eaat1-GAL4 driver and UAS-NICD, were crossed to homozygous UAS-E(spl)mα. RNAi stock to knock down E(spl)mα on the background of Notch overactivation. Hypoxia tolerance of the progeny was determined by the eclosion rate at a 5% O_2 hypoxic condition. Compared with the Notch overactivation control (EN/w1118), knocking down E(spl)mα on the Notch overactivation background abolished hypoxia tolerance ($P < 0.01$; t test). Bars represent the mean $\pm$ SEM ($n = 3$) for each group/treatment.

with a different phenotype. Their analysis (different from ours) does not point to any specific regions with a genome-wide signature of selection. Rather, they suggest that the adaptation is due to "incomplete sweep models." By contrast, our experiments and analysis provide a more definitive evidence of a selective sweep. Specifically, we have identified 24 distinct hypoxia-selected regions containing a total of 188 genes in the hypoxia-selected population compared with the naive ones. Interestingly, 20 of these regions (>80%) were located on chromosome X. Because Drosophila males are hemizygous on the X chromosome with no possibility for recombination, recessive X-linked alleles are readily available for selection in males. A consequence may be a more rapid accumulation of favorable mutations as well as stronger purifying selection of deleterious recessive alleles on the X chromosome relative to autosomes [52, 53, 54].

The hypoxia-selected regions in the hypoxia-selected flies contain several genes that belong to the Notch signaling pathway and/or regulate the activity of Notch. For example, both Hairless and HDAC4 are important in Notch signaling because downstream genes in the Notch pathway are activated through binding of the NICD to the DNA-binding complex CSL [55]. NICD competes for CSL binding with a corepressor complex, which consists of Hairless, CtBP, and Groucho, and additionally recruits histone deacetylases (HDAC) [56]. Such enrichment of polymorphic differences in Notch pathway-related genes and the evidence of Notch activation revealed by expression profiling suggested a potentially important role of Notch in hypoxia tolerance. This role of Notch was further investigated using Notch mutants, RNAi-mediated knockdown, and pharmacological inhibitory reagents. These experiments demonstrated that Notch activation confers hypoxia survival. In addition, a remarkable hypoxia survival was observed in flies with NICD overexpression in specific neuronal and/or glial cells, demonstrating the importance of maintaining the integrity of neuronal function in hypoxia tolerance for the whole organism. Although the precise underlying mechanisms regarding the mode of action of Notch remain elusive, we show in this work that the specific NICD activation in particular cells in the central nervous system confers survival and protects from hypoxia-induced death. Indeed, previous studies have demon-

strated an anti-apoptotic function of Notch [57, 58, 59, 60, 61, 62, 63, 64] and the current study provides evidence demonstrating that the activation of Notch in highly differentiated cells can maintain its survival-promoting property. Furthermore, a recent whole-exome sequencing study of 50 Tibetan subjects revealed that the frequency of a noncoding intronic SNP located at the EPAS1/HIF2α locus is significantly higher in the high altitude-adapted Tibetan subjects than in the lowlander controls, indicating possible involvement of HIF2α in hypoxia tolerance during evolution over many generations [65]. It is interesting to note that several studies have demonstrated the interaction between Notch and HIF under hypoxic condition [66, 67, 68, 69]. In the current study, in addition to proving that Notch is critical for hypoxia tolerance, we have also demonstrated with our newly generated lines overexpressing NICD that the role of Notch in hypoxia tolerance requires its downstream gene E(spl)mα. A previous study has shown that E(spl)mα is activated by Notch and functions in a negative feedback loop to accurately adjust Notch signaling [70], which may be important in distinguishing a cell survival property of Notch from its function of inhibiting cell differentiation. Accumulating evidence demonstrates that hypoxia induces Notch activation in mammals, including humans. For example, chronic constant hypoxia up-regulates Notch1 expression in mouse heart [71] and Notch1 regulates melanoma development by protecting cells from hypoxia-induced cell death [72]. Furthermore, Notch3 plays a major role in the development of hypoxic pulmonary hypertension [73] in both humans and rodents. The role of Notch during hypoxia discovered in this study is bound to be important not only in Drosophila but also in humans, raising the distinct possibility of Notch signaling as a potential target for translational and therapeutic strategies.

4.5 Materials and Methods

4.5.1 *Drosophila* stocks and culture

To generate the hypoxia-tolerant Drosophila strain, we pooled 27 wild-type isogenic lines (kindly provided by Andrew Davis) to form the parental population

and subjected them to long-term laboratory selection [35]. Certain interparental genetic variations of hypoxia tolerance were determined in the parental lines that include a signicant variation in eclosion rates under hypoxia and recovery time from anoxic stupor [35]. F1 embryos of this pooled population were collected and cultured at different levels of hypoxia (8%, 6%, or 4% O_2). We found that 6% O_2 dramatically decreased their survival rate, and 4% O_2 was lethal. Under 8% O_2, the majority of the embryos (>80%) completed their development and reached the adult stage. Therefore, we initiated the hypoxia-selection experiment at 8% O_2, and this O_2 concentration was gradually decreased by ~1% every 35 generations to maintain selection pressure. After >30 generations of selection, we obtained flies that break through the lethal hypoxia limit and tolerate 4% of O_2 perpetually. To test whether this hypoxia-tolerant trait is heritable, a subset of embryos obtained from the hypoxia-selected flies was collected and cultured under normoxia for several consecutive generations and then reintroduced back into the same hypoxic environment (i.e., 4% O_2); again, the majority (>80%) of these flies completed their development and could be maintained in this extreme condition perpetually. One hundred male and 100 female flies from the 180th generation of hypoxia-selected or control populations were collected and genomic DNA was isolated for sequencing analyses.

The UAS-NICD and 4XSu(H)-lacZ stocks were provided by J. Posakony. All GAL4 driver lines and N[Ax-tsl], N[fa-1], and N[spl-1] mutants were obtained from Bloomington Drosophila Stock Center at Indiana University. Drosophila stocks were cultured on standard cornmeal/yeast media. To assay for cell type specificity of NICD overexpression homozygous UAS-NICD flies were crossed to Eaat1-GAL4 flies. To test for NICD transcriptional up-regulation, double-homozygous (E/E; N/N) males were crossed to 4×Su(H)-lacZ virgin females. The UAS-RNAi-mα stock was obtained from the Vienna Stock Collection.

4.5.2 Whole-genome resequencing

Genomic DNA was isolated from a pool of 100 male and 100 female adult flies collected from hypoxia-selected populations or generation-matched control

populations by standard phenol:chloroform extraction followed by treatment with DNase-free RNase. DNA quality was assessed by spectrophotometry (260/280 and 260/230) and gel electrophoresis. A total of 3 μg was sheared DNA (Covaris) and was used to construct a library for paired-end sequencing. The DNA fragments were subjected to end repair using the End-IT DNA End-Repair Kit (Epicentre) and then ligated to Illumina PE adapters. The adapter-ligated products were purified on Qiaquick spin columns (Qiagen) and PCR amplified with high-fidelity DNA Polymerase in 12 cycles using Illuminas PE primer set. Cluster generation was performed using the Illumina cluster station and cluster generation kit v2. The 54 + 54 paired-end sequencing was performed using genome analyzer II (Illumina) and sequencing kit v3. The fluorescent images were processed to sequences using the Illumina base-calling pipeline (GA Pipeline-1.4.0). The D. melanogaster reference genome, together with the annotation of genes and repeats, was downloaded from the University of California (Santa Cruz, CA) (UCSC) database (http://genome.ucsc.edu/).

4.5.3 Data analysis

The next-generation sequencing data for each of the pools were derived from 200 flies descended from 27 parental strains. Neither individual genotypes nor the number of individuals sampled at a region could be determined, precluding use of standard analysis tools to identify differences between control and hypoxia-tolerant populations. We therefore identified interesting genomic regions using allelic frequencies that differed between control and hypoxia-tolerant flies, representing regions of potential selection. This analysis used Maq v.0.7.1 [74] under its default parameters to map reads from the four populations (H1, H2, C1, and C2) to the D. melanogaster reference genome downloaded from FlyBase (http://www.flybase.org).

4.5.4 Test for selection in pooled data

Common tests of selection primarily work through measuring loss of diversity in selected haplotypes [15, 17]. In pooled samples, we do not have such data available — for a SNP, the only measure of diversity we can calculate is the SNP's frequency. From the raw mappings to D. melanogaster reference release 5.23, we performed a few more processing steps to generate accurate frequencies. We used Maq's rmdup to remove duplicate reads and Maq's cns2snp to identify variant positions. To calculate frequencies, we used a method developed by Holt et al. [75] that measures the frequency of an allele as the fraction of reads covering the locus showing that allele, weighted by their quality scores. We viewed the reference sequence as an outgroup, and thus, any deviation from the reference as a derived allele. To ensure robust frequency estimates, we considered only sites with at least 10× coverage of reads that had a mapping quality of at least 40. This meant that only 93,306,140 loci in H1, 79,726,429 loci in H2, 76,569,897 loci in C1, and 40,708,770 loci in C2 were considered as having enough information to accurately gauge allelic frequencies. We considered only derived alleles with frequencies between 10% and 90% as being polymorphic. This determination resulted in 292,410 SNPs in H1, 288,952 SNPs in H2, 274,200 SNPs in C1, and 97,895 SNPs in C2.

4.5.5 DAPT treatment

Adult flies were collected from each control and hypoxia-selection chamber (10 vials per chamber, 10 flies per vial, 5 vials of male, and 5 vials of female) (n = 100 for each chamber). DAPT was dissolved with ethanol and diluted with 5% sucrose solution to reach a final experimental concentration of 0.25 or 0.50 mM (final ethanol concentration <1%). Five percent sucrose with 1% ethanol was used as a control. The DAPT or control solution was applied in 150 μL on a filter paper in each vial every other day. The dead flies were counted every 24 h to determine life span. The median and maximum life spans were calculated using GraphPad Prism 4 (GraphPad Software), and the statistical significance was calculated by Student's t test.

4.5.6 Hypoxia tolerance and vulnerability tests

The survival rate of Notch mutants in hypoxia was determined by culturing them in a computer-controlled environmental chamber. After 3 wk in culture, the numbers of eclosed and total pupae were counted. The ratio between elcosed pupae and the total number of pupae was calculated and presented as eclosion rate. The UAS-NICD stock was crossed with specific GAL4 transgenic flies to determine the effect of specific spatiotemporal NICD overexpression on hypoxia tolerance. Each cross contained 10 virgin female homozygous UAS-NICD flies and 5 male homozygous GAL4 transgenic flies and allowed them to lay eggs for 48 h in normoxia. The flies were then moved to a control vial (for another 48 h before being discarded) and the vial with the eggs was moved to a 5% oxygen chamber with a 12-h dark and 12-h light cycle with a temperature of 22 $\pm$ 1 °C. In parallel, the parental line without the cross wasere tested in normoxia as a control. After 4 wk, both sets of flies were assayed for the number of pupal cases that were empty or full. Six vials of each condition were completed in two different experiments for a minimum of 500 pupal cases scored to calculate the eclosion rate for each condition. Adult survival during hypoxia was evaluated by counting the number of newly eclosed experimental and control flies and transferring them to a new individual vial. The following day, the number of adult dead flies in the adult-only vial was subtracted from the previous day's total and the newly eclosed flies from the original vial were counted and added to the adult-only vial to avoid a cumbersome number of vials for all tests done. Hence, the number of flies in the adult-only vial could go down (due to adult death), but later increase (due to the addition of newly eclosed adults from the original vial to the adult-only vial).

The statistical significance of eclosion rate between mutants, NICD- over-expressed flies, and controls was calculated by unpaired t test.

4.6 Acknowledgments

We thank O. Gavrialov, M. Y. Hsiao, Y. Lu-Bo, J. Wang, and N. Morgan for technical assistance. Confocal microscopy was performed in the University of

California at San Diego Neuroscience Microscopy Shared Facility (P30 NS047101). This work was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (5P01HD032573); the National Institute of Neurological Disorders and Stroke (5R01NS037756); the National Heart, Lung, and Blood Institute (5R33HL087375); the National Human Genome Research Institute (5R01HG004962); the National Science Foundation (DBI-0641037 and NSF-III-#0810905); and an American Heart Association Award (0835188N).

Chapter 4 is in full adapted from *Zhou D, Udpa N*, Gersten M*, Visk DW*, Bashir A, Xue J, Frazer KA, Posakony JW, Subramaniam S, Bafna V, Haddad GG. Experimental selection of hypoxia-tolerant Drosophila melanogaster. Proc Natl Acad Sci U S A. (2011);108(6):2349-2354.* The dissertation author was a second author of this manuscript, and was responsible for the "coarse-grained" computational analysis.

Chapter 5

Evolutionarily conserved hypoxia tolerance genes in Ethiopian highlanders

5.1 Abstract

Although it has long been proposed that there are genetic factors that contribute to adaptation or mal-adaptation to high altitude, these have remained largely unproven. Recent advances in high-throughput sequencing technologies have made it feasible to analyze patterns of genetic variation across the genome in human populations. Since these studies only surveyed a small portion of the genome (i.e., exons and flanking regions) or a focused group of candidate genes, the interpretation of these results is believed to be limited. We have focused our studies on Ethiopian populations since these have been previously found to be well adapted to high altitudes. We report here our results of the first whole genome re-sequencing-based analysis identifying genes that can modulate high altitude adaptation of native Ethiopian human residents (living at ~3500m on Bale Mountain or Chennek plateau in Ethiopia). We used cross-population tests to identify regions with significant loss of diversity, indicative of a selective sweep. Only one region was significant for both the sampled Ethiopian populations. This 208kbp

gene-rich region (located on chromosome 19) contains 8 protein coding genes and spans 135 SNPs. We then tested whether particular genes discovered in these individuals played a role in hypoxia tolerance in Drosophila. Three genes were found to improve the tolerance of flies to low oxygen by 2 to 4 fold: cic (ortholog of human CIC), CG11055 (ortholog of human LIPE) and Paf-AHalpha (ortholog of human PAFAH1B3). This study reveals evolutionary conserved genes that can modulate hypoxia tolerance. In addition, many of the results could not be found using exome sequencing or microarray-based studies, highlighting the importance of whole genome sequencing.

5.2 Introduction

Humans have occupied high altitude regions for thousands of years [76]. It is estimated that now more than 140 million people live and work at altitudes above 2500 m [77]. The hypoxic conditions prevalent at high altitudes present a challenge for survival. Previous studies have shown that the three large high altitude populations (i.e., the Andeans, the Himalayans and the Ethiopians) have each adapted uniquely to cope with their inhospitable hypoxic environments [78, 79]. It has also been suggested that the Ethiopians are the best adapted for life under such conditions since they show the least evidence of chronic mountain sickness (CMS), a mal-adaptation syndrome to high-altitude life that exists in other populations, especially the Andeans [80, 81, 79]. Although it has long been proposed that there are genetic factors that contribute to adaptation or mal-adaptation to high altitude, these have remained largely unknown or unproven [82]. Recent advances in high-throughput sequencing technologies have made it feasible to analyze patterns of genetic variation across the genome in human populations. To date, two sequencing-based partial genomic analyses have been performed in high altitude populations [83, 65]. Since these studies only surveyed a small portion of the genome (i.e., exons and flanking regions) or a focused group of candidate genes, the interpretation of these results is limited.

We have previously performed the first whole genome-based, unbiased anal-

ysis of genes that contributed to hypoxia adaptation in a hypoxia-tolerant Drosophila strain, which was generated through laboratory evolution (see Chapter 4). This study allowed us to examine the complete genome at single nucleotide resolution, and to detect fine changes in the allele frequency spectrum consistent with natural selection. We extend the analytical strategy developed in our previous study, and present here the results of a whole genome resequencing-based analysis identifying genes that modulate high altitude adaptation in humans. We focused our study on 17 high altitude (~3500m) native Ethiopian residents, as Ethiopian highlander populations have been found to be well adapted to high altitudes [78, 79]. Specifically, we sequenced 10 individuals of Oromo heritage living on Bale Mountain (labeled Oromos), and 7 individuals residing on the Chennek Plateau in the Simen Mountains (labeled Amhara). Furthermore, we tested whether particular genes highlighted by our study play a role in adaptation to low O_2 conditions, by using RNAi to target their respective orthologs in Drosophila.

5.3 Results

We sequenced the whole genome of each individual using Illumina's HiSeq 2000 platform to a mean genome-wide depth of ~18× per individual. We mapped the reads to the hg19 human reference using BWA [84], and performed variant calling using the GATK pipeline [85, 86] (Figure 5.1). We used ADMIXTURE analysis [87] to identify the closest populations from the 1000 Genomes Project, release 20100804 [88]. This analysis (Figure 5.2) shows that our Ethiopian populations share common genetic ancestry, and are largely an admixture of two ancestral groups. Over half of the ancestry shows highest similarity to African populations, particularly the Luhya (LWK), located in neighboring Kenya. The remainder is largely shared with individuals of non-Finnish, European ancestry. As a result, for lowlander controls, we used variant calls from low coverage whole-genome sequencing of 67 Luhya (LWK). For an out-group we used 90 northern European ancestry (CEU) individuals. Due to differences in coverage between the control populations and our Ethiopian sequence data, we filtered low coverage or poor quality loci prior

to testing for selection (see Methods).

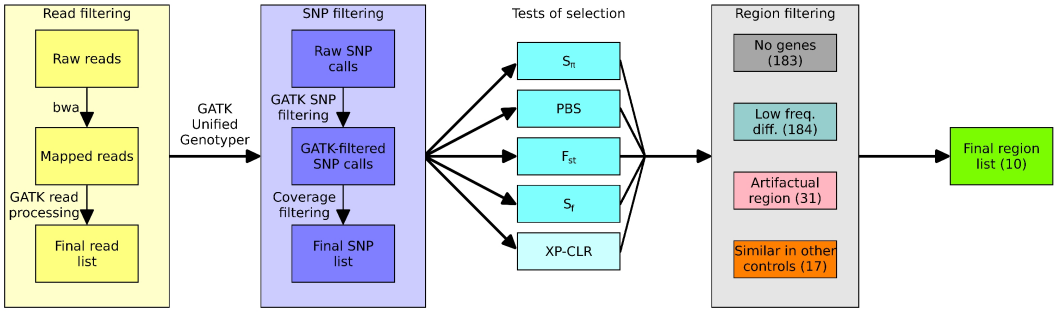


Figure 5.1: **Computational analysis work flow.** The raw reads were mapped using BWA, followed by indel realignment, duplicate marking, and quality score recalibration using GATK's pipeline. Variants were then called and filtered using UnifiedGenotyper (also from GATK). After applying additional variant filters to account for the differences in coverage between the case and control populations, we ran several complementary tests to identify 425 regions as candidates for positive selection. 415 of these were filtered using four filters customized to the challenges of our sequencing framework, leading to ten prioritized regions.

Hypoxic stress manifests as a decrease in survival rates of average, unadapted individuals. Therefore, alleles that confer an adaptive advantage will increase in frequency (along with their linked neighbors), a process known as a "selective sweep". We searched for regions with evidence of such a sweep: a loss of genetic diversity in the region and a corresponding decrease in the scaled mutation rate, $\theta(= 4N\mu)$. We computed four cross-population test statistics (denoted S_{f}, S_π, F_{st}, and PBS) that measure this loss in diversity. The first three tests have been defined in Chapter 3, while the PBS statistic is a three-population variant of F_{st} that can thus distinguish selection on cases from selection on controls (see Methods). Cross-population tests provide a control for locus-specific variability in local mutation rates, enabling a direct comparison of the effective population size as a measure of selection (S_{f} and S_π). They also allow for an estimation of branch lengths and bottlenecks relative to the point of divergence between populations (F_{st} and PBS). Through extensive simulations, we showed that the power of these tests varies depending upon the selection coefficient and time since selection

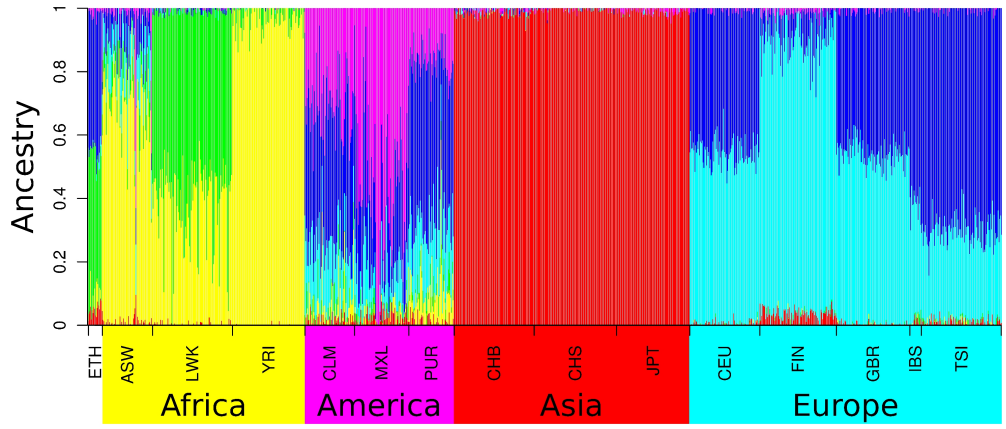


Figure 5.2: **ADMIXTURE analysis with 6 clusters on the Ethiopian highlanders, along with the 1000 Genomes populations.** The highlanders ancestry is a mixture of traditionally African and traditionally European genotypes, represented by the green and dark blue segments, respectively. Within the African 1000 Genomes populations, the nearest population geographically as well as ancestrally appears to be the Luhya (LWK) population. We thus selected this population as our control. Similarly, the section sharing ancestry with European populations appears closer to the southern and western Europeans than the Finnish population. As a result, as outgroup for the PBS test, we selected the CEU population.

(Figure 5.3). As these parameters are unknown, we considered regions that were significant in any of these tests.

We initially identified 425 regions spanning 36.9Mbp that were significant in at least one test under a 0.1% genome-wide false discovery rate. Due to our relatively small samples, our SNP frequency estimates have increased variance (Figure 5.4). Consequently, we filtered out regions that did not show evidence of testable, strong selection. Specifically, we removed 183 regions lacking known genes, and subsequently, 184 regions that did not show a pronounced ($\geq 20\%$) increase in the dominant haplotype frequency (see Methods). Additionally, since we used low coverage 1000 Genomes populations as our controls, we filtered out 31 regions with artifacts in variant calling that escaped our variant filtration steps. Finally, to ensure our regions represent selection on a Amhara/Oromos only phenotype, we

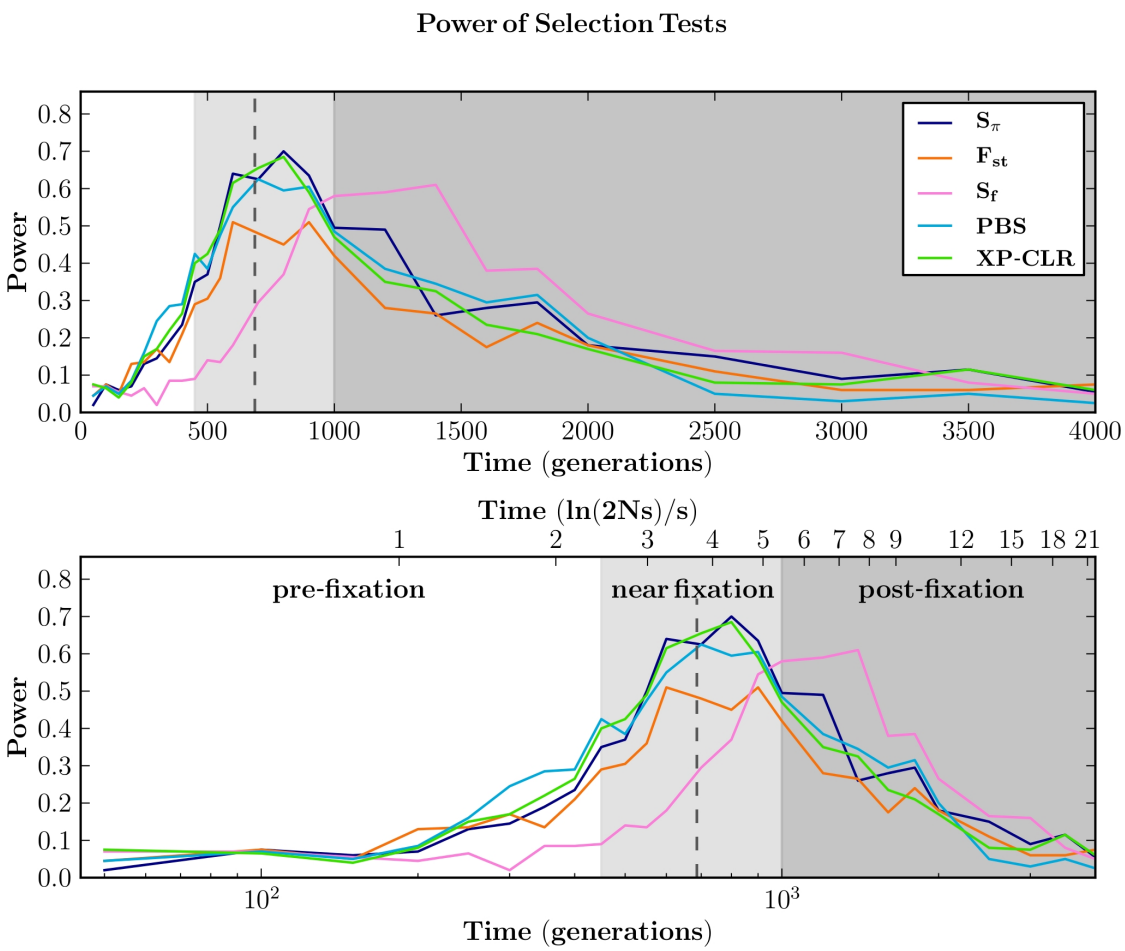


Figure 5.3: **Power as function of time for neutrality tests used in this study (S_π, F_{st}, S_f, and PBS).** A) The x-axis scales linearly in terms of generations since selection start. B) Power as function of logarithmically scaled time for the neutrality tests used in this study (S_π, F_{st}, S_f, and PBS). We also show the x-axis in units of $\ln(2Ns)/s$ (top axis), which can define the regimes as a function of selection pressure. We observe three major regimes, corresponding to the state of the beneficial haplotype in the case population: before the haplotype has significantly risen in frequency ("pre-fixation"), as the haplotype dominates the case population ("near fixation"), and after the haplotype has gone to fixation, while the frequency spectrum gradually reverts to neutrality ("post-fixation".

filtered out 17 regions showing similar variant frequencies in other 1000 Genomes and dbSNP control populations. Ten regions remained after these filtration steps (Table 5.1).

Of the ten regions, only one showed up as significant in both the Amhara

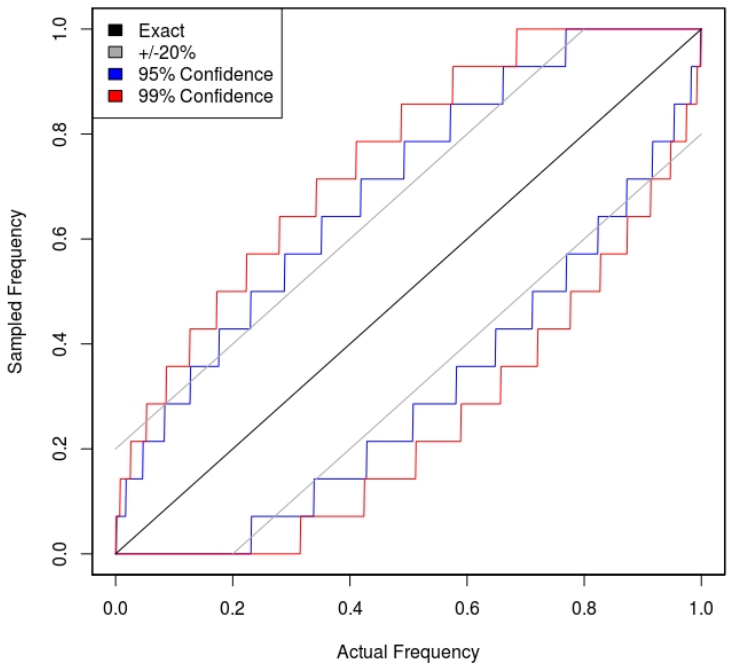


Figure 5.4: **The impact of sampling $n = 14$ haplotypes (the sample size of our Amhara population) from a population on estimated allele frequencies.** For a given intermediate frequency, a difference of $\approx 20\%$ is within the 95% confidence interval. As a results, we use this frequency difference as a cutoff, prioritizing regions containing haplotype blocks with a greater frequency differential between the case and the control populations.

and the Oromos populations. This 208kbp gene-rich region on chromosome 19 contains a block of 135 "*differential*" SNPs with a significant change in frequency relative to the control populations (Figure 5.5). Specifically, the mean frequencies are 48% (Oromos), 42% (Amhara), 16% (LWK), and 1% (CEU). The 8 genes in the region (Table 5.1) point to many intriguing candidates. For example, the differential SNPs include two missense mutations in the LIPE gene (rs7246232 and rs16975750). While these mutations have not previously been linked to a known phenotype, LIPE is associated with gestational hypertension (and consequent placental ischemia) [89]. It belongs to the lipase family, which is known to play a role in hypoxia via lipolysis, triglyceride metabolism, and energy storage [90]. Other genes

in this region include CIC, which is a transcriptional suppressor involved in early organ development, CNFN (associated with hematopoiesis [91]), CXCL17 (associated with angiogenesis [92]), and PAFAH1B3 (associated with coronary artery disease [93] and organ development [94]). Thus, our results point to a cluster of putative hypoxia response genes. As these genes are associated with phenotypes such as lipid metabolism, transcription regulation, or angiogenesis, they illustrate the potential for a variety of adaptive mechanisms to high altitude in humans.

The remaining nine regions contained several other gene candidates (Table 5.1), including genes linked to blood clotting, heart failure and hypoxia response. For instance, the 110kbp region on chromosome 13 that is significant for the Amhara population in the S_π test contains Endothelin receptor B (EDNRB or ETB) gene (Figure 5.6). This gene encodes a receptor for Endothelin, a potent vasoactive peptide, which activates signaling cascades that promote blood vessel constriction [95]. It is a known target for drugs (e.g., Bosentan) prescribed for altitude sickness [96]. In the Amhara population, this gene has 52 fixed, or near-fixed, SNPs (spanning ~170kbp) upstream of the promoter region, of which 20 are in a 10kbp region containing several transcription-factor binding sites (Dataset S1). This haplotype block is also present in the controls, at 36% frequency in LWK and 66% frequency in CEU.

To provide further evidence of the role of these genes in hypoxia, we used Drosophila as a model organism to test the hypothesis that the differential regulation of expression of the orthologs in flies has an effect on tolerance or susceptibility to survival in low O_2 conditions. The fixation (or near fixation) of SNP variations in the candidate genes may cause either gain- or loss-of-function changes. For conceptual reasons (up-regulation of genes could be problematic if the gene is not expressed in a particular tissue) and practical reasons, we first used the UAS-RNAi/GAL4 system in Drosophila to analyze whether the down-regulation/knockdown of the Drosophila orthologs representing the human candidate genes located in the selected region on chromosome 19 affects hypoxia tolerance. When the fly orthologs were tested, three genes were found to improve markedly the tolerance of flies to low oxygen. These genes were cic (ortholog

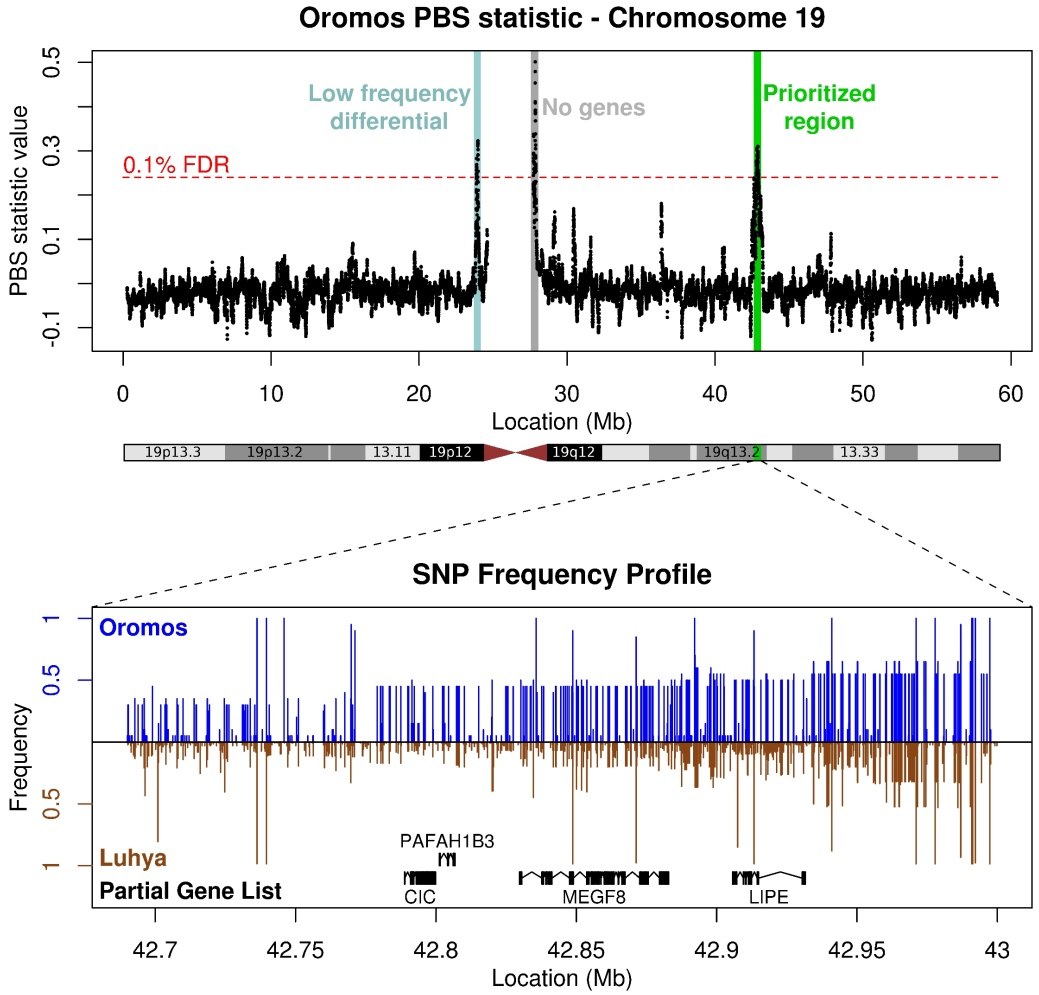


Figure 5.5: **PBS statistic across chromosome 19 in the Oromos population compared to both Luhya (LWK) and European (CEU) populations** The red line represents a genome-wide, 0.1% FDR. Three distinct regions exceed this cutoff, two of which are near the centromere and were thus filtered out. Below is the SNP frequency profile of the remaining chromosome 19 region in the Oromos (blue) and LWK (brown, inverted) populations. Some relevant genes in the region are shown in black below the frequency profiles. As can be seen, variant frequencies in this region are much higher in the highlander populations than in the lowlander populations.

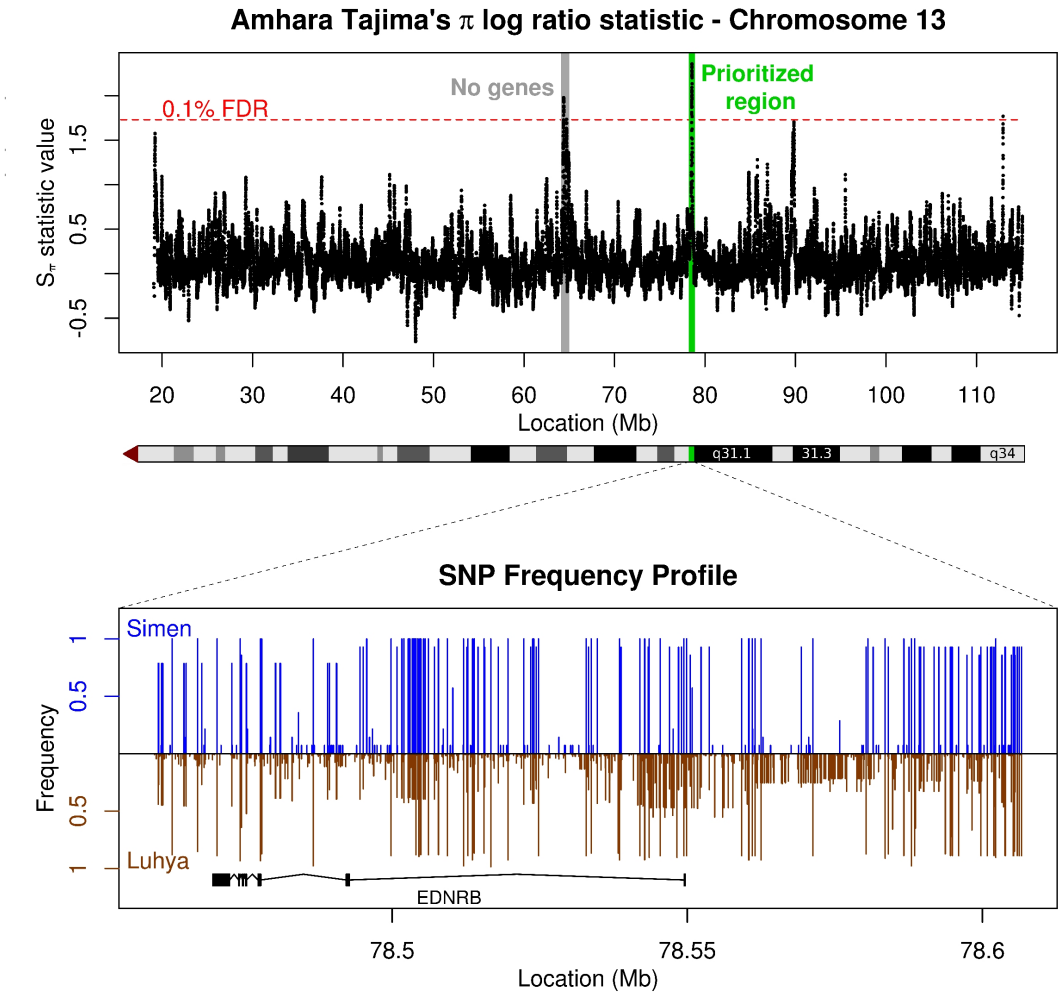


Figure 5.6: **Evidence supporting EDNRB as a gene candidate.** A) S_π statistic values across chromosome 13 in the Amhara population, compared to the Luhya (LWK) population. The red line represents a genome-wide, 0.1% FDR. Two distinct regions exceed this cutoff, one of which contains no genes and was thus filtered out. B) SNP frequency profile of the significant region in the Amhara (blue) compared to Luhya (brown, inverted) populations. As can be seen, variant frequencies in this region are much higher in the Amhara population than in the lowlander controls.

of human CIC), CG11055 (ortholog of human LIPE) and CG8962/Paf-AHalpha (ortholog of human PAFAH1B3). We found that there was an increase in survival and this varied from about 40% to 80%, hence constituting 2 to 4 fold in-

Table 5.1: **List of significant genomic regions found in Amhara and Oromos populations**

	Region	S	O	Tests	Genes Located in the Region
1	chr1:27.9M		✓	PBS	AHDC1, FGR
2	chr2:26.4M	✓		Fst, PBS	FAM59B, HADHA, HADHB
3	chr4:17.6M	✓		Fst, PBS	GPM6A
4	chr11:61.1M		✓	PBS	VWCE, DDB1, DAK, CYBASC3, TMEM138
5	chr11:87.0M		✓	Fst	TMEM135
6	chr13:78.5M	✓		S_π	EDNRB
7	chr18:51.9M		✓	Fst	STARD6, C18orf54
8	chr19:42.8M	✓	✓	PBS	CIC*, PAFAH1B3*, PRR19, CNFN, TMEM145, MEGF8, LIPE*, CXCL17
9	chrX:45.0M		✓	S_π	CXorf36
10	chrX:118.4M		✓	PBS	PGRMC1

*Genes regulating hypoxia tolerance with experimental validation in Drosophila

crease in survival rate over control flies in the same hypoxic environment (Figure 5.7). CIC has been shown to function as a repressor of receptor tyrosine kinase (RTK)-responsive genes. Following activation of RTK signaling, CIC repression is removed, allowing the expression of targeted genes downstream. CIC is well conserved from Drosophila to humans and is mostly known in determining cell fate and cell proliferation [97, 98]. Of interest is that there is a cross-talk between RTK and Notch pathways, including core components of the RTK pathway and other major pathways such as TGFb, Jak/Stat and Wnt [99]. This is remarkable as we had previously shown that the Notch pathway is crucial for hypoxia tolerance in Drosophila (Chapter 4). LIPE, a hormone-sensitive lipase, is important in lipolysis and in mobilization of fatty acids and glycerol from fat cells. PAF, a platelet-activating factor is a potent lipid mediator and is involved in a variety of physiologic events. Its deacetylation induces a loss of activity that is catalyzed by PAF-AH, a platelet-activating factor acetylhydrolase. Type I PAF-AH has 2 subunits (alpha and beta) and plays a role in cellular functions such as induction of nuclear movement and control of microtubule organization.

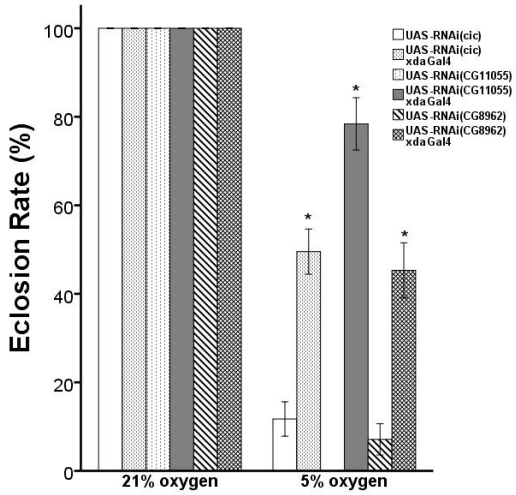


Figure 5.7: **RNAi-mediated knockdown of orthologs of candidate genes enhanced hypoxia tolerance in Drosophila.** The available UAS-RNAi lines for cic (Drosophila ortholog of human gene CIC), CG11055 (Drosophila ortholog of human gene LIPE) and CG8962 (Drosophila ortholog of human gene PAFAH1B3) were crossed with the daughterless (da)-GAL4, a driver strain that expresses GAL4 ubiquitously. The level of hypoxia tolerance was determined by measuring eclosion rate in atmosphere chamber containing 5% O_2. The UAS-RNAi stocks without cross were used as negative control (open bars). Two different UAS-RNAi lines targeting each candidate gene were used in each experiment to minimize off-target effects. Each bar represents the mean ± SEM value of 3 separate tests, *: $p<0.05$.

5.4 Discussion

The notion that high altitude adaptation is heritable has dominated high altitude medicine for some time. However, it was only in recent years that attempts to identify the genetic basis of this adaptation have been made. These studies used genotyping or exome sequencing, but not whole genome sequencing (WGS). The relatively sparse sampling of the genome makes it harder to identify large-scale shifts in the allele frequency spectrum associated with natural selection. Consequently, these studies focused on variants in candidate genes. Moore et al. reported one of the first genome-wide scans for selection using a set over 11,000 SNPs genome-wide [100]. They identified a subset of variants that contained SNPs located in genes involved in the hypoxia inducible factor (HIF) pathway, which included nitric oxide synthase 2, alpha-1B-adrenergic receptor,

Endothelin 1, and HIF-prolyl hydroxylase 3. An extended analysis was also reported and, again, a subset of candidate genes involved in the HIF pathway was identified that included vascular endothelial growth factor, tenascin C, cadherin 1, endothelin receptor A, and EGLN1, which is involved in red blood cell production through the down-regulation of HIF targets, including EPO. Other studies have also included Tibetans using tests of neutrality based on extended haplotype homozygosity (iHS) and cross-population extended haplotype homozygosity testing (XP-EHH) [101]. These authors focused loci for EGLN1, heme oxygenase 2, and peroxisome proliferator-activated receptor alpha and PTEN. Similarly, Beall et al. identified a signal of positive selection in a sample of high-altitude Tibetans at the EPAS1 locus [102]. Similar work was also carried out in Ethiopia [103] and a set of candidate genes was identified in the high altitude population, including CBARA1, VAV3, ARNT2 and THRB. Although most of these genes have not been identified in previous studies in the high altitude Tibetan or Andean populations, two of them (THRB and ARNT2) play a role in the HIF-1 pathway, a pathway implicated in the previous Tibetan and Andean studies.

There is an important trade-off when comparing WGS to exome sequencing or genotyping studies. Specifically, WGS is usually performed on a much smaller number of individuals, but provides for a near-complete sampling of variant sites. For selection signatures, this is critical. For instance, consider the high frequency block found near EDNRB gene (Figure 5.8). With WGS, this region corresponds to the highest peak in the chromosome, with a block of 52 variants that are fixed in Amhara, but only 36% in LWK. However, the Nimblegen (Madison, WI) 2.1M exon capture array targets only two high-frequency variants in this region, none within the block. As for genotyping, the ~1M Affymetrix (Santa Clara, CA) Genome-Wide Human SNP Array 6.0 only samples 2 of the 52 sites in the block, resulting in a much weaker signal (see Methods). A similar argument holds for every region identified in our study, where we find clear and strong peaks for our tests at a low genomic FDR.

The drawback of sampling fewer individuals is that observed frequency differences may arise from sampling. To account for this, we determined the sampling

variance of a SNP at a given frequency that can be expected when sampling 14-20 haplotypes from a population (Figure 5.4). When sampling 7 individuals (i.e. 14 haplotypes), haplotype frequency changes greater than 20% can be detected with high confidence. Thus, we prioritized regions showing frequency differential $\geq 20\%$ between case and control. Despite this limitation, simulations show that our test statistics achieve between 67-95% power on 14-20 haplotypes, compared to a much larger sample of 400 haplotypes (Figure 5.9). This indicates that sampling 14-20 haplotypes is sufficient for capturing features of selection.

Our study identifies a number of candidate genes for hypoxia tolerance, which were not previously reported. To further validate our approach, we tested the impact that down regulating our candidate genes (using RNAi knock-down) has on hypoxia tolerance in a Drosophila model system. Several orthologs, when knocked-down, lead to a much higher eclosion rate (2 to 4 fold) in low O_2 conditions relative to controls. Thus, our study demonstrates that genes identified from WGS in humans indeed affect hypoxia tolerance in a model system. This study provides evidence for the importance of these genes to survival in low O_2 conditions and lends further credence to our analysis.

5.5 Materials and Methods

5.5.1 Ethics statement

The study was approved by the institutional review board and the ethics committee of each institution. Written informed consent was obtained from each participant in accordance with institutional requirements and the Declaration of Helsinki Principles. All subjects in the study are over 18 years of age, and gave informed consent at each examination. Study protocols were approved by the Institutional Review Board of the University of California at San Diego.

5.5.2 Sample description

Ten Oromos subjects from the South and seven Amhara subjects from the north of Ethiopia were chosen to reflect differences in ancestral adaptation to life at high altitude. The Oromos generally have a darker skin color and a less slender build. They appear more muscular and are generally shorter than the Amhara people. They have lived at high altitude for 600 to 700 years, a much shorter time as compared to the millennia of the Amhara people [104]. The subjects were examined and a history was taken. Only males who are free of disease were selected. After informed written consent (on file) given in their local language and obtained by local physicians, venous blood was obtained in the field, stored, and transported in suitable containers to allow extraction of sufficient quantities of DNA from both study populations.

5.5.3 DNA extraction, library construction and sequencing

Genomic DNA was isolated using Blood DNA extraction kit (Qiagen, Valencia, CA) and randomly fragmented. Fragments of the desired length were gel-purified. Adapter ligation and DNA cluster preparation were performed using the library preparation kit according to manufacturer's instruction (Illumina, San Diego, CA). Whole genome sequencing was performed using Illumina's HiSeq 2000 platform on all 17 individuals to a mean, per-sample depth of $\approx 18\times$.

5.5.4 Read alignment, score recalibration and variant calling

We aligned the reads to the human reference genome (hg19) using BWA [84] with default parameter settings. We adjusted the alignments using the GATK indel-realignment, the Picard read duplicate marking, and the GATK quality score recalibration modules [85, 86] under default parameter settings, as defined by the GATK manual (version 2). We finally called and filtered the SNPs using the GATK UnifiedGenotyper tool under default settings. The sequencing was free

of any mapping bias in coverage, mapping percentage, or variant counts for all individuals. As an independent test, we also identified variants using the SoapSNP pipeline [105]. The SoapSNP variants were generally a super-set of the GATK variants, with 25% more calls (9,508,898 vs. 7,594,936 for Amhara, and 10,284,853 vs. 8,144,023 for Oromos). This is attributed mainly to less restrictive filtering.

5.5.5 Variant filters

The coverage difference between the cases ($\approx 20\times$) and controls ($\approx 4\times$) led to differences in processing the called variants. To adjust for these differences, we filtered our call set using three steps. First, we observed several variants in clustered genomic loci that were discarded by the variant caller in the case populations. This happens due to various sequencing and mapping artifacts, such as strand bias, low sequence complexity, or structural variations. Due to the low coverage, variants in these loci are not always discarded in the controls. We thus removed from consideration any region comprising 10 consecutive SNPs that were filtered out using GATK in our case population. Second, following the protocol used by the 1000 Genomes project, we filtered out any site with a mean coverage higher than twice the median genome-wide coverage as being likely to be caused by a duplication [88]. This removes variants found in repetitive regions, such as centromeric sequence. We also filtered out any site with less than $2\times$ coverage per person in the case population as being too poorly covered to accurately call SNPs. Finally, we removed sites that had an excess of heterozygotes, compared to expectations from Hardy-Weinberg equilibrium. We tested this using a test from Emigh [106] describing the heterozygote probability as:

$$P_{Aa} = \frac{n!}{n_{AA}!\, n_{Aa}!\, n_{aa}!} \times \frac{n_A!\, n_a!}{(2n)!} \times 2^{n_{Aa}}$$

Variants with P-value under 0.05 were discarded. After filtering variants based on the three filters described above, we remained with 7,555,907 SNPs in the Amhara population and 8,069,425 SNPs in the Oromos population. See Figure 5.1 for an overview of the computational workflow.

5.5.6 Control populations

To identify appropriate controls, we used low-coverage whole-genome sequencing calls taken from the 1000 Genomes Project [88] populations. We ran the ADMIXTURE program [87] on 13,928 sampled sites to identify the population most closely related to the highlanders. As seen in Figure 5.2, the Ethiopian individuals consist largely of African ancestry, but possess a more substantial European component compared to any other African population considered. The closest population consists of 67 Luhya (LWK) individuals from Webuye, Kenya, and was thus chosen as the control for all cross-population tests of selection (see below). As an outgroup for the PBS test (see below), we used 90 European (CEU) individuals in order to capture variation in the highlanders shared with individuals of European ancestry.

5.5.7 Identifying regions under positive selection

Under positive selection, haplotypes containing the beneficial mutation (as well as linked, neutral mutations hitchhiking along) rapidly increase in frequency, leading to a loss of genetic diversity in the region surrounding the mutation [107] (see illustration in Figure 5.10). This loss of diversity, or selective sweep, decreases with distance from the beneficial mutation due to recombination. The loss of allelic diversity and the corresponding skew in the allele frequency spectrum can be used to detect loci important for adaptation to the selective stress [107]. We use cross population tests to adjust for interesting frequency profiles that are shared between our case and control populations. These are likely due to events (such as bottlenecks, genetic drift, or even selection for a different phenotype) occurring before our case and control populations separated, and thus, are unlikely to be causal for hypoxia tolerance. Thus, population-specific selection can be measured by comparing the estimated scaled mutation rate $\theta = 4N_0\mu$ in a given region to that of a control region. For a region, a large decrease in θ in our case population compared to our controls indicates a non-neutrally evolving case population, consistent with positive selection.

5.5.8 Tests of selection

First, we ran two cross-population tests comparing the Amhara or Oromos populations (case) against the 1000 Genomes Luhya population (control). These tests are based on two common estimators of θ: the summed non-fixed frequencies estimator, denoted θ_{f}, and the average pairwise heterozygosity estimator, denoted θ_{π} [24]. For a given region, a high log ratio of θ_{π} (θ_{f}) in the control relative to the case population is indicative of selection (Chapter 3). We label these log ratio statistics as S_{π} for the average heterozygosity estimator and S_{f} for the summed frequency estimator, such that:

$$S_{\mathrm{f}} = \log\left[\frac{\theta_{\mathrm{f,\ control}}}{\theta_{\mathrm{f,\ case}}}\right] \qquad S_{\pi} = \log\left[\frac{\theta_{\pi,\ \mathrm{control}}}{\theta_{\pi,\ \mathrm{case}}}\right]$$

Another class of tests for selection is based on the fixation index, or F_{st}, between two populations [108]. This class aggregates differential SNP frequencies across two populations. For instance, Hudson [28] defines this measure as:

$$F_{\mathrm{st}} = 1 - \frac{\pi_{\mathrm{within}}}{\pi_{\mathrm{between}}}$$

where π_{within} represents the within-population average heterozygosity and π_{between} represents the between-population average heterozygosity. As two populations diverge, the variability between the populations increases much more than the variability within each the population, and the statistic approaches one. The fixation index roughly correlates to the evolutionary branch length T between two populations [109] as:

$$T = -\log(1 - F_{\mathrm{st}})$$

This approach is not directional, however. As a results, a significant statistic value may indicate a selective sweep in either the case or the control population. To address this, Shriver et al. [110] and Yi et al. [65] developed the concept of the population branch statistic (PBS). This combines the pairwise branch lengths of three populations as follows:

$$\text{PBS} = \frac{T^{CN} + T^{CO} - T^{NO}}{2}$$

Where C represents a case population, N represents an evolutionarily close control population, and O represents a distant out-group. We calculated the PBS test statistic with our case population defined as either the Amhara or Oromos population, our control as the Luhya population, and our out-group as the CEU population. Additionally, we compared the results of the above tests with XP-CLR [111], a method that attempts to detect large linkage blocks with high frequency differential as indicative of positive selection.

For S_{f}, S_{π}, F_{st}, and PBS, we use genomic windows of size 50 kbp, overlapping at 2 kbp intervals. For each test, we define the top 0.1% genome-wide value as the genomic-control cutoff to determine the windows of interest. For the XP-CLR test statistic, we found that using a 0.1% genome-wide threshold seemed too stringent. When testing Amhara vs. Luhya, using a 0.1% threshold led to exactly 5 non-overlapping regions exceeding the threshold, all of which containing highly repetitive sequence except for the HLA region, which has a high mutation rate. Relaxing the threshold to 0.3% genome-wide yielded a comparable number of regions to that found by our other tests, but since XP-CLR uses variable size genomic windows (normally much larger than the 50 kbp used in the other tests), the list of implicated genes was dominated by XP-CLR results. Hence, we used XP-CLR only for secondary validation. For instance, the EDNRB gene region on chromosome 13 was found to be significant using XP-CLR under a 0.3% threshold.

5.5.9 Population simulations and power estimation

We generated simulated populations using the *mpop* forward simulator [112] and Hudson's *ms* coalescent simulator[113]. For a given set of parameters μ, r, s, τ (mutation rate, recombination rate, selection coefficient and time since selection, respectively) we generated 200 simulated populations. First, we initiated each instance with a source population of $N_0 = 1000$ diploids from a neutral coalescent process, using Hudson's ms simulator. We then sampled with replacement from the source population into three separate populations of size N_0 each, labelled

case, *cont1*, and *cont2*. We evolved these populations separately using the mpop forward simulator, where only the case population had a locus under positive selection. Individuals carrying the advantageous allele had higher likelihood ($1+s$, for a homozygous carrier) to reproduce at each generation. The other populations (cont1 and cont2) continued to evolve neutrally. After τ generations, a random sample ($n = 100$ diploids) was taken from each of the three populations, and cross-population neutrality tests were applied. Genomic regions of size 50 kbp were simulated, with mutation and recombination rates set to $\mu = 2.4 * 10^{-7}$ and $r = 3.784 * 10^{-8}$ per base, per generation. The selection coefficient used for these simulations was $s = 0.02$, and the number of generations since selection τ ranged between $[50, 4000]$.

The power of a given test at 5% False Positive Rate (FPR) was determined as the fraction of test statistic values exceeding a certain cutoff when applied to the case vs. cont1 samples. The cutoff value was set to the top 5% of the null distribution, obtained by applying the same test to the corresponding cont1 vs. cont2 samples.

5.5.10 Power of different tests under varying model parameters

The different tests for selection described above all aim to find regions with marked differences in allele frequencies across case and control populations. However, the specific signal observed is highly influenced by different factors, such as the selection coefficient and the time since selection. In general, the allelic divergence in a region is a function of the local mutation and recombination rates. Under a Wright Fisher model of neutral evolution, the expected distribution of allele frequencies (the site frequency spectrum) is known. Specifically, the expected number of alleles with frequency i, where $i = (1, \ldots, n-1)$, is given by θ/i [24]. Under selective pressure, the site frequency spectrum begins to shift [16]. Initially, the haplotypes carrying the beneficial alleles rapidly increase in frequency, reducing the overall divergence (a so-called selective sweep). As the beneficial allele becomes fixed in the population, the divergence is as its lowest point, and the

signal of selection is the strongest. As time passes, de novo mutations and recombination events gradually restore variability to the region. Initially, there is an increase in low frequency alleles, which then reach intermediate and high frequencies, finally drowning out the selection signal. Thus, there are three major regimes for a population under positive selection: "pre-fixation", where the beneficial haplotype starts to rise in frequency; "near fixation", where the haplotype approaches fixation; and "post-fixation", where the haplotype is fixed in the population, and de novo mutations slowly restore diversity to the population. In these regimes, the four tests show different relative strengths in detecting positive selection (see Figure 5.3). Importantly, although the selection coefficient in the example shown was set to $s = 0.02$, performance of the different tests also varies under different selection pressures, where some tests dominate in weaker selection and others in stronger selection.

$\mathbf{S_f}$ test

The S_f test sums over non-fixed frequencies in a region, effectively down-weighting low frequency alleles [25]. The result of this is that the S_f test is optimized for the post-fixation regime, since in this regime, the bulk of the signal comes from the reduced number of intermediate and high frequency de novo mutations in the case population. As the de novo mutations only approach these frequencies many generations after fixation, S_f is able to detect selection for longer periods of time after fixation has occurred than other tests. In addition, S_f excels at capturing stronger selection pressures, where the beneficial allele goes to fixation relatively quickly and thus, reaches the post-fixation regime at an earlier time. As Figure 5.3 shows, once in the post-fixation regime, the peak power for S_f is sustained for more generations as compared to other tests.

$\mathbf{S_\pi}$ test

The S_π test is similar to S_f in that it is sensitive to strong selection pressures, as well as long times since selection start. However, S_π is based on average heterozygosity, which weights allelic differences identically (in other words, S_π re-

turns the same value if the derived allele is defined as either the major or the minor allele). This essentially folds the frequency spectrum [25], leading to two major benefits. First, an approach such as S_{f} depends heavily on the idea that the ancestral allele is the reference nucleotide. If this is not true, for instance, a variant at 10% frequency can be mistaken for a variant at 90% frequency instead, heavily impacting the statistic value. For a folded spectrum, however, this is not the case. In addition, unlike S_{f}, S_{π} can detect the loss of diversity due to a loss of intermediate frequency alleles, causing it to pick up selection prior to fixation (where there is an abundance of high frequency alleles). However, since it folds the spectrum, S_{π} cannot distinguish high frequency from low frequency variants, and thus only has high power until de novo mutations reach intermediate frequencies.

$\mathbf{F_{st}}$ and PBS tests

As mentioned previously, under positive selection, as the beneficial haplotype dominates the case population, the variability within this population decreases. This can be tested directly using the relative allele frequency spectra as in S_{f} and S_{π}, but there is additional information in the site-specific frequency differences across the case and control populations. For instance, let us consider a variant at 20% frequency in the control population. In the case population, this variant lies on a beneficial haplotype, and is sampled at frequency 80%. Under S_{π}, this variant contributes equally to both the case and the control statistics, while in S_{f}, this variant contributes negatively to the overall statistic. However, there is clearly a sharp rise in frequency, representing an increased branch length between the case and control populations in the phylogenetic tree, which may be indicative of selection. Since the F_{st} test measures the site-specific variability between populations (π_{between}), it would be able to detect such situations. Importantly, the scenario described above is consistent with selection occurring on standing variation, where the beneficial haplotype is present in non-negligible frequencies in the control population. However, the undirected nature of the branch lengths presents disadvantages. For instance, a significant F_{st} value could also indicate positive selection in the 'control' population. This is addressed in the PBS test by

calculating *population specific* branch lengths using multiple controls.

5.5.11 Region filtration

We assume that the genetic basis for the adaptation influences relatively few loci genome-wide. As a result, for a cross population test, the null distribution of two neutrally evolving populations can be approximated by the observed case vs. control distribution. As described above, we set significance thresholds corresponding to the top 0.1% genome-wide value for each statistic. For the Amhara population, these values were 0.16 (PBS), 0.18 (F_{st}), 1.73 (S_{π}), and 2.0 (S_{f}). For the Oromos, these were 0.24 (PBS), 0.22 (F_{st}), 1.58 (S_{π}), and 1.83 (S_{f}). Overall, there were 425 distinct regions that passed at least one of these tests. We thus implemented a number of filters (see Figure 5.1) in order to shortlist candidates that showed strong signals of selection. These are described below.

No genes in the region

We filtered out candidate regions that did not have a gene (as defined by RefSeq release 45, downloaded January 14, 2011) within their boundaries. Such regions may contain important regulatory variations, however, for an initial pass, we focused our efforts on regions for which there are more readily accessible methods to identify and validate causal linked genes. 183 regions were filtered out in this step.

Low frequency differential

A region under strong selection is characterized by changes in variant frequencies that cannot be explained by a neutral model. These regions manifest as blocks of SNPs where at least one allele confers a beneficial advantage and increases in frequency. The frequencies of other, nearby alleles in linkage disequilibrium (LD) with these correspondingly increase as well. We take advantage of this fact by looking for regions with multiple SNPs present in a block structure (at comparatively high frequencies) in the case population. For a case population sample of size n, we

iterate over all possible frequency values f, where $f = (1/n, 2/n, \ldots, (n-1)/n)$. For a given value of f, we then isolate all variants in the region with frequency within $1/n$ from f. From these, we define an f-frequency block as a subset of $\geq$10 consecutive SNPs. For each f-frequency block, we calculate the frequency differential, defined by the absolute difference in mean frequency between the case population and the closer of the LWK and CEU controls. We filter out regions where the maximum block differential in the region is less than 20%. We use a 20% frequency differential as the cutoff since, due to our relatively low sample size, differences in frequency smaller than 20% cannot be reliably distinguished from deviations due to sampling (see Figure 5.4 and Discussion). 184 regions were filtered out in this step.

Artifactual regions

We further eliminated candidates where the signal was caused by inconsistent SNP calls. These are regions where the frequency differential from the previous filter was directly caused by filtered variants in the case population that were not filtered in the control population. Although we attempted to prevent this using our variant filters (see above), some candidates (e.g., some of the HLA regions) avoided them. To identify these, we take as input, the set of f-frequency blocks with differentials $\geq$20%. We then extract all variants from in the case population, and replace the frequencies directly with those from our unfiltered SNP list. We then recalculate block frequency differential, and filter out regions where the maximum block differential in the region is less than 20%. 31 regions were filtered out in this step.

Similarity to other controls

Finally, to ensure that the regions represent selection for a Amhara/Oromos-only phenotype, we expanded our controls to include additional populations. These controls include other lowlander dbSNP [114] and 1000 Genomes populations [88]. If another population showed frequencies within 20% (our sampling error, as mentioned above) of the optimal haplotype block in our case population, we filter out

the region. 17 regions were filtered out of this step, leaving 10 regions, which we label as prioritized.

5.5.12 Power of whole genome sequencing

In our study, we performed high coverage ($15 - 20\times$), whole genome sequencing on 7 (Amhara) or 10 (Oromos) individuals. Alternative approaches to whole genome sequencing would include exome sequencing or genotyping. As these approaches are currently less expensive, this may allow for sampling more individuals. In Figure 5.9, we show the impact of sample size on power, using simulated populations. The simulation procedure was similar to the one described above, with 500 neutral initial populations, and selection coefficient fixed at $s = 0.02$. As previously described (Figure 5.3), there appear to be three general regimes of selection ("pre-fixation", "near fixation", and "post-fixation"), where different tests vary in their relative performance across regimes. We focus on a single test, S_π, and sample from each of the three regimes ($t = 450, 700, 1000, 1500$ generations after selection starts). In order to identify the effect of decreasing the sample size on power, we vary the sample sizes from $n = (2, \ldots, 40)$. As a gold standard for maximal attainable power, we used a large sample size of $n = 400$. Although sequencing more individuals would improve the sensitivity, as seen in Figure 5.9, sampling 14 or 20 haplotypes yields between 67-95% power compared to our gold standard. Notably, we see that sampling fewer individuals has the greatest impact in the "pre-fixation" regime. This is due to two factors. First, sampling fewer individuals leads to higher variance in the observed frequencies (see Figure 5.4). Second, the pre-fixation regime is when the frequency differential of the beneficial haplotype block compared to controls is at its lowest. Despite this, we still can detect the majority of cases of positive selection in our simulations.

We also tested the power of whole genome sequencing in comparison to other technologies, due to its unique ability to capture all allelic variation in a region. Figure 5.8 shows one of our statistics (S_π) applied to chromosome 13, which contains one of our top hits (the EDNRB gene region). We compared the variants captured in our study to those captured by two alternative approaches:

whole exome sequencing and genotyping. To mimic the effects of whole exome sequencing, we masked variants not targeted by the Nimblegen (Madison, WI) 2.1M exon capture array. For comparison with genotyping studies, we masked variants not included in the ~1M Affymetrix (Santa Clara, CA) Genome-Wide Human SNP Array 6.0. As shown in Figure 5.8, with whole genome sequencing the strongest signal chromosome wide is located in the EDNRB gene region. In contrast, genotyping shows a significantly weaker signal in the region, while whole exome sequencing shows no signal at all. For situations where a large portion of the signal is in intergenic or intronic sequence, whole genome sequencing provides a major advantage over other technologies.

5.5.13 Drosophila stocks and test of hypoxia tolerance

The Drosophila stocks carrying UAS-RNAi transgene were obtained from the Vienna Drosophila RNAi center (Vienna, Austria) (stock number: 25995 and 103805 carrying UAS-RNAi(cic); 22358 and 109336 carrying UAS-RNAi(CG11055); 29003 and 107333 carrying UAS-RNAi(CG8962/Paf-Aha), 42462 and 8018 carrying UAS-RNAi(CG7466)). The da-Gal4 driver (stock number: 8641) was obtained from Bloomington stock center (Bloomington, Indiana, USA).

Hypoxia tolerance of Drosophila crosses with specific RNAi-mediated knockdown was carried out as described in Chapter 4. Fifteen virgin female flies homozygous for UAS-RNAi were crossed with 10 male flies homozygous for da-GAL4 and allowed to lay eggs for 24 hours in normoxia. The vials with the eggs were transferred into a computer-controlled atmosphere chamber supplied with 5% oxygen balanced with nitrogen, with 12 hour-dark and 12 hour-light cycle at $22 \pm 1°C$. The Gal4 driver and UAS-RNAi stocks alone without crossing were included in parallel as controls. After three weeks of culturing, the vials were assayed for the number of pupal cases that were empty or full to calculate the eclosion rate. Six vials of each condition were completed in 2 different experiments for a minimum of 200 pupal cases scored for each condition/cross. The eclosion rate was presented as percentage of empty pupae in all scored pupal cases.

5.6 Acknowledgments

We thank Yu-hsin Hsiao, Orit Poulsen, and Shewadeg Gebru for technical assistance. This study is supported by NSF-CCF-1115206, 5RO1-HG004962 and U54 HL108460 to VB, and NIH awards 1P01HL098053 and 5P01HD32573 to GGH.

Chapter 5 is in full material submitted for publication from *Udpa N*, Ronen R*, Zhou D*, Liang J, Stobdan T, Appenzeller O, Yin Y, Du Y, Guo L, Cao R, Wang Y, Jin X, Huang C, Jia W, Cao D, Guo G, Claydon VE, Hainsworth R, Gamboa JL, Zibenigus M, Zenebe G, Xue J, Liu S, Frazer KA, Li Y, Bafna V, Haddad GG. Whole Genome Sequencing of Ethiopian Highlanders Reveals Evolutionarily Conserved Hypoxia Tolerance Genes.* The dissertation author was a joint first author of this paper, responsible for much of the computational analysis.

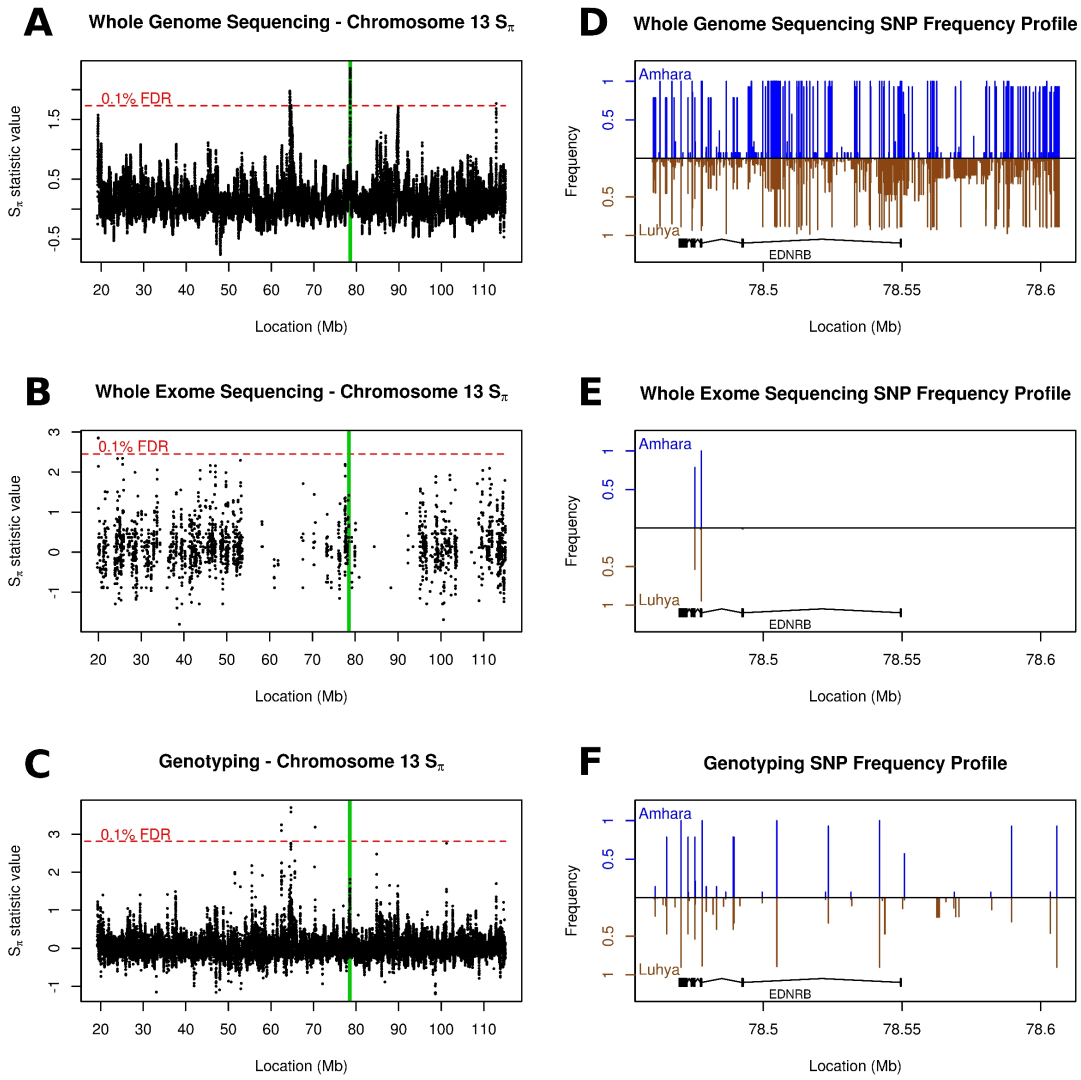


Figure 5.8: **Impact of whole genome sequencing on selection signals.** A-C) S_π statistic values across chromosome 13 in the Amhara population, compared to the Luhya (LWK) population, using variants present in our whole genome sequencing study (part A), the subset of variants in the targets from whole exome sequencing (part B), and the ~1M subset of variants from genotyping (part C). The red line represents a genome-wide, 0.1% FDR. The region highlighted in green contains the genomic location of the EDNRB gene. D-F) SNP frequency profiles of the region in green (representing EDNRB) in the Amhara (blue) compared to Luhya (brown, inverted) populations for whole genome sequencing (part D), whole exome sequencing (part E), and genotyping (part F). As can be seen, the strong signal present using whole genome sequencing is reduced drastically in genotyping and is entirely absent using exome sequencing.

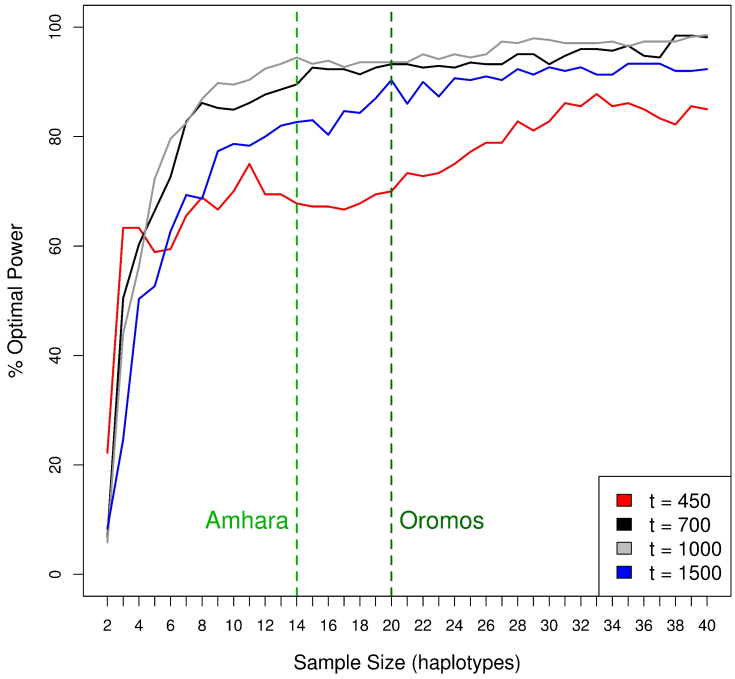


Figure 5.9: **The impact of sequenced sample size on power, using S_π as an exemplar test.** 500 populations were simulated, and selection coefficient was fixed at $s = 0.02$. Sample size is shown in haplotypes, and ranges form $n = (2, \ldots, 40)$. Optimal power at a given time was determined using a large sample size ($n = 400$). The populations were sampled at four time points ($t = 450, 700, 1000, 1500$), representing each of the observed regimes ("pre-fixation", "near-fixation" [including both $t = 700$ and $t = 1000$], and "post-fixation"). Although we see an increase in power as more haploptypes are sampled, sampling 14 or 20 haplotypes (our Amhara and Oromos populations, respectively) yields 67-95% of the optimal power.

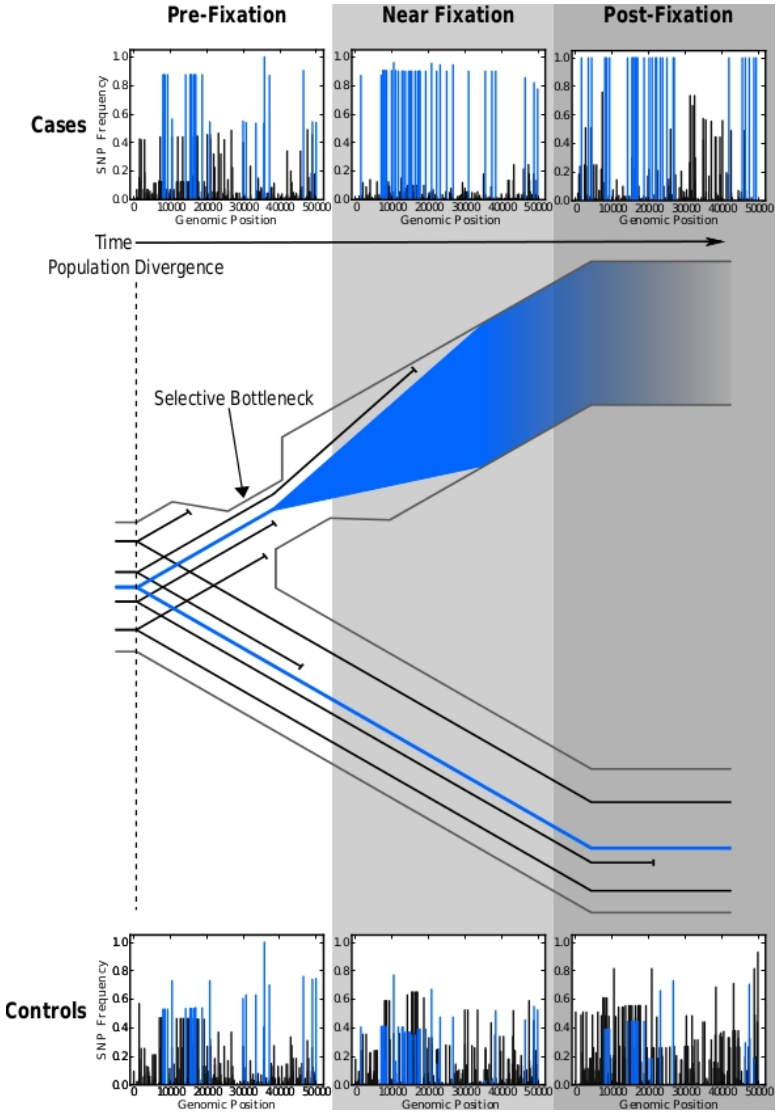


Figure 5.10: **Illustration of a selective bottleneck in one of two diverged populations, leading to a loss of genetic diversity.** The haplotype carrying the beneficial allele (shown in blue), along with other neighbouring (linked) alleles, becomes dominant in the case population at the expense of other haplotypes that die out (the "pre-fixation", and "near fixation" regimes). This leads to decreased genetic diversity in this population, which is characterized by a skew in the site frequency spectrum (top) relative to neutrality (bottom). As time progresses, genetic diversity is gradually restored to the region via de novo mutation (the "post-fixation" regime).

Chapter 6

The genetic basis of chronic mountain sickness (Monge's disease) in Andean highlanders

6.1 Abstract

We will now work on a similar problem as far as setup, although with some subtle, but important variations. As mentioned previously, the hypoxic conditions at high altitudes present a challenge for survival, causing pressure for adaptation. Interestingly, many high-altitude denizens (particularly in the Andes) are maladapted, with a condition known as chronic mountain sickness (CMS), or Monge's disease. To decode the genetic basis of this disease, we sequenced and compared the whole genomes of 20 Andean subjects (10 CMS and 10 non-CMS). We discovered 11 regions genome-wide with significant differences in haplotype frequencies consistent with selective sweeps. In these regions, two genes (an erythropoiesis regulator, *SENP1*, and an oncogene, *ANP32D*) had a higher transcriptional response to hypoxia in CMS individuals relative to non-CMS. We further found that downregulating the orthologs of these genes in flies dramatically enhanced survival rates under hypoxia, demonstrating that suppression of *SENP1* and *ANP32D* plays an essential role in hypoxia tolerance. Our study provides an unbiased framework to

identify and validate the genetic basis of adaptation to high altitudes, and identifies potentially targetable mechanisms for CMS treatment.

6.2 Introduction

More than 140 million humans have permanently settled on high altitude regions in various locations around the world, such as the Ethiopian plateau in East Africa, the Tibetan plateau in Asia and the Andes Mountains in South America. These geographically distinct populations have adapted uniquely to cope with high altitude hypoxia. For example, a higher hemoglobin concentration and oxygen saturation were detected among Andean highlanders as compared with Tibetans at the same altitude, but there was no difference between Ethiopian highlanders and sea-level residents in these two traits [115]. Furthermore, Tibetans have higher plasma concentrations of nitric oxide metabolites than North Americans [116], and their infants have higher birth weight and higher arterial oxygen saturation than Han Chinese infants at the same altitude [117, 118]. A statistical analysis of 4 quantitative traits (i.e., resting ventilation, hypoxic ventilatory response, oxygen saturation, and hemoglobin concentration) provided further evidence that the phenotypic adaptive responses to high-altitude hypoxia are different between the Tibetan and Andean populations [119]. Such differences in the patterns of hypoxia tolerant phenotypes suggest that distinct genetic mechanisms underlie hypoxia adaptation in different high altitude human populations. Despite the fact that the genetic contribution to human adaptation to high altitude has been proposed for a long time, the research is still at an early stage and additional evidence is critical for our understanding adaptation at high altitude [120, 121, 101, 65, 122, 123, 124]. In fact, some of the individuals living at high altitudes are mal-adapted, and present symptoms of Chronic Mountain Sickness.

Chronic mountain sickness (CMS) or Monges disease is characterized by severe polycythemia and an array of neurologic symptoms including headache, fatigue, somnolence, and depression [125]. Often, patients with CMS suffer from strokes and myocardial infarctions in early adulthood because of increased blood

viscosity. Previous studies have shown that CMS is common in Andeans, occasionally found in Tibetans and absent from the Ethiopian population living on East African high altitude plateau [126, 79]. Therefore, the Andean high altitude population provides us with the opportunity to dissect the genetic mechanisms underlying high altitude adaptation by comparing the genetic variations between patients with CMS and those adapted subjects without CMS.

To address this, we sequenced the whole genomes of 20 individuals (10 with CMS, and 10 without) residing in Cerro de Pasco, Peru. Unlike genotyping arrays or exome sequencing, whole genome sequencing captures the entire spectrum of variation in a region, thus providing complete characterization of the site frequency spectrum (SFS) and allowing maximal information for discovering selective sweeps.

6.3 Materials and Methods

6.3.1 Subjects and clinical characterization

All subjects are adult males residing in the Andean mountain range, in Cerro de Pasco, Peru, at elevation of over 4300m. Chronic mountain sickness (CMS), or Monges disease, is diagnosed using CMS scores. Individuals with CMS score > 15 were selected as CMS subjects, and those with CMS score < 5 were chosen as non-CMS subjects. Both blood samples (for DNA samples used in whole genome sequencing) and skin biopsies (for fibroblast cell cultures used in cell-based assays) were collected. Subjects were volunteers, and each subject gave informed, written consent. The UCSD institutional review board approved the protocol.

6.3.2 DNA extraction, library construction and sequencing

The process from this point up to SNP calling is identical to the previous chapter, and is summarized below primarily for convenience's sake. Genomic DNA was isolated using Blood DNA extraction kit (Qiagen, Valencia, CA) and randomly fragmented. Fragments of the desired length were gel-purified. Adapter ligation

and DNA cluster preparation were performed using the library preparation kit according to manufacturer's instruction (Illumina, San Diego, CA). Whole genome sequencing was performed on all 20 individuals using the Illumina HiSeq2000 platform to a mean, per-sample depth of $20 - 40\times$.

6.3.3 Read alignment, score recalibration and variant calling

We aligned the reads to the human reference genome (hg19) using BWA16 with default parameter settings. We adjusted the alignments using GATK indel-realignment, Picard read duplicate marking, and GATK quality score recalibration modules [85, 86] under default parameter settings, as defined by the GATK manual (version 2). We finally called and filtered SNVs using the GATK UnifiedGenotyper under default parameter settings. The sequencing was free of any mapping bias in coverage, mapping percentage, or variant counts for all subjects. See Figure 6.1 for an overview of the computational workflow.

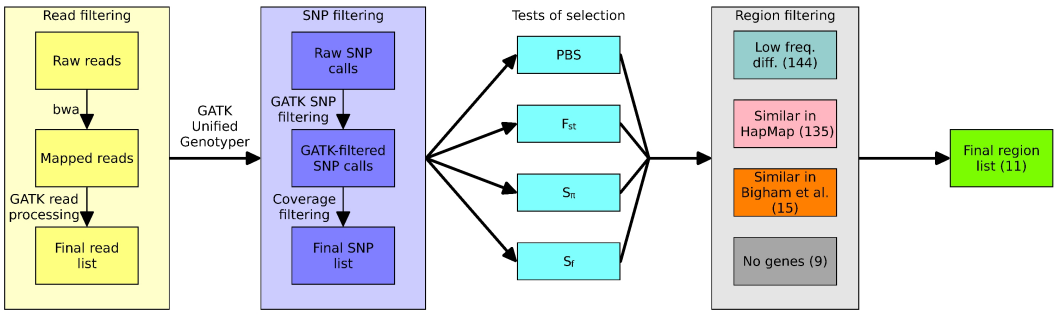


Figure 6.1: **Computational analysis workflow.** The raw sequence reads were mapped using BWA, followed by indel realignment, duplicate marking, and quality score recalibration using the GATK pipeline. Variants were then called and filtered using the GATK UnifiedGenotyper. After applying additional variant filters to account for the differences in coverage between our subjects and the control populations, we applied four complementary tests of selection to identify 314 regions as candidates for undergoing hypoxia-specific selective stress. Of these, we prioritized 11 regions for further experimental validation.

6.3.4 Tests of selection

We applied the same four cross-population tests of neutrality as in Chapter 5. S_{f} and S_{π} look at the differences in the site frequency spectrum associated with selection (see Chapter 3). F_{st} and PBS are designed to look for frequency differentials consistent with the longer phylogenetic branch lengths tied to non-neutrality. Please see Chapter 5 for more information on the definition and relative behaviors of these tests.

6.3.5 Human fibroblast cell culture, hypoxia treatment and real-time qPCR assay

Primary fibroblast cells were derived from CMS or non-CMS skin biopsies and expended in DMEM medium supplement with 10% fetal calf serum, 2.5% penicillin/streptomycin and 1% fungizone antibiotic (Life Technologies, CA). When reaching 75-80% confluence, the cells were treated with 1.5% O_2 for 24hrs. Untreated cultures were used as controls. After treatment, the cells were first washed with (the chemical) PBS (Cellgro, VA) and then treated with TrypLE$^{\mathrm{TM}}$ Express (Life Technologies, CA) for 5-10 minutes at 37°C to detach the cells. The detached/trypsinized cells were washed with fresh culture media, and centrifuged at $200\times g$ for 3 minutes. The pelleted cells were frozen at -80°C until RNA extraction.

Total RNA was extracted using the NucleoSpin® RNA II Kit (Clontech, CA) and eluted with 40μl of RNase-free water. RNA concentrations were measured with a NanoDrop 1000 (Thermo Scientific, DE). cDNA was synthesized with 1μg of total RNA using the SuperScript® III First-Strand Synthesis System according to the manufactures instructions (Life Technologies, CA). Real-time qPCR was performed in duplicates in 20 μl reaction volume on a MicroAmp® Fast Optical 96-Well Reaction Plate (Life Technologies, CA). Each reaction contained 1μl of cDNA, 2μl of 3μM forward and reverse primer mix, 10μl of Power SYBR® Green PCR Master Mix (Life Technologies, CA), and 7μl of water. The real-time PCRs were run on a 7900HT Fast Real-Time PCR System (Life Technologies, CA) using the following conditions: 95°C for 10 min followed by 40 cycles of 95°C for 15 sec

and 60°C for 1 min. GAPDH was used as internal control for normalization.

6.3.6 Fly lines and culture

The candidate genes were obtained from the human study and their fly orthologs were identified using FlyBase (www.flybase.org). Publicly available RNAi stock lines for each candidate gene (if possible, duplicate or triplicate lines per gene) were obtained from Vienna Drosophila Research Stock Center. The w1118 was used as background control. To ubiquitously knock down the candidate gene in the F1 progeny, the da-GAL4 driver was obtained from Bloomington Drosophila Stock Center at Indiana University. All stock lines were raised at room temperature and maintained on standard cornmeal.

6.3.7 In-vivo hypoxia tolerance test

The virgin females (n=9) da-GAL4 were crossed to different UAS-RNAi lines (n=6) or vice versa. Sufficient time (~3 days) was given for the flies to mate/cross. These are referred to as "cross". The vials were kept under ambient conditions for the flies to lay a sufficient number of fertilized eggs. After 48 hours, adults were transferred to a new vial. The original vials were then transferred to a computer-controlled hypoxia chamber, maintained at 5% oxygen on a 12/12 hours light/dark cycle at room temperature. The adults were discarded after 48 hours from the second batch of vials and these vials were then kept at ambient oxygen conditions (~21% oxygen) to be used as "control". After 21 days, the ratio of empty pupae (eclosed) to total pupae formed (eclosed + uneclosed) in each vial was calculated to determine the percentage eclosion rate. Simultaneously, the w1118, da-GAL4 and RNAi were self-crossed to be used as controls. Each set was performed in triplicate and the entire experiment was repeated to check for consistency. The differences in eclosion rate for the crosses under 21% and 5% O_2 were calculated using a chi-squared test, and between the (UAS-RNAi x da-Gal4) cross and the UAS-RNAi alone (stock control) using an unpaired t-test. A P-value of <0.05 was considered significant.

6.4 Results

We sequenced the genomes using the Illumina HiSeq 2000 platform to a mean depth of 34× per individual, mapped the reads to the hg19 reference using BWA [84], and called SNVs using the GATK pipeline [85, 86] (Figure 6.1).

6.4.1 Lowlander control populations

Using cross-population tests of selection, we looked for sweeps in the non-CMS individuals relative to the CMS individuals. However, since both groups belong to the same population, it may be difficult to determine if elements of a polygenic response are partially inherited in the CMS subjects. To address this issue, we additionally compared our non-CMS highlander population against the nearest 1000 Genomes (lowlander) controls [88]. We applied ADMIXTURE analysis [87] to 10,363 variant sites to identify the population most closely related to our Andean highlander subjects. As can be seen in Figure 6.2, our Andean subjects show varying amounts of shared ancestry with the three American populations (CLM, MXL, PUR). The closest population consists of 66 Mexican (MXL) individuals from Los Angeles, California, and was thus chosen as the lowlander control for all cross-population tests of selection. As outgroup (for the PBS test), we used a distant population consisting of 67 Luhya (LWK) individuals from Webuye, Kenya. Importantly, our highlander subjects and control populations had considerable differences in coverage, leading to discrepancies in variant calling.

6.4.2 Variant filters

To adjust for these differences, we filtered our call set using three steps. First, we observed several variants in clustered genomic loci that were discarded by the variant caller in the (CMS and non-CMS) study populations. This happens due to sequencing and mapping artifacts such as strand bias, low sequence complexity, or structural variations. We considered a region as suspicious if 10 consecutive SNPs were filtered out by GATK in our study populations, and filtered out any SNPs present in these regions in the controls. Second, following the

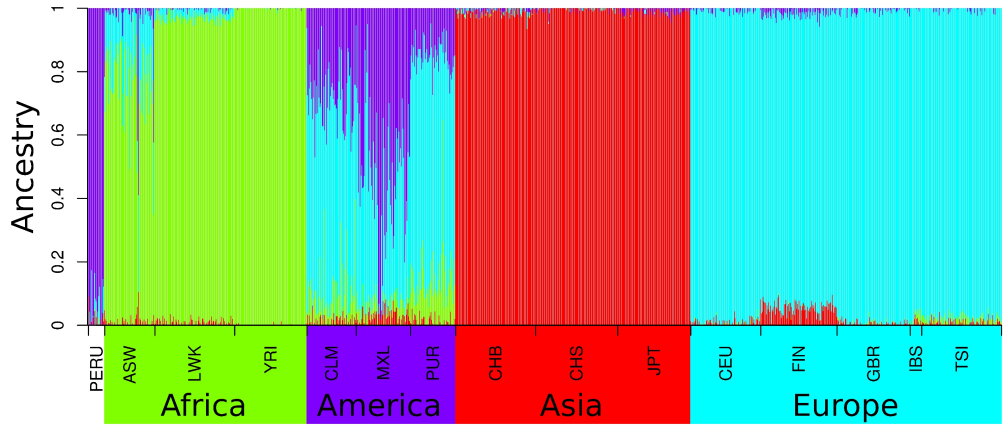


Figure 6.2: **ADMIXTURE analysis (4 clusters) applied to the 1000 Genomes populations jointly with our Andean highlander subjects.** The Andean highlanders (leftmost) show no signs of significant shared ancestry with Europeans or Africans, but do show varying degrees of shared ancestry with the American populations (CLM, MXL, PUR). As lowlander controls for our cross-population tests of selection, we used the closest population (MXL). As an outgroup for the PBS test, we used a distant African population (LWK).

protocol used by the 1000 Genomes project, we removed any site with a mean coverage higher than twice the median genome-wide coverage as likely to be caused by duplication [88]. This removes variants found in repetitive regions, such as centromeric sequence. We also removed any site with less than 2× coverage per person in the study population as being too poorly covered to accurately call. Finally, we removed sites with an excess of heterozygotes, using a test from Emigh [106] describing the heterozygote probability as:

$$P_{Aa} = \frac{n!}{n_{AA}!\, n_{Aa}!\, n_{aa}!} \times \frac{n_A!\, n_a!}{(2n)!} \times 2^{n_{Aa}}$$

We discarded variants with P-value under 0.05. After applying the above filters, a total of 5,937,347 variants in the CMS subjects and 5,777,092 variants in the non-CMS subjects remained.

6.4.3 Identifying regions under positive selection

As mentioned throughout this thesis, under positive directional selection, any haplotype harboring a beneficial mutation, as well as linked hitchhiking mutations, rapidly increases in frequency. This leads to a characteristic loss of genetic diversity centered on the beneficial mutation known as a selective sweep [24]. Importantly, the loss of genetic diversity and the consequent skew in the site frequency spectrum (SFS) can be used to detect loci important for adaptation to selective stress [24, 25]. We use cross-population tests to adjust for events shared between case and control populations (such as population bottlenecks, genetic drift, or even directional selection acting on unrelated phenotypes). These are likely due to events that took place before the divergence of our case and control populations, and thus unlikely related to hypoxia tolerance. Population-specific selection can be measured as a large decrease in diversity in the case population compared to controls [24]. This is usually captured as skews in the site frequency spectrum (SFS) of a region under selection. However, the SFS (and thus, the performance of different tests of selection) is significantly affected by many parameters, including the selective pressure (s) affecting the beneficial allele, as well as the length of time (t) for which the allele has been under selection (Chapter 3). For a complex phenotype such as hypoxia tolerance, we expect that multiple loci throughout the genome may simultaneously undergo selective sweeps, each under a distinct selection pressure and for a distinct time period. For this reason, we apply several tests of selection: S_f, S_π, F_{st}, and PBS. As the tests are powerful under different regimes (weak/strong and early/late) of selection, and as we have no prior knowledge of the regime we are after, we consider regions found as significant in any of the above tests as potentially interesting.

We also assume that the genetic basis for adaptation to hypoxia influences relatively few loci genome-wide. As a result, for a cross-population test, the null distribution of two neutrally evolving populations can be approximated by the observed case vs. control distribution. We set significance thresholds corresponding to the top 0.1% genome-wide value for each statistic. For the non-CMS versus MXL tests, these values were 0.11 (PBS), 0.19 (F_{st}), 2.93 (S_π), and 3.87 (S_{f}). For

the non-CMS versus CMS tests, these were 0.17 (PBS), 0.31 (F_{st}), 2.18 (S_π), and 3.23 (S_f). This set of analyses identified 314 regions spanning 29.67Mbp that were significant in at least one test under a 0.1% genome-wide false discovery rate.

6.4.4 Region filtration

Similarly to Chapter 5, we wanted to validate interesting gene candidates experimentally. As a result, we developed a series of automated prioritization criteria in order to shortlist candidates that showed the strongest signals of selection. Due to the presence of other samples of the same population, though, we changed our prioritization steps from the previous setup.

Low frequency differential

A region under strong selection should have multiple variants present with a high frequency differential between case and control populations. To identify this, we iterate over all possible case frequency values f, where $f = (1/n, 2/n, ..., (n-1)/n)$ (for a case population sample of size n haplotypes). For a given value of f, we isolate all variants in the region with frequency within $1/n$ from f. From these, we define an f-frequency block as a subset of $\geq$10 consecutive SNPs. For each f-frequency block, we calculate the frequency differential, defined by the absolute difference in mean frequency between the non-CMS population and the associated control population (either CMS or the closer of the MXL/CEU populations). We prioritize regions where the maximum block differential in the region is greater than 20%. We set the threshold at 20%, as this is the expected sampling variance of a variant at a given frequency when sampling 20 haplotypes (corresponding to 10 CMS or 10 non-CMS subjects) from a population (see Figure 5.4). 170 regions were considered prioritized under this criterion.

Similarity to other controls

In order to ensure that the prioritized regions represent selection for high altitude adaptation (rather than other phenotypes, potentially shared with different populations), we expanded our controls to include additional populations.

These controls included the CMS individuals we sequenced, the MXL/CEU 1000 Genomes populations, and other lowlander HapMap populations [127]. We prioritize regions where the observed haplotype block has frequency differential of greater than 20% (our sampling error, as mentioned above) compared to all other sampled controls. 35 regions were considered prioritized under this (as well as the previous) criterion.

Leveraging existing genotype data

We also used variant calls from a previous genotyping study by Bigham et al. [122] to further prioritize candidate regions. This data from this study provided us with two advantages. First, the authors performed genotyping on 49 Andean highlanders with no symptoms of CMS (including 24 from the same population we sequenced, in Cerro de Pasco, Peru). This helped us refine our sample frequencies and identify any false signals caused by sampling. Second, they genotyped 39 lowlanders of Native American ancestry from Southern Mexico, providing us with an additional lowlander control population. This population is both geographically closer to our Andean highlander subjects and does not show any signs of admixture with Europeans [122].

For a given region, we extracted all variants sampled from the previously identified f-frequency block that were also sampled by Bigham et al. For these, we refined our non-CMS frequencies by taking an average (weighted by sample size) over the highlander frequencies from both studies. Due to the increased population size (total of 59 subjects), the expected error due to sampling was reduced to less than 10%. As a result, for a given region, if the revised block frequency of the adapted population was greater than 10% compared to all controls, we considered the region prioritized. We note that previously prioritized regions that had no variants sampled by Bigham et al. were unaffected by this criterion. 20 regions remained under consideration after this step.

No genes in the region

Finally, we prioritized candidate regions that had at least one gene (as defined by RefSeq release 45, downloaded January 14, 2011) within their boundaries. Although regions that do not overlap known genes may contain important regulatory variations, for an initial pass, we focused our efforts on regions for which there are more readily accessible methods to identify and validate causal linked genes. However, we did attempt to identify important regulatory variation in the significant non-genic regions by determining variants within transcription factor binding sites, as defined by TRANSFAC or ENCODE.

The 11 final regions all had haplotypes that were much higher in frequency in the non-CMS population than in many controls (including our CMS population as well as several sequenced and genotyped lowlander populations), and are presented in Table 6.1. These include many plausible candidates, including genes involved in oxidative stress response (*DGKK*, *DUOX1*, *DUOX2*, *DUOXA1*, and *DUOXA2*), response to reactive oxygen species (*GABRA3*), cell metabolism, and signaling (*PFKM*, *SENP1*, and *ANP32D*).

Two genomic regions (both on chromosome 12) appeared in the top 0.1% in both non-CMS vs. MXL and non-CMS vs. CMS tests. One of them is a 144kbp region at chr12:48411360-48555360 that contains a block of 66 "differential" SNPs with mean frequencies of 99% in non-CMS, 66% in CMS, 58% in MXL and 14% in CEU. The other region spans 156kbp at chr12:48751360-48907360 and contains a block of 114 "differential" SNPs with mean frequencies of 99% in non-CMS, 53% in CMS, 47% in MXL and 5% in CEU (Figure 6.3). Three genes (*SENP1*, *PFKM*, and *ASB8*) are located in the first region, and two genes (*ANP32D* and *C12orf54*) are located in the second region. Strikingly, some of these genes have been shown to regulate CMS-related phenotypes in mammals. Specifically, mice carrying a deficient *PFKM* allele exhibit severe cardiac and hematological disorders, muscle hypoxia and hyper-vascularization, impaired oxidative metabolism, fiber necrosis, and exercise intolerance [128]. Previous studies have also found that SENP1-/- led to erythropoiesis defect in mice [129, 130]. Furthermore, SENP1 regulates the activities of several cell signaling pathways through desumoylation of key medi-

Table 6.1: **List of the top 11 regions prioritized in non-CMS relative to all lowlander controls**

Genomic Region	Gene Symbol	Tests
chr3:33254596-33314596	SUSD5	$S_{\pi,\mathrm{CMS}}$
chr6:58244452-58392452	GUSBP4	$S_{\pi,\mathrm{MXL}}$
chr6:157504452-157554452	ARID1B	$F_{\mathrm{st,MXL}}$
chr10:101014523-101092523	CNNM1	$\mathrm{PBS}_{\mathrm{MXL}}$
chr11:118147948-118199948	CD3E	$\mathrm{S}_{\pi,\mathrm{CMS}}$
chr12:48411360-48555360	SENP1*,PFKM,ASB8	$S_{\pi,\mathrm{MXL}}$, $S_{\mathrm{f,MXL}}$, $S_{\pi,\mathrm{CMS}}$, $S_{\mathrm{f,CMS}}$
chr12:48751360-48907360	ANP32D*,C12orf54	$S_{\pi,\mathrm{MXL}}$, $S_{\pi,\mathrm{CMS}}$
chr15:45338058-45436058	SORD,DUOX2,DUOXA2, DUOXA1,DUOX1	$F_{\mathrm{st,MXL}}$
chr19:19665844-19747844	PBX4,LPAR2,GMIP	$\mathrm{PBS}_{\mathrm{MXL}}$
chrX:50147676-50197676	DGKK	$S_{\pi,\mathrm{CMS}}$
chrX:151275676-151421676	MAGEA5,MAGEA10, GABRA3	$\mathrm{PBS}_{\mathrm{CMS}}$, $F_{\mathrm{st,CMS}}$

*Genes regulating hypoxia tolerance with transcriptional changes in fibroblasts and experimental validation in Drosophila

ators. For example, SENP1 enhances ASK1-JNK activation and cell apoptosis through desumoylation of HIPK1 in a ROS-dependent manner [131]. In addition, SENP1-dependent desumoylation also regulates the stability and activity of HIF1α and GATA1 transcription factors [129, 130] that play important roles in regulating physiological responses to hypoxia including erythrogenesis, angiogenesis and metabolic adaptation [132, 133].

6.4.5 In vitro and in vivo validation of candidate genes

We extended our investigation to study the functional impact of the SNP variants and candidate genes identified by our current analysis. We did this using human fibroblast cells [134] derived from four of the CMS and non-CMS subjects as well as in vivo using a Drosophila model. Of the top 5 candidate genes identified from our Andean samples, 2 genes have orthologs identified in the Drosophila genome (Flybase gene symbol *CG32110* for human gene *SENP1*, and *Mapmodulin*

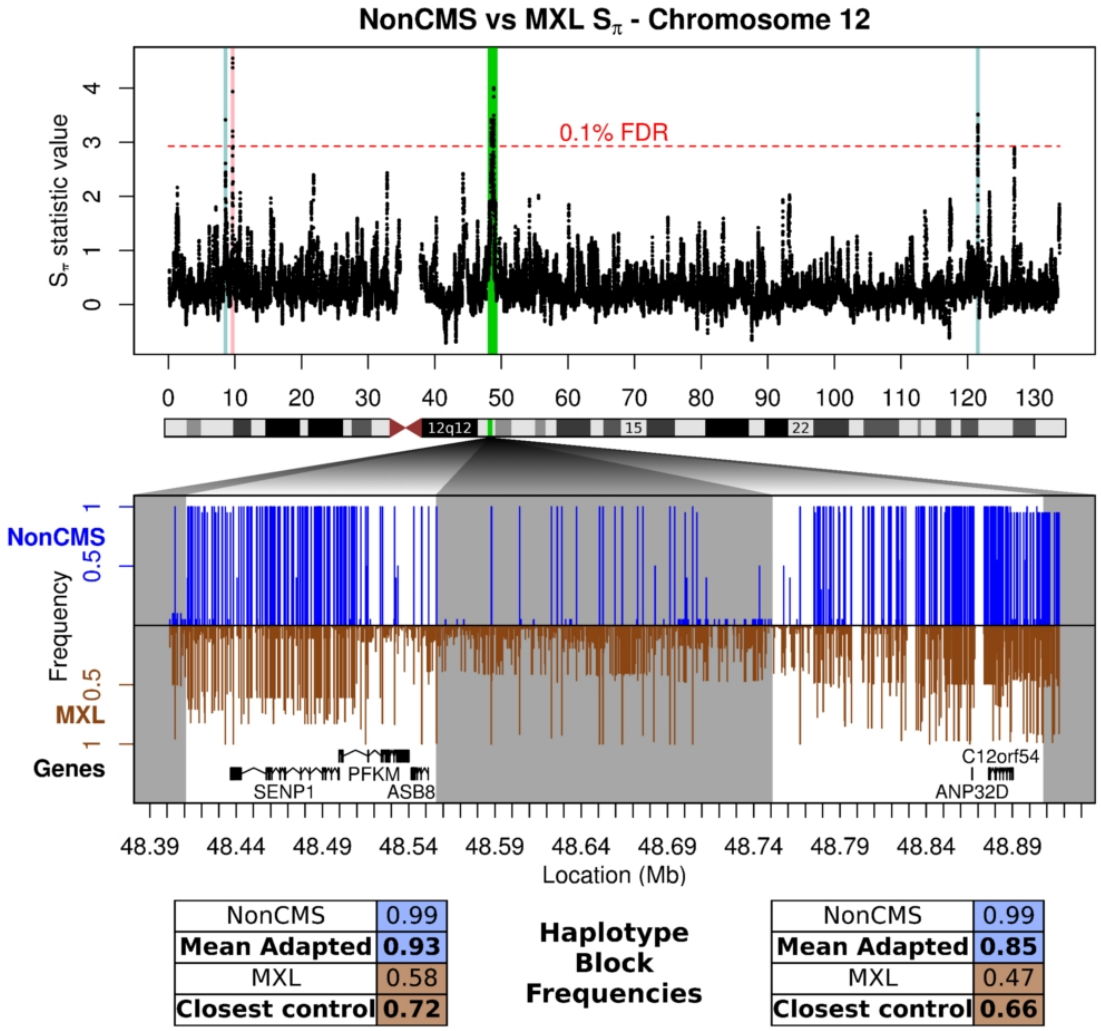


Figure 6.3: **Profile of the only two candidate regions that are significant in non-CMS vs. CMS and non-CMS vs. MXL tests.** (Top) One of the statistics in which both of these regions are significant ($S_{\pi,\mathrm{MXL}}$) is plotted across chromosome 12. Five distinct regions exceed the 0.1% FDR threshold the two highlighted in light blue do not have a major frequency differential between the non-CMS and MXL populations, while the one highlighted in pink is similar in other controls. The remaining two are considered prioritized and highlighted in green. The SNP frequencies in the area encompassing these two, part of q13.11, are plotted in the middle. In this plot, the two prioritized regions are highlighted in white, while other regions are shaded in gray. As can be seen, in both regions, there is an almost complete fixation of a haplotype in the non-CMS population that is at a much lower frequency in all lowlander and mal-adapted controls.

for human gene *ANP32D*) [135, 136].

We measured the transcriptional response of the candidate genes to hypoxia challenge using real-time qPCR. Interestingly, the expression levels of *SENP1* and *ANP32D* were significantly higher in the CMS cells as compared to the non-CMS cells following hypoxia treatment. In contrast, the expression of *PFKM* was down-regulated in both non-CMS and CMS cells (Figure 6.4). These results suggested that as compared to room-air cultured cells, the suppression of SENP1, ANP32D, and/or PFKM in non-CMS cells is beneficial for high altitude adaptation. As a corollary, when comparing the transcriptional response of CMS cells to that of non-CMS cells, the up-regulation of SENP1 and ANP32D in CMS cells might reflect mechanisms that underlie mal-adaptation in the CMS individuals at high altitude. We then proceeded to test this hypothesis *in vivo* in *Drosophila melanogaster*.

Since many human disease genes are evolutionarily conserved in Drosophila melanogaster [48, 137, 138], flies have been extensively used as a powerful in vivo model to dissect genetic mechanisms that contribute to human disease, including aging [139, 140], neurologic and cardiac disease [141, 142, 143], cancer [144, 145] as well as mechanisms underlying hypoxia tolerance or susceptibility [146]. In the current study, we took advantage of a Drosophila GAL4/UAS-RNAi system [46, 45, 147] to knockdown the transcript levels of these orthologs individually, mimicking the transcriptional suppression of these candidate genes in the non-CMS samples in an attempt to determine their potential role in adaptation to high altitude hypoxia. In addition, this strategy is also relevant to test the opposite hypothesis that the up-regulation of *SENP1* and *ANP32D* following hypoxia in the CMS samples is deleterious for survival in a hypoxic environment. The UAS-RNAi x GAL4 crosses were first cultured in normoxia to determine the effect of RNAi-mediated knockdown of each candidate gene on development. All the crosses developed normally with eclosion rates of over 95%, demonstrating that down-regulation of the candidate genes had no significant effect on development in normoxia. The flies resulting from these crosses were then tested under a hypoxic condition (5% O_2) by scoring the eclosion rate, an index of completion of development and survival. This hypoxic condition has been previously proven to be critical for distinguishing

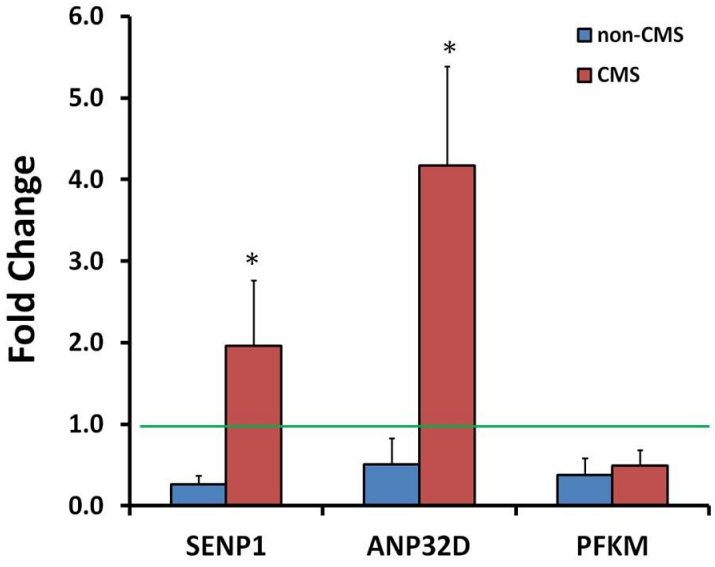


Figure 6.4: **Hypoxia response of top candidate genes in non-CMS and CMS cells.** Fibroblast cells were derived from skin biopsies obtained from the non-CMS and CMS subjects. Two non-CMS and 2 CMS cell lines were treated with 1.5% O_2 for 24 hours. The expression levels of SENP1, ANP32D and PFKM were measured by using quantitative real-time PCR. Same cell lines cultured under room air condition were used as normoxia controls. Hypoxia treatment induced a significant down regulation of SENP1, ANP32D and PFKM in non-CMS cells. In contrast, hypoxia treatment up-regulated the expression of SENP1 and ANP32D in the CMS cells, an opposite effect to the changes observed in the non-CMS cells (*$P<0.05$, Wilcoxon rank test). Each bar represents mean ± SEM of 2 measurements in duplicate.

hypoxia-tolerant flies from others [36, 148]. In order to minimize false positive results induced by off-target effects of RNAi or insertion effects of a particular UAS-RNAi transgene, we only included Drosophila orthologs with multiple available UAS-RNAi transgenic lines in this analysis. As shown in Figure 6.5, a dramatic enhancement of hypoxia tolerance was observed in both *CG32110/SENP1* and *Mapmodulin/ANP32D* when these genes were knocked down, demonstrating that down-regulation of the orthologs of *SENP1* and *ANP32D* is indeed beneficial for survival in severe hypoxic conditions in vivo.

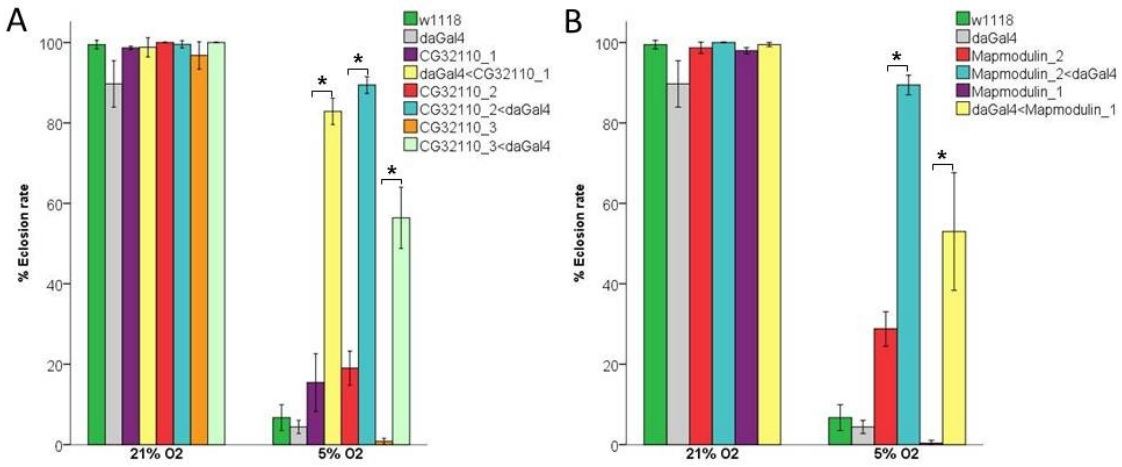


Figure 6.5: **Down-regulation of human SENP1 and ANP32D ortholog in Drosophila enhances survival under hypoxic condition.** da-Gal4 driver was used for ubiquitously knocking down the candidate gene by crossing with respective UAS-RNAi lines. Their eclosion rates were observed at 21% and 5% O_2 environment. (A) CG32110-RNAi when crossed with da-Gal4 significantly increases the eclosion rate at 5% O2 environment. The results were consistent in three RNAi lines targeting the same human SENP1 ortholog gene CG32110 (*:P<0.005, unpaired t-test). (B) The difference were also significant for the eclosion rate (%) of the F1 progeny for the two Mapmodulin-RNAi fly lines (*:P<0.005, unpaired t-test). Each bar represents mean ± SEM of eclosion rate. The w1118 and da-Gal4 stocks were tested and used as background controls.

6.5 Discussion

We present here the first genome-wide study of genetic adaptation in humans that confirms the effect of relevant genotypes on expression and further validates their role in model organisms. Using the enhanced power of whole genome sequencing, we identified a number of putative regions showing strong signals of selective sweep. We find that two genes in these regions, *ANP32D* and *SENP1*, have significantly increased expression in the CMS patients compared to the non-CMS individuals. Consequently, we hypothesized that down-regulating these genes could be beneficial in coping with hypoxia. We found that flies with these genes down-regulated had a remarkably enhanced survival rate under hypoxia.

There are several implications of our study. First, similarly to Chapter 5, we

performed high-coverage, whole genome sequencing (WGS) of 20 individuals. Traditionally, genome-wide scans for selection generally involve sampling the genome through genotyping or whole exome sequencing. However, there is an important trade-off when using these experimental designs. Specifically, WGS provides for a complete sampling of variant sites, albeit (usually) on a much smaller number of individuals. This completeness is critical for detecting selection. For instance, 10 of the 11 prioritized regions identified in our study, where we find strong peaks for our tests at a low genomic FDR using WGS, are not discovered when restricting to sites sampled by the alternative approaches.

Second, our study reveals many mechanistic insights on human adaptation (and mal-adaptation) to hypoxia. Since CMS patients are polycythemic (with hematocrit $> 65\%$), their blood becomes much more viscous. In turn, this increased viscosity jeopardizes blood flow to major organs, sometimes to the degree of ischemia, leading in some patients to myocardial infarction and stroke [3, 149, 150]. SENP1 is known to regulate erythropoiesis [130], and indeed SENP1-/- mice die of anemia in early life [129]. This gives credence to the idea that the increased expression of SENP1 plays a role in the basic pathogenesis of polycythemia in CMS patients. In contrast, in spite of the fact that there is little known about ANP32D and the PP32 phosphatase gene family, ANP32D functions as an oncogene. We raise here the possibility that this particular gene alters cellular metabolism in a fashion that is similar to that of cancer cells, especially given that such cells can flourish in low oxygen conditions. In conclusion, a better understanding of the mechanisms underlying hypoxia tolerance in high altitude human populations will, in all likelihood, elucidate the pathogenesis of other conditions at sea level, including congenital heart disease, obstructive sleep apnea and cancer.

6.6 Acknowledgments

We thank Ms. Mary Hsiao, Ms. Orit Poulsen for technical assistance, and Dr. Abigail W. Bigham for providing Latin American highlander and lowlander genotype data. This study is supported by NSF-CCF-1115206, 5RO1-HG004962

and U54 HL108460 to VB, and NIH award 1P01HL098053 and 5P01HD32573 to GGH. The authors do not have any competing interests associated with this manuscript.

Chapter 6 is in full material submitted for publication from *Zhou D*, Udpa N*, Ronen R*, Stobdan T, Liang J, Appenzeller O, Zhao HW, Yin Y, Du Y, Guo L, Cao R, Wang Y, Jin X, Huang C, Jia W, Cao D, Guo G, Gamboa JL, Villafuerte F, Callacondo D, Xue J, Liu S, Frazer KA, Li Y, Bafna V*, Haddad GG*. Genetic mechanisms underlie chronic mountain sickness (Monge's disease) in Andean highlanders.* The dissertation author was a joint first author of this paper, responsible for much of the computational analysis.

Bibliography

[1] Collins FS, Morgan M, Patrinos A (2003) The Human Genome Project: lessons from large-scale biology. Science 300: 286–290.

[2] Dewitt ND, Yaffe MP, Trounson A (2012) Building stem-cell genomics in California and beyond. Nat Biotechnol 30: 20–25.

[3] Monge CC, Whittembury J (1976) Chronic mountain sickness. Johns Hopkins Med J 139 SUPPL: 87–89.

[4] Brinza D, Schultz M, Tesler G, Bafna V (2010) RAPID detection of gene-gene interactions in genome-wide association studies. Bioinformatics 26: 2856–2862.

[5] Prabhu S, Pe'er I (2012) Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. Genome Res 22: 2230–2240.

[6] Indyk P, Motwani R (1998) Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the thirtieth annual ACM symposium on Theory of computing. New York, NY, USA: ACM, STOC '98, pp. 604–613. doi:10.1145/276698.276876. URL `http://doi.acm.org/10.1145/276698.276876`.

[7] Paturi R, Rajasekaran S, Reif J (1995) The light bulb problem. Information and Computation 117: 187–192.

[8] Fraley C, Raftery AE, Murphy TB, Scrucca L (2012) mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report 597.

[9] Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, et al. (2012) The Drosophila melanogaster Genetic Reference Panel. Nature 482: 173–178.

[10] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.

[11] Zeng M, Kikuchi H, Pino MS, Chung DC (2010) Hypoxia activates the K-ras proto-oncogene to stimulate angiogenesis and inhibit apoptosis in colon cancer cells. PLoS ONE 5: e10966.

[12] Strub BR, Parkes TL, Mukai ST, Bahadorani S, Coulthard AB, et al. (2008) Mutations of the withered (whd) gene in Drosophila melanogaster confer hypersensitivity to oxidative stress and are lesions of the carnitine palmitoyltransferase I (CPT I) gene. Genome 51: 409–420.

[13] Asha H, Nagy I, Kovacs G, Stetson D, Ando I, et al. (2003) Analysis of Ras-induced overproliferation in Drosophila hemocytes. Genetics 163: 203–215.

[14] Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37: 413–417.

[15] Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

[16] Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.

[17] Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133: 693–709.

[18] Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.

[19] Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, et al. (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7: e1001322.

[20] Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89: 82–93.

[21] Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, et al. (2010) A covering method for detecting genetic associations between rare variants and common phenotypes. PLoS Comput Biol 6: e1000954.

[22] Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5: e1000384.

[23] Durrett R (2002) Probability models for DNA sequence evolution. New York, NY: Springer-Verlag, second edition.

[24] Fu YX (1995) Statistical properties of segregating sites. Theor Popul Biol 48: 172–197.

[25] Achaz G (2009) Frequency spectrum neutrality tests: one for all and all for one. Genetics 183: 249–258.

[26] Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7: 256–276.

[27] Fu YX, Huai H (2003) Estimating mutation rate: how to count mutations? Genetics 164: 797–805.

[28] Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583–589.

[29] Fiston-Lavier AS, Singh ND, Lipatov M, Petrov DA (2010) Drosophila melanogaster recombination rate calculator. Gene 463: 18–20.

[30] Kingman JFC (1982) On the genealogy of large populations. Journal of Applied Probability 19: 27–43.

[31] Muller HJ (1928) The Measurement of Gene Mutation Rate in Drosophila, Its High Variability, and Its Dependence upon Temperature. Genetics 13: 279–357.

[32] Lilliefors H (1967) On the Kolmogorov-Smirnov Test for normality with mean and variance unknown. J Am Stat Assoc 62: 399–402.

[33] Campbell RB (2007) Coalescent size versus coalescent time with strong selection. Bull Math Biol 69: 2249–2259.

[34] Leips J, Mackay TFC (2000) Quantitative trait loci for life span in Drosophila melanogaster: interactions with genetic background and larval density. Genetics 155: 1773–1788.

[35] Zhou D, Xue J, Lai JC, Schork NJ, White KP, et al. (2008) Mechanisms underlying hypoxia tolerance in Drosophila melanogaster: hairy as a metabolic switch. PLoS Genet 4: e1000221.

[36] Zhou D, Xue J, Chen J, Morcillo P, Lambert JD, et al. (2007) Experimental selection for Drosophila survival in extremely low O(2) environment. PLoS ONE 2: e490.

[37] McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–654.

[38] Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11: 863–874.

[39] Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.

[40] Artavanis-Tsakonas S, Grimwade BG, Harrison RG, Markopoulou K, Muskavitch MAT, et al. (1983) The Notch locus of Drosophila melanogaster: A molecular analysis. Developmental Genetics 4: 233–254.

[41] Kidd S, Young MW (1986) Transposon-dependent mutant phenotypes at the Notch locus of Drosophila. Nature 323: 89–91.

[42] Shellenbarger DL, Mohler JD (1975) Temperature-sensitive mutations of the notch locus in Drosophila melanogaster. Genetics 81: 143–162.

[43] Hartley DA, Xu TA, Artavanis-Tsakonas S (1987) The embryonic expression of the Notch locus of Drosophila melanogaster and the implications of point mutations in the extracellular EGF-like domain of the predicted protein. EMBO J 6: 3407–3417.

[44] Dovey HF, John V, Anderson JP, Chen LZ, de Saint Andrieu P, et al. (2001) Functional gamma-secretase inhibitors reduce beta-amyloid peptide levels in brain. J Neurochem 76: 173–181.

[45] Brand AH, Perrimon N (1993) Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. Development 118: 401–415.

[46] Fischer JA, Giniger E, Maniatis T, Ptashne M (1988) GAL4 activates transcription in Drosophila. Nature 332: 853–856.

[47] Baker KD, Thummel CS (2007) Diabetic larvae and obese flies-emerging studies of metabolism in Drosophila. Cell Metab 6: 257–266.

[48] Bier E (2005) Drosophila, the golden bug, emerges as a tool for human genetics. Nat Rev Genet 6: 9–23.

[49] Lu B, Vogel H (2009) Drosophila models of neurodegenerative diseases. Annu Rev Pathol 4: 315–342.

[50] Ocorr K, Perrin L, Lim HY, Qian L, Wu X, et al. (2007) Genetic control of heart function and aging in Drosophila. Trends Cardiovasc Med 17: 177–182.

[51] Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, et al. (2010) Genome-wide analysis of a long-term evolution experiment with Drosophila. Nature 467: 587–590.

[52] Charlesworth B, Coyne JA, Barton NH (1987) The relative rates of evolution of sex chromosomes and autosomes. The American Naturalist 130: pp. 113-146.

[53] Singh ND, Davis JC, Petrov DA (2005) Codon bias and noncoding GC content correlate negatively with recombination rate on the Drosophila X chromosome. J Mol Evol 61: 315–324.

[54] Takahashi KH, Tanaka K, Itoh M, Takano-Shimizu T (2009) Reduced X-linked rare polymorphism in males in comparison to females of Drosophila melanogaster. J Hered 100: 97–105.

[55] Bray SJ (2006) Notch signalling: a simple pathway becomes complex. Nat Rev Mol Cell Biol 7: 678–689.

[56] Nagel AC, Krejci A, Tenin G, Bravo-Patino A, Bray S, et al. (2005) Hairless-mediated repression of notch target genes requires the combined activity of Groucho and CtBP corepressors. Mol Cell Biol 25: 10433–10441.

[57] Androutsellis-Theotokis A, Leker RR, Soldner F, Hoeppner DJ, Ravin R, et al. (2006) Notch signalling regulates stem cell numbers in vitro and in vivo. Nature 442: 823–826.

[58] Ciofani M, Zuniga-Pflucker JC (2005) Notch promotes survival of pre-T cells at the beta-selection checkpoint by regulating cellular metabolism. Nat Immunol 6: 881–888.

[59] Eliasz S, Liang S, Chen Y, De Marco MA, Machek O, et al. (2010) Notch-1 stimulates survival of lung adenocarcinoma cells during hypoxia by activating the IGF-1R pathway. Oncogene 29: 2488–2498.

[60] Jogi A, Ora I, Nilsson H, Lindeheim A, Makino Y, et al. (2002) Hypoxia alters gene expression in human neuroblastoma cells toward an immature and neural crest-like phenotype. Proc Natl Acad Sci USA 99: 7021–7026.

[61] Osborne BA, Minter LM (2007) Notch signalling during peripheral T-cell activation and differentiation. Nat Rev Immunol 7: 64–75.

[62] Osipo C, Golde TE, Osborne BA, Miele LA (2008) Off the beaten pathway: the complex cross talk between Notch and NF-kappaB. Lab Invest 88: 11–17.

[63] Perumalsamy LR, Nagala M, Sarin A (2010) Notch-activated signaling cascade interacts with mitochondrial remodeling proteins to regulate cell survival. Proc Natl Acad Sci USA 107: 6882–6887.

[64] Sahlgren C, Gustafsson MV, Jin S, Poellinger L, Lendahl U (2008) Notch signaling mediates hypoxia-induced tumor cell migration and invasion. Proc Natl Acad Sci USA 105: 6392–6397.

[65] Yi, X and Liang, Y and Huerta-Sanchez, E and Jin, X and Cuo, Z X and Pool, J E and Xu, X and Jiang, H and Vinckenbosch, N and Korneliussen, T S and Zheng, H and Liu, T and He, W and Li, K and Luo, R and Nie, X and Wu, H and Zhao, M and Cao, H and Zou, J and Shan, Y and Li, S and Yang, Q and Asan and Ni, P and Tian, G and Xu, J and Liu, X and Jiang, T and Wu, R and Zhou, G and Tang, M and Qin, J and Wang, T and Feng, S and Li, G and Huasang and Luosang, J and Wang, W and Chen, F and Wang, Y and Zheng, X and Li, Z and Bianba, Z and Yang, G and Wang, X and Tang, S and Gao, G and Chen, Y and Luo, Z and Gusang, L and Cao, Z and Zhang, Q and Ouyang, W and Ren, X and Liang, H and Zheng, H and Huang, Y and Li, J and Bolund, L and Kristiansen, K and Li, Y and Zhang, Y and Zhang, X and Li, R and Li, S and Yang, H and Nielsen, R and Wang, J and Wang, J (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329: 75–78.

[66] Gustafsson MV, Zheng X, Pereira T, Gradin K, Jin S, et al. (2005) Hypoxia requires notch signaling to maintain the undifferentiated cell state. Dev Cell 9: 617–628.

[67] Pistollato F, Rampazzo E, Persano L, Abbadi S, Frasson C, et al. (2010) Interaction of hypoxia-inducible factor-1alpha and Notch signaling regulates medulloblastoma precursor proliferation and fate. Stem Cells 28: 1918–1929.

[68] Wang R, Zhang YW, Zhang X, Liu R, Zhang X, et al. (2006) Transcriptional regulation of APH-1A and increased gamma-secretase cleavage of APP and Notch by HIF-1 and hypoxia. FASEB J 20: 1275–1277.

[69] Zheng X, Linke S, Dias JM, Zheng X, Gradin K, et al. (2008) Interaction with factor inhibiting HIF-1 defines an additional mode of cross-coupling between the Notch and hypoxia signaling pathways. Proc Natl Acad Sci USA 105: 3368–3373.

[70] Apidianakis Y, Nagel AC, Chalkiadaki A, Preiss A, Delidakis C (1999) Overexpression of the m4 and malpha genes of the E(spl)-complex antagonizes notch mediated lateral inhibition. Mech Dev 86: 39–50.

[71] Fan C, Iacobas DA, Zhou D, Chen Q, Lai JK, et al. (2005) Gene expression and phenotypic characterization of mouse heart after chronic constant or intermittent hypoxia. Physiol Genomics 22: 292–307.

[72] Bedogni B, Warneke JA, Nickoloff BJ, Giaccia AJ, Powell MB (2008) Notch1 is an effector of Akt and hypoxia in melanoma development. J Clin Invest 118: 3660–3670.

[73] Li X, Zhang X, Leathers R, Makino A, Huang C, et al. (2009) Notch3 signaling promotes the development of pulmonary arterial hypertension. Nat Med 15: 1289–1297.

[74] Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851–1858.

[75] Holt KE, Teo YY, Li H, Nair S, Dougan G, et al. (2009) Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA. Bioinformatics 25: 2074–2075.

[76] Rupert JL, Hochachka PW (2001) The evidence for hereditary factors contributing to high altitude adaptation in Andean natives: a review. High Alt Med Biol 2: 235–256.

[77] Moore LG, Niermeyer S, Zamudio S (1998) Human adaptation to high altitude: regional and life-cycle perspectives. Am J Phys Anthropol Suppl 27: 25–64.

[78] Beall CM (2007) Two routes to functional adaptation: Tibetan and Andean high-altitude natives. Proc Natl Acad Sci USA 104 Suppl 1: 8655–8660.

[79] Beall CM, Decker MJ, Brittenham GM, Kushner I, Gebremedhin A, et al. (2002) An Ethiopian pattern of human adaptation to high-altitude hypoxia. Proc Natl Acad Sci USA 99: 17215–17218.

[80] Appenzeller O, Claydon VE, Gulli G, Qualls C, Slessarev M, et al. (2006) Cerebral vasodilatation to exogenous NO is a measure of fitness for life at altitude. Stroke 37: 1754–1758.

[81] Beall CM (2003) High-altitude adaptations. Lancet 362 Suppl: s14–15.

[82] Beall CM (2007) Detecting natural selection in high-altitude human populations. Respir Physiol Neurobiol 158: 161–171.

[83] Pagani L, Ayub Q, MacArthur DG, Xue Y, Baillie JK, et al. (2012) High altitude adaptation in Daghestani populations from the Caucasus. Hum Genet 131: 423–433.

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

[85] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498.

[86] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.

[87] Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655–1664.

[88] Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.

[89] Bernard N, Girouard J, Forest JC, Giguere Y (2007) The combination of ApoCIII, hepatic lipase and hormono sensitive lipase gene polymorphisms suggests an association with susceptibility to gestational hypertension. J Hum Genet 52: 244–254.

[90] Lampidonis AD, Rogdakis E, Voutsinas GE, Stravopodis DJ (2011) The resurgence of Hormone-Sensitive Lipase (HSL) in mammalian lipolysis. Gene 477: 1–11.

[91] Weber GJ, Choe SE, Dooley KA, Paffett-Lugassy NN, Zhou Y, et al. (2005) Mutant-specific gene programs in the zebrafish. Blood 106: 521–530.

[92] Iqbal J, Weisenburger DD, Chowdhury A, Tsai MY, Srivastava G, et al. (2011) Natural killer cell lymphoma shares strikingly similar molecular features with a group of non-hepatosplenic gamma-delta T-cell lymphoma and is highly sensitive to a novel aurora kinase A inhibitor in vitro. Leukemia 25: 348–358.

[93] Chen CH (2004) Platelet-activating factor acetylhydrolase: is it good or bad for you? Curr Opin Lipidol 15: 337–341.

[94] Arai H, Koizumi H, Aoki J, Inoue K (2002) Platelet-activating factor acetylhydrolase (PAF-AH). J Biochem 131: 635–640.

[95] Clozel M, Gray GA, Breu V, Loffler BM, Osterwalder R (1992) The endothelin ETB receptor mediates both vasodilation and vasoconstriction in vivo. Biochem Biophys Res Commun 186: 867–873.

[96] McCulloch KM, MacLean MR (1995) EndothelinB receptor-mediated contraction of human and rat pulmonary resistance arteries and the effect of pulmonary hypertension on endothelin responses in the rat. J Cardiovasc Pharmacol 26 Suppl 3: S169–176.

[97] Astigarraga S, Grossman R, Diaz-Delfin J, Caelles C, Paroush Z, et al. (2007) A MAPK docking site is critical for downregulation of Capicua by Torso and EGFR RTK signaling. EMBO J 26: 668–677.

[98] Tseng AS, Tapon N, Kanda H, Cigizoglu S, Edelmann L, et al. (2007) Capicua regulates cell proliferation downstream of the receptor tyrosine kinase/ras signaling pathway. Curr Biol 17: 728–733.

[99] Hurlbut GD, Kankel MW, Artavanis-Tsakonas S (2009) Nodal points and complexity of Notch-Ras signal integration. Proc Natl Acad Sci USA 106: 2218–2223.

[100] Moore LG, Shriver M, Bemis L, Hickler B, Wilson M, et al. (2004) Maternal adaptation to high-altitude pregnancy: an experiment of nature–a review. Placenta 25 Suppl A: 60–71.

[101] Simonson TS, Yang Y, Huff CD, Yun H, Qin G, et al. (2010) Genetic evidence for high-altitude adaptation in Tibet. Science 329: 72–75.

[102] Beall CM, Cavalleri GL, Deng L, Elston RC, Gao Y, et al. (2010) Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. Proc Natl Acad Sci USA 107: 11459–11464.

[103] Scheinfeldt LB, Soi S, Thompson S, Ranciaro A, Woldemeskel D, et al. (2012) Genetic adaptation to high altitude in the Ethiopian highlands. Genome Biol 13: R1.

[104] Marcus HG (2002) A history of Ethiopia. Berkeley and Los Angeles, CA: University of California Press, updated edition.

[105] Li R, Li Y, Fang X, Yang H, Wang J, et al. (2009) SNP detection for massively parallel whole-genome resequencing. Genome Res 19: 1124–1132.

[106] Emigh TH (1980) A comparison of tests for Hardy-Weinberg equilibrium. Biometrics 36: 627-642.

[107] Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. (2006) Positive natural selection in the human lineage. Science 312: 1614–1620.

[108] Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet 10: 639–650.

[109] Cavalli-Sforza L (1969) Human diversity. Proc 12th Int Cong Genet 2: 16.

[110] Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, et al. (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics 1: 274–286.

[111] Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Res 20: 393–402.

[112] Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19: 826–837.

[113] Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

[114] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308–311.

[115] Beall CM (2006) Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. Integr Comp Biol 46: 18–24.

[116] Erzurum SC, Ghosh S, Janocha AJ, Xu W, Bauer S, et al. (2007) Higher blood flow and circulating NO products offset high-altitude hypoxia among Tibetans. Proc Natl Acad Sci USA 104: 17593–17598.

[117] Moore LG (2001) Human genetic adaptation to high altitude. High Alt Med Biol 2: 257–279.

[118] Niermeyer S, Yang P, Shanmina, Drolkar, Zhuang J, et al. (1995) Arterial Oxygen Saturation in Tibetan and Han Infants Born in Lhasa, Tibet. New England Journal of Medicine 333: 1248-1252.

[119] Beall CM (2000) Tibetan and Andean patterns of adaptation to high-altitude hypoxia. Hum Biol 72: 201–228.

[120] Lorenzo VF, Yang Y, Simonson TS, Nussenzveig R, Jorde LB, et al. (2009) Genetic adaptation to extreme hypoxia: study of high-altitude pulmonary edema in a three-generation Han Chinese family. Blood Cells Mol Dis 43: 221–225.

[121] Bigham AW, Mao X, Mei R, Brutsaert T, Wilson MJ, et al. (2009) Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. Hum Genomics 4: 79–90.

[122] Bigham A, Bauchet M, Pinto D, Mao X, Akey JM, et al. (2010) Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet 6.

[123] Xu S, Li S, Yang Y, Tan J, Lou H, et al. (2011) A genome-wide search for signals of high-altitude adaptation in Tibetans. Mol Biol Evol 28: 1003–1011.

[124] Ji LD, Qiu YQ, Xu J, Irwin DM, Tam SC, et al. (2012) Genetic adaptation of the hypoxia-inducible factor pathway to oxygen pressure among eurasian human populations. Mol Biol Evol 29: 3359–3370.

[125] Leon-Velarde F, McCullough RG, McCullough RE, Reeves JT (2003) Proposal for scoring severity in chronic mountain sickness (CMS). Background and conclusions of the CMS Working Group. Adv Exp Med Biol 543: 339–354.

[126] Xing G, Qualls C, Huicho L, Rivera-Ch M, River-Ch M, et al. (2008) Adaptation and mal-adaptation to ambient hypoxia; Andean, Ethiopian and Himalayan patterns. PLoS ONE 3: e2342.

[127] Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.

[128] Garcia M, Pujol A, Ruzo A, Riu E, Ruberte J, et al. (2009) Phosphofructo-1-kinase deficiency leads to a severe cardiac and hematological disorder in addition to skeletal muscle glycogenosis. PLoS Genet 5: e1000615.

[129] Yu L, Ji W, Zhang H, Renda MJ, He Y, et al. (2010) SENP1-mediated GATA1 deSUMOylation is critical for definitive erythropoiesis. J Exp Med 207: 1183–1195.

[130] Cheng J, Kang X, Zhang S, Yeh ET (2007) SUMO-specific protease 1 is essential for stabilization of HIF1alpha during hypoxia. Cell 131: 584–595.

[131] Li X, Luo Y, Yu L, Lin Y, Luo D, et al. (2008) SENP1 mediates TNF-induced desumoylation and cytoplasmic translocation of HIPK1 to enhance ASK1-dependent apoptosis. Cell Death Differ 15: 739–750.

[132] De Maria R, Zeuner A, Eramo A, Domenichelli C, Bonci D, et al. (1999) Negative regulation of erythropoiesis by caspase-mediated cleavage of GATA-1. Nature 401: 489–493.

[133] Prabhakar NR, Semenza GL (2012) Adaptive and maladaptive cardiorespiratory responses to continuous and intermittent hypoxia mediated by hypoxia-inducible factors 1 and 2. Physiol Rev 92: 967–1003.

[134] Boutin AT, Weidemann A, Fu Z, Mesropian L, Gradin K, et al. (2008) Epidermal sensing of oxygen is essential for systemic hypoxic response. Cell 133: 223–234.

[135] McQuilton P, St Pierre SE, Thurmond J, Gelbart W, Brown N, et al. (2012) FlyBase 101–the basics of navigating FlyBase. Nucleic Acids Res 40: D706–714.

[136] Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, et al. (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics 12: 357.

[137] Pandey UB, Nichols CD (2011) Human disease models in Drosophila melanogaster and the role of the fly in therapeutic drug discovery. Pharmacol Rev 63: 411–436.

[138] Gilbert LI (2008) Drosophila is an inclusive model for human diseases, growth and development. Mol Cell Endocrinol 293: 25–31.

[139] Michno K, van de Hoef D, Wu H, Boulianne GL (2005) Modeling age-related diseases in Drosophila: can this fly? Curr Top Dev Biol 71: 199–223.

[140] Grotewiel MS, Martin I, Bhandari P, Cook-Wiens E (2005) Functional senescence in Drosophila melanogaster. Ageing Res Rev 4: 372–397.

[141] Lu B (2009) Recent advances in using Drosophila to model neurodegenerative diseases. Apoptosis 14: 1008–1020.

[142] Lessing D, Bonini NM (2009) Maintaining the brain: insight into human neurodegeneration from Drosophila melanogaster mutants. Nat Rev Genet 10: 359–370.

[143] Diop SB, Bodmer R (2012) Drosophila as a model to study the genetic mechanisms of obesity-associated heart dysfunction. J Cell Mol Med 16: 966–971.

[144] Polesello C, Roch F, Gobert V, Haenlin M, Waltzer L (2011) Modeling cancers in Drosophila. Prog Mol Biol Transl Sci 100: 51–82.

[145] Vidal M, Cagan RL (2006) Drosophila models for cancer research. Curr Opin Genet Dev 16: 10–16.

[146] Zhou D, Visk DW, Haddad GG (2009) Drosophila, a golden bug, for the dissection of the genetic basis of tolerance and susceptibility to hypoxia. Pediatr Res 66: 239–247.

[147] Dietzl G, Chen D, Schnorrer F, Su KC, Barinova Y, et al. (2007) A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. Nature 448: 151–156.

[148] Azad P, Zhou D, Zarndt R, Haddad GG (2012) Identification of genes underlying hypoxia tolerance in Drosophila by a P-element screen. G3 (Bethesda) 2: 1169–1178.

[149] Ergueta J, Spielvogel H, Cudkowicz L (1971) Cardio-respiratory studies in chronic mountain sickness (Monge's syndrome). Respiration 28: 485–517.

[150] West JB (2012) High-altitude medicine. Am J Respir Crit Care Med 186: 1229–1237.