View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Washington University St. Louis: Open Scholarship

Washington University in St. Louis
Washington University Open Scholarship

All Theses and Dissertations (ETDs)

7-31-2012

Identification of Deleterious and Disease Alleles in a General Population and Preterm Labor Patients

Sung Gook Chun
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/etd

Recommended Citation

Chun, Sung Gook, "Identification of Deleterious and Disease Alleles in a General Population and Preterm Labor Patients" (2012). *All Theses and Dissertations (ETDs)*. 948.
https://openscholarship.wustl.edu/etd/948

This Dissertation is brought to you for free and open access by Washington University Open Scholarship. It has been accepted for inclusion in All Theses and Dissertations (ETDs) by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Computational and Systems Biology

Dissertation Examination Committee:
Justin Fay, Chairperson
Barak Cohen
Elaine Mardis
Robi Mitra
Kelle Moley
Mike Province

Identification of Deleterious and Disease Alleles in a General Population and Preterm Labor Patients

by

Sung Gook Chun

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August, 2012

Saint Louis, Missouri

Copyright by

Sung Gook Chun

2012

ABSTRACT OF THE DISSERTATION

Identification of Deleterious and Disease Alleles
in a General Population and Preterm Labor Patients

by

Sung Gook Chun

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2012

Associate Professor Justin C. Fay, Chairperson

With the recent advance in sequencing technology, there have been growing interests in developing new methods to predict disease-causing alleles in a personal genome by integrating functional evidences from sequence conservation, genome-wide association studies and the transcriptional regulatory network. However, even in protein-coding regions, it is not well understood how often and by what mechanism deleterious alleles disrupting strong sequence conservation can become common in population frequency and affect complex traits in humans. Moreover, in non-coding regions, even for known disease-causing genes, it is not clear how sequence conservation can be combined with functional genomic data to predict underlying disease-causing variants.

To address the first question, I developed a new likelihood ratio test for sequence conservation to predict deleterious missense alleles in the human genome. By applying the new test to three personal genomes, I find that the presence of only 10% of common deleterious SNPs can be explained by false positives due to multiple hypothesis testing, violation of evolutionary model assumptions, recent gene duplication and relaxation of selective constraints on biological processes. Next, by applying the likelihood ratio test to a general human population, I find that both computationally predicted deleterious SNPs and known disease-associated alleles are enriched within genomic regions that have been influenced by positive selection in the recent past. The observed pattern agrees with the prediction that deleterious alleles can dragged along to higher-than-expected allele frequencies due to the genetic linkage with beneficial alleles by the hitchhiking effect.

Second, I developed an integrative strategy to predict disease-causing non-coding variants in *FSH receptor*, a gene known to be associated with preterm birth, as a proof of principle. I sequenced protein-coding and conserved non-coding regions in preterm and term mothers, and conducted fine-mapping and transcription factor binding site analysis to narrow down the causal non-coding variants. Here, I find that in non-coding regions the causal variants can be resolved better by accounting for the expected effects of binding site mutations on the transcription regulatory network in addition to sequence conservation.

These results indicate that the comparative genomics will provide the new opportunity to explore deleterious and disease-causing genetic variation at an unprecedentedly high resolution across the genome and in a population especially if functional genomics can be integrated with comparative genomics.

Acknowledgements

At every turning point of my Ph.D. studies, I was indebted to so many great people who kindly made themselves available and gave me their encouragement and support. Without their help, this dissertation would have not been possible.

First, I would like to thank my mentor Justin Fay for the training, encouragement, and guidance that he provided throughout my Ph.D. He encouraged and challenged me to be self-motivated, and helped develop my career as an independent scientist. In addition, I would like to thank my thesis committee for providing me with productive and stimulating insight. I truly enjoyed working with them and highly value their constructive criticism. I would like to thank Kelle Moley for her clinical mentorship while I was in the Markey Pathophysiology Pathway, which was indeed an eye-opening experience to me. I would like to thank Lou Muglia for providing me the enthusiasm and encouragement throughout our collaboration on the preterm birth project.

I would like to thank all members of the Fay lab for their support, friendship, encouragement, and compassion. In particular, I am very grateful to Scott Doniger for providing his sincere foresight and help when I joined the Fay lab. I thank Betsy Engle and Devi Swain Lenz for their invaluable advice and many comments, particularly on my scientific presentations and writing. I also thank Betsy for encouraging me many times during the dissertation writing process, and Devi for providing editorial help while finishing my thesis. Juyoung Huh taught me almost everything I now know about laboratory work. Without her support, the experimental pieces of my research project would have not been possible. Additionally I would like to thank Maia Dorsett and Juyoung Huh for all the help with parenthood, all of which meant a lot to my family.

I also thank my colleagues in the Mitra and Druley lab for generously sharing their time, expertise, reagents, and equipment. Todd Druley and Francesco Vallania spent countless hours with me explaining the exquisite details of their experimental protocols. Moreover, I am deeply indebted to Francesco for his friendship.

I need to thank my fellow classmates in the Computational Biology Program for their friendship and compassion. Carlo Lapid and Igor Nikolskiy took me to indoor climbing every now and then. It literally brightened up my Sundays. I am also deeply indebted to Gurmukh Sahota for his countless advice on fatherhood and research, and his lessons on how to tie a tie, while I shadowed Kelle Moley in her reproductive endocrinology clinic for the Markey Pathway. Nick and Kat Lyle invited me to their parties on many occasions and took care of my often too frisky cat, Meow, while I visited my family in Korea.

I want to thank Juyeon Son and Minhoon Kim in Wisconsin for their friendship. They have been like a family to me all these years, and the thought that they live within driving distance and in the same time zone made me feel comfortable in St. Louis.

Last but not the least, I thank my family and Hyun's family for their tremendous support in taking care of my son Nate. I also have to show my deepest gratitude and respect to my wife Hyun, for raising Nate single-handedly and keeping our family in one piece across the Pacific. Without her courage and inspiration, I could not have started my Ph.D., let alone complete it.

To my wife,

Hyun Gyung Im

Table of Contents

ABSTRACT OF THE DISSERTATION ii
Acknowledgements iv
Table of Contents vii
List of Tables x
List of Figures xii
Abbreviations xiv
CHAPTER 1: Background 1
    Background on deleterious mutations 2
    Comparative genomics identifies deleterious variants 3
    Common deleterious polymorphism in the human genome 5
    Preterm birth 7
    This work 10
CHAPTER 2: Identification of deleterious mutations within three human genomes 12
  ABSTRACT 13
  INTRODUCTION 14
  RESULTS 17
    Identification of deleterious mutations in three human genomes 17
    An abundance of common deleterious mutations 19
    Estimation of the false positive rate 20
    Estimation of the false negative rate 23
    Comparison to SIFT and PolyPhen predictions 24
  DISCUSSION 26
  METHODS 34
    Comparative genomic dataset 34
    Likelihood ratio test 35
    Identifying deleterious mutations 36
    False positive rate 38
    False negative rate 39
    Allele frequency 40
    Recent gene duplication 40
    SIFT and PolyPhen predictions 40
    Estimation of rare deleterious mutations 41
  ACKNOWLEDGEMENTS 42
  FIGURES 43
  TABLES 46
  SUPPORTING INFORMATION 48
CHAPTER 3: Evidence for hitchhiking of deleterious mutations within the human genome 53
  ABSTRACT 54
  INTRODUCTION 55

RESULTS 57
  Simulated effect of hitchhiking on deleterious mutations 57
  Classification of deleterious and neutral nonsynonymous SNPs in humans 58
  Enrichment of deleterious SNPs in regions showing evidence of hitchhiking..... 59
  Hitchhiking regions show a similar enrichment of rare, intermediate and common deleterious SNPs 61
  Deleterious SNPs in regions showing population-specific patterns of hitchhiking 62
  Deleterious SNPs within and around genes under positive selection 63
  Disease-associated alleles within hitchhiking regions 64
  Disease-phenotype classification 66
  Genome clustering of deleterious SNPs 67
DISCUSSION 68
  Computer Simulations 74
  Classification of neutral and deleterious SNPs 75
  Correlation of SNP density with recombination rate 76
  SNPs in hitchhiking and non-hitchhiking regions 78
  Disease-associated alleles in hitchhiking and non-hitchhiking regions 81
  Genome clustering of deleterious SNPs 82
ACKNOWLEDGEMENTS 83
FIGURES 84
TABLES 91
SUPPORTING INFORMATION 93
CHAPTER 4: Fine-mapping of the genetic association of *FSH receptor* with preterm birth in the Finnish population 103
ABSTRACT 104
INTRODUCTION 105
RESULTS 106
  Identification of genetic variants in coding and non-coding regions 106
  Genetic variants altering protein sequence 108
  No enrichment of rare variants in cases 109
  Fine-map the association of common variants 110
  Candidate causal variants in risk and protective haplotypes 112
DISCUSSION 114
MATERIALS AND METHODS 119
  Human subjects 119
  Candidate regions 120
  Pooled high-throughput sequencing 121
  Identification of single nucleotide variants 122
  Comparison of allele frequencies between cases and controls 124
  Haplotype analysis 125
  Computational prediction of transcription factor binding sites 125
FIGURES 127
TABLES 131

    SUPPORTING INFORMATION 134
CHAPTER 5: Progress and future directions 144
References 152
Curriculum Vitae 179

List of Tables

Table 2.1. Summary of deleterious mutations found in three individuals and the reference genome.

Table 2.2. Deleterious mutations are enriched in rare frequency classes.

Table S2.1. A sequence of filters used to identify deleterious mutations in the Venter genome.

Table 3.1. The ratio of deleterious to neutral SNPs at low, intermediate and high frequencies.

Table 3.2. Deleterious SNPs within and around genes under positive selection.

Table S3.1. Characteristics of hitchhiking regions.

Table S3.2. The ratio of deleterious to neutral SNPs in iHS population-specific hitchhiking regions.

Table S3.3. Ratio of deleterious to neutral SNPs in three populations panels for regions identified by multiple tests of selection.

Table S3.4. Frequency of disease-associated alleles in hitchhiking and non-hitchhiking regions.

Table S3.5. Disease or phenotype associated SNPs within hitchhiking regions.

Table S3.6. Genes in hitchhiking regions with deleterious SNPs.

Table S3.7. Genomic regions enriched for deleterious SNPs.

Table 4.1. Summary statistics of sequencing and identified genetic variants.

Table 4.2. Nonsynonymous SNVs identified in subjects.

Table 4.3. Rare SNVs observed in either cases or controls, exclusively.

Table S4.1. Genomic regions selected for sequencing.

Table S4.2. PCR parameters used to amplify candidate regions.

Table S4.3. Rare SNVs observed in either cases or controls, exclusively, defined by minor allele frequency < 5%.

Table S4.4. Rare SNVs exclusively found in cases and located at conserved or nonsynonymous sites.

List of Figures

Figure 1.1. Classification of genetic variation by the direction of natural selection.

Figure 1.2. Heuristic to predict deleterious or neutral nonsynonymous variants based on sequence conservation.

Figure 2.1. Venn diagram of deleterious mutations identified in J. Craig Venter, James D. Watson and a Han Chinese individual.

Figure 2.2. Characteristics of deleterious mutations.

Figure 2.3. Comparison of SIFT, PolyPhen and the likelihood ratio test (LRT) predictions.

Figure S2.1. The fraction of mutations that are deleterious for different quality intervals.

Figure S2.2. Phylogenetic tree of species used and their synonymous rate (dS) of evolution.

Figure S2.3. The percentage of mutations predicted to be deleterious as a function of the total synonymous substitution rate (dS).

Figure S2.4. The estimated number of deleterious mutations as a function of the false discovery rate (FDR).

Figure 3.1. The effect of hitchhiking on neutral and deleterious polymorphism as a function of the rate of recombination.

Figure 3.2. Regions of low recombination are enriched for deleterious SNPs.

Figure 3.3. Rates of deleterious and neutral SNPs in hitchhiking and non-hitchhiking regions.

Figure 3.4. The ratio of deleterious to neutral nonsynonymous SNPs declines as a function of distance to the nearest hitchhiking region.

Figure 3.5. Enrichment of disease-associated alleles in hitchhiking relative to non-hitchhiking regions.

Figure S3.1. The effect of hitchhiking on neutral and deleterious polymorphism as a function of the rate and strength of advantageous and deleterious mutations.

Figure S3.2. The ratio of deleterious to neutral SNPs is associated with the rate of recombination.

Figure S3.3. Hitchhiking regions are enriched for deleterious SNPs.

Figure 4.1. Comparison of allele frequencies between pooled sequencing and Affymetrix genotyping.

Figure 4.2. Fine-map of the association of common variants.

Figure 4.3. Haplotype structure of common SNPs within the fine-mapped interval in normal Finns.

Figure 4.4. Haplotype frequencies in case and control pools.

Figure S4.1. Sequence context of candidate causal variant, rs12052281.

Figure S4.2. Sequence context of candidate causal variant, rs72822025.

Figure S4.3. Linkage disequilibrium between risk haplotype and positively selected p.S680N in Sub-Saharan Africans (YRI).

Abbreviations

ASW, African ancestry in the southwestern USA

CEU, Utah residents with ancestry from northern and western Europe from Centre d'Etude du Polymorphisme Humain

CHB, Han Chinese in Beijing, China

DEL, Deleterious nonsynonymous

dN, Nonynonymous substitution rate per nonsynonymous site

dS, Synonymous substitution rate per synonymous site

FDR, False Discovery Rate

FIN, Finnish in Finland

FSH, Follicle Stimulating Hormone

FSHR, FSH Receptor

GWAS, Genome-Wide Association Study

InDel, Insertion/Deletion

JPT, Japanese in Tokyo, Japan

LLR, Log Likelihood Ratio

LRT, Likelihood Ratio Test

NEU, Neutral nonsynonymous

OMIM, Online Mendelian Inheritance in Man

PCR, Polymerase Chain Reaction

QTL, Quantitative Trait Locus

SNP, Single Nucleotide Polymorphism

SNV, Single Nucleotide Variant

SYN, Synonymous

TF, Transcription Factor

YRI, Yoruba in Ibadan, Nigeria

CHAPTER 1: Background

Background on deleterious mutations

Deleterious mutations are defined by their reduction of fitness to survive and reproduce relative to the wild-type allele. Since deleterious alleles lower an individual's reproductive output, deleterious alleles are selected against by negative selection, and thus have low allele frequencies in the absence of other interfering factors. However deleterious mutations can be maintained in a population. The main mechanism maintaining deleterious mutations in a population is the balance between spontaneous germ-line mutation and negative selection.

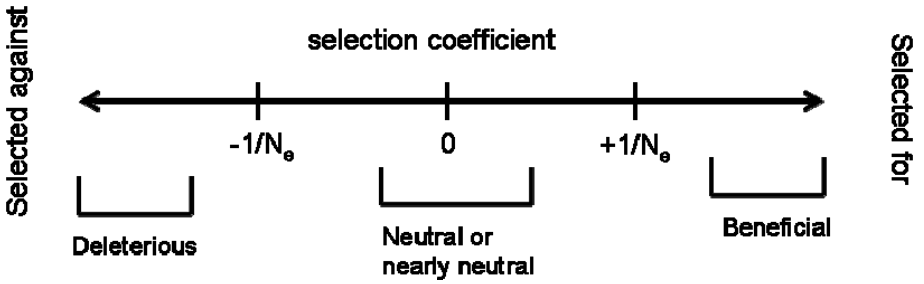


Figure 1.1. Classification of genetic variation by the direction of natural selection. The deleterious, neutral, and beneficial alleles are classified by the selection coefficient, which is the fitness of the variant allele relative to the wild-type. The N_e stands for the effective population size.

Although the majority of deleterious mutations are hidden by rare allele frequencies and heterozygosity, the genetic load of deleterious mutations can be observed indirectly by a number of means. For instance in the offspring of consanguineous marriage, many rare recessive deleterious mutations can become homozygous and cause an increase of infant mortality and morbidity [1]. Contrarily, the hybrid between pure lines has an increased fitness compared to the original parental lines since deleterious alleles become hidden heterozygously in the hybrid [2].

The fitness impact of deleterious alleles often takes the form of genetic disorders in humans. The majority of Mendelian disorders clearly reduce the fitness of individuals [3], and many complex genetic disorders are also attributed to aggregate burden of rare deleterious alleles, as previously shown in serum cholesterol phenotype and many others [4-6].

In contrast, beneficial alleles, which increase fitness, are selected for and rapidly spread in population. When the fitness change is much smaller than $\sim 1/N_e$, where N_e is the effective population size, the variant alleles behave nearly neutrally and influenced more by genetic drift than by natural selection (Figure 1.1) [7].

Comparative genomics identifies deleterious variants

The recent advances in comparative genomics enable the identification of deleterious variants to single-site resolution. The basis of two main approaches is that the functional constraints at each site is observed through sequence conservation across multiple species, and a mutation disrupting the patterns of sequence conservation is highly likely to be deleterious in terms of the fitness.

One of the two approaches, used by software such as SIFT [8] and PolyPhen [9], is to transform protein sequence alignments to a position-specific scoring matrix, which represents how likely each of 20 amino acid residues is evolutionarily allowed at each position (Figure 1.2). From the position-specific scoring matrix, a heuristic classifier determines whether a mutation disrupts the site-specific amino acid preference based on somewhat arbitrarily defined scoring cut-offs.

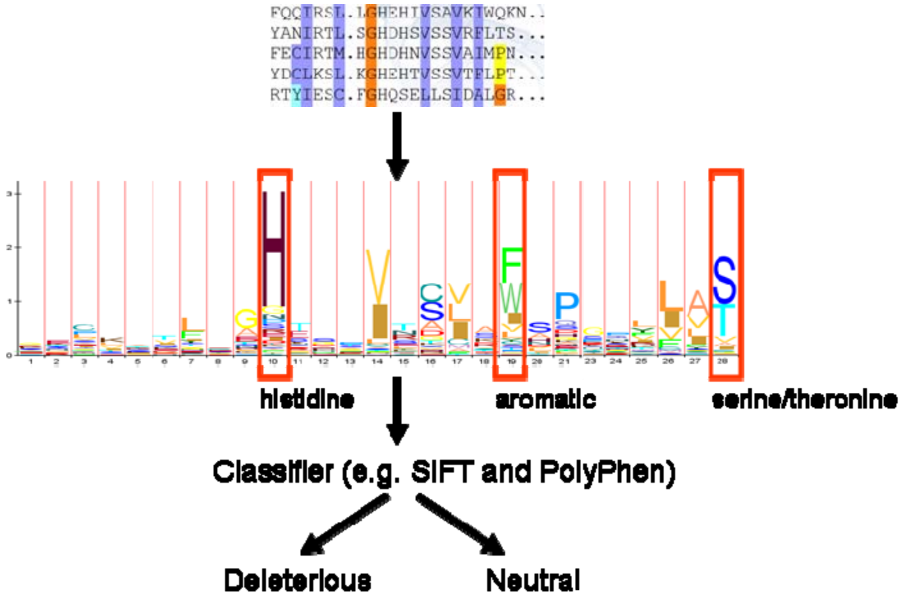

FQQTRSL.LGHEHIVSAVKTWQKN..
YANIRTL.SGHDHSVSSVRFLTS...
FECIRTM.HGHDHNVSSVAIMPN...
YDCLKSL.KGHEHTVSSVTFLPT...
RTYIESC.FGHQSELLSIDALGR...
histidine
aromatic
serine/theronine
Classifier (e.g. SIFT and PolyPhen)
Deleterious
Neutral

Figure 1.2. Heuristic to predict deleterious or neutral nonsynonymous variants based on sequence conservation. Protein-sequence-based approach applied to distinguish deleterious from neutral variants by comparative genomics software such as SIFT and PolyPhen. The position-specific scoring matrix are represented here as a sequence logo. The conserved sites stand out as high information-content peaks, highlighted in red boxes, and at each position, the kinds of evolutionarily allowed amino acid residues are annotated below.

The other approach is a nucleotide-sequence-based test, mainly used for non-coding regions [10; 11]. Because the functional constraints in non-coding regions are not as simple as the biochemical similarity observed in protein-coding regions, sequence conservation is tested by taking advantage of a probabilistic model of DNA sequence evolution and the estimated rate of neutral substitutions. Until recently, to our knowledge, the nucleotide-sequence-based tests have not been commonly applied to protein-coding regions.

As a hybrid between the above two approaches, a likelihood ratio test (LRT) for significant sequence conservation has been recently developed to identify deleterious mutations in protein-coding regions for the yeast genome [12]. The LRT directly accounts for the evolution of codons similarly to the protein-sequence-based methods but also relies on a probabilistic model of DNA sequence evolution as the nucleotide-sequence-based methods. The codon evolution model utilized by the LRT[13] has been well characterized through its extensive use in numerous evolutionary genetic studies since it was devised almost two decades ago.

Common deleterious polymorphism in the human genome

In agreement with the idea that the majority of variants disrupting sequence conservation are under negative selection, variants predicted as deleterious by comparative genomics have significantly lower allele frequencies compared to variants predicted as neutral [9]. However, at the same time, there exist an abundant number of deleterious variants segregating at common allele frequencies in human population. For instance, in the HapMap, ~50% and ~20% out of single nucleotide polymorphisms (SNPs) predicted as deleterious by PolyPhen have minor allele frequencies over 5% and even 25%, respectively [14]. The observed abundance of common deleterious SNPs is not simply because rare variants are more difficult to ascertain for the HapMap than common variants. A similar abundance of common deleterious SNPs was observed using the LRT within three personal genomes, in which both rare and common variants were ascertained without bias by sequencing [15].

The abundance of such common deleterious SNPs has been noted in public SNP databases as early as 2001 and 2002 [9; 16]. At the time, however, a systematic study could not be undertaken because the number of known nonsynonymous variants was still relatively small, and the allele frequency of deposited SNPs was heavily affected by ascertainment biases. Nonetheless, it was debated whether common deleterious SNPs identified in the SNP repositories were false positives or not. Sunyaev et al. argued that they were not entirely false positives based on the false positive rate estimated by using ancestral mammalian substitutions as negative controls [9]. On the other hand, Ng and Henikoff argued that common deleterious SNPs ascertained in a small number of healthy individuals are mostly false positives because complementation assays do not find a detectable phenotypic effect for ~20% SNPs disrupting sequence conservation [16]. Clearly, the sensitivity of complementation assay is too low to detect negative selection, which can operate at the fitness change greater than $\sim 1/N_e$. Humans have a relatively small N_e, but it is still in the order of ~10,000. Thus, for most of common deleterious SNPs, the fitness deficit can be as tiny as 0.01%.

A recent population simulation study estimates that 25-40% of deleterious SNPs predicted by PolyPhen are effectively neutral or weakly deleterious with the fitness reduced by less than 0.1% [17]. This study utilized the unbiased allele frequency spectrum, which became available with the recent advances in high-throughput sequencing technology, and assumed a reasonably demographic history that closely fit the observed data. The estimated fitness impact of 0.1% or less is small enough to allow them to reach very high allele frequency in human populations. Although this explains why some deleterious mutations can reach common allele frequencies, it does not explain why mutations disrupting strong sequence conservation have such tiny fitness effects.

Preterm birth

Preterm birth, a delivery of a fetus prior to the completion of gestational week 37, is one of the major unsolved healthcare problems. In the United States, the rate of preterm birth has been steadily increasing for the past two decades and reached 12.3% of all pregnancies in 2008 [18; 19]. Despite this alarming prevalence, current treatment options are not effective in preventing or delaying preterm birth [20][Smith 09], and once born, premature infants are at a high risk of mortality and life-long morbidity [21].

Current genetic epidemiological evidence suggests that preterm birth is a typical complex genetic disorder, defined by the complex interactions between multiple genetic and environmental risk factors. The most conclusive evidence for the genetic contribution to preterm birth comes from multiple twin studies that report higher correlation of preterm birth outcomes between monozygotic compared to dizygotic twins [22-25]. Interestingly, these studies estimate that the heritability (i.e. the fraction of observed phenotypic variation explained by additive phenotypic effects of alleles transmitted to the next generation) is consistently high for maternal genes (17-36%) but weak or nonexistent for paternal genes, which may act through fetus.

Which genetic mechanism can explain these patterns of heritability for preterm birth? It can be explained simply by genes acting in the pregnant mother, but it is also possible that preterm birth is inherited by other genetic mechanisms such as maternally imprinted genes acting in the fetus, mitochondrial inheritance or vertical transmission of the vaginal and uterine microbiome from mother to fetus [18].

With respect to the genetic imprinting in the fetus, the lack of paternal heritability is somewhat unexpected, considering that the parental conflict hypothesis suggests the contribution of both paternal as well as maternal heritability to preterm birth. Theorized to explain the evolutionary origin of genetic imprinting, the parental conflict hypothesis posits that paternal genes in the fetus are selected in the direction to increase the fitness of the offspring at the expense of the mother, whereas maternal genes are selected for the direction to diversify resources to multiple offspring [26]. If the length of gestation in humans follows the parental conflict hypothesis indeed, paternally inherited genes in the fetus may favor relatively longer gestation and thus protect the fetus from preterm birth. In contrast, maternally inherited genes in the fetus may favor relatively shorter gestation and thus predispose the fetus to preterm birth. The discrepancy between the expectation and observed lack of paternal heritability may be indicating that perhaps the parental conflict may be more complicated than we think, even if the imprinting is indeed involved in the regulation of gestational length.

The potential role of the vaginal and uterine microbiome is also an interesting possibility, since the sign of infection and inflammation is often observed in preterm birth [27; 28]. Even in this case, the human genome could have co-evolved with the microbiome in order to conserve maternal resources by terminating a compromised pregnancy or pregnancy under adverse environmental conditions. As a consequence, the nuclear genome may have genetic variants predisposing women to preterm delivery by lowering a threshold for a stress response to the microbiome.

An alternative theory is that preterm birth may be an extreme phenotype of heritable genetic variation in gestational length that is spontaneously determined [29]. In support of this idea, about a half of preterm birth cases seem spontaneous without identifiable cause [21] . In addition, the history of post-term as well as preterm deliveries strongly predicts the gestational length of subsequent pregnancies [30].

Interestingly, chimpanzees seem to have very low susceptibility to preterm birth compared to humans [31]. Why did humans become susceptible to a high rate of preterm birth despite the detrimental fitness cost to prematurely born infants? One theory is that humans may have evolved to give birth earlier than chimpanzees as the human ancestor evolved to have a larger brain and a narrower pelvis, which increased the risk of cephalopelvic disproportion and threatened the survival of both mother and fetus, particularly in post-term birth [29]. Another interesting possibility is that positive selection may have increased the allele frequency of genetic variants promoting the risk for preterm birth by the hitchhiking effect. The African American population has a higher risk of preterm birth than other populations even after correcting for the socioeconomic confounders [32; 33]. Interestingly, there is evidence that *FSH receptor* (*FSHR*), which is known to be associated with the risk for preterm birth, was positively selected in Sub-Saharan Africans [29; 34].

This work

The goal of my thesis is to understand what evolutionary model can explain the small fitness effects of an abundant number of common deleterious SNPs found in human population. This question is implicated with why certain common disease alleles have higher-than-expected allele frequencies.

In chapter 2, I adapt the likelihood ratio test, previously developed for the yeast genome, to the human genome to identify deleterious nonsynonymous variants in humans. By applying this test to three personal genome sequences, I count the number of deleterious nonsynonymous SNPs carried in each person and begin to address explanation of the abundance of common deleterious SNPs found in the three personal genomes. Specifically, I test the possibility of false positives due to multiple hypothesis testing, violation of model assumptions, recent gene duplication, and relaxation of selective constraints on a specific functional category of genes.

In chapter 3, I explore the hypothesis that positive selection may have influenced the allele frequency of deleterious SNPs and disease alleles in the human population by the hitchhiking effect. When the recombination rate is low enough, the apparent fitness of a deleterious allele is determined not only by its own fitness but also by the combined fitness of other linked alleles.

In chapter 4, I fine-map the candidate causal variants underlying the known genetic association of *FSH receptor* with preterm birth in the Finnish population by sequencing protein-coding and conserved non-coding regions in this locus. Here, I also examine whether positive selection on *FSH receptor* has influenced the allele frequency of candidate causal variants promoting the risk of preterm birth.

CHAPTER 2: Identification of deleterious mutations within three human genomes

Sung Chun[1] and Justin C. Fay[1,2]

[1]Computational Biology Program, Washington University, St. Louis, MO

[2]Department of Genetics, Washington University, St. Louis, MO

This work was done in collaboration with Justin Fay. My contribution was design of the experiments, execution of the experiments and analysis of the data. The web server for this algorithm is available from the Fay lab website. This chapter is a reprint of the manuscript originally published in Genome Research in 2009. Large supplemental data are available from the Fay lab website.

ABSTRACT

Each human carries a large number of deleterious mutations. Together, these mutations make a significant contribution to human disease. Identification of deleterious mutations within individual genome sequences could substantially impact an individual's health through personalized prevention and treatment of disease. Yet, distinguishing deleterious mutations from the massive number of non-functional variants that occur within a single genome is a considerable challenge. Using a comparative genomics dataset of 32 vertebrate species we show that a likelihood ratio test can accurately identify a subset of deleterious mutations that disrupt highly conserved amino acids within protein coding sequences and that are likely to be unconditionally deleterious. The likelihood ratio test is also able to identify known human disease alleles and performs as well as two commonly used heuristic methods, SIFT and PolyPhen. Application of the likelihood ratio test to three human genomes reveals 796-837 deleterious mutations per individual, ~40% of which are estimated to be at less than 5% allele frequency. However, the overlap between predictions made by the likelihood ratio test, SIFT and PolyPhen is low; 76% of predictions are unique to one of the three methods and only 5% of predictions are shared across all three methods. Our results indicate that only a small subset of deleterious mutations can be reliably identified but that this subset provides the raw material for personalized medicine.

INTRODUCTION

Mutations that impact an organism's ability to survive and reproduce are deleterious and must be eliminated by natural selection in order to ensure the long term survival of a species [35]. Removal of deleterious mutations from the gene pool requires a substantial number of genetic deaths and incurs a considerable reproductive cost [36]. However, many deleterious mutations persist for hundreds of generations or more before they are removed since their effects are largely masked in the heterozygous state [37].

The presence of deleterious mutations within the human population has a significant impact on human health. Inbreeding causes an increase in child morbidity and mortality and suggests that each human carries a sufficient number of deleterious mutations that if made homozygous would together result in premature death [1]. In addition, most mutations that cause monogenic diseases are clearly deleterious. Diseases with a complex genetic basis are also likely to be affected by deleterious mutations. In support of this possibility, rare variants that have been associated with complex human diseases are often predicted to be deleterious [5; 6]. However, even when rare variants have been associated with human disease there is considerable uncertainty as to which rare variants are responsible for the association.

A number of methods have been developed to identify deleterious and/or disease-causing mutations within protein coding sequences. These methods predict whether an amino acid altering mutation is deleterious or disease-causing based on physico-chemical properties [38], population frequency [39; 40], protein structure [9; 41; 42], and cross-species conservation [8; 9; 42]. For a comprehensive review of methods see Ng and

Henikoff [43]. While these methods can identify 40-90% of disease-causing mutations, the rate of false positives is uniformly high, 10-20% [43]. The high rate of false positives may be caused by most predictions relying on some aspect of sequence conservation, which is difficult to accurately model. One factor that confounds evolutionary models is that not all disease-causing mutations are conserved, presumably due to compensatory changes elsewhere in the protein [44]. A further complication is that mutations in highly conserved sequences do not always produce phenotypes that are easily noticeable [6; 45]. Regardless of the cause, the accuracy of cross-species conservation depends on both the assumptions and parameterizations of evolutionary models that relate sequence conservation to fitness or function [43; 46].

Genome sequencing of a large number of closely related species makes it possible to develop better parameterized evolutionary models that more accurately predict human deleterious mutations. Closely related species minimize the frequency of compensatory changes that enable functional sites to diverge. Conversely, a large number of species maximizes the phylogenetic distance among taxa required to accurately distinguish selectively constrained sites from neutral sites that have not yet diverged [47]. A number of models use putatively neutral classes of sites to distinguish functionally constrained and neutral sites based on closely related genomes, e.g. [10; 12; 48]. Although evolutionary models may not identify all disease-causing mutations, they provide a probabilistic framework in which the subset of disease-causing mutations that disrupt highly conserved amino acid positions can be accurately identified given enough phylogenetic information.

Prediction of deleterious mutations within individual human genomes has the potential to impact both the prevention and treatment of disease at an individual level. Although a number of human genomes have been sequenced, the number of nonsynonymous variants predicted to impact protein function varies widely. Using SIFT [49], 14% (1,455) of nonsynonymous variants within the Venter genome were predicted to impact protein function [50]. Using PolyPhen [51], 7.3% (~770) of nonsynonymous variants within the Watson genome were predicted to impact protein function [52]. However, comparison of these predictions is difficult since they are based on different data, different models and use different methods to control for sequencing errors, a potentially important source of false positives [50].

To identify and characterize deleterious mutations present within an individual genome we examined three recently sequenced human genomes [52-54]. Using a likelihood ratio test [12], we identified 796-837 amino acid altering mutations per genome that disrupt highly conserved amino acids. Comparison with two other methods, SIFT and PolyPhen, revealed only a small amount of overlap among the three methods and suggests that multiple methods should be used when trying to identify deleterious mutations in humans.

RESULTS

Identification of deleterious mutations in three human genomes

Mutations that alter evolutionarily conserved sequences are likely to be deleterious and have a negative impact on fitness. To distinguish functionally constrained and unconstrained amino acid positions we generated a comparative genomics dataset using 32 vertebrate species (Methods). Multiple alignments of orthologous protein coding sequences were generated for 18,993 human genes. The synonymous substitution rate was estimated to be 12.2 across all species and 4.7 across the eutherian clade, placental mammals, similar to previous studies (Methods). The large amount of divergence among these species implies that unconstrained amino acid positions should rarely be conserved across species and mutations that alter conserved amino acids are likely to be deleterious.

To identify deleterious mutations present within an individual human genome we examined nonsynonymous variants present within the genomes of J. Craig Venter, James D. Watson, and a Han Chinese male [52-54]. To eliminate sequencing errors we only used a subset of high quality alleles, a Phred quality score of 60 or greater for the Venter and Chinese genome and a variant score of 70 or greater for the Watson genome [52]. After eliminating 24-26% of variants that occur in regions without a sufficient number of aligned orthologs for accurate inference of functional constraint, less than ten eutherian species, we analyzed 5,417-5,707 nonsynonymous variants per individual.

Using a likelihood ratio test, we identified between 796 and 837 deleterious mutations in the three diploid genomes, ~15% of those tested ($P < 0.001$, Table 2.1). The likelihood ratio test compares the probability of the data under a conserved model, allowing for any level of selective constraint, relative to a neutral model, where there is no difference between the nonsynonymous and synonymous substitution rate (Methods). Most of the deleterious mutations, 59-62%, were individual-specific (Figure 2.1). In addition, the majority of deleterious mutations, 76-83%, were present in a heterozygous state. However, the true frequency of heterozygotes is likely higher since homozygotes are more easily identified than heterozygotes and some heterozygotes may be misclassified as homozygotes [52-54]. In addition to deleterious mutations present within the three individual genomes, we identified another set of 838 deleterious mutations that occur in the reference genome in comparison to either the Venter, Watson or Chinese genomes (Methods). Out of the 838 deleterious mutations, 474 were specific to the reference genome and were not present in either the Venter, Watson or Chinese genome.

Consistent with previous studies [55], the frequency of deleterious mutations is lower on the X chromosome compared to the autosomes, 10.5% relative to 18.1%, respectively ($P < 0.05$, Fisher's Exact Test). Only 14 out of 1,928 deleterious mutations were found on the X relative to 119 out of 8,761 neutral variants. These results indicate that partially recessive deleterious mutations are more rapidly eliminated from the X chromosome.

An abundance of common deleterious mutations

Most deleterious mutations are maintained at low population frequencies due to negative selection [4]. However, 435/1928 (23%) of the deleterious mutations are present in more than one of the three genomes, suggesting that at least some of the deleterious mutations may be common (Figure 2.1). Consistent with negative selection, the fraction of nonsynonymous alleles that are deleterious is lower for those that are shared among the three individuals, 19%, relative to those that are individual-specific, 26-32% (Table 2.2). To more accurately quantify the frequency of deleterious mutations we used the HapMap phase II and III panels and found that 1121/1928 deleterious mutations (58%) are common, greater than 5% allele frequency in at least one of the three HapMap panels. Surprisingly, many deleterious mutations have reached intermediate to high frequencies. Out of 1,928 deleterious mutations typed by the HapMap project, 925 (48%) are at frequencies greater than 20%, 472 (24.5%) are at frequencies greater than 50% and 163 (8.5%) are at frequencies greater than 80%.

Deleterious mutations can become common if their effects are buffered by recently duplicated genes. Degeneration of recently duplicated genes is expected under both nonfunctionalization and subfunctionalization models of gene duplication [56]. Tabulating all deleterious mutations that occur in recently duplicated proteins (Methods), 70 out of 1928 deleterious mutations (3.6%) occur in duplicated genes (Figure 2.2A). Although there is a significant enrichment of deleterious relative to neutral alleles in duplicated genes (Chi-square test, $P < 0.01$), only 7 of the common deleterious alleles occur in duplicated genes. Thus, most common deleterious mutations are not buffered from selection by gene duplication.

If negative selection is weak, a substantial number of deleterious mutations can become common by random genetic drift. To examine whether common deleterious alleles are under weaker selection than those that are rare we calculated the frequency of deleterious mutations in perfectly conserved sites, the frequency of deleterious mutations that caused radical amino acid substitutions, defined by a BLOSUM62 score of less than negative two, and the frequency of deleterious mutations for which the deleterious allele was observed in one or more non-eutherian species. Combining the data from all three genomes, rare deleterious mutations are more likely to occur in perfectly conserved sites, are more often radical and are less likely to be present in non-eutherian species (Chi-square test, $P < 0.01$, Figure 2.2B). This result suggests that rare deleterious mutations are under stronger negative selection than common alleles, consistent with both empirical and theoretic studies [4; 17; 57]. However, some of the common deleterious mutations may also be false positives.

Estimation of the false positive rate

A number of factors could lead to false positive prediction of deleterious mutations. To estimate the rate of false positives we examined: multiple hypothesis testing, sequencing errors and model assumptions. Given an uncorrected P-value cutoff of 0.001 and the number of tests (Table 2.1), we estimate 5-6 false positive predictions per individual due to multiple hypothesis testing. A P-value cutoff of 0.01 results in a prediction of 1,120-1,197 deleterious mutations within the three genomes but is also expected to include a higher number of false positives, 54-57 per individual.

Sequencing errors can also result in false positive predictions. Given a Phred quality score cutoff of 60, 0.47 false positives are expected due to sequencing errors in the Venter genome (Methods). For the reference genome 53 false positives are expected assuming three nucleotide substitution errors per 1 Mbp [58]. The rate of sequencing errors is more difficult to know for the Watson and Chinese genomes since a complex series of quality filters were used and quality values derived from new sequencing technologies may not be entirely accurate, e.g. [59].

To empirically estimate the impact of sequence errors on prediction of deleterious mutations, we tested whether the frequency of deleterious mutations is affected by the quality score cutoff. Consistent with a minor effect of sequencing errors, the fraction of deleterious mutations in the Venter genome is nearly constant for Phred values greater than 60 (Figure S2.1A). High-quality nonsynonymous heterozygous variants from each genome were split into two groups of roughly equal numbers using their quality scores. In each case the proportion of variants called deleterious was slightly higher (0.3-3.6%) in the group with higher compared to lower quality values (Figure S2.1). This result suggests that few of the deleterious mutations identified by the likelihood ratio can be attributed to sequencing errors.

False positive predictions can also arise if the assumptions of the likelihood ratio test are violated. The likelihood ratio test assumes a neutral substitution rate estimated from synonymous sites and so false positive predictions may occur for variants that occur in positions with lower than average mutation rates. Two significant sources of variation in mutation rates are methylated CpG sites and regional variation across the genome [60; 61]. To examine the effect of regional variation across the genome we estimated the number deleterious mutations using a synonymous substitution rate of 10.1, two standard deviations below the mean. The standard deviation due to regional variation in mutation rates was obtained from the coefficient of variation (8.75%) reported for mouse-rat divergence in ancestral repeats at a scale of 1 Mb [61]. To account for overestimation of the synonymous substitution rate due to methylated CpG sites all CpG-prone sites were eliminated and the synonymous rate was found to be reduced by 12% in the mammalian clade and by 5% overall. Because regional variation in mutation rates generated a greater reduction in the synonymous rate, 1 - 10.1/12.2 (17%), we used a total synonymous rate of 10.1 and found 621 deleterious mutations remain significant in the Venter genome. This implies that only 22% (621/796) of highly conserved amino acid positions could be due to low mutation rates rather than negative selection. The true percentage is much lower since the only a small fraction of sites tested, ~2.5%, are expected to have a mutation rate two standard deviations below the genome average.

The false positive rate can also be estimated using a set of negative controls. Previous studies estimated the rate of false positives using nonsynonymous SNPs present in humans [16] or substitutions between humans and other mammals [51]. Given that a significant number of deleterious mutations and sequencing errors may be present within any sequenced genome, we generated a set of negative controls using 39,028 nonsynonymous substitutions that occurred after orangutan but before chimpanzee split off from the lineage leading to humans (Methods). The likelihood ratio test identified 2,633 substitutions (6.7%) as deleterious ($P < 0.001$). This is lower than that estimated for both PolyPhen (9%)[9] and SIFT (20%)[16], but is still much higher than that caused by sequencing errors or variation in mutation rate. When the negative controls were subdivided into three classes by the severity of amino acid changes using Grantham's distance [38], the likelihood ratio test identified more conservative substitutions (8.3%), compared to moderate (6.1%) or radical substitutions (5.1%).

Estimation of the false negative rate

The false negative rate was estimated using known disease-causing missense mutations in the OMIM database (Methods). Out of 5,493 mutations, 3,947 (71.9%) are significant by the likelihood ratio test ($P < 0.001$). The false negative rate of the likelihood ratio test (28%) is similar to that of both SIFT (31%) [16] and PolyPhen (31%) [9]. Of the 1,546 disease mutations that were not identified, 644 (42%) occur at positions with little or no comparative genomic data (fewer than 10 eutherian mammals or a synonymous substitution rate less than or equal to two). Although the likelihood ratio test does not explicitly account for the type of amino acid change, the false negative rate is lower for radical disease-causing mutations (25.2%) than for moderate or conservative disease-causing mutations (30.2 and 29.7%, respectively). Such differences are not due to varying availability of comparative genomic data. A similar fraction of mutations lack sufficient comparative genomic data in all three classes of amino acid changes.

Comparison to SIFT and PolyPhen predictions

A variety of methods have been developed to specifically identify mutations that cause human disease. Two of the most commonly used methods, SIFT [49] and PolyPhen [51], use heuristic measures of cross-species conservation along with the type of amino acid change to predict human disease mutations. SIFT uses a median conservation score rather than synonymous sites to measure protein conservation. PolyPhen also uses a normalized cross-species conservation score and combines this with a variety of protein structural features when available. Both methods use non-redundant protein databases and so make use of a much more diverse set of species than the likelihood ratio test.

Compared to the 796 deleterious mutations identified by the likelihood ratio test (LRT), SIFT predicted 890 intolerable mutations and PolyPhen predicted 768 possibly damaging and 555 probably damaging mutations in the Venter genome (Methods). The overlap between these predictions is low but significantly greater than chance (Figure 2.3). Out of all predictions, 18%, 30% and 28% were specific to PolyPhen, SIFT and the LRT, respectively, and 93 mutations (5%) were predicted by all three methods. The overlap of all three methods is greater than the 5.9 mutations (0.3%) expected by chance. Each of the methods predicted a similar fraction of mutations that were not predicted by the other two methods; 57%, 59% and 61% of predictions made by PolyPhen, SIFT and the LRT, respectively, were not predicted by either of the two other methods.

The low overlap among predictions made by the LRT, SIFT and PolyPhen is not due to differences in coverage. Out of the 7,534 high-quality mutations present within the Venter genome, PolyPhen, SIFT and the LRT generated predictions for 6,746 (90%), 5,401 (72%) and 5,645 (75%) mutations and all three methods generated predictions for 4,303 (57%) mutations. Most of these differences are likely the result of each method requiring a sufficient number and diversity of aligned sequences in order to make a prediction and each method using a different set of sequences and alignments. Despite the differences in coverage, the overlap among the methods remains low when only the 4,303 mutations with predictions made by all three methods were compared (Figure 2.3).

To determine whether the low overlap can be attributed to alignment differences, SIFT was run using the same alignments as those used for the likelihood ratio test. Figure 2.3B shows that tuning SIFT's median conservation score to generate a similar number of deleterious predictions as the LRT improved the overlap between SIFT and the LRT from 269 (19%) to 404 (34%). Thus, many but not all of the differences between SIFT and the LRT are due to differences in the sequences and/or alignments used to identify evolutionary conserved amino acids.

The remaining differences between SIFT and the LRT predictions can be attributed to a number of factors. Out of 392 mutations that were predicted deleterious by the likelihood ratio test but not SIFT, 226 (58%) occurred in highly conserved proteins and were not identified as intolerable by SIFT due to the high median conservation score of the protein. SIFT uses the a median conservation score of a protein to eliminate predictions based on highly similar proteins. Because the LRT uses synonymous site divergence to calibrate conservation many cases are likely false negative predictions made by SIFT. In 184/392 (47%) cases the amino acid predicted to be deleterious by the LRT was found in one or more vertebrate species outside of the eutherian mammals.

Although this indicates that these alleles are probably neutral in distantly related species, the likelihood ratio test implies that amino acid conservation within the eutherian mammals is significant and so the allele may be deleterious in humans.

A number of factors are also likely to contribute to the 404 mutations predicted to be intolerable by SIFT that were predicted neutral by the LRT. A total of 126/404 (31%) cases showed marginal significance by the likelihood ratio test, ($0.001 < P < 0.01$). Some of these cases may be due to the lower power of the LRT since SIFT differentiates between conservative and radical amino acid changes, a factor known to be predictive of function [62]. The LRT may also have lower power for sites with a significant number of missing species. A total of 133/404 (33%) cases occurred in alignments for which there were not enough species to apply the likelihood ratio test, as defined by a total synonymous substitution rate less than four. Because the LRT alignments are based mostly on closely related mammalian species, some of these cases are likely false positive predictions made by SIFT. However, these predictions may also be a consequence of forcing SIFT to use alignments from closely related species and would not be predicted by SIFT when run using its own set of alignments.

DISCUSSION

Identification of deleterious mutations within individual genomes has the potential to directly impact both health and reproductive decisions. The wealth of comparative genomics data now available makes it possible to rapidly identify all mutations that disrupt highly conserved amino acid positions and that are likely to be deleterious. Here, we have evaluated the ability of a likelihood ratio test (LRT) to identify deleterious mutations within three human genomes. We identified a similar number of deleterious mutations, 796-837, in three human genomes and showed that only a handful are expected to be false positives due to sequencing errors or multiple hypothesis testing. While the likelihood ratio test performs as well as two other commonly used methods, SIFT and PolyPhen, the overlap among predictions made by different methods is disturbingly low. Analysis of these differences indicates that both the algorithms as well as the alignments used to identify conserved sites make a significant contribution to the low overlap among predictions. Our results suggest that multiple methods should be used to reliably identify deleterious mutations for association with human disease.

The likelihood ratio test is conceptually distinct from other comparative genomic methods. To our knowledge, all previous methods designed to identify deleterious mutations (see [43] for review) rely on heuristic procedures to distinguish sites within a protein that are conserved from those that are not conserved. This is achieved by selecting sequences that are not too closely or too distantly related to the sequence of interest and comparing the degree of conservation at the site of interest to other sites in the protein. The advantage of this approach is that the phylogenetic relationship and evolutionary distance among the sequences is not required. However, there are also limitations to this approach since no distinction is made between distantly related proteins that are highly conserved and closely related proteins that have evolved rapidly. Although proteins can always be selected that show a desired degree of similarity, compensatory changes are more likely to have occurred in distantly related proteins. In comparison, the likelihood ratio test was designed to explicitly model phylogenetic relationships using a probabilistic framework and to minimize compensatory changes, which are thought to be common [44] and can cause false negatives, by using closely related vertebrate species: mammals, chicken, frog and fish.

The likelihood ratio test also differs from other comparative genomic methods in that all amino acid changes are treated the same rather than weighting radical and conservative amino acid changes differently. While this is expected to reduce the power of the likelihood ratio test, empirically, both the false positive and the false negative rates of the likelihood ratio test are lower for radical relative to conservative amino acid changes. This may be a consequence of the genetic code and other mutational parameters being correlated with the ratio of radical to conservative amino acid changes [63].

The likelihood ratio test performs similar to two other commonly used comparative methods, SIFT [9] and PolyPhen [16]. We found that the likelihood ratio test was able to identify 72% of known disease-causing mutations, slightly higher than that reported for PolyPhen (69%) and SIFT (69%). However, it should be noted that these numbers are not directly comparable since accuracy depends on how it is measured. For example, SIFT and PolyPhen detected 70% and 72% of deleterious mutations when applied to the same protein mutation database [64]. Many, 644/1,546 (42%), of the disease mutations that were not identified by the likelihood ratio test can be attributed to the absence of sufficient comparative genomic data, i.e. not enough species. The remaining cases may be attributed to compensatory changes that allowed sites that cause human disease to change in non-human species.

The false positive rate of the likelihood ratio test (LRT) is estimated to be lower than that estimated for SIFT and PolyPhen. While few false positives are expected due to sequencing errors and multiple hypothesis testing, we found 6.7% of negative controls were called deleterious by the likelihood ratio test. The rate of false positives is lower than that estimated for both SIFT (20-30%) and PolyPhen (9-28%)[9; 16; 64]. However, some of these differences may be due to the use of different negative controls. Regardless of slight methodological differences, the frequency of putatively neutral mutations called deleterious (6.7%) is only half the fraction predicted to be deleterious in the Venter, Watson and Chinese genomes (14-15%), suggesting a false discovery rate of nearly 50% within individual genomes. Yet, the rate of false positives estimated from the negative controls may be an overestimate since not all human ancestral substitutions may be neutral. Some ancestral substitutions may be weakly deleterious mutations that became fixed by genetic drift, a process which can be exacerbated by a small effective population size. Other ancestral substitutions may be advantageous mutations that alter the function of previously conserved amino acids due to a change in the environment. Thus, ancestral substitutions provide an upper bound on the number of false positive predictions.

The potentially high rate of false positives may explain the large number of common alleles predicted to be deleterious in the three genomes. Similar to the negative controls, some of the common alleles may occur in sites that have been under negative selection in primates and other mammals but have recently become neutral along the human lineage. However, it is also possible that some deleterious mutations have increased in frequency due to hitchhiking along with recent positively selected mutations. Further work will be needed to address the impact of positive selection on the number and frequency deleterious mutations.

Despite similar performance of the LRT, SIFT and PolyPhen, the overlap among predictions made by the three methods is low, 5% (Figure 2.3). The majority of differences are not due to cases where one or more methods did not make a prediction due to limited or insufficient data; the overlap remains low for cases where all three methods generated predictions (Figure 2.3). Inspection of cases where the three methods disagreed revealed two general explanations for the disagreements. First, distantly related species can have a strong influence on the prediction of deleterious mutations and each method uses a different set of distantly related species. Distantly related species tend to have a large effect since fewer sites are conserved and they are more likely to carry the deleterious allele due to compensatory changes. The impact of distantly related species is significant since each method measures conservation using a different set of distantly related species. Both SIFT and PolyPhen can use any homologous protein sequence and generate alignments using different non-redundant databases of protein sequences. In contrast, the likelihood ratio test only uses a few non-mammalian species: chicken, frog and five fish species (Figure S2.2). One of the goals of developing the likelihood ratio test was to avoid using distantly related species which are more likely to contain compensatory changes and which may produce variable results depending on arbitrary decisions as to which distantly related species to use. However, the use of closely related species has its own set of disadvantages (see below).

A second class of disagreements are sites that are perfectly conserved within each of the different alignments but are slightly above or below cutoffs used by the LRT, SIFT or PolyPhen. These borderline cases may also depend on which sequences are included in the alignment because SIFT and PolyPhen use a site-specific score that is normalized to conservation within the rest of the protein and the P-value of the LRT explicitly depends on which species are included since this determines the expected rate of change as measured by the synonymous substitution rate. One of the goals of developing the LRT was to accurately account for each species' contribution to the likelihood using the synonymous substitution rate. The drawback of this approach is that the increase in accuracy depends on a number of parameters that must be estimated from the data.

The likelihood ratio test's use of closely related species may result in false positive and false negative predictions not made by either SIFT or PolyPhen. In order to use closely related species the likelihood ratio test uses synonymous sites to estimate the neutral substitution rate. However, the accuracy of this estimate depends on many factors. While we showed that CpG sites and large-scale regional variation in mutation rates are unlikely to have large effects on the prediction of deleterious mutations, other types of mutational variation were not accounted for, e.g. [65]. Even slight changes in the estimated neutral substitution rate will affect some predictions. Consequently, despite the advantages of using closely related species, many borderline cases may be false positives or false negatives. Borderline cases may be more accurately resolved by using additional closely related genomes or by inclusion of distantly species.

Is the power of the likelihood ratio test limited by the amount of comparative genomic data? For perfectly conserved sites, the fraction of mutations called deleterious does not increase with the number of species used to identify conserved sites (Figure S2.3). However, the fraction of mutations predicted to be deleterious increases with the number of species used for sites that are not perfectly conserved (Figure S2.3). This suggests that additional species will only increase the power of the LRT to detect sites that are not perfectly conserved. Increasing the power to detect semi-conserved sites may be useful given the large number of human disease alleles that occur at sites that are not perfectly conserved.

What fraction of deleterious mutations were not identified? Not considering false negatives due to compensatory changes and other modeling assumptions, deleterious mutations could be missed due to a lack of alignments or low power associated with the available alignments. To estimate the number of deleterious mutations that did not reach a P-value of less than 0.001, i.e. low power, we estimated the number of true positives as a function of the false discovery rate. Figure S2.4 shows that as the P-value cutoff is lowered, the estimated number of true positives plateaus to approximately 1,100 mutations, suggesting that nearly 75% of deleterious mutations were identified at a P-value cutoff of 0.001. The number of deleterious mutations missed due to a lack of sufficient alignments is more difficult to know. Only 74-76% of mutations were tested in the Venter, Watson and Chinese genomes due to a paucity of mammalian homologs. However, a smaller proportion of these may be deleterious since they are more likely to occur in new, recently duplicated genes. An alternative method of estimating the number of deleterious mutations is based on allele frequency and avoids deficiencies due to a lack of alignments [40]. Applying the allele frequency estimate to the Venter genome indicates that the likelihood ratio test identified 62% of rare deleterious mutations (Methods). However, many factors other than the presence of suitable alignments may contribute to the difference between the allele frequency and the LRT estimates.

Despite different sequencing technologies we found a very similar number of deleterious mutations within three human genomes, 796-837. Previous reports predicted ~770 [52] and 1,455 [50] missense mutations within the Watson and Venter genome, respectively. However, these differences can be almost entirely attributed to differences between the methods used to predict deleterious mutations and/or the quality score cutoffs used to eliminate sequencing errors. Using the same set of high-quality variants tested by the LRT, SIFT predicted 890, 769 and 861 and PolyPhen predicted 555, 501 and 496 deleterious mutations in the Venter, Watson and the Chinese genomes, respectively. While the coefficient of variation is greater than one for both SIFT and PolyPhen, it is less than one for the LRT. The standard deviation of the the number of deleterious mutations identified by the LRT is 20.5 and is less than that expected assuming a Poisson distribution, 28.6. This supports the idea that truncating selection mediated by synergistic epistasis facilitates the removal of deleterious mutations [66]. However, not all features of deleterious mutations were similar among the three genomes. The fraction of deleterious or neutral mutations within recently duplicated genes was much smaller in the Chinese compared to the other two genomes. This difference may be reflective of the greater difficulty of identifying SNPs in duplicated sequences using short-read sequencing technologies.

In conclusion, the abundance of mutations that disrupt highly conserved amino acid positions within three healthy human genomes implies that in most cases their phenotypic effects are either small or that the mutations are recessive. However, many of the mutations may produce large effects when homozygous. Although only a small number of mutations were predicted deleterious by the LRT, SIFT and PolyPhen, the use of all three methods should provide an excellent source of candidates for association with human disease. Finally, since just over half of all deleterious mutations were found to be common, our results support the possibility that rare variants make a significant contribution to complex human diseases [4; 67].

METHODS

Comparative genomic dataset

Multiple sequence alignments of protein coding sequences were generated from 32 vertebrate species (Figure S2.2). Orthologous protein sequences were downloaded from Ensembl (ftp://ftp.ensembl.org/pub/release-49/emf/ensembl_compara/homologies/), originally inferred by TreeBest [68], aligned using MUSCLE [69] and then back-translated into nucleotide alignments. All known selenocysteine residues were masked. After removing alignments with too few orthologous, less than 10 eutherian mammals, there were 18,993 alignments with an average of 16.5 species per alignment.

Likelihood ratio test

The likelihood ratio test was used to compare the null model that each codon is evolving neutrally, with no difference in the rate of nonsynonymous (*dN*) to synonymous (*dS*) substitution, to the alternative model that the codon has evolved under negative selection with a free parameter for the dN/dS ratio. The log likelihood ratio (*LLR*) of the conserved relative to the neutral model is:

$$LLR = \log \frac{L(D|T,\theta,dN=\hat{C}\,dS)}{L(D|T,\theta,dN=dS)}$$

where *D* is an alignment of a single codon, *T* is a phylogenetic tree, *dN* and *dS* are the nonsynonymous and synonymous substitution rates of the codon and $\hat{C}$ is the maximum likelihood estimate of dN/dS. The synonymous rate and θ, the parameters of the rate matrix, were estimated from the concatenated set of all codons without gaps, as described in the next paragraph. P-values were obtained by comparing twice the log likelihood ratio to a χ^2 distribution with one degree of freedom. The likelihood ratio test was implemented using the MG94 codon model [13] combined with an HKY85 model [70] to account for unequal base frequencies and differences in the rates of transitions and transversions. This was accomplished within a maximum likelihood framework using HYPHY [71].

The synonymous substitution rate was estimated from gap-free concatenated alignments of 1,227 genes completely conserved across all 32 species, a total of 54 kb, using the same model as that described above. The estimated *dS* values for the human to the mouse-rat ancestor (0.46), mouse lineage (0.10), and rat lineage (0.13) are similar to previous estimates, 0.457, 0.095, and 0.101, respectively [72].

To make the predictions of the likelihood ratio test available to the community for every potential nonsynonymous variant in the human genome, we applied it to 10,073,284 codons for which there were a sufficient number of aligned species. A total of 5,828,045 sites were significant ($P < 0.001$ and $dN/dS < 1$). Note that a substantial number of nonsynonymous variants at these positions may not be predicted deleterious if the mutant allele being tested is present within one or more of the eutherian mammals. The complete dataset can be obtained by request from the authors or downloaded from the corresponding author's website (http://www.genetics.wustl.edu/jflab/).

Identifying deleterious mutations

A complete catalog of SNPs were obtained for J. Craig Venter and a Han Chinese male from their respective websites (http://www.jcvi.org/cms/research/projects/huref/ and http://yh.genomics.org.cn), and for James D. Watson directly from Dr. David Wheeler. Nonsynonymous and synonymous SNPs were identified using known genes in Ensembl release 49. Coding SNPs in ambiguous reading frames, due to overlap of adjacent genes or frame shifts between known splice variants, or in known pseudogenes were excluded.

To avoid sequencing errors, stringent quality filters were applied. Since quality scores were not available for each allele in the Venter genome these were independently tabulated by mapping SNPs in the Venter genome to individual sequencing reads (ftp://ftp.ncbi.nih.gov/pub/TraceDB/Personal_Genomics/Venter/) and obtaining Phred quality values for each allele from the sum of the quality scores supporting each allele.

For each SNP, a 1000-bp flanking sequence around the SNP was extracted from the reference genome (NCBI build 36) and queried with MegaBlast against the Venter reads. Of the high scoring blast hits (E-value < 1e-100), only reads with perfect alignment across the 40-bp flanking each SNP were retained. Out of the original nonsynonymous SNPs, 22% failed to support a combined Phred score of greater than or equal to 60 and at least two supporting reads. By comparison to experimentally validated SNPs [54], this high-quality set of SNP has a lower rate of false positive (0 out of 15) but a higher rate of false negatives (4 out of 19) compared to the list of SNPs originally reported in Venter. The SNPs in the Chinese genome were also filtered using a Phred quality cutoff of 60. SNPs in the Watson genome are associated with a variant score that is similar yet not identical to a Phred quality score [52]. Based on the distribution of quality variant scores we used a quality variant cutoff of 70 for the Watson genome (Figure S2.1B).

Deleterious mutations were predicted by nonsynonymous SNPs that disrupt significantly constrained codons defined by the likelihood ratio test ($P < 0.001$) and a number of subsequent filters (Table S2.1). First, positions with low power, less than 10 eutherian mammals, were eliminated. Second, a small number of sites with dN significantly greater than dS were discarded. Finally, positions where the derived deleterious allele occurred in another eutherian species were eliminated. Deleterious mutations were assigned to either the tested or reference genome depending on whether the reference or variant allele resulted in a lower dN/dS ratio. The total number of deleterious mutations includes all heterozygous or homozygous positions that differ from the reference genome except for homozygous positions where the reference allele rather than the variant allele was inferred to be deleterious. The number of deleterious mutations in the reference genome is the non-redundant set identified by comparison with the Venter, Watson and Chinese genomes.

False positive rate

The false positive rate due to sequencing errors was estimated using Phred quality scores. Using a Phred quality cutoff of 60, 0.027 false positive SNPs per 1 Mb are expected in the Venter genome. This calculated is based on an average Phred quality score of 75.7. A total of 10.4 million codons were tested and 56.2% of these, 17.6 Mb, were estimated to be significantly constrained codons ($P < 0.001$), based on a random sample of 6,357 codons. Using the estimated rate of sequencing errors and the total number of constrained codons, we expect a total of 0.47 false positive SNPs due to sequencing errors. Using the same approach, 1.06 false positives are estimated to occur in the Chinese genome. Using an error rate of 3 nucleotide substitutions per 1 Mbp for the reference genome [58], a total of 52.8 false positives are expected.

To empirically estimate the rate of false positives, the frequency of deleterious mutations was estimated for heterozygous sites with lower and higher quality values. Only heterozygous sites were used since homozygous variants are less likely to be deleterious but are more likely to have high quality values. For the Venter, Watson and Chinese genome the quality value splits were 157, 129 and 97, respectively (Figure S2.1). In each case, the estimated number of false positives was zero since the fraction of deleterious mutations was higher using alleles with higher quality values.

The false positive rate was also estimated using a set up negative controls. A negative control set was defined by ancestral amino acid substitutions which were inferred by maximum parsimony to have occurred along the lineage leading to humans after the split with the orangutan lineage but before the split with the chimpanzee lineage. We identified 39,028 amino acids that are the same between human and chimpanzee and between orangutan and macaque but differ between human and orangutan. The likelihood ratio test was applied to the negative control set as if these ancestral substitutions were missense mutations present in the human population. The sequences of the primate species used to identify the set of negative controls (human, chimp, orangutan and macaque) were excluded from the likelihood ratio test. To examine different types of amino acid changes the negative controls were subdivided into three classes by the severity of amino acid changes using the Grantham scale [38]. Conservative amino acid changes were defined by a Grantham score of 50 or below, moderate changes by a Grantham score of 51 to 100 and radical changes by a score of greater than 100.

False negative rate

The false negative rate was estimated using disease-causing missense mutations in OMIM database. The coordinates of OMIM allelic variants were converted from protein-based residue positions to genomic coordinates. Each OMIM gene was mapped to a single representative RefSeq protein which provided the best match to the positions and amino acids of curated OMIM alleles. To eliminate potential mapping errors, we filtered out amino acid variants which are more than one mutation away from the mapped codon in the reference genome. Out of 9,231 missense variants in OMIM, we mapped 5,493 variants (59.5%) to the genome and tested each of the using the likelihood ratio test. Identification of conservative, moderate and radical amino acid changes was quantified using the same set of Grantham scores as that used for the analysis of false positives.

Allele frequency

Filtered non-redundant allele frequencies in three HapMap analysis panels (CEU, CHB, and YRI) were downloaded from http://www.hapmap.org (Phase II+III, release 26) and used to identify SNPs with at least 5% frequency in the CEU, CHB, or YRI panels.

Recent gene duplication

Recently duplicated genes were defined by human paralogs with greater than 95% protein identity. Percent identity was calculated from human paralogs within the MUSCLE-generated multiple alignments of homologs from Ensembl. A total of 1,232 human genes were identified as having at least one recently duplicated paralog in the human genome.

SIFT and PolyPhen predictions

SIFT 3.0 was downloaded and run locally to predict deleterious mutations in Venter using two different modes. First, SIFT was run using an independent set of alignments built using the TrEMBL 39.8 protein sequence database. Note that while TrEMBL contains many human sequence variants SIFT eliminates highly similar sequences [16]. SIFT was able to generate alignments and predictions without errors for 6,539 out of 7,534 nonsynonymous variants. Filtering by the median conservation cutoff of 3.5 as in [50], we obtained 5,401 predictions and a total of 890 variants predicted to affect protein function. Second, SIFT was run with the same set of alignments used by the likelihood ratio test. The median conservation score was set to 4.0 so that the number of SNPs predicted to impact protein function was similar to the number predicted by the likelihood ratio test.

PolyPhen 1.15 was obtained from Shamil Sunyaev and run locally to generate predictions for the Venter genome. SwissProt 56.8 and the latest version of the BLAST NR, PDB, and DSSP databases were used as input. PolyPhen generated predictions for 6,746 nonsynonymous variants and called 768 possibly damaging and 555 probably damaging. Because PolyPhen does not generate allele-specific predictions, some of these predictions may be for Venter-reference differences where the derived, deleterious allele is present in the reference but not Venter. Both SIFT and the LRT avoided this issue because they both generate allele-specific predictions.

Estimation of rare deleterious mutations

To estimate the number of deleterious mutations using allele frequencies we compared the ratio of nonsynonymous (N) to synonymous (S) variants for rare versus common frequency classes [40]. Because allele frequencies were not always available and nonsynonymous alleles may be more often selected for genotyping, we used double hits in the dbSNP database as a proxy for rare versus common alleles, where double hits are defined by SNPs with at least two submissions. Within the three human genomes, the N/S ratio of double hit variants was 0.84-0.85 and non-double hit variants was 1.01-1.08, leading to an average estimate of 689 (18.6%) rare deleterious mutations. After accounting for the fact that only 48% of all deleterious mutations predicted by the likelihood ratio test are present as double hits within dbSNP and so do not contribute to the frequency calculation, we estimate that 62% of all rare deleterious mutations were identified by the likelihood ratio test (0.52*816/689).

ACKNOWLEDGEMENTS

We thank members of the Fay lab and S. Sunyaev for helpful suggestions and S. Sunyaev for sending us the PolyPhen algorithm. This work was supported by a National Institute of Health grant (GM-080669) and an Alfred P. Sloan Research Fellowship to J.C.F.

FIGURES

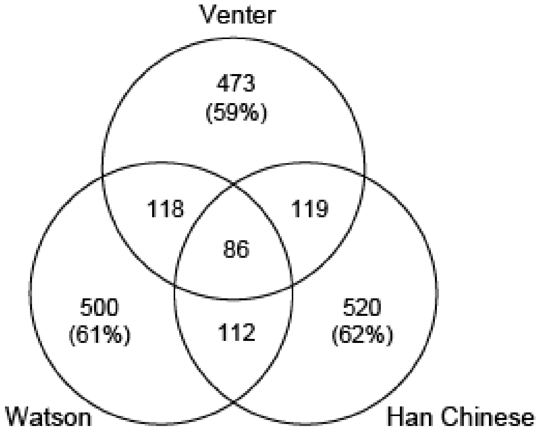


Figure 2.1. Venn diagram of deleterious mutations identified in J. Craig Venter, James D. Watson and a Han Chinese individual. The percentage of individual-specific deleterious mutations found in each genome is shown in parentheses.

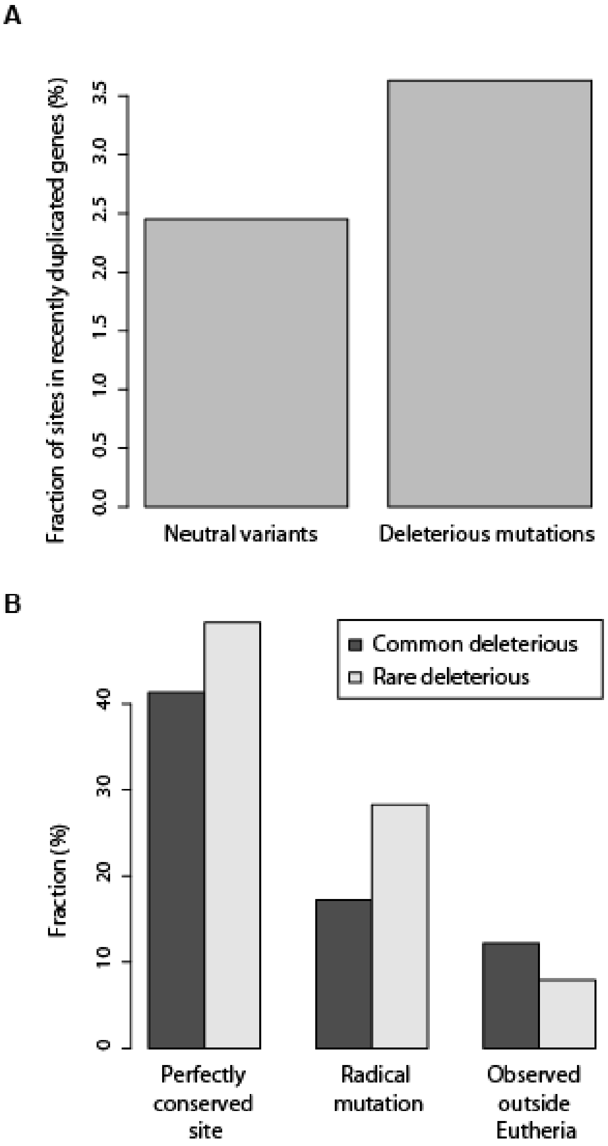
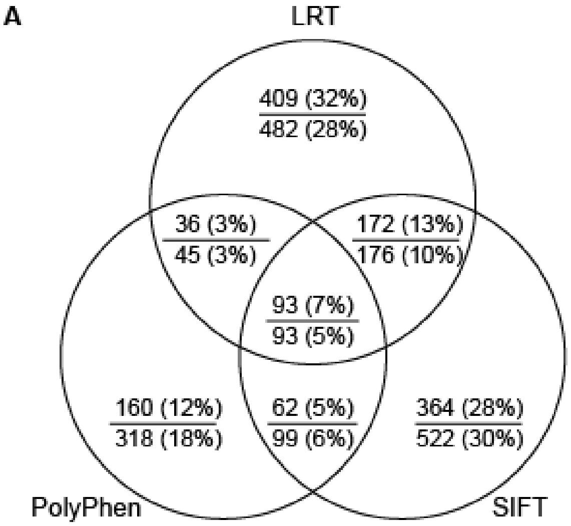


Figure 2.2. Characteristics of deleterious mutations. (**A**) Deleterious mutations (n = 1,928) are more likely to occur in recently duplicated genes relative to neutral variants (n = 8,287). (**B**) Mutations at perfectly conserved sites, mutations that cause radical amino acid changes, defined by BLOSUM62 <= -2, and mutations to amino acids that not are observed outside of eutherian mammals are more frequent among rare (n = 807) compared to common deleterious mutations (n = 1,121).



B

		LRT	
		Deleterious	Neutral
SIFT	Neutral	392 (7%)	4445 (79%)
	Deleterious	404 (7%)	404 (7%)

Figure 2.3. Comparison of SIFT, PolyPhen and the likelihood ratio test (LRT) predictions. (**A**) Venn diagram of the number of predictions made by the three methods. Probably damaging mutations were used for PolyPhen. Numbers below and above each line are for the complete set of 7,534 high-quality variants present within the Venter genome and a subset of 4,303 where all three methods generated a prediction, respectively. (**B**) Overlap between the LRT and SIFT predictions based on the same alignments.

TABLES

Table 2.1. Summary of deleterious mutations found in three individuals and the reference genome.

		Tested		Deleterious	
Genome	High-quality variants	Number	Heterozygotes (%)[b]	Number[a]	Heterozygotes (%)[b]
J. Craig Venter	7,534	5,645	52	796 (14%)	78%
James D. Watson	7,353	5,417	49	816 (15%)	76%
Han Chinese	7,462	5,707	58	837 (15%)	83%
Reference	n.a.	10,689	n.a.	838 (8%)	n.a.

[a] The percentage of tested mutations that are deleterious is shown in parenthesis.

[b] The frequency of heterozygotes were derived from genotype calls in the original publications.

Table 2.2. Deleterious mutations are enriched in rare frequency classes.

	Rare alleles		Common alleles[a]	
Gnome	Tested	Deleterious[b]	Tested	Deleterious[b]
J. Craig Venter	766	213 (28%)	3066	583 (19%)
James D. Watson	940	303 (32%)	2632	513 (19%)
Han Chinese	703	186 (26%)	3438	651 (19%)

[a] Common alleles include those that are shared between any of the three genomes.

[b] Percent deleterious is shown in parentheses.

SUPPORTING INFORMATION

Table S2.1. A sequence of filters used to identify deleterious mutations in the Venter genome.

Filter	Sites	Genes	Coding sequence (Mbp)	Heterozygotes (%)
Nonsynonymous SNP	9708	5758	12.4	56
High quality SNP	7534	4879	10.7	53
SNPs with alignments	5645	3756	8.9	52
Derived allele in Venter	3832	2755	6.7	75
dN not equal to dS ($P < 0.001$)	1012	886	2.3	76
dN < dS	942	827	2.2	77
Not observed in other eutherian mammals	796	700	1.9	78

Filters were applied cumulatively from top to bottom.

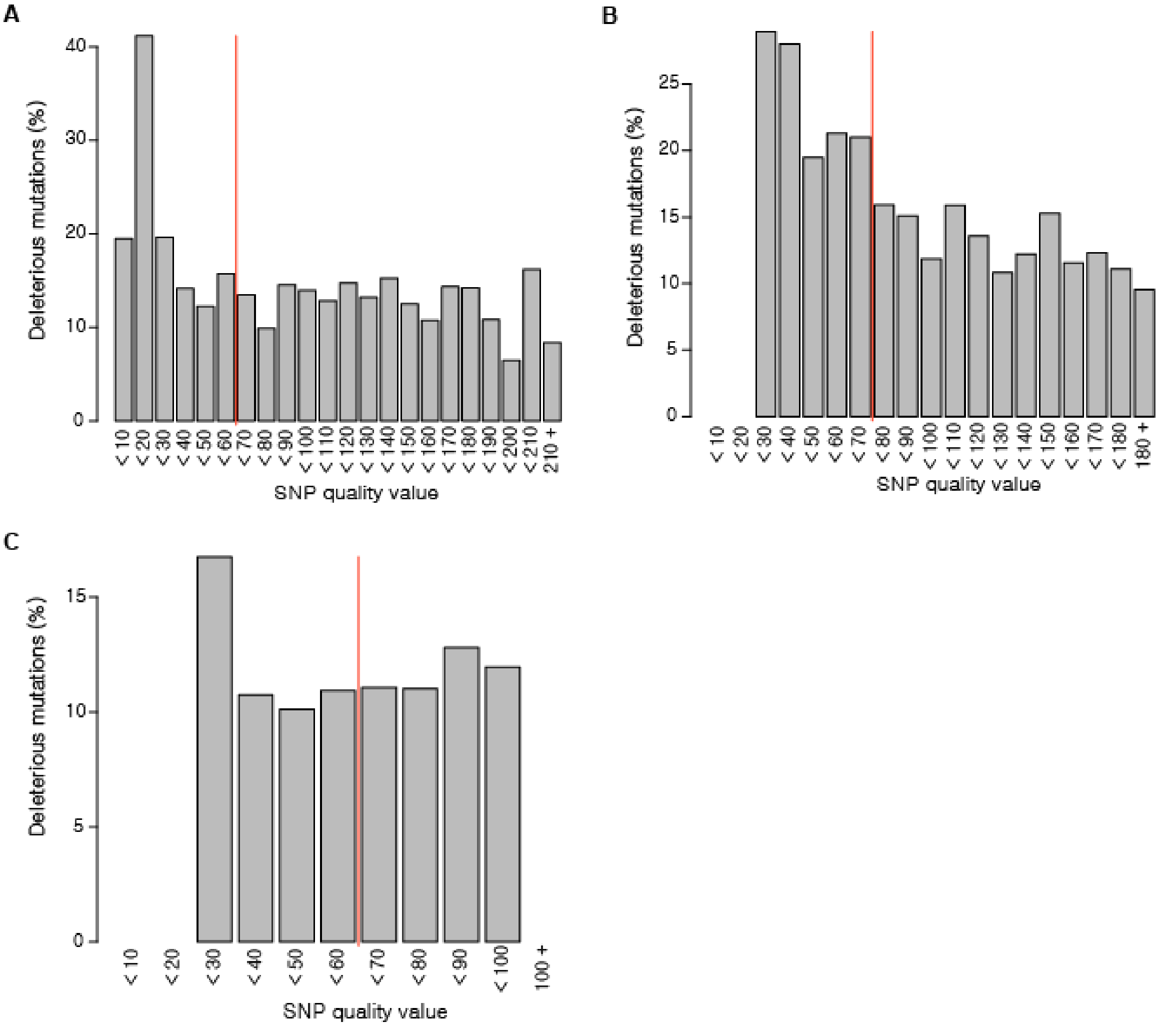


Figure S2.1. The fraction of mutations that are deleterious for different quality intervals within the Venter (A), Watson (B) and Han Chinese (C) genome. The quality value cutoff for high-quality SNPs is marked by a red vertical line.

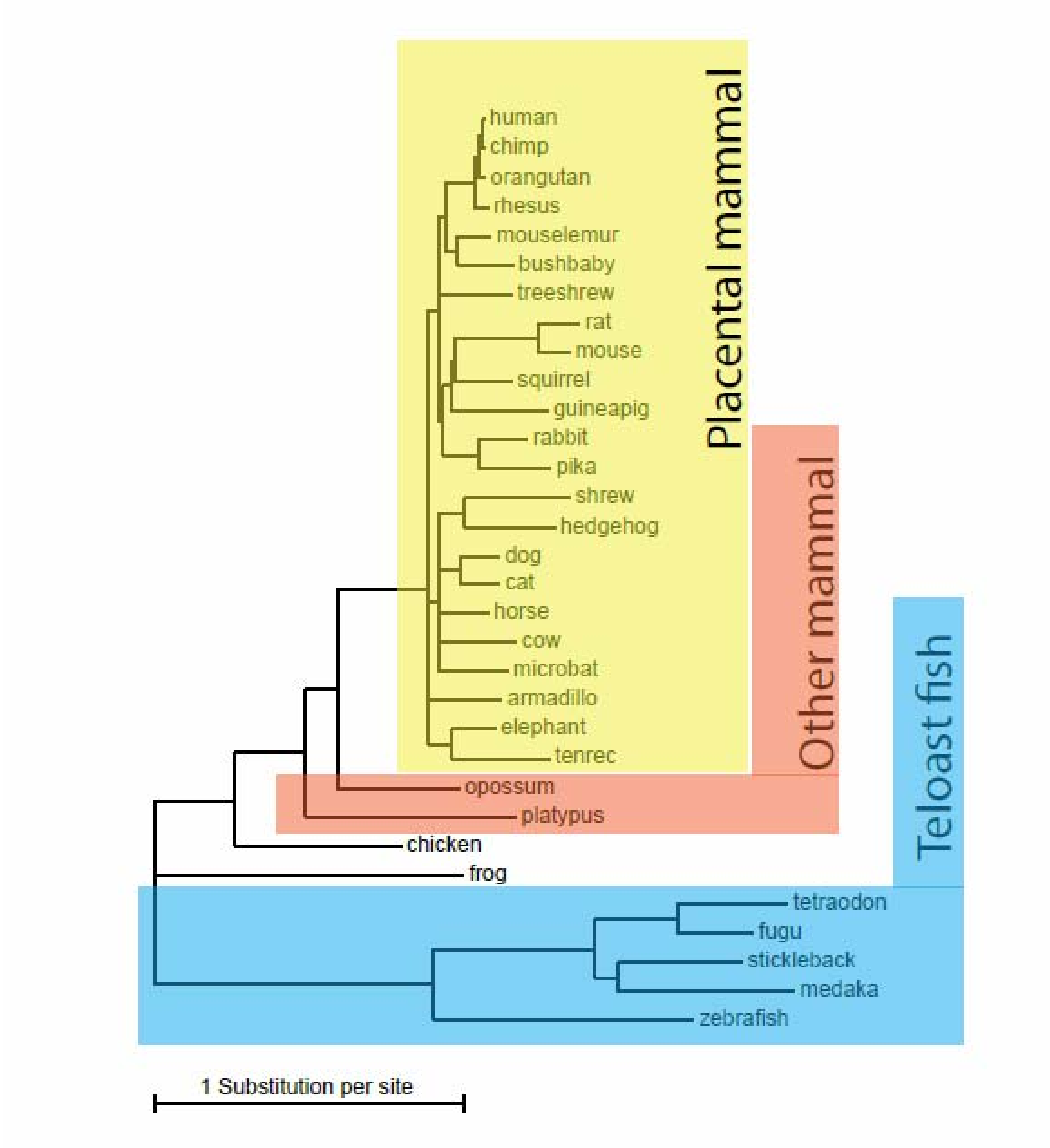


Figure S2.2. Phylogenetic tree of species used and their synonymous rate (dS) of evolution.

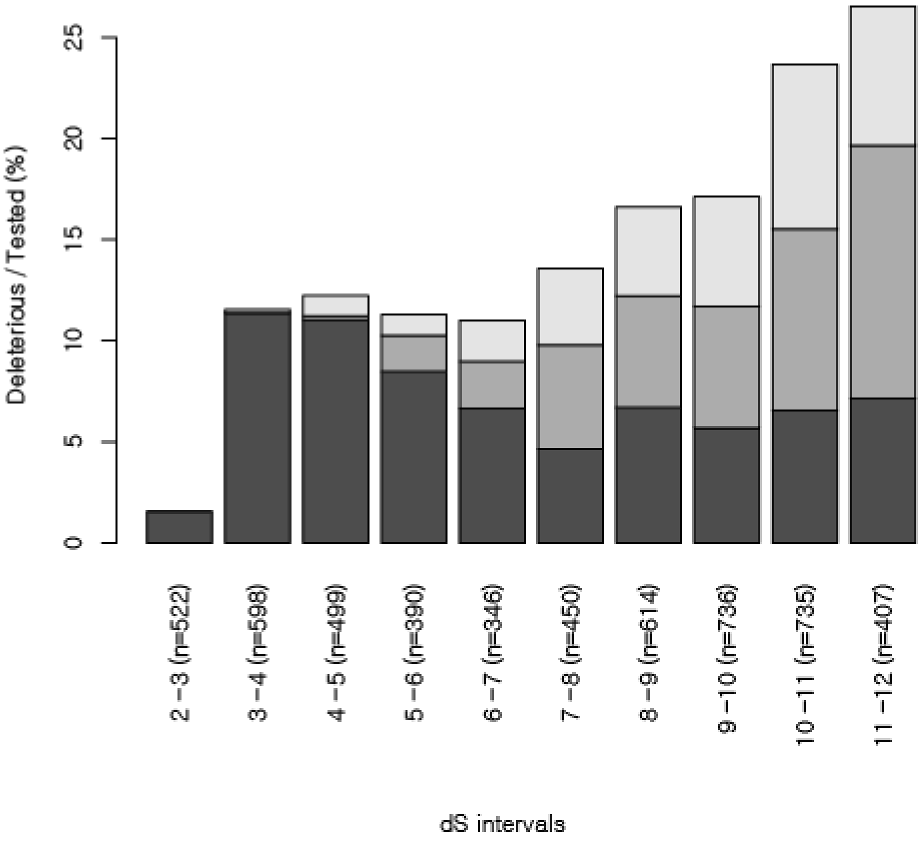


Figure S2.3. The percentage of mutations predicted to be deleterious as a function of the total synonymous substitution rate (dS). Each bar represents a different dS interval and sample sizes are denoted by n. Dark gray shows deleterious mutations at perfectly conserved sites, medium gray shows sites where all eutherian mammals are perfectly conserved but at least one vertebrate outside of Eutheria is different, and light gray shows deleterious mutations at all other types of sites.

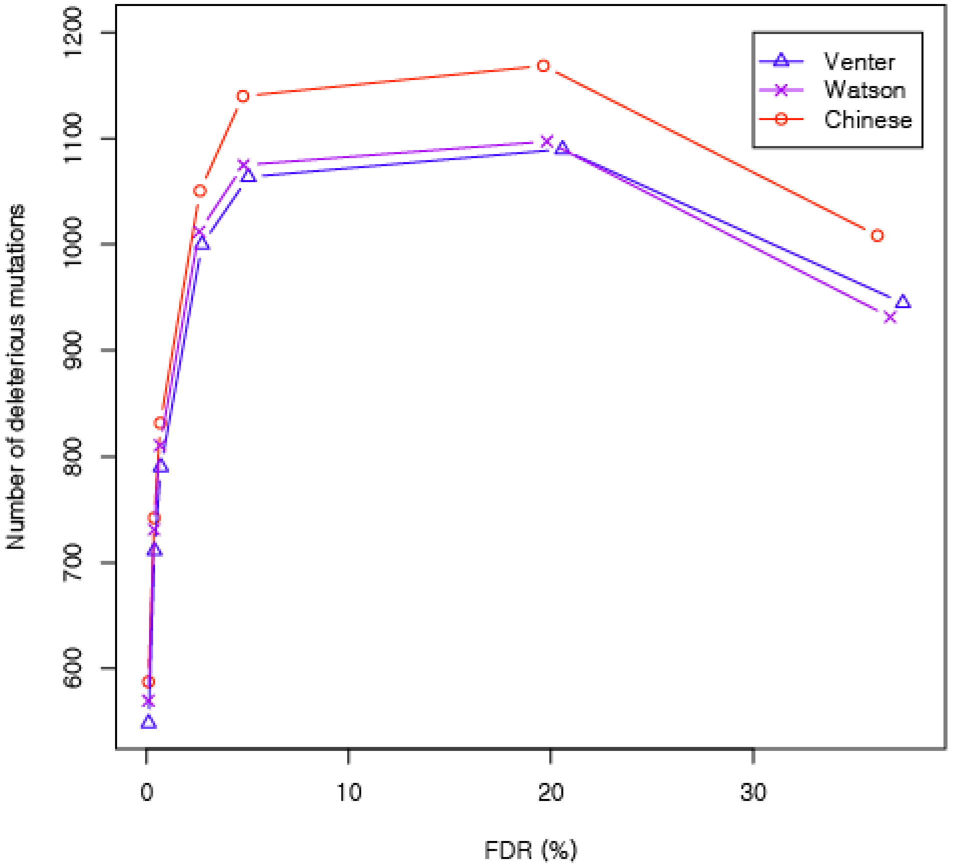


Figure S2.4. The estimated number of deleterious mutations as a function of the false discovery rate (FDR). The number of deleterious mutations was estimated by the number of mutations predicted to be deleterious at a P-value cutoff of 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1 minus the number of false positive predictions expected due to multiple testing. The false discovery rate was calculated by the estimated number of deleterious mutations (true positives) divided by the total number of mutations predicted at each P-value cutoff. For example, at a P-value cutoff of 0.1, 1,509/5,645 mutations were predicted to be deleterious in the Venter genome, 0.1*5,645 = 564 of these are expected to be false positives. The leads to a false discovery rate of 564/1,509 = 37%.

CHAPTER 3: Evidence for hitchhiking of deleterious mutations within the human genome

Sung Chun[1] and Justin C. Fay[1,2]

[1] Computational and Systems Biology Program, Washington University, St. Louis, MO

[2] Department of Genetics and Center for Genome Sciences and Systems Biology, Washington University, St. Louis, MO

This work was done in collaboration with Justin Fay. My contribution was design of the experiments, execution of the experiments and analysis of the data. This chapter is a reprint of the manuscript originally published in PLoS Genetics 2011. Large supplemental tables are available from the Journal website.

ABSTRACT

Deleterious mutations present a significant obstacle to adaptive evolution. Deleterious mutations can inhibit the spread of linked adaptive mutations through a population and, conversely, adaptive substitutions can increase the frequency of linked deleterious mutations and even result in their fixation. To assess the impact of adaptive mutations on linked deleterious mutations we examined the distribution of deleterious and neutral amino acid polymorphism in the human genome. Within genomic regions that show evidence of recent hitchhiking, we find fewer neutral but a similar number of deleterious SNPs compared to other genomic regions. The higher ratio of deleterious to neutral SNPs is consistent with simulated hitchhiking events and implies that positive selection eliminates some deleterious alleles and increases the frequency of others. The distribution of disease-associated alleles is also altered in hitchhiking regions. Disease alleles within hitchhiking regions have been associated with auto-immune disorders, metabolic diseases, cancers, and mental disorders. Our results suggest that positive selection has had a significant impact on deleterious polymorphism and may be partly responsible for the high frequency of certain human disease alleles.

INTRODUCTION

The continuous removal of deleterious mutations is essential to maintaining a species' reproductive output and even its existence. While deleterious mutations incur a considerable fitness cost [1], they are not always effectively removed from a population. Deleterious mutations are more difficult to remove from small populations and their accumulation can lead to further reductions in population size and eventually to extinction, a process called mutational meltdown [73-75]. Sexual recombination facilitates the elimination of deleterious mutations [76] and the lack of recombination on the *Y* sex chromosome may have contributed to its degeneration through the accumulation of deleterious mutations [77].

In humans, many deleterious mutations have reached high population frequencies. Each human is estimated to carry on the order of 1,000 deleterious mutations in their genome [9; 15; 40]. Although most deleterious mutations are rare, a significant fraction is common in the population. For example, 19% of deleterious mutations identified in three human genomes are common enough to be shared among them [15]. However, the cause and consequence of common deleterious mutations have been difficult to determine.

A number of factors may contribute to the large number of common deleterious mutations in humans. Most genome-wide methods used to identify deleterious mutations are based on the alteration of sites that are significantly conserved across species [43; 78]. As such, lineage-specific changes in selective constraint provide one explanation for common alleles that alter highly conserved sites.

Changes in selective constraint can be caused by changes in population size, the environment, or other genetic changes [79]. Because the efficacy of selection is a function of effective population size, a reduction in population size can result in reduced constraints on sites that are conserved in other species [80]. Many common deleterious mutations in humans can be attributed to the small effective population size of humans and recent human population bottlenecks [17; 81]. However, changes in constraint can also be mediated by genetic or environmental changes. For example, the thrifty gene hypothesis posits that the high frequency of diabetes risk alleles is a consequence of their being previously advantageous during periods of food scarcity [82]. Relaxed constraints may also arise due to certain types of genetic changes, such as gene duplication or compensatory mutations. The observation that human disease alleles are often present in mouse supports the notion that the selective constraints on a site are not always static but can change with the genetic or environmental background [44]. However, not all common deleterious mutations may result from species-specific differences in selective constraint.

Positive selection can influence the frequency of deleterious mutations directly, through genetic hitchhiking, or indirectly, through a reduction in effective populations size mediated by an increase in the variance of reproductive success [83]. As a consequence, positive selection can increase the rate at which deleterious mutations accumulate, particularly when the effect of the advantageous mutation outweighs the effects of linked deleterious mutations [84-87]. Hitchhiking of deleterious mutations along with advantageous mutations may have contributed to the degeneration of the *Y* sex chromosome [86; 88] and the increased number of deleterious mutations present in domesticated species [89; 90].

In this study, we examined the effect of positive selection on linked deleterious polymorphism in the human genome. We compared the abundance of deleterious and neutral nonsynonymous single nucleotide polymorphisms (SNPs) in regions showing evidence of hitchhiking to other genomic regions. While hitchhiking is expected to remove neutral variation from a population [91], we find that the rate of deleterious SNPs is not reduced, resulting in an enrichment of deleterious relative to neutral SNPs in hitchhiking regions. Our results imply that positively selected mutations may often influence the frequency of linked deleterious mutations.

RESULTS

Simulated effect of hitchhiking on deleterious mutations

To characterize the effect on positive selection on linked deleterious mutations we conducted simulations under a Wright-Fisher model. Subsequent to a single hitchhiking event, the rate of neutral and deleterious polymorphism was reduced as a function of the rate of recombination (Figure 3.1A). Despite the overall reduction in the number of deleterious polymorphisms, at intermediate rates of recombination, hitchhiking caused an increase in the number of high frequency deleterious polymorphisms, as measured by θ_H (Figure 3.1A), similar to its effect on neutral polymorphism [92]. Compared to deleterious polymorphism, hitchhiking caused a greater reduction in neutral polymorphism, resulting in an enrichment of deleterious relative to neutral polymorphism. The enrichment was greatest for high compared to intermediate and low frequency polymorphism, as measured by θ_H, θ_π, and θ_W, respectively (Figure 3.1B). Because the reduction in fitness due to deleterious polymorphism remained constant during hitchhiking, the cost of increasing the frequency of some deleterious alleles to high frequency must be offset by the elimination of other deleterious alleles.

To examine the average effect of multiple hitchhiking events we also simulated populations under a continuous influx of advantageous mutations. Similar to single hitchhiking events, recurrent hitchhiking reduced the rate of neutral and deleterious polymorphism (Figure 3.1C), and increased the ratio of deleterious to neutral polymorphism (Figure 3.1B). While the degree to which hitchhiking caused an enrichment of deleterious polymorphism depended on the strength of positive and negative selection and the rate of advantageous and deleterious mutation (Figure S3.1), our simulations indicate that hitchhiking may often have a measurable impact on the ratio of deleterious to neutral polymorphism segregating in natural populations.

Classification of deleterious and neutral nonsynonymous SNPs in humans

To examine the impact of positive selection on deleterious polymorphism in humans we classified nonsynonymous SNPs from the 1000 Genomes Project [93] as neutral or deleterious using a likelihood ratio test based on cross-species conservation (Materials and Methods). Although not all classifications may be correct, the likelihood ratio test classifies 72% of human disease mutations as deleterious and only 6.7% of nonsynonymous substitutions between species as deleterious [15]. Out of 48,558 autosomal nonsynonymous SNPs tested, 14,094 (29.0%) were predicted to be deleterious, of which 2,263 (16.1%) have a derived allele frequency of over 10%. Using a cutoff of 10%, the fraction of SNPs called deleterious is 17.8% for common alleles compared to 33.0% for rare alleles, consistent with the expected effects of negative selection.

Enrichment of deleterious SNPs in regions showing evidence of hitchhiking

Hitchhiking is expected to have a stronger effect on linked variation in regions of low recombination [91]. While the spread of a positively selected allele through a population causes a reduction in the amount of linked neutral variation, it may interfere with the elimination of linked deleterious mutations. Consistent with this hypothesis, the rate of synonymous and neutral nonsynonymous SNPs decreases in regions of low recombination, whereas the rate of deleterious SNPs remains nearly constant (Figure 3.2A). As a consequence, the ratio of deleterious to neutral and deleterious to synonymous SNPs is significantly correlated with the rate of recombination ($P = 3.1 \times 10^{-15}$ and $P < 2.0 \times 10^{-16}$, respectively, Figure 3.2B). The association remains significant when accounting for the frequency of conserved codons and biased gene conversion ($P = 2.1 \times 10^{-7}$ and $P = 9.3 \times 10^{-6}$, respectively, Figure S3.2), which are also correlated with the rate of recombination. However, this correlation is also expected due to background selection, which reduces the efficacy of selection against deleterious mutations [94; 95].

In contrast to background selection, which exerts more uniform effects across the genome [96], hitchhiking can generate strong local effects. Furthermore, hitchhiking can have large effects in regions of both low and high recombination whereas background selection is expected to have much smaller effects in regions of high recombination [97].

To determine whether deleterious SNPs have been influenced by recent episodes of positive selection, we examined genomic regions showing evidence of hitchhiking based on multiple tests of selection [98]. In hitchhiking regions defined by two or more tests of selection, we found a significantly higher ratio of deleterious to neutral SNPs compared to other genomic regions (Figure 3.3 and Table S3.1). The elevated ratio of deleterious to neutral SNPs within hitchhiking cannot be explained by a reduced rate of recombination or a higher density of conserved sites; the difference between hitchhiking and non-hitchhiking regions remained significant using a logistic regression model with these factors as covariates ($P = 6.3 \times 10^{-5}$, Figure S3.3). The increase in the ratio of deleterious to neutral SNPs in hitchhiking relative to non-hitchhiking regions is 1.09-fold for regions identified by two or more tests of selection and increases to 1.87-fold for regions identified by all nine tests of selection. The increase in the ratio of deleterious to neutral SNPs in hitchhiking regions is due to a decrease in the number of neutral SNPs rather than an increase in the number of deleterious SNPs (Figure 3.3B and 3.3C). With the exception of the composite likelihood ratio test (CLR) [99], all of the methods used to detect hitchhiking identify regions with a higher ratio of deleterious to neutral SNPs (Figure 3.3D). Thus, the increase in the relative abundance of deleterious SNPs in hitchhiking regions does not appear to be associated with any specific test of selection.

The effects of hitchhiking are expected to decline as a function of recombinational distance from the site under selection [91]. To examine the decay in the number of deleterious SNPs associated with hitchhiking, we used iHS [100] and Rsb [101] defined hitchhiking regions. iHS is better at detecting incomplete hitchhiking events [100], where the advantageous mutations is still segregating in the population, whereas Rsb is better at detecting complete or nearly-complete episodes of selection [101]. The frequency of deleterious SNPs decreases as a function of distance from iHS defined hitchhiking region ($P = 2 \times 10^{-7}$, Figure 3.4A). Compared to iHS regions, the frequency of deleterious SNPs shows a more modest decline with distance from the Rsb defined hitchhiking regions ($P = 0.018$, Figure 3.4B). This difference could result from Rsb detecting older hitchhiking events providing additional time for negative selection to eliminate linked deleterious mutations or due to a weaker influence of hitchhiking outside of Rsb defined regions, which are twice as large as iHS defined regions (Table S3.1).

Hitchhiking regions show a similar enrichment of rare, intermediate and common deleterious SNPs

As the rate of recombination decreases, hitchhiking causes a larger increase in the ratio of deleterious to neutral SNPs for common compared to low frequency SNPs (Figure 3.1). To determine whether hitchhiking regions show a similar pattern, we compared the ratio of deleterious to neutral SNPs as a function of allele frequency. Similar to the simulation results, the ratio of deleterious to neutral SNPs declines with increasing allele frequency. However, the ratio of deleterious to neutral SNPs in hitchhiking regions is not significantly different among three frequency classes (Table 3.1). We observed the same pattern using HapMap SNPs (data not shown) indicating that low coverage sequencing errors in the 1000 Genomes Project is unlikely to explain this result. Although the absence of differences in the ratio of deleterious to neutral SNPs across allele frequencies is somewhat surprising, it is consistent with simulations with a high rate of recombination or strong negative selection (Figure 3.1 and Figure S3.1).

Deleterious SNPs in regions showing population-specific patterns of hitchhiking

Many of the methods used to detect hitchhiking were independently applied to populations of different ancestry. Although some hitchhiking events may be specific to European, African, or Asian populations, e.g. [100], the power to detect hitchhiking is expected to differ among populations even when an adaptive mutation is fixed in all populations [102; 103]. We examined the enrichment of deleterious SNPs in iHS defined hitchhiking regions in the European, African, and Asian samples. Surprisingly, we found no enrichment of deleterious SNPs in African and Asian defined hitchhiking regions (Table S3.2). Despite these population-specific differences revealed by iHS, the ratio of deleterious to neutral SNPs is elevated in hitchhiking regions defined by multiple methods in the African, European and Asian samples (Table S3.3).

Deleterious SNPs within and around genes under positive selection

For most hitchhiking regions the target of selection is not known. We identified ten hitchhiking regions from the literature for which there is evidence for the target of selection. The putative targets are *LCT* [104; 105], *SLC45A2* [106], *TYRP1* [107], *HERC2* [107], *KITLG* [107], *SLC24A5* [108], *TYR* [109], *EDAR* [106], *PCDH15* [107] and *LEPR* [107]. Within these genes the ratio of deleterious to neutral SNPs (1.83) is higher than in non-hitchhiking regions (0.41) (Fisher's Exact Test P = 0.0023, Table 3.2). The deleterious SNPs include 5/6 nonsynonymous SNPs that are putative targets of selection. Within the 1 Mbp regions flanking these genes, there is also a higher ratio of deleterious to neutral SNPs (0.69) relative to that in non-hitchhiking regions (0.41) (Fisher's Exact Test P = 0.034).

Positive selection at *SLC45A2* and *TYR* is particularly interesting since linked deleterious SNPs have been associated with human disease. The putative target of selection on *TYR* is a nonsynonymous SNP (S192Y) that has an allele frequency of 42% in the European sample (CEU) and is associated with the absence of freckles in Europeans [109]. Another nonsynonymous SNP in *TYR* (R402Q), 106 kb away, is classified as deleterious, has a frequency of 21% in CEU and is associated with mild ocular albinism and risk for cutaneous melanoma and basal cell carcinoma [110; 111]. The putative target of selection on *SLC24A5* is a nonsynonymous SNP (A111T) that is associated with skin pigmentation and is nearly fixed in European populations but is at low frequency in African and Asian populations [108]. Positive selection on this allele may have influenced the frequency of deleterious SNPs in *FBN1*, 265 kb downstream of

SLC24A5. *FBN1* has five deleterious SNPs in HapMap CEU, all of which are present at low frequency in CEU, 0.5-1.4%, but are absent from both the African or Asian HapMap samples. Three of these deleterious SNPs cause Marfan syndrome [112; 113] and one has been found in patients with Marfan syndrome or related phenotypes [114].

Disease-associated alleles within hitchhiking regions

Hitchhiking may have also influenced SNPs that are associated with human disease. This might occur by increasing the frequency of rare, disease-causing mutations or by increasing the frequency of more common, disease-risk alleles. To investigate this possibility we compared the abundance of disease-associated alleles in hitchhiking and non-hitchhiking regions.

Within known disease genes in OMIM, there are 9,481 mutations that have been associated with human disease, of which 1,722 were common enough to be typed in the HapMap project and can be considered SNPs. The ratio of all OMIM variants in hitchhiking relative to non-hitchhiking regions (0.053) is lower than that of the number of OMIM morbid genes (0.071), consistent with the elimination of variation within hitchhiking regions (Table S3.4). However, the ratio of common OMIM variants in hitchhiking to non-hitchhiking regions, 0.079, is significantly higher than that of rare variants, 0.047 (Fisher's Exact Test, $P < 10^{-5}$, Figure 3.5). This difference is opposite to that found for neutral HapMap SNPs, which are skewed towards rare alleles in hitchhiking relative to non-hitchhiking regions. Furthermore, the minor allele frequencies of OMIM SNPs is slightly higher in hitchhiking compared to non-hitchhiking regions (Wilcoxon Rank Sum Test, P = 0.03). Similar to OMIM SNPs, the ratio of disease-associated SNPs in hitchhiking relative to non-hitchhiking is higher for common compared to rare alleles identified in the 1000 Genomes Project, although the difference is not significant (Figure 3.5, Fisher's Exact Test, P = 0.20). For the 1000 Genomes Project data, the mean frequency of common disease alleles in hitchhiking regions (0.25) is higher than that in non-hitchhiking regions (0.20), although the difference is not significant (Wilcoxon Rank Sum Test, P = 0.80). Thus, hitchhiking regions appear to be characterized by an increase in the number common disease-associated SNPs rather than by an increase in the number of rare, disease-associated variants.

To examine the abundance of common, risk-associated alleles within hitchhiking regions, we used alleles that have been associated with human disease from genome-wide association studies (GWAS) [115] and from a literature survey (see Materials and Methods). Consistent with a previous study [115], the ratio of risk-alleles identified by GWAS in hitchhiking to non-hitchhiking regions, 0.059, is not greater than that expected based on the number of genes, 0.068 (Table S3.4). However, nonsynonymous risk alleles, which are likely enriched for functional variants, have a higher hitchhiking to non-hitchhiking ratio than that of other risk-alleles (Figure 3.5, Fisher's Exact Test, P = 0.02). Although risk alleles in hitchhiking regions do not have significantly higher allele frequencies than those in nonhitchhiking regions (Wilcoxon Rank Sum Test, P = 0.63), the proportion of risk alleles with odds ratios over 2.0 in hitchhiking regions (18.9%) is significantly higher than that in non-hitchhiking regions (11.5%) (Fisher's Exact Test, P =

0.03). For disease-associated nonsynonymous SNPs identified in a literature survey, the ratio of SNPs in hitchhiking to non-hitchhiking regions is lower than that of neutral SNPs (Table S3.4).

Disease-phenotype classification

To identify which types of diseases hitchhiking may have influenced, we examined disease-associated SNPs and genes with deleterious SNPs within hitchhiking regions. Classification of the 126 OMIM SNPs within hitchhiking regions by phenotype (Table S3.5) revealed a number of SNPs involved auto-immune disorders (21 SNPs), energy metabolism (16 SNPs), and a variety of mental, neurological, and neurodevelopmental disorders (25 SNPs). Classification of the 461 genes (Table S3.6) within hitchhiking regions that contain deleterious SNPs by their disease association revealed a number that have been associated with cardiovascular (N = 21), immune (N = 19), metabolic (N = 18), neurological (N = 12) and psychiatric disease (N = 10), and cancer (N = 17), according to the Genetic Association Database classification [116]. Classification of the 12 nonsynonymous SNPs identified by GWAS and the three nonsynonymous SNPs identified from the literature revealed five associated with auto-immune disease, three associated with metabolic disease, and two associated with cancer. However, none of these disease classifications are significantly different from those outside of hitchhiking regions.

Genome clustering of deleterious SNPs

Most deleterious SNPs lie outside of currently defined hitchhiking regions. However, this does not exclude the possibility that they were influenced by positive selection. The overlap among methods used to detect hitchhiking is low [98], and some hitchhiking events may not be detected by any of the methods. For example, a beneficial mutation may initially spread slowly through a population while it becomes disentangled from linked deleterious mutations. In this scenario, patterns of hitchhiking may be weak or absent, similar to those that occur when positive selection acts on standing genetic variation [117]. To characterize genomic regions enriched for deleterious SNPs, we split the genome into 1 Mbp windows and selected the top 2% of windows with the highest rate of deleterious SNPs per kb of coding sequence.

Regions enriched for deleterious SNPs have a high ratio of deleterious to neutral nonsynonymous SNPs, 0.66, much higher than the genome average, 0.41. Together, these 43 regions contain 7.4% of deleterious SNPs (Tables S3.7). 17 of these regions show evidence of hitchhiking, ten with evidence from three or more tests of selection. In addition, one region may have been influenced by positive selection on *DARC* [118], even though it does not overlap with hitchhiking regions defined by multiple tests of selection [98]. Ten of the regions contain deleterious SNPs in multiple duplicated olfactory receptor or keratin genes. Of the remaining 21 regions, 16 have deleterious SNPs in more than two genes. While loss of constraint may explain the accumulation of deleterious SNPs in some genes, particularly those that are duplicated, it is less likely to explain deleterious SNPs in multiple linked genes with disparate functions.

DISCUSSION

Deleterious mutations have a significant impact on a species' ability to survive, reproduce and adapt to new environments [73-75]. In humans, there is an abundance of common nonsynonymous SNPs that disrupt sites highly conserved across species and likely to be deleterious [15]. By examining the genome distribution of nonsynonymous SNPs classified as either neutral or deleterious, we found a greater reduction in neutral compared to deleterious polymorphism within genomic regions likely to have been influenced by hitchhiking. This observation combined with hitchhiking simulations suggests that while many deleterious SNPs are eliminated due to hitchhiking, a substantial number of rare deleterious mutations must also increase to frequencies common enough to be considered polymorphic. Our results imply that positive selection is not responsible for the abundance of common deleterious SNPs across the human genome but is relevant to understanding the distribution and dynamics of deleterious mutations as well as certain disease alleles.

Despite evidence for a hitchhiking effect, most common deleterious SNPs are unlikely to have been influenced by positive selection and are better explained by a change in selective constraint, mediated by a population bottleneck [81] or environment change [119]. Only 11.5% of deleterious SNPs occur in regions showing evidence of hitchhiking (Table S3.1). However, this does not exclude the possibility that positive selection has influenced the frequency of some deleterious SNPs outside of hitchhiking regions. Hitchhiking regions were defined by the overlap of two or more methods of detecting selection and are unlikely to include all regions influenced by hitchhiking [98]. In support of this possibility, we identified a number of genomic regions that contain an exceptionally high ratio of deleterious to neutral SNPs. Although some of these regions include multiple duplicated genes, which could explain the large number of SNPs predicted to be deleterious, one of the regions includes a gene thought to have been under selection, *DARC* [118], and many of the regions contain deleterious SNPs in genes with disparate functions.

Within hitchhiking regions, we found an elevated ratio of deleterious to neutral SNPs caused by a reduction in the number of neutral SNPs. The elevated ratio of deleterious to neutral SNPs is consistent with simulations of both single and recurrent hitchhiking events across a range of parameters (Figure 3.1 and Figure S3.1) and can be explained by the difference in the frequency distribution of deleterious and neutral SNPs prior to hitchhiking. During a hitchhiking event neutral and deleterious alleles increase or decrease in frequency depending on their original configuration with the advantageous mutation. However, rare alleles are more likely to be deleterious and common alleles are more likely to be neutral. Thus, positive selection removes many common alleles, which tend to be neutral, and increases the frequency of many rare alleles, which tend to be deleterious, resulting in an increase in the ratio of deleterious to neutral SNPs. However, the simulated hitchhiking events showed two patterns that were not observed in the human data. First, hitchhiking caused a reduction in the number of deleterious SNPs. Second, hitchhiking caused a much larger increase in the ratio of deleterious to neutral SNPs at high frequencies relative to that at low frequencies. The significance of these differences is hard to evaluate since many factors known to influence hitchhiking were not examined, e.g. dominance, population structure, changes in population size and selection on new mutations versus standing genetic variation. Furthermore, hitchhiking simulations with high rates of recombination or strong selection against deleterious mutations tended to show patterns that are more consistent with those observed in humans (Figure 3.1 and Figure S3.1). Although some theory results have recently been obtained [84], further work will be needed to understand the effects of hitchhiking on deleterious mutations in humans.

A number of factors besides hitchhiking may contribute to the increased ratio of deleterious to neutral SNPs. Background selection is expected to increase the ratio of deleterious to neutral SNPs, particularly within regions of low recombination (Figure 3.1). While the rate of recombination can explain some of the difference between hitchhiking and non-hitchhiking regions, the ratio of deleterious to neutral SNPs is significantly higher in hitchhiking regions even after controlling for differences in recombination rate between hitchhiking and non-hitchhiking regions. Given the slightly lower rates of recombination in hitchhiking regions, the logistic regression model predicts hitchhiking regions should have a ratio of deleterious to neutral SNPs of 0.462, which is only slightly higher than that in non-hitchhiking regions, 0.441, and less than that observed, 0.531. It is conceivable that background selection may exert much weaker effects over shorter intervals that are not related to regional rates of recombination. However, weak background selection would have to exert a stronger influence within hitchhiking compared to non-hitchhiking regions, making it difficult to attribute the increased ratio of deleterious to neutral SNPs within these regions to background selection alone.

Another factor that complicates the analysis of differences between hitchhiking and non-hitchhiking regions is how hitchhiking regions were defined. Hitchhiking regions were defined by genome scans for patterns of variation expected to occur as a result of positive selection. However, some regions identified in genome scans for selection are likely neutral outliers that by chance show patterns of variation similar to those created by hitchhiking. This was one of our main motivations for using hitchhiking regions defined by two or more genome scans for selection. Although a contribution from neutral outliers cannot be excluded, the observation that the ratio of deleterious to neutral SNPs is 1.87-fold higher in regions identified by all nine genome scans and 1.68-fold higher in regions containing genes known to have been under positive selection suggests that hitchhiking makes a significant contribution to the elevated ratio of deleterious to neutral SNPs.

Similar to deleterious SNPs, common, disease-associated SNPs are enriched in hitchhiking compared to non-hitchhiking regions. In contrast, the number of rare, disease-associated mutations in hitchhiking relative to non-hitchhiking regions is lower than that of OMIM morbid genes. This difference can be explained by hitchhiking. Since most rare disease mutations occur on different chromosomes, hitchhiking will increase the frequency of one or a small number of disease mutations but decrease or eliminate the majority of rare disease mutations. However, the difference between rare and common disease-associated alleles is complicated by the heterogeneous evidence used to define disease-associated mutations in OMIM and the fact that common mutations are more likely to be associated with disease than rare mutations. The effect of hitchhiking on GWAS SNPs is more complex since most GWAS SNPs may be neutral. The ratio of GWAS SNPs in hitchhiking to non-hitchhiking regions is lower than that of all genes or neutral SNPs (Table S3.4). The lower frequency of GWAS SNPs in hitchhiking to non-hitchhiking regions is consistent with a previous study [115][50] and may be caused by the removal of common SNPs and reduced power of linkage disequilibrium-based tests of association. Consistent with this possibility, the hitchhiking to non-hitchhiking ratio of GWAS SNPs that are nonsynonymous, and thus more likely to be causative, is higher than that of all GWAS SNPs.

Our results also bear on the incidence of certain human diseases [60; 120] and disease alleles [121], which in some cases are higher than what one might expect based on disease severity. While genetic drift and population bottlenecks are likely to contribute to common disease alleles, balancing selection has also been invoked in some instances. For example, the high frequency of the delta F508 mutation in *CFTR* has been hypothesized to be the result of a heterozygote advantage due to cholera resistance [122; 123]. Mutations in *G6PD* and *Beta-globin* have been hypothesized to provide a heterozygote advantage due to malaria resistance [121]. Another explanation for why some disease alleles are so common is the ancestral-susceptibility hypothesis, under which derived alleles associated with human disease were advantageous to ancestral lifestyles and environmental conditions [119]. Similarly, under the less is more model, loss of function mutations that were previously disadvantageous can become advantageous [124]. In support of this model, we found five out of six nonsynonymous SNPs that are putative targets of positive selection are highly conserved across species and so classified as deleterious.

However, our results also provide evidence for an alternative explanation for the frequency of common disease-associated alleles: the frequency of certain disease alleles is increased due to hitchhiking with linked advantageous mutations. A number of previous observations support this explanation. The MHC locus has been associated with over 40 human genetic diseases [125], and multiple lines of evidence suggest long-term balancing selection [126]. A mutation in *HFE* that causes hemochromatosis is 150 kb away from a hitchhiking region and may have increased in frequency due to hitchhiking [127-129]. Hitchhiking has also been implicated in the increased frequency of a common risk haplotype for diabetes, hypertension and celiac disease [130] and another risk haplotype for Crohn's disease [131]. Intriguingly, the delta F508 mutation in *CFTR* is one of the most common disease-causing alleles in Caucasians, with an estimated allele frequency of 1.4% [132], and *CFTR* occurs within a hitchhiking region. Four of the HapMap nonsynonymous SNPs within *CFTR* are classified as deleterious, one of which has been associated with infertility [133]. One of the regions with the strongest evidence for hitchhiking (7 tests) also has one of the highest ratios of deleterious to neutral SNPs (16/22, Table S3.7). Within this region, 8/16 deleterious SNPs occur in *BLK*, *NEIL2*, and *CTSB*, and there are three disease alleles in the Human Gene Mutation Database [134], with frequencies of 0.8%, 5.3% and 44% based on the 1000 Genomes Project. The frequency of these deleterious/disease alleles may have been influenced by positive selection in the region.

The interaction between positive and negative selection makes it difficult to isolate and understand the effects of each individually. In the presence of deleterious mutations, the effect of hitchhiking on linked neutral variation may be reduced compared to that which would occur in the absence of deleterious mutations, similar to patterns created by soft sweeps [117]. Conversely, hitchhiking increases the frequency of some deleterious mutations and decreases the frequency of others such that the distribution of deleterious mutations is significant different from that expected in the absence of hitchhiking. Furthermore, the recent expansion in human population size combined with population subdivision may amplify or reduce the influence of hitchhiking on deleterious SNPs. This will make it valuable to examine the extent to which deleterious alleles are enriched in hitchhiking regions in other species, particularly domesticated species where the strength of selection was likely strong and for which targets of selection are in some cases known.

MATERIALS AND ETHODS

Computer Simulations

The effects of hitchhiking on deleterious and neutral polymorphism were simulated using a Wright-Fisher model [135]. Simulated populations had a size, N, of 1000 diploid individuals. Mutations were distributed into the population assuming an infinite sites model with a Poisson rate of $2Nu$, where u is the mutation rate per chromosome. A Poisson number of recombination events was generated in the population with a rate of Nr, where r is the rate of recombination per individual. Chromosomes in the next generation were sampled based on the fitness of the individual from which they were derived. Fitness was calculated by the multiplicative effects of each non-neutral allele, $1+hs$ for heterozygous sites and $1+s$ for homozygous sites, where s is selection coefficient and h is the degree of dominance. The dominance coefficient was 0.5 for all simulations. For each set of parameters, simulations were run for $20N$ generations before sampling. For a single hitchhiking event, an advantageous mutation was generated in the center of the chromosome and sampled at the end of hitchhiking conditional on its fixation. For multiple hitchhiking events, advantageous mutations were generated at a constant rate uniformly across the chromosome and samples were taken in intervals of N generations. θ_W, $\theta\pi$ and θ_H were estimated using a sample size of 100 chromosomes as described in [92].

Classification of neutral and deleterious SNPs

Low-coverage SNP calls for CEU, CHB+JPT, and YRI samples were downloaded from the 1000 Genomes Project (release 2010_07) [93], and all tri-allelic sites were filtered out. Coding SNPs were identified based on their genomic coordinates in the NCBI reference genome (build 36) and Ensembl known genes (release #49). After eliminating SNPs on the sex chromosomes, SNPs in known pseudogenes or gene fragments, and sites monomorphic across CEU, CHB+JPT and YRI samples, there were

47,730 synonymous and 48,558 nonsynonymous SNPs within coding regions with multi-species alignments used by the likelihood ratio test (see below).

Nonsynonymous SNPs were classified as neutral or deleterious using a previously implemented likelihood ratio test (LRT) for conservation across multiple species [15]. The LRT is based on 18,993 multiple sequence alignments from 32 vertebrate species. Positions with less than 10 aligned eutherian mammals were excluded from the analysis due to low power of the LRT. At each codon in the alignment, the LRT calculates the likelihood of the data under a neutral model, where the nonsynonymous substitution rate (*dN*) equals the synonymous substitution rate (*dS*), relative to a conserved model, where *dN* can deviate from *dS*. For these calculations, *dS* is set to an average rate of 12.2 substitutions per site across the entire tree based on an estimate from gap-free concatenated alignments of 1,227 genes (54 kb) with data from all species. Nonsynonymous SNPs were predicted to be deleterious if: 1) the codon is significantly conserved by the LRT ($P < 0.001$), 2) *dN* is less than *dS*, and 3) the derived amino acid is not present at orthologous positions in other eutherian mammals.

Correlation of SNP density with recombination rate

The density of SNPs was measured as a function of local recombination rate using CEU, CHB+JPT, and YRI SNPs from the 1000 Genomes Project. Following previous work [136], recombination rates were estimated from non-overlapping 400 kb windows by dividing the genetic map distance of the two most distant SNPs by their physical distance. The genetic map, estimated by LDhat [137], was obtained from the 1000

Genomes Project. Windows that were less than 10 Mb away from the end of centromeres and telomeres, windows without a pair of SNPs greater than 360 kb apart, and windows with no aligned coding sequence were excluded. The remaining 3,666 windows were assigned into ten equal-sized bins by their recombination rates, and the number of synonymous, nonsynonymous deleterious and nonsynonymous neutral SNPs was counted per kb of aligned coding sequence in each bin. To account for the the proportion of codons that are conserved, which is correlated with both the rate of recombination and the number of G or C nucleotides within codon (Figure S3.2A), codons in each recombination bin were subdivided into four classes by the number of GC nucleotides within the human codon ($j = 0, 1, \ldots, 3$). In cases of polymorphic codons, GC content of the ancestral codon were counted. A total of 6,248,078 codons were classified as significantly conserved or not by the LRT at a P-value cutoff of 0.001. The relationship between recombination and the ratio of deleterious to neutral SNPs was assessed using the logistic regression model:

$$logit\left(\frac{DEL_{i,j}}{DEL_{i,j}+NEU_{i,j}}\right)=\log\left(\frac{DEL_{i,j}}{NEU_{i,j}}\right)=\beta_0+\beta_1\cdot r_i+\beta_2\cdot s_{i,j}$$

where $DEL_{i,j}$, and $NEU_{i,j}$, are the number of deleterious and neutral nonsynonymous SNPs, respectively, r_{i} is the average recombination rate of windows in bin i, and $s_{i,j}$ adjusts for differences in the number of potentially deleterious sites. $s_{i,j}$ was estimated by:

$$s_{i,j}=\frac{fcon_{i,j}}{fcon_j}\cdot fdel_j$$

where $fcon_{i,j}$ is the fraction of conserved codons out of all aligned codons with j GC nucleotides in bin i, $fcon_j$ is the mean of $fcon_{i,j}$ over all $i = 1, \ldots, 10$, and $fdel_j$ is the fraction of deleterious out of all tested nonsynonymous SNPs with the same j GC nucleotides.

To account for biased gene conversion, which has been previously proposed to explain a higher rate of GC-biased disease alleles in regions of higher recombination [138], we re-examined the relationship between the ratio of deleterious to neutral SNPs and recombination after excluding 13,995 AT-to-GC mutating SNPs potentially affected by biased gene conversion. SNPs within codons with zero GC nucleotides were also eliminated due to their relatively small number (N = 335). Using the logistic regression model that accounts for the variation in the number of potentially deleterious sites, the regression coefficient β_1 of recombination rate remained similar (-0.097 to -0.101) and highly significant ($P = 9.3 \times 10^{-6}$).

SNPs in hitchhiking and non-hitchhiking regions

Hitchhiking regions were defined by genomic intervals that were identified by two or more out of nine tests for hitchhiking, using intervals rounded to the nearest multiple of 10 kbp [98]. To compare different methods, we examined regions that were identified by one method and overlapped with any other method. Non-hitchhiking regions were defined as autosomal regions excluding hitchhiking regions as defined above. The density of deleterious and neutral nonsynonymous SNPs was measured relative to the accessible portion of aligned coding regions used for likelihood ratio test. The accessible genome, which satisfies minimum read depth required for SNP calling, was obtained from the 1000 Genomes Project for CEU, CHB+JPT, and YRI [93], and their union was used for the combined analysis of all samples. The difference between the SNP density within hitchhiking and non-hitchhiking regions was tested by a two-proportion z-test.

To test whether a higher ratio of deleterious to neutral SNPs in hitchhiking relative to non-hitchhiking regions is caused by a higher recombination rate or a larger number of potentially deleterious sites in hitchhiking regions, the 400-kb genomic windows which were already binned by the rate of recombination and the number of GC nucleotides in a codon were further classified into hitchhiking and non-hitchhiking groups. After removing windows near centromeres and telomeres, there were 388 windows identified by three or more tests of hitchhiking that were assigned to the hitchhiking group ($h = 1$), and 2,917 windows without any hitchhiking regions that were assigned to the non-hitchhiking group ($h = 0$). The data were fit to the following logistic regression model:

$$logit\left(\frac{DEL_{i,j,h}}{DEL_{i,j,h}+NEU_{i,j,h}}\right)=\log\left(\frac{DEL_{i,j,h}}{NEU_{i,j,h}}\right)$$
$$=\beta_0+\beta_1 \cdot r_i+\beta_2 \cdot s_{i,j,h}+\beta_3 \cdot h$$

where r_{i} is the rate of recombination, $s_{i,j,h}$ adjusts for the density of conserved codons and h is an indicator variable for hitchhiking windows.

To study the decay of the ratio of deleterious to neutral SNPs as a function of distance from hitchhiking regions, we used regions identified in CEU by iHS [100] and Rsb [101]. For iHS, the top 5% of scanned genomic windows (a total of 127.6 Mb) were used as hitchhiking regions, as described below. For Rsb, we used regions identified in CEU in comparison to both YRI and CHB+JPT (a total of 119.7 Mb). Deleterious and neutral nonsynonymous SNPs outside iHS and Rsb regions were assigned into bins of non-overlapping 200-kb windows by their distance from the nearest hitchhiking region. The ratio of deleterious to neutral SNPs was modeled as a function of the distance (d_k) of each window in bin k to the nearest hitchhiking region using logistic regression:

$$logit\left(\frac{DEL_k}{DEL_k + NEU_k}\right) = \log\left(\frac{DEL_k}{NEU_k}\right) = \beta_0 + \beta_1 \cdot d_k$$

Population specific patterns of hitchhiking were examined using regions identified by multiple tests of selection and by iHS alone. Regions identified by multiple tests of selection were not differentiated by which population showed evidence of selection and so represent a composite view of hitchhiking [98]. iHS regions were identified in CEU, CHB+JPT, and YRI, using empirical cutoffs of 0.25%, 1%, and 5%. To identify iHS hitchhiking regions using HapMap Phase II data, iHS scores of individual SNPs (HapMap Phase II) were downloaded (http://hg-wen.uchicago.edu/selection/), and for each 100-kb non-overlapping genomic window the signal of selection was evaluated by the fraction of SNPs with iHS scores above +2 or below -2, as in Voight et al. [100]. Windows were grouped into bins by the number of SNPs within the window using increments of 25 SNPs. Empirical cutoffs were applied separately to each bin. Windows with less than 10 SNPs and bins with less than 100 windows (less than 400 for the 0.25% cutoff) were excluded.

Disease-associated alleles in hitchhiking and non-hitchhiking regions

Disease-associated alleles were obtained from OMIM (http://www.ncbi.nlm.nih.gov/omim), a catalog of published GWAS studies (http://www.genome.gov/26525384) and Google Scholar searches of the literature. For OMIM, dbSNP IDs (release #132) with OMIM links were downloaded (ftp://ftp.ncbi.nih.gov/snp/database/organism_data/human_9606/OmimVarLocusIdSNP.bcp.gz). Excluding InDels, unmapped variants, and variants on sex chromosomes, 10,775 OMIM variants were re-mapped to the reference genome using UCSC's LiftOver program. All OMIM variants included in HapMap Phase II (release #24) were considered common enough to be SNPs with the exception of those with minor allele frequency of zero. Average minor allele frequency across CEU, CHB+JPT, and YRI was compared between hitchhiking and non-hitchhiking regions. For allele frequency, HapMap Phase II+III (release #26) data were used [139]. For disease SNPs identified in the 1000 Human Genomes project [93], common and rare variants were distinguished by their mean allele frequencies across CEU, CHB+JPT, and YRI using a 5% allele frequency cutoff. SNPs without allele frequencies were set to an allele frequency of zero.

Disease-risk alleles were obtained from a catalog of published Genome-Wide Association Studies (GWAS) [115]. Excluding 115 regions without associated SNPs and 10 regions with multi-SNP haplotype associations, we obtained 3,383 non-redundant autosomal risk alleles with the strongest trait association at each locus from a total of 585 published studies. Allele frequencies in control population and odds ratios were available for 2,504 and 1,253 risk alleles, respectively. Reported risk allele frequencies were averaged over control populations if the risk allele was identified in more than two studies. However, reported odds ratios were not pooled over different studies and traits even if the risk allele was reported in multiple studies.

To examine common deleterious and neutral SNPs reported in the literature, we used Google Scholar (http://scholar.google.com) and the dbSNP rs number as the search term. The set of tested SNPs was based on 790 deleterious SNPs and 369 neutral nonsynonymous SNPs with an allele frequency of greater than 30% in the HapMap CEU panel. SNPs within known olfactory receptors were excluded. Neutral SNPs were matched to the frequency distribution of deleterious SNPs by acceptance-rejection sampling. As a result, derived allele frequencies are not significantly different between the two sets (Wilcoxon Rank Sum Test, P = 0.79). For each SNP, we searched for reported phenotype associations based on population association or cell-based functional assays. To minimize potential human biases, dbSNP identifiers of deleterious and neutral SNPs were mixed together and Google Scholar search results were manually examined without knowledge of SNP classification. Patents, eQTL associations, conference and poster abstracts, and journals without full-text access were excluded. SNP associations had to be significant after a multiple testing correction. SNP association studies with sample size less than 200 were also not included.

Genome clustering of deleterious SNPs

To identify genomic regions with exceptionally high rates of deleterious SNPs per coding sequence, 1-Mb sliding windows were scanned across all autosomes with a step size of 0.5Mb. Assuming that the rate of deleterious SNPs per accessible coding sequence is constant across the genome, a Poisson distribution was used to evaluate the excess number of deleterious SNPs in each window. The expected number of deleterious SNPs per window was set to the product of the genome average (0.51 deleterious SNPs per 1kb accessible CDS) and the length of accessible coding sequence in the window. Out of 3,549 windows with at least two deleterious SNPs, 70 (2%) with the highest P-value were selected ($P < 4.5 \times 10^{-4}$). After excluding regions that were consecutive to or overlapped another region with a smaller P-value, we retained 43 regions.

ACKNOWLEDGEMENTS

We thank members of the Fay lab, Katherine Geist and three anonymous reviewers for their comments.

FIGURES

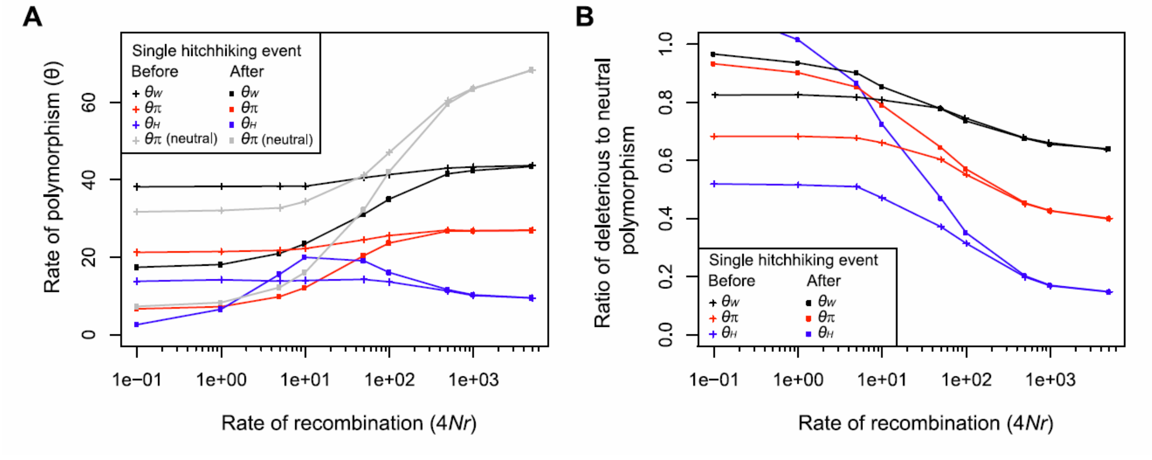


Figure 3.1. The effect of hitchhiking on neutral and deleterious polymorphism as a function of the rate of recombination. The rate of low, intermediate and high frequency deleterious polymorphism measured by θ_W (black), θ_π (red) and θ_H (blue), respectively, before (crosses) and after (squares) a single hitchhiking event (A) and in the presence (crosses) and absence (squares) of multiple hitchhiking events (C). Average heterozygosity (θ_π) of neutral polymorphism is shown in gray. The ratio of deleterious to neutral polymorphism before (crosses) and after (squares) a single hitchhiking event (B) and in the absence (crosses) and presence (squares) of multiple hitchhiking events (D). All panels show the mean of 500 simulations for which $4Nu_n = 70$, $4Nu_d = 70$, $4Ns_d = -10$ and $4Ns_a = 100$, where N is the population size, u is the mutation rate, s is the selection coefficient, and subscripts n, a and d refer to neutral, advantageous and deleterious mutations. In panel A, a single hitchhiking event occurs at the center of the chromosome. In panel B, $4Nu_a = 0.5$ and multiple hitchhiking events occur across the entire chromosome.

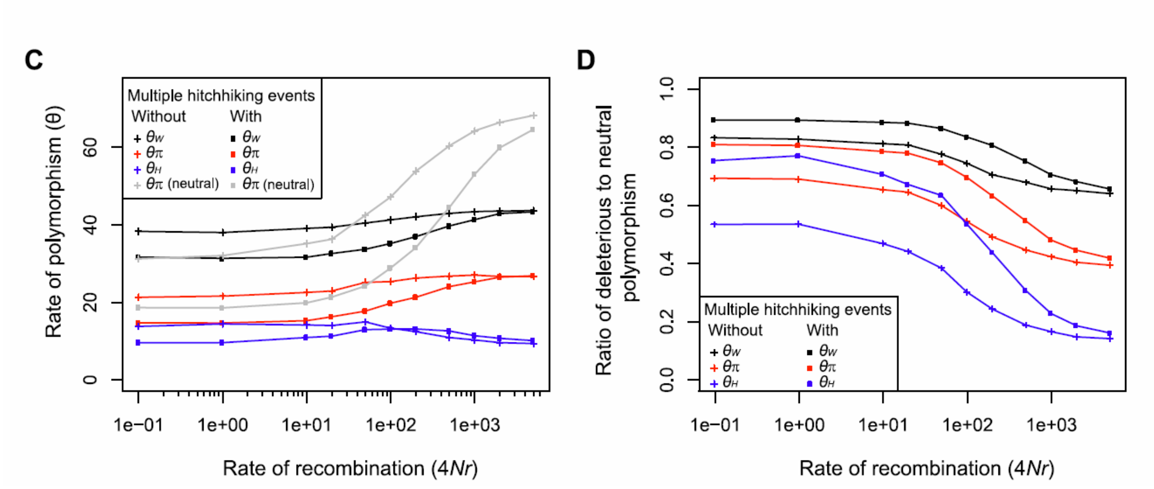


Figure 3.1. (Continued)

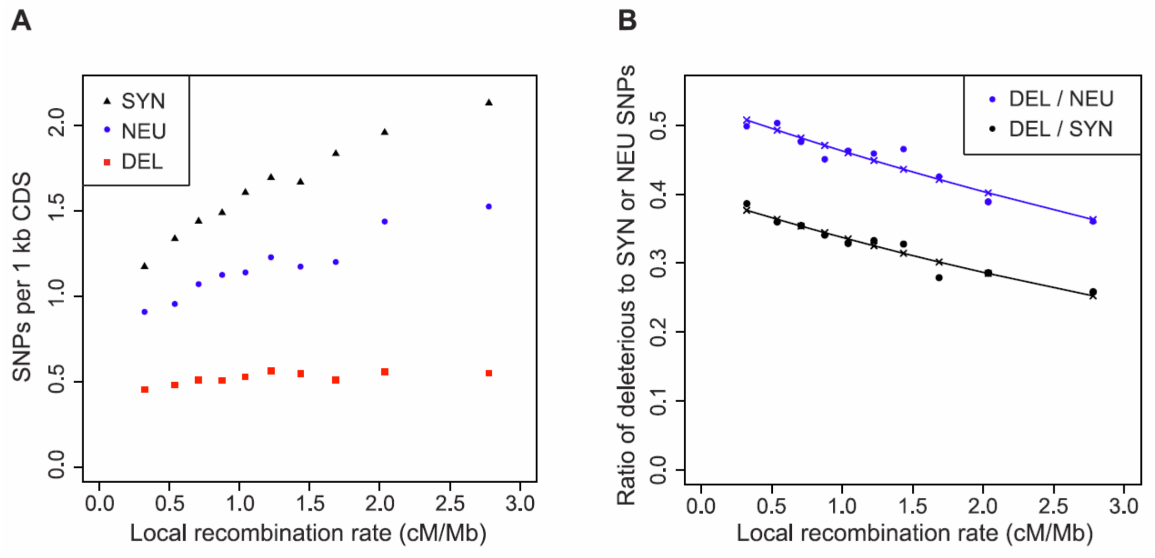


Figure 3.2. Regions of low recombination are enriched for deleterious SNPs. The number of synonymous (SYN) and neutral nonsynonymous (NEU) and deleterious (DEL) SNPs per kb of coding sequence (A) and the ratio of deleterious to synonymous or neutral nonsynonymous SNPs (B) as a function of the local recombination rate. The rate of neutral and deleterious SNPs was normalized by the number of sites that were testable by the likelihood ratio test. Lines show the results of logistic regression.

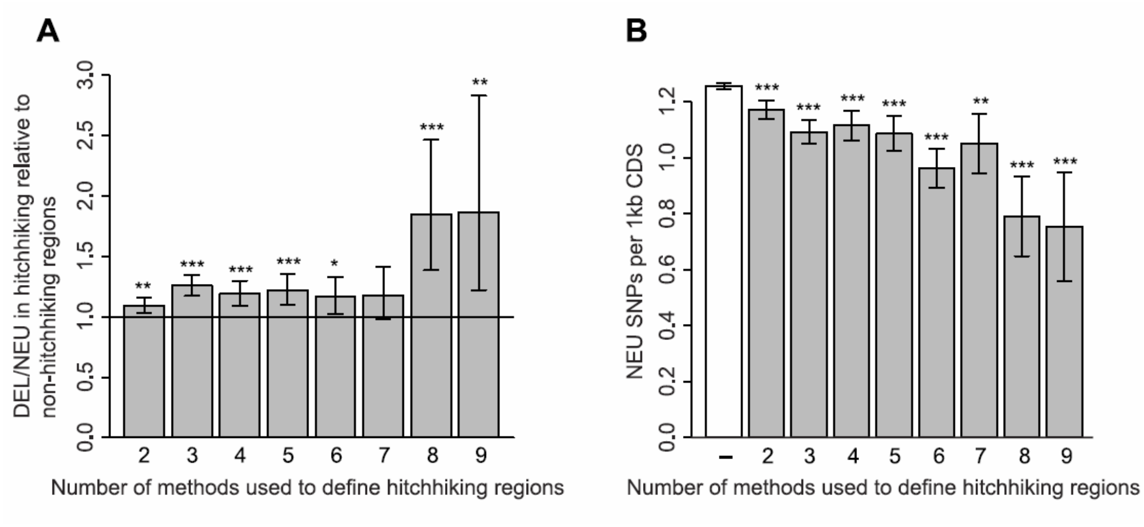


Figure 3.3. Rates of deleterious and neutral SNPs in hitchhiking and non-hitchhiking regions. The ratio of deleterious (DEL) to neutral (NEU) SNPs is higher in hitchhiking relative to non-hitchhiking regions (A). The rate of neutral SNPs is reduced (B) and the rate of deleterious SNPs remains relatively constant (C) in hitchhiking compared to non-hitchhiking regions. The x-axes in panels A-C denotes the minimum number of methods used to define hitchhiking regions. Non-hitchhiking regions are labeled by a dash(–). The ratio of deleterious to neutral SNPs is higher in hitchhiking to non-hitchhiking regions for the majority of tests of selection (D). Tajima's D was used by two studies: [140][1] and [141][2]. Bars show 90% confidence intervals, one, two and three stars indicate $P < 0.05$, $P < 0.01$, and $P < 0.001$ based on a one-sided Fisher's Exact Test.

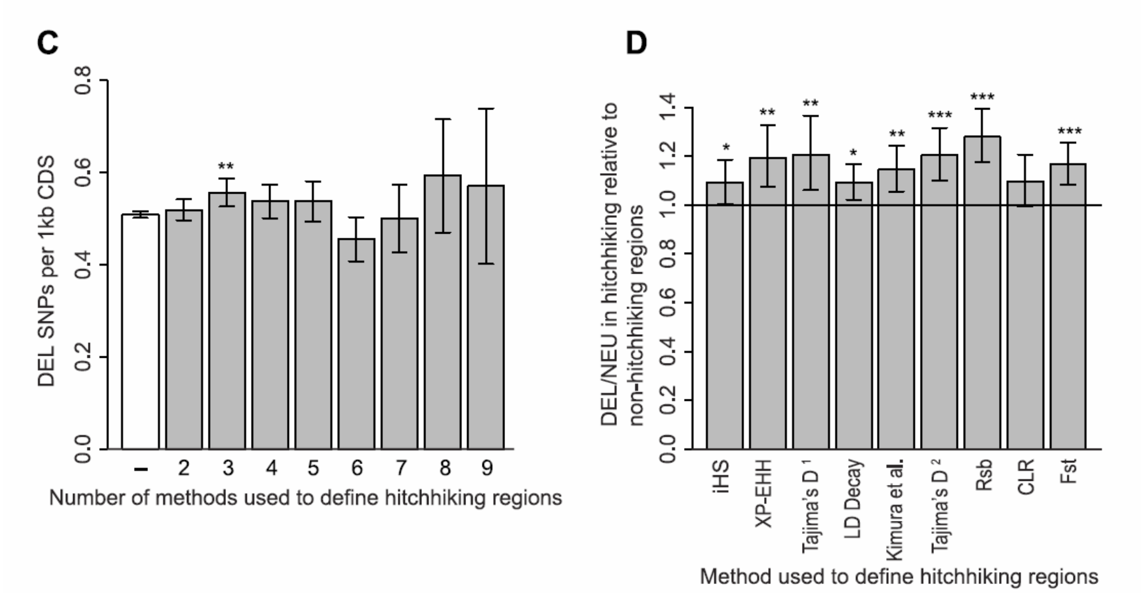


Figure 3.3. (Continued)

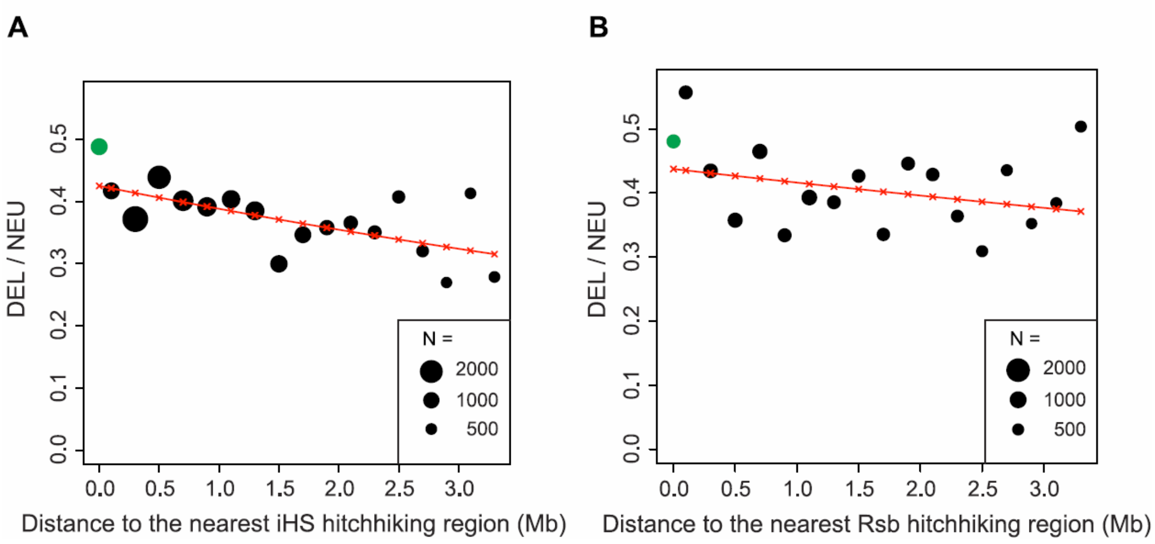


Figure 3.4. The ratio of deleterious to neutral nonsynonymous SNPs declines as a function of distance to the nearest hitchhiking region. Hitchhiking regions were defined using the European population by iHS (A) or Rsb (B). Sample size is indicated by circle size. Green circles represent iHS and Rsb hitchhiking regions.

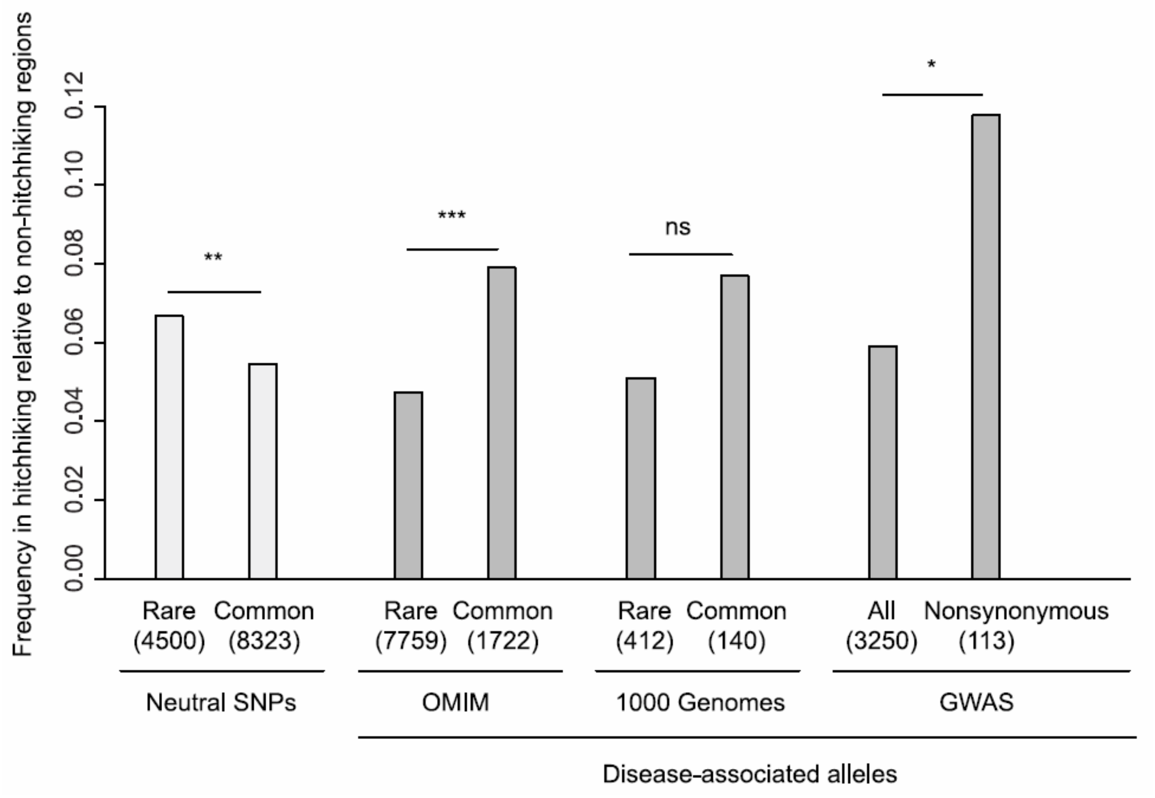


Figure 3.5. Enrichment of disease-associated alleles in hitchhiking relative to non-hitchhiking regions. Each category shows the number within hitchhiking to non-hitchhiking regions, where hitchhiking regions were defined by the overlap of three or more tests of selection. Neutral SNPs are from HapMap Phase II. Disease alleles in1000 Genomes columns are based on the Human Gene Mutation Database. The sample size of each category is shown in parentheses. Bars show one-sided Fisher's Exact test comparisons, not significant (ns), $P < 0.05$ (*), $P < 0.01$ (**), and $P < 0.001$ (***).

TABLES

Table 3.1. The ratio of deleterious to neutral SNPs at low, intermediate and high frequencies.

Allele frequency	Deleterious / Neutral		Fold increase in hitchhiking regions (95% CI)
	Hitchhiking	Non-hitchhiking	
Low (0 – 0.008)	650 / 1013 (0.64)	5469 / 9109 (0.60)	1.07 (0.96 – 1.19)
Intermediate (0.008 – 0.059)	470 / 1050 (0.45)	4241 / 10376 (0.41)	1.10 (0.97 – 1.23)
High (0.059 – 1.0)	329 / 1206 (0.27)	2935 / 11710 (0.25)	1.09 (0.95 – 1.24)

Hitchhiking regions are defined by two or more tests of selection.

Table 3.2. Deleterious SNPs within and around genes under positive selection.

	Within target gene		Target gene +/- 1 Mbp of flanking region		
Putative target of selection	**Deleleterious [1]**	**Neutral**	**Deleterious [1]**	**Neutral**	**Genes with deleterious SNPs [2]**
LCT, noncoding (lactose tolerance)	1	2	4	6	*R3HDM1 (2), LCT, MCM6*
SLC45A2, L374F (pigmentation)	1 (1)	0	3 (1)	8	*ADAMTS12, SLC45A2, C1QTNF*
TYRP1, noncoding (pigmentation)	0	0	1	3	*MPDZ*
OCA2-HERC2, noncoding (pigmentation)	3	2	3	2	*OCA2 (3)*
KITLG, noncoding (pigmentation)	0	0	1	5	*CEP290*
SLC24A5, A111T (pigmentation)	1 (1)	0	2 (1)	3	*SLC24A5, SLC12A1*
TYR, S192Y (pigmentation)	2 (1)	0	6 (1)	1	*GRM5, TYR (2), FOLHB1 (3)*
EDAR, V370A (thicker hair)	1 (1)	1	2 (1)	5	*GCC2, EDAR*
PCDH15, D435A (unknown)	0	0	0	0	
LEPR, K109R (metabolism)	2 (1)	1	2 (1)	2	*LEPR(2)*
Total	11 (5)	6	24 (5)	35	

The number of deleterious and neutral nonsynonymous SNPs were tabulated using the European sample (CEU), except for EDAR, PCDH15 and LEPR, which were tabulated using the Asian sample (CHB+JPT).

A111T in *SLC24A5*, one deleterious SNP in OCA2, and three neutral SNPs in flanking regions of *LCT*, *SLC45A2*, and *SLC24A5* are Fixed or nearly fixed in CEU.

[1] Putatively functional nonsynonymous SNPs under selection are in parentheses.

[2] The number of deleterious SNPs within flanking genes is in parenthesis if greater than one.

SUPPORTING INFORMATION

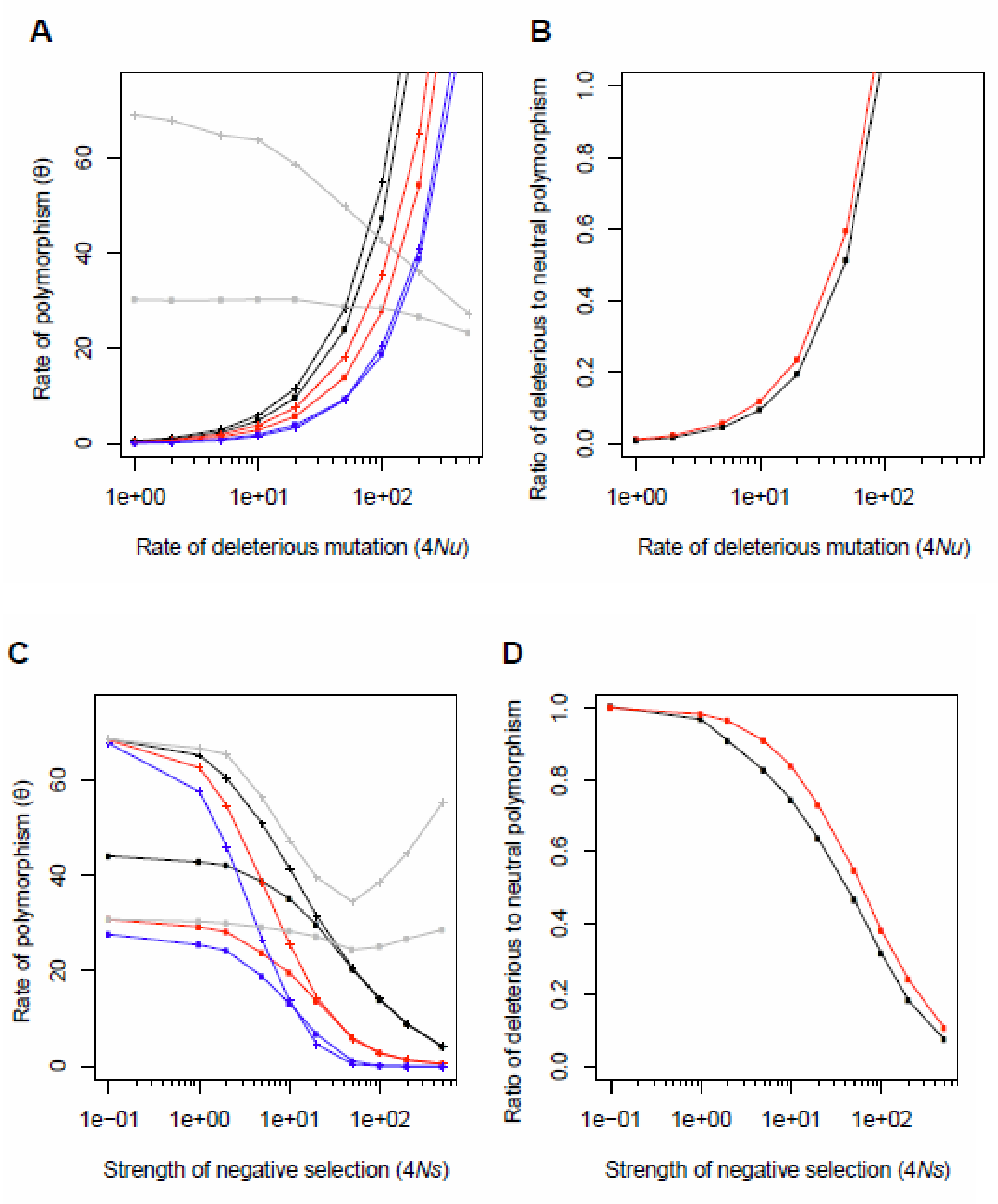


Figure S3.1. The effect of hitchhiking on neutral and deleterious polymorphism as a function of the rate and strength of advantageous and deleterious mutations.

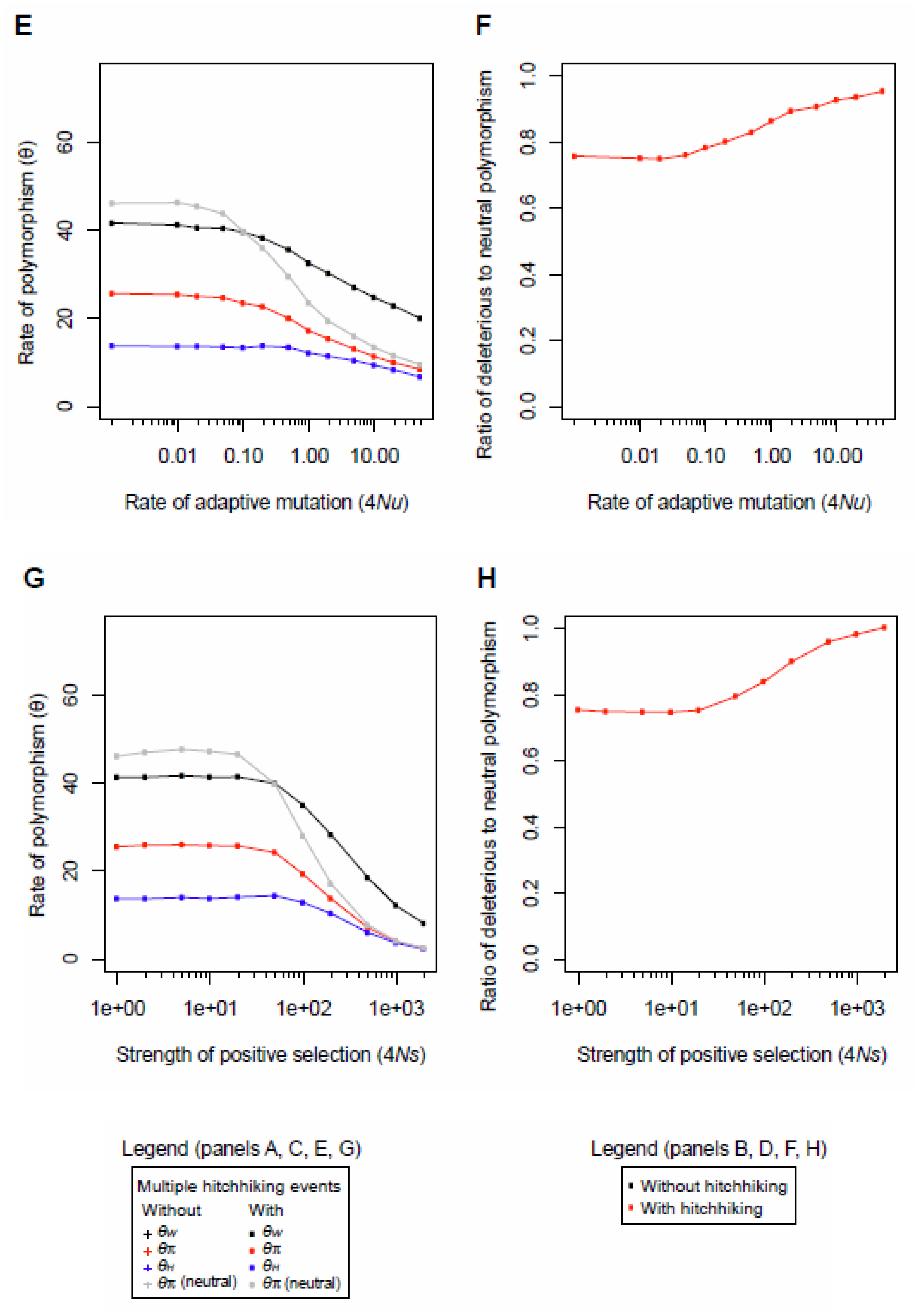


Figure S3.1. (Continued)

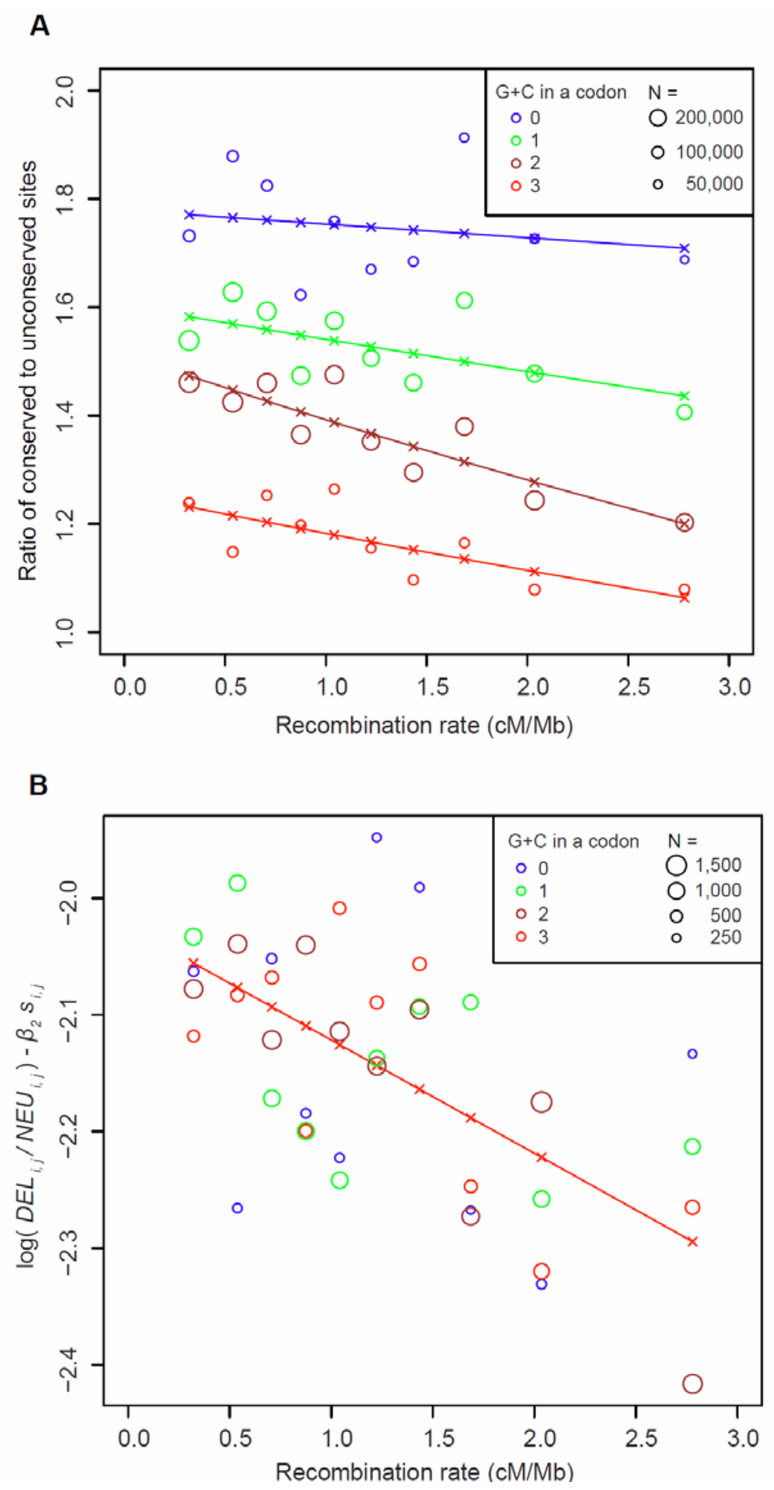


Figure S3.2. The ratio of deleterious to neutral SNPs is associated with the rate of recombination.

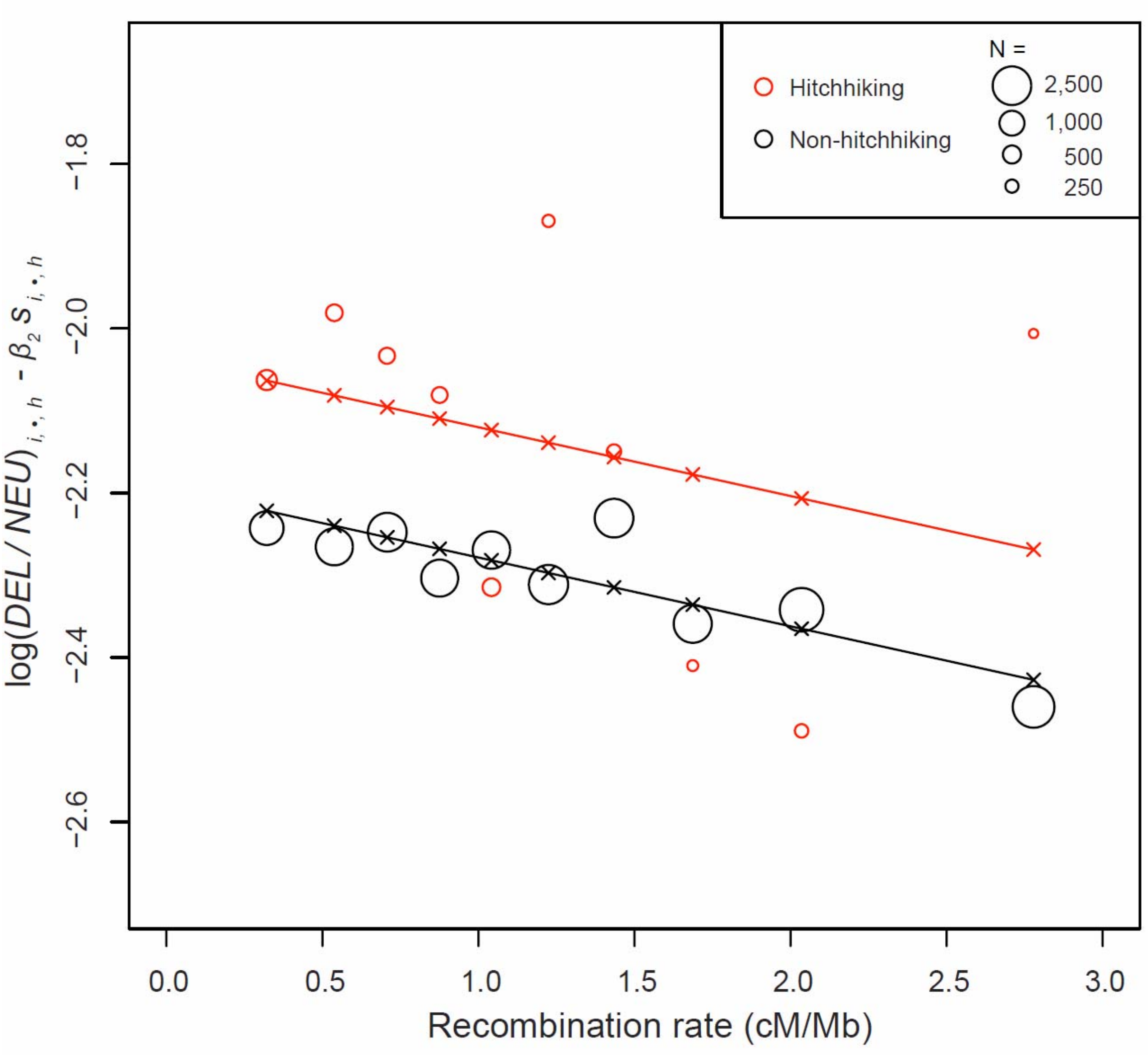


Figure S3.3. Hitchhiking regions are enriched for deleterious SNPs.

Table S3.1. Characteristics of hitchhiking regions.

Region	Number of Regions	Total size (Mb)	Median size (kb)	Deleterious		Neutral		DEL / NEU	DEL/NEU in hitchhiking relative to non-hitchhiking regions
				Count	per 1kb CDS	Count	per 1kb CDS		
2 or more	722	250	240	1449	0.52	3269	**1.17** ***	0.44	**1.09** **
3 or more	271	150	500	921	**0.56** **	1808	**1.09** ***	0.51	**1.26** ***
4 or more	128	93	665	582	0.54	1207	**1.12** ***	0.48	**1.19** ***
5 or more	69	60	790	406	0.54	820	**1.09** ***	0.50	**1.22** ***
6 or more	43	41	900	243	0.46	514	**0.96** ***	0.47	**1.17** *
7 or more	20	21	1015	126	0.50	264	**1.05** **	0.48	1.18
8 or more	8	10	1190	63	0.59	84	**0.79** ***	0.75	**1.85** ***
9	3	3	1020	31	0.57	41	**0.75** ***	0.76	**1.87** **
iHS	279	106	240	617	**0.55** *	1395	1.24	0.44	**1.09** *
XP-EHH	135	73	460	394	0.54	814	**1.12** ***	0.48	**1.19** **
Tajima's D (Carlson et al.)	50	39	730	270	0.51	553	**1.05** ***	0.49	**1.20** **
LD Decay	351	132	240	969	0.52	2189	**1.17** ***	0.44	**1.09** *
Kimura et al	325	120	250	634	0.50	1365	**1.08** ***	0.46	**1.15** **
Tajima's D (Kelley et al.)	136	61	340	540	**0.56** *	1107	**1.14** **	0.49	**1.20** ***
Rsb	221	130	510	608	**0.56** *	1173	**1.08** ***	0.52	**1.28** ***
CLR	129	80	540	445	0.54	1002	**1.21** *	0.44	1.10
Fst	360	137	245	793	0.49	1678	**1.04** ***	0.47	**1.17** ***
Non-hitchhiking regions	730	2618	2160	12645	0.51	31195	1.26	0.41	1.00

Regions are defined by a single method, two or more methods or by non-hitchhiking regions that were not identified by any of the 9 methods.

A significant increase in the rate of deleterious SNPs, decrease in the rate of neutral SNPs, and increase in the ratio of deleterious to neutral SNPs was tested within hitchhiking relative to non-hitchhiking regions and are marked in bold, $P < 0.05$ (*), $P < 0.01$ (**), $P < 0.001$ (***).

Table S3.2. The ratio of deleterious to neutral SNPs in iHS population-specific hitchhiking regions.

		Deleterious / Neutral		**Fold increase in**
Panel	**iHS cutoff**	**Hitchhiking**	**Non-hitchhiking**	**hitchhiking regions**
CEU	5%	351 / 719 (0.49)	5918 / 15869 (0.37)	**1.31** ***
	1%	55 / 100 (0.55)	6214 / 16488 (0.38)	**1.46** *
	0.25%	23 / 31 (0.74)	6246 / 16557 (0.38)	**1.97** *
CHB+JPT	5%	213 / 579 (0.37)	4920 / 12998 (0.38)	0.97
	1%	43 / 107 (0.40)	5090 / 13470 (0.38)	1.06
	0.25%	8 / 11 (0.73)	5125 / 13566 (0.38)	1.93
YRI	5%	288 / 965 (0.30)	7175 / 22326 (0.32)	0.93
	1%	45 / 218 (0.21)	7418 / 23073 (0.32)	0.64
	0.25%	12 / 56 (0.21)	7451 / 23235 (0.32)	0.67

* One-sided Fisher exact test $P < 0.05$

Table S3.3. Ratio of deleterious to neutral SNPs in three populations panels for regions identified by multiple tests of selection.

		Deleterious		Neutral			DEL/NEU in hitchhiking relative to
Panel	**Region**	**Count**	**per 1kb CDS**	**Count**	**per 1kb CDS**	**DEL / NEU**	**non-hitchhiking regions**
CEU	2 or more	643	0.23	1551	0.56	0.41	**1.11** *
	3 or more	397	0.24	865	0.52	0.46	**1.23** ***
	4 or more	244	0.23	568	0.52	0.43	**1.15** *
	5 or more	172	0.23	394	0.52	0.44	1.17
	6 or more	99	0.19	246	0.46	0.40	1.08
	7 or more	51	0.20	123	0.49	0.41	1.11
	8 or more	26	0.24	41	0.39	0.63	**1.69** *
	9	13	0.24	23	0.42	0.57	1.51
	Non-hitchhiking	5626	0.23	15037	0.61	0.37	1.00
CHB+JPT	2 or more	482	0.17	1144	0.41	0.42	**1.13** *
	3 or more	306	0.18	591	0.36	0.52	**1.38** ***
	4 or more	189	0.17	366	0.34	0.52	**1.38** ***
	5 or more	128	0.17	223	0.30	0.57	**1.53** ***
	6 or more	86	0.16	132	0.25	0.65	**1.74** ***
	7 or more	36	0.14	74	0.29	0.49	1.30
	8 or more	17	0.16	29	0.27	0.59	1.57
	9	7	0.13	15	0.28	0.47	1.25
	Non-hitchhiking	4651	0.19	12433	0.50	0.37	1.00
YRI	2 or more	788	0.28	2206	0.79	0.36	**1.13** **
	3 or more	492	0.30	1222	0.74	0.40	**1.27** ***
	4 or more	310	0.29	812	0.75	0.38	**1.21** **
	5 or more	215	0.28	545	0.72	0.39	**1.25** **
	6 or more	117	0.22	338	0.63	0.35	1.09
	7 or more	68	0.27	179	0.71	0.38	1.20
	8 or more	34	0.32	56	0.53	0.61	**1.92** **
	9	18	0.33	30	0.55	0.60	**1.90** *
	Non-hitchhiking	6675	0.27	21085	0.85	0.32	1.00

DEL/NEU ratios in hitchhiking relative to non-hitchhiking regions greater than one are marked in bold, Fisher's Exact Test $P < 0.05$ (*), $P < 0.01$ (**), $P < 0.001$ (***).

Table S3.4. Frequency of disease-associated alleles in hitchhiking and non-hitchhiking regions.

Class	Hitchhiking	Non-Hitchhiking	Ratio [1]
All genes [2]	1203	17601	0.068
OMIM morbid genes	105	1475	0.071
Rare neutral variants [3]	285	4215	0.068
Common neutral SNPs [3]	435	7888	0.055
OMIM rare variants [4]	351	7408	0.047
OMIM SNPs [4]	126	1596	0.079
1000 genomes rare variants [5]	20	392	0.051
1000 genomes SNPs [5]	10	130	0.077
GWAS SNPs [6]	181	3069	0.059
GWAS NSN SNPs	12	101	0.119
Literature SNPs [7]	3	103	0.029

[1] Ratio is the number within hitchhiking to non-hitchhiking regions, where hitchhiking regions are defined by the overlap of three or more tests of selection.

[2] Ensemble 49, known protein-coding genes, no pseudogene, no gene fragment, only autosomal genes.

[3] Neutral nonsynonymous variants in HapMap. Common SNPs are those with an allele frequency greater than or equal to 5%.

[4] Disease-associated variants that are rare or common enough to be typed in the HapMap project.

[5] 1000 human genomes variants that have been associated with human disease in the Human Gene Mutation Database. SNPs are those with an allele frequency greater than or equal to 5%.

[6] SNPs associated with human disease from genome-wide association studies ($p < 1.0 \times 10^{-5}$).

[7] Nonsynonymous SNPs which have been reported in the literature to be functional by disease-association or a functional assay.

The following supplementary tables are available from the PLoS genetics website (http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1002240).

Table S3.5. Disease or phenotype associated SNPs within hitchhiking regions.

Table S3.6. Genes in hitchhiking regions with deleterious SNPs.

Table S3.7. Genomic regions enriched for deleterious SNPs.

CHAPTER 4: Fine-mapping of the genetic association of *FSH receptor* with preterm birth in the Finnish population

Sung Chun[1], Jevon Plunkett[2], Kari Teramo[3], Louis J. Muglia[4] and Justin C. Fay[1,5]

[1] Computational and Systems Biology Program, Washington University, St. Louis, MO

[2] Program in Human and Statistical Genetics, Washington University, St. Louis, MO

[3] University of Helsinki, Finland

[4] Cincinnati Children's Hospital Medical Center

[5] Department of Genetics and Center for Genome Sciences and Systems Biology, Washington University, St. Louis, MO

This work was done in collaboration with Jevon Plunkett, Kari Termo, Louis Muglia and Justin Fay. Jevon Plunkett, Kari Termo and Louis Muglia recruited the human subjects and provided the DNA samples. My contribution is to design and execution of the experiments and analysis of the data.

ABSTRACT

Preterm birth is a complex genetic disorder of birth timing regulation, of which components and their interactions are not well understood. Recently, one of such component, *FSH receptor* (*FSHR*), was discovered by the genetic association with preterm birth in Finnish mothers. However, it is not clear what role *FSHR* plays in determining the timing of birth in term and risk for preterm deliveries. As a first step to address this question, we fine-mapped the candidate causal variants underlying the genetic association of *FSHR* by sequencing a total of 44-kb regions, including protein-coding and conserved non-coding sequences, in 127 preterm and 135 term Finnish mothers. Overall, we identified 288 single nucleotide variants and 65 insertion/deletions of 1-2 bp across all subjects; however, no common SNP in the protein-coding region is associated with preterm birth. To narrow the causal variants down in non-coding regions, we conducted fine-mapping and haplotype analysis and determined that both protective (P=0.02, OR=0.48, 95% CI = 0.23-0.94) as well as risk promoting (P=0.02, OR = 1.87, 95% CI = 1.08-3.29) haplotypes spanning intron 1 and 2 underlie the association of *FSHR* with preterm birth. In these haplotypes, two SNPs, rs12052281 and rs72822025, are predicted to cause risk of and protection from preterm birth by disrupting putative binding sites for ZEB1 and Elf3 transcription factors, respectively. These transcription factors have been implicated in regulatory function in human parturition, but previously not implicated with *FSHR* regulation. Although our claims need further functional validation, they provide a testable hypothesis on the mechanism of *FSHR* in regulation of birth timing in human.

INTRODUCTION

Preterm birth is a complex genetic disorder of which mechanisms and pathophysiology are little understood. However recent evidence has begun to accumulate that some forms of preterm birth may be an extreme phenotype of heritable genetic variation in the length of gestation [29]. Multiple twin studies reported that the length of gestation is partly genetic with the maternal heritability of 15-40% [22-25]. Prior history of post-term as well as preterm delivery is a strong predictor of the length of gestation of subsequent pregnancies [30]. Moreover, preterm birth clusters among siblings [142], across generations in kinship [143; 144] and even by races [32; 33].

Identifying the genes underlying the genetic risk for preterm birth has shed light on the components of the pathway determining the timing of human parturition [29; 145]. In particular, the genetic association of *FSH receptor* (*FSHR*) with preterm birth has been identified in Finnish and African Americans, although the details of mechanism by which *FSHR* predisposes women to preterm birth is not clear yet [29]. While the main function of *FSHR* in females is the regulation of ovarian function and follicular development, its expression in the uterus or cervix hints that it may have a potential function therein [146-148]. One possibility is that *FSHR* may regulate the transition of the myometrium from the quiescent to contractile state on the onset of labor. In support of this idea, the serum levels of its ligand, FSH, progressively increase toward the end of pregnancy [149], and FSH has been shown to modify the electrical signaling property of the myometrium *in vitro* [148].

In order to predict causal variants underlying the association of *FSHR* with preterm birth, we sequenced coding and conserved non-coding regions in Finnish preterm and full-term mothers. The Finnish population were chosen over the African Americans for a number of reasons, such as the higher signal of association with preterm birth [29], the genetic homogeneity due to a founder effect [150], and lower environmental risks for preterm birth. In the Finnish cohort, we find that neither a common nonsynonymous SNP nor the aggregate burden of rare variants is associated with preterm birth. By fine-mapping and haplotype analysis, the predicted causal non-coding variants are narrowed down to risk and protective haplotypes spanning intron 1 and 2. After computationally characterizing the effects of predicted transcription factor binding sites in these haplotypes, we propose that common SNPs disrupting putative binding sites for ZEB1 and Elf3 transcriptional factors may increase or decrease the risk for preterm birth by affecting the differential expression of *FSHR*.

RESULTS

Identification of genetic variants in coding and non-coding regions

To identify potentially causal variants underlying the genetic association of *FSHR* with preterm birth, 67 candidate regions were sequenced in 127 preterm and 135 term Finnish mothers (Materials and Methods), covering a total of 17 kb of sequence within candidate regions and an additional 27 kb of sequence flanking the candidate regions. The candidate regions include *FSHR* protein-coding regions as well as any non-coding regions likely to be functional. Non-coding regions were selected based on experimental evidence from the literature, sequence conservation across placental mammals using PhastCon [151], or rapid evolution along the human lineage. Rapidly evolved sequences are of particular interest since they may have influenced changes in the length of gestation during human evolution [29].

Using next-generation sequencing technology, we applied a pooled high-throughput sequencing protocol [152; 153] to sequence the target regions in equimolar pools of cases and controls. Although this protocol cannot assay individual genotypes, single nucleotide variants (SNVs) and 1 – 2 bp insertions/deletions (InDels) can be efficiently identified with the high sensitivity and specificity along with their allele frequencies. Across both cases and controls, we identified a total of 281 high-quality variants and an additional 72 lower-quality variants (Table 4.1). The lower quality of some variants can be attributed to either their low frequencies or low read coverage. Nearly all of low-quality variants (98.2%) are very rare, estimating to occur in one or two alleles in the pool, and four lower-quality variants were detected at 0.93% of sites that were sequenced in less than 30 reads per allele, the read coverage needed to saturate the power to detect variants in pooled samples [152]. However, not all rare variants are lower-quality; 49.2% of singleton and doubleton variants pass our high-quality thresholds.

The power and accuracy of variant identification were validated in three ways. First, we added plasmid controls to our library at a singleton allele frequency, and recovered all singletons with no false positive. Second, a large proportion of SNVs identified in our subjects are variants also known in the dbSNP database (release 135) or the 1000 Genomes Project [93]; the proportions of known SNVs are substantially higher among both of the high- and lower-quality SNVs (93.0% and 38.1%, respectively) than among putative sequencing errors, which could not satisfy our variant detection criteria (1.0%). Third, the estimated allele frequencies were cross-checked using 24 SNPs that were previously genotyped with Affymetrix microarrays in partly overlapping human subjects [29]. The frequencies estimated by pooled sequencing agree well with those obtained by genotyping (Pearson's correlation coefficient = 0.99) (Figure 4.1).

Genetic variants altering protein sequence

Although we did not find any frame-shift, splice-site, or nonsense variants in *FSHR*, we identified four nonynonymous variants (Table 4.2). All of them were previously reported in dbSNP and have high variant quality scores in our data, and thus are unlikely to be false positives. Among them, two alleles, p.S680N and p.A307T, reach high enough allele frequencies to be tested for the direct allelic association with preterm birth. In particiular, p.S680N is potentially functional, since it was previously found to be associated with the the length and dynamics of menstrual cycle [34]. However, neither p.S680N or p.A307T show significant association with preterm birth (Methods, two-proportion Z-test, $P = 1.0$ and 0.86, respectively).

Although the sample size was too small to detect the direct association of rare variants, we observe two potentially functional nonsynonymous rare variants, p.A189V and p.R162K in our samples. p.A189V was detected as a singleton unique to the case pool, and based on an allele frequency of 0.4%, it is likely to be carried heterozygously in a single individual. The p.A189V disrupts an evolutionarily conserved amino acid ($P < 10^{-8}$ by the likelihood ratio test [15] and probably damaging by PolyPhen-2 [154]) and is a well-characterized loss-of-function mutation previously reported to cause ovarian failure in homozygotes [155].

Another variant p.R162K is over-represented two-fold in controls (2.6%) relative to cases (1.3%). Although the difference is not significant, the sequence conservation suggests that p.R162K may be potentially deleterious. Although PolyPhen-2 predicts it as benign, the mutated residue is marginally conserved evolutionarily according to the likelihood ratio test ($P=0.0017$), and despite the biochemical similarity between arginine and lysine, lysine was not observed at orthologous positions of any other 18 placental mammals aligned at this site.

No enrichment of rare variants in cases

Although an individual rare variant can explain only a small fraction of the risk for common genetic disorder, rare variants can still make a substantial aggregate contribution to the risk [4; 5]. To test if rare variants in *FSHR* are implicated to the risk of preterm birth, we compared the distribution of rare SNVs, defined by minor allele frequency less than 1%, between cases and controls (Table 4.3).

In the protein-coding regions, only one nonsynonymous rare SNV, p.A189V, is of high variant quality and unique to case individuals. In candidate non-coding regions, three additional high-quality rare SNVs were found at conserved nucleotides, defined by PhyloP ($P < 0.05$) [156], which are likely to be functional. In total, combining conserved non-coding and nonsynonymous sites within candidate regions, an equal number of rare high-quality SNVs are uniquely observed in cases and controls ($N=2$). The lack of enrichment of rare variants in cases cannot be attributed to the lower sensitivity of variant detection or the slightly smaller sample size for the case cohort compared to the control. When the sensitivity for variant detection is maximized at the expense of specificity by lowering the cut-offs of variant quality scores, only two additional lower-quality rare SNVs can be identified, and all of which are unique to controls. In addition, when the sequences flanking candidate regions, which are likely to deplete of functional sites, are used to control for the differences of the sensitivity and sample size, the number of rare SNVs at conserved non-coding or nonsynonymous sites in candidate regions is not significantly different between cases and controls (Fisher's Exact Test, P = 1.00 and 0.67 for the variants of high-quality and of both high- and lower-quality, respectively).

Even when the rare variant class is extended to those of minor allele frequency up to 5%, we cannot find any significant enrichment of rare SNVs in cases compared to controls (Supplementary table 3). At a minor allele frequency between 1% and 5%, there are three high-quality and two lower-quality SNVs at conserved non-coding sites of candidate regions. However, these variants are not significantly over-represented in cases compared to controls.

Fine-map the association of common variants

A previous study by Plunkett et al. identified three tag SNPs, rs11686474, rs11680730 and rs12473815, in linkage disequilibrium to be significantly associated with preterm birth with odds ratios of 1.76 – 1.82 in the Finnish cohort used in our study [29]. In our current study, we observe a slight decline in the odds ratios of the tag SNPs to 1.58 – 1.59. The decrease of odds ratios is caused by subtle differences of allele frequency estimates between the two studies, as this study only includes a subset of subjects from the earlier report. In our current study, the average minor allele frequency of the three tag SNPs, is estimated to be 40.1% in cases and 29.7% in controls, whereas the previous study reports the frequency of 41.8% in cases and 29.0% in controls.

To fine-map the association of *FSHR* with preterm birth, 169 common variants in sequenced regions were tested for the allelic association with preterm birth (Figure 4.2). The common variants are defined by a minor allele frequency > 5%, and include 17 InDels as well as 152 SNPs. The candidate regions encompass 39% (17 kb) of sequenced regions and harbored 30.8% (N=52) of common variants, and the flanking sequences contain the rest of common variants. If a potentially causal variant exists, it will associate with preterm birth at least as strongly as the associated tag SNPs. Of 169 tested variants, 11 SNPs had association stronger than the three associated tag SNPs (two proportion Z-test, $P <= 0.05$). All 11 associated SNPs are non-coding and localize in 103 kb region spanning intron 1 and 2. The SNP with the highest association (rs12052281) is located in a conserved non-coding element in the intron 2 (two-proportion Z-test, $P=0.026$).

To explore the linkage disequilibrium structure within the fine-mapped interval bounded by the 11 associated SNPs, we examined the haplotypes of 93 normal Finns (FIN) [93], which were computationally phased using MaCH/Thunder [157] by the 1000 Genomes Project Consortium (Figure 4.3, Methods). Within a total of 9 kb sequenced region inside the 103 kb fine-mapped interval, 39 SNPs are known to segregate at a minor allele frequency over 5% and clustered in four major haplotypes in the 1000 Genomes Project data (FIN). These four haplotypes constitute 77% of chromosomes in FIN, and the rest are either rare haplotypes below a 5% frequency or could not be directly tagged by a known allele within the sequenced regions. All of the four haplotypes are in linkage disequilibrium with the SNPs that were originally discovered to be associated with preterm birth by Plunkett et al. (D' = 1.0) [29] (r^2 =0.28, 0.06, 0.32 and 0.38, respectively).

To estimate the frequencies of these haplotypes in cases and controls, we utilized 21 SNPs tagging one of the four haplotypes with $r^2 > 0.9$ in FIN (Figure 4.4, Methods). Based on the genetic homogeneity of the Finnish population, we assume that the linkage disequilibrium is consistent across FIN and our Finnish preterm and term cohorts. Haplotype 1 shows a significant risk promoting effect (Fisher's exact test, P = 0.0074, OR = 2.03, 95% CI = 1.17 – 3.53), and haplotype 2 shows a significant protective effect (P = 0.010, OR = 0.42, 95% CI = 0.21 – 0.85). The frequencies of haplotype 3 and 4 are not significantly different between cases and controls. To test if the effects of haplotype 1 and 2 are independent from each other, we re-examined the association of haplotype 1 with preterm birth after excluding haplotype 2 from the gene pool, and vice versa. The odds ratios for both haplotypes (OR = 1.87 and 0.48, respectively) remain similar after this correction and are significantly different from 1.0 (Fisher's exact test, P = 0.020 and 0.023, respectively). The tag SNPs previously used by Plunkett et al. (rs11686474-rs11680730-rs12473815) [29] captures both haplotype 1 and 2 simultaneously, with one allele tagging the risk promoting haplotype and the other tagging the protective haplotype.

Candidate causal variants in risk and protective haplotypes

Assuming the simple genetic model of co-dominance and no epistasis, a potentially causal risk variant should be exclusively carried by the risk promoting haplotype and not by other haplotypes for which we found protective or no effect. This criterion may not hold under more complex scenarios such that the casual risk variant is completely recessive, or that the risk phenotype is rescued by *cis*-epistatic interactions in the other haplotype background. In the risk promoting haplotype, six variants satisfying this criterion are segregating in the 9-kb sequenced regions within the fine-mapped interval (Figure 4.3) and thus can be examined for the potential direct causality. The p-values of association are not highly informative for further narrowing down of the causal variants within the risk promoting haplotype, since all of variants within the haplotype are in strong linkage disequilibrium and their allele frequency estimates are inexact due to pooling and sampling variation. Thus, we instead screened the 9-kb sequenced regions in the fine-mapped interval for evolutionarily conserved sites disrupting a putative transcription factor binding site by utilizing databases of transcription factor binding motifs, TRANSFAC [158], JASPAR [159] and UniProbe [160]. Out of the six candidate variants, we identify two in conserved non-coding elements defined by PhastCon, and only one out of the two, rs12052281, is found at conserved nucleotides defined by PhyloP ($P < 0.05$).

The risk allele of rs12052281 (G) is a derived allele and predicted to impact the binding for ZEB1 transcriptional repressor by 12.5% (Methods) [161]. A previous study identified ZEB1 as a key suppressor of the genes involved in uterine contraction in both humans and mice [162]. FSH, the ligand recognized by FSHR, has been known to modulate the electrical signaling property of the uterine muscle [148], thus the weaker binding for ZEB1 of the rs12052281 G allele may increase the risk for preterm birth through premature de-repression of *FSHR*. The up-regulation of *FSHR* may work in conjunction with rising serum FSH levels toward the end of pregnancy [149].

Similarly, the protective haplotype contains seven variants, which are not shared with haplotypes of risk or no effect (Figure 4.3). Although none of those variants is located within sequence elements or at nucleotides conserved across placental mammals, rs72827283 and rs72822025 are within the sequence element identified to be conserved across primates by PhastCon. In particular, the protective rs72822025 A allele is at an Elf3 binding motif characterized by protein-binding micorarray and predicted to reduce the binding energy by 27.1% [163]. In mice, Elf3 is known to be up-regulated during late pregnancy and activate prostaglandin synthesis pathway in the uterus through transcriptional activation of COX-1 [164].

DISCUSSION

As a necessary first step to follow up on the previously reported genetic association of *FSHR* with preterm birth, we conducted a sequencing-based fine-mapping study to narrow down the candidate causal variants underlying this genetic disorder in the Finnish population. We rule out the possibility of causal variants in the protein-coding region, and map the candidate causal variants to risk promoting and protective haplotypes spanning intron 1 and 2. To further narrow down candidate causal variants, we scanned for putative binding sites for transcription factors known to be differentially regulated toward the end of pregnancy and computationally characterized the expected effects of binding site mutations. Based on these results, we predict that the timing of parturition might be modulated by transcriptional regulation of *FSHR* by ZEB1 and Elf3 transcription factors in human.

In this study, we do not observe any evidence that preterm birth cases are enriched with rare variants in *FSHR,* although in the literature some evidence suggests otherwise. In *FSHR*, there are a number of well studied rare mutations causing the reduced fertility or ovarian hyperstimulation syndrome [165], both of which are known risk factors of preterm birth [166; 167]. Although the rare deleterious variant p.189V that is unique to our preterm subjects and known to reduce fertility may fit this category, the total contribution of rare variants to preterm birth seems to be very small based on the observation that in nonsynonymous or conserved non-coding sites the three case-specific high-quality rare SNVs constitute merely 3.6% of combined allele frequency in the case pool (Supplementary table 3 and 4). The three case-specific rare SNVs include a non-coding variant with MAF of 0.8% located at chr2:49,202,765 in a putative binding site for peroxisome proliferator-activated receptor alpha (PPARA), which has been implicated with preterm birth complicated with infection [168] or alternatively in a putative binding site for vitamin D receptor (VDR), which is also known to be differentially expressed in preterm placental tissue [169]. Another rare variant with MAF of 2.4% is located at chr2:49,202,863 in a putative binding site for SOX transcription factors, which was observed to be differentially regulated in the uterus of mice and to a lesser degree in humans [164].

To search for causal variants efficiently, we focused on protein-coding regions and conserved non-coding elements. Conserved intronic and intergenic regions in *FSHR* are known to be enriched with *cis*-regulatory elements [170; 171]. On the other hand, non-conserved regions, although potentially interesting, are often repetitive sequences making specific PCR amplification difficult and labor-intensive. ~53% of non-conserved genic regions are classified as repeats by RepeatMasker [172], whereas only ~13% of sequenced regions are repetitive. Moreover, it is often difficult to justify the functional predictions made at non-conserved non-coding sequences in the absence of direct experimental evidence. Nonetheless, we cannot rule out the possibility that the causal variants might be outside of our sequenced regions, which include only conserved regions and their flanking sequences.

We predicted that causal common SNPs might disrupt putative binding sites for ZEB1 and Elf3 transcription factors. The most critical piece of information supporting this prediction came from the observation that the effects of mutations agree with the direction of regulatory changes in FSHR transcription expected by risk and protective alleles. Given that the serum FSH levels rise toward the end of pregnancy [149], we hypothesized that up-regulation of FSHR predisposes to preterm birth and down-regulation of FSHR protects from preterm birth. That being said, however, it is not clear yet whether FSHR expression indeed goes up toward the end of pregnancy and particularly in preterm labor. In the mouse myometrium, there seems to be no significant difference in FSHR mRNA levels throughout pregnancy [162]. However, the physiology of pregnancy and parturition in primates and particularly in humans is highly diverged from that of other mammals [173]. Moreover, the putative ZEB1 binding site in *FSHR* seems non-functional in mice, because the predicted binding energy of mice sequence is only ~34% of the human wild-type. To resolve this conflict, the function and transcriptional regulation of *FSHR* will need to be studied directly in humans.

Can the predicted causal variants explain the signal of association observed in other populations [29]? Although Plunkett et al. did not genotype any SNP corresponding to the protective haplotype in the US populations, they included rs3788982, which tags our candidate risk variant rs12052281 with r^2 of 0.93 and 0.79 in CEU and YRI, respectively (SNAP, HapMap release 22)[174]. However, in the European American population, the risk promoting haplotype shows a slightly protective effect (OR=0.79), although it was not significant (P = 0.09). The lack of significant effect of the risk promoting haplotype in this population might be due to epistatic interactions between genes or gene-by-environment interaction unique to the European American population. Or, perhaps, the risk variant may seem as if to exert a relatively stronger effect in a population with a higher frequency of protective variants such as in the Finnish population. The Finnish population has a much higher frequency of the segregating protective variant rs72822025 (16.4%) than CEU (4.4%) [93]. Unlike US whites, in US blacks the risk promoting haplotype showed an expected risk effect (OR=1.43, P=0.17), although it is not significant. However, rs11686474, which has a weak r^2 of 0.43 with the candidate risk variant rs12052281, shows a much stronger effect (OR=1.73, P=0.004) than rs3788982 that is expected to directly tag the causal risk variant. It is possible that rs12052281 might be tagging a third yet unknown causal variant segregating in US blacks because the protective rs72822025 is absent in Sub-Saharan Africans (YRI) and very rare in African Americans (~2% in ASW from the 1000 Genomes Project) [93].

What could have affected the frequency of the common variant predisposing to preterm birth? Interestingly, iHS detected a moderate signature of recent partial selective sweep in *FSHR* in YRI (top 5%) [100]. At the center of this sweep signal is the nonsynonymous SNP p.S680N, and the haplotype containing the derived allele S^{680} was positively selected (iHS score = -1.96). Since S^{680}/S^{680} genotype is associated with a two-day longer menstrual cycle in females but not with the fertility itself or ovarian aging, it has been proposed that a slightly lower lifetime reproductive output may be beneficial under certain environmental conditions that pregnancy itself is unfavorable [34]. Nonetheless, the selective sweep of S^{680} does not seem to have directly increased the frequency of adjacent preterm risk haplotype in Africans by the hitchhiking effect. In Africans, S^{680} is present in the haplotype background different from the 79-kb away preterm risk promoting rs12052281 G allele (Figure S4.3). The lack of linkage disequilibrium between S^{680} and the rs12052281 G allele ($r^2 = 0.02$) may be due to incorrect prediction of causal variants, the decay of ancient sweep event by recombination, indirect effect of selective sweep, or epistasis between S^{680} and the rs12052281 G allele. Selective sweep decreases the effective population size at around p.S680N due to an increase in the variation of fitness, which in effect can diminish the strength of negative selection against preterm risk variant. The epistasis between S^{680} and the rs12052281 G allele may select against the allelic linkage of longer menstrual cycle and higher preterm birth risk (antagonistic epistasis), or alternatively, select for the allelic linkage of longer menstrual cycle and longer gestation (synergistic epistasis). Perhaps, when harsh environmental conditions make a higher chance of pregnancy unfavorable, the combination of genetic and environmental risk factors for preterm birth might be much worse for the fitness of newborns than having fewer but full-term off-spring. Similarly, when resources are plentiful enough to allow for pregnancy at advanced maternal age, the gene-by-environment intersection between the rs12052281 G allele and the maternal age, which is a known risk factor of preterm birth by itself, will raise the risk for prematurity thus impact the survival of infants born to older mothers.

To sum up, this study provides clues on the role of *FSHR* in regulation of the onset of labor in term and preterm birth in human. Although our claims need to be subject to functional validation, it will guide future mechanistic and population genetic studies.

MATERIALS AND METHODS

Human subjects

The human subjects investigated in this study largely overlap with the Finnish mothers in which the association of *FSHR* was originally identified [29]. Out of 127 preterm and 135 term mothers investigated in this work, 96 cases and 70 controls were shared with the previous study, and the rest were unique to this study. The human subject study was approved by Institutional Review Boards and Ethics Committees at all participating institutions, and informed consent for the genetics research was obtained properly. The inclusion criteria for preterm mothers was the non-atrogenic singleton pregnancy with less than 37 completed weeks of gestation without a sign of trauma, infection, or drug abuse. The term mothers who delivered at least two children spontaneously after 37 gestational weeks were recruited as controls. There was no difference in the average maternal age between cases (30.6 years) and controls (31.4 years) (Wilcoxon test, P=0.17). The sample genomic DNA was collected from peripheral bloods or saliva using standard methods.

Candidate regions

The following regions, a total of 17 kb, were selected as candidate regions to identify causal variants: all exons (NM_000145), 50-bp exon-intron junctions, core promoter region (-1 to -225 relative to the translational start site) [175], 3 SNPs found significantly associated with preterm birth in African Americans (rs11686474, rs11680730 and rs12473815) [29], 15 non-coding elements rapidly evolved along the human lineage [29] and conserved non-coding elements (Supplementary table 1). Conserved non-coding elements, a total of 8.6 kb (N=269), were identified within the transcribed region and 5 kb upstream and downstream, based on the sequence conservation across 32 placental mammals using PhastCon [151]. A PhastCon element was selected as a candidate if it is longer than 50 bp by itself or a part of a cluster of PhastCon elements which are separated by less than 200 bp and together span more than 50 bp. In addition to PhastCon regions, we also included conserved functional non-coding elements from the literature. One transcriptional silencer [171] and seven distal transcriptional regulatory elements [170] were previously identified in rat, and their sequences are well conserved to human. The genomic coordinates of the rat non-coding elements were transferred to the human genome with UCSC LiftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver).

The candidate regions were amplified in 67 PCR amplicons (44 kb). PCR primers were designed using Primer 3 (http://frodo.wi.mit.edu/primer3/input.htm) with the minimum amplicon size of 300 bp and other parameters previously described [176; 177]. To avoid allele-specific PCR failure, PCR primers were selected within polymorphism-free segments utilizing the 1000 Genomes Project pilot data [93].

Pooled high-throughput sequencing

Human subjects were assorted into four groups by case/control status and by whether an individual is shared with [29] or exclusive to this study. For each group, Illumina sequencing library was prepared following the pooled high-throughput sequencing protocol [152; 153]. Briefly, genomic DNA samples in each group were fluorescently quantified using SYBR Gold (Invitrogen) staining technique [153], and mixed to an equimolar pool. To average out stochastic noise, we prepared two technical replicates of pooled genomic samples and repeated all the following steps.

In each pool, the candidate regions were amplified by PCR with PfuUltra High-Fidelity DNA polymerase (Stratagene) in presence of 1M betaine (Sigma-Aldrich). In each PCR reaction, $0.3N$ ng of pooled genomic DNA was added as templates, where N is the number of pooled individuals. While the number of PCR cycles was fixed to 28, other PCR parameters were optimized for each amplicon (Supplementary table 2). All PCR products were purified on QIAquick spin columns (Qiagen), validated on agarose gel and then quantified again by SYBR Gold staining.

To control for the sensitivity and specificity of pooled sequencing, we prepared positive and negative controls as described previously [152]. The negative control was

1.9 kb region of M13mp18 plasmid (NEB), and the positive controls were 335-bp synthetic sequences derived from *TP53* (shared by F. L. M. Vallania). Seven positive control plasmids carrying a total of 13 known mutations were spiked into unmutated plasmid DNA at the lowest allele frequency of each pool (1/2*N*). Both positive and negative controls were PCR-amplified similarly as candidate regions.

Next, amplicons of candidate regions and controls were pooled in an equimolar ratio so that all regions were sequenced at an even read coverage. The pooled amplicons were randomly ligated into >10 kb concatemers and then sonicated into 100-500 bp fragments with Bioruptor XL (Diagenode) following [153]. From the sonicated fragments, Illumina sequencing library was generated using the Genomic DNA Sample Prep Kit (Illumina). Each library was tagged with an index unique to each subject group. To minimize run-specific variation in error rates, all four sequencing libraries were multiplexed and sequenced on a single Illumina HiSeq lane in a single-read 42-cycle mode.

Identification of single nucleotide variants

For each pool, 42-bp sequence reads were mapped, aligned, calibrated and then scanned for single nucleotide variants (SNVs) and 1-2 bp insertions/deletions (InDels) using SPLINTER (version 6t) [152]. Specifically, out of 35.8 – 40.4 million reads, 30.0 – 35.2 miilion (84 – 88%) were aligned to the reference sequence (hg18) allowing two or less edits per read in the alignment. Then, to calibrate sequencing error rates, a second-order SPLINTER error model was generated from the reads aligned to the negative control sequence. Overall, the sequencing error rates were higher for the last 19 sequencing cycles than for the initial 23 cycles and spike up also for 4 intermittent sequencing cycles. After masking out all such error-prone cycles, only high-quality basecalls (19 nt per read) were utilized to identify SNVs.

The power to detect SNVs goes up with an increasing read coverage but saturates above ~30 reads per site per allele [152]. We obtained 74.7 – 198.7 reads per site per allele on average for each pool, which is well above the saturation point. However, the read coverage can still be under 30 reads per site per allele at a subset of sequenced sites. In sequenced regions, only 2.1% of sites were covered by less than 30 reads in at least one of four pools. This could be due to random sampling variation or difficulty in alignment. These regions are located in two rapidly evolving elements, a rat distal *cis*-regulatory element (ECR6) and variable poly-dA region (11-17 bp) in core promoter.

The optimal cut-offs for SNV calls were determined by positive and negative controls spiked into sequencing library. In all pools, SNV quality scores of positive controls were well separated from those of negative controls, thus there exist a range of score cut-offs discriminating positive from negative controls with 100% accuracy. We defined SNV sets at high and lower variant quality thresholds to maximize the specificity and the sensitivity, respectively. For each pool, the high quality cut-off was defined by the lowest variant quality score of positive controls whereas the lower quality cut-off was defined by the highest variant quality score of negative controls. After excluding a tri-allelic variant, we obtained a total of 281 high-quality SNVs and 72 lower-quality SNVs across all pools. The allele frequency estimated by SPLINTER was rounded off to the nearest multiple of singleton allele frequency in each pool. As a quality control, the proportion of known SNVs was evaluated using the combined dataset of dbSNP (release 135) and the 1000 Genomes Project pilot [93].

Comparison of allele frequencies between cases and controls

The significant difference of allele frequency between cases and controls were tested for all common variants of minor allele frequency over 5% by using two-proportion Z-test:

$$z = \frac{p_A - p_B}{\sqrt{\sigma_s + s_e^{case} + s_e^{control}}}$$

$$p_0 = \frac{N_A p_A + N_B p_B}{N_A + N_B}$$

$$\sigma_s = p_0 \cdot (1 - p_0) \cdot (\frac{1}{N_A} + \frac{1}{N_B})$$

where p_A and p_B are the allele frequencies in cases and controls, respectively, N_A and N_B are the number of alleles in case and control pools, respectively, p_0 is the pooled allele frequency over p_A and p_B, σ_s is the variance component due to random sampling, and the statistic z follows the standard normal distribution. To account for the error of allele frequency estimation as an additional source of variation independently of random sampling, we inferred the error of allele frequency estimation s_e from the mean squared error of observed allele frequencies compared to the expected values by using 24 SNPs genotyped in Finnish cohorts with Affymetrix arrays [29]:

$$s_e = \frac{1}{24}\sum_{i=1}^{24}(f_{o,i} - f_{e,i})^2$$

where $f_{o,i}$ and $f_{e,i}$ are the observed and expected allele frequencies of SNP i, respectively. The s_e was estimated to be 0.00081 for cases and 0.00032 for controls.

Haplotype analysis

The haplotype structure of the Finnish population was obtained from the 1000 Genomes Project (release 20120316) [93], in which genotypes of 93 individuals (FIN) were phased across the genome using MaCH/Thunder [157]. Ancestral alleles inferred from three sister primate genomes were also obtained from the 1000 Genomes Project. Conserved sites were identified using PhyloP scores across placental mammals [156], which were downloaded from UCSC genome browser.

The haplotype frequencies were estimated in our case and control subjects by averaging the allele frequencies of multiple tag SNPs, which were selected by the tight linkage disequilibrium between the SNP and the haplotype. Specifically, we looked for the SNPs with r^2 greater than 0.9 in normal Finns from the 1000 Genomes Project (FIN). For the haplotype 1, excluding one outlier, nine tag SNPs were utilized to quantify the haplotype frequency. The frequencies of haplotype 2, 3 and 4 were estimated using six, three and three tag SNPs, respectively.

Computational prediction of transcription factor binding sites

We examined three databases of transcription factor (TF) binding profiles: TRANSFAC [158], JASPAR [159] and UniProbe [160]. TRANSFAC contains the most comprehensive collection of TF binding profiles whereas JASPAR has fewer but higher-quality profiles, and UniProbe encompasses mostly zinc finger TF binding motifs, which were derived from *in vitro* protein-binding microarray experiments. In TRANSFAC (version 10.2), the human reference sequences with ancestral alleles were scanned for binding sites of vertebrate TFs by using ECR browser (http://ecrbrowser.dcode.org/) [178]. In JASPAR, CORE Vertebrata collection of binding matrices was examined using their web server (http://jaspar.cgb.ki.se/) with the default parameters. In UniProbe, human and mouse TFs were scanned with the default setting (http://the_brain.bwh.harvard.edu/uniprobe/), but SNPs with more than 20 hits of predicted binding sites were filtered out for potential non-specificity. For JASPAR and UniProbe, the reference sequences were examined with both derived as well as ancestral alleles in order to explore the gain as well as loss of binding sites. The binding sites predicted with the equivalent score for both ancestral and derived alleles were excluded. The binding energy for ZEB1 and Elf3 was calculated for each allele using ConSite [179] with the default parameter values and published binding site profiles [161; 163]

FIGURES

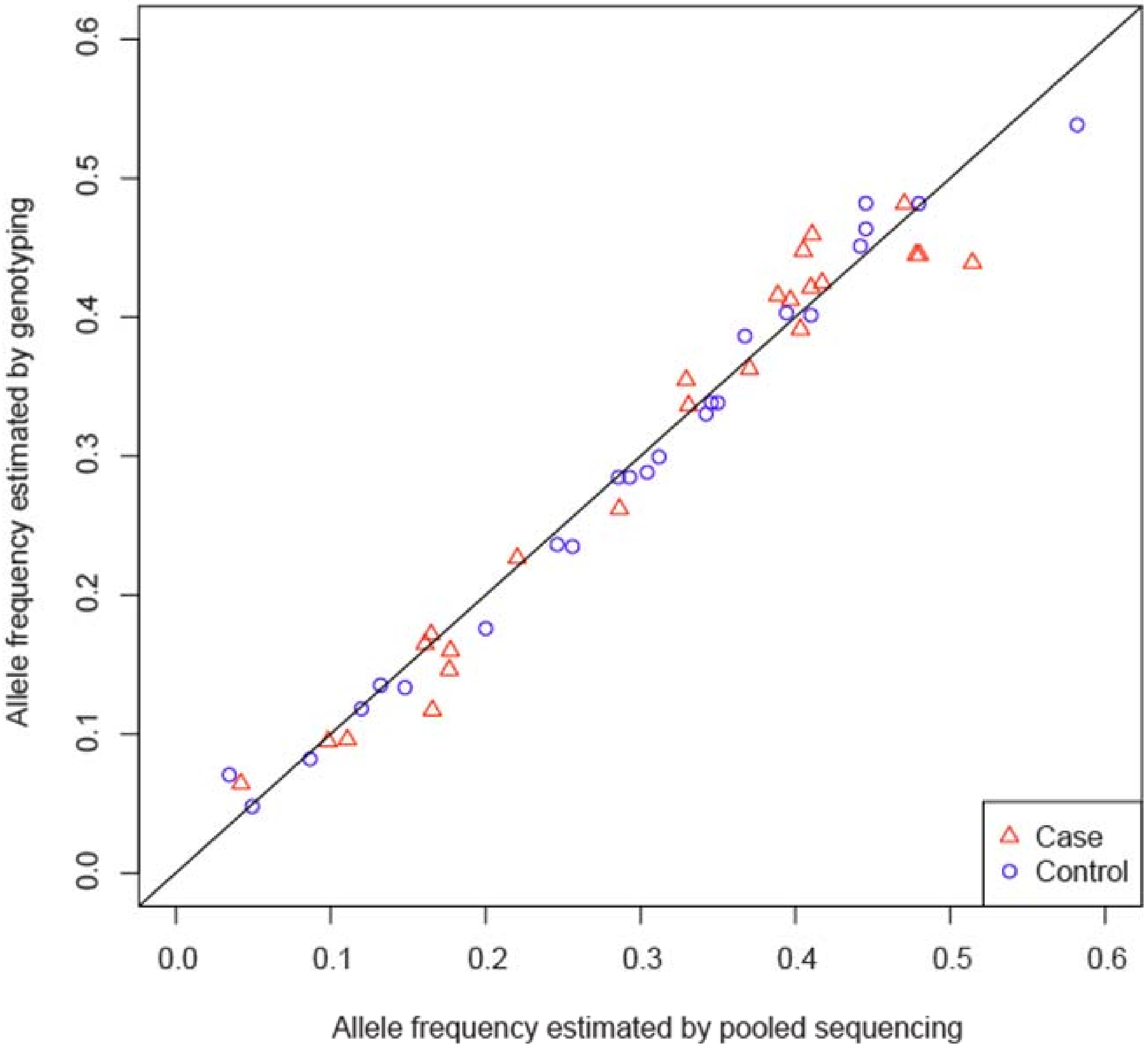


Figure 4.1. Comparison of allele frequencies between pooled sequencing and Affymetrix genotyping. The allele frequencies in case and control pools are indicated by red triangles and blue circles, respectively, for 24 SNPs shared between the two assays. The allele frequencies estimated by Affymetrix genotyping were obtained from [29].

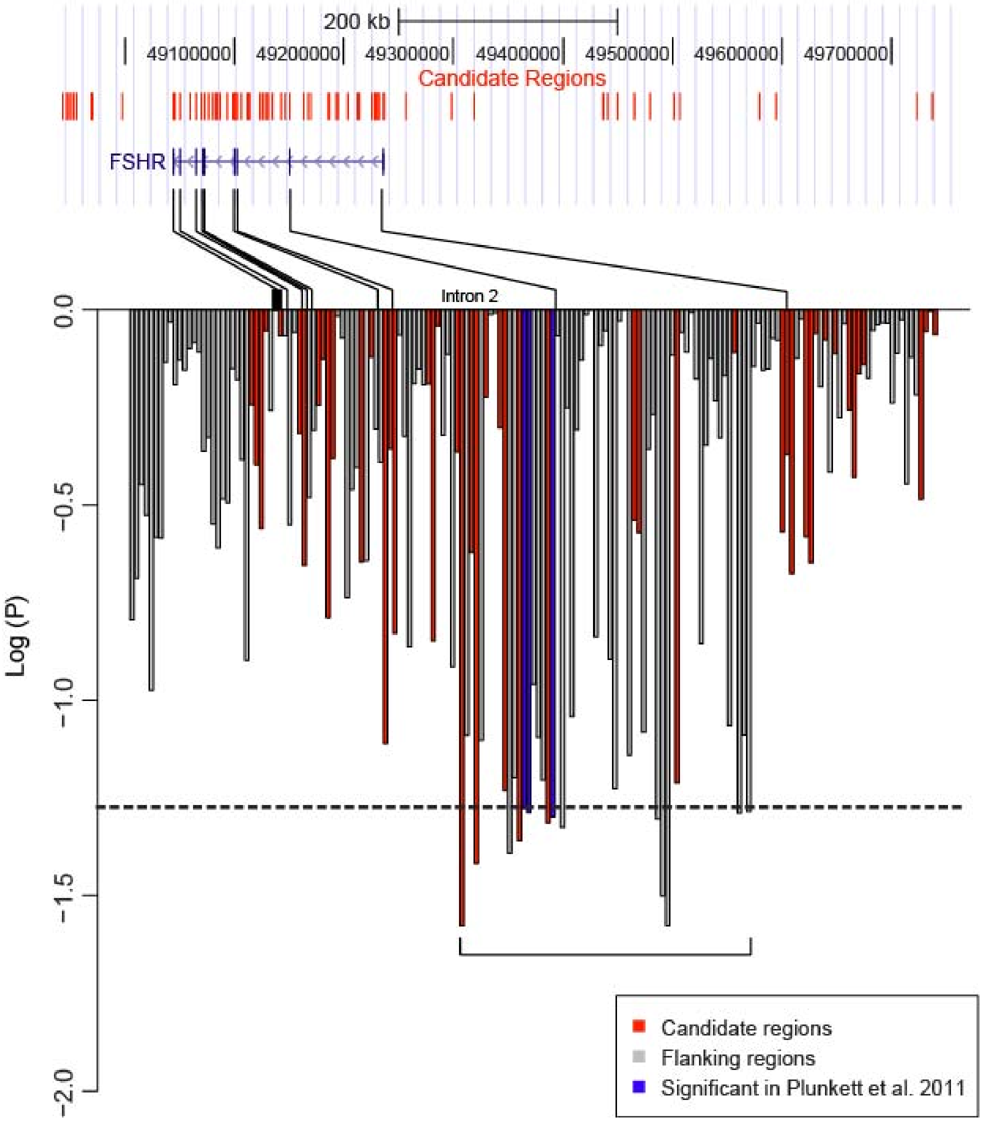


Figure 4.2. Fine-map of the association of common variants. Each of 169 variants with a minor allele frequency greater than 5% was tested for the allelic association with pre-term birth by two-proportion Z-test. Candidate regions and 52 SNPs therein are marked in red. SNPs in flanking regions are shown in gray. The dashed horizontal line indicates the least significant P value of the three tag SNPs (blue) found significant in [29].

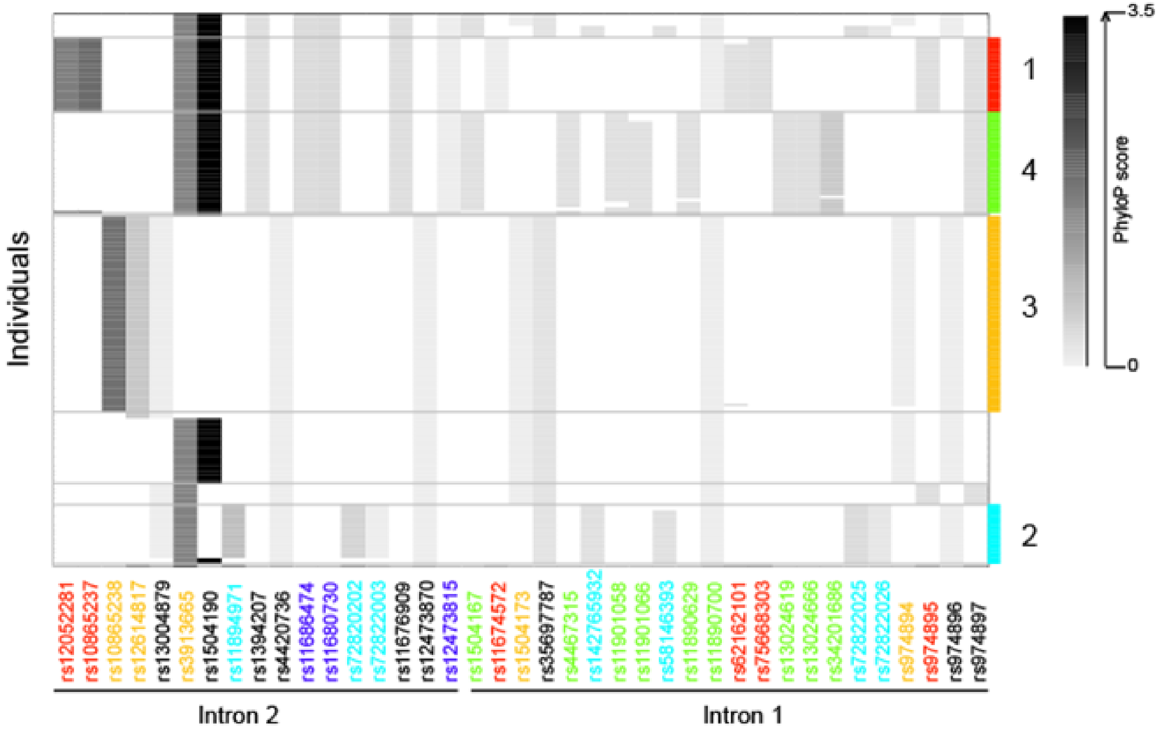


Figure 4.3. Haplotype structure of common SNPs within the fine-mapped interval in normal Finns. Each row corresponds to a phased chromosome in 93 normal Finnish individuals from the 1000 Genomes Project (FIN). Each column represents one of the 39 SNPs that are segregating in the sequenced regions and also in the fine-mapped interval spanning intron 1 and 2. Derived alleles at conserved sites are marked in a gray scale (black being the most highly conserved site by PhyloP). The ancestral alleles are shown in white. The four major haplotypes are numbered and color-tagged on the right. The SNPs carried exclusively by one of the four haplotypes are also highlighted in the color of the haplotype on the bottom. The three tag SNPs evaluated in [29] (rs11686474, rs11680730 and rs12473815) are marked in dark blue.

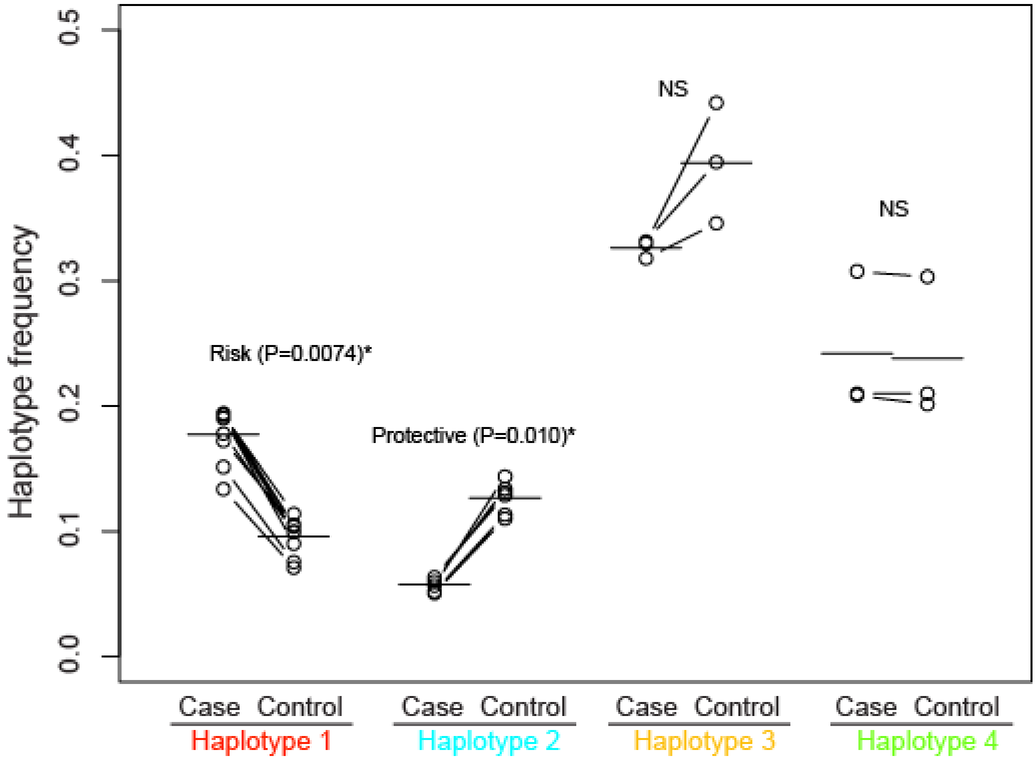


Figure 4.4. Haplotype frequencies in case and control pools. The haplotype frequencies were estimated from allele frequencies of tag SNPs with $r^2 > 0.9$ (N=9, 6, 3 and 3 for haplotype 1, 2, 3 and 4, respectively). The horizontal lines indicate the average allele frequencies among tag SNPs. The difference in allele frequency between cases and controls was tested by Fisher's Exact Test. NS denotes “not significant.”

TABLES

Table 4.1. Summary statistics of sequencing and identified genetic variants.

				High-quality variants[1]			Low-quality variants[1]		
Pool	Alleles (2N)	Reads	Coverage[2]	SNV	InDel	All [3]	SNV	InDel	All [3]
Case #1	192	34,940,358	74.7	198	42	240 (30)	29	16	45 (44)
Case #2	62	30,017,566	198.7	191	38	229 (60)	19	16	35 (34)
Control #1	150	32,513,156	89.0	202	41	243 (38)	30	13	43 (42)
Control #2	120	35,190,219	120.4	197	36	233 (33)	27	16	43 (43)
Total	524	132,661,299	103.9	236	45	281	52	20	72

[1] The number includes single nucleotide variants found in flanking regions as well as in candidate regions.

[2] Reads per site per allele.

[3] In parenthesis is the number of variants with the estimated allele count less than or equal to 2 in a pool.

Table 4.2. Nonsynonymous SNVs identified in subjects.

Variant	Variant quality	Amino acid change	Functional Prediction [1]	Minor allele frequency (%)		OR	P allelic [2]
				Case	Control		
rs6166	High	S680N	Neutral	48.0	48.0	1.00	1.00
rs6165	High	A307T	Neutral	50.2	51.2	0.96	0.86
rs121909658	High	A189V	Deleterious	0.4	0.0	n.a.	not tested
rs111883853	High	R162K	Marginal	1.3	2.6	0.49	not tested

[1] Functional predictions were made using a likelihood ratio test and PolyPhen-2. All predictions agreed between two methods except R162K, which was classified as marginally deleterious by the likelihood ratio test but as neutral by PolyPhen-2.

[2] P-values were calculated by two-proportion Z-test. Rare variants with minor allele frequency < 5% were not tested for the lack of statistical power.

Table 4.3. Rare SNVs observed in either cases or controls, exclusively.

Variant quality	Region [1]	Case	Control	P [2]
high	conserved or NSN	2	2	1.00
	flanking	13	11	
high + low	conserved or NSN	2	4	0.67
	flanking	23	21	

The rare variants are defined by minor allele frequency below 1% and exclusivity to either cases or controls.

[1] Rare variants were counted in “conserved or NSN (nonsynonymous)” sites within candidate regions or in “flanking” regions around the candidate regions. Conserved sites were defined by PhyloP ($P < 0.05$).

[2] The number of rare variants in cases vs. controls was compared between conserved or nonsynonymous sites in candidate regions and flanking regions by Fisher's Exact Test.

SUPPORTING INFORMATION

Table S4.1. Genomic regions selected for sequencing.

Region	Size (bp)	Feature
chr2:49127504-49127804	301	Conserved non-coding element [1]
chr2:49186258-49186604	347	Conserved non-coding element [1]
chr2:49105504-49105820	317	Conserved non-coding element [1]
chr2:49184226-49184925	700	Conserved non-coding element [1]
chr2:49101927-49102250	324	Conserved non-coding element [1]
chr2:48946672-48946999	328	SNP genotyped in Plunkett et al. 2011 [2]
chr2:49079672-49080023	352	Conserved non-coding element [1]
chr2:49297537-49298053	517	Rapidly evolving element [2]
chr2:49501372-49501725	354	Rapidly evolving element [2]
chr2:49255600-49255954	355	Rapidly evolving element [2]
chr2:48954581-48954938	358	SNP genotyped in Plunkett et al. 2011 [2]
chr2:49082112-49082470	359	Conserved non-coding element [1]
chr2:49083589-49083948	360	Conserved non-coding element [1]
chr2:48967579-48967943	365	SNP genotyped in Plunkett et al. 2011 [2]
chr2:49141174-49141662	489	SNP genotyped in Plunkett et al. 2011 [2,+]
chr2:48945687-48946057	371	SNP genotyped in Plunkett et al. 2011 [2]
chr2:49578471-49578843	373	Rapidly evolving element [2]
chr2:49098850-49099224	375	Conserved non-coding element [1], SNP genotyped in Plunkett et al. 2011 [2]
chr2:49440600-49440977	378	Rapidly evolving element [2]
chr2:49449628-49450010	383	Rapidly evolving element [2]
chr2:49227165-49227552	388	Conserved non-coding element [1]
chr2:49224931-49225319	389	Conserved non-coding element [1]
chr2:49145539-49145930	392	SNP genotyped in Plunkett et al. 2011 [2,+]
chr2:48951377-48951769	393	SNP genotyped in Plunkett et al. 2011 [2]
chr2:49071132-49071524	393	Exon
chr2:49085824-49086217	394	Conserved non-coding element [1]
chr2:49049211-49049606	396	Exon
chr2:48996812-48997208	397	Rapidly evolving element [2]
chr2:48941767-48942166	400	Rat distal regulatory element (ECR3) [3]
chr2:49100625-49101024	400	Exon
chr2:49112246-49112645	400	Conserved non-coding element [1]
chr2:49479368-49479796	429	Rapidly evolving element [2]
chr2:49464454-49464900	447	Rapidly evolving element [2]
chr2:49436827-49437300	474	Rapidly evolving element [2]
chr2:49148749-49149228	480	Exon
chr2:49736964-49737445	482	Rat distal regulatory element (ECR7) [3]
chr2:49074978-49075497	520	Conserved non-coding element [1]
chr2:49722580-49723159	580	Rat distal regulatory element (ECR6) [3]

chr2:49318574-49319159	586	Rapidly evolving element [2], rat distal regulatory element (ECR1) [3]
chr2:49212715-49213312	598	Conserved non-coding element [1]
chr2:49124571-49125229	659	Conserved non-coding element [1]
chr2:49202385-49203054	670	Conserved non-coding element [1]
chr2:49130312-49131010	699	Conserved non-coding element [1]
chr2:48968471-48969170	700	Rat distal regulatory element (ECR2) [3]
chr2:49113392-49114091	700	Conserved non-coding element [1]
chr2:49193762-49194461	700	Conserved non-coding element [1]
chr2:49133560-49134292	733	Conserved non-coding element [1]
chr2:48949522-48950275	754	SNP genotyped in Plunkett et al. 2011 [2]
chr2:49191694-49192463	770	Conserved non-coding element [1]
chr2:49097566-49098553	988	Exon, conserved non-coding element [1]
chr2:49121579-49122363	785	Conserved non-coding element [1]
chr2:49230140-49230939	800	Rat silencer element [4], conserved non-coding element [1]
chr2:49594515-49595314	800	Rat distal regulatory element (ECR5) [3]
chr2:49069125-49070027	903	Exon, conserved non-coding element [1]
chr2:49166133-49167001	869	Conserved non-coding element [1]
chr2:49092586-49093471	886	Conserved non-coding element [1]
chr2:49058506-49059405	900	Conserved non-coding element [1]
chr2:49506892-49507814	923	Rapidly evolving element [2], rat distal regulatory element (ECR4) [3]
chr2:49211052-49212049	998	Rapidly evolving element [2], conserved non-coding element [1]
chr2:49129080-49130079	1,000	Conserved non-coding element [1]
chr2:49169123-49170221	1,099	Rapidly evolving element [2], conserved non-coding element [1]
chr2:49161645-49163214	1,570	Conserved non-coding element [1]
chr2:49063460-49064650	1,191	Exon, conserved non-coding element [1]
chr2:49234734-49236033	1,300	Exon, 5' UTR, core promoter [5], conserved non-coding element [1]
chr2:49109868-49111753	1,886	Conserved non-coding element [1]
chr2:49227532-49229341	1,810	Rapidly evolving element [2], conserved non-coding element [1]
chr2:49042669-49044948	2,280	Exon, 3' UTR, conserved non-coding element [1]
Total	44,017	

Coordinates are relative to the NCBI reference genome build 36.

[1] Non-coding element conserved across placental mammals defined by PhastCon.

[2] Plunkett et al. (2011) PLoS Genetics.

[3] Distal regulatory element identified in rat (Hermann et al. (2007) Molecular and Cellular Endocrinology).

[4] Transcriptional silencer element (DHS3) identified in rat (Hermann and Heckert (2005) Molecular Endocrinology).

[5] The region from -225 to -1 relative to the translational start site (Gromoll et al. (1994) Molecular and Cellular Endocrinology).

[+] SNPs found significant in US replicate populations in Plunkett et al. (2011) PLoS Genetics.

Table S4.2. PCR parameters used to amplify candidate regions.

			28 cycles				
Region	Size (bp)	Denaturation	Denaturation	Annealing	Extension	Extension	Note
chr2:49127504-49127804	301	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49186258-49186604	347	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49105504-49105820	317	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49184226-49184925	700	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49101927-49102250	324	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:48946672-48946999	328	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49079672-49080023	352	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49297537-49298053	517	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49501372-49501725	354	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49255600-49255954	355	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:48954581-48954938	358	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49082112-49082470	359	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49083589-49083948	360	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:48967579-48967943	365	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49141174-49141662	489	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:48945687-48946057	371	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49578471-49578843	373	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49098850-49099224	375	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49440600-49440977	378	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49449628-49450010	383	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	1
chr2:49227165-49227552	388	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49224931-49225319	389	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49145539-49145930	392	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:48951377-48951769	393	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49071132-49071524	393	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49085824-49086217	394	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49049211-49049606	396	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:48996812-48997208	397	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	1
chr2:48941767-48942166	400	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49100625-49101024	400	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49112246-49112645	400	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49479368-49479796	429	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49464454-49464900	447	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49436827-49437300	474	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49148749-49149228	480	94C x 2 min	94C x 30 sec	64C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49736964-49737445	482	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49074978-49075497	520	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49722580-49723159	580	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49318574-49319159	586	94C x 2 min	94C x 30 sec	64C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49212715-49213312	598	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49124571-49125229	659	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49202385-49203054	670	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49130312-49131010	699	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:48968471-48969170	700	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49113392-49114091	700	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49193762-49194461	700	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49133560-49134292	733	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:48949522-48950275	754	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49191694-49192463	770	94C x 2 min	94C x 30 sec	64C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49097566-49098553	988	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49121579-49122363	785	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49230140-49230939	800	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49594515-49595314	800	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49069125-49070027	903	94C x 2 min	94C x 30 sec	64C x 30 sec	68C x 90 sec	65C x 10 min	
chr2:49166133-49167001	869	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49092586-49093471	886	94C x 2 min	94C x 30 sec	64C x 30 sec	68C x 3 min	65C x 10 min	
chr2:49058506-49059405	900	94C x 2 min	94C x 30 sec	64C x 30 sec	68C x 3 min	65C x 10 min	

chr2:49506892-49507814	923	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min
chr2:49211052-49212049	998	94C x 2 min	94C x 30 sec	58C x 30 sec	68C x 3 min	65C x 10 min
chr2:49129080-49130079	1,000	94C x 2 min	94C x 30 sec	60C x 30 sec	70C x 5 min	65C x 10 min
chr2:49169123-49170221	1,099	94C x 2 min	94C x 30 sec	60C x 30 sec	70C x 5 min	65C x 10 min
chr2:49161645-49163214	1,570	94C x 2 min	94C x 30 sec	60C x 30 sec	70C x 5 min	65C x 10 min
chr2:49063460-49064650	1,191	94C x 2 min	94C x 30 sec	60C x 30 sec	70C x 5 min	65C x 10 min
chr2:49234734-49236033	1,300	94C x 2 min	94C x 30 sec	64C x 30 sec	68C x 3 min	65C x 10 min
chr2:49109868-49111753	1,886	94C x 2 min	94C x 30 sec	60C x 30 sec	70C x 5 min	65C x 10 min
chr2:49227532-49229341	1,810	94C x 2 min	94C x 30 sec	62C x 30 sec	65C x 7 min	65C x 10 min
chr2:49042669-49044948	2,280	94C x 2 min	94C x 30 sec	62C x 30 sec	65C x 7 min	65C x 10 min

[1] To remove non-specific bands, PCR products of correct size were extracted from agarose gel.

Table S4.3. Rare SNVs observed in either cases or controls, exclusively, defined by minor allele frequency < 5%.

Variant quality	Region [1]	Case	Control	P [2]
high	conserved or NSN	3	4	1.00
	flanking	15	16	
high + low	conserved or NSN	2	6	0.28
	flanking	23	25	

The rare variants are defined by minor allele frequency below 5% and exclusivity to either cases or controls.

[1] Rare variants were counted in “conserved or NSN (nonsynonymous)” sites within candidate regions or in “flanking” regions around the candidate regions. Conserved sites were defined by PhyloP ($P < 0.05$).

[2] The number of rare variants in cases vs. controls was compared between conserved or nonsynonymous sites in candidate regions and flanking regions by Fisher's Exact Test.

Table S4.4. Rare SNVs exclusively found in cases and located at conserved or nonsynonymous sites.

Genomic coordinate [1]	Major allele	Minor allele	Case AF	Region	Predicted Function
chr2:49,063,768	G	A	0.4%	Protein-coding	Deleterious nonsynonymous p.A189V
chr2:49,202,765	G	A	0.8%	Conserved non-coding	Putative binding site of PPARA and VDR
chr2:49,202,863	G	C	2.4%	Conserved non-coding	Putative binding site of SOX

Conserved sites were defined by PhyloP ($P < 0.05$).

[1] Coordinates are relative to the NCBI reference genome build 36.

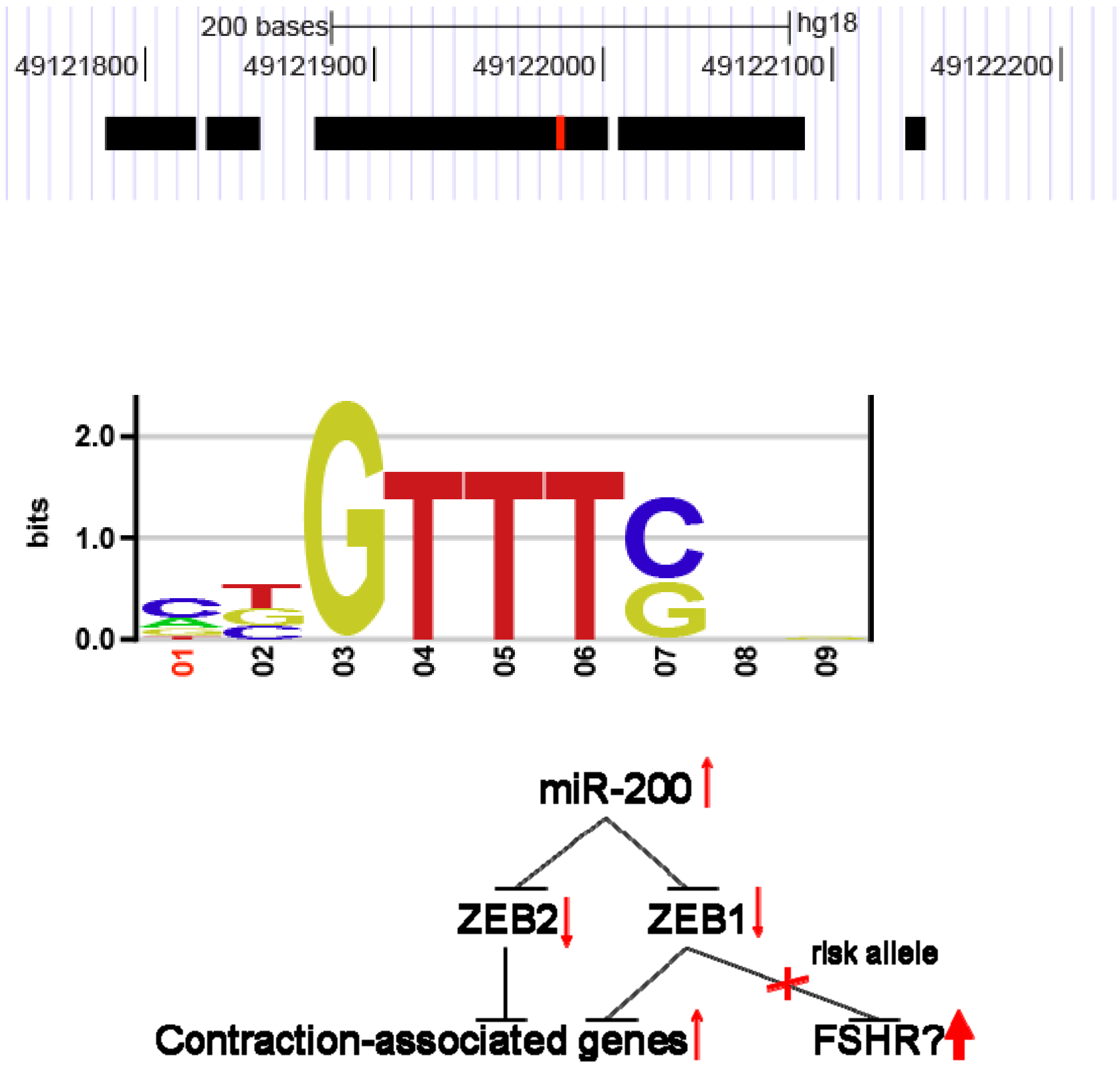


Figure S4.1. Sequence context of candidate causal variant, rs12052281. Conserved non-coding elements around the SNP are indicated by black bars, and the location of candidate causal variant is marked in red. In the center, the DNA-binding site profile of ZEB1 transcription factor is represented as a sequence logo. The derived allele at rs12052281 mutates C at the position 01 to G. The genetic model on the bottom depicts the predicted direction of transcriptional change toward the end of pregnancy.

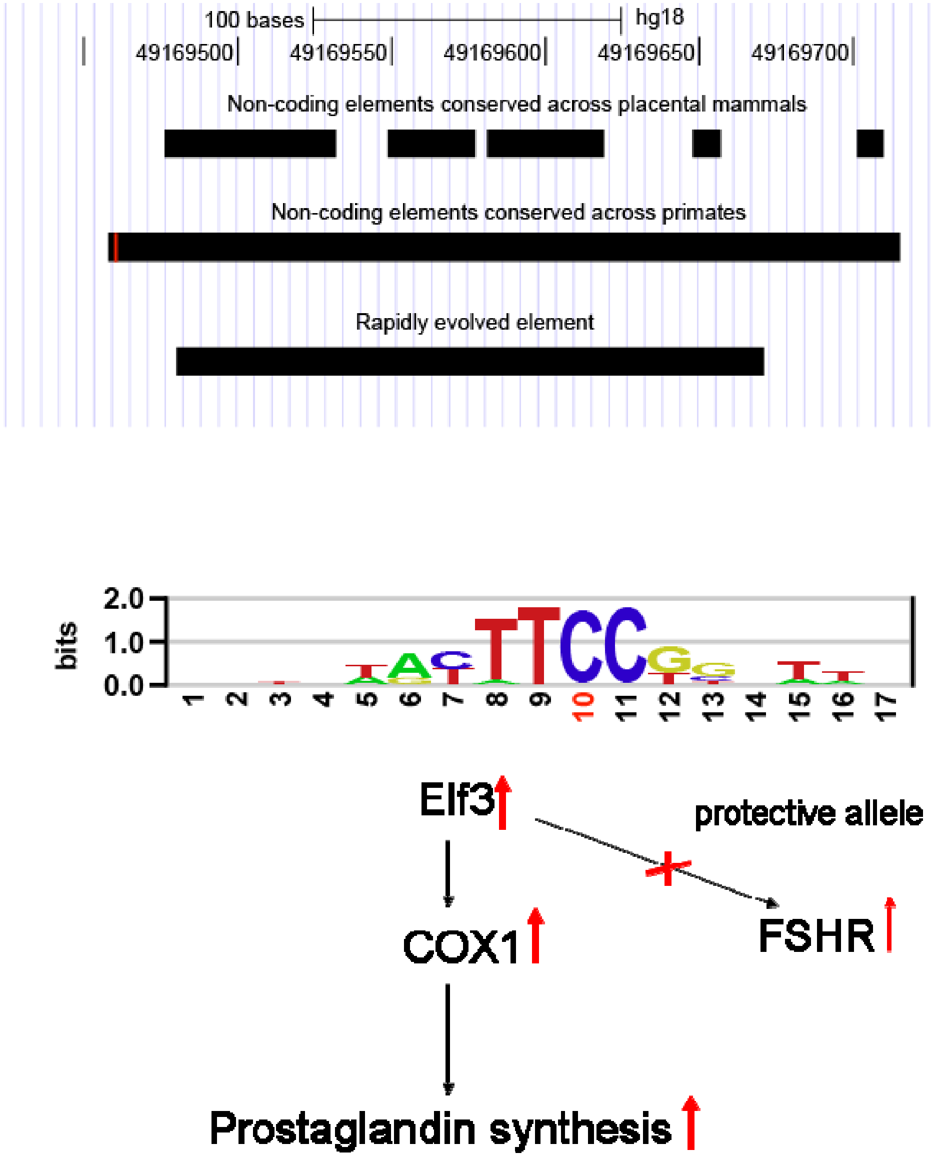


Figure S4.2. Sequence context of candidate causal variant, rs72822025. Conserved and rapidly evolving non-coding elements around the SNP are indicated by black bars, and the location of candidate causal variant is marked in red. In the center, the DNA-binding site profile of Elf3 transcription factor is represented as a sequence logo. The derived allele at rs72822025 mutates C at the position 10 to T. The genetic model on the bottom depicts the predicted direction of transcriptional change toward the end of pregnancy.

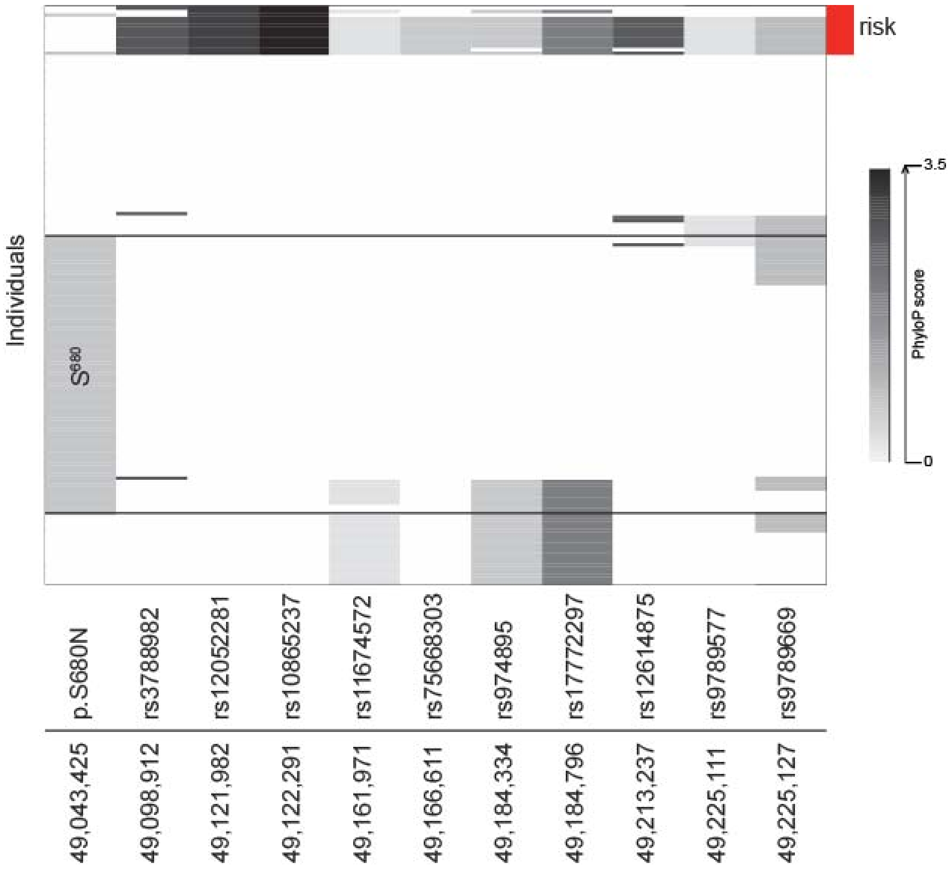


Figure S4.3. Linkage disequilibrium between risk haplotype and positively selected p.S680N in Sub-Saharan Africans (YRI). Each row corresponds to a phased chromosome in 83 normal individuals from the 1000 Genomes Project (YRI). Each column represents one of the 10 SNPs included in the risk haplotype of preterm birth or the positively selected p.S680N. Derived alleles at conserved sites are marked in a gray scale (black being the most highly conserved site by PhyloP). The ancestral alleles are shown in white. The risk haplotype is marked in red on the right. The positively selected allele at p.S680N is the derived allele S^{680}. The rs12052281 is the candidate causal risk variant disrupting a putative ZEB1 binding site.

CHAPTER 5: Progress and future directions

In this thesis, new comparative genomics software called a likelihood ratio test (LRT) was developed to predict deleterious nonsynonymous variants within the human genome. The LRT was a useful analytical framework to systematically investigate the selective constraints in protein-coding regions. Based on a long-studied generative probabilistic model with explicitly defined modeling assumptions, the LRT directly accounts for the protein-sequence evolution similarly to heuristic methods such as SIFT and PolyPhen, but at the same time, it can provide the interpretive simplicity of nucleotide-based phylogenetic tests mainly used for non-coding regions. Taking advantage of the LRT, this thesis provided systematic accounts on the functional constraints imposed in protein-coding regions from the method development as well as evolutionary perspectives.

In terms of future method development, an important but often neglected issue is the frequent disagreement of predictions made by distinct algorithms. For instance, when the LRT and two pre-existing methods, SIFT and PolyPhen-2 [154], were applied to the same personal genome, out of all deleterious variants predicted in the person, less than ~50% could be identified by an individual method, and only 5-7% could be identified by all three methods (chapter 2). One practical solution to this issue is to take the consensus among multiple methods, as recently approached by CONDEL algorithm [180]. However, the consensus approach does not circumvent the necessity of good understanding on the features of hard-to-predict cases. Rather, it is exactly the opposite; the merit of consensus methods is limited by how well the consensus priors can be fine-tuned depending on queries. In this respect, addressing why distinct algorithms give distinct predictions, as studied in chapter 2, can be a good starting point to improve the consensus priors. For example, for methods that does not distinguish orthologs from paralogs (e.g. SIFT), it may be difficult to identify functional constraints in a large gene family that acquired numerous functionally diverged paralogs through multiple rounds of gene duplication. Also, it will be interesting to examine which of the protein sequence similarity and the taxonomic relationship of species is a better proxy for the conservation of functional constraints on protein sequences. Many methods, such as SIFT and PolyPhen, identify homologous proteins by protein sequence similarity, but the rate of protein sequence evolution can vary widely by proteins (see discussion section of chapter 2 for more details).

The other areas of particular interests in terms of method improvement are the lineage-specific loss of constraints and the inference of functionally equivalent amino acid changes from closely related species. First, while comparative genomics methods can accommodate the loss of functional constraints in non-ancestral species albeit with reduced statistical power, they often find it difficult to handle the loss of constraints occurred along the ancestral lineage leading to the human. An example of this is the olfactory receptors, of which selective constraints were relaxed after the divergence between human and chimpanzee [181]. The majority of polymorphisms found in these genes are selectively neutral in humans even if they seem to disrupt sequences strongly conserved across mammals. Contrary to olfactory receptors, however, it is often challenging to identify ancient or subtle loss of constraints such as the losses shared across primates, in a subset of genes in a pathway or in only a part of gene. Second, while evidence suggests that distant orthologs often mislead the inference of functionally equivalent changes of amino acids due to functional divergence (see [182] for MTHFR complementation assay), many existing algorithms put higher weights on distantly related sequences than on closely related ones. Although this approach is optimal for identifying sequence conservation, functionally equivalent amino acids need to be inferred with the opposite weighting scheme, namely that closely related species should receive a higher weight than distant one. To this end, the LRT separates the test for significant sequence conservation from the subsequent step to filter out amino acid changes observed in any of 23 placental mammals as being functionally equivalent. Although only ~10% of variants at significantly conserved sites were affected by this filter, it is important to note the current strategy may be too aggressive for a number of reasons: 1) it does not account for frequent sequencing errors found in draft genomes, 2) distantly related mammals get the same weight as the closely related primates, 3) the depth of aligned mammalian sequences differs by genes and by sites, and 4) as more diverse mammalian genomes become sequenced in near future, more codons are expected to have experienced functional divergence in at least one species. The necessity and opportunity to address the above mentioned challenges will continue to rise with the advances of genome sequencing technology.

The LRT was also conducive to the evolutionary analysis of why human population has an abundant number of common SNPs that disrupt strong sequence conservation thus highly likely to be deleterious (chapter 2). Why so many human SNPs could escape the strong negative selection maintained sequence conservation in other vertebrates? Interestingly, only a small fraction of these common deleterious SNPs (~10%) can be explained by false positives due to multiple hypothesis testing, violation of model assumptions, recent gene duplication, or relaxation of selective constraints on specific category of genes. One of the violated model assumptions worth mentioning is the uniformity of synonymous substitution rate across the genome. In chapter 2, in order to examine the effect of relaxing this assumption, the genome-wide synonymous rate was reduced by two standard deviations as expected for less than 2.5% of genome. This reduction made LRT to predict 22% fewer deleterious SNPs, which could be enriched with false positives caused by the assumption of uniform synonymous rate. Interestingly, the deleterious SNPs corresponding to the 22% were only slightly enriched with common deleterious SNPs (not significant, odds ratios = 1.06 and 1.16 for common deleterious SNPs of derived allele frequency > 5% and > 30%, respectively). This enrichment can only explain 7-12 out of 1,121 common deleterious SNPs, defined by derived allele frequency > 5%, within three personal genomes examined in chapter 2. The largest source of probable false positives was the olfactory receptors (N=80), which is explained by the relaxation of selective constraints. 80 out of 85 deleterious SNPs predicted in olfactory receptors have allele frequency > 5%. However, genes with common deleterious SNPs were not enriched in any other functional categories.

What can explain the rest of common deleterious SNPs? One possibility is the genome-wide relaxation of negative selection due to the reduction of effective population size during the human evolution. The long-term effective population size of human-chimpanzee common ancestor is estimated to be 5-9 times higher than that of the human

[183]. Our empirical data also support this possibility. The substitutions of deleterious changes are twice as frequent along the human lineage (deleterious-to-neutral nonsynonymous ratio, DEL/NEU = 0.14) as they are along the human-chimpanzee common ancestor lineage (DEL/NEU = 0.07). Based on that the DEL/NEU ratio among common SNPs, defined by derived allele frequency > 5%, is ~0.24 for both CEU and YRI [93], ~30% common deleterious SNPs are expected to have fitness impact small enough to allow them to be fixed along the human lineage. It could be potentially interesting to compare the patterns of deleterious substitutions along the lineages of small effective population sizes, such as orangutan and domesticated species [184], with those along the other lineages. That being said, however, the small fitness consequence of some common deleterious SNPs does not necessarily imply that they are unlikely to be interesting in terms of phenotype. Many genetic changes of medical importance were fixed along the human lineage [120].

Another model that can explain common deleterious SNPs is the hitchhiking effect (chapter 3). Because of genetic linkage, the apparent fitness of a deleterious allele is determined not only by its own fitness but also by the combined fitness of all linked alleles. As expected by the hitchhiking effect, the genome-wide distribution of positively selected alleles and deleterious polymorphisms suggests that positive selection on beneficial allele can mitigate the fitness deficit of linked deleterious alleles and drag them along to higher-than-expected allele frequencies in humans. However, the hitchhiking effect does not seem to increase the overall abundance of deleterious polymorphisms since it also eliminates many deleterious variants that are not linked with beneficial alleles.

In this study, the hitchhiking effect was indirectly analyzed by using the close physical distance within the genome as a proxy for the genetic linkage. Ideally, direct examination of whether deleterious alleles are carried in positively selected haplotypes or not can establish stronger support for the hitchhiking hypothesis, and furthermore distinguish the primary effect of hitchhiking from the secondary effect caused by a reduced effective population size due to positive selection. Although currently available data have limited confidence on long-range haplotype phases especially for low-frequency variants, in near future, trio sequencing will provide more reliable phase information for both rare and common variants so that the hitchhiking effect can be directly evaluated in a haplotype level. In the meanwhile, further work will have to rely on population simulation.

In chapter 4, the known genetic association of *FSH receptor* (*FSHR*) with preterm birth was fine-mapped in a non-coding region in the Finnish population, and candidate causal variants were predicted therein by computationally scanning for transcription factor (TF) binding sites. While sequence conservation was useful in identifying functional non-coding elements, most valuable clues to pin-point candidate causal variants came from the prior knowledge of TFs differentially regulated in presumable target tissue toward the end of pregnancy, DNA-binding profiles of these TFs and which direction of expression change of these TFs and FSHR would increase/decrease the risk for preterm birth. By combining all these information, causal variants were predicted to be common SNPs disrupting putative ZEB1 and Elf3 binding sites in *FSHR*, although functional studies will be needed to establish this claim.

Further studies may find it very interesting to sequence African Americans and patients received assisted reproductive technology due to subfertility. The candidate causal variants predicted in this work in the Finnish population do not agree well with the signal of association observed previously in African Americans [29]. In addition, because the patients with reduced fertility have a higher risk for preterm birth [167; 185], these individuals may reveal the rare variants association, which could not be detected in the Finnish preterm mothers conceived normally. A previous sequencing study on body mass suggested that the subjects with extreme phenotypes are enriched rare causal variants, thus can provide higher statistical power to detect rare variant association [6; 186].

The recent advances of sequencing technology have unlocked the ability to identify the selective constraints in an increasingly high resolution. As repeatedly shown by multiple studies, each human carries a large number of deleterious variants in an individual genome, and at the same time, a large fraction of human genome evolved under the influence of recent positive selection. This will provide the unprecedented challenges and opportunities to explore the dependence between natural selection acting at nearby sites and to inquire why certain disease alleles have reached higher-than-expected frequencies in human population.

References

1 Morton NE, Crow JF, Muller HJ: **AN ESTIMATE OF THE MUTATIONAL DAMAGE IN MAN FROM DATA ON CONSANGUINEOUS MARRIAGES**. *Proc Natl Acad Sci U S A* 1956, **42**:855-863.

2 Shull GH: **What Is "Heterosis"?**. *Genetics* 1948, **33**:439-446.

3 Miller MP, Kumar S: **Understanding human disease mutations through the use of interspecific genetic variation**. *Hum Mol Genet* 2001, **10**:2319-2328.

4 Kryukov GV, Pennacchio LA, Sunyaev SR: **Most rare missense alleles are deleterious in humans: implications for complex disease and association studies**. *Am J Hum Genet* 2007, **80**:727-739.

5 Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: **Multiple rare alleles contribute to low plasma levels of HDL cholesterol**. *Science (80-)* 2004, **305**:869-872.

6 Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hebert S, Doelle H, Ersoy B, Kryukov G, Schmidt S, Yosef N, Ruppin E, Sharan R, Vaisse C, Sunyaev S, Dent R, Cohen J, McPherson R, Pennacchio LA: **Medical sequencing at the extremes of human body mass**. *Am J Hum Genet* 2007, **80**:779-791.

7 Ohta T: **The Nearly Neutral Theory of Molecular Evolution**. *Annu Rev Ecol Syst* 1992, **23**:263-286.

8 Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions**. *Genome Res* 2001, **11**:863-874.

9 Sunyaev S, Ramensky V, Koch I, Lathe W3, Kondrashov AS, Bork P: **Prediction of deleterious human alleles**. *Hum Mol Genet* 2001, **10**:591-597.

10 Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S: **Analysis of sequence conservation at nucleotide resolution**. *PLoS Comput Biol* 2007, **3**:e254.

11 Cooper GM, Stone EA, Asimenos G, , Green ED, Batzoglou S, Sidow A: **Distribution and intensity of constraint in mammalian genomic sequence**. *Genome Res* 2005, **15**:901-913.

12 Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang S, Fay JC: **A catalog of neutral and deleterious polymorphism in yeast**. *PLoS Genet* 2008, **4**:e1000183.

13 Muse SV, Gaut BS: **A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome**. *Mol Biol Evol* 1994, **11**:715-724.

14 Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L: **Natural selection has driven population differentiation in modern humans**. *Nat Genet* 2008, **40**:340-345.

15 Chun S, Fay JC: **Identification of deleterious mutations within three human genomes**. *Genome Res* 2009, **19**:1553-1561.

16 Ng PC, Henikoff S: **Accounting for human polymorphisms predicted to affect protein function**. *Genome Res* 2002, **12**:436-446.

17 Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R,

Clark AG, Bustamante CD: **Assessing the evolutionary impact of amino acid mutations in the human genome**. *PLoS Genet* 2008, **4**:e1000083.

18 Muglia LJ, Katz M: **The enigma of spontaneous preterm birth**. *N Engl J Med* 2010, **362**:529-535.

19 Martin JA, Hamilton BE, Sutton PD, Ventura SJ, Mathews TJ, Osterman MJK: **Births: final data for 2008**. *Natl Vital Stat Rep* 2010, **59**:1, 3-71.

20 Smith V, Devane D, Begley CM, Clarke M, Higgins S: **A systematic review and quality assessment of systematic reviews of randomised trials of interventions for preventing and treating preterm birth**. *Eur J Obstet Gynecol Reprod Biol* 2009, **142**:3-11.

21 Committee on Understanding Premature Birth and Assuring Healthy Outcomes BOHSP: *Preterm birth: causes, consequences, and prevention*. Washington, DC: National Academies; 2006.

22 York TP, Strauss JF3, Neale MC, Eaves LJ: **Estimating fetal and maternal genetic contributions to premature birth from multiparous pregnancy histories of twins using MCMC and maximum-likelihood approaches**. *Twin Res Hum Genet* 2009, **12**:333-342.

23 Treloar SA, Macones GA, Mitchell LE, Martin NG: **Genetic influences on premature parturition in an Australian twin sample**. *Twin Res* 2000, **3**:80-82.

24 Kistka ZA, DeFranco EA, Ligthart L, Willemsen G, Plunkett J, Muglia LJ, Boomsma DI: **Heritability of parturition timing: an extended twin design analysis**. *Am J Obstet Gynecol* 2008, **199**:43.e1-5.

25 Clausson B, Lichtenstein P, Cnattingius S: **Genetic influence on birthweight and gestational length determined by studies in offspring of twins**. *BJOG* 2000, **107**:375-381.

26 Moore T, Haig D: **Genomic imprinting in mammalian development: a parental tug-of-war**. *Trends Genet* 1991, **7**:45-49.

27 Watts DH, Krohn MA, Hillier SL, Eschenbach DA: **The association of occult amniotic fluid infection with gestational age and neonatal outcome among women in preterm labor**. *Obstet Gynecol* 1992, **79**:351-357.

28 Onderdonk AB, Hecht JL, McElrath TF, Delaney ML, Allred EN, Leviton A: **Colonization of second-trimester placenta parenchyma**. *Am J Obstet Gynecol* 2008, **199**:52.e1-52.e10.

29 Plunkett J, Doniger S, Orabona G, Morgan T, Haataja R, Hallman M, Puttonen H, Menon R, Kuczynski E, Norwitz E, Snegovskikh V, Palotie A, Peltonen L, Fellman V, DeFranco EA, Chaudhari BP, McGregor TL, McElroy JJ, Oetjens MT, Teramo K, Borecki I, Fay J, Muglia L: **An evolutionary genomic approach to identify genes involved in human birth timing**. *PLoS Genet* 2011, **7**:e1001365.

30 Kistka ZA, Palomar L, Boslaugh SE, DeBaun MR, DeFranco EA, Muglia LJ: **Risk for postterm delivery after previous postterm delivery**. *Am J Obstet Gynecol* 2007, **196**:241.e1-6.

31 Varki A, Altheide TK: **Comparing the human and chimpanzee genomes: searching for needles in a haystack**. *Genome Res* 2005, **15**:1746-1758.

32 Adams MM, Elam-Evans LD, Wilson HG, Gilbertz DA: **Rates of and factors associated with recurrence of preterm delivery**. *JAMA* 2000, **283**:1591-1596.

33 Kistka ZA, Palomar L, Lee KA, Boslaugh SE, Wangler MF, Cole FS, DeBaun MR, Muglia LJ: **Racial disparity in the frequency of recurrence of preterm birth**. *Am J Obstet Gynecol* 2007, **196**:131.e1-6.

34 Greb RR, Grieshaber K, Gromoll J, Sonntag B, Nieschlag E, Kiesel L, Simoni M: **A common single nucleotide polymorphism in exon 10 of the human follicle stimulating hormone receptor is a major determinant of length and hormonal dynamics of the menstrual cycle**. *J Clin Endocrinol Metab* 2005, **90**:4866-4872.

35 Lynch M, Burger R, Butcher D, Gabriel W: **The mutational meltdown in asexual populations**. *J Hered* 1993, **84**:339-344.

36 MULLER HJ: **Our load of mutations**. *Am J Hum Genet* 1950, **2**:111-176.

37 Simmons MJ, Crow JF: **Mutations affecting fitness in Drosophila populations**. *Annu Rev Genet* 1977, **11**:49-78.

38 Grantham R: **Amino acid difference formula to help explain protein evolution**. *Science (80-)* 1974, **185**:862-864.

39 Ohta T: **Slightly deleterious mutant substitutions in evolution**. *Nature* 1973, **246**:96-98.

40 Fay JC, Wyckoff GJ, Wu CI: **Positive and negative selection on the human genome**. *Genetics* 2001, **158**:1227-1234.

41 Wang Z, Moult J: **SNPs, protein structure, and disease**. *Hum Mutat* 2001, **17**:263-270.

42 Chasman D, Adams RM: **Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation**. *J Mol Biol* 2001, **307**:683-706.

43 Ng PC, Henikoff S: **Predicting the effects of amino acid substitutions on protein function**. *Annu Rev Genomics Hum Genet* 2006, **7**:61-80.

44 Kondrashov AS, Sunyaev S, Kondrashov FA: **Dobzhansky-Muller incompatibilities in protein evolution**. *Proc Natl Acad Sci U S A* 2002, **99**:14878-14883.

45 Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D, Altman RB, Davis RW, Nislow C, Giaever G: **The chemical genomic portrait of yeast: uncovering a phenotype for all genes**. *Science (80-)* 2008, **320**:362-365.

46 Care MA, Needham CJ, Bulpitt AJ, Westhead DR: **Deleterious SNP prediction: be mindful of your training data!**. *Bioinformatics* 2007, **23**:664-672.

47 Eddy SR: **A model of the statistical power of comparative genome sequence analysis**. *PLoS Biol* 2005, **3**:e10.

48 Stone EA, Cooper GM, Sidow A: **Trade-offs in detecting evolutionarily constrained sequence by comparative genomics**. *Annu Rev Genomics Hum Genet* 2005, **6**:143-164.

49 Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function**. *Nucleic Acids Res* 2003, **31**:3812-3814.

50 Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter JC: **Genetic variation in an individual human exome**. *PLoS Genet* 2008, **4**:e1000160.

51 Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey**. *Nucleic Acids Res* 2002, **30**:3894-3900.

52 Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song X, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing**. *Nature* 2008, **452**:872-876.

53 Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J: **The diploid genome sequence of an Asian individual**. *Nature* 2008, **456**:60-65.

54 Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AWC, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC,

Remington KA, Abril JF, Gill J, Borman J, Rogers Y, Frazier ME, Scherer SW, Strausberg RL, Venter JC: **The diploid genome sequence of an individual human**. *PLoS Biol* 2007, **5**:e254.

55 Fay JC, Wyckoff GJ, Wu C: **Testing the neutral theory of molecular evolution with genomic data from Drosophila**. *Nature* 2002, **415**:1024-1026.

56 Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization**. *Genetics* 2000, **154**:459-473.

57 Keightley PD, Eyre-Walker A: **Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies**. *Genetics* 2007, **177**:2251-2261.

58 Schmutz J, Wheeler J, Grimwood J, Dickson M, Yang J, Caoile C, Bajorek E, Black S, Chan YM, Denys M, Escobar J, Flowers D, Fotopulos D, Garcia C, Gomez M, Gonzales E, Haydu L, Lopez F, Ramirez L, Retterer J, Rodriguez A, Rogers S, Salazar A, Tsai M, Myers RM: **Quality assessment of the human genome sequence**. *Nature* 2004, **429**:365-368.

59 Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB: **Quality scores and SNP detection in sequencing-by-synthesis systems**. *Genome Res* 2008, **18**:763-770.

60 : **Initial sequence of the chimpanzee genome and comparison with the human genome**. *Nature* 2005, **437**:69-87.

61 Gaffney DJ, Keightley PD: **The scale of mutational variation in the murid genome**. *Genome Res* 2005, **15**:1086-1094.

62 Stone EA, Sidow A: **Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity**. *Genome Res* 2005, **15**:978-986.

63 Dagan T, Talmor Y, Graur D: **Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection**. *Mol Biol Evol* 2002, **19**:1022-1025.

64 Bromberg Y, Rost B: **SNAP: predict effect of non-synonymous polymorphisms on function**. *Nucleic Acids Res* 2007, **35**:3823-3835.

65 Hodgkinson A, Ladoukakis E, Eyre-Walker A: **Cryptic variation in the human mutation rate**. *PLoS Biol* 2009, **7**:e1000027.

66 Crow JF, Kimura M: **Efficiency of truncation selection**. *Proc Natl Acad Sci U S A* 1979, **76**:396-399.

67 Pritchard JK: **Are rare variants responsible for susceptibility to complex diseases?**. *Am J Hum Genet* 2001, **69**:124-137.

68 Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E: **EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates**. *Genome Res* 2009, **19**:327-335.

69 Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**:1792-1797.

70 Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA**. *J Mol Evol* 1985, **22**:160-174.

71 Pond SLK, Frost SDW, Muse SV: **HyPhy: hypothesis testing using phylogenies**. *Bioinformatics* 2005, **21**:676-679.

72 Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A: **Distribution and intensity of constraint in mammalian genomic sequence**. *Genome Res* 2005, **15**:901-913.

73 Gabriel W, Lynch M, Burger R: **Muller's ratchet and mutational meltdowns**. *Evolution Int J Org Evolution* 1993, **47**:1744-1757.

74 Lynch M, Gabriel W: **Mutational load and the survival of small populations**. *Evolution Int J Org Evolution* 1990, **44**:1725-1737.

75 Lande R: **Risk of population extinction from fixation of new deleterious mutations**. *Evolution Int J Org Evolution* 1994, **48**:1460-1469.

76 Felsenstein J: **The evolutionary advantage of recombination**. *Genetics* 1974, **78**:737-756.

77 Charlesworth B: **Mutation-selection balance and the evolutionary advantage of sex and recombination**. *Genet Res* 1990, **55**:199-221.

78 Jordan DM, Ramensky VE, Sunyaev SR: **Human allelic variation: perspective from protein function, structure, and evolution**. *Curr Opin Struct Biol* 2010, **20**:342-350.

79 Fay JC, Wu C: **Sequence divergence, functional constraint, and selection in protein evolution**. *Annu Rev Genomics Hum Genet* 2003, **4**:213-235.

80 Fay JC, Wu CI: **The neutral theory in the genomic era**. *Curr Opin Genet Dev* 2001, **11**:642-646.

81 Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, Clark AG, Bustamante CD: **Proportionally more deleterious genetic variation in European than in African populations**. *Nature* 2008, **451**:994-997.

82 NEEL JV: **Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"?**. *Am J Hum Genet* 1962, **14**:353-362.

83 Barton NH: **Linkage and the limits to natural selection**. *Genetics* 1995, **140**:821-841.

84 Hartfield M, Otto SP: **Recombination and hitchhiking of deleterious alleles**. *Evolution Int J Org Evolution* 2011, **65**:2421-2434.

85 Johnson T, Barton NH: **The effect of deleterious alleles on adaptation in asexual populations**. *Genetics* 2002, **162**:395-411.

86 Bachtrog D, Gordo I: **Adaptive evolution of asexual populations under Muller's ratchet**. *Evolution Int J Org Evolution* 2004, **58**:1403-1413.

87 Birky CWJ, Walsh JB: **Effects of linkage on rates of molecular evolution**. *Proc Natl Acad Sci U S A* 1988, **85**:6414-6418.

88 Rice WR: **Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome**. *Genetics* 1987, **116**:161-167.

89 Cruz F, Vila C, Webster MT: **The legacy of domestication: accumulation of deleterious mutations in the dog genome**. *Mol Biol Evol* 2008, **25**:2331-2336.

90 Lu J, Tang T, Tang H, Huang J, Shi S, Wu C: **The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication**. *Trends Genet* 2006, **22**:126-131.

91 Smith JM, Haigh J: **The hitch-hiking effect of a favourable gene**. *Genet Res* 1974, **23**:23-35.

92 Fay JC, Wu CI: **Hitchhiking under positive Darwinian selection**. *Genetics* 2000, **155**:1405-1413.

93 : **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**:1061-1073.

94 Stephan W, Charlesworth B, McVean G: **The effect of background selection at a single locus on weakly selected, partially linked variants**. *Genet Res* 1999, **73**:133-146.

95 Charlesworth B, Morgan MT, Charlesworth D: **The effect of deleterious mutations on neutral molecular variation**. *Genetics* 1993, **134**:1289-1303.

96 Nordborg M, Charlesworth B, Charlesworth D: **The effect of recombination on background selection**. *Genet Res* 1996, **67**:159-174.

97 Reed FA, Akey JM, Aquadro CF: **Fitting background-selection predictions to levels of nucleotide variation and divergence along the human autosomes**. *Genome Res* 2005, **15**:1211-1221.

98 Akey JM: **Constructing genomic maps of positive selection in humans: where do we go from here?**. *Genome Res* 2009, **19**:711-722.

99 Kim Y, Stephan W: **Detecting a local signature of genetic hitchhiking along a recombining chromosome**. *Genetics* 2002, **160**:765-777.

100 Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome**. *PLoS Biol* 2006, **4**:e72.

101 Tang K, Thornton KR, Stoneking M: **A new approach for using genome scans to detect recent positive selection in the human genome**. *PLoS Biol* 2007, **5**:e171.

102 Bierne N: **The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population**. *Evolution Int J Org Evolution* 2010, **64**:3254-3272.

103 Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, Pritchard JK: **Signals of recent positive selection in a worldwide sample of human populations**. *Genome Res* 2009, **19**:826-837.

104 Lewinsky RH, Jensen TGK, Moller J, Stensballe A, Olsen J, Troelsen JT: **T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro**. *Hum Mol Genet* 2005, **14**:3945-3953.

105 Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN: **Genetic signatures of strong recent positive selection at the lactase gene**. *Am J Hum Genet* 2004, **74**:1111-1120.

106 Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C,

Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT, Galver LM, Fan J, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok P, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L, Mak W, Song YQ, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PIW, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Johnson TA, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J: **Genome-wide detection and characterization of positive selection in human populations**. *Nature* 2007, **449**:913-918.

107 Grossman SR, Shlyakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O, Lander ES, Schaffner SF, Sabeti PC: **A composite of multiple signals distinguishes causal variants in regions of positive selection**. *Science (80-)* 2010, **327**:883-886.

108 Lamason RL, Mohideen MPK, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE, Sinha S, Moore JL, Jagadeeswaran P, Zhao W, Ning G, Makalowska I, McKeigue PM, O'donnell D, Kittles R, Parra EJ, Mangini NJ, Grunwald DJ, Shriver MD, Canfield VA, Cheng KC: **SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans**. *Science (80-)* 2005, **310**:1782-1786.

109 Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, Jakobsdottir M, Steinberg S,

Palsson S, Jonasson F, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K: **Genetic determinants of hair, eye and skin pigmentation in Europeans**. *Nat Genet* 2007, **39**:1443-1452.

110 Gudbjartsson DF, Sulem P, Stacey SN, Goldstein AM, Rafnar T, Sigurgeirsson B, Benediktsdottir KR, Thorisdottir K, Ragnarsson R, Sveinsdottir SG, Magnusson V, Lindblom A, Kostulas K, Botella-Estrada R, Soriano V, Juberias P, Grasa M, Saez B, Andres R, Scherer D, Rudnai P, Gurzau E, Koppova K, Kiemeney LA, Jakobsdottir M, Steinberg S, Helgason A, Gretarsdottir S, Tucker MA, Mayordomo JI, Nagore E, Kumar R, Hansson J, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K: **ASIP and TYR pigmentation variants associate with cutaneous melanoma and basal cell carcinoma**. *Nat Genet* 2008, **40**:886-891.

111 Hutton SM, Spritz RA: **A comprehensive genetic study of autosomal recessive ocular albinism in Caucasian patients**. *Invest Ophthalmol Vis Sci* 2008, **49**:868-872.

112 Schrijver I, Liu W, Brenn T, Furthmayr H, Francke U: **Cysteine substitutions in epidermal growth factor-like domains of fibrillin-1: distinct effects on biochemical and clinical phenotypes**. *Am J Hum Genet* 1999, **65**:1007-1020.

113 Nijbroek G, Sood S, McIntosh I, Francomano CA, Bull E, Pereira L, Ramirez F, Pyeritz RE, Dietz HC: **Fifteen novel FBN1 mutations causing Marfan syndrome detected by heteroduplex analysis of genomic amplicons**. *Am J Hum Genet* 1995, **57**:8-21.

114 Liu WO, Oefner PJ, Qian C, Odom RS, Francke U: **Denaturing HPLC-identified novel FBN1 mutations, polymorphisms, and sequence variants in Marfan syndrome and related connective tissue disorders**. *Genet Test* 1997-1998, **1**:237-242.

115 Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits**. *Proc Natl Acad Sci U S A* 2009, **106**:9362-9367.

116 Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database**. *Nat Genet* 2004, **36**:431-432.

117 Hermisson J, Pennings PS: **Soft sweeps: molecular population genetics of adaptation from standing genetic variation**. *Genetics* 2005, **169**:2335-2352.

118 Tournamille C, Colin Y, Cartron JP, Le Van Kim C: **Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals**. *Nat Genet* 1995, **10**:224-228.

119 Di Rienzo A, Hudson RR: **An evolutionary framework for common diseases: the ancestral-susceptibility model**. *Trends Genet* 2005, **21**:596-601.

120 Olson MV, Varki A: **Sequencing the chimpanzee genome: insights into human evolution and disease**. *Nat Rev Genet* 2003, **4**:20-28.

121 Reich DE, Lander ES: **On the allelic spectrum of human disease**. *Trends Genet* 2001, **17**:502-510.

122 Wiuf C: **Do delta F508 heterozygotes have a selective advantage?**. *Genet Res* 2001, **78**:41-47.

123 Gabriel SE, Brigman KN, Koller BH, Boucher RC, Stutts MJ: **Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model**. *Science (80-)* 1994, **266**:107-109.

124 Olson MV: **When less is more: gene loss as an engine of evolutionary change**. *Am J Hum Genet* 1999, **64**:18-23.

125 Thomson G: **HLA disease associations: models for the study of complex human genetic disorders**. *Crit Rev Clin Lab Sci* 1995, **32**:183-219.

126 Hughes AL, Yeager M: **Natural selection and the evolutionary history of major histocompatibility complex loci**. *Front Biosci* 1998, **3**:d509-16.

127 Distante S, Robson KJH, Graham-Campbell J, Arnaiz-Villena A, Brissot P, Worwood M: **The origin and spread of the HFE-C282Y haemochromatosis mutation**. *Hum Genet* 2004, **115**:269-279.

128 Toomajian C, Kreitman M: **Sequence variation and haplotype structure at the human HFE locus**. *Genetics* 2002, **161**:1609-1623.

129 Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R: **Localizing recent adaptive evolution in the human genome**. *PLoS Genet* 2007, **3**:e90.

130 Soranzo N, Spector TD, Mangino M, Kuhnel B, Rendon A, Teumer A, Willenborg C, Wright B, Chen L, Li M, Salo P, Voight BF, Burns P, Laskowski RA, Xue Y, Menzel S, Altshuler D, Bradley JR, Bumpstead S, Burnett M, Devaney J, Doring A, Elosua R, Epstein SE, Erber W, Falchi M, Garner SF, Ghori MJR, Goodall AH, Gwilliam R, Hakonarson HH, Hall AS, Hammond N, Hengstenberg C, Illig T, Konig IR, Knouff CW, McPherson R, Melander O, Mooser V, Nauck M, Nieminen MS, O'Donnell CJ,

Peltonen L, Potter SC, Prokisch H, Rader DJ, Rice CM, Roberts R, Salomaa V, Sambrook J, Schreiber S, Schunkert H, Schwartz SM, Serbanovic-Canic J, Sinisalo J, Siscovick DS, Stark K, Surakka I, Stephens J, Thompson JR, Volker U, Volzke H, Watkins NA, Wells GA, Wichmann H, Van Heel DA, Tyler-Smith C, Thein SL, Kathiresan S, Perola M, Reilly MP, Stewart AFR, Erdmann J, Samani NJ, Meisinger C, Greinacher A, Deloukas P, Ouwehand WH, Gieger C: **A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium**. *Nat Genet* 2009, **41**:1182-1190.

131 Huff CD, Witherspoon DJ, Zhang Y, Gatenbee C, Denson LA, Kugathasan S, Hakonarson H, Whiting A, Davis CT, Wu W, Xing J, Watkins WS, Bamshad MJ, Bradfield JP, Bulayeva K, Simonson TS, Jorde LB, Guthery SL: **Crohn's disease and genetic hitchhiking at IBD5**. *Mol Biol Evol* 2012, **29**:101-111.

132 Morral N, Bertranpetit J, Estivill X, Nunes V, Casals T, Gimenez J, Reis A, Varon-Mateeva R, Macek MJ, Kalaydjieva L: **The origin of the major cystic fibrosis mutation (delta F508) in European populations**. *Nat Genet* 1994, **7**:169-175.

133 Anguiano A, Oates RD, Amos JA, Dean M, Gerrard B, Stewart C, Maher TA, White MB, Milunsky A: **Congenital bilateral absence of the vas deferens. A primarily genital form of cystic fibrosis**. *JAMA* 1992, **267**:1794-1797.

134 Cooper DN, Stenson PD, Chuzhanova NA: **The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms**. *Curr Protoc Bioinformatics* 2006, **Chapter 1**:Unit 1.13.

135 Ewans W: *Mathematical Population Genetics*. Springer; 2004.

136 Cai JJ, Macpherson JM, Sella G, Petrov DA: **Pervasive hitchhiking at coding and regulatory sites in humans**. *PLoS Genet* 2009, **5**:e1000336.

137 McVean G, Awadalla P, Fearnhead P: **A coalescent-based method for detecting and estimating recombination from gene sequences**. *Genetics* 2002, **160**:1231-1241.

138 Necsulea A, Popa A, Cooper DN, Stenson PD, Mouchiroud D, Gautier C, Duret L: **Meiotic recombination favors the spreading of deleterious mutations in human populations**. *Hum Mutat* 2011, **32**:198-206.

139 Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PIW, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE, Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF, Zhang Q, Ghori MJR, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE: **Integrating common and rare genetic variation in diverse human populations**. *Nature* 2010, **467**:52-58.

140 Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA: **Genomic regions exhibiting positive selection identified from dense genotype data**. *Genome Res* 2005, **15**:1553-1565.

141 Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM: **Genomic signatures of positive selection in humans and the limits of outlier approaches**. *Genome Res* 2006, **16**:980-989.

142 Plunkett J, Borecki I, Morgan T, Stamilio D, Muglia LJ: **Population-based estimate of sibling risk for preterm birth, preterm premature rupture of membranes, placental abruption and pre-eclampsia**. *BMC Genet* 2008, **9**:44.

143 Wilcox AJ, Skjaerven R, Lie RT: **Familial patterns of preterm delivery: maternal and fetal contributions**. *Am J Epidemiol* 2008, **167**:474-479.

144 Ward K, Argyle V, Meade M, Nelson L: **The heritability of preterm delivery**. *Obstet Gynecol* 2005, **106**:1235-1239.

145 Plunkett J, Doniger S, Morgan T, Haataja R, Hallman M, Puttonen H, Menon R, Kuczynski E, Norwitz E, Snegovskikh V, Palotie A, Peltonen L, Fellman V, DeFranco EA, Chaudhari BP, Oates J, Boutaud O, McGregor TL, McElroy JJ, Teramo K, Borecki I, Fay JC, Muglia LJ: **Primate-specific evolution of noncoding element insertion into PLA2G4C and human preterm birth**. *BMC Med Genomics* 2010, **3**:62.

146 Shemesh M, Mizrachi D, Gurevich M, Stram Y, Shore LS, Fields MJ: **Functional importance of bovine myometrial and vascular LH receptors and cervical FSH receptors**. *Semin Reprod Med* 2001, **19**:87-96.

147 Mizrachi D, Shemesh M: **Follicle-stimulating hormone receptor and its messenger ribonucleic acid are present in the bovine cervix and can regulate cervical prostanoid synthesis**. *Biol Reprod* 1999, **61**:776-784.

148 Hascalik S, Celik O, Tagluk ME, Yildirim A, Aydin NE: **Effects of highly purified urinary FSH and human menopausal FSH on uterine myoelectrical dynamics**. *Mol Hum Reprod* 2010, **16**:200-206.

149 Padmanabhan V, Sonstein J, Olton PL, Nippoldt T, Menon KM, Marshall JC, Kelch RP, Beitins IZ: **Serum bioactive follicle-stimulating hormone-like activity increases during pregnancy**. *J Clin Endocrinol Metab* 1989, **69**:968-977.

150 Kere J: **Human population genetics: lessons from Finland**. *Annu Rev Genomics Hum Genet* 2001, **2**:103-128.

151 Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res* 2005, **15**:1034-1050.

152 Vallania FLM, Druley TE, Ramos E, Wang J, Borecki I, Province M, Mitra RD: **High-throughput discovery of rare insertions and deletions in large cohorts**. *Genome Res* 2010, **20**:1711-1718.

153 Druley TE, Vallania FLM, Wegner DJ, Varley KE, Knowles OL, Bonds JA, Robison SW, Doniger SW, Hamvas A, Cole FS, Fay JC, Mitra RD: **Quantification of rare allelic variants from pooled genomic DNA**. *Nat Methods* 2009, **6**:263-265.

154 Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**:248-249.

155 Aittomäki K, Lucena JL, Pakarinen P, Sistonen P, Tapanainen J, Gromoll J, Kaskikari R, Sankila EM, Lehväslaiho H, Engel AR, Nieschlag E, Huhtaniemi I, de la Chapelle A: **Mutation in the follicle-stimulating hormone receptor gene causes hereditary hypergonadotropic ovarian failure**. *Cell* 1995, **82**:959-968.

156 Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies**. *Genome Res* 2010, **20**:110-121.

157 Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR: **Low-coverage sequencing: implications for design of complex trait association studies**. *Genome Res* 2011, **21**:940-951.

158 Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes**. *Nucleic Acids Res* 2006, **34**:D108-10.

159 Bryne JC, Valen E, Tang ME, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A: **JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update**. *Nucleic Acids Res* 2008, **36**:D102-6.

160 Newburger DE, Bulyk ML: **UniPROBE: an online database of protein binding microarray data on protein-DNA interactions**. *Nucleic Acids Res* 2009, **37**:D77-82.

161 Ikeda K, Kawakami K: **DNA binding through distinct domains of zinc-finger-homeodomain protein AREB6 has different effects on gene transcription**. *Eur J Biochem* 1995, **233**:73-82.

162 Renthal NE, Chen C, Williams KC, Gerard RD, Prange-Kiel J, Mendelson CR: **miR-200 family and targets, ZEB1 and ZEB2, modulate uterine quiescence and contractility during pregnancy and labor**. *Proc Natl Acad Sci U S A* 2010, **107**:20828-20833.

163 Wei G, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR, Yan J, Talukder S, Turunen M, Taipale M, Stunnenberg HG, Ukkonen E, Hughes TR, Bulyk ML, Taipale J: **Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo**. *EMBO J* 2010, **29**:2147-2160.

164 Bethin KE, Nagai Y, Sladek R, Asada M, Sadovsky Y, Hudson TJ, Muglia LJ: **Microarray analysis of uterine gene expression in mouse and human pregnancy**. *Mol Endocrinol* 2003, **17**:1454-1469.

165 Lussiana C, Guani B, Mari C, Restagno G, Massobrio M, Revelli A: **Mutations and polymorphisms of the FSH receptor (FSHR) gene: clinical implications in female fecundity and molecular biology of FSHR protein and gene**. *Obstet Gynecol Surv* 2008, **63**:785-795.

166 Schenker JG: **Clinical aspects of ovarian hyperstimulation syndrome**. *Eur J Obstet Gynecol Reprod Biol* 1999, **85**:13-20.

167 Romundstad LB, Romundstad PR, Sunde A, von Düring V, Skjaerven R, Gunnell D, Vatten LJ: **Effects of technology or maternal factors on perinatal outcome after assisted fertilisation: a population-based cohort study**. *Lancet* 2008, **372**:737-743.

168 Holdsworth-Carson SJ, Permezel M, Riley C, Rice GE, Lappas M: **Peroxisome proliferator-activated receptors and retinoid X receptor-alpha in term human gestational tissues: tissue specific and labour-associated changes**. *Placenta* 2009, **30**:176-186.

169 Fischer D, Schroer A, Lüdders D, Cordes T, Bücker B, Reichrath J, Friedrich M: **Metabolism of vitamin D3 in the placental tissue of normal and preeclampsia complicated pregnancies and premature births**. *Clin Exp Obstet Gynecol* 2007, **34**:80-84.

170 Hermann BP, Hornbaker KI, Maran RRM, Heckert LL: **Distal regulatory elements are required for Fshr expression, in vivo**. *Mol Cell Endocrinol* 2007, **260-262**:49-58.

171 Hermann BP, Heckert LL: **Silencing of Fshr occurs through a conserved, hypersensitive site in the first intron**. *Mol Endocrinol* 2005, **19**:2112-2131.

172 [http://www.repeatmasker.org]

173 Ratajczak CK, Fay JC, Muglia LJ: **Preventing preterm birth: the past limitations and new potential of animal models**. *Dis Model Mech* 2010, **3**:407-414.

174 Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW: **SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap**. *Bioinformatics* 2008, **24**:2938-2939.

175 Gromoll J, Dankbar B, Gudermann T: **Characterization of the 5' flanking region of the human follicle-stimulating hormone receptor gene**. *Mol Cell Endocrinol* 1994, **102**:93-102.

176 Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers**. *Methods Mol Biol* 2000, **132**:365-386.

177 Mitra RD, Butty VL, Shendure J, Williams BR, Housman DE, Church GM: **Digital genotyping and haplotyping with polymerase colonies**. *Proc Natl Acad Sci U S A* 2003, **100**:5926-5931.

178 Ovcharenko I, Nobrega MA, Loots GG, Stubbs L: **ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes**. *Nucleic Acids Res* 2004, **32**:W280-6.

179 Sandelin A, Wasserman WW, Lenhard B: **ConSite: web-based prediction of regulatory elements using cross-species comparison**. *Nucleic Acids Res* 2004, **32**:W249-52.

180 González-Pérez A, López-Bigas N: **Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel**. *Am J Hum Genet* 2011, **88**:440-449.

181 Gimelbrant AA, Skaletsky H, Chess A: **Selective pressures on the olfactory receptor repertoire since the human-chimpanzee divergence**. *Proc Natl Acad Sci U S A* 2004, **101**:9019-9022.

182 Marini NJ, Thomas PD, Rine J: **The use of orthologous sequences to predict the impact of amino acid substitutions on protein function**. *PLoS Genet* 2010, **6**:e1000968.

183 Chen FC, Li WH: **Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees**. *Am J Hum Genet* 2001, **68**:444-456.

184 Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T: **Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection**. *Genome Res* 2011, **21**:349-356.

185 Ludwig M: **Are adverse outcomes associated with assisted reproduction related to the technology or couples' subfertility?**. *Nat Clin Pract Urol* 2009, **6**:8-9.

186 Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR: **Power of deep, all-exon resequencing for discovery of human trait genes**. *Proc Natl Acad Sci U S A* 2009, **106**:3871-3876.

Curriculum Vitae

Sung Gook Chun

Contact Information:

Sung Gook Chun

Center for Genome Sciences and Systems Biology (Justin Fay lab)

4444 Forest Park Ave.

St. Louis, MO 63108

Education:

2006-2012	Washington University in St. Louis, Missouri PhD in Computational and Systems Biology
1996-2000	Korea Advanced Institute of Science and Technology, South Korea BS in Computer Science with Mathematics minor

Employment:

2004-2005	Researcher, Bioinformatics Research Laboratories, Co., South Korea
2000-2003	Software Engineer, Alticast Corp., South Korea

Honor:

2011	Walter M. Fitch student award contestant at the 2011 annual meeting of Society for Molecular Biology and Evolution
1996-2000	KAIST merit-based scholarship

Other experience:

2008-2010	Lucille P. Markey Special Emphasis Pathway in Human Pathobiology
Fall 2007	Teaching assistant to Michael Brent (Computational Molecular Biology)
1996-2000	College UNIX programming club

Publication:

1. **Chun, S.** and Fay, J.C. Identification of deleterious mutations within three human genomes. *Genome Research* (2009) 19:1553-1561.
2. **Chun, S.** and Fay, J.C. Evidence for hitchhiking of deleterious mutations within the human genome. *PLoS Genetics* (2011) 7(8): e1002240.
3. Lawson, H.A., Nikolskiy, I., **Chun, S**., McLellan, M.D., Fay, J.C., Mardis, E.R. and Cheverud, J.M. Whole-genome sequences of LG/J and SM/J inbred mouse strains. *Submitted*.
4. **Chun, S.**, Plunkett, J., Teramo, K., Muglia, L.J. and Fay, J.C. Fine-mapping of the genetic association of *FSH receptor* with preterm birth in the Finnish population. *Manuscript in preparation.*

Conference presentations:

1. **Chun, S.** and Fay, J.C. *Invited Speaker.* Hitchhiking regions are enriched with deleterious mutations within the human genome. Walter M. Fitch student talk at the 2011 annual meeting of the Society for Molecular Biology and Evolution, Kyoto, Japan.
2. **Chun, S.** and Fay, J.C. Hitchhiking regions are enriched with deleterious mutations within the human genome. Contributed talk at the Symposium Evolution 2011, Norman, Oklahoma.
3. **Chun, S.** and Fay, J.C. Low overlap among predictions of deleterious mutations within three human genomes using SIFT, PolyPhen, and a new likelihood ratio test. Poster presented at the 2nd meeting on Personal Genomes at Cold Spring Harbor Laboratory, New York.
4. **Chun, S.** and Fay, J.C. Identification of deleterious mutations within three human genomes. Contributed talk at the 2009 annual meeting of the Society for Molecular Biology and Evolution, Iowa City, Iowa.