THE EVOLUTIONARY AND CLINICAL SIGNIFICANCE OF REGULATORY AND MITOCHONDRIAL GENETIC VARIANTS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Kaixiong Ye

January 2015

THE EVOLUTIONARY AND CLINICAL SIGNIFICANCE OF REGULATORY AND MITOCHONDRIAL GENETIC VARIANTS

Kaixiong Ye, Ph. D.

Cornell University 2015

Genetic adaptations to local environment during evolution shaped the human genome. Identifying evolutionarily important genetic variants is clinically significant because the mismatch between our slow-evolving genome and the cultural upheaval underlies many diseases of civilization. Recent sequencing technology advances start a Genomic era and promise to elucidate the genetic basis of human health and disease. While most attention had been drawn to protein-coding genes in the nuclear genome, regulatory regions and mitochondrial genome were much less studied. My research aims to investigate the evolutionary and clinical significance of these two categories.

Starting my research, I summarized the latest advances in understanding the role of nutrition in human genome evolution and introduced to the Nutrition field population genomics approaches to identify dietary adaptions. My first research project examined the adaptation of regulatory variants to 42 environmental factors. With a newly developed environmental correlation method, I found that expression QTLs are enriched in signals of environmental adaptation and regulatory adaptation are especially important for some environmental factors, like climate, and for some biological pathways, including immune and metabolic pathways.

My second project was a case study to test a hypothesis that an Asian-common

haplotype was adaptive to the plant-based diet in Asia by enhancing non-heme iron absorption. With an iron absorption study involving 57 Asian women volunteers, I found that consistent to our hypothesis homozygous carriers of the adaptive haplotype absorbed 22% more non-heme iron than non-carriers. Intriguingly, I also observed that compared to Caucasian women, Asian women absorbed more non-heme iron even in face of higher iron store.

My third project utilized the next-generation sequencing data from the 1000 Genomes Project and investigated the prevalence and clinical relevance of mitochondrial DNA (mtDNA) heteroplasmy, the presence of multiple versions of mtDNA in a cell. I found that about 90% healthy individuals carry at least one heteroplasmies and at least 20% individuals harbor disease-associated heteroplasmies. I demonstrated that heteroplasmic mutations are highly pathogenic and subject to weak negative selection. My research suggests that heteroplasmic mutations may drift to high frequency across life-span and contribute to age-related diseases.

BIOGRAPHICAL SKETCH

Kaixiong Ye grew up in Shantou, a Southern city in China. He graduated from Jin Shan Middle School in 2004 and traveled half of the China to start his undergraduate study in Wuhan University in central China. In his undergraduate years, Kaixiong was obsessed with Darwin's Theory of Evolution and intrigued by its power in explaining the diversity of life. After reading several versions of Darwin's biography, he decided to embark on his own journey of evolutionary research. He firstly joined an Ecological Evolution lab and participated in ecological research in the mountains of the Yunnan–Guizhou Plateau. Unimpressed by the research techniques as old as Darwin himself, Kaixiong decided to try something new and joined a Human Genetics lab in the Kunming Institute of Zoology at the southwest corner of China. It was in this Human Genetics lab under the supervision of Prof. Bing Su that Kaixiong first tasted the charm and fun of scientific research and later decided to pursue an advanced degree on Human Genetics.

Before attending graduate school at Cornell University, Kaixiong worked for half a year as a bioinformatician in Beijing Genomics Institute at Shenzhen. This experience exposed him to the then most innovative sequencing technology and provided him intensive programing training. In the past five years, Kaixiong has been working under the mentorship of Dr. Zhenglong Gu to investigate the evolutionary and clinical significance of regulatory variants and mitochondrial DNA mutations.

To my parents

(叶育宣，叶苧卿)

for your love and support

And to my brother and sisters

(叶凯升，叶凯丽，叶凯丹)

for your support and encouragement

and for taking care of Dad and Mom

ACKNOWLEDGMENTS

Pursuing a Ph.D.'s degree is not easy. I feel very fortunate that I have met many respectful mentors and lovely friends who made the challenging process enjoyable.

Firstly, I would like to thank my advisor, Dr. Zhenglong Gu, for offering me the opportunity to work with him for the past five years, and for his continuous effort to help me succeed in any of my professional endeavors. It was under his mentorship that I truly developed into an independent and confident researcher. I also want to thank my committee members, Drs. James Booth, Jason Mezey, Kimberly O'Brien and Tom Brenna for their academic guidance. Especially, I learned a lot from Drs. O'Brien and Brenna through our collaborative projects. Moreover, I would like to thank Dr. Alon Keinan and his research group. My weekly involvement in the Keinan lab meeting and my regular discussion with them were extremely useful for my research.

Secondly, I am very grateful for my labmates and Cornell friends. The Gu lab members, especially Huifeng Jiang, Lin Xu, Brandon Barker, and Xiaoxian Guo, have been very supportive and helpful in my research and daily life. All my Cornell friends, such as Yue Yu, the Larson family, Liuqi Gu, were truly amazing and they made my Cornell life much more enjoyable and memorable.

Thirdly, I would like to thank my life-long friends, especially Lei Cao, Xiong Ji, and He Fu, who were not in Ithaca but have been hugely supportive whenever I need them.

Last but not least, I feel deeply indebted to my parents, younger brother and sisters for their unconditional love and support. I also feel especially grateful and fortunate to meet my girlfriend, Yuan Si, who brings me joy and happiness every day.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1 – Recent Advances in Understanding the Role of Nutrition in Human Genome Evolution[1]

1.1. Abstract

Dietary transitions in the human history have been suggested to play important roles in the evolution of mankind. Genetic variations caused by adaptation to diet during human evolution could have significant health consequences in current society. The advance of sequencing technologies and the rapid accumulation of genome information provide unprecedented opportunity to comprehensively characterize genetic variations in human populations and to unravel the genetic basis of human evolution. Series of selection detection methods, based on various theoretical models and exploiting different aspects of selection signatures, have been developed. Their applications at the species and population levels have respectively led to the identification of human specific selection events that distinguish human from non-human primates and local adaptation events that contribute to human diversity. Scrutiny of candidate genes has revealed paradigms of adaptations to specific nutritional components while genome-wide selection scans have verified the prevalence of diet-related selection events and provided many more candidates awaiting further investigation. Understanding the role of diet in human evolution is fundamental for the development of evidence-based, genome-informed nutritional practices in the era of personal genomics.

[1] Published on *Advances in Nutrition*. See Appendix A for inclusion authorization.

1.2. Introduction

1.2.1. The role of diet in human evolution

Food represents one of the most important environmental factors for human. Genetic adaptations to the diet consumed along human history have sculpted the human genome and influenced a variety of human traits. Local adaptations to regionally specific dietary components might have been one critical shaping force of the human genome, driving population differentiation and laying the genetic basis for human diversity. Genomic adaptations to environmental factors through increasing frequency of advantageous mutations in human population usually take hundreds of generations (thousands of years), whereas societal, cultural, and dietary transformations in the human society are ever-accelerating. Mal-adaptations of the lagging genome to rapid dietary shifts may underlie a wide range of so-called civilization diseases, such as diabetes, obesity, cardiovascular diseases and cancers. Extensive efforts have been invested into the examination of genomic adaptations to diet in the hope of elucidating the genetic basis of complex disorders and tailoring strategies of disease management through personalized medicine and nutrition (1-3).

The objective of this review is to briefly overview the shaping effects of food on the human genome and to discuss insights learned from evolutionary research into genome-informed clinical and nutritional practices. Along the way, we will summarize the main dietary transitions in human history, the corresponding selective events on the genome, approaches utilized to detect selection signals, and knowledge gained about the genetic basis of complex traits and diseases. Comprehensive reviews could

be found elsewhere of evidences for adaptations to diet (2, 3), or more generally of human demographic history and evolutionary adaptations (4-6). So are specialized reviews of approaches for detecting selection and their specific advantages and limitations (7-9).

1.2.2. Dietary transitions during human evolution

There are three major dietary transitions in human history (Figure 1.1.). Upon the split of human-chimpanzee about 4.6-6.2 mya (million year ago) (10), the early hominid (*Ardipithecus ramidus,* about 6-4 mya) was suggested to reside in a predominantly wooded habit and exploit a generalized plant-based diet (11, 12). Around 4 mya *Australopithecus,* the predecessor of *Homo,* began to transfer to more open landscapes and to consume harder and more abrasive foods, such as large seeds with tough shells and underground storage organs. This dietary pattern was supported by the adaptively enlarged teeth and thickened enamel (11). *Homo* arose about 2.5 mya and began to use stone tools, leading to an increased consumption of animal source foods through scavenging and hunting (11, 13). Another significant technological innovation in this period, around 1.6 mya onward, was the controlled use of fire (14), which also contributed to this dietary shift. This dual dietary strategy, providing essential nutrients from animal source foods and energy mainly from plant source foods, made possible the evolution of human as a large, active and highly social primate with an unraveled complex brain (15).

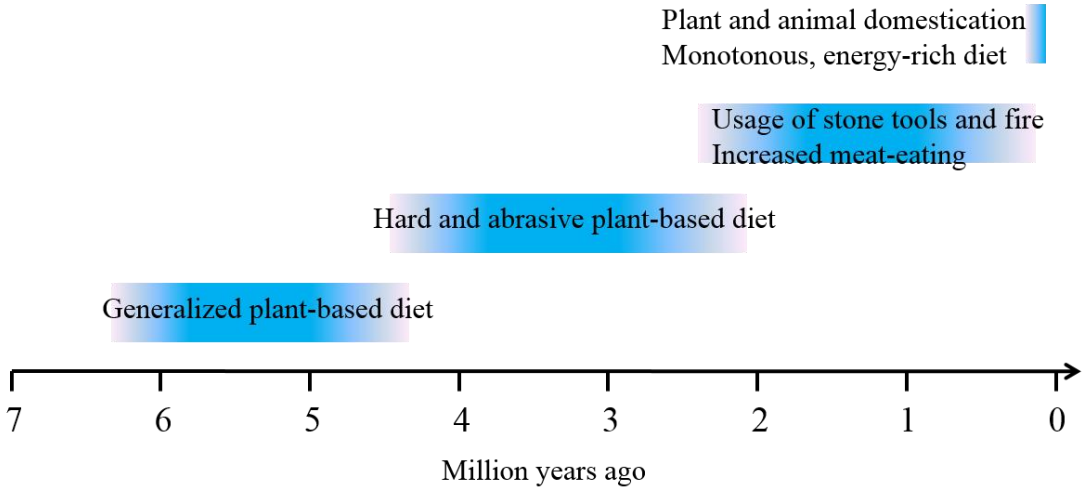


Figure 1.1. Major dietary transitions in human history. The dietary pattern of each period is depicted with a box with blurry boundaries representing the transitional processes.

Around 200 kya (thousand year ago), anatomically modern humans (*Homo sapiens*) originated in Africa and expanded to occupy other parts of the world replacing the local archaic populations (16, 17). The Neolithic Revolution, characterized by domestication of plants and animals, occurred about 10 kya and stimulated the most dramatic dietary transformation in human history. Varying environmental resources on different continents led to domestications of local species, and thus adoptions of regionally specific dietary patterns. In spite of the pronounced explosion of food production, only a few wild species were domesticated and has since then been heavily relied on (18). Some agricultural societies might draw 90% of their energy input from domesticated plants (19). Overall, this dietary transformation resulted in loss of nutritional diversity, excess of caloric availability and sedentary lifestyle, contributing to most of the health challenges in modern society (20-22).

Adaptations to dietary transitions during human evolution have left specific signatures in our genome. Identifying these signatures in the hope of unraveling human evolutionary history and improving human health has been an active research direction in the fields of evolutionary genomics, population genetics, and molecular evolution.

1.2.3. Genetic variation and its functional significance

Genetic variation underlies the tremendous range of phenotypic diversity and disease susceptibility in human populations. Any two individuals are estimated to be different in 1- 3% of their genomes (23). Each person is estimated to carry about 250 - 300 genes with loss-of-function variants, 50 - 100 of which are implicated in inherited disorders (24). With the rapid advance of sequencing technologies, especially the

advent of next and next-next generation sequencing platforms, extensive effort have been invested into unraveling human genetic variations, and into understanding their general features, forces shaping their pattern, and most importantly, their clinical consequence. Genome-wide association studies (GWAS) have been successful in linking some genetic variations to complex traits and common diseases. In spite of confronting difficulty in explaining all genetic impacts, it still holds the promise of elucidating the genetic architecture of human health, providing interesting hypotheses and pointing possible directions for medical research (25).

Genetic variations include single nucleotide polymorphism (SNP), structure variation (SV) and chromosomal abnormality. SNP is the simplest, most common and most well-studied variation. It is estimated that there are in total 10-15 million common SNPs with frequency higher than 1% in human genomes and 3 Millions are estimated to be present in each individual (24, 26-28). SVs, according to the molecular mechanisms of their origin, could be classified into deletion, insertion, reversion, translocation, duplication and their complicated combinations. Some SVs result in different copy numbers of sequence unit and are called copy number variations (CNV). The importance of SV has been underestimated until the recent development of sequencing technologies, which makes its characterization possible. It is estimated that CNVs cover 12-30% of the human genome (29, 30). Not all genetic variations have functional effects. In other words, their existences will not make their carriers sicker or healthier. Only functional variations that influence the fitness of their carriers are subject to natural selection. Functional variations in coding regions may directly change protein structure or its function while those in non-coding regions may

interfere with the regulation of gene expression. Usually, coding SNPs are classified into synonymous SNPs and non-synonymous SNPs. Synonymous SNPs do not change the peptide sequence because two codons encode the same amino acids. Non-synonymous SNPs may change the amino acid (missense SNPs) or introduce a stop codon (nonsense SNPs). Non-synonymous SNPs are more likely to have phenotypic effects.

The pattern of genetic variation in human populations is the product of interactions among demographic history (e.g. effective population size, population structure, and migration), gene-specific factors (mutation and recombination rate), selection pressure and random process (also called genetic drift) (31). Mutation rate determines how many variations are introduced into the genome. Under simple demographic models without selection, mutation rate equals to evolutionary rate. Recombination rate is another shaping force of variation pattern, especially when coupled with selection and demographic history. Recombination rate is the main parameter determining the degree of linkage disequilibrium between adjacent variants and the coupling of their evolutionary fate. Regions identified as targets of selection show a drastically reduced level of recombination, probably due to the fact that low recombination rate decelerates the deterioration of selection signatures (8, 32). For the same reason, regions of high recombination rate attenuate the sweeping effect of selection and tend to have high variability, though there is also a hypothesis that recombination could be mutagenic (33). Demographic history is an important determinant of variation patterns in different populations (34). Demographic events, such as population bottleneck and population growth, can create variation patterns that mimic those of selection. Thus it

is always important to take into account the effect of demographic history when interpreting potential selection signals (31, 35). Selection only occurs on functional variations and thus creates dramatically different patterns of evolution and population dynamics between functional and non-functional variants, which usually is regarded as important indicators of selection (9).

1.2.4. Modes of natural selection

Natural selection can be broadly defined as the process in which beneficial heritable traits increase in frequency while unfavored ones decrease. The reflection of this process at the molecular level is the spread of advantageous alleles and the purification of deleterious alleles in the population. Natural selection happens in a variety of modes. According to the trajectory of frequency shift for the allele under study, selection can be classified into three categories: positive selection, negative selection and balancing selection. Positive selection refers to the process that a beneficial allele spreads in the population and ultimately reaches fixation. In contrast, negative selection depicts the process that a deleterious allele decreases in frequency and ultimately becomes extinct. Balancing selection occurs under two scenarios. Under one scenario, the presence of multiple conflicting selective pressures, instead of a single selective pressure or multiple selective pressures in a single direction, renders the targeted allele both beneficial and deleterious. Ultimately the frequency of the targeted allele will fluctuate around a specific level. The other scenario is called heterozygous advantage under which the heterozygote is favored and both alleles are maintained in the population at the equilibrium frequency.

Positive selection is of special interest because it underlies regional adaptation and development of novel traits. Positive selection alone has a wide range of modes, leaving distinct signatures along the selected genomic region (Figure 1.2. and 1.3.) (36). Two extensively studied models of positive selection are complete sweep and partial sweep, both of which emphasize the onset of selection on a newly mutated allele, driving it to rapidly increase in frequency. Complete sweep describes the scenario in which the allele reaches fixation in the population while partial sweep refers to the situation that the allele only reaches high frequency but not fixation. Under both scenarios, the mutation of the advantageous allele happens in a specific genomic background (haplotype) and is linked with specific alleles of nearby variable sites. The strong selection on the favored allele rapidly drags the haplotype to high frequency or fixation in such short time that mutation and recombination do not drastically degrade the selected haplotype, leaving specific selection signatures, including long-range haplotype with high frequency, skewed site frequency spectrum (excess of rare variants, or excess of high-frequency derived alleles), reduced variation level in the selected region, etc (7, 9).

In contrast with complete sweep and partial sweep, both of which are also called hard sweep, soft sweep depicts another situation under which selection operates on standing

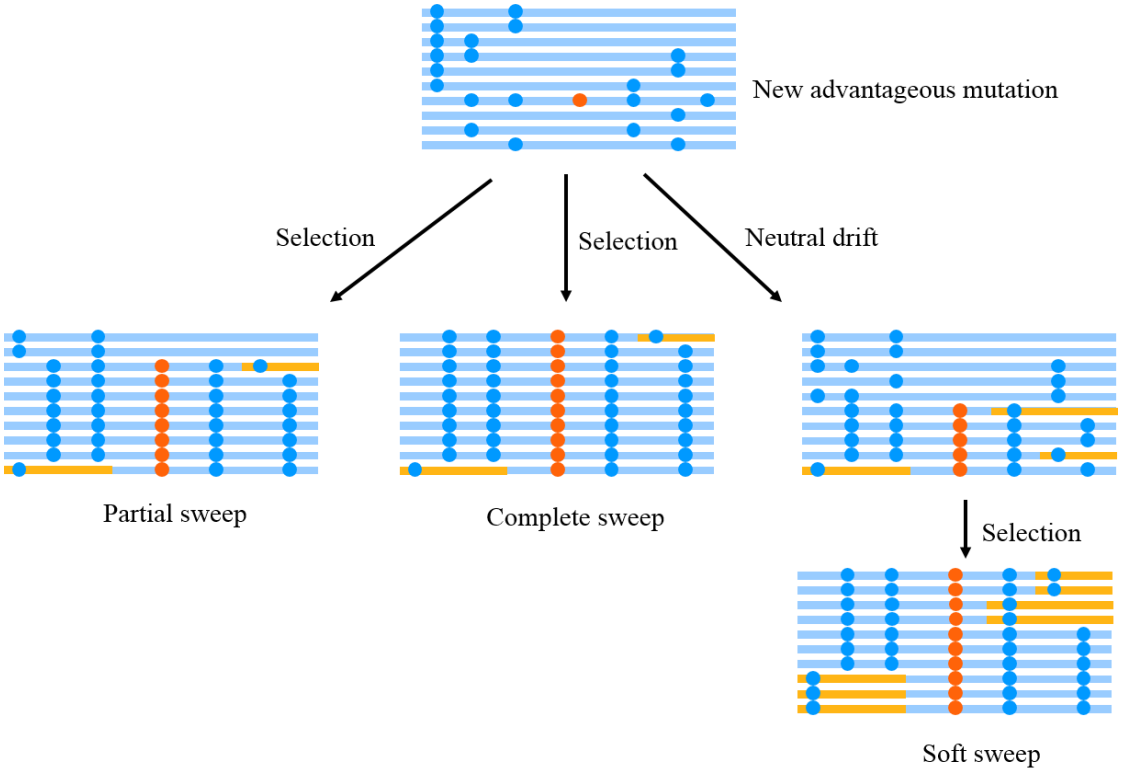


Figure 1.2. Models of partial sweep, complete sweep and soft sweep. Each horizontal line represents a haplotype defined by different combinations of alleles of adjacent variable sites. Blue lines are haplotype present before selection and new haplotype generated from recombination are depicted with an additional color (yellow). The derived allelle (mutation) of a variable site is illustrated as a dot. The orange dot represents the beneficial mutation under selection. Hard sweep, both partial and complete sweep, emphasizes the onset of selection on a new advantageous mutation. Haplotypes, old or new, carrying the selected mutation will rapidly increase in frequency to high frequency (partial sweep) or to fixation (complete sweep). The time to reach high frequency or fixation is so short that not many recombination events occur. Soft sweep occurs on a standing variation, rather than a new mutation. Before the onset of selection, the mutation is neutral and its frequency fluctuates randomly. In this period of neutral drift, recombination creates new haplotypes carrying the mutation. Once environmental change renders the mutation beneficial, all haplotypes carrying this mutation will rapidly spread in the population. Although, the selected mutation also reach high frequency or fixation as did under hard sweep, the haplotype homogeneity has been degraded.

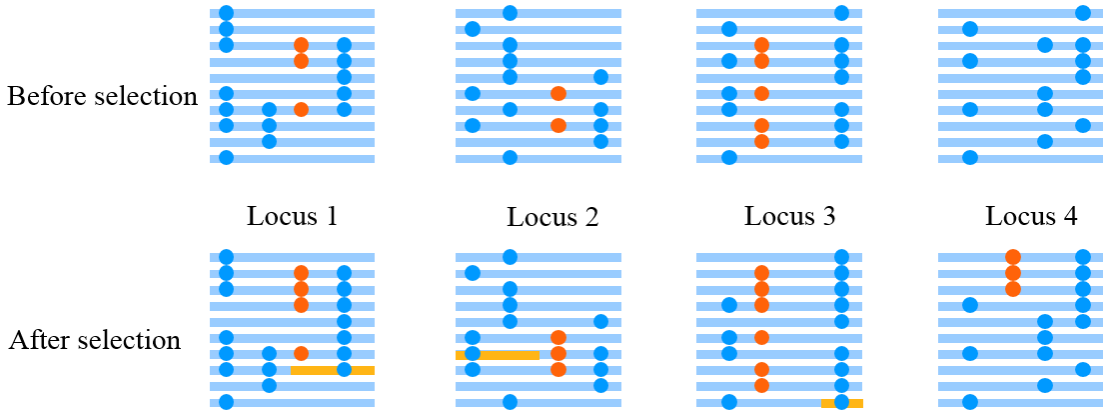


Figure 1.3. Polygenic adaptation model. Multiple loci (4 in this case) are responsible for a single trait. Most loci (loci 1, 2, 3) harbor mutations later becoming beneficial while there are also loci (locus 4) gaining new advantageous mutations after the onset of selection. The selected mutations are present at different frequency in the population before selection as a result of random drift. Upon environmental change, selection acts on all these beneficial mutations and elevates their frequency. However, each mutation only needs to have a slight increase in frequency.

variations or simultaneously on multiple alleles of a locus. Standing variations refer to variations that have been present and evolved neutrally before the action of selection when environmental shifts render them adaptive. The initial neutral drifting phase allows mutation and recombination to erode the homogeneous genetic background upon which the new mutations arise. When multiple alleles of a locus are similarly adaptive, all of them will be selected but none of them will be able to reach fixation. Under this situation the selected alleles have limited effects on the linked sites (37).

Selective sweeps, both hard and soft sweep, emphasize the drastic frequency change before and after selection. However, adaptation could also happen through subtle allele frequency shifts of many loci, which is recently proposed as a model of polygenic adaptation (36-38). For complex traits controlled by multiple genes of small effect, a new optimum phenotype could be reached by simultaneous subtle allele frequency shifts across many loci without dramatic allele frequency changes or fixation events. The relative importance of hard sweep, soft sweep and polygenic adaptation to human evolution is still difficult to estimate since most studies have been focused on hard sweep and it was only possible in the past decade to perform genome-wide assessment of natural selection events. Moreover, both soft sweep and polygenic adaptation only received attention recently. Theoretical modeling and empirical genome-wide examination are required to understand their relative importance.

1.2.5. Neutrality tests for detecting natural selection

A series of neutrality tests have been developed to identify selection signatures that are distinct from those of neutral evolution. Different tests exploit different aspects of selection signatures, have specific underlying assumptions (about selection mode, strength, scale, etc) and have varying power for different selective modes. Human specific adaptation events after the split of human with other primates can be detected by comparative genomic analysis between human and non-human primates while regional adaptation events along with human global migration can be analyzed using human population genomic data from populations of different habitats. For each neutrality test, it is essential to assess the significance of the signal, that is, to differentiate a true selection signal from those by neutral processes. For this purpose, neutral signals are usually gained from simulations with equivalent demographic models without selection or from empirical distributions of functionally neutral loci. Due to the complexity of human demographic history, it is usually difficult to bring up an accurate realistic model. With the availability of genome-wide variation data, however, using empirical neutral distribution is more practical and convenient. Neutrality tests have been applied for specific genes and for genome-wide selection scans. The candidate gene approach is hypothesis-driven and usually provides detailed evolutionary stories while genome-wide scans are unbiased, which can reveal a complete picture of adaptation events and generate hypotheses for further candidate gene approach. A large number of loci under selection have been revealed with the application of these methods.

1.2.5.1. Comparative analysis

Without selection, the evolutionary rate of functional mutations (e.g. non-synonymous mutations, regulatory mutations) should be the same as that of non-functional ones (e.g. synonymous mutations). However, if there were repeated selective events on multiple functional mutations in a gene or a regulatory region and drove these mutations to fixation, we would find an elevated evolutionary rate of functional mutations compared with that of neutral ones (dN/dS test) (39-41). In the absence of selection the polymorphism levels within species and sequence divergence between species should be positively correlated. Similarly, the ratio of nonsynonymous to synonymous changes between species should be equal to that within species. Departures from these neutral expectations are signals of natural selection and could be tested by HKA test and MK test, respectively (42, 43). A creative comparison of the evolutionary rate at the promoter regions with that at neutral intron regions made a surprising discovery that nutrition-related genes have experienced positive selection at the regulatory regions (41). It is worthwhile to point out that classical examples of metabolic adaptation, such as lactose digestion (44) and starch metabolism (45), experienced their selections on gene regulation and adapted to varying food sources by changing gene expression levels. Selection on regulatory regions may have played an important role in the evolution of metabolic genes, though its relative importance compared to selection on coding regions during human evolution are still difficult to assess.

1.2.5.2. Site frequency spectrum -based tests

After a complete selection sweep, the variation levels along the selected regions were reduced because most variations were wiped out due to fixation of the selected alleles. While new mutations began to accumulate in this region, they were only present in low frequency given limited time after the selection event. Thus the site frequency spectrum (SFS) skewed with an excess of low-frequency polymorphisms. Several tests, such as Tajima's D test (46) and Fu and Li's tests (47), have been developed for this type of signature left by complete sweep. Under the scenario of partial sweep, the rapid increase of a beneficial allele to high frequency drags alleles of nearby loci to similarly high frequencies (a phenomenon called hitchhiking). If the derived alleles of linked loci happened to be in linkage with the favored allele, the hitchhiking effect will result in an excess of high-frequency derived alleles, which would otherwise remain at low frequency given their young age. This type of skewed SFS can be detected with Fay and Wu's H test (48). Since complete re-sequencing data is needed to capture the accurate SFS with extremely rare variations, SFS-based tests are usually applied in candidate gene studies. For instance, *CYP3A4* and *CYP3A5* genes have a marked excess of rare variants in Asian and European populations as revealed by negative Tajima's D. *CYP3A4* gene also has an excess of high-frequency derived alleles in both populations (49). Taken together with other evidences, these two genes are suggested to be adaptive to climate-related selection, contributing to the current population disparity in the prevalence of salt-sensitive hypertension (49).

1.2.5.3. Haplotype-based tests

When a new advantageous mutation arose, it did so in a specific haplotype background. Positive selection on the beneficial variant dragged the haplotype to high frequency in such short time that mutation and recombination did not substantially break down the haplotype. Thus a long-range haplotype with an usually high frequency in a population is a signature of positive selection (50, 51). To control for regional heterogeneity of recombination rate, several derived tests were developed to compare the haplotype of interest with other haplotypes in the same region (50, 52). While these tests are powerful for partial sweeps, they lose power for complete sweeps or when the selected allele approaches fixation because there are few alternative alleles to serve as control. A comparison between populations may solve this problem because when the selected allele reaches fixation in one population, it remains polymorphic in the others (53). A genome-wide selection scan using haplotype-based tests in three main continental populations (African, European, and Asian) found genes with exceptionally long common haplotypes in one or more populations. Examples include *LCT* (a gene responsible for lactose digestion) in Europeans, *GBA* (a gene associated glycogen-storage disorder) in East Asians, *NKX2-2* (a gene associated with carbohydrate metabolism and blood sugar regulation) in Europeans, and *CYP* genes (involving in detoxification of foreign compounds) in Asians and Europeans (52).

1.2.5.4. Population differentiation and Environmental correlation

The global environmental disparity creates an adaptive landscape of drastic heterogeneity. The presence or absence of certain environmental factors confers

selective pressure only in some populations but not others. Besides dichotomous environmental factors, continuous ones with geographical clines confer a gradient of selective strength, leading to gradual adaptive frequency changes. Furthermore, similar selective pressure in different populations, either geographically distant or close, might result in parallel adaptive frequency shifts. These spatial patterns of variation as a result of adaptation to varying environmental variables can be detected by population differentiation or environmental correlation methods. Population differentiation measures the difference of allele frequency between populations. Large population differentiation usually indicates the presence of natural selection and local adaptation. Wright's fixation index (F_{ST}) and its derivatives have been used for this purpose (54, 55). *ADH1B* in alcohol metabolism (56), *HFE* in iron metabolism (57), *LCT* in lactose digestion (58), among many other metabolic genes, have variations with unusually high level of population differentiation.

A direct correlation between allele frequency and environmental variables is also a signature of adaptation to various degrees of selective pressure. This method is called environmental correlation. One advantage of this method is that it is applied with the knowledge of selective pressure whereas all other approaches described above only examine the sequence information. Another advantage is that it is powerful even for subtle allele frequency shifts, making it especially useful for selection on standing variations or polygenic adaptation model. One confounding factor in environmental correlation is population structure, as geographically close populations tend to share similar frequency while those of distant populations are different. Although there are parallel allele frequency shifts in distant populations due to similar selective pressure,

these changes might not be apparent unless the regional averages are controlled for (37, 59). A Bayesian linear model was recently developed to assess the evidence of allele frequency and environmental variable correlation while controlling for the covariance of allele frequencies between populations as a result of population structure (60). Two genome-wide selection scans with this new method utilized 61 global populations and assessed the correlations between ~ 650,000 SNPs and 20 environmental factors. They discovered strong signals of adaptation to polar eco-region, foraging subsistence, diet rich in roots and tubers, and several climate factors (e.g. temperature, precipitation, and solar radiation) (61, 62). Especially, genes involved in metabolism were found to play an important role in adaptation to these environmental factors (61, 62). For instance, pathways of starch and sucrose metabolism, and folate biosynthesis are enriched with strong signals of adaptation to a diet rich in roots and tubers, which are poor dietary sources for folic acids (61). Intriguingly, genes underlying complex diseases, including common metabolic disorders (e.g. type 2 diabetes, obesity, hypertension, and dyslipidemia), show a significant overlap with these adaptive signals (61-63).

Applications of these statistical methods in candidate gene investigations and/or genome-wide selection scans have revealed a large number of genes, functional groups, and pathways under selection during human evolution. Well-studied cases of adaptation to food or diet-related practices, which are discussed in details hereafter, include regulatory variants of the lactase gene adaptive to milk consumption, copy number variation of amylase gene adaptive to starchy food, enhanced *ADH1B* activity preventive from alcohol overconsumption, increased sensitivity of bitter taste receptor

gene *TAS2R16* avoiding ingestion of plant toxins, etc. Beside individual genes, certain pathways or functional groups are also enriched with genes or polymorphisms under selection. Genome-wide scans for recent selection signatures found that selective signals are enriched in starch/sucrose metabolism and folate biosynthesis pathway (61), carbohydrate, steroid, phosphate metabolism and vitamin/cofactor transport (52), protein and DNA metabolism (51), etc. Ancient selection signatures were found enriched in the promoter regions of nutrition-related genes (41), nucleoside, nucleotide and nucleic acid metabolism (64), amino acid metabolism (40), etc. It is advisable to bear in mind that different approaches of selection detection utilize different types of data, exploit different aspects of selection signatures, and detect selective events at varying levels and timescale. Between species comparison (dN/dS, MK test, HKA test) examine selective events older than 1 mya at the early stage of human evolution while population genomics analyses (haplotype-based tests, frequency spectrum-based tests, population differentiation and environmental correlation) will detect more recent adaptation events, usually less than 10 kya (4, 8, 9).

To illustrate the action of natural selection and approaches to identify its signatures, we will discuss several well-studied cases of adaptation of metabolism and perception to dietary components or diet-related practices. We need to point out that our discussion here is intended to be representative but surely not exhaustive, and we apologize for missing some interesting studies in this area.

1.3. Classic examples of dietary adaptation

1.3.1. Metabolism

1.3.1.1. Lactase persistence

Lactase, an enzyme expressed in intestine, is required for the digestion of lactose in milk. Most mammals, including human, lose the ability to digest lactose after weaning due to reduced lactase expression. The persistence of the ability to digest milk lactose into adulthood is called lactase persistence (LP), which has a genetic basis and is inherited as a dominant trait. The prevalence of LP is high among northern Europeans (>90% in Swedes and Danes), decreases across southern Europe and the Middle East (~50% in Spanish, French, and pastoralist Arab populations), and is low in the African and Asian agriculturalists (~5-20% in West African and ~1% in Chinese). Notably, LP is also prevalent in African pastoralist groups, like the Tutsi (~90%) and Fulani (~50%) (44). Interestingly, LP has different genetic basis in these populations. Two SNPs C/T-13910 and G/A-22018 (rs4988235 and rs182549), identified in the *cis*-regulatory elements of the gene encoding lactase (*LCT*) are highly associated with LP in European populations and C/T-13910 is proposed to be the causal regulatory site (65). However, the SNP C/T-13910 is only present with low frequency in a few West African pastoralist populations and absent in others. Recent studies in eastern African populations identified three new SNPs to be significantly associated with LP and enhance transcription of the *LCT* promoter *in vitro* (44). More recent study in Middle Easterns also revealed a population-specific haplotype responsible for LP (66). Evolutionary analyses (population divergence, long range haplotype, etc.) detected significant selection signatures on these functional variants

and age estimates of mutations (~8000-9000 years ago in Europeans, ~2700-6800 years ago in Eastern Africans, ~4000 years ago in Saudi Arabia) are consistent with the timeframe of animal domestication in each region, indicating convergent adaptations in different regions to similar selective pressures of milk consumption (44, 58, 66).

1.3.1.2. Starch digestion

The gene *AMY1*, encoding salivary amylase for starch digestion, shows extensive variations in copy number among individuals and between human populations. The copy number of the gene is positively correlated with its protein expression. Populations consuming high-starch diets, such as agricultural populations of European Americans, Japanese, and Hadza hunter-gatherers, have higher copy number of *AMY1* than low-starch populations, like hunter-gatherers in the rainforests and near the Arctic Circle. Comparative analyses with other primates suggest that the additional copy of *AMY1* was gained in the human lineage. The low amount of nucleotide divergence among different gene copies might be a result of recent origin that fell within the timeframe of modern human origins (~200,000 years ago). Taken together, the copy number variations of *AMY1* among different populations are suggested to be a result of regional adaptation to diets with varying starch content (45).

1.3.1.3. Alcohol metabolism

Ethanol is oxidized to acetaldehyde by the enzyme alcohol dehydrogenase (*ADH*), and acetaldehyde is subsequently oxidized to acetic acid by aldehyde dehydrogenase

(*ALDH*). The metabolism of ethanol varies widely among individuals due to genetic variations in these two genes. One of the best studied polymorphisms influencing ethanol metabolism is *ADH1B* Arg47His (rs1229984). The derived allele *ADH1B**47His changes the pKa of alcohol dehydrogenase from 8.5 to 10.0, leading to 40 - 100 fold increase in K_m and V_{max} of alcohol metabolism (67). The global distribution of *ADH1B**47His reveals that this allele reaches high frequencies only in western and eastern Asia, but is nearly absent in other regions (68). Population differentiation test and long-range haplotype test both provide evidences of positive selection in East Asian populations (56, 67, 69). Interestingly, molecular dating suggests that the emergence of this allele (7,000~10,000 years ago) coincides with the origin of rice domestication in East Asia and that there is a strong correlation between *ADH1B**47 His frequency and the age of rice domestication (67). Thus, it was proposed that the rise of *ADH1B**47His frequency was an adaptation to rice domestication and the subsequent production and consumption of fermented food or beverages. *ADH1B**47His enables the rapid accumulation of acetaldehyde after alcohol consumption. Acetaldehyde is toxic and can cause a flushing reaction, which is proposed to be protective from alcohol overconsumption (67). Consistently, *ADH1B**47His has long been shown to be protective against alcoholism (70, 71).

1.3.2. Perception and bitter taste

Our perceptions of food, including vision, olfaction, and taste, shape our dietary preference and influence our health. Humans overall have remarkably reduced sensory capabilities compared with other mammals. In accordance with the phenotypic

reduction, genes related with olfaction, taste pheromone, and hearing have experienced relaxation of selective constraint and loss of function during human evolution (72-76). However, in contrast to this general trend, some perception-related genes are still targets of natural selection. The long-wavelength opsin responsible for red color vision was suggested to be under positive selection during human evolution (77). The gain of full trichromatic color vision was proposed to contribute to the deterioration of genes related to olfaction and pheromone because the reliance on visual signals to communicate social and reproductive status might reduce the importance of chemical signals (73, 78). Moreover, bitter-taste receptor gene *TAS2R38* has signatures of balancing selection (79) while *TAS2R16* is suggested to be target of positive selection (80).

Taste is of specific interest because it is the major determinant of food selection. Our perception of bitter taste is mediated by a highly variable family of seven-transmembrane G protein-coupled receptor genes (*TAS2Rs*), which had experienced family expansion by tandem gene duplication during mammalian evolution (81, 82). There are 38 putatively functional *TAS2R* genes and 5 pseudogenes with disrupted open reading frames (ORFs), organized in 3 clusters and located in chromosome 5, 7, and 12 (83). Population and comparative genetic analyses of 25 *TAS2R* genes among human and other primates suggest that this group of genes have undergone relaxation of selective constraints and in the process of loss of function through pseudogenization (75, 84). Evidences supporting relaxed selective constraints can be drawn from equal evolutionary rates of nonsynonymous to synonymous polymorphisms within human species, equal rates of nonsynonymous to synonymous substitutions between human

and other mammals, and the presence of nonsense mutation at medium frequency and pseudogenes (75, 84). It was suggested that the dietary shift at the early stage of human evolution with increased intake of animal-based food reduced the ingestion of plant toxins and the controlled use of fire helped detoxify poisonous compounds, so the importance of bitter-tasting was reduced and related genes were free to deteriorate (75).

In spite of the general trend of relaxed selection, one bitter taste receptor gene, *TAS2R16* shows signatures of adaptation. *TAS2R16* has an excess of high-frequency derived alleles as measured by Fay and Wu's H statistics in worldwide populations (80). The derived allele at one of its nonsynonymous mutations, K172N (rs846664), has been showed to increase sensitivity to toxic β-glucopyranosides, which are ubiquitous in nature and synthesized by over 2,500 plants and insects as a mean of protection against predators. Molecular dating revealed that this allele arose 78,700-791,000 years ago, in the Middle Pleistocene and before the expansion of early humans out of Africa. Thus it was suggested that K172N drove the positive selection of *TAS2R16* at the early stage of human evolution because it protected its carriers from consuming cyanogenic plant toxins (80). Another bitter taste receptor gene known to be under selection is *TAS2R38*, which is responsible for the classic phenotype of phenylthiocarbamide (PTC) tasting. *TAS2R38* has two dominating intermediate-frequency haplotypes corresponding to PTC -taster and -nontaster phenotypes. These two haplotypes have similar frequencies in African, European and Asian populations, and have little population differentiation (low Fst) among them, indicating similar evolutionary history. SFS-based tests (Tajima's D, Fu and Li's D and F) revealed an

excess of intermediate-frequency variants, suggesting a role of balancing selection. However, the selective pressures are yet to be found (79, 85).

1.4. Insight into nutritional practices from the evolutionary research

Evolutionary analyses of genetic adaptations to food and dietary-related behaviors not only complete our knowledge of human evolution but also increase our understanding of the genetic bases of complex traits or diseases and thus facilitate clinical and nutritional practices. Many genetic loci adaptive to specific food sources or dietary practices are also involved in pathological processes. For instance, the ancestral allele of K172N in *TAS2R16* contributes to higher risk of alcohol dependence because of its decreased sensitivity to bitter-taste stimuli, including alcoholic beverages (86, 87). Similarly, the non-taster genotype of *TAS2R38* reduces bitter sensitivity and correlates with higher alcohol consumption (87, 88). Furthermore, the understanding of human nature of bitterness avoidance as an adaptive mechanism in the evolutionary past emphasizes the importance of comprehensive examination of food properties and rational dietary choice because not all bitter substances are poisonous and harmful. For instance, plant-derived phytonutrients (phenols, flavonoids, isoflavones, terpenes, glucosinolates, etc.) are usually bitter and aversive to consumers, but they have chemoprotective effects against cancer and cardiovascular diseases. The bitter taste has been a barrier for the consumption of phytonutrients and renders them removed from products in food industry. The conflicting demands of taste and health call for innovative solutions to incorporate bitter but beneficial components into our diet (89).

More generally, evolutionary research provides an evolutionary perspective of complex diseases and nutritional practices. “Thrifty gene hypothesis” is one of the leading evolutionary theories explaining the explosion of metabolic syndromes, including type 2 diabetes and obesity, in modern society. This hypothesis proposed the existence of thrifty genes that promote more efficient food utilization, fat deposition and rapid weight gain during times of food abundance and confer their carriers higher chance of survival at times of famine. These genes were adaptive to the ancient hunter-gatherer lifestyle with frequent cycles of feast and famine, but are maladaptive in modern society with excess of energy availability, which can result in the prevalence of metabolic disorders (20). Another similar theory, the “Carnivore Connection Hypothesis” proposes that insulin resistance was probably advantageous to accommodate low glucose intake during the early human evolution when the Ice Ages dominated and restricted the ancient human diet to be low in carbohydrate and high in protein. However, the carbohydrate-rich diet after the advent of agriculture and the more recent high glycaemic index diet after industrial revolution rendered insulin resistance deleterious, contributing to the prevalence of diabetes (21, 22). In contrast to the dramatic societal, cultural and dietary transformations during the past 10,000 years, our genome remains almost the same as that of our Paleolithic ancestors. While our genome had been shaped by millions of years' selection to be adaptive to the dietary pattern in the Paleolithic era, it is maladaptive to our modern dietary habits. This maladaption of the genome to modern diet may be the underlying evolutionary cause of “civilization diseases”. Based on this theory, a dietary regimen called “the Paleolithic Diet” has been proposed to shift our diet to resemble that of our Paleolithic

ancestors to meet the limitation of our genome and to prevent us from complex diseases (90, 91). Although these theories about complex diseases and nutritional practices are still controversial (92), they provided valuable evolutionary perspectives and guidance for further investigations.

1.5. Future directions

The hunt for diet-adaptive genes is still at an early age. The rapid advance in sequencing technologies and the plummet of sequencing cost will stimulate an explosion of genomic information in the foreseeable future. Genome sequences from indigenous groups will be especially valuable in this endeavor because of their genetic homogeneity and characteristic dietary patterns. For instance, some indigenous populations, such as the Inuit and the Maasai, traditionally consuming an animal-based diet, will be great subjects for studying adaptation to high-fat/low carbohydrate diet. Generally, the increase of genomic sequences, coupled with an accumulation of dietary and phenotypic information, will provide an unprecedented opportunity to detect even small frequency shifts in response to environmental factors, including dietary components (59). The increase of population genomics data in non-human primates will also facilitate the elucidation of adaptation events at the early stage of human evolution. Another field of specific interest is the study of human gut microbiome. Investigation into the symbiotic relationship between gut microbiome and the human host will not only shed light on human dietary history but also uncover their co-adaptations to human diet (93, 94).

Research of human evolutionary adaptations to dietary changes has tremendous implications in human health, especially at the era of personal genomics. Incorporating this type of knowledge with genome-wide association study and functional investigations will provide comprehensive understanding of the evolutionary and molecular causes of complex metabolic diseases. Individualized nutritional practices

will be subsequently made possible to tailor optimal strategies for health management by taking into account personal genomic information.

1.6. Acknowledgements

The authors are grateful to Ryan A. Coots, Liuqi Gu for their valuable comments. This work was supported by startup funds from Cornell University, ILSI Future Leader Award in Nutrition, NSF grant DEB-0949556 and NIH 1R01AI085286-01 awarded to Dr. Zhenglong Gu.

1.7. References

1. Feero WG, Guttmacher AE, Collins FS. Genomic medicine--an updated primer. N Engl J Med. 2010;362:2001-11.

2. Luca F, Perry GH, Di Rienzo A. Evolutionary adaptations to dietary changes. Annu Rev Nutr. 2010;30:291-314.

3. Babbitt CC, Warner LR, Fedrigo O, Wall CE, Wray GA. Genomic signatures of diet-related shifts during human origins. Proc Biol Sci. 2011;278:961-9.

4. Kelley JL, Swanson WJ. Positive selection in the human genome: from genome scans to biological significance. Annu Rev Genomics Hum Genet. 2008;9:143-60.

5. Shi H, Su B. Molecular adaptation of modern human populations. Int J Evol Biol. 2011;2011:484769.

6. Akey JM. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. 2009;19:711-22.

7. Nielsen R. Molecular signatures of natural selection. Annu Rev Genet. 2005;39:197-218.

8. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. Nat Rev Genet. 2007;8:857-68.

9. Oleksyk TK, Smith MW, O'Brien SJ. Genome-wide scans for footprints of natural selection. Philos Trans R Soc Lond B Biol Sci. 2010;365:185-205.

10. 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;68:444-56.

11. White TD, Asfaw B, Beyene Y, Haile-Selassie Y, Lovejoy CO, Suwa G, WoldeGabriel G. Ardipithecus ramidus and the paleobiology of early hominids. Science. 2009;326:75-86.

12. Hernandez FM, Vrba ES. Plio-Pleistocene climatic change in the Turkana Basin (East Africa): evidence from large mammal faunas. J Hum Evol. 2006;50:595-626.

13. Heinzelin JD, Clark JD, White T, Hart W, Renne P, WoldeGabriel G, Beyene Y, Vrba E. Environment and behavior of 2.5-million-year-old Bouri hominids. Science. 1999;284:625-9.

14. Roebroeks W, Villa P. On the earliest evidence for habitual use of fire in Europe. Proc Natl Acad Sci U S A. 2011;108:5209-14.

15. Milton K. The critical role played by animal source foods in human (Homo) evolution. J Nutr. 2003;133:3886S-3892S.

16. White TD, Asfaw B, DeGusta D, Gilbert H, Richards GD, Suwa G, Howell FC. Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature. 2003;423:742-7.

17. Liu H, Prugnolle F, Manica A, Balloux F. A geographically explicit genetic model of worldwide human-settlement history. Am J Hum Genet. 2006;79:230-7.

18. Diamond J. Evolution, consequences and future of plant and animal domestication. Nature. 2002;418:700-7.

19. Fairweather-Tait SJ. Human nutrition and food research: opportunities and challenges in the post-genomic era. Philos Trans R Soc Lond B Biol Sci. 2003;358:1709-27.

20. NEEL JV. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet. 1962;14:353-62.

21. Miller JC, Colagiuri S. The carnivore connection: dietary carbohydrate in the evolution of NIDDM. Diabetologia. 1994;37:1280-6.

22. Colagiuri S, Brand MJ. The 'carnivore connection'--evolutionary aspects of insulin resistance. Eur J Clin Nutr. 2002;56 Suppl 1:S30-5.

23. Venter JC. Multiple personal genomes await. Nature. 2010;464:676-7.

24. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061-73.

25. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747-53.

26. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest. 2008;118:1590-605.

27. Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, Carter NP, Church DM, Felsenfeld A, Guyer M, Lee C, et al. Completing the map of human genetic variation. Nature. 2007;447:161-5.

28. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, et al. The diploid genome sequence of an Asian individual. Nature.

2008;456:60-5.

29. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10:451-81.

30. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. Global variation in copy number in the human genome. Nature. 2006;444:444-54.

31. Tishkoff SA, Verrelli BC. Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet. 2003;4:293-340.

32. Keinan A, Reich D. Human population differentiation is strongly correlated with local recombination rate. PLoS Genet. 2010;6:e1000886.

33. Lercher MJ, Hurst LD. Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet. 2002;18:337-40.

34. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. Whole-genome patterns of common DNA variation in three human populations. Science. 2005;307:1072-9.

35. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, et al. Darwinian and demographic forces affecting human protein coding genes. Genome Res. 2009;19:838-49.

36. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol. 2010;20:R208-15.

37. Hancock AM, Alkorta-Aranburu G, Witonsky DB, Di Rienzo A. Adaptations to new environments in humans: the role of subtle allele frequency shifts. Philos Trans R Soc Lond B Biol Sci. 2010;365:2459-68.

38. Pritchard JK, Di Rienzo A. Adaptation - not by sweeps alone. Nat Rev Genet. 2010;11:665-7.

39. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 2005;3:e170.

40. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science. 2003;302:1960-3.

41. Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. Promoter regions of many neural- and nutrition-related genes have experienced positive

selection during human evolution. Nat Genet. 2007;39:1140-4.

42. Hudson RR, Kreitman M, Aguade M. A test of neutral molecular evolution based on nucleotide data. Genetics. 1987;116:153-9.

43. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351:652-4.

44. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007;39:31-40.

45. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 2007;39:1256-60.

46. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123:585-95.

47. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993;133:693-709.

48. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000;155:1405-13.

49. Thompson EE, Kuttab-Boulos H, Witonsky D, Yang L, Roe BA, Di Rienzo A. CYP3A variation and the evolution of salt-sensitivity variants. Am J Hum Genet. 2004;75:1059-69.

50. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419:832-7.

51. Wang ET, Kodama G, Baldi P, Moyzis RK. Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci U S A. 2006;103:135-40.

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

53. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449:913-8.

54. Weir B S CCC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984;38:1358-70.

55. Weir BS, Hill WG. Estimating F-statistics. Annu Rev Genet. 2002;36:721-50.

56. Han Y, Gu S, Oota H, Osier MV, Pakstis AJ, Speed WC, Kidd JR, Kidd KK. Evidence of positive selection on a class I ADH locus. Am J Hum Genet. 2007;80:441-56.

57. Toomajian C, Kreitman M. Sequence variation and haplotype structure at the human HFE locus. Genetics. 2002;161:1609-23.

58. 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;74:1111-20.

59. Novembre J, Di Rienzo A. Spatial patterns of variation due to natural selection in humans. Nat Rev Genet. 2009;10:745-55.

60. Coop G, Witonsky D, Di Rienzo A, Pritchard JK. Using environmental correlations to identify loci underlying local adaptation. Genetics. 2010;185:1411-23.

61. Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, Sukernik R, Utermann G, Pritchard J, Coop G, et al. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proc Natl Acad Sci U S A. 2010;107 Suppl 2:8924-30.

62. Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, Di Rienzo A. Adaptations to climate-mediated selective pressures in humans. PLoS Genet. 2011;7:e1001375.

63. Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, Di Rienzo A. Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genet. 2008;4:e32.

64. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005;437:1153-7.

65. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I. Identification of a variant associated with adult-type hypolactasia. Nat Genet. 2002;30:233-7.

66. Enattah NS, Jensen TG, Nielsen M, Lewinski R, Kuokkanen M, Rasinpera H, El-Shanti H, Seo JK, Alifrangis M, Khalil IF, et al. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. Am J Hum Genet. 2008;82:57-72.

67. Peng Y, Shi H, Qi XB, Xiao CJ, Zhong H, Ma RL, Su B. The ADH1B Arg47His polymorphism in east Asian populations and expansion of rice

domestication in history. BMC Evol Biol. 2010;10:15.

68. Li H, Mukherjee N, Soundararajan U, Tarnok Z, Barta C, Khaliq S, Mohyuddin A, Kajuna SL, Mehdi SQ, Kidd JR, et al. Geographically separate increases in the frequency of the derived ADH1B*47His allele in eastern and western Asia. Am J Hum Genet. 2007;81:842-6.

69. Li H, Gu S, Cai X, Speed WC, Pakstis AJ, Golub EI, Kidd JR, Kidd KK. Ethnic related selection for an ADH Class I variant within East Asia. PLoS One. 2008;3:e1881.

70. Thomasson HR, Edenberg HJ, Crabb DW, Mai XL, Jerome RE, Li TK, Wang SP, Lin YT, Lu RB, Yin SJ. Alcohol and aldehyde dehydrogenase genotypes and alcoholism in Chinese men. Am J Hum Genet. 1991;48:677-81.

71. Chen CC, Lu RB, Chen YC, Wang MF, Chang YC, Li TK, Yin SJ. Interaction between the functional polymorphisms of the alcohol-metabolism genes in protection against alcoholism. Am J Hum Genet. 1999;65:795-807.

72. Gilad Y, Man O, Paabo S, Lancet D. Human specific loss of olfactory receptor genes. Proc Natl Acad Sci U S A. 2003;100:3324-7.

73. Gilad Y, Przeworski M, Lancet D. Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. PLoS Biol. 2004;2:E5.

74. Zhang J, Webb DM. Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates. Proc Natl Acad Sci U S A. 2003;100:8337-41.

75. Wang X, Thomas SD, Zhang J. Relaxation of selective constraint and loss of function in the evolution of human bitter taste receptor genes. Hum Mol Genet. 2004;13:2671-8.

76. Nance WE, Kearsey MJ. Relevance of connexin deafness (DFNB1) to human evolution. Am J Hum Genet. 2004;74:1081-7.

77. Verrelli BC, Lewis CJ, Stone AC, Perry GH. Different selective pressures shape the molecular evolution of color vision in chimpanzee and human populations. Mol Biol Evol. 2008;25:2735-43.

78. Liman ER, Innan H. Relaxed selective pressure on an essential component of pheromone transduction in primate evolution. Proc Natl Acad Sci U S A. 2003;100:3328-32.

79. Wooding S, Kim UK, Bamshad MJ, Larsen J, Jorde LB, Drayna D. Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. Am J Hum Genet. 2004;74:637-46.

80. Soranzo N, Bufe B, Sabeti PC, Wilson JF, Weale ME, Marguerie R, Meyerhof W, Goldstein DB. Positive selection on a high-sensitivity allele of the human bitter-taste receptor TAS2R16. Curr Biol. 2005;15:1257-65.

81. Shi P, Zhang J, Yang H, Zhang YP. Adaptive diversification of bitter taste receptor genes in Mammalian evolution. Mol Biol Evol. 2003;20:805-14.

82. Conte C, Ebeling M, Marcuz A, Nef P, Andres-Barquin PJ. Evolutionary relationships of the Tas2r receptor gene families in mouse and human. Physiol Genomics. 2003;14:73-82.

83. Bachmanov AA, Beauchamp GK. Taste receptor genes. Annu Rev Nutr. 2007;27:389-414.

84. Fischer A, Gilad Y, Man O, Paabo S. Evolution of bitter taste receptors in humans and apes. Mol Biol Evol. 2005;22:432-6.

85. Wooding S, Bufe B, Grassi C, Howard MT, Stone AC, Vazquez M, Dunn DM, Meyerhof W, Weiss RB, Bamshad MJ. Independent evolution of bitter-taste sensitivity in humans and chimpanzees. Nature. 2006;440:930-4.

86. Hinrichs AL, Wang JC, Bufe B, Kwon JM, Budde J, Allen R, Bertelsen S, Evans W, Dick D, Rice J, et al. Functional variant in a bitter-taste receptor (hTAS2R16) influences risk of alcohol dependence. Am J Hum Genet. 2006;78:103-11.

87. Wang JC, Hinrichs AL, Bertelsen S, Stock H, Budde JP, Dick DM, Bucholz KK, Rice J, Saccone N, Edenberg HJ, et al. Functional variants in TAS2R38 and TAS2R16 influence alcohol consumption in high-risk families of African-American origin. Alcohol Clin Exp Res. 2007;31:209-15.

88. Duffy VB, Davidson AC, Kidd JR, Kidd KK, Speed WC, Pakstis AJ, Reed DR, Snyder DJ, Bartoshuk LM. Bitter receptor gene (TAS2R38), 6-n-propylthiouracil (PROP) bitterness and alcohol intake. Alcohol Clin Exp Res. 2004;28:1629-37.

89. Drewnowski A, Gomez-Carneros C. Bitter taste, phytonutrients, and the consumer: a review. Am J Clin Nutr. 2000;72:1424-35.

90. Eaton SB, Konner M. Paleolithic nutrition. A consideration of its nature and current implications. N Engl J Med. 1985;312:283-9.

91. Eaton SB, Eaton SR, Konner MJ. Paleolithic nutrition revisited: a twelve-year retrospective on its nature and implications. Eur J Clin Nutr. 1997;51:207-16.

92. Speakman JR. Thrifty genes for obesity, an attractive but flawed idea, and an alternative perspective: the 'drifty gene' hypothesis. Int J Obes (Lond). 2008;32:1611-7.

93. Hehemann JH, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature. 2010;464:908-12.

94. Ley RE, Peterson DA, Gordon JI. Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell. 2006;124:837-48.

Chapter 2 – Human Expression QTLs are Enriched in Signals of Environmental Adaptation[2]

2.1. Abstract

Expression QTLs (eQTLs) have been found to be enriched in trait-associated SNPs. However, whether eQTLs are adaptive to different environmental factors and its relative evolutionary significance compared with non-synonymous SNPs (NS SNPs) are still elusive. Compiling environmental correlation data from three studies for more than 500,000 SNPs and 42 environmental factors, including climate, subsistence, pathogens and dietary patterns, we performed a systematic examination of the adaptive patterns of eQTLs to local environment. Compared with intergenic SNPs, eQTLs are significantly enriched in the lower tail of a transformed rank statistic in the environmental correlation analysis, indicating possible adaptation of eQTLs to the majority of 42 environmental factors. The mean enrichment of eQTLs across 42 environmental factors is as great as, if not greater than, that of NS SNPs. The enrichment of eQTLs, while significant across all levels of recombination rate, is inversely correlated with recombination rate, suggesting the presence of selective sweep or background selection. Further pathway enrichment analysis identified a number of pathways with possible environmental adaption from eQTLs. These pathways are mostly related with immune function and metabolism. Our results indicate that eQTLs might have played an important role in recent and ongoing human adaptation and are of special importance for some environmental factors and

[2] Published on *Genome Biology and Evolution*. See Appendix A for inclusion authorization

biological pathways.

2.2. Introduction

There has been a long-standing debate about the evolutionary significance of regulatory mutations and their prevalence in adaptation to local environments (1-5). Most of our current knowledge on the evolutionary patterns of mutations and their phenotypic consequences have been mainly gained from mutations in coding regions, whose identification and functional assessment are much easier than regulatory variation (1, 4, 6). However, evidence of regulatory adaptation during human evolution, including recent and ongoing adaptation to local environment, has been mounting (1). One classic example is the parallel adaptations of multiple regulatory mutations in *LCT* for milk consumption in North Europeans (7, 8), East Africans (9) and Saudi Arabians (10). Another case in point is the near fixation of a regulatory mutation in the promoter of *DARC* in the sub-Saharan African populations, which abolishes the expression of *DARC* in erythrocytes and benefits its homozygous carriers with complete resistance to malarial parasites, *Plasmodium vivax* (11, 12). Such studies have highlighted the importance of regulatory mutations in shaping human physiology, behavior and cognition (1, 13, 14).

The difficulty of accurate and comprehensive identification of regulatory mutations has hindered evolutionary studies on their functional importance. The recent expansion of expression quantitative trait loci (eQTLs) that are identified by the genome-wide association studies (GWAS) between genotypes and gene expression levels provides a unique opportunity for genome-wide evolutionary assessments of the underlying regulatory variants (15-22). As proxies of regulatory variants, eQTLs have

been shown to be enriched in SNPs associated with complex traits and diseases (6, 23, 24). However, the significance of genome-wide eQTLs in the evolutionary context has remained nearly unexplored. Preliminary analysis found that eQTLs tend to overlap with signals of incomplete selective sweeps as detected by "integrated haplotype score" (iHS) (25), warranting further verifications and systemic examinations with varying selection-detecting approaches, which capture different aspects or types of selection signals (26, 27).

Environmental correlation is a way of detecting adaptation by testing whether the spatial distribution of the frequency of an allele could be explained by an environmental factor. This selection-detection method has a special advantage of providing an ecological context that exerts selection pressure and is capable of detecting different types of selection, including hard and soft selective sweeps, and adaptation by subtle allele frequency shift (26-29). A specific method of environmental correlation based on a Bayesian linear model was recently developed (30) and applied to more than 500,000 SNPs and more than 35 environmental factors (27, 31, 32). This method effectively quantifies the correlation between allele frequency and environmental factor while controlling for population structure. For each SNP and each environmental factor, this method yields a Bayes Factor (BF), which indicates the strength of evidence that the environmental factor influences the frequency of the SNP in local populations. With this method, genic and non-synonymous (NS) SNPs have been shown to be enriched in signals of adaptation to a wide spectrum of environmental factors, including climate, subsistence, dietary patterns, and pathogens (27, 31, 32). This method, with the currently available eQTL

dataset, provides a great opportunity to investigate the importance of regulatory variants in recent and ongoing human environmental adaptation and to compare qualitatively and quantitatively their relative importance to NS SNPs. In this study, we conducted such analyses and present strong evidence that regulatory mutations played important roles in recent and ongoing human adaptation to local environment.

2.3. Material and Methods

Environmental correlation data. We collected the environmental correlation data from three large-scale studies, which span more than 550,000 SNPs and 42 environmental factors (36 individual environmental factors and 6 environmental categories, Figure 2.1 A and B). We referred to these three studies as Hancock *et al.* PNAS (27), Hancock *et al.* Plos (31) and Fumagalli *et al.* Plos (32), and labeled environmental factors from each study with prefixes of "1-", "2-", and "3-", respectively. Hancock *et al.* PNAS and Hancock *et al.* Plos used the same sample of 61 worldwide human populations, including 52 Human Genome Diversity Project panel populations, 4 HapMap phase III populations and 5 additionally genotyped populations. Hancock *et al.* PNAS includes 595,891 SNPs and 11 environmental factors (1-Polar domain, 1-Humid temperate domain, 1-Dry domain, 1-Humid tropical domain, 1-Foraging, 1-Horticulture, 1-Pastoralism, 1-Agriculture, 1-Cereals, 1-Roots and tubers, 1-Fats, meat, and milk). Hancock *et al.* Plos contains 623,318 SNPs and 11 environmental factors (2-Latitude, 2-Minimum temperature (Winter), 2-Maximum temperature (Summer), 2-Precipitation rate (Summer), 2-Precipitation rate (Winter), 2-Short wave radiation flux (Summer), 2-Short wave radiation flux (Winter), 2-Relative humidity (Summer), 2-Relative humidity (Winter), 2-Absolute latitude, 2-Longitude). Fumagalli *et al.* Plos used 55 human populations by joining data from the Human Genome Diversity Project and HapMap Phase III. It has 552,134 SNPs and 14 environmental factors (3-Distance from the sea; 3-Virus diversity; 3-Bacteria diversity; 3-Protozoa diversity; 3-Helminths diversity; 3-Relative humidity; 3-Temperature (annual mean); 3-Precipitation rate (annual mean); 3-Net short wave

radiation flux; 3-Gathering; 3-Hunting; 3-Fishing; 3-Animal husbandry; 3-Agriculture). The climate-related environmental factors in Fumagalli *et al.* Plos are annual means while those in Hancock *et al.* Plos are separated into Summer and Winter components. The subsistence-related environmental factors in Fumagalli *et al.* Plos are continuous, representing the percentage of time spent on a specific activity while those in Hancock *et al.* PNAS are binary.

All three studies applied the same Bayesian linear model method (30) to test the association between a SNP and an environmental factor. For each SNP and each environmental factor, this method yields a Bayes Factor (BF), which indicates the strength of evidence that the environmental factor influences the frequency of the SNP in local populations. To further examine the statistical significant of a BF, a transformed rank statistic (also known as empirical *p* value) was calculated based on its rank in BFs for a group of SNPs in the same ascertainment panel and within the same allele frequency bin. We obtained the data of transformed rank statistics for each SNP and environmental factors from dbCLINE (http://genapps2.uchicago.edu:8081/dbcline/main.jsp) or directly from the authors, Fumagalli *et al*. To summarize the evidence of association for each SNP with the six categories of variables (1-Ecoregion, 1-Subsistence, 2-Climate, 3-Subsistence, 3-Climate and 3-Pathogen), we followed the method used in the original studies by assigning the minimum of transformed rank statistics across all individual variables in the category (27, 31).

eQTLs data. eQTLs data were downloaded from http://eqtl.uchicago.edu/cgi-

bin/gbrowse/eqtl/ which is a compilation of eQTLs identified in eight large-scale studies in different human tissues (22). All eQTLs were used when the analyses were restricted to eQTLs itself. But when we did comparison between eQTLs and genic (or NS) SNPs, only eQTLs associated with RefSeq-supported protein-coding genes (see definition below) were used. We denoted these two groups of eQTLs as "eQTLs-all" and "eQTLs-for comparison", respectively.

Definitions of genomic regions. Gene annotations based on hg18 were downloaded from UCSC Genome Browser database. Only autosomal genes were retrieved because environmental correlation data were only available for autosomal SNPs. In total, there were 24,814 autosomal genes, 18,396 of which were coding genes while the rest were non-coding genes. For genes with multiple isoforms, the longest transcript was used. If there were multiple transcripts of the same length, one was arbitrarily chosen. Based on these 24,814 autosomal genes, we defined intergenic SNPs as those at least 50 kb away from known genes, either coding or non-coding. There were 16,914 autosomal coding genes with support from RefSeq and thus they were used in the definition of genic SNPs and NS SNPs. Genic SNPs were defined as those within 5 kb of a gene. Annotation to SNPs is also downloaded from UCSC. A SNP is called NS based on the longest transcript used in the analysis. To compare the enrichment ratio of eQTLs with that of genic SNPs (or NS SNPs), only eQTLs for these 16,914 genes were used. Further, to control for the potential difference between genes with and without eQTLs, we restricted this comparison for only genes with eQTLs. We denoted genic SNPs for this group of gene as "e-genic SNPs" and NS SNPs as "e-NS SNPs". All data are available upon request.

Enrichment of eQTLs, genic and NS SNPs in the lower tail of the transformed rank statistic distribution. To examine if there is an excess of tested SNPs in the lower tail of the transformed rank statistic distribution, which is the empirical distribution of the transformed rank statistics for genome-wide SNPs, we calculated an enrichment ratio (ER), which is the ratio of the percentage of tested SNPs in the lower tail to the corresponding percentage of intergenic SNPs. The tested SNPs could be genic SNPs, NS SNPs or eQTLs while the intergenic SNPs were treated as the neutral control. Although not all intergenic SNPs are neutral, the presence of adaptive SNPs in the control could only lead to a smaller ER, which will make our results more conservative. An ER larger than 1 indicates that there is an excess of tested SNPs in the lower tail compared with intergenic SNPs. To avoid the arbitrary choice of cutoffs in the definition of the lower tail, we used three cutoffs, 5%, 1% and 0.5%. To assess the significance of an observed excess, we applied a block bootstrap resampling procedure which corrected for the non-independence of nearby loci because of linkage disequilibrium (LD). To this end, we broke the genome into 500-kb non-overlapping segments and bootstrap resampled a number of segments to make a pseudo-genome of the same length as the real genome. For the pseudo-genome, we calculated the enrichment ratio just as we did for the empirical genome. We performed 1000 bootstrap replicates to obtain a confidence interval for the observed excess. We considered an observed excess significant if at least 95% of the bootstrap replicates produced enrichment ratios larger than 1.

One specific feature for eQTLs, different from genic SNPs and NS SNPs, is that they tend to cluster together because they are identified with association analysis. If the

clustering of eQTL overlaps with the clustering of significant correlative signals indicated by the Bayes Factor, the block bootstrap may not be powerful to control for this effect because each time a block of SNPs are drawn rather than a single SNP. To totally break down the association among eQTLs for a gene, we performed the following random simulation. For genes with multiple eQTLs, only one eQTL was randomly chosen and used for the calculation of the enrichment ratio. We repeated this procedure 1000 times and we considered an observed excess significant if 95% of the simulations reveal enrichment ratio larger than 1. This random simulation procedure gave essentially the same pattern of significance and thus for most analyses we only used the block bootstrap procedure.

Enrichment of eQTLs, genic and NS SNPs in the higher tail of prediction accuracy (Q^2) in relative to intergenic SNPs. Prediction accuracy data were obtained from Fumagalli *et al.* and data were available for 552,136 SNPs for each of four categories of environmental factors (all environmental factors, climate, subsistence and pathogen) (32). Following the method used in the original study, we divided the distribution of Q^2 into bins. The number of bins was chosen to ensure that each bin has a large enough number for block bootstrap and the calculation of enrichment. Enrichment here is defined as the proportion of tested SNPs (genic, NS, eQTLs and intergenic) in the bin divided by the total proportion of tested SNPs. Intergenic SNPs are used as control for other tested SNPs (genic, NS, and eQTLs). For example, when calculating the enrichment for eQTLs in comparison to intergenic SNPs, the total proportion of eQTLs is the number of eQTLs divided by the total number of eQTLs and intergenic SNPs while the proportion of eQTLs in a specific bin is the number of

eQTLs in the bin divided by the total number of eQTLs and intergenic SNPs in the bin. We applied the same block bootstrap procedure as described in the previous section to estimate the confidence interval of each enrichment value.

Statistical tests comparing ERs. To compare the enrichment ratios of different groups of SNPs, we did the analysis at two levels: for each individual environmental factor/category and for all environmental factors. When doing analysis for a specific environmental factor, we used ERs from 1,000 bootstraps. For example, when we compared eQTLs with NS SNPs for the climate category, we had 1,000 pairs of ERs from 1,000 bootstraps and we performed a paired one-tailed two-sample t test.

When doing the analysis for all environmental factors as a whole to detect the general trend, we cannot directly apply t test because observations for different environmental factors are not independent of each other. Therefore to test if there is difference between the means of ERs for two groups of SNPs, we developed a generalized paired Z-test by incorporating the correlations among different environmental factors. The test was developed as follows. Let Y_{ij} denote the response for environmental factor $j = 1, \dots, n$ in group $i = 1, 2$. In our case, $n = 42$. We can apply a mixed effect model:

$$Y_{ij} = u + G_i + E_j + e_{ij}$$

where G_i is the fixed effect of group i, E_j is the random effect of environmental factor j and e_{ij} is the random error. The random errors for environmental factors within the same group are correlated. With the assumption of multivariate normal

distribution, we have $e_{ij} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{R})$. The correlation matrix ($\boldsymbol{R}$) among different environmental factors was estimated from all SNPs with the transformed rank statistics for all 42 environmental factors. By pairing G_{1j} with G_{2j} to take into account the dependence between the two observations for each environmental factor, we have a simpler model:

$$D_j = \delta + \epsilon_j$$

where $\delta = G_2 - G_1$ and $\epsilon_j = e_{2j} - e_{1j}$ with $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \sim N(\mathbf{0}, 2\sigma^2 \boldsymbol{R})$. The observed differences for each environmental factor are $\boldsymbol{D} = (D_1, \ldots, D_n)$. The mean difference $\overline{D}$ has variance:

$$var(\overline{D}) = \frac{1}{n^2}\left(\sum_{j=1}^{n} var(D_j) + \sum_{j \neq k}^{n} cov(D_j, D_k) \right) = \frac{2\sigma^2}{n} + \frac{2\sigma^2}{n^2} \sum_{j>k}^{n} r_{jk}$$

δ and σ^2 can be estimated using maximum likelihood estimation from observed differences $\boldsymbol{D}$ while taking into account correlations among environmental factors. Assuming multivariate normal distribution for $\boldsymbol{D}$, we have the following probability density function:

$$f(\boldsymbol{D}) = \frac{1}{(2\pi)^{n/2} |2\sigma^2 \mathbf{R}|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{D} - \mathbf{1}\delta)^T (2\sigma^2 \mathbf{R})^{-1} (\boldsymbol{D} - \mathbf{1}\delta) \right)$$

Where **1** is a vector of n 1s. By taking derivatives of the log-likelihood function with respect to δ and σ^2 and setting the resulted derivatives to be zero, we solved the equations and got:

$$\hat{\delta} = \frac{\mathbf{1}^T \boldsymbol{R}^{-1} \boldsymbol{D}}{\mathbf{1}^T \boldsymbol{R}^{-1} \mathbf{1}} \text{ and } \hat{\sigma}^2 = \frac{1}{2n} \left(\boldsymbol{D} - \mathbf{1}\hat{\delta}\right)^T \boldsymbol{R}^{-1} (\boldsymbol{D} - \mathbf{1}\hat{\delta})$$

A test for difference between two groups can be based on a Z statistic,

$$Z = \frac{\overline{D}}{SE_{\overline{D}}}$$

Where $\overline{D} = \hat{\delta}$ and $SE_{\overline{D}} = \sqrt{var(\overline{D})} = \sqrt{\frac{2\sigma^2}{n}(1 + \frac{1}{n}\sum_{j>k}^{n} r_{jk})}$. The p value of an observed Z statistic was found from a standard normal distribution.

Controlling for recombination rate. Recombination rate data were downloaded from Hapmap (/downloads/recombination/2008-03_rel22_B36/rates). These data were calculated from 4 continental populations and thus were only approximation of recombination rates in global populations. Combining three datasets (Hancock *et al.* PNAS, Hancock *et al.* Plos, and Fumagalli *et al.* Plos), there were 639,663 SNPs, out of which 633,340 had point estimate of recombination rate. These SNPs were grouped into 5 equal-sized bins, corresponding to median recombination rate of 0.0354, 0.187, 0.445, 1.016, and 4.236 cM/Mb. 5 bins were chosen to ensure that for each type of variation, there was enough number of SNPs in each bin for the calculation of enrichment ratio. For different types of variations, the numbers of SNPs for each bin were different. Each bin of recombination rate was denoted as the log(median recombination rate). The proportion of significant SNPs in the lower tail (defined with three cutoffs, 5%, 1%, and 0.5%) was calculated for each bin and each type of variation, including intergenic SNPs. The enrichment ratio of a specific type of

variation (genic, NS, or eQTLs) for each environmental factor in each bin of recombination rate was calculated using corresponding intergenic SNPs in the bin as control. To demonstrate the general trend of the effect of recombination rate on the proportion of significant SNPs (or ER), simple linear regressions were performed between the median of recombination rate and the mean of the proportion of significant SNPs (or the mean of ER) in each bin.

In addition to using population-based estimates of recombination rate, we further confirmed our analysis with a pedigree-derived sex-averaged recombination map from deCODE (33), from which recombination rate for each SNP was calculated with a 500 Kb window centered on the SNP.

Canonical pathway enrichment. To determine whether there is an enrichment of eQTLs in a canonical pathway, for each environmental factor (or category), we did the following analyses. For each environmental factor, we calculated ERs separately for both eQTLs and genic SNPs. However, it is of note that ERs calculated here is different from those for genome-wide pattern. The control we used here was not intergenic SNPs. Rather, for eQTLs we used all other eQTLs not in the pathway as control, and for genic SNPs we used all other genic SNPs not in the pathway. We did the analyses for three different cutoffs (5%, 1%, and 0.5%). The significance of enrichment was assessed with 1000 block bootstraps as described before. Pathways of interest are those that have significant enrichment signals for eQTLs across three cutoffs but have no consistent significant enrichment signals for genic SNPs.

Figure 2.1. eQTLs are enriched in signals of environmental adaptation. The enrichment ratios (ERs) of eQTLs to intergenic SNPs in the 5% tail of the transformed rank statistics for (A) 36 individual environmental factors and (B) six environmental categories. Mean and standard error for each enrichment ratio are estimated from 1,000 whole-genome block bootstraps. Six environmental categories are defined by grouping individual environmental factors as indicated at the left. Colors of blue, orange, magenta and green represent ecoregion, subsistence, climate, and pathogen, respectively. (C) The box plot of empirical ERs of eQTLs to intergenic SNPs in the tail of the transformed rank statistics for 42 environmental factors under three tail cutoffs. The ERs for each environmental factor under three cutoffs are connected with dashed lines. Environmental categories are represented by dash lines with the same colors as Fig. 1B. The red squares, connected by a red line, are the means of all 42 empirical ERs under three tail cutoffs. The p values are for generalized paired Z tests of whether eQTLs with more stringent tail cutoffs have higher mean ERs. (D) Progressive enrichment of eQTL SNPs in comparison to intergenic SNPs for increasing values of prediction accuracy (Q^2) of 14 environmental factors. Red line indicates the median value of enrichment of eQTLs from 1,000 block bootstraps while blue line is that for intergenic SNPs. Pink region denotes the 90^{th} confidence interval for eQTL enrichment.

A

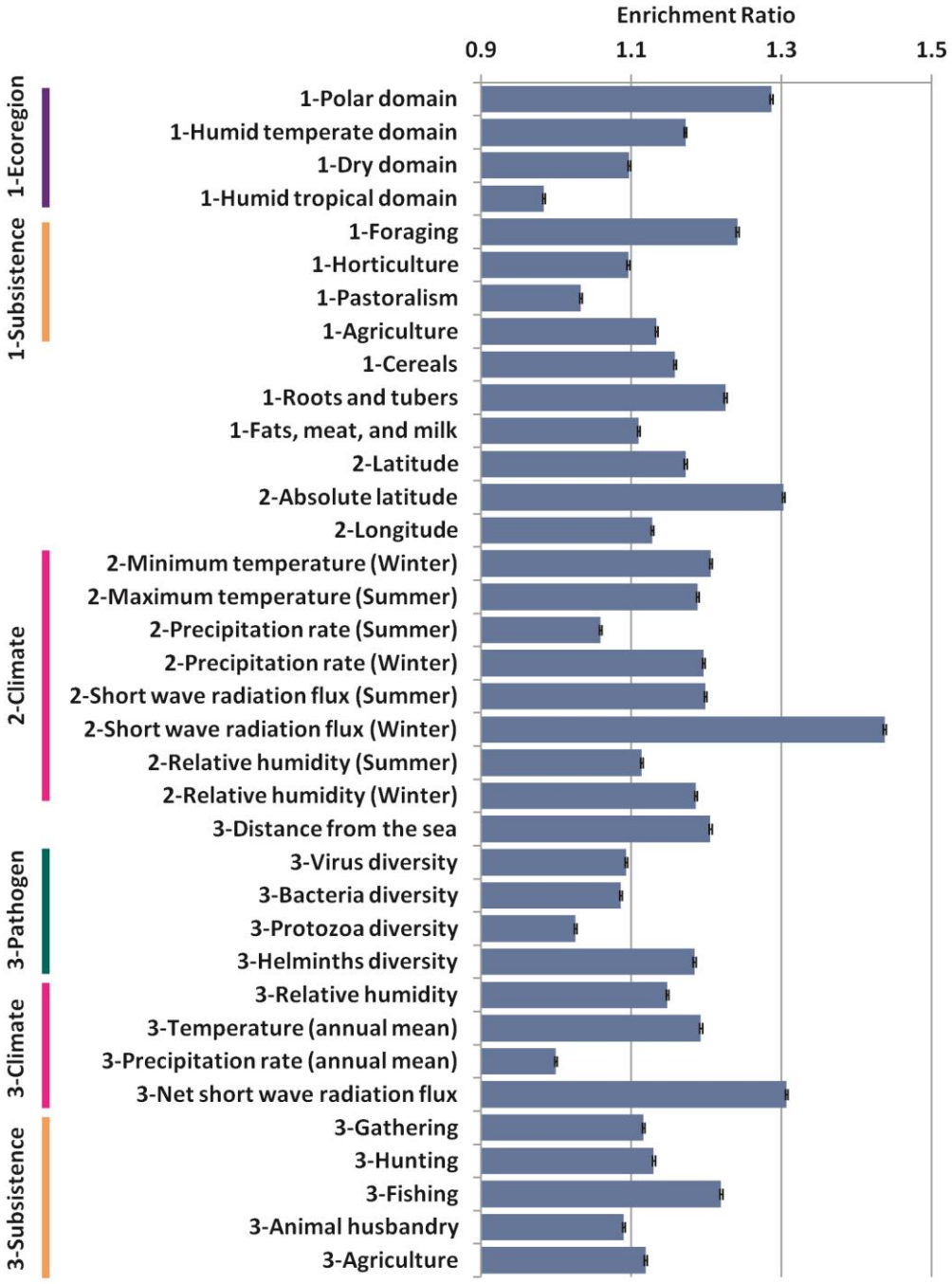
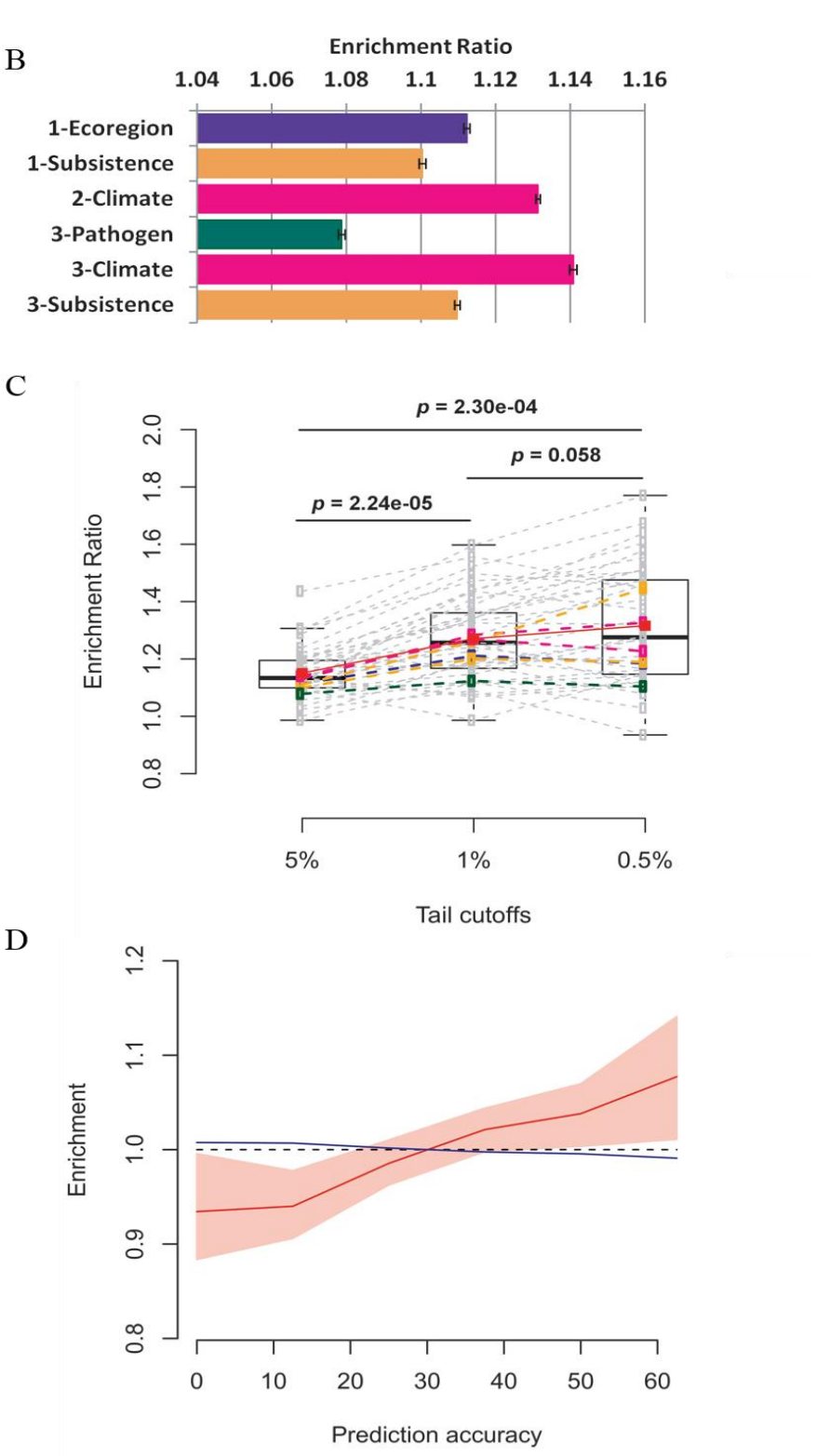

Enrichment Ratio
0.9
1.1
1.3
1.5
1-Ecoregion
1-Subsistence
2-Climate
3-Pathogen
3-Climate
3-Subsistence
1-Polar domain
1-Humid temperate domain
1-Dry domain
1-Humid tropical domain
1-Foraging
1-Horticulture
1-Pastoralism
1-Agriculture
1-Cereals
1-Roots and tubers
1-Fats, meat, and milk
2-Latitude
2-Absolute latitude
2-Longitude
2-Minimum temperature (Winter)
2-Maximum temperature (Summer)
2-Precipitation rate (Summer)
2-Precipitation rate (Winter)
2-Short wave radiation flux (Summer)
2-Short wave radiation flux (Winter)
2-Relative humidity (Summer)
2-Relative humidity (Winter)
3-Distance from the sea
3-Virus diversity
3-Bacteria diversity
3-Protozoa diversity
3-Helminths diversity
3-Relative humidity
3-Temperature (annual mean)
3-Precipitation rate (annual mean)
3-Net short wave radiation flux
3-Gathering
3-Hunting
3-Fishing
3-Animal husbandry
3-Agriculture

Figure 2.1. (Continued)



2.4. Results

We extracted environmental correlation data from three large-scale studies, which span more than 550,000 SNPs and 42 environmental factors (36 individual environmental factors and 6 environmental categories, Figure 2.1. A and B). We referred to these three studies as Hancock *et al.* PNAS (27), Hancock *et al.* Plos (31) and Fumagalli *et al.* Plos (32), and labeled environmental factors from each study with prefixes of "1-", "2-", and "3-", respectively. Climate and subsistence related environmental factors are included in two of the studies but they represent different aspects of the same environmental variables. For instance, climate-related factors in Study 2 are separated into winter and summer components while those in Study 3 are annual means. Subsistence-related factors in Study 1 are treated as binary variables while those in Study 3 are continuous and representing the percentage of time a population spent on a specific activity. For each SNP, we retrieved its transformed rank statistic (also known as empirical *p* value), for each of the 42 environmental factors. Although the transformed rank statistics of a SNP are correlated among environmental factors, to fully exploit the information available, we included all 42 environmental factors in our analysis while accounting for their correlations. A statistic called enrichment ratio (ER), the ratio of the proportion of tested SNPs to that of intergenic neutral SNPs in the lower tail of the transformed statistic, was used to test if a group of SNPs is enriched in signals of adaptation to a specific environmental factor (27, 31). Groups of SNPs important for adaptation will be enriched in the lower tail and have ERs larger than 1. The more important a group of SNPs is in adaptation, the larger the ER.

We combined eQTLs that are identified in eight large-scale studies in different human tissues (15-22). By integrating the eQTLs data and the environmental correlation data, we examined whether eQTLs are enriched in signals of environmental correlation at the genome-wide level and compared their enrichment pattern to that of NS SNPs.

2.4.1. Genome-wide enrichment of eQTLs in environmental correlation

The ERs of eQTLs to intergenic SNPs were calculated for each of the 36 individual environmental factors and 6 environmental categories. The significance of each ER is assessed by a whole-genome block bootstrap procedure. To avoid the arbitrary use of tail cutoff, we used 5%, 1%, and 0.5% to define the lower tails of the statistic distribution. For the 36 individual environmental factors (Figure 2.1. A, Table 2.1.), 32 (89%) have ER significantly larger than 1 under at least one tail cutoff and 18 (50%) have significant ERs across all three tail cutoffs. Under the tail cutoff of 5%, short wave radiation flux in winter has the largest ER of 1.44, suggesting that the enrichment of eQTLs in the 5% lower tail is 44% higher than neutral expectation. Consistently, the annual mean of net short wave radiation flux from Study 3 has the second largest ER of 1.31, 31% higher than neutral expectation. Individual factors in categories other than climate also exhibit significant ERs, such as polar domain (1.29), foraging (1.24), helminths diversity (1.18), roots and tubers (1.22) and absolute latitude (1.30). Taken together, eQTLs are observed to be enriched in the lower tail of the transformed rank statistics for a majority of environmental factors examined here, including ecoregion, climate, subsistence, pathogen and dietary patterns.

Table 2.1. The enrichment of eQTLs in all environmental categories/factors.Three tail cutoffs, 5%, 1% and 0.5%, were used. The level of significance was estimated by whole-genome block bootstrap. Red and orange indicates respectively >99% and >95% of bootstrap replicates having ER values larger than 1.

Environmental Categories/Factors	eQTLs:Neutral		
	5%	1%	0.5%
3-Climate	1.1406	1.2691	1.2268
3-Net short wave radiation flux	1.3065	1.4726	1.5038
3-Temperature (annual mean)	1.1894	1.2569	1.1024
3-Relative humidity	1.1483	1.1843	1.2093
3-Precipitation rate (annual mean)	0.9995	1.1948	1.142
2-Climate	1.1319	1.284	1.3266
2-Short wave radiation flux (Winter)	1.4366	1.5559	1.4475
2-Minimum temperature (Winter)	1.2057	1.3326	1.4065
2-Short wave radiation flux (Summer)	1.2004	1.3515	1.5135
2-Precipitation rate (Winter)	1.1949	1.4361	1.3725
2-Maximum temperature (Summer)	1.1906	1.4418	1.5538
2-Relative humidity (Winter)	1.1858	1.346	1.4773
2-Relative humidity (Summer)	1.1134	1.2807	1.4606
2-Precipitation rate (Summer)	1.058	1.1153	1.141
1-Ecoregion	1.1136	1.2112	1.1842
1-Polar domain	1.2887	1.5238	1.6738
1-Humid temperate domain	1.1698	1.1774	1.0785
1-Dry domain	1.0993	1.2055	1.1919
1-Humid tropical domain	0.9861	1.116	1.1018
3-Subsistence	1.1109	1.2616	1.447
3-Fishing	1.2219	1.3886	1.6412
3-Hunting	1.1325	1.3412	1.5805
3-Agriculture	1.1205	1.4114	1.5115
3-Gathering	1.1139	1.4411	1.6061
3-Animal husbandry	1.0915	0.9859	1.1466
1-Subsistence	1.1003	1.2004	1.1895
1-Foraging	1.2389	1.4975	1.452
1-Agriculture	1.1346	1.0678	1.1056
1-Horticulture	1.0965	1.117	1.0284
1-Pastoralism	1.0348	1.1878	1.2684
3-Pathogen	1.0783	1.1232	1.1038
3-Helminths diversity	1.1834	1.2409	1.1648
3-Virus diversity	1.0937	1.1475	1.2714
3-Bacteria diversity	1.0845	1.119	1.162
3-Protozoa diversity	1.0253	1.0893	0.935
Diet			
1-Roots and tubers	1.2243	1.1673	1.2797
1-Cereals	1.1574	1.2269	1.3386
1-Fats, meat, and milk	1.1086	1.2567	1.2327
Others			
2-Absolute latitude	1.3015	1.5976	1.7703
3-Distance from the sea	1.2094	1.3253	1.3211
2-Latitude	1.1695	1.3607	1.4756
2-Longitude	1.1254	1.0756	1.1201

For all six environmental categories examined (Figure 2.1. B, Table 2.1.), the ERs of eQTLs are consistently and significantly larger than 1. Under the tail cutoff of 5%, when summarizing over the summer and winter components of all related environmental factors from Study 2, the climate category has an empirical ER of 1.13. Consistently, when summarizing over the annual means of all related factors from Study 3, the empirical ER for the climate category is 1.14. For the subsistence category, when all related factors are treated as binary, the summarized ER is 1.10 and it is 1.11 when all related factors are treated as continuous. Moreover, the ERs for the categories of ecoregion and pathogens are 1.11 and 1.08, respectively. Over all 42 environmental factors, the mean empirical ERs of eQTLs are 1.15, 1.27, and 1.31, respectively, for the three tail cutoffs (Figure 2.1. C). They are significantly larger than 1 (p values are 1.02×10^{-13}, 3.36×10^{-14}, and 3.09×10^{-10}, respectively with generalized one-tailed paired Z tests). As previously observed for genic and non-synonymous SNPs (27, 31), ERs of eQTL SNPs increase with more stringent tail cutoffs, suggesting that in general eQTLs are enriched in signals of environmental adaptation.

One observation worth pointing out is that for the same type of climate-related environmental factors, such as temperature, humidity and precipitation rate, separating them into summer and winter components provides stronger evidence of enrichment for eQTLs. For instance, while the annual mean of temperature has significant eQTLs enrichment under two tail cutoffs, the minimum temperature in winter and the maximum temperature in summer have significant signals over all three tail cutoffs (Table 2.1.). Another interesting case is precipitation. While the annual mean of precipitation rate has significant signal only in the 1% tail, the precipitation rate in

winter has significant enrichment under all three cutoffs and the precipitation rate in summer has no significant results (Table 2.1.). This pattern suggests the presence of season-specific selection pressures from different environment factors during human evolution.

Another statistic, prediction accuracy (Q^2), from a different implementation of environmental correlation provides us an opportunity to verify our observation that eQTLs are enriched in signals of environmental adaptation. Prediction accuracy, developed by Fumagalli *et al.* (32), is a measure of how well a group of environmental variables predict the global frequency distribution of an allele. The prediction accuracy for SNPs that are adaptive to the tested environmental factor is expected to be higher than neutral SNPs. We retrieved prediction accuracy data for more than 550,000 SNPs for 14 climate, subsistence and pathogen –related environmental factors that were used in Fumagalli *et al.* (32). When all environmental factors were combined as a single predictor, we observed a significant enrichment of eQTLs compared to intergenic SNPs in the highest bin of prediction accuracy (Figure 2.1. D). The median value for the re-sampled distribution of the enrichment statistic is 1.08 ($p < 0.05$). These observations verified that environmental adaptation has shaped the differential allele frequency distribution of eQTLs among human populations. If these 14 environmental factors are separated into three categories (climate, subsistence, pathogen), we observed a similar trend of progressive enrichment of eQTLs for higher values of prediction accuracy.

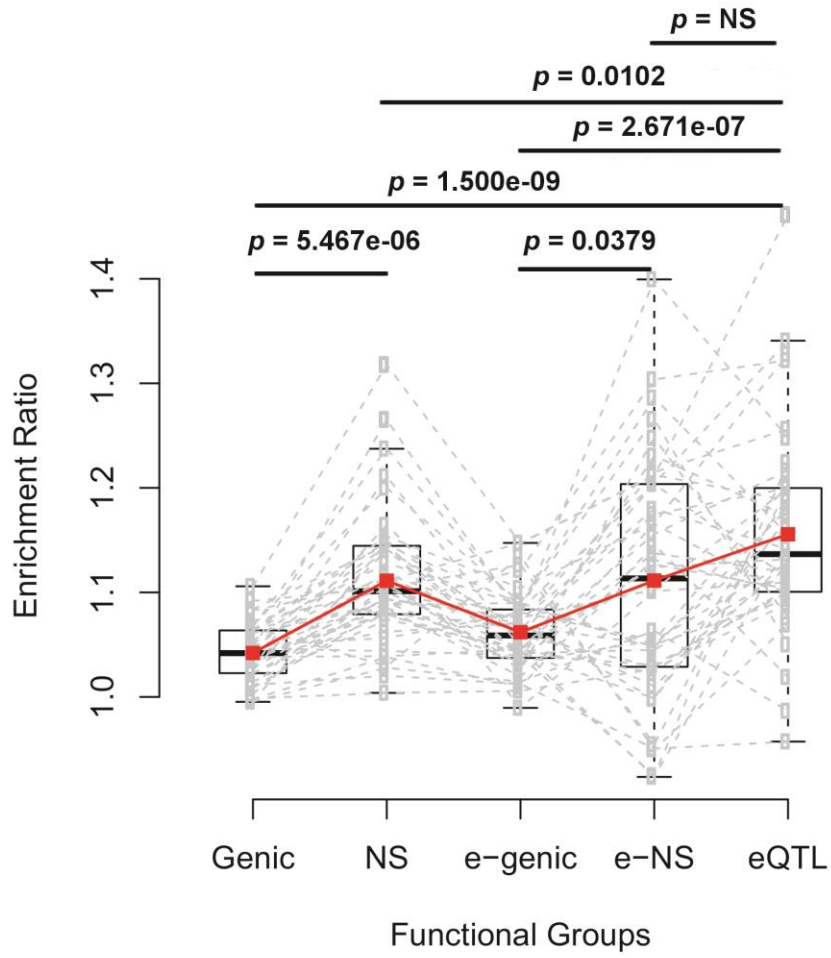


Figure 2.2. Enrichment ratios of different types of SNPs. Five types of SNPs were analyzed, including genome-wide genic and NS SNPs (Genic and NS), genic and NS SNPs for genes with eQTLs (e-genic and e-NS), and eQTLs whose associated genes were included in our analyses. For each type of SNPs, ERs were calculated for all 42 environmental factors. Each dashed line connects ERs calculated from one environmental factor. The red line connects mean ERs of each SNP group. The p values are for generalized paired Z tests of whether the mean ERs of one group is larger than that of the other group. The tail cutoff is 5%. The patterns are similar for cutoffs of 1% and 0.5%.

2.4.2. eQTLs and NS SNPs in environmental adaptation

Our data allowed us to investigate whether regulatory variations are qualitatively and quantitatively distinct from coding variations in environmental adaptations. To compare the relative prevalence of eQTLs and NS SNPs in environmental adaptation, we tested if the ER of eQTLs is significantly larger than that of NS SNPs. As shown in Figure 2.2., under the tail cutoff of 5%, the ER of eQTLs is 1.16, which is significantly higher than that of genic SNPs, 1.04 (p = 1.500e-09). The mean ER of eQTLs is also significantly higher than that of genome-wide NS SNPs, 1.11 (p = 0.0102). Because not all genes have eQTLs, it is possible that genes with eQTLs may have different ERs from genome-wide patterns, thus leading to a biased comparison. To correct for this, we further calculated ERs for genic and NS SNPs only for the genes with eQTLs (denoted as e-genic and e-NS SNPs). Under the same cutoff the comparison between eQTLs and e-NS SNPs supports the above trends but the ER difference between these two categories of SNPs is not statistically significant. Similar patterns were observed under the other two tail cutoffs (1% and 0.5%). Taken together, these results indicate that in general eQTLs are as prevalent as, if not more prevalent than, NS SNPs in environmental adaptation.

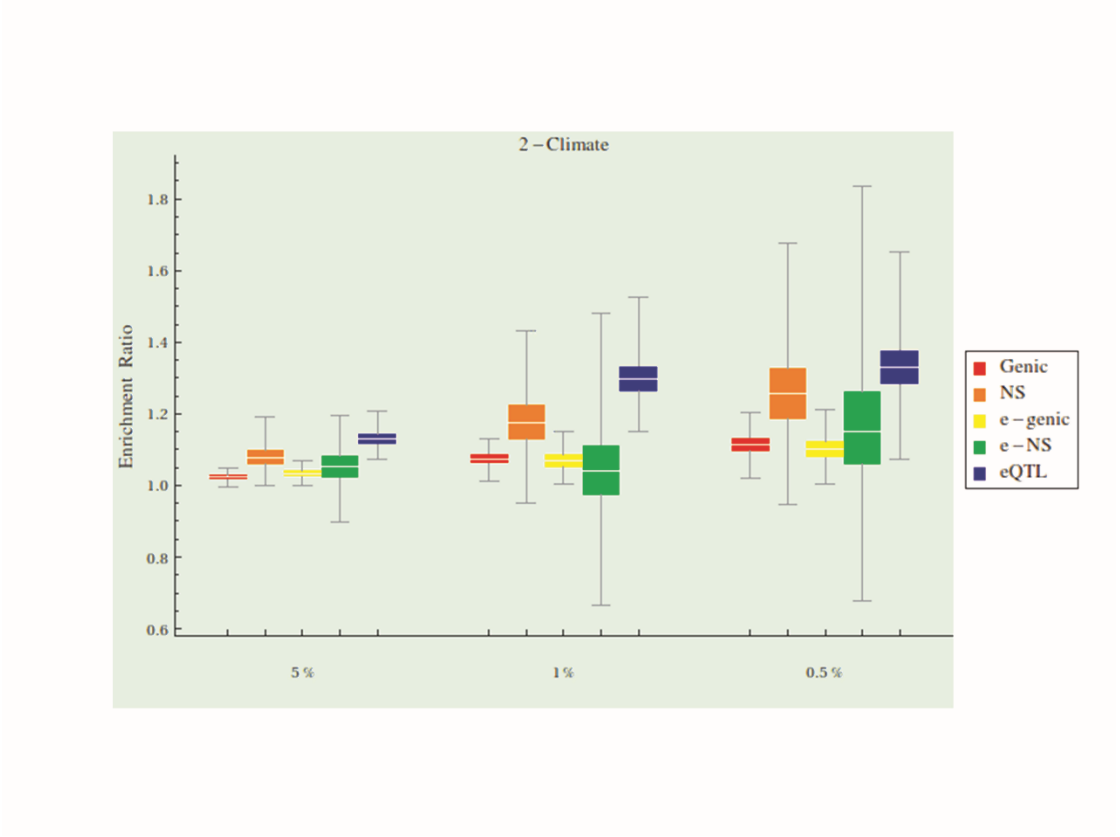


Figure 2.3. eQTLs are more likely to be adaptive than NS SNPs for climate. This is a box plot of enrichment ratios from 1,000 whole-genome block bootstraps for five groups of SNPs under three tail cutoffs. Enrichment was examined at the level of each environmental factor/category. “2-Climate” is presented here. Another seven categories/factors were also significant, including cereals, latitude, short wave radiation flux in summer, short wave radiation flux in winter, relative humidity in winter, virus diversity, and fishing.

We further identified environmental factors to which eQTLs are more likely to be adaptive than NS SNPs. To this end, we required that the ER of eQTLs be significantly larger than those of genic, NS, e-genic, and e-NS SNPs across all three tail cutoffs. Among 42 environmental factors examined, eight have consistent and significant signals (Figure 2.3.). They include cereals, latitude, short wave radiation flux in summer, short wave radiation flux in winter, relative humidity in winter, virus diversity, fishing and the climate category summarizing over seasonal components. Although these eight factors encompass categories of climate, subsistence, pathogens, and dietary patterns, four of them are climate-related and climate is the only environmental category identified. It is noteworthy that these climate-related factors are continuous, as are the other three environmental factors except cereals. The prevalence of eQTLs over NS SNPs in adaptation to continuous environmental factors underlies the possibility that gene expression, because of its continuous nature, is more suitable than protein function to be fine-tuned to meet the dynamic range of continuous environmental factors (1).

2.4.3. Greater enrichment of adaptive eQTLs in regions of low recombination

Signals of background selection and selective sweep tend to extend longer in genomic regions of low recombination. Therefore, in those regions, the selective signal on a causal locus is more likely to be captured by nearby linked loci (34, 35). If the underlying regulatory variations tagged by eQTLs are adaptive, eQTLs in region of low recombination are expected to show stronger footprint of adaptation. To explore the effect of recombination rate on the ERs of eQTLs, we divided all SNPs into five

bins of equal number of SNPs based on their recombination rates and calculated ERs for each of 42 environmental factors in each bin. Linear regression analyses were performed between the natural logarithm of median recombination rate and the mean of ERs over 42 environmental factors. Significant inverse correlations were observed between local recombination rate and ERs of eQTLs SNPs (Figure 2.4.). Similar results are also observed for the genic, NS and synonymous SNPs.

Over all 42 environmental factors, the mean ERs of eQTLs in the 5% tail are significantly larger than 1 across all five bins of recombination rate (Figure 2.4.). Similar patterns are observed under the other two cutoffs. Furthermore, we compared the ERs of different types of SNPs in the five recombination rate bins. Considering the significance over all three tail cutoffs, ERs of eQTLs are larger than that of genic SNPs for the first three recombination rate bins while ERs of NS SNPs are larger than that of genic SNPs for the first two bins, indicating the genome-wide patterns of higher ERs of eQTLs and NS SNPs than genic SNPs are mainly driven by SNPs from the regions of low recombination.

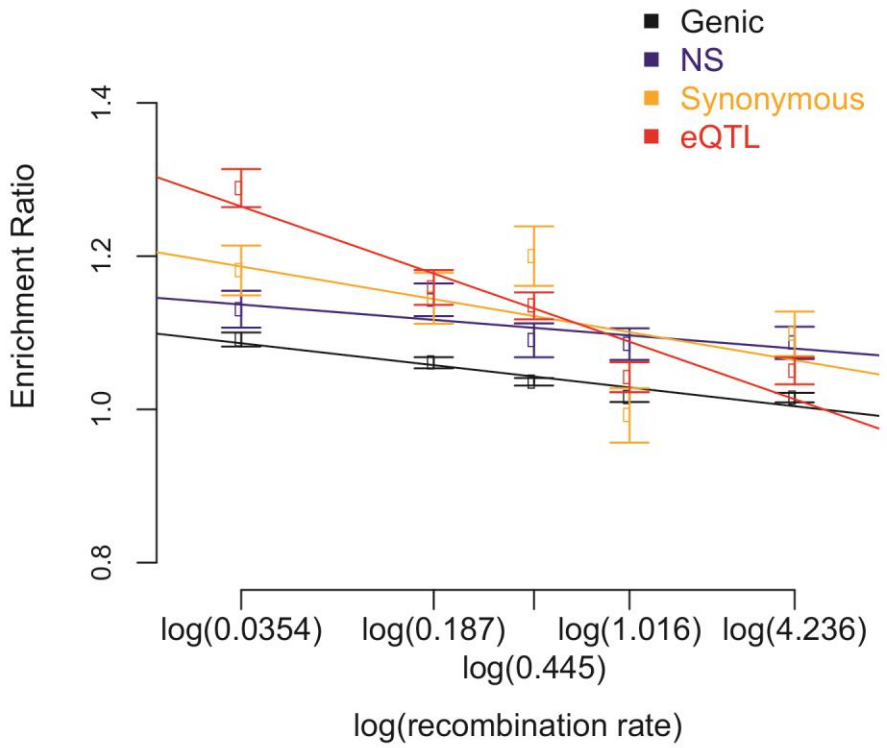


Figure 2.4. Negative correlation between recombination rate and enrichment ratio. Results were shown for eQTLs, genic, NS and synonymous SNPs. The tail cutoff is 5%. The patterns are similar for cutoffs of 1% and 0.5%. Only a subset of eQTLs associated with RefSeq-supported protein-coding genes (eQTLs-for comparison) were used in these analyses. The results for all eQTLs (eQTLs-all) have similar patterns.

Table 2.2. A subset of biological pathways with eQTLs-specific enrichment

Environmental factors / Pathways	Enrichment Ratio[a]					
	genic SNPs in pathway : other genic SNPs			eQTLs in pathway : other eQTLs		
	Tail cutoffs:					
	5%	1%	0.5%	5%	1%	0.5%
1-Subsistence						
Hematopoietic cell lineage	1.24	1.49	1.64	1.91	2.28	2.72
Intestinal immune network for IgA production	1.29	1.78	1.92	1.87	2.81	2.64
Genes involved in Costimulation by the CD28 family	1.12	1.17	0.95	1.80	2.57	2.55
Genes involved in PD-1 signaling	1.47	2.18	1.93	1.92	2.84	2.50
Genes involved in Signaling in Immune system	1.07	1.18	1.31	1.44	1.85	2.27
2-Climate						
Long-term depression	0.98	0.85	0.82	1.39	2.31	2.44
Asthma	1.33	2.00	2.97	1.56	1.87	1.82
Systemic lupus erythematosus	1.13	1.23	1.31	1.50	1.79	1.77
1-Dry domain						
Genes involved in biological oxidations	0.93	0.57	0.94	2.11	3.88	6.02
2-Short wave radiation flux (Summer)						
Genes involved in peroxisomal lipid metabolism	1.96	3.97	4.98	2.86	9.36	17.30
2-Precipitation rate (Winter)						
ErbB signaling pathway	1.33	1.71	1.74	2.83	5.62	6.24
TGF beta signaling pathway	1.10	1.15	1.30	2.16	4.82	9.10
Renal cell carcinoma	1.45	1.30	1.20	2.17	4.14	4.59
MAPK/ERK pathway	1.00	1.33	1.40	2.03	5.16	7.76
Hemostasis	1.01	1.10	0.84	1.81	3.47	3.28
3-Bacteria diversity						
Genes involved in lipids and lipoproteins metabolism	1.15	1.11	1.01	1.80	3.17	3.87

[a]The level of significance was estimated by whole-genome block bootstrap. Red and orange indicates respectively >99% and >95% of bootstrap replicates having enrichment ratios larger than 1.

2.4.4. Enrichment of eQTLs in environmental adaptation for specific biological functions

To identify biological pathways undergoing regulatory environmental adaptation, for each environmental factor, we identified biological pathways with consistent and significant enrichment of eQTLs, but not genic SNPs, in the lower tail of the transformed rank statistic. As indicated in Table 2.2., most of the significant pathways we identified are mainly related to immune system, cellular signaling and metabolism. For instance, eQTLs associated with genes involved in biological oxidations are significantly enriched in the lower tail of the transformed rank statistic for the environmental factor of dry domain with significant ERs of 2.11, 3.88 and 6.02, respectively for three tail cutoffs. In comparison, genic SNPs in the same pathway are not significantly enriched with ERs of 0.93, 0.57, and 0.94, respectively. For the pathway of peroxisomal lipid metabolism, eQTLs are enriched for short wave radiation flux in summer with significant ERs of 2.86, 9.36 and 17.30, much higher than those of the genic SNPs. For genes involved in signaling in immune system, eQTLs are significantly enriched in signals of environmental correlation for the subsistence category with binary individual factors, whose ERs are 1.44, 1.85 and 2.27.

The prevalence of immune-related genes with significant regulatory adaptation is consistent with previous findings that active binding sites for immune-related TFs are among the most highly eQTL-enriched regions in the genome (36). It has been found that eQTLs of genes interacting with HIV proteins tend to overlap with signals of

incomplete selective sweep as measured by his (25). One of the cases is *HLA-C*, which has a cluster of eQTLs overlapped with signals of selection in both Europeans and Asians (25). Interestingly, we found the same cluster of eQTLs have significant signals of association with multiple environmental factors. One eQTL (rs6931332) has strong evidence of environmental association signals (transformed rank statistic = 0.0025 for the climate category with seasonal components, 0.03 for the climate category with annual means, 0.0016 for the subsistence category with continuous individual factors and 0.034 for pathogen). Another eQTLs (rs9264942), which has been found to be associated HIV-1 viral load (37), has significant association with the climate category with seasonal components (transformed rank statistic = 0.026) and suggestive evidence of association with pathogen (0.068). These findings suggest that *HLA-C* has regulatory variants that might be adaptive to certain climate or pathogen related environmental factor(s), which further provided protective effect against HIV infection in the near past.

Lipid metabolism is the most prominent metabolic pathway to exhibit consistently significant ERs for eQTLs but not genic SNPs. In addition to the above mentioned example of peroxisomal lipid metabolism with enrichment signal to short wave radiation flux in summer, the pathway of metabolism of lipids and lipoproteins also shows significant eQTLs enrichment to the environmental factor of bacteria diversity with ERs of 1.80, 3.17 and 3.87. These observations are consistent with the fact that lipid metabolism plays a wide variety of roles in numerous signaling and regulatory process and host-pathogen interactions (38, 39). Our analysis also highlights top candidate genes in these two pathways for future detailed examination. For

peroxisomal lipid metabolism, the top three SNPs with most significant correlative signal with short wave radiation flux in summer are found in ACOX1 (rs8065144, transformed rank statistic = 8.11e-4), SCP2 (rs11206043, 0.0018), and AMACR (rs35414, 0.0021). And for metabolism of lipids and lipoproteins, the top three correlative signals with bacteria diversity are located in NCOR2 (rs701078, 0.00047), LSS (rs2280957, 0.0016) and SGPP2 (rs4673024, 0.0021).

2.5. Discussion

2.5.1. The observed enrichment patterns of eQTLs are biologically meaningful

In this study we are able to show that compared with intergenic neutral SNPs, eQTLs are significantly more likely to show association with 42 environmental factors. Our results are unlikely to be statistical artifacts because of the following reasons. First, these enrichments are consistent across environmental variables, and also across different methods for assessing environmental correlations. Second, our results of eQTLs enrichment are also consolidated by the observation that similar environmental factors from different studies (27, 31, 32) exhibit consistently significant patterns. For instance, the environmental category of climate, either summarized over the summer and winter components or the annual means, exhibits similar level of enrichment across three tail cutoffs. So does the category of subsistence, either summarized over binary individual factors or continuous ones. Similarly, consistent patterns are observed at the level of individual environmental factors from different studies (Figure 2.1. A, Table 2.1.). Third, we observed a greater enrichment of adaptive eQTLs in regions of lower recombination, which is consistent with theoretical expectations. And fourth, we observed enrichment of eQTLs in environmental adaptation for specific functions, indicating that the patterns we observed are of biological significance.

2.5.2. Regulatory variants is underlying the significant enrichment of eQTLs

The statistical association between eQTLs and environmental variables appears to be strong. However, NS SNPs, rather than regulatory variants, could be the underlying

explanation if a large proportion of eQTLs are themselves NS SNPs or they are in strong linkage disequilibrium (LD) with NS SNPs. Firstly, we ruled out the possibility that the enrichment of eQTLs is a direct effect of NS SNPs because only ~3% of eQTLs are NS SNPs and excluding them has no influence on the observed patterns. Secondly, to exclude the indirect effect of NS SNPs as the underlying driver, we examined the LD patterns between eQTLs and NS SNPs. Under the environmental correlation framework, an eQTL could only capture the adaptive effect of a NS SNP if a strong LD between the two is present in most human populations tested. However, LD structure is different across populations and it has only been well-studied for a subset of populations included in the environmental correlation analysis. Therefore, we performed a preliminary and conservative analysis using LD data from Hapmap 3 (40). We excluded eQTLs that are in strong LD ($r^2 > 0.8$) with any NS SNP in any of the 11 populations, which account for ~20% of all eQTLs, much higher than the number of eQTLs (~2%) that are in strong LD with any NS SNPs in all 11 populations. By further excluding eQTLs that are also NS SNPs, we obtained a group of eQTLs that are less dependent on NS SNPs than all eQTLs. However, as observed for all eQTLs, this group of eQTLs is also significantly enriched in signal of environmental correlation (p = 1.76e-09, 2.21e-08, 2.84e-07 respectively for three tail cutoffs).

Moreover, although around 66% of eQTLs locate within 5 kb of a gene and ~49% are within introns, the difference between the ER for eQTLs located in genic region and that for eQTLs in non-genic regions is not consistently significant. In addition, if the enrichment of eQTLs is only caused by their LD with NS SNPs, the enrichment of

eQTLs should be smaller than that of NS SNPs, as the trend observed for synonymous SNPs. However, under all three tail cutoffs, the mean/median ER of eQTLs across 42 environmental factors is slightly higher than that of NS SNPs, although the difference is not significant. Taken together, we rule out the possibility that the enrichment of eQTLs is an indirect effect of NS SNPs through their LD with eQTLs. Therefore, the most parsimonious explanation for the enrichment of eQTLs in signals of environmental correlation is the adaptive effect of regulatory variants, which could be eQTLs themselves or the underlying causal variants in strong LD with eQTLs.

2.5.3. eQTLs are as prevalent as, if not more prevalent than, NS SNPs in local adaptation

We also observed a general trend that the degree of enrichment over all environmental factors is higher for eQTLs than NS SNPs, although this trend is only significant under the tail cutoff of 5%. However, the interpretation of this observation is not straight-forward. Firstly, these two groups of SNPs are not independent of each other and strong LD between these two may be the underlying cause of similar enrichment. To relieve the complication of LD, we obtained a group of eQTLs that are less dependent on NS SNPs as defined above and similarly a group of NS SNPs that are less dependent on eQTLs. The same trend is observed that the less dependent eQTLs have higher enrichment ratio than the less dependent NS SNPs. And this difference is also significant under the tail cutoff of 5% ($p = 0.0035$). These additional observations suggest that the similar degree of enrichment in signals of environmental correlation is unlikely to be explained by their LD with each other.

Secondly, recombination rate may confound the comparison of enrichment ratios

between two SNPs groups. As we demonstrated previously, SNPs in regions of low recombination rate tend to have higher enrichment ratio. The recombination rate of genome-wide NS SNPs (median 0.29) is not significantly different than that of eQTLs (median 0.30) (Wilcoxon test, $p = 0.62$). However, the recombination rate of NS SNPs for genes with eQTLs (median 0.27) is significantly lower than that of eQTLs ($p = 0.0022$), indicating that our identification of environmental factors, to which eQTLs adaptation is more prevalent than NS SNPs, is conservative because recombination rate make the comparison bias towards NS SNPs.

Therefore, the observed slightly higher enrichment ratio for eQTLs than NS SNPs suggests that eQTLs, at least the current collection of eQTLs, are as prevalent as, if not more prevalent than, NS SNPs in local environmental adaptation. Consistent with this conclusion, a new study recently published while our manuscript was under review utilized part of the datasets used in our study (Hancock *et al.* Plos) and found that loci under local adaptation are 10-fold more likely to overlap with eQTLs than NS SNPs, supporting the important role of regulatory variants in human adaptation (41). In spite of our effort to compile eQTLs identified from different tissues and populations, they may represent only a small subset of all regulatory variants that are functional in human regulatory networks. It is unclear how many eQTLs remain to be identified and how these newly identified eQTLs will affect the patterns observed in our study.

2.5.4. Different functional aspects of eQTLs may impact enrichment patterns

***Tissue-specific eQTLs*.** eQTLs for specific tissues may show unique enrichment

patterns for different environmental factors. The eQTLs data we used were mainly identified in lymphoblastoid cell lines (LCL) and monocytes. So it is possible that the enrichment patterns we observed only reflect the properties of these two cell types. For instance, enrichment analysis for tissue-specific eQTLs of LCL, monocytes and liver revealed that LCL-specific eQTLs tend to exhibit enrichment to a larger number of environmental factors than the other two. And several environmental factors only show significant enrichment of LCL-specific eQTLs. Moreover, although eQTLs identified in multiple tissues tend to have higher enrichment ratio than eQTLs identified in a single tissue ($p = 0.032$), interpreting the role of tissue specificity is not straight-forward because eQTLs identified in multiple tissues also tend to locate at regions of low recombination rate ($p < 2.2e-16$). Therefore, the presence of specific evolutionary pattern for tissue-specific eQTLs needs future investigation when more tissue-specific eQTLs become available.

***Cis- and trans eQTLs*.** Our analyses here included all eQTLs, without separating them into cis- and trans-eQTLs. It is a very important question to examine if cis- and trans-eQTLs exhibit different evolutionary properties. However, most eQTLs identified by far are cis-eQTLs. Trans-eQTLs are much less identified, probably due to the fact that they are much less frequent or they exert much smaller effects than cis-eQTLs. And a much larger burden of multiple-testing also prevents the identification of trans-eQTLs (42). In the dataset of eQTLs used in our analysis, only a small fraction (<5%) are trans-eQTLs, precluding our analyses with trans-eQTLs. Due to the same reason, our observations may only reflect the evolutionary patterns of cis-eQTLs. A larger collection of trans-eQTLs is needed to characterize their evolutionary significance

using the method applied in our study.

The number of genes regulated by an eQTL. Master regulators (eQTLs regulating multiple genes) (42) may be functionally more important and exhibit different evolutionary properties from those regulating only one gene. It is observed that the ERs across 44 environmental factors are higher for master regulator than those for eQTLs associated with only one gene (p = 7.96e-5). However, master regulators also tend to locate in regions of lower recombination (p < 2.2e-16), confounding the interpretation of the higher ERs.

The effect size of an eQTL. The effect size of an eQTL refers to the degree to which the eQTL could influence the expression level of the associated gene. If an eQTL has higher effect size, it may be more likely to play a role in adaptation. To explore if eQTLs with larger effect size exhibit higher ERs, we took advantage of an eQTLs dataset (22) that provides the effect size and p value of association for each eQTL. Since the effect size of an eQTL is inversely correlated with the p value of association (ρ = -0.31, p = 2.2e-16), a statistic called expression score (24) was developed to measure how likely an eQTL is a true positive and how strong its effect is. The more likely an eQTL is a true positive and the larger its effect size, the higher its expression score. Similar to the cases of tissue-specific eQTLs and master regulators, although eQTLs with higher expression score show stronger enrichment, they are also associated with lower recombination rate.

2.5.5. Environmental correlation study assists the elucidation of regulatory adaptation

Our analyses highlight a number of environmental factors and biological pathways for

which eQTLs are of special evolutionary importance. The underlying biological mechanisms making eQTLs important for environmental adaptation need further investigation. After identifying the adaptation signals and ecological context, the natural next step is to elucidate the underlying mechanistic processes of how a regulatory change can result in phenotypic differences (42). There are a growing number of cases of regulatory adaptation whose molecular and evolutionary mechanism have been elucidated. For example, a causal eQTL (rs9493857) regulating the expression level of SGK1, a key gene in response to environmental stress, was found to be associated with multiple environmental factors, including latitude (43). This SNP (rs9493857) is not present in the dataset used in our study, but an available proxy SNP (rs4896028, r2 with rs9493857 is 0.736 in Europeans) is significantly associated with short wave radiation flux in summer (transformed rank statistic = 0.016). Combining the functional context of eQTLs and the ecological context of environmental correlation, many more cases of regulatory adaptation could be illustrated in the near future.

2.6. Conclusions

Our evolutionary analyses with eQTLs reveal that regulatory variations are as prevalent as, if not more prevalent than, NS SNPs, in driving recent and ongoing human adaptation to local environment. The importance of regulatory variations is more prominent for continuous environmental factors, such as climate, possibly due to the fine-tuning property of gene expression. Moreover, regulatory changes played an important role in some biological pathways, especially those related with signaling, immune and metabolic functions, for their local adaptation. Combining the functional context of eQTLs and the ecological implication of environmental correlation provides important insights for future elucidation of the mechanism and selection pressure of regulatory adaptation.

2.7. Acknowledgements

We would like to thank Dr. James Booth for his help in developing the generalized Z test. We are grateful to valuable comments from Drs. Alon Keinan and Andrew Clark and their lab members. We also want to thank Drs. Matteo Fumagalli and Anna Di Rienzo for providing their data. This work was supported by startup funds from Cornell University and ILSI Future Leader in Nutrition Award to Dr. Zhenglong Gu.

2.8. References

1. Wray GA (2007) *Nat Rev Genet* **8,** 206-216.

2. King MC & Wilson AC (1975) *Science* **188,** 107-116.

3. Zheng W, Gianoulis TA, Karczewski KJ, Zhao H, & Snyder M (2011) *Annu Rev Genomics Hum Genet* **12,** 327-346.

4. Carroll SB (2005) *PLoS Biol* **3,** e245.

5. Wittkopp PJ & Kalay G (2012) *Nat Rev Genet* **13,** 59-69.

6. Dermitzakis ET (2008) *Adv Genet* **61,** 295-306.

7. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, & Jarvela I (2002) *Nat Genet* **30,** 233-237.

8. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, & Hirschhorn JN (2004) *Am J Hum Genet* **74,** 1111-1120.

9. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, & Osman M, *et al.* (2007) *Nat Genet* **39,** 31-40.

10. Enattah NS, Jensen TG, Nielsen M, Lewinski R, Kuokkanen M, Rasinpera H, El-Shanti H, Seo JK, Alifrangis M, & Khalil IF, *et al.* (2008) *Am J Hum Genet* **82,** 57-72.

11. Hamblin MT & Di Rienzo A (2000) *Am J Hum Genet* **66,** 1669-1679.

12. Tournamille C, Colin Y, Cartron JP, & Le Van Kim C (1995) *Nat Genet* **10,** 224-228.

13. Prabhakar S, Noonan JP, Paabo S, & Rubin EM (2006) *Science* **314,** 786.

14. Haygood R, Fedrigo O, Hanson B, Yokoyama KD, & Wray GA (2007) *Nat Genet* **39,** 1140-1144.

15. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, & Suver C, *et al.* (2008) *PLoS Biol* **6,** e107.

16. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, & Nath P, *et al.* (2007) *Nat Genet* **39,** 1494-1499.

17. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, & Koller D, *et al.* (2007) *Nat Genet* **39,** 1217-1224.

18. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, & Pritchard JK (2008) *PLoS Genet* **4,** e1000214.

19. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, & Pritchard JK (2010) *Nature* **464,** 768-772.

20. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, & Dermitzakis ET (2010) *Nature* **464,** 773-777.

21. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez AM, & Sekowska M, *et al.* (2009) *Science* **325,** 1246-1250.

22. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, & Rossmann H, *et al.* (2010) *PLoS One* **5,** e10693.

23. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, & Dermitzakis ET (2010) *PLoS Genet* **6,** e1000895.

24. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, & Cox NJ (2010) *PLoS Genet* **6,** e1000888.

25. Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, & Pritchard JK (2009) *Mol Biol Evol* **26,** 649-658.

26. Ye K & Gu Z (2011) *Adv Nutr* **2,** 486-496.

27. Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, Sukernik R, Utermann G, Pritchard J, & Coop G, *et al.* (2010) *Proc Natl Acad Sci U S A* **107 Suppl 2,** 8924-8930.

28. Pritchard JK, Pickrell JK & Coop G (2010) *Curr Biol* **20,** R208-R215.

29. Hancock AM, Alkorta-Aranburu G, Witonsky DB, & Di Rienzo A (2010) *Philos Trans R Soc Lond B Biol Sci* **365,** 2459-2468.

30. Coop G, Witonsky D, Di Rienzo A, & Pritchard JK (2010) *Genetics* **185,** 1411-1423.

31. Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, & Di Rienzo A (2011) *PLoS Genet* **7,** e1001375.

32. Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, & Nielsen R (2011) *PLoS Genet* **7,** e1002355.

33. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A,

Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, & Kristinsson KT, *et al.* (2010) *Nature* **467,** 1099-1103.

34. Keinan A & Reich D (2010) *PLoS Genet* **6,** e1000886.

35. Cai JJ, Macpherson JM, Sella G, & Petrov DA (2009) *PLoS Genet* **5,** e1000336.

36. Gaffney DJ, Veyrieras JB, Degner JF, Pique-Regi R, Pai AA, Crawford GE, Stephens M, Gilad Y, & Pritchard JK (2012) *Genome Biol* **13,** R7.

37. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M, Zhang K, Gumbs C, Castagna A, & Cossarizza A, *et al.* (2007) *Science* **317,** 944-947.

38. van der Meer-Janssen YP, van Galen J, Batenburg JJ, & Helms JB (2010) *Prog Lipid Res* **49,** 1-26.

39. Wenk MR (2006) *Febs Lett* **580,** 5541-5551.

40. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, & Peltonen L, *et al.* (2010) *Nature* **467,** 52-58.

41. Fraser HB (2013) *Genome Res*.

42. Gilad Y, Rifkin SA & Pritchard JK (2008) *Trends Genet* **24,** 408-415.

43. Luca F, Kashyap S, Southard C, Zou M, Witonsky D, Di Rienzo A, & Conzen SD (2009) *PLoS Genet* **5,** e1000489.

Chapter 3 – Natural Selection on HFE in Asian Populations Contributes to Enhanced Non-heme Iron Absorption[3]

3.1. Abstract

HFE, a major regulator of iron (Fe) homeostasis, has been suggested to be under positive selection in both European and Asian populations. While the genetic variant under selection in Europeans has been relatively well-studied, the adaptive variant in Asians and its functional consequences are still unknown. Using data from the International HapMap Project, we confirmed positive selection on *HFE* in Asian populations and identified a candidate adaptive haplotype that is common in Asians (54.71%) but rare in Europeans (5.98%) and Africans (4.35%). The T allele at tag SNP rs9366637 (C/T) captured 95.8% of this Asian-common haplotype. Using gene expression profiles from lymphoblastoid cell lines of 85 Asians in the HapMap Project, a significantly reduced *HFE* expression was observed in individuals carrying T/T at rs9366637 ($p < 0.025$) compared to C/C and C/T. We recruited 57 women of Asian descent and measured Fe absorption using stable isotopes in those homozygous at rs9366637. We observed nearly a 20% higher absorption (p for trend =0.099) in women homozygous for the Asian-common haplotype (T/T, 17.1 ± 6.3%, n = 11) compared to the control genotype (C/C, 14.1 ± 6.2%, n = 10). Additionally, compared with a group of age-matched Caucasian women (n=18), Asian women exhibited both significantly elevated Fe status (p = 9.4e-3 for hemoglobin and p = 0.045 for serum

[3] This manuscript is currently under review for publication. Kaixiong Ye and Chang Cao contribute equally to this work.

ferritin) and Fe absorption (p = 5.3e-4). Our results indicate population differences in iron homeostasis and suggest that natural selection on *HFE* may have contributed to elevated Fe absorption in Asians.

3.2. Introduction

Adaptations to dietary intake during human evolution have shaped the human genome and resulted in ethnic variability in nutrient utilization and risk of diseases (1-3). Classic examples of this process have been established for the prevalence of lactase persistence in Northern Europeans and East Africans as an adaptation to milk consumption (4), high copy number of *AMY1* in populations ingesting starch-rich diets (5), and the Asian alcohol flush reaction which evolved as an adaptive response to alcohol consumption after rice domestication (6). The recent advent of high-throughput genotyping and sequencing technology enables genome-wide scans for signals of positive selection and generates many hypotheses that await functional testing and confirmation (7, 8). This is particularly relevant in that the incompatibility between ancient genetic adaptation and modern dietary environment could underlie certain metabolic diseases in the current society (1-3).

Iron (Fe) is an essential micronutrient involved in oxygen transport, oxidative metabolism and immune function (9, 10). Iron deficiency (ID) is one of the most widespread micronutrient deficiencies worldwide and may lead to ID anemia, causing chronic fatigue, reduced work productivity, impaired immune response, poor pregnancy outcome, and delayed physical and cognitive development in infants (11-14). The risk of ID is especially high for populations consuming predominantly plant-based diets, with little dietary sources of heme Fe (15, 16). Genetic variations enhancing non-heme Fe absorption from plant-based diets would have been especially beneficial in these populations and subject to positive natural selection, yet no such

genetic variations have been revealed to date. Furthermore, while population differences in Fe status and the prevalence of ID are known to occur (12, 17), the potential role of genetic variation underlying these differences has remained largely uncharacterized. In the modern Fe-replete dietary environment, identifying genetic variation enhancing Fe absorption is especially important for future prevention of Fe overload and its related diseases, including type 2 diabetes (18-20).

HFE is one of the major regulators of Fe homeostasis and was the first gene found to be implicated in hereditary hemochromatosis (HH) (21, 22). A non-synonymous mutation of *HFE*, C282Y (rs100562), is responsible for more than 80% of HH found in Europeans (21). While this mutation has a frequency of 5-14% in northern European populations, it is nearly absent outside of Europe (23, 24). The relatively high frequency of this mutation in European populations has been suggested to be a result of recent positive selection whereby this currently disease-causing mutation provided resistance to Fe-requiring pathogens during human evolution (24-27). The sequence variation and haplotype structure at *HFE* are quite different among continental populations, and interestingly, Asian populations possess a high-frequency haplotype, referred to as the Asian-common haplotype, that is rarely observed among European or African populations (28). This haplotype may have been driven to high frequency by positive selection if it provided a selective advantage. Consistently, a signal of positive selection on *HFE* has been suggested in Chinese populations based on patterns of SNP allele frequency around the *HFE* gene (29). However, the possibility of local adaptation of *HFE* in Asia requires further confirmation and the underlying adaptive variants need to be revealed.

We hypothesize that this Asian-common haplotype that has been found to exist at high frequency in Asian populations is associated with improved Fe stores and has a functional impact on absorption of non-heme Fe. To test this hypothesis, we performed multiple evolutionary tests and unraveled a regulatory variant in *HFE* with adaptive signals in Asian populations. The impact of this regulatory variant on Fe status and absorption was explored in a group of young Asian women and compared to data on non-heme Fe absorption in Caucasian women.

3.3. Materials and Methods

3.3.1. Evolutionary analysis

Haplotype structure analysis was conducted with haplotype data downloaded from Hapmap 3 (30). Single Nucleotide Polymorphisms (SNPs) within 20 Kb upstream or downstream of gene *HFE* and with a minimum population frequency of 5% were used. For SNPs that are in perfect linkage disequilibrium ($r^2 = 1$) with each other, only one SNP was randomly chosen to remove redundant information. In total, 10 SNPs were selected (rs9295684, rs6942196, rs2794719, rs9366637, rs2071303, rs1800708, rs1572982, rs17596719, rs6918586, and rs1150658). Haplotype information was available for 170 individuals from Beijing, China and Tokyo, Japan (Coded as CHB+JPT), 85 Chinese in Metropolitan Denver, Colorado (CHD), 117 Utah residents with Northern and Western European ancestry from the CEPH collection (CEU) and 115 Yoruba in Ibadan, Nigeria (YRI). Haplotype network analysis was performed using the median-joining algorithm of Network 4.6.0.0 (31). The root was inferred assuming that the chimpanzee allelic state at each SNP was ancestral.

Evolutionary analysis on *HFE* was performed through Haplotter (32) and HGDP Selection Browser (33). Haplotter utilized a selection detection test, iHS, to detect signals of positive selection using HapMap data (32). HGDP Selection Browser presented the global distribution of SNPs in 53 populations (33).

3.3.2. Gene expression analysis

The relationship between the rs9366637 genotype and the *HFE* expression level was assessed using the genotype and gene expression data of Asian individuals in the

International Hapmap Project. Genotype data were retrieved from Hapmap 3 (30). The expression level of *HFE* was retrieved from a previous study, which used a commercial whole-genome expression array (Sentrix Human-6 Expression BeadChip version 1, Illumina) to quantify the transcriptional profiles of Epstein-Barr virus–transformed lymphoblastoid cell lines from individuals genotyped in the HapMap Consortium (34). Expression signal values were $\log_2$- transformed and normalized first by quantile normalization across the four replicates of an individual, followed by median normalization across all individuals to allow comparison of expression values across individuals (34). In total, we have both genotype and *HFE* expression data for 43 Chinese individuals from Beijing, China and 42 Japanese individuals from Tokyo, Japan.

3.3.3. Subjects in genotyping and absorption study

Fifty-seven women of Asian descent were recruited through advertisement on the Cornell University campus in Ithaca, NY from August 2012 to May 2013. Women were eligible for the study if they met the following criteria: 1) non-pregnant and 18-35 years old; 2) of Asian descent with both maternal and paternal grandparents from Asia; 3) not taking any vitamin or mineral supplements for at least 1 month before the study and during the 2-week dosing study interval; 4) without pre-existing medical problems including malabsorption, blood disorders, ulcers, inflammatory diseases, asthma or conditions that might impact inflammation or Fe status; and 5) not taking any prescribed medications known to affect Fe homeostasis. Informed written consent was obtained from each woman, and the study was approved by the Institutional Review Board at Cornell University. A sample size of 60 participants was

selected based on the frequency of the tag SNP rs9366637 in Asian populations (30) and this sample size was expected to yield roughly 15 women homozygous for either allele. The sample size of about 15 in each group of homozygous individuals was selected based on previous studies examining genetic effects on Fe absorption (35, 36).

For the genotype and Fe status screening study, Asian women came to the Human Metabolic Research Unit (HMRU) at Cornell University and a venous blood sample (10 mL) was collected to determine *HFE* haplotype and baseline Fe status, including hemoglobin (Hb), serum ferritin (SF), serum transferrin receptor (sTfR), and hepcidin. Women carrying genotypes of C/C or T/T at SNP rs9366637, were then asked to further participate in an iron absorption study, which required two additional visits to the HMRU. On the day of the absorption study, designated as the dosing day, fasted (≥ 1.5 h) women arrived at the HMRU at which time a baseline weight was obtained and a venous blood sample was collected by either finger stick or venipuncture (< 3 mL) based on subject preference. Participants were then given an oral dose of ^{57}Fe tracer as ferrous sulfate (6.3 mg ^{57}Fe; total Fe dose of 6.6 mg) flavored with raspberry syrup (Humco, Texarkana, TX). The dose was administered by syringe, which was pre- and post-weighted to calculate the amount of Fe tracer consumed. After consumption, participants remained fasted for an additional 1.5 hours after which time they were provided a standardized lunch of vegetable soup, pretzels and water. Two weeks after the ingestion of the Fe tracer, designated as the post-dosing visit, participants returned to the HMRU and a 10 mL venous blood samples was collected for analysis of red blood cell ^{57}Fe enrichment and serum Fe status indicators.

3.3.4. Laboratory analysis

DNA was extracted from the whole blood samples of the 57 Asian participants using a commercially available Genomic DNA Purification Kit (Promega Corporation, Fitchburg, WI). A 439 bp long segment centering on rs9366637 was PCR amplified (forward primer: 5'-ATGGTACACTGGGCTTTGGT-3'; reverse primer: 5'-TAGTGCTGAGAAAACCCGCTT-3') and sequenced by a Sanger sequencing platform (ABI 3730xl). The genotype at rs9366637 (C/C, C/T, T/T) was visually identified from chromatograms using Chromas Lite 2.1.1 software.

Hb and hematocrit were analyzed with a hematology analyzer (Beckman Coulter, Fullerton, CA). SF was measured by a commercially available enzyme immunoassay procedure (Ramco Laboratories Inc, Stafford, TX). The concentration of sTfR was measured with an enzyme-linked immunosorbent assay (ELISA; Ramco Laboratories Inc, Stafford, TX). Total body iron (TBI) was calculated using the ratio of serum transferrin receptor to serum ferritin as described in a previous study: TBI (in mg/kg) = – [log10 (sTfR/SF) – 2.8229]/0.1207 (37). Serum folate, vitamin B-12, and C-reactive protein (CRP) were measured by Immulite 2000 immunoassay system (Siemens Medical Solutions Diagnostics, Los Angeles, CA). Hepcidin was determined with a commercial Enzyme Immunoassay Kit **(**S-1337**;** Bachem, San Carlos, CA).

3.3.5. Isotope preparation and sample analysis

Fe isotope (^{57}Fe at 95% enrichment) was purchased as the metal from Trace Sciences International (Richmond Hill, Canada). The tracer was converted into a sterile, pyrogen-free solution of ferrous sulfate following the methods used by Kastenmayer et

al. (38). The isotopic composition of the tracer solution was validated with the use of a ThermoQuest Triton TI Magnetic Sector Thermal Ionization Mass Spectrometer (ThermoQuest Corporation, Bremen, Germany).

Whole-blood samples (0.5 mL) were digested with 4 mL Ultrex nitric acid in a polytetrafluoroethylene beaker. Samples were then dried on a hot plate and dissolved in 6N ultrapure hydrochloric acid (JT Baker, Phillipsburg, NJ). Fe was extracted with the use of modified anion exchange chromatography as previously described (39). Extracted samples were reconstituted in 3% nitric acid and loaded onto a rhenium filament (H Cross Co, Weehawken, NY) with 4 μL of silica gel (Sigma-Aldrich Inc, St Louis, MO) and 4 μL of phosphoric acid (0.7 N). Isotopic ratio of ^{57}Fe to ^{56}Fe ($^{57/56}Fe$) was measured by thermal ionization mass spectrometry and compared to the natural abundance of $^{57/56}Fe$ (0.02317). Relative SDs of $^{57/56}Fe$ in analyzed samples averaged 0.04%.

To control for the known effect of Fe stores on Fe absorption, percent Fe absorption was normalized to each woman's measured SF concentration as follows: normalized percent absorption = raw percent absorption * SF / 40 (40).

3.3.6. Iron absorption in Caucasian women

Eighteen women of Caucasian descent, aged 18-32 y, were recruited on the same university campus starting in the spring of 2007 with similar enrollment criteria as our recruitment for Asian women. Fe status markers and other molecular characteristics were measured with the same experimental approaches as in Asian women except that hepcidin was measured by a competitive ELISA (Intrinsic LifeSciences, La Jolla,

CA). Non-heme Fe absorption was assayed from a total Fe dose of 7.6 mg of $FeSO_4$ given in 1.5 mL of flavored syrup. The same meals were fed on the dosing day as those detailed above. Of the total Fe load administered, 0.9 mg was provided as ^{58}Fe. The detailed experimental procedure and subject characteristics have been described previously (39). Genotyping was not done in these samples but it is estimated that 88.36% of this cohort (16 individuals) were C/C, 11% (2 individuals) were C/T, and 0.36% (0 individual) were T/T based on the known frequency of this haplotype among Caucasians (30).

3.3.7. Iron status and absorption data analysis

All statistical analyses were performed with R 3.1.0 (41). Variables that were not normally distributed (SF, hepcidin) were transformed using a natural logarithm. Potential differences in Fe status and Fe absorption as a function of genotypes were tested with a one-tailed two-sample Wilcoxon rank sum test, which is a non-parametric test without making assumptions about the distribution of variables (42). Our alternative hypothesis was that individuals carrying the T/T genotype would have higher Fe status (as determined by Hb, SF, sTfR and TBI) and higher percent Fe absorption than those carrying the C/C genotype. Possible differences in physical characteristics (age, weight, BMI, folate, vitamin B-12, CRP, etc.) among women with different genotypes, and between Asians and Caucasians were tested using the two-tailed, two-sample Wilcoxon rank sum test. Simple linear regression analysis was used to determine possible relationships between Fe status (SF, sTfR, TBI, and Hb), percent Fe absorption and serum hepcidin concentrations. Comparisons of Fe status and Fe absorption between Asian and Caucasian women were performed using one-tailed

two-sample Wilcoxon rank sum test with the alternative hypothesis that Asians have higher Fe status and normalized Fe absorption. Individuals with missing data for a parameter were excluded only for analysis on that specific parameter.

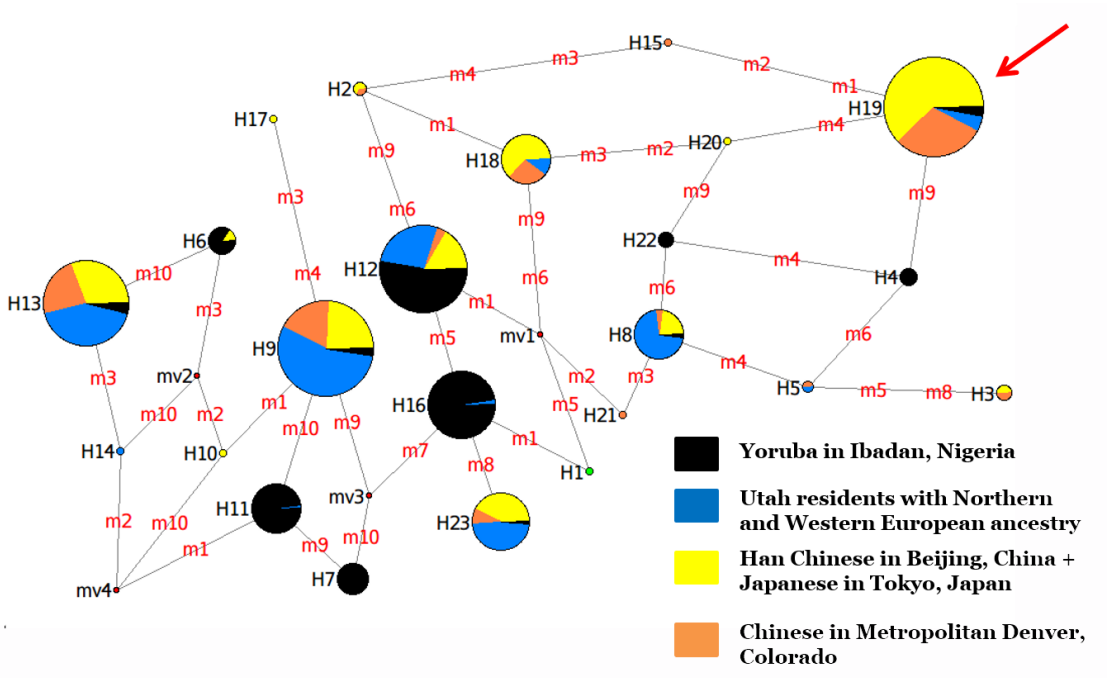


Figure 3.1. The haplotype structure of *HFE*. SNPs within 20 Kb of *HFE* were used in this analysis. Data were retrieved from Hapmap 3. Each node represents one haplotype and the size of each node is proportional to the frequency of the haplotype among all haplotypes observed. Each node is also a pie chart and each sector represent the contribution of each population. The line between two nodes represents the evolutionary connection between the two haplotypes and the mutations to convert the two haplotypes are indicated on the line. The red arrow points to the Asia-prevalent haplotype, H19. The green node represents the haplotype (H1) observed in Chimpanzees.

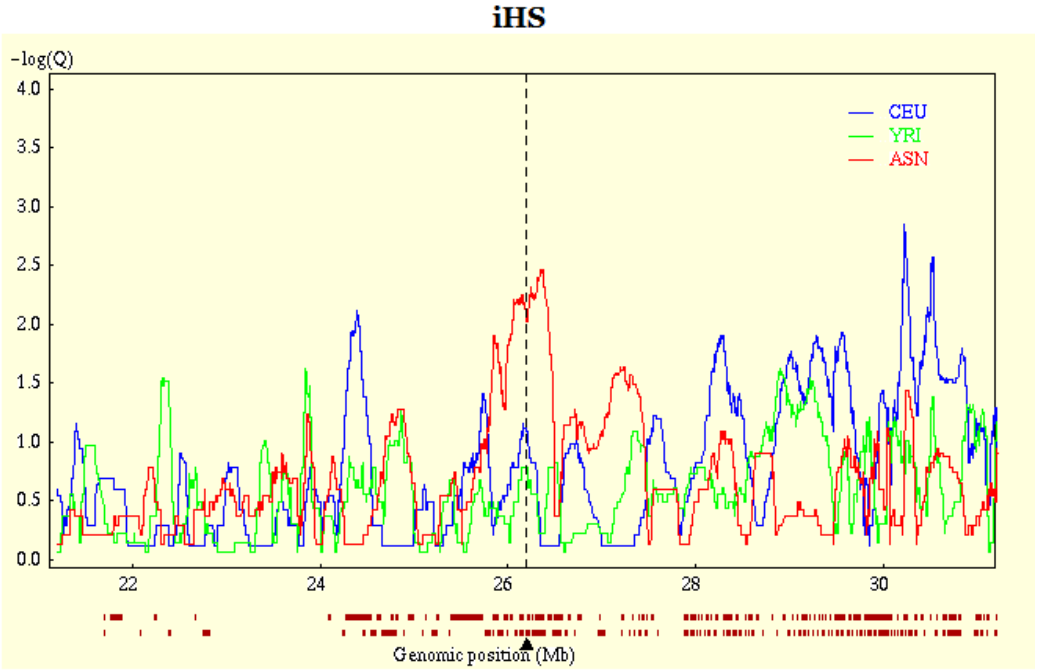


Figure 3.2. Signals of positive selection on *HFE* in Asian populations. Positive selection was evaluated with the iHS method in three continental population samples. The x axis represents genomic position with nearby genes indicated as red boxes at the bottom. The target gene, *HFE*, is indicated with a black triangle and a vertical dashed line. The y axis represents the negative logarithm of empirical *p* values for evidence of positive selection. ASN (in red) represents a combination of two samples, coded as CHB and JPT, referring respectively to Han Chinese in Beijing, China and Japanese in Tokyo, Japan. CEU (in blue) refers to Utah residents with Northern and Western European ancestry. YRI (in green) refers to Yoruba in Ibadan, Nigeria.

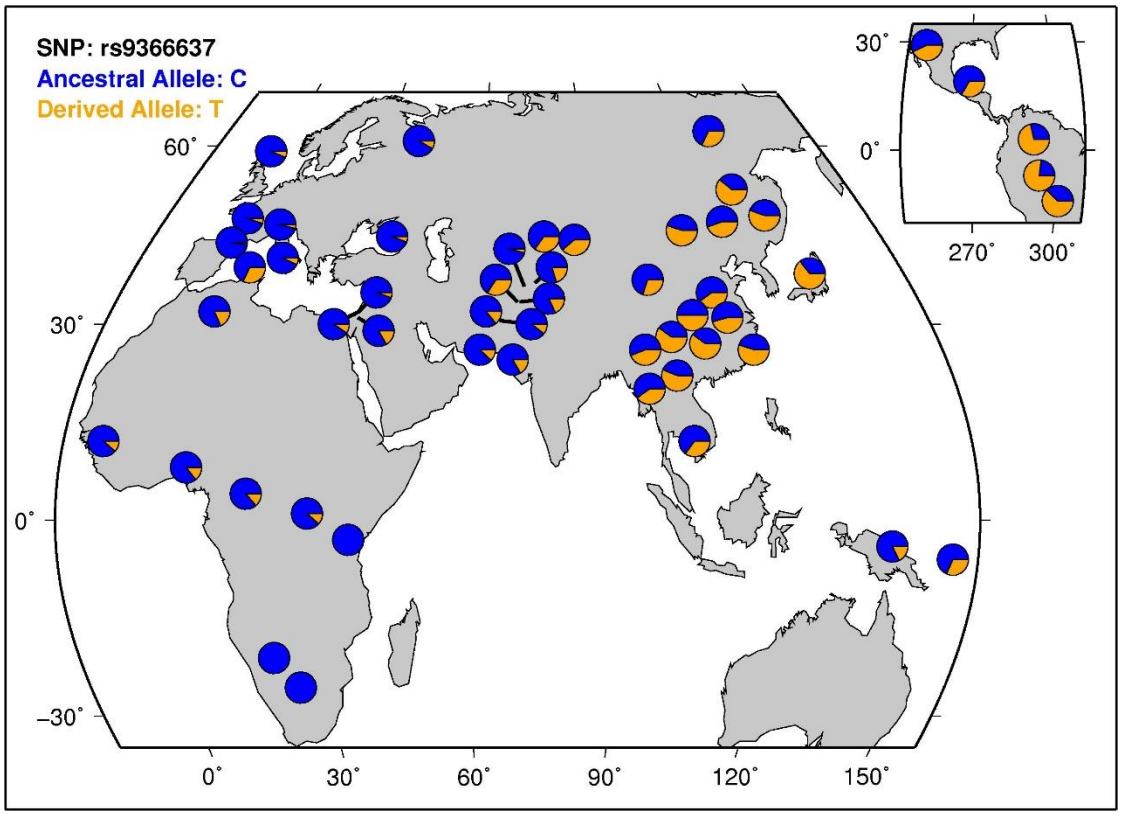


Figure 3.3. The frequency distribution of the tag SNP rs9366637 in global populations. The derived allele T is on the Asian-common haplotype.

3.4. Results

3.4.1. Evolutionary analysis and gene expression analysis

Haplotype analysis on *HFE* in 487 individuals from three continents revealed 22 haplotypes (Figure. 3.1.). One haplotype was identified that had a frequency of 52.35~54.71% in Asian samples, 5.98% in Northern European samples and 4.35% in West African samples. This Asian-common haplotype could be tagged by a single SNP, rs9366637 (ancestral allele C, derived allele T). Among haplotypes carrying T at rs9366637 in all samples, 95.80% were the Asian-common haplotype. Evolutionary analysis using the iHS method revealed a signal of positive selection on *HFE* only among the Asian samples but this signal was absent in the European or African samples (empirical *p* value = 0.011, Figure 3.2.). Specifically, the tag SNP, rs9366637, has an extreme iHS value of -2.89, suggesting that its derived allele, T, is on an unusually long haplotype, probably as a result of positive selection. Consistently, the frequency of T is high in Asian populations but low in Africans and Europeans (Figure 3.3.). The population frequency divergence is also reflected by the F_{ST} statistic, which measures population differentiation and ranges from 0 to 1 with higher values corresponding to higher degrees of differentiation. While the genome-wide mean F_{ST} value across all inter-population comparisons is about 0.11 (43), the value for rs9366637 is 0.23 between Asian and European populations, and 0.279 between Asian and African populations.

The relationship between the genotype of rs9366637 and the *HFE* expression level was assessed using the genotype and gene expression data of Asian individuals in the

International Hapmap Project. Individuals carrying the genotype C/T or T/T had a significantly lower expression level of *HFE* in lymphoblastoid cell lines (p = 0.02, 0.025, respectively, Figure 3.4.), suggesting that the Asian-common haplotype may carry a regulatory variant(s) that down-regulate the expression of *HFE*.

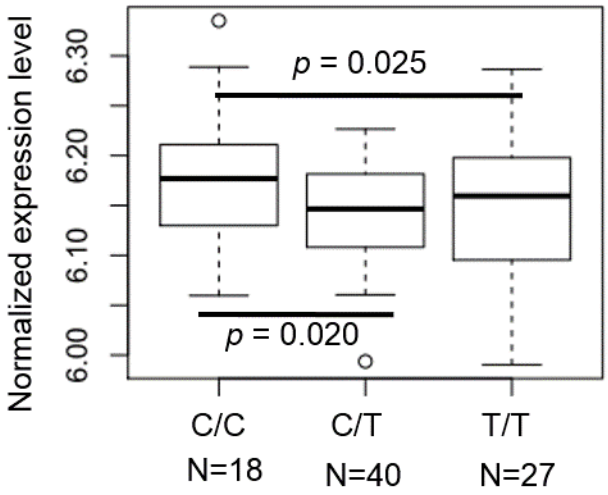


Figure 3.4. The Asian-common *HFE* haplotype is associated with lower expression level. Normalized expression level of *HFE* was measured in the lymphoblastoid cell lines of 43 Chinese individuals from Beijing, China and 42 Japanese individuals from Tokyo, Japan. Individuals with the C/C genotype had significantly higher *HFE* expression when compared to those with the C/T or T/T genotype (Student's t test, $p<0.05$).

3.4.2. Genotype and iron status screening

Women enrolled in the genotype and Fe status screening study were all of Asian descent. Of the 57 women studied, 82% were of Chinese descent (n=47), 11% of South Korean descent (n=6), 4% of Vietnamese descent (n=2), 2% of Japanese (n=1) and 2% of Thai descent (n=1). Of these 57 women, 13 individuals (23%) were C/C at locus rs9366637, 11 T/T (19%) and 33 C/T (58%). The observed frequency of allele T, 48.2%, is close to that observed in Asian samples from the Hapmap project (52.35~54.71%). General characteristics of the study participants are shown in Table 3.1. No significant differences were evident in subject characteristics as a function of genotype (C/C, T/T and C/T). All women had normal serum concentrations of folate (>5 ng/mL), vitamin B-12 (>200 pg/mL) and CRP (< 10 mg/L). Ten women (17.5%) had biochemical evidence of ID: two had depleted Fe stores (SF < 12 ug/L), seven had early functional Fe deficiency (sTfR > 8.5 mg/L) and one had anemia (Hb < 12 g/dL). There was no overlap among these ten individuals. Only one individual had TBI < 0 mg/kg (-1.54 mg/kg) and this individual also exhibited elevated sTfR (8.7 mg/L). Relationships among Fe status indicators were tested. As expected, significant positive correlations were evident between Hb and SF (p = 0.035, R^2 = 0.06). Serum hepcidin concentrations were also significantly positively associated with several Fe status indicators (Figure 3.5.): Hb (p = 0.048, R^2 = 0.05), SF (p = 8.52e-08, R^2 = 0.40), and TBI (p = 5.96e-04, R^2 = 0.18). The only Fe status marker with suggestive evidence of a difference between C/C and T/T genotypes was sTfR, which was higher in T/T women (p for trend = 0.06).

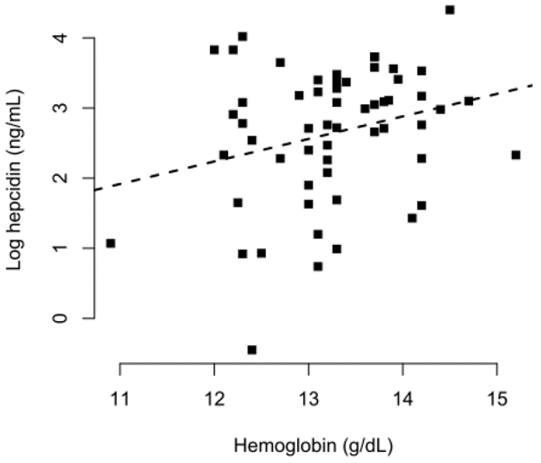
Table 3.1. General characteristics and iron status indicators of the 57 study participants as a function of their genotype at rs9366637

Variable	Entire Study Population (n=57)	C/C Genotype (N=13)	C/T Genotype (N=33)	T/T Genotype (N=11)
Age (y)	22.7 ± 3.5 (18-34)	24.0 ± 4.8 (18-34)	22.7 ± 3.3 (19-31)	21.2 ± 1.8 (18-24)
Weight (kg)	53.72± 7.4 (39-75)	53.9 ± 7.8 (44-75)	54.0 ± 7.9 (39-72)	52.7 ± 5.7 (45-61)
BMI (kg/m^2)	20.8 ± 2.5 (16.5-31.2)	21.7 ± 3.2 (18.2-31.2)	20.7 ± 2.3 (16.5-26.7)	19.9 ± 2.1 (16.5-24.2)
Hemoglobin (g/dL)	13.3 ± 0.8 (10.9-15.2)	13.2 ± 0.6 (12.3-14.2)	13.4 ± 0.8 (12-15.2)	13.0 ± 1.0 (10.9-14.4)
Folate (ng/mL)	15.5 ± 4.5 (7.7-28.8)	16.8 ± 4.6 (9.6-23.9)	14.7 ± 4.0 (7.7-23.1)	16.2 ± 5.7 (8.1-28.8)
Vitamin B-12 (pg/mL)	603.2 ± 255.7 (204.5-1780.5)	527.3 ± 223.8 (204.5-956)	616.0 ± 188.2 (251-994)	654.4 ± 426.3 (308.5-1780.5)
C-reactive protein (mg/L)	0.5 ± 0.6 (<0.2-3.0)	0.8 ± 0.9 (<0.2-3.0)	0.4 ± 0.4 (<0.2-2.4)	0.6 ± 0.8 (<0.2-2.6)
Serum ferritin (ug/L)	47.0 ± 35.0 (6.5-183.5)	38.0 ± 27.7 (10.5-109.1)	51.0 ± 36.2 (14.8-183.5)	45.6 ± 39.7 (6.5-152.8)
Serum transferrin receptor (mg/L)	4.8 ± 2.8 (1.9-15.6)	4.0 ± 1.7 (1.9-7.5)	4.9 ± 3.1 (2.4-15.6)	5.5 ± 2.8 (2.0-11.7)
Total body iron (mg/kg)	6.2 ± 3.1 (-1.5-11.6)	6.1 ± 3.3 (1.0-11.6)	6.6 ± 2.9 (-1.5-11.6)	5.3 ± 3.4 (0.2-11.4)
Serum hepcidin (ng/mL)	20.1 ± 15.6 (0.6-81.7)	16.3 ± 11.4 (2.1-36.0)	23.1 ± 17.0 (2.5-81.7)	15.7 ± 14.2 (0.6-46.3)

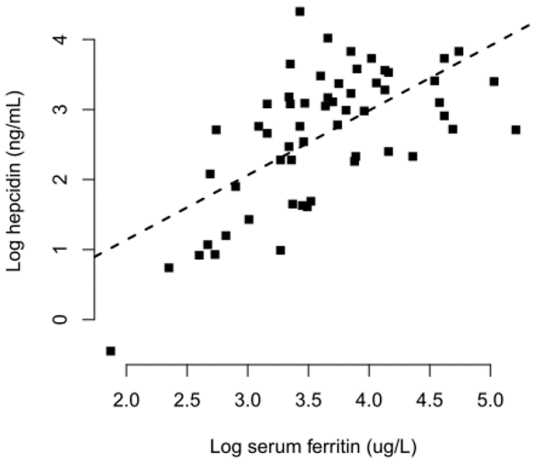
Data are expressed as the mean ± SD with the range in parentheses. No significant differences were found between genotypes.

Figure 3.5. The correlation between hepcidin and iron status markers. Iron status markers include A) hemoglobin, B) serum ferritin, and C) total body iron in 57 Asian women participating in the genotype and iron status screening study. Hepcidin and serum ferritin were transformed by natural logarithm. Black dashed lines represent the linear regression line. The correlations between iron status markers and hepcidin were significant: hemoglobin (p=0.048, R^2=0.05), serum ferritin (p=8.52e-08, R^2=0.40), and total body iron (p=5.96e-04, R^2=0.18).

A


Log hepcidin (ng/mL)
0
1
2
3
4
11
12
13
14
15
Hemoglobin (g/dL)

B


Log hepcidin (ng/mL)
0
1
2
3
4
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Log serum ferritin (ug/L)

C

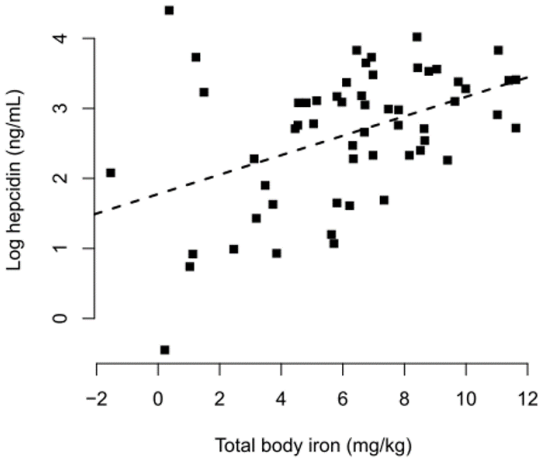

Log hepcidin (ng/mL)
0
1
2
3
4
−2
0
2
4
6
8
10
12
Total body iron (mg/kg)

Table 3.2. Iron status indicators of study participants on the dosing day

Variable	Total	C/C	T/T
Hemoglobin (g/dL)	13.5 ± 1.6 (10.7-17.4, N=13)	14.38 ± 1.93 (12.4-17.4, N=5)	12.86 ± 1.13 (10.7-14.4, N=8)
Serum ferritin (ug/L)	36.2 ± 32.0 (6.3-141.3, N=21)	35.1 ± 27.69 (8.6-97.3, N=10)	37.27 ± 36.80 (6.3-141.3, N=11)
Serum transferrin receptor (mg/L)	4.5 ± 2.4 (2.0-11.3, N=18)	3.36 ± 1.50 (2.01-6.32, N=9)	5.58 ± 2.76 * (2.48-11.33, N=9)
Total body iron (mg/kg)	5.3 ± 3.6 (0.3-11.9, N=18)	6.31 ± 3.81 (0.7-11.9, N=9)	4.21 ± 3.23 (0.32-11.38, N=9)
Serum hepcidin (ng/mL)	13.4 ± 14.2 (0.9-52.3, N=14)	12.9 ± 19.6 (1.1-52.3, N=6)	13.76 ± 9.78 (0.89-26.5, N=8)

Data are expressed as the mean ± SD with the range in parentheses. Missing data are due to lack of sample volume in samples obtained by finger stick. * indicates $p < 0.05$ in one-tailed Wilcoxon rank sum test between the two genotypes.

Table 3.3. Iron status indicators in Asian women participating in the iron absorption study

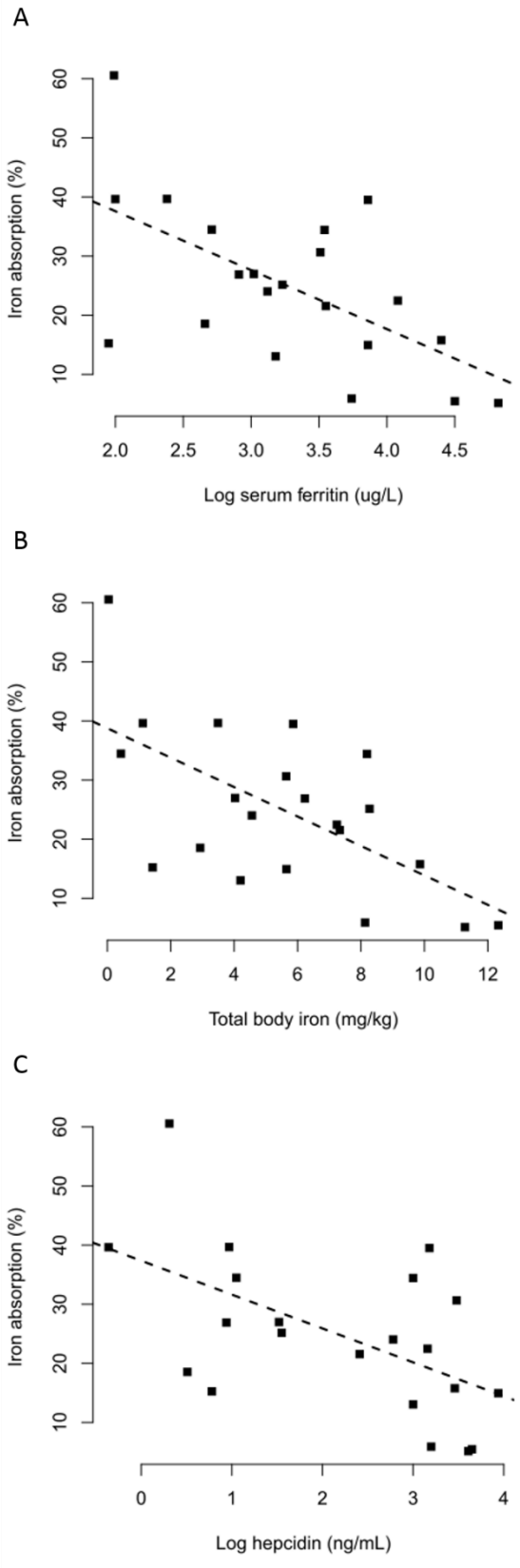
Variable	Total (N=21)	C/C (N=10)	T/T (N=11)
Hemoglobin (g/dL)	13.1 ± 1.1 (11.1-14.7)	12.8 ± 0.8 (11.3-14.1)	13.3 ± 1.2 (11.1-14.7)
Folate (ng/mL)	20.3 ± 8.3 (9.2-41.8)	21.6 ± 8.5 (12.2-41.8)	19 ± 8.3 (9.2-34.7)
Vitamin B-12 (pg/mL)	651.3 ± 373.9 (235.5-1692.5)	549.3 ± 244.2 (235.5-925.5)	744.1 ± 453.9 (316-1692.5)
C-reactive protein (mg/L)[a]	0.3 ± 0.3 (<0.2-1.0)	0.4 ± 0.3 (<0.2-1.0)	0.3 ± 0.2 (<0.2-0.7)
Serum ferritin (ug/L)	36.5 ± 30.4 (7-124)	36.3 ± 28.3 (7-89.7)	36.7 ± 33.6 (7.4-124)
Serum transferrin receptor (mg/L)	4.0 ± 1.7 (1.7-8.9)	3.2 ± 1.2 (1.7-5.0)	4.76 ± 1.81* (2.7-8.9)
Total body iron (mg/kg)	5.6 ± 3.4 (0.0-12.3)	6.4 ± 3.8 (0.0-12.3)	4.9 ± 3.1 (0.4-11.3)
Serum hepcidin (ng/mL)	16.9 ± 15.2 (0.7-51.5)	17.8 ± 14.2 (1.4-38.5)	16.0 ± 16.7 (0.7-51.5)
Absorption (%)	24.8 ± 13.7 (5.1-60.6)	23.3 ± 16.4 (5.5-60.6)	26.1 ± 11.3 (5.1-39.7)
Normalized absorption (%)[b]	15.6 ± 6.3 (3.3-28.1)	14.0 ± 6.2 (3.3-25.4)	17.1 ± 6.3 (6.2-28.1)

Data are expressed as the mean ± SD with the range in parentheses in indicators measured in blood obtained 2-weeks post-dosing. * indicates $p < 0.05$ in one-tailed Wilcoxon rank sum test between the two genotypes.

[a] The sample size for C-reactive protein is 6 for C/C and 8 for T/T due to lack of blood samples.

[b] Percent absorption was normalized to a fixed serum ferritin concentration of 40 ug/L.

Figure 3.6. The correlation between iron absorption and iron status markers. Iron status markers include A) serum ferritin, B) total body iron, C) hepcidin in 21 Asian women who participated in a stable Fe isotope absorption study. Concentrations of serum ferritin, total body iron and hepcidin were measured in blood obtained 2 weeks after the stable iron isotope was administered. Hepcidin and serum ferritin were transformed by natural logarithm. The black dashed line represents the linear regression line. The correlations were all significant: serum ferritin (p=0.004, R^2=0.33), total body iron (p=0.003, R^2=0.35) and hepcidin (p=0.009, R^2=0.27).


A
Iron absorption (%)
10
20
30
40
50
60
2.0
2.5
3.0
3.5
4.0
4.5
Log serum ferritin (ug/L)
B
Iron absorption (%)
10
20
30
40
50
60
0
2
4
6
8
10
12
Total body iron (mg/kg)
C
Iron absorption (%)
10
20
30
40
50
60
0
1
2
3
4
Log hepcidin (ng/mL)

3.4.3. Iron absorption analysis

Of the 57 Asian women who participated in the genotype screening study, all C/C and T/T women were invited to participate in a Fe absorption study. Of those invited, 10 women with the C/C genotype (8 Chinese, 1 Vietnamese, and 1 Korean) and 11 women with the T/T genotype (9 Chinese and 2 Korean) consented to participate in the study. The average time interval between the screening and the absorption study was 68 days. Paired measures of Fe status markers (Hb, SF, sTfR, TBI, serum hepcidin) obtained at screening, on the day the Fe tracer was administered and 2-weeks post-dosing did not significantly differ within this group of women (Table 3.2. and 3.3.). No significant differences in Hb, SF, TBI and serum hepcidin were evident between the two genotypes. However, women with the T/T genotype exhibited significantly higher sTfR concentrations in serum obtained at dosing and 2-weeks post-dosing (p = 0.026 and 0.021, respectively) compared to women with the C/C genotype. In women that returned for the dosing study, the associations between serum hepcidin and Fe status indicators became even more significant: Hb (p = 0.018, R^2=0.22), SF (p = 4.24e-08, R^2 = 0.79), and TBI (p = 9.96e-05, R^2 = 0.53), probably resulting from the lower standard deviation in Fe status evident in the Fe absorption cohort versus that observed in the larger group of 57 women.

As expected, the average percent Fe absorption in these 21 women was highly variable (ranging from 5.1% to 60.6%) and was inversely associated with Fe status indicators, including SF (p = 0.004, R^2 = 0.33, Figure 3.6. A) and TBI (p = 0.003, R^2 = 0.35, Figure 3.6. B). Fe absorption was also negatively associated with serum hepcidin (p =

0.009, $R^2 = 0.27$, Figure 3.6. C). In order to control for the well-known effect of Fe status on intestinal Fe absorption efficiency, Fe absorption data were normalized to a fixed serum ferritin concentration of 40 ug/L (40). Mean normalized Fe absorption data are presented in Table 3.3. After normalization, elevated percent Fe absorption was evident in T/T group, although the difference only approached significance (*p* for trend = 0.099). The absolute difference in percent Fe absorption between the two genotypes was 3.1%, which reflects an average increase of 22% in women with T/T genotype when compared with C/C genotype in the face of similar Fe stores.

3.4.4. Iron absorption between Asians and Caucasians

Fe status and absorption of non-heme Fe in our sample of women of Asian descent were compared to data obtained from 18 healthy, similarly aged Caucasian women recruited from the same university campus using similar enrollment criteria and investigated with the same experimental approaches (39). Comparison of characteristics between the groups are presented in Table 3.4. The two groups did not significantly differ with respect to age, concentrations of folate, vitamin B-12 and CRP but women of Caucasian descent had significantly higher BMI ($p = 0.003$). Also, we observed that women of Asian descent had significantly higher Fe status as evidenced by the 5% higher Hb concentrations ($p = 0.0094$) and the 24% higher SF concentrations ($p = 0.045$). There was no significant association between BMI and Hb (or SF) in either cohort or in the combined study sample.

Percent Fe absorption normalized to a fixed SF concentration (40 ug/L) was significantly higher among Asians independent of genotype when compared to the

Caucasian cohort (Figure 3.7.). Caucasian women absorbed 12.02 ± 15.29% of the supplemental non-heme Fe, while Asian women with C/C genotype absorbed 16% more ($p = 0.0078$) and those with T/T genotype absorbed 42% more ($p = 0.0018$). Averaged over the two genotypes, Asian women absorbed 30% more non-heme Fe than Caucasians ($p = 0.00053$). An association between BMI and normalized percent Fe absorption approached significance ($p = 0.058$) in the combined study sample but after controlling for the ethnic identity, there was no significant association between the two variables ($p = 0.19$).

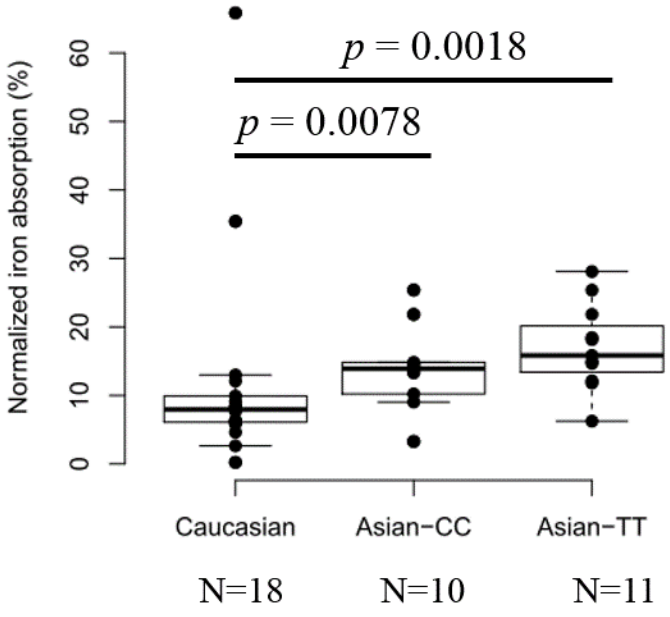


Figure 3.7. Higher iron absorption in Asian than in Caucasian women. Percent iron absorption normalized to a fixed serum ferritin concentration is presented from data obtained in 21 Asian and 18 Caucasian women. The Asian women were split into two groups based on their genotype at SNP rs9366637. Significantly higher percent iron absorption was evident in Asians of both genotypes.

Table 3.4. Subject characteristics in Asian and Caucasian samples

Variable	Asian (N=57)	Caucasian (N=18)
Age (y)	22.7 ± 3.5 (18-34)	22.5 ± 3.1 (18-32)
BMI (kg/m^2)	20.8 ± 2.6 (16.5-31.2)	22.8 ± 2.8** (18.7-30.3)
Hemoglobin (g/dL)	13.3 ± 0.8 (10.9-15.2)	12.6 ± 1.3** (11-15.6)
Folate (ng/mL)	15.5 ± 4.5 (7.7-28.8)	18.4 ± 5.5 (11.5-31.9)
Vitamin B-12 (pg/mL)	603.2 ± 255.7 (204.5-1780.5)	594.6 ± 255.6 (241-1088)
C-reactive protein (mg/L)	0.5 ± 0.6 (<0.2-3.02)	2.9 ± 4.5 (<0.2-15)
Serum ferritin (ug/L)	47.0 ± 35.0 (6.5-183.5)	38.0 ± 34.1* (5.7-119.7)
Serum transferrin receptor (mg/L)	5.4 ± 4.7 (1.9-26.4)	4.7 ± 1.3 (2.8-7.1)
Total body iron (mg/kg)	6.2 ± 3.1 (-1.5-11.6)	5.1 ± 3.4 (-1.5-11.7)
Serum hepcidin (ng/mL)	20.1 ± 15.6 (0.64-81.72)	NA

Data are expressed as the mean ± SD with the range in parentheses. * indicates $p < 0.05$, ** indicates $p < 0.01$ (one-tailed Wilcoxon rank sum test for hemoglobin, serum ferritin, serum transferrin receptor, total body iron, hepcidin; two-tailed Wilcoxon rank sum test for other parameters)

3.5. Discussion

This study utilized a multidisciplinary approach combining evolutionary analyses and biomedical mass spectrometry to identify genetic variants in *HFE* that may have a functional impact on non-heme Fe absorption in Asian populations. Evolutionary analyses identified an Asian-common haplotype in the *HFE* gene that may have been under positive selection during evolution. Gene expression analyses found significantly lower transcription of *HFE* among carriers of this haplotype. In support of a possible evolutionary advantage, Asian subjects homozygous for this haplotype exhibited higher non-heme Fe absorption after adjusting to a fixed amount of storage Fe. Additionally, Asian women exhibited both significantly higher Fe status and significantly higher stores-adjusted Fe absorption when compared to age-matched Caucasian women, supportive of a population differential genetic basis for Fe homeostasis.

Our evolutionary analyses using genotype data from multiple continental populations confirmed the presence of an Asian-common *HFE* haplotype and positive selection signals on *HFE* in Asian populations (28, 29). We hypothesized that this frequency shift was due to the selection pressure from the traditional plant-based diet in Asia, which is known to be low in bioavailable Fe (44, 45). Genetic mutations that enhance the ability to absorb non-heme Fe from a plant-based diet may have increased the ability to maintain adequate Fe status and thus incurred positive natural selection. In support of this hypothesis, individuals carrying the Asian-common haplotype exhibited reduced *HFE* expression. A reduced expression of *HFE* theoretically would

lower the expression of hepcidin and consequently lead to enhanced Fe absorption (9). A limitation with our study design is that while *HFE* mainly functions in hepatocytes (46), our gene expression data were evaluated using lymphoblastoid cell lines and may not represent the expression pattern in other tissues.

To directly examine the impact of *HFE* haplotypes on Fe absorption, a stable Fe isotope absorption study was undertaken. Young women homozygous for the Asian-common haplotype exhibited elevated non-heme Fe absorption when compared to women homozygous for other haplotypes after controlling for SF concentration. However, the difference observed only approached significance likely due to our limited sample size. Power analysis based on our data suggested that a sample size of 54 individuals per genotype would be required to detect this difference as significant with 80% power. While the 3% absolute difference in absorption may not appear substantial, it is actually equivalent to a 14~20 ug/L decrease in SF concentration based on the relationship between SF and percent Fe absorption observed in this and other previous studies (47). Considering that there are no physiological regulatable routes of Fe excretion and excess Fe accumulation continues across the lifespan (48), a 22% increase in percent Fe absorption might have significant health implications if maintained across the lifecycle.

Fe status is one of the major determinants of non-heme Fe absorption, with either SF or serum hepcidin explaining approximately 30% of the variation in intestinal Fe absorption (39, 49, 50). High Fe status leads to increased hepcidin expression, which inhibits intestinal Fe absorption and suppresses the release of Fe from body stores (9).

On the other hand, ID or increased erythropoiesis reduces hepcidin expression to allow for enhanced Fe mobilization and utilization (9, 51). Testing for potential differences in biochemical indicators of Fe status between the two genotypes, only one marker, sTfR, was significantly impacted. Elevated sTfR was observed in individuals homozygous for the Asian-common haplotype. The sTfR is a truncated version of the membrane-bound transferrin receptor (TfR1) and its concentration is directly proportional to the amount of TfR1 on cell membranes, especially those of erythroid precursors (51). TfR1 binds not only to Fe-loaded transferrin but also to HFE on the membrane of hepatocyte. Since TfR1 has a higher affinity for transferrin, when Fe-loaded transferrin is abundant, HFE is released from TfR1 and binds to transferrin receptor 2 (TfR2). The interaction between HFE and TfR2 is required to up-regulate hepcidin and subsequently down-regulate Fe absorption (22). Increased TfR1 on the cell membrane, as suggested by elevated sTfR, may prevent HFE from interacting with TfR2 thereby enhancing Fe absorption. Elevated sTfR has been associated with increased Fe absorption in some (51-54), but not all studies (35, 39, 55) examining these relationships. Significant associations between sTfR and serum hepcidin have typically not been reported (39, 55, 56). In our data, we did not observe any significant correlations between sTfR and Fe absorption or serum hepcidin. The effect of elevated sTfR on non-heme Fe absorption is still unclear. Additionally, the observed hepcidin concentration did not significantly differ between the two genotypes, probably due to the large variation and our limited sample size. Larger sample sizes will be needed to identify the underlying mechanisms responsible for increased Fe absorption in individuals homozygous for this Asian-common haplotype.

Our population comparison analysis identified significantly higher Fe status (Hb and SF) and absorption in Asian females when compared to age-matched healthy Caucasian females recruited from the same city. To our knowledge, this is the first study comparing iron absorption between Asian and Caucasian. These findings support other published data on population differences in Fe status (12, 17). Specifically, individuals of African descent tend to have higher Fe status when compared to whites and Hispanics of comparable age and sex (17). Similarly the United States Centers for Disease Control and Prevention has set race-specific guidelines that define anemia using a lower Hb concentration in African Americans (57). Our analysis revealed that in spite of improved Fe status, Asian women still had significantly higher SF-normalized Fe absorption. This is very surprising given that ID is prevalent in Asian populations (44, 45). Enhanced non-heme Fe absorption may have been a selective advantage that was once beneficial under conditions of limited Fe availability, but may now prove detrimental in the modern dietary environment. Iron absorption enhancing genetic variants, coupled with Fe-rich diets, may contribute to the increased risk of Fe-related complex diseases, such as type 2 diabetes. Higher SF concentrations have been linked to a higher disease risk for diabetes (18-20) and Asian Americans are 30-50% more likely to suffer from this disorder than their Caucasian counterparts (58). Future studies are needed to assess the impact of population differences in genes associated with Fe homeostasis and to assess their functional impact on Fe absorption and Fe-related complex disorders.

3.6. Acknowledgements

We thank Tera Kent for technical laboratory support, and Kevin Klatt, Drs. Patrick Stover, Andy Clark and Frank Hu for commenting on the manuscript. This work was supported by the 2011 seed grant from Center for Vertebrate Genomics, Cornell University.

3.7. References

1. Ye K & Gu Z (2011) Recent advances in understanding the role of nutrition in human genome evolution. *Adv Nutr* 2(6): 486-496.

2. Babbitt CC, Warner LR, Fedrigo O, Wall CE & Wray GA (2011) Genomic signatures of diet-related shifts during human origins. *Proc Biol Sci* 278(1708): 961-969.

3. Luca F, Perry GH & Di Rienzo A (2010) Evolutionary adaptations to dietary changes. *Annu Rev Nutr* 30: 291-314.

4. Tishkoff SA, *et al* (2007) Convergent adaptation of human lactase persistence in africa and europe. *Nat Genet* 39(1): 31-40.

5. Perry GH, *et al* (2007) Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 39(10): 1256-1260.

6. Peng Y, *et al* (2010) The ADH1B Arg47His polymorphism in east asian populations and expansion of rice domestication in history. *BMC Evol Biol* 10: 15-2148-10-15.

7. Vitti JJ, Grossman SR & Sabeti PC (2013) Detecting natural selection in genomic data. *Annu Rev Genet* 47(1): 97-120.

8. Grossman SR, *et al* (2013) Identifying recent adaptations in large-scale genomic data. *Cell* 152(4): 703-713.

9. Viatte L & Vaulont S (2009) Hepcidin, the iron watcher. *Biochimie* 91(10): 1223-1228.

10. Munoz M, Garcia-Erce JA & Remacha AF (2011) Disorders of iron metabolism. part 1: Molecular basis of iron homoeostasis. *J Clin Pathol* 64(4): 281-286.

11. Hare D, Ayton S, Bush A & Lei P (2013) A delicate balance: Iron metabolism and diseases of the brain. *Front Aging Neurosci* 5: 34.

12. UNSCN (2010) 6th report on the world nutrition situation

13. Fretham SJ, Carlson ES & Georgieff MK (2011) The role of iron in learning and memory. *Adv Nutr* 2(2): 112-121.

14. Brownlie T,4th, Utermohlen V, Hinton PS & Haas JD (2004) Tissue iron deficiency without anemia impairs adaptation in endurance capacity after aerobic training in previously untrained women. *Am J Clin Nutr* 79(3): 437-443.

15. Zimmermann MB, Chaouki N & Hurrell RF (2005) Iron deficiency due to consumption of a habitual diet low in bioavailable iron: A longitudinal cohort study in moroccan children. *Am J Clin Nutr* 81(1): 115-121.

16. Zimmermann MB & Hurrell RF (2007) Nutritional iron deficiency. *Lancet* 370(9586): 511-520.

17. Zacharski LR, Ornstein DL, Woloshin S & Schwartz LM (2000) Association of age, sex, and race with body iron stores in adults: Analysis of NHANES III data. *Am Heart J* 140(1): 98-104.

18. Acton RT, *et al* (2006) Relationships of serum ferritin, transferrin saturation, and HFE mutations and self-reported diabetes in the hemochromatosis and iron overload screening (HEIRS) study. *Diabetes Care* 29(9): 2084-2089.

19. Sun L, *et al* (2008) Ferritin concentrations, metabolic syndrome, and type 2 diabetes in middle-aged and elderly chinese. *J Clin Endocrinol Metab* 93(12): 4690-4696.

20. Sun L, *et al* (2013) Elevated plasma ferritin is associated with increased incidence of type 2 diabetes in middle-aged and elderly chinese adults. *J Nutr* 143(9): 1459-1465.

21. Feder JN, *et al* (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet* 13(4): 399-408.

22. Gao J, *et al* (2009) Interaction of the hereditary hemochromatosis protein HFE with transferrin receptor 2 is required for transferrin-induced hepcidin expression. *Cell Metab* 9(3): 217-227.

23. Lucotte G & Dieterlen F (2003) A european allele map of the C282Y mutation of hemochromatosis: Celtic versus viking origin of the mutation?. *Blood Cells Mol Dis* 31(2): 262-267.

24. Distante S, *et al* (2004) The origin and spread of the HFE-C282Y haemochromatosis mutation. *Hum Genet* 115(4): 269-279.

25. Toomajian C, Ajioka RS, Jorde LB, Kushner JP & Kreitman M (2003) A method for detecting recent selection in the human genome from allele age estimates. *Genetics* 165(1): 287-297.

26. Weinberg ED (2008) Survival advantage of the hemochromatosis C282Y mutation. *Perspect Biol Med* 51(1): 98-102.

27. Moalem S, Percy ME, Kruck TP & Gelbart RR (2002) Epidemic pathogenic selection: An explanation for hereditary hemochromatosis?. *Med Hypotheses* 59(3): 325-329.

28. Toomajian C & Kreitman M (2002) Sequence variation and haplotype structure at the human HFE locus. *Genetics* 161(4): 1609-1623.

29. Williamson SH, *et al* (2007) Localizing recent adaptive evolution in the human genome. *PLoS Genet* 3(6): e90.

30. International HapMap 3 Consortium, *et al* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7311): 52-58.

31. Bandelt HJ, Forster P & Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16(1): 37-48.

32. Voight BF, Kudaravalli S, Wen X & Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4(3): e72.

33. Pickrell JK, *et al* (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 19(5): 826-837.

34. Stranger BE, *et al* (2007) Population genomics of human gene expression. *Nat Genet* 39(10): 1217-1224.

35. Roe MA, *et al* (2005) Iron absorption in male C282Y heterozygotes. *Am J Clin Nutr* 81(4): 814-821.

36. Hunt JR & Zeng H (2004) Iron absorption by heterozygous carriers of the HFE C282Y mutation associated with hemochromatosis. *Am J Clin Nutr* 80(4): 924-931.

37. Cook JD, Flowers CH & Skikne BS (2003) The quantitative assessment of body iron. *Blood* 101(9): 3359-3364.

38. Kastenmayer P, *et al* (1994) A double stable isotope technique for measuring iron absorption in infants. *Br J Nutr* 71(3): 411-424.

39. Young MF, *et al* (2009) Serum hepcidin is significantly associated with iron absorption from food and supplemental sources in healthy young women. *Am J Clin Nutr* 89(2): 533-538.

40. IAEA HUMAN HEALTH SERIES No.21 (2012) *Assessment of Iron Bioavailability in Humans Using Stable Iron Isotope Techniques*, (International Atomic Energy Agency, pp 43-44.

41. R Core Team (2014) R: A language and environment for statistical computing. R foundation for statistical computing, vienna, austria. URL: Http://Www.R-project.org/.

42. Bauer DF (1972) Constructing confidence sets using rank statistics. *Journal of the American Statistical Association* 67(339): 687-690.

43. Barreiro LB, Laval G, Quach H, Patin E & Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. *Nat Genet* 40(3): 340-345.

44. Mason J, Mannar V & Mock N (1999) Controlling micronutrient deficiencies in asia. *Asian Development Review* 17: 66-95.

45. Ma G, *et al* (2008) Iron and zinc deficiencies in china: What is a feasible and cost-effective strategy?. *Public Health Nutr* 11(6): 632-638.

46. Vujic Spasic M, *et al* (2008) Hfe acts in hepatocytes to prevent hemochromatosis. *Cell Metab* 7(2): 173-178.

47. O'Brien KO, Zavaleta N, Caulfield LE, Yang DX & Abrams SA (1999) Influence of prenatal iron and zinc supplements on supplemental iron absorption, red blood cell iron incorporation, and iron status in pregnant peruvian women. *Am J Clin Nutr* 69(3): 509-515.

48. Institute of Medicine (2001) *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc,* (National Academy Press,

49. Roe MA, Collings R, Dainty JR, Swinkels DW & Fairweather-Tait SJ (2009) Plasma hepcidin concentrations significantly predict interindividual variation in iron absorption in healthy men. *Am J Clin Nutr* 89(4): 1088-1091.

50. Collings R, *et al* (2013) The absorption of iron from whole diets: A systematic review. *Am J Clin Nutr* 98(1): 65-81.

51. Li H & Ginzburg YZ (2010) Crosstalk between iron metabolism and erythropoiesis. *Adv Hematol* 2010: 605435.

52. Hunt JR (2003) High-, but not low-bioavailability diets enable substantial control of women's iron absorption in relation to body iron stores, with minimal adaptation within several weeks. *Am J Clin Nutr* 78(6): 1168-1177.

53. Skikne BS (2008) Serum transferrin receptor. *Am J Hematol* 83(11): 872-875.

54. Tussing-Humphreys L, Pusatcioglu C, Nemeth E & Braunschweig C (2012) Rethinking iron regulation and assessment in iron deficiency, anemia of chronic disease, and obesity: Introducing hepcidin. *J Acad Nutr Diet* 112(3): 391-400.

55. Young MF, *et al* (2010) Utilization of iron from an animal-based iron source is greater than that of ferrous sulfate in pregnant and nonpregnant women. *J Nutr* 140(12): 2162-2166.

56. Dallalio G, Fleury T & Means RT (2003) Serum hepcidin in clinical specimens. *Br J Haematol* 122(6): 996-1000.

57. Institute of Medicine (1993) *Iron deficiency anemia: recommended guidelines for the prevention, detection, and management among U.S. children and women of childbearing age,* (The National Academies Press, Washington, DC),

58. Lee JW, Brancati FL & Yeh HC (2011) Trends in the prevalence of type 2 diabetes in asians versus whites: Results from the united states national health interview survey, 1997-2008. *Diabetes Care* 34(2): 353-357.

Chapter 4 – Extensive Pathogenicity of Mitochondrial Heteroplasmy in Healthy Human Individuals[4]

4.1. Abstract

A majority of mitochondrial DNA (mtDNA) mutations reported to be implicated in diseases are heteroplasmic, a status with co-existing mtDNA variants in a single cell. Quantifying the prevalence of mitochondrial heteroplasmy and its pathogenic effect in healthy individuals could further our understanding of its possible roles in various diseases. 1085 human individuals from 14 global populations have been sequenced by the 1000 Genomes Project to a mean coverage of ~2000X on mtDNA. Using a combination of stringent thresholds and a maximum likelihood method to define heteroplasmy, we demonstrated that ~90% of the individuals carry at least one heteroplasmy. At least 20% individuals harbor heteroplasmies reported to be implicated in disease. Mitochondrial heteroplasmy tend to show high pathogenicity, and is significantly over-represented in disease-associated loci. Consistent with their deleterious effect, heteroplasmies with derived allele frequency larger than 60% within an individual show a significant reduction in pathogenicity, indicating the action of purifying selection. Purifying selection on heteroplasmies can also be inferred from non-synonymous and synonymous heteroplasmy comparison and the unfolded site frequency spectra for different functional sites in mtDNA. Nevertheless, in comparison to population polymorphic mtDNA mutations, the purifying selection is

[4] Published on *Proceedings of the National Academy of Sciences*. See Appendix A for inclusion authorization.

much less efficient in removing heteroplasmic mutations. The prevalence of mitochondrial heteroplasmy with high pathogenic potential in healthy individuals, along with the possibility of these mutations drifting to high frequency inside a subpopulation of cells across life-span, emphasizes the importance of managing mitochondrial heteroplasmy to prevent disease progression.

4.2. Significance Statement

There are hundreds to thousands of copies of mitochondrial DNA (mtDNA) in each human cell in contrast to only two copies of nuclear DNA. High-frequency pathogenic mtDNA mutations have been found in patients with classic mitochondrial diseases, pre-mature aging, cancers and neurodegenerative diseases. In this study we investigated the distribution of heteroplasmic mutations, their pathogenic potential and their underlying evolutionary forces using genome sequence data from the 1000 Human Genome Project. Our results demonstrated the prevalence of low-frequency high-pathogenic-potential mtDNA mutations in healthy human individuals. These deleterious mtDNA mutations, when reaching high frequency, could provide a likely source of mitochondrial dysfunction. Managing the expansion of deleterious mtDNA mutations could be a promising means of preventing disease progression.

4.3. Introduction

Hundreds to thousands of copies of mitochondrial DNA (mtDNA) are present in each single human cell, in contrast to only two copies of nuclear DNAs. These mtDNAs can differ from each other as a result of inherited or somatic mutations. The co-existence of multiple mtDNA variants in a single cell or among cells within an individual is called heteroplasmy (1). Mitochondrial heteroplasmy has been shown to be implicated in a large spectrum of human diseases. Besides classical mitochondrial diseases such as mitochondrial myopathy, myoclonic epilepsy with ragged red fibers (MERRF), and mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes (MELAS), mitochondrial heteroplasmy also plays roles in complex disorders including type 2 diabetes mellitus, aging, cancer, and late-onset neurodegenerative diseases (1-7). Of the over 500 mtDNA point mutations reported so far that are implicated in disease, about 55% of them were observed at known heteroplasmic sites (7). The co-existence of mutant and wild-type mtDNAs requires the pathogenic mutation to reach a frequency threshold before it could evince itself as clinical phenotypes (mitochondrial threshold effect). (4, 8).

Mitochondrial heteroplasmy is common in healthy human populations. Before the application of next-generation sequencing (NGS) technologies, most studies focused on the mtDNA control region and revealed that 6~11.6% of the population carry heteroplasmy in this region (9-11). The advent of NGS technologies enables the inquiry of mitochondrial heteroplasmy at the genome-wide scale. Several studies using these approaches allowed detection of medium- and high-frequency heteroplasmy with

minor allele frequency (MAF) higher than 9%, and it was found that 25~65% of the general population have at least one heteroplasmy across the entire mitochondrial genome (12-14). However, deeper sequencing depth at the order of thousands is required for confident identification of low-frequency heteroplasmy (MAF in the range of 1%~10%) (15, 16). Without considering these low-frequency heteroplasmy, the population prevalence of mitochondrial heteroplasmy is underestimated (12-14). On the other hand, a preliminary study with ultra-deep sequencing (4158~20,803X) of two ~300bp mtDNA regions was able to find heteroplasmies with very low frequency (>0.2%) in all tested healthy samples (17). Further investigation with a large sample size and deep sequencing coverage across the whole mitochondrial genome is needed to reveal the universal prevalence of mtDNA heteroplasmy in healthy human populations.

Despite the widespread presence of heteroplasmy in the healthy population, its pathogenic potential has not been well characterized and the population prevalence of pathogenic heteroplasmy might be underestimated. It has been recognized that mitochondrial heteroplasmy across the genome increases with age (9, 18-20) and acquires unique patterns in tumors (21, 22). A recent epidemiological study indicates that pathogenic mtDNA mutations might be more common in the general population than previously appreciated (23). This study investigated 10 common pathogenic mtDNA mutations in over 3000 healthy individuals and revealed that at least 1 in 200 individuals harbors a mutation that could potentially cause disease (23). This is much higher than epidemiological estimates of the prevalence of mtDNA diseases, which is only ~1 in 5000 (24). This discrepancy is likely due to the mitochondrial threshold

effect because most of the pathogenic mutations exist as heteroplasmy and are compensated by the wild-type mtDNA (4, 23). The population prevalence of pathogenic mtDNA mutations should be much higher if more reported pathogenic mutations are examined in a large population sample.

Characterizing the pathogenic potential of mitochondrial heteroplasmy in healthy individuals and its underlying evolutionary forces will further our understanding of the roles of mtDNA variations in aging, tumorigenic and neurodegenerative processes. In this study, we addressed this issue by analyzing deep sequencing data of mtDNA for 1085 healthy individuals sampled from 14 global populations in the 1000 Human Genomes Project (25). Firstly, we quantified the prevalence of mitochondrial heteroplasmy, especially disease-associated heteroplasmy, in this healthy cohort. We further characterized the pathogenicity of mitochondrial heteroplasmy with computationally predicted and experimentally reported pathogenic effects. Moreover, we scrutinized the patterns of genomic distribution and site frequency spectrum for mitochondrial heteroplasmy and elucidated the major evolutionary forces underlying these patterns. We demonstrated that pathogenic mitochondrial heteroplasmy is prevalent in healthy individuals, likely due to insufficient purifying selection in removing them. The implication of our results in health management was also discussed.

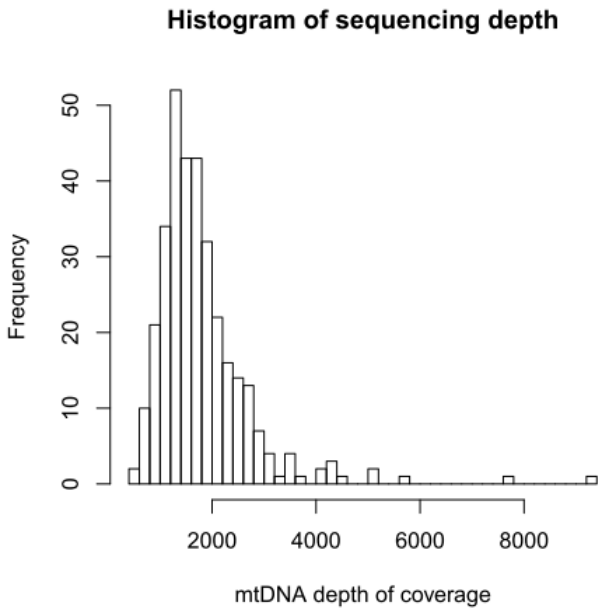


Figure 4.1. The histogram of sequencing depth for mtDNA in 1085 individuals.

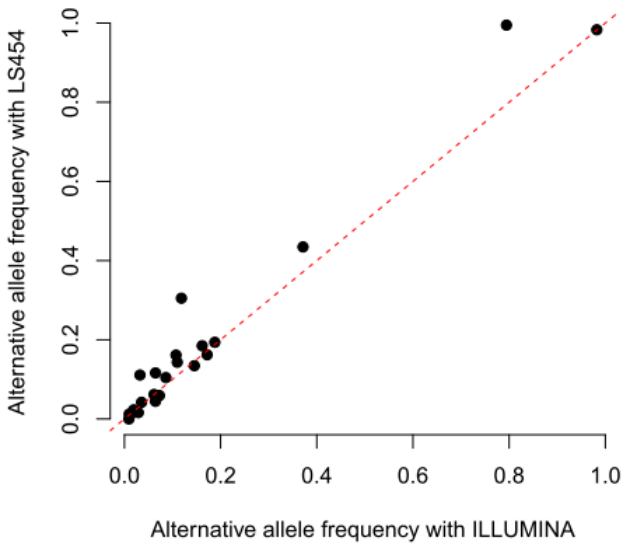


Figure 4.2. Consistency of heteroplasmy identified by ILLUMINA and LS 454. Nine individuals were sequenced by both methods. Alternative allele frequencies for heteroplasmies identified in each method were compared.

4.4. Results

4.4.1. Mitochondrial heteroplasmy is prevalent in the normal human population.

The average depth of coverage for 1085 individuals sequenced by ILLUMINA or SOLID in the 1000 Genomes Project is 1805X (Figure 4.1.), allowing the identification of low frequency heteroplasmies. We applied a combination of stringent thresholds to define heteroplasmy with high confidence and estimated the frequency of heteroplasmy with a maximum likelihood method. In total, we identified 4342 heteroplasmies. There were 9 individuals included in our analysis that were additionally sequenced by LS454, providing an opportunity of verifying the accuracy of our computational pipeline. For the 22 heteroplasmies identified in these 9 individuals with the ILLUMINA data, all of them were observed in the LS454 data with similar frequency (Figure 4.2.), reassuring the reliability of our computational procedure. With the 4342 heteroplasmies identified in 1085 individuals, 973 individuals (89.68%) have at least 1 heteroplasmy. In an extreme case, an individual (HG00740) carried 71 heteroplasmies (Figure 4.3.A). The population prevalence of heteroplasmy depends on the criteria of defining heteroplasmy. The higher the cutoff for MAF, the lower the prevalence. However, even with MAF cutoffs of 5% and 10%, heteroplasmy is observed in 63.50% and 44.42% of the individuals, respectively (Figure 4.4.).

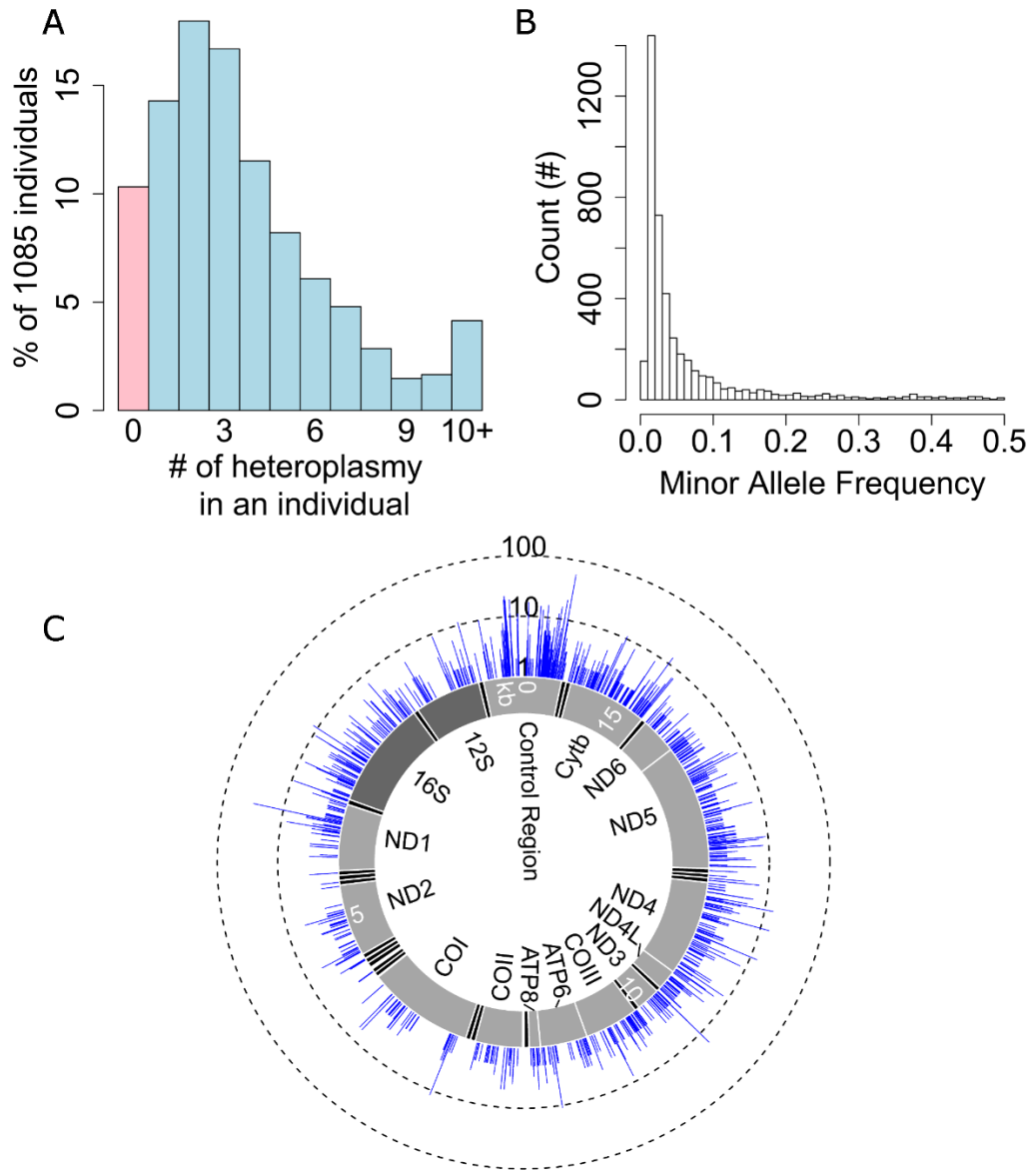


Figure 4.3. Distribution of heteroplasmy in the sample. A. The percentage of individuals carrying a specific number of heteroplasmy. The category of individuals who do not carry any heteroplasmy is highlighted in pink. B. Histogram for minor allele frequency of heteroplasmy. C. The genomic distribution of heteroplasmies and their incidences in the sample. The inner layer represents the mitochondrial genome with tRNA genes highlighted in black, rRNA genes in dark grey and protein-coding genes in light grey. The blue layer indicates the number of individuals (in a total of 1085) carrying heteroplasmy at a specific site. The number of individuals is shown at a common logarithm scale.

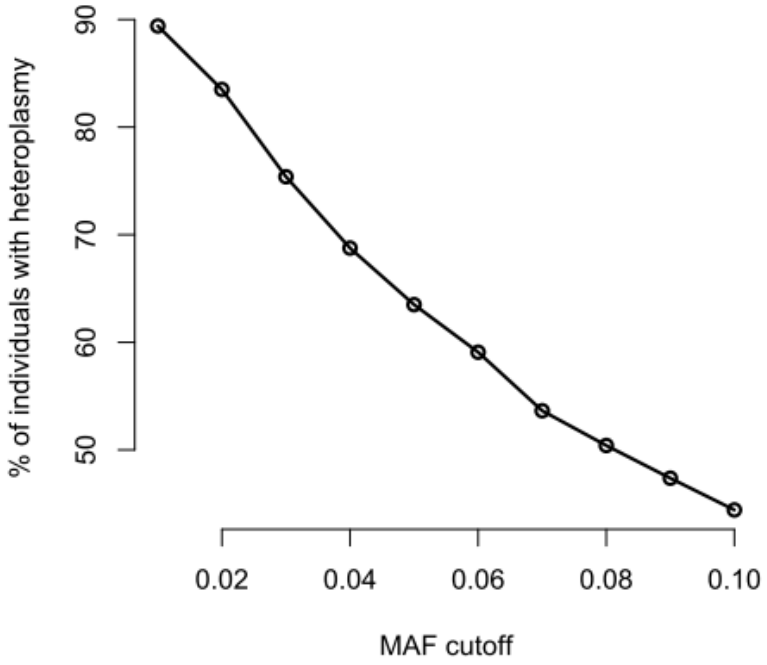
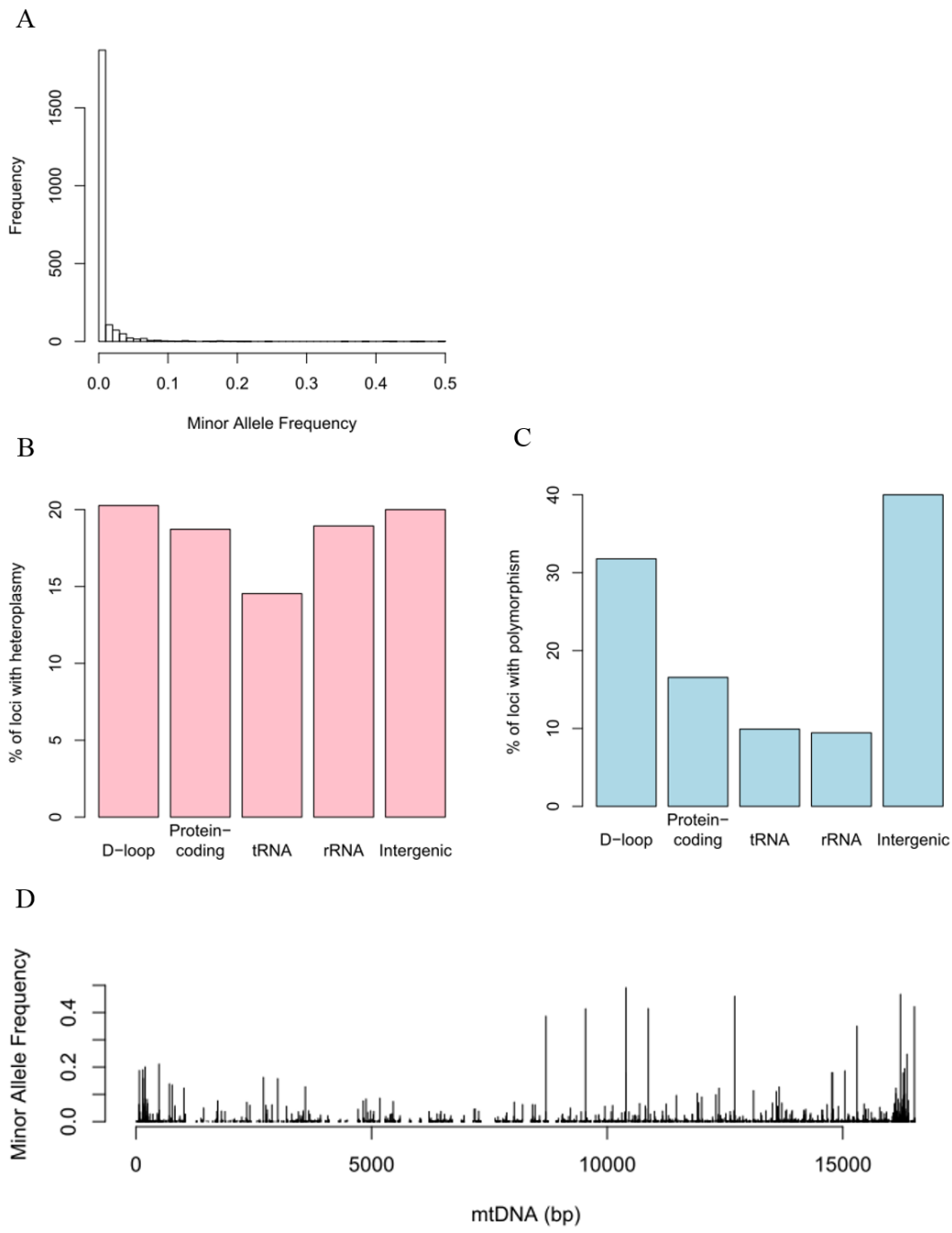


Figure 4.4. The prevalence of heteroplasmy with different MAF cutoff in definition of heteroplasmy.

Figure 4.5. The distribution of heteroplasmy and polymorphisms in mtDNA. A. The histogram for minor allele frequency of polymorphism. B. The prevalence of heteroplasmy in each genomic region; C. The prevalence of polymorphism in each genomic region; D. The genomic distribution of polymorphisms and their minor allele frequency in the sample of 1085 individuals.


A
Frequency
0
500
1000
1500
0.0
0.1
0.2
0.3
0.4
0.5
Minor Allele Frequency
B
% of loci with heteroplasmy
0
5
10
15
20
D−loop
Protein−coding
tRNA
rRNA
Intergenic
C
% of loci with polymorphism
0
10
20
30
40
D−loop
Protein−coding
tRNA
rRNA
Intergenic
D
Minor Allele Frequency
0.0
0.2
0.4
0
5000
10000
15000
mtDNA (bp)

The majority of heteroplasmies are present at low frequency (Figure 4.3.B). The median ML estimated MAF is 2.71%. The skew to low frequency is similar to the site frequency spectrum of population polymorphism but less severe (Figure 4.4. A). These heteroplasmies were observed at 2531 mtDNA sites across different regions in mtDNA, and 1757 (69.42%) of these sites are heteroplasmic in only 1 individual (Figure 4.3. C and 4.4. B). Among all heteroplasmic sites, 36.67% were also observed to be polymorphic in the population (permutation test, p <1.00e-5, Figure 4.4. C and D). There is a positive correlation between the population incidence of heteroplasmy and the population MAF of polymorphic sites (linear regression $R^2 = 0.2358$, $p <$ 2.20e-16). In a previous study, a relative mutation rate for each site in the mitochondrial genome was defined as the absolute frequency of mutation occurrence in a phylogenetic tree constructed with global human samples (26). Using this dataset, we showed that heteroplasmic sites have significantly higher relative mutation rates than homoplasmic sites (Wilcoxon rank-sum test, $p < 2.20e-16$, Figure 4.6.A), and that the incidence of heteroplasmy is positively correlated with relative mutation rate (linear regression $R^2 = 0.3702$, $p < 2.20e-16$, Figure 4.6.B). These observations further confirmed the reliability of our pipeline in identifying heteroplasmies and indicated that high mutation rate might be a major driving force for the population prevalence of heteroplasmy.

4.4.2. Mitochondrial heteroplasmy is over-represented in disease-associated sites.

Of the 4342 detected heteroplasmies, 301 (7.11%) are reported to be disease-associated (7) and 210 individuals (19.35% out of 1085) carried at least 1 disease-

associated heteroplasmy. These observations prompted us to further investigate the disease implication for these heteroplasmies. Among the 13639 mtDNA sites that satisfied quality control criteria and were examined in our study, 399 (2.93%) are disease-associated (7). However, the corresponding number is 147 (5.81%) among the 2531 heteroplasmic sites, which is significantly higher than expected by chance (Chi-squared test, p = 2.52e-12). The percentage of disease-associated sites among population polymorphic sites (6.30%) is also significantly more than random expectation (Chi-squared test, p = 1.44e-14) but is comparable to that of heteroplasmic sites (Figure 4.7. A). For the two disease categories that have the highest number of associated sites, mitochondrial myopathy and mitochondrial encephalomyopathy, heteroplasmy is over-represented even when compared with polymorphism. Among all the sites examined, 64 (0.47%) have been reported to be associated with mitochondrial myopathy, and 52 (0.38%) with mitochondrial encephalomyopathy. Only one site is shared between the two diseases. Heteroplasmic sites are 2.02 times (95% CI: 1.40~2.68) more likely to be associated with mitochondrial myopathy than given by random expectation, and 2.97 (95% CI: 1.60~8.65) times more likely than polymorphic sites. Similarly, for mitochondrial encephalomyopathy, heteroplasmic sites are 1.87 (95% CI: 1.19~2.57) times more likely to be disease-associated than random expectation, and 1.97 (95% CI: 1.04~4.97) times more likely than polymorphic sites (Figure 4.7. A). Additionally, among the heteroplasmic sites we identified, 10 are associated with Leber hereditary optic neuropathy, 4 with deafness or sensorineural hearing loss, 2 with Leigh syndrome, 3 with cardiomyopathy, 3 with diabetes mellitus, and 2 with Alzheimer's or Parkinson's disease.

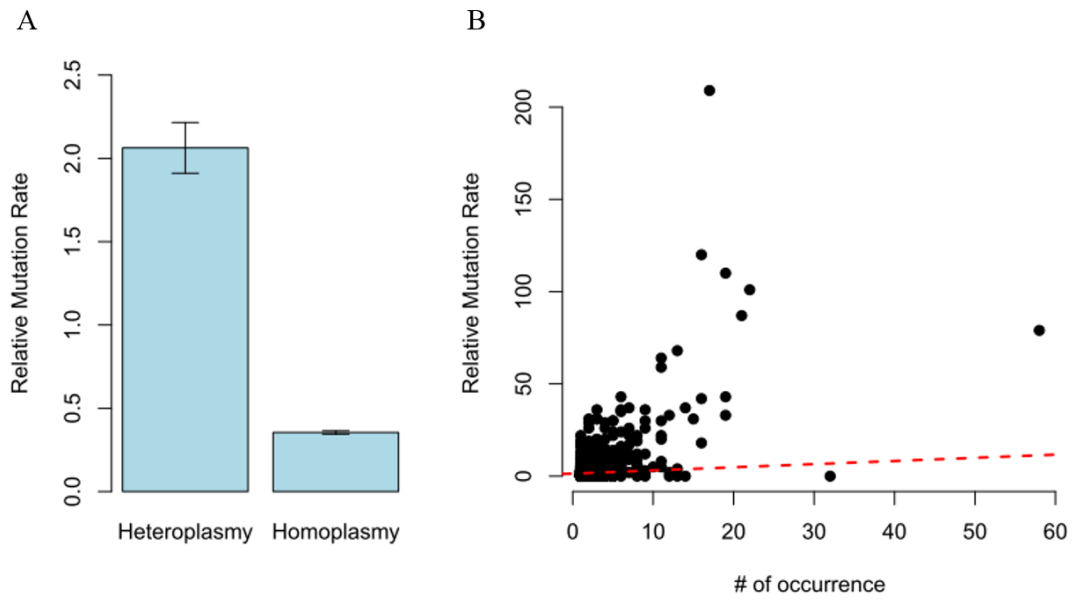
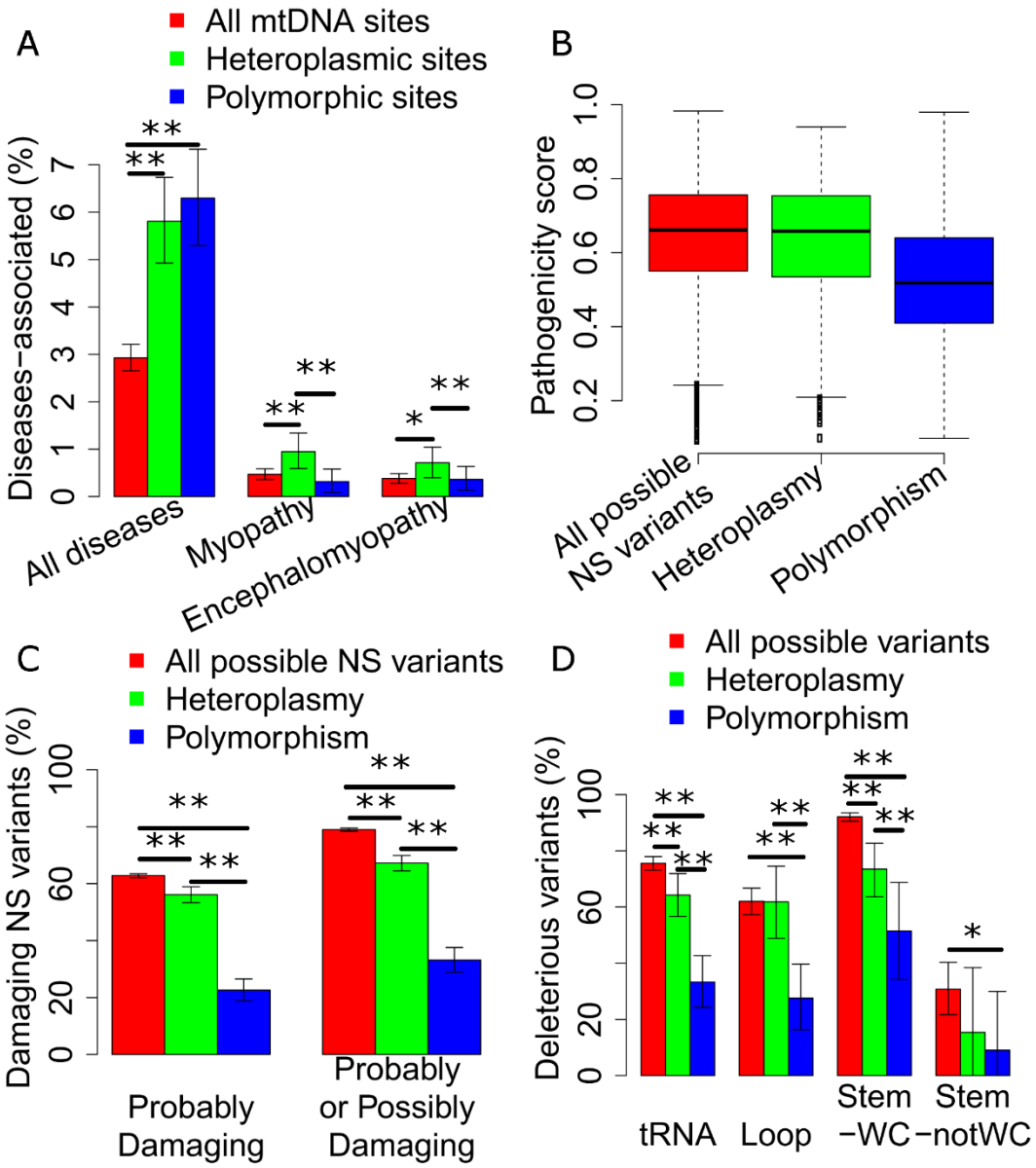


Figure 4.6. Mutation rate in mtDNA and heteroplasmy. A. The barplot of relative mutation rate for heteroplasmic and homoplasmic loci. Error bar represents one standard error. B. The positive correlation between relative mutation rate and the number of occurrence in the population. Each black dot represents a heteroplasmic locus. And the red dashed line indicates the linear regression.

Figure 4.7. Mitochondrial heteroplasmy is highly pathogenic. **A.** The percentage of loci associated with diseases in all, heteroplasmic, and polymorphic sites. "All diseases" represents all diseases included in MITOMAP. Myopathy and encephalomyopathy are the two disease categories that have the highest number of mitochondrial loci reported to be associated with. **B**. The boxplot of MutPred pathogenicity scores for all possible NS variants in the mitochondrial genome, NS heteroplasmies and polymorphisms. Heteroplasmies occurred in multiple individuals were counted only once. **C**. The percentage of PolyPhen-2 predicted damaging variants in all possible NS variants, NS heteroplasmies and polymorphisms. **D**. The percentage of predicted deleterious tRNA variants in all possible variants, heteroplasmy and polymorphism. "tRNA" represents all regions of tRNA genes, including loop and stem regions. "Loop" represents the loop region. "Stem-WC" refers to the Watson-Crick pairing positions in the stem region. "Stem-notWC" refers to those that are not Watson-Crick paired. The error bar represents 95% CI from 10,000 bootstraps. Permutation *p* values were indicated as ** for $p < 0.01$ and * for $p < 0.05$.


A
All mtDNA sites
Heteroplasmic sites
Polymorphic sites
Diseases−associated (%)
0 1 2 3 4 5 6 7
All diseases
Myopathy
Encephalomyopathy
B
Pathogenicity score
0.2 0.4 0.6 0.8 1.0
All possible NS variants
Heteroplasmy
Polymorphism
C
All possible NS variants
Heteroplasmy
Polymorphism
Damaging NS variants (%)
0 20 60 100
Probably Damaging
Probably or Possibly Damaging
D
All possible variants
Heteroplasmy
Polymorphism
Deleterious variants (%)
0 20 60 100
tRNA
Loop
Stem −WC
Stem −notWC

4.4.3. Non-synonymous and tRNA heteroplasmy is highly pathogenic.

To explore the pathogenicity of mitochondrial heteroplasmy on a broader basis, we applied computational methods to predict the deleterious effect of non-synonymous (NS) and tRNA mutations. For NS heteroplasmy, we first defined pathogenicity scores for all possible NS changes in the mtDNA with the MutPred algorithm (27, 28). The pathogenicity score ranges from 0 to 1 with a higher pathogenicity score indicating greater likelihood of being pathogenic. For all possible 24206 NS changes in the mtDNA, the average pathogenicity score is 0.64 (sd = 0.15). For all 1184 NS heteroplasmies in the dataset, the average score is 0.63 (sd = 0.16), similar to random NS mutations. In contrast, the average pathogenicity score for all 467 population polymorphisms is 0.52 (sd = 0.16), significantly lower than that of heteroplasmies and all possible variants (p = 4.24e-36 and 5.96e-55 respectively. Figure 4.7. B). To quantify the pathogenic potential of heteroplasmic variants in comparison with population polymorphic ones, we could choose a cutoff of pathogenicity score and define NS changes with scores higher than this cutoff as pathogenic. To avoid the arbitrary use of cutoffs, we applied a series of cutoffs from 0.6 to 0.8 and in general heteroplasmy is 1.87~4.26 times more likely to be pathogenic than polymorphism (Figure 4.8.). As a verification, the pathogenicity of NS variants was also predicted using PolyPhen-2 (29, 30). PolyPhen-2 yielded comparable predictions with MutPred (Figure 4.9.). The percentage of damaging heteroplasmic variants is significantly lower than random expectation, but significantly higher than that of polymorphic ones (Figure 4.7. C).

We further investigated the pathogenicity of heteroplasmy in tRNA genes. We used the pathogenic prediction for all possible variants in tRNA genes from a previous study which utilized evolutionary information in functional assessment (31). The percentage of pathogenic variants for heteroplasmy is 64.23%, significantly lower than that for all possible variants (75.58%, Chi-squared test, p = 0.0023) but significantly higher than that for polymorphism (33.33%, Chi-squared test, p = 2.63e-06). In other words, tRNA heteroplasmies are 1.93 (95% CI: 1.51~2.63) times more likely to be pathogenic than polymorphisms. Similar trends were observed when we separated tRNAs into three regions: loop, Watson-Crick pairing positions in stem, and non-Watson-Crick pairing positions in stem (Figure 4.7. D).

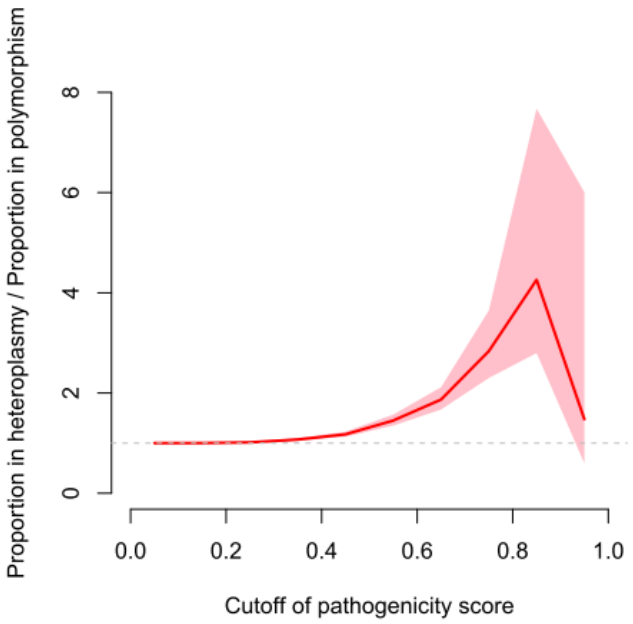


Figure 4.8. The relative risk of heteroplasmy being pathogenic when compared with polymorphism. A pathogenic mutation is defined with varying cutoff of pathogenicity score. The red line is the empirical observation while the pink region represent the 95% bootstrap confidence interval.

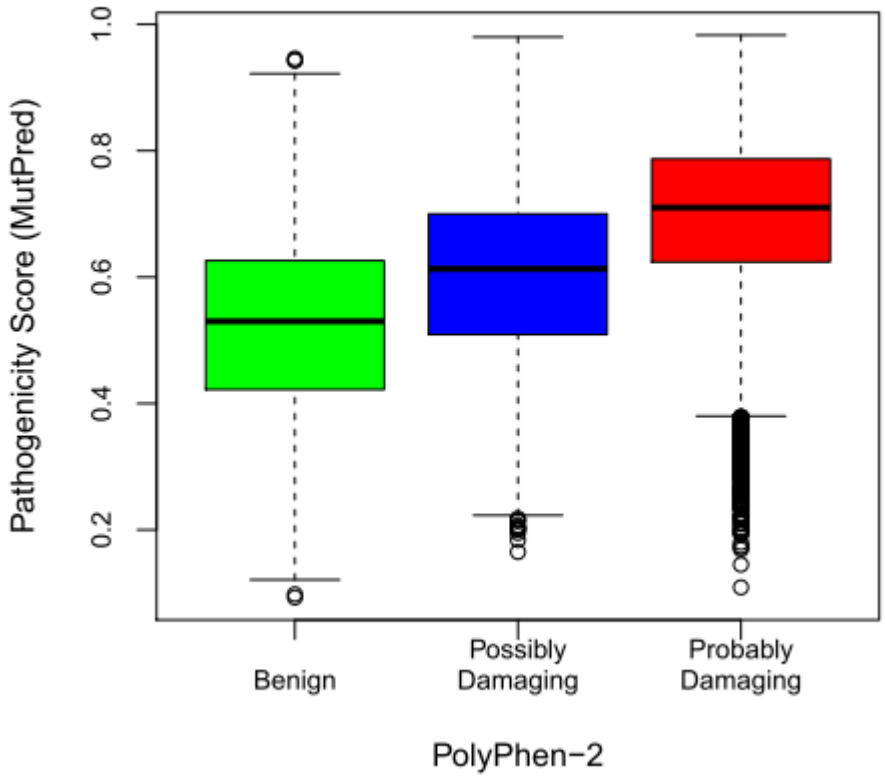


Figure 4.9. Consistent pathogenicity as predicted by MutPred and PolyPhen-2. The MutPred pathogenicity scores for the three functional categories predicted by PolyPhen-2.

4.4.4. Mitochondrial heteroplasmy is subject to purifying selection.

The high pathogenicity of heteroplasmies and their strong association with diseases suggest that heteroplasmy might be under purifying selection. To test this hypothesis, we investigated the genome-wide distribution of heteroplasmies, their unfolded site frequency spectra and the relationship between pathogenicity and derived allele frequency (DAF).

We first examined synonymous and NS variants in protein-coding genes: 5.78% of all possible NS changes in mtDNA were observed with heteroplasmy, which is significantly lower than that of synonymous changes (8.10%, Chi-squared test, p = 2.01e-10, Figure 4.10. A), indicating that NS heteroplasmies are subject to purifying selection. We further examined the site frequency spectrum of heteroplasmy. We found that the distribution of DAF for heteroplasmies in the control region is comparable to that for synonymous heteroplasmies. In comparison to these two types of sites, the distributions of DAF for NS, tRNA, and rRNA heteroplasmies are significantly shifted toward lower frequencies (Wilcoxon rank-sum test, $p < 9.42e-16$, Figure 4.10. B). Furthermore, heteroplasmies within the rRNA stem tend to have a lower DAF frequency than those in the loop and heteroplasmies in the tRNA stem and anticodon loop regions have significantly lower DAF than those in other tRNA regions (Figure 4.11.). Intriguingly, heteroplasmy at disease-associated sites also exhibit significantly lower DAF than that of synonymous heteroplasmy (Wilcoxon rank-sum test, p = 2.07e-10. Figure 4.10. C). Taken together, these results suggest purifying selection is acting on functional heteroplasmies to keep them at low

frequency.

The effect of purifying selection on removing deleterious heteroplasmy suggests a possible reverse correlation between the level of pathogenicity and the frequency of a heteroplasmy. Consistent with this expectation, as depicted in Figure 4.10. D, heteroplasmies with low derived frequency inside an individual tend to have high pathogenicity scores. This negative relationship can be modeled with a logistic function (R^2 = 0.9794, p < 9.76e-06). From the regression, we inferred that the pathogenicity scores are comparable among heteroplasmies with DAF less than 60% and declines as DAF exceeds 60%. This pattern indicates that pathogenic heteroplasmies must reach high frequency before they are selected against, and 60% in general might be a good estimation of the threshold for pathogenic heteroplasmic mutations to express deleterious effect. Consistent with the impact of purifying selection on removing deleterious heteroplasmy, our results also show that while heteroplasmies observed in a few individuals (1~4) have comparable pathogenicity scores (mean=0.64, sd=0.16), those observed in more than 5 individuals have significantly lower pathogenicity scores (mean=0.43, sd=0.25, Wilcoxon rank-sum test, p = 8.74e-04). When we examined the DAF of pathogenic tRNA heteroplasmies, we also found that 81.90% of heteroplasmies with DAF less than 5% are pathogenic, while only 30% of heteroplasmies with DAF larger than 95% are pathogenic (Fisher's exact test, p = 0.0010).

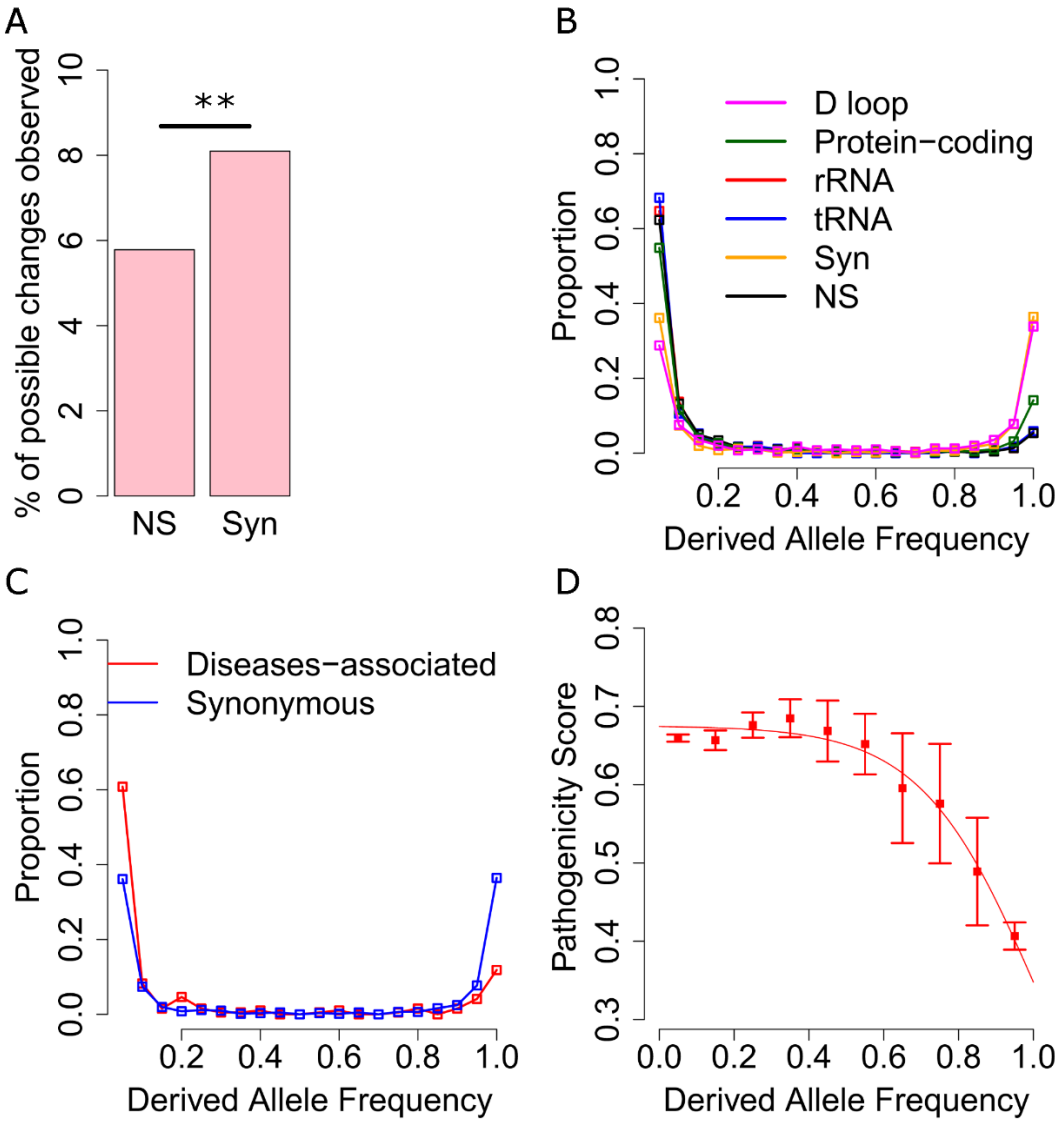


Figure 4.10. Purifying selection on mitochondrial heteroplasmy. **A**. The prevalence of synonymous and NS heteroplasmies, which is defined as the percentage of all possible (synonymous or NS) changes that is observed to be heteroplasmic. ** represents p = 2.01e-10 in Chi-squared test. **B.** The distribution of derived allele frequency (DAF) for heteroplasmies in different mtDNA genomic regions. **C.** The distribution of DAF for disease-associated and synonymous heteroplasmies. **D.** The average pathogenicity score in each bin of DAF. Error bar represents 1 standard error. The red line represents model-fitting with a logistic function of $y = \frac{0.67}{1+e^{\frac{1-x}{-0.16}}}$. R^2 = 0.9794, p < 9.76e-06.

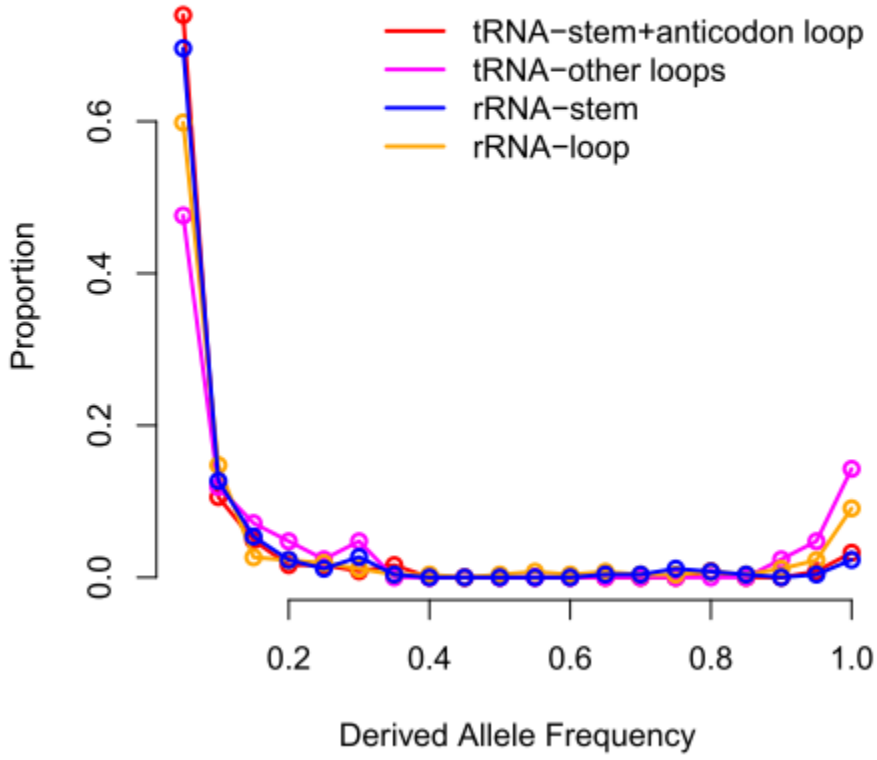


Figure 4.11. The distribution of derived allele frequency for heteroplasmies in different regions of tRNA and rRNA.

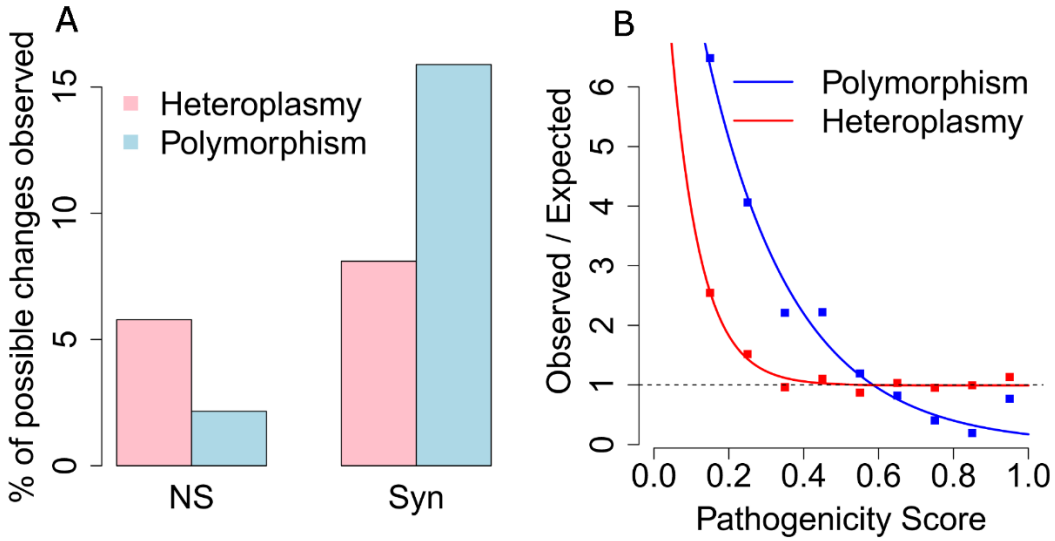


Figure 4.12. Less efficient purifying selection on mitochondrial heteroplasmy than on polymorphism. **A**. The prevalence of synonymous and NS heteroplasmies in comparison with that of synonymous and NS polymorphisms. **B.** The selection function for heteroplasmy (or polymorphism) defined by dividing the observed distribution of pathogenicity scores for heteroplasmy by the expected distribution of pathogenicity scores from all possible NS variants. The dash line represents the expected value, 1, for selection function under neutral evolution. The exponential fit for polymorphism is $y = 12e^{-\frac{x}{0.23}}$. R^2 = 0.9758, p = 4.71e-06. The exponential function for heteroplasmy is $y = 10e^{-\frac{x}{0.079}} + 0.99$. R^2 = 0.9650, p = 4.65e-06.

4.4.5. Purifying selection is less efficient on heteroplasmy than on polymorphism.

Although heteroplasmic sites show evidence of purifying selection, we hypothesized that purifying selection on heteroplasmy is much weaker than that on polymorphism due to the low frequencies of most heteroplasmies inside individual cells. Indeed, consistent with this hypothesis, difference between the percentages of synonymous and NS variants is much bigger for polymorphisms than heteroplasmies (Chi-squared test, $p < 2.2e-16$, Figure 4.12. A). To further quantitatively compare the effect of natural selection on these two types of mtDNA variants, we defined a selection function by dividing the observed distribution of pathogenicity scores for all NS heteroplasmies (or polymorphisms) by the expected distribution of pathogenicity scores from all possible NS variants. This quantitative method has been previously applied to mitochondrial polymorphisms (28). In the absence of natural selection, mutations are similar to random draws from all possible changes in the genome, so the selection function is expected to be equal to a constant, 1. Consistent with previous study (28), the selection function of polymorphism can be modeled by a simple function of exponential decay ($R^2 = 0.9758$, $p = 4.71e-06$, Figure 4.12. B). The parameterizations in our data are similar to the previous study (28). We also confirmed that the observed value for polymorphisms with very high pathogenicity scores (>0.9) deviates from the exponential fit, indicating that forces other than purifying selection might have acted on these variants (28).

The selection function for heteroplasmy also follows an exponential decay ($R^2 = 0.9650$, $p = 4.65e-06$, Figure 4.12. B). Interestingly, in contrast to polymorphism, it

has an additional constant very close to 1, indicating that purifying selection is too weak to effectively remove pathogenic heteroplasmies, likely due to their low frequency inside the cells. Using the selection functions for both polymorphisms and heteroplasmies, the relative effect of selection on two different amino acid variations could be assessed by a ratio of the exponential functions for the two pathogenicity scores (28). For example, a population polymorphic variant with a pathogenicity score of 0.8 is subject to about 2 times stronger purifying selection than a polymorphic variant with a score of 0.6. In comparison, the strength of purifying selection on two heteroplasmies with pathogenicity scores of 0.8 and 0.6 is almost the same. This quantitative comparison further confirms that purifying selection on heteroplasmy is much less efficient than that on population polymorphism in removing deleterious mutations.

Table 4.1. Comparison of criteria for calling heteroplasmy

	He et al. 2010 (21)	Li et al. 2010 (13)	Goto et al. 2011, (16)	Picardi & Pesole, 2012, (32)	This study
Sequencing	ILLUMINA ~16,700X	ILLUMINA 36 & 76 bp ~67 & ~211X	ILLUMINA ~1,170X	ILLUMINA, Agilent, NimbleGen	ILLUMINA, SOLID
Mapping	Eland	MIA	BWA	GSNAP	GSNAP remapping
Mismatches	<= 3 in 36 bp	Default	Default	Default	Default or 7%
Reads	remove low-quality reads	remove duplicate reads & low-quality reads	--	--	Reads mapped to mtDNA in 1000G Project
Mapping	--	--	Unique	Unique	Unique
Base quality	>=23 for all bases in the read	>= 20 on site; >= 15 for 5 bp flanking	>= 30 on site	>=20 on site	>=20 on site
Minimum depth	>= 10 distinct reads	--	>= 100X HQ depth on each strand	>=20 X >=2 reads	>= 10X HQ depth on each strand
Double-strand validation	>=3 reads on each strand	>= 1 read on each strand	>=100 HQ depth on each strand; >= 2% raw frequency[a] on each strand;	--	>= 2 reads on each strand; >=1% raw frequency on each strand
Minor allele frequency[b]	>=1.6%	>=10%	>= 2% on each strand	--	>=1% on each strand
Log-likelihood ratio	--	--	--	>=5	>=5

a. Raw frequency for each locus was calculated as the fraction of the allele among all observed alleles. This is in contrast to frequency estimated with maximum likelihood method which takes into account sequencing error.
b. The minor allele frequency used in all studies are based on raw frequency.

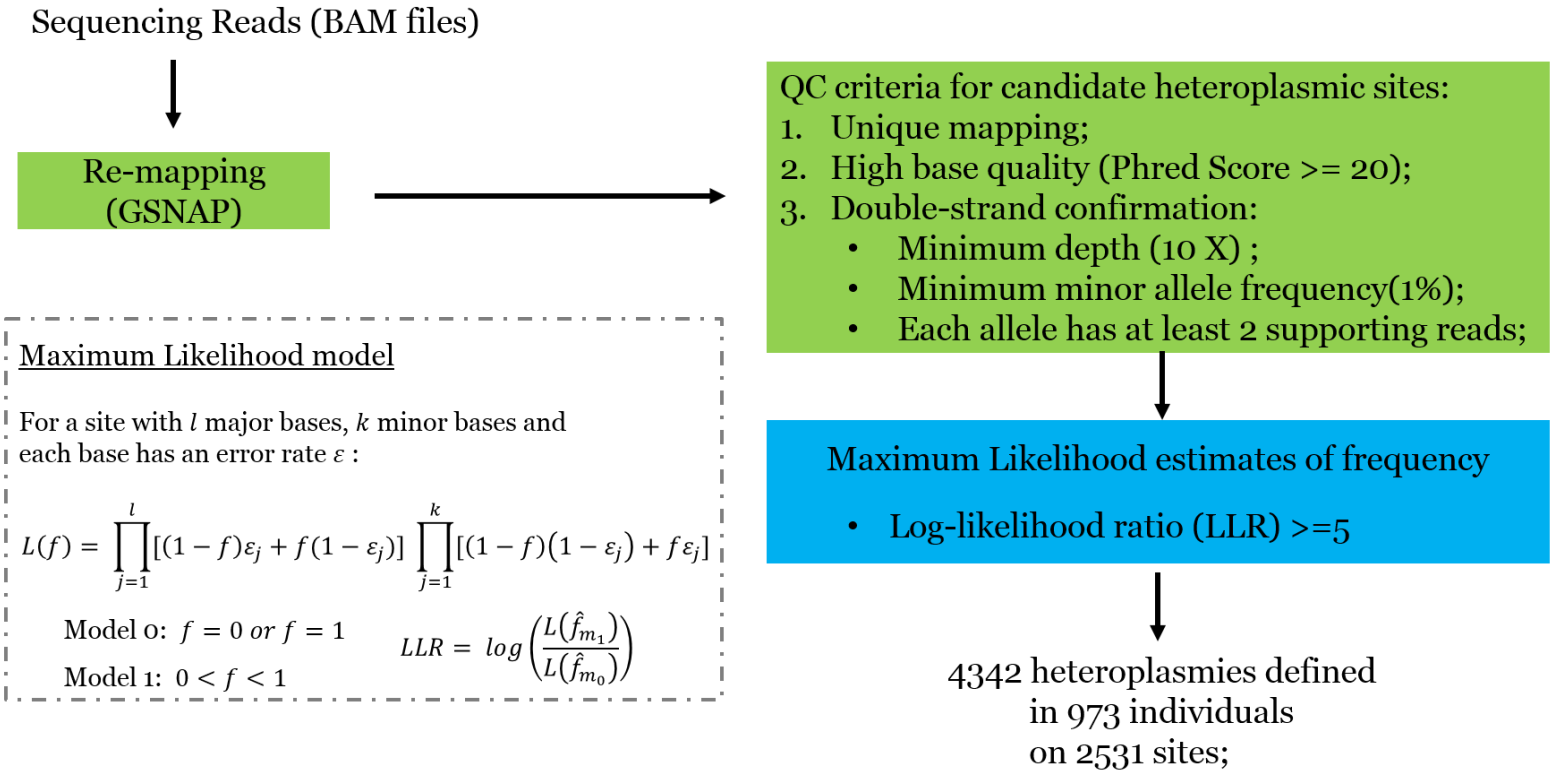


Figure 4.13. Computational pipeline for heteroplasmy identification.

4.5. Discussion

Next-generation sequencing technologies enable the detection of mitochondrial heteroplasmy at the genome-wide level with unprecedented resolution. However, specificity of detection and accuracy of quantification can only be achieved when sequencing errors and technical artifacts are carefully controlled for. A set of criteria for detecting heteroplasmy with modern sequencing technologies have been developed in a few pioneering studies (13, 16, 21, 32). Integrating criteria that have been proven to be effective (Table 4.1, Figure 4.13.), our computational pipeline filtered low-quality bases and unreliable mappings, especially minimizing the complications of nuclear mitochondrial sequences (NumtS) (33). It also used double-stranded validation which required heteroplasmy to be detected in both strands with support from multiple reads. Furthermore, it estimated the frequency of heteroplasmy with a maximum likelihood method by taking into account sequencing error and yields a log likelihood ratio (LLR) indicating the confidence of true positive heteroplasmy. The applications of these tested criteria ensure the correct detection and accurate quantification of heteroplasmy. The reliability of our computational pipeline was confirmed by examining 9 individuals sequenced by both ILLUMINA and LS454 (Figure 4.2.). Moreover, the biologically meaningful patterns of mitochondrial heteroplasmy observed in our study also augment the reliability of our computational pipeline. Additionally, we did not observe consistent and significant population or gender difference in heteroplasmy patterns (Figure 4.14. and 4.15.).

Figure 4.14. Similar heteroplasmy pattern across different human populations. Inter-population comparisons of: **A.** the percentage of individuals carrying at least one heteroplasmy; **B.** the percentage of individuals carrying at least one disease-associated heteroplasmy; **C.** the number of heteroplasmy per individual; **D.** derived allele frequency; **E.** pathogenicity score of non-synonymous heteroplasmies. The error bars in A and B represent 95% CI from 10^5 bootstraps of individuals. For A and B, pairwise comparisons were performed with permutation test. For C. D and E, pairwise comparison were performed with Wilcoxon rank-sum test. Bonferroni corrections were performed with 92 tests including the comparison of male and female. None of the population achieve significance in all comparisons with other populations.

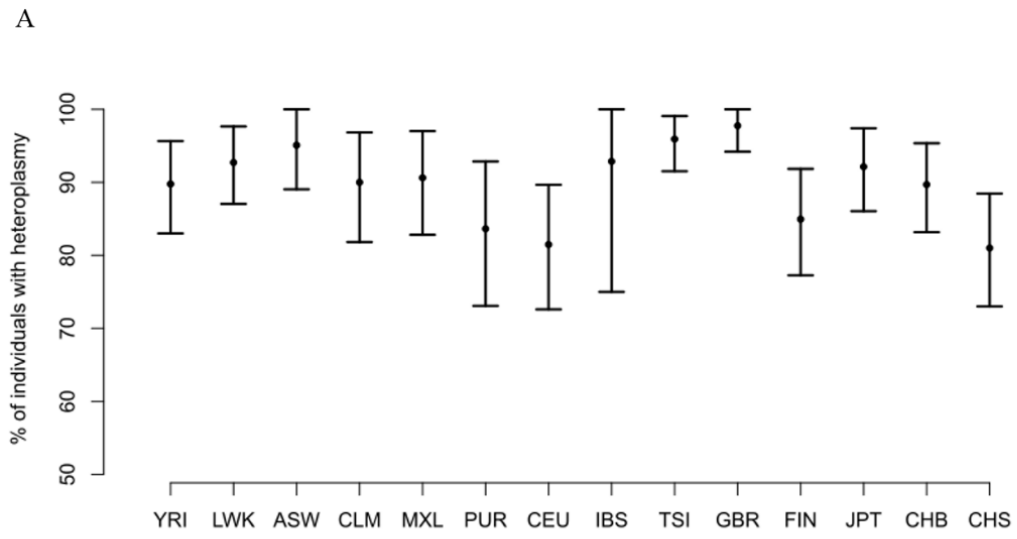
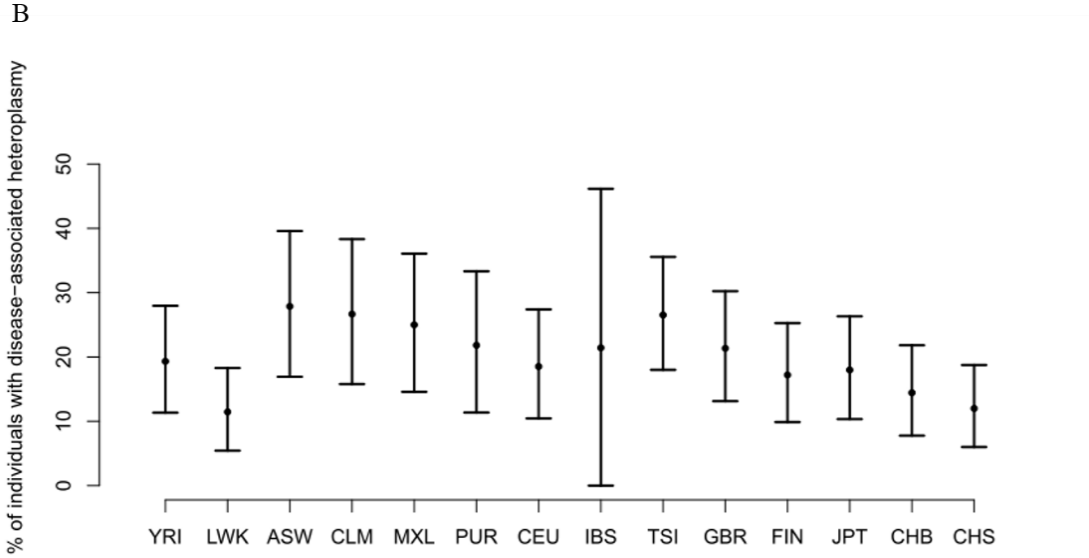


Figure 4.14. (Continued)

B



C

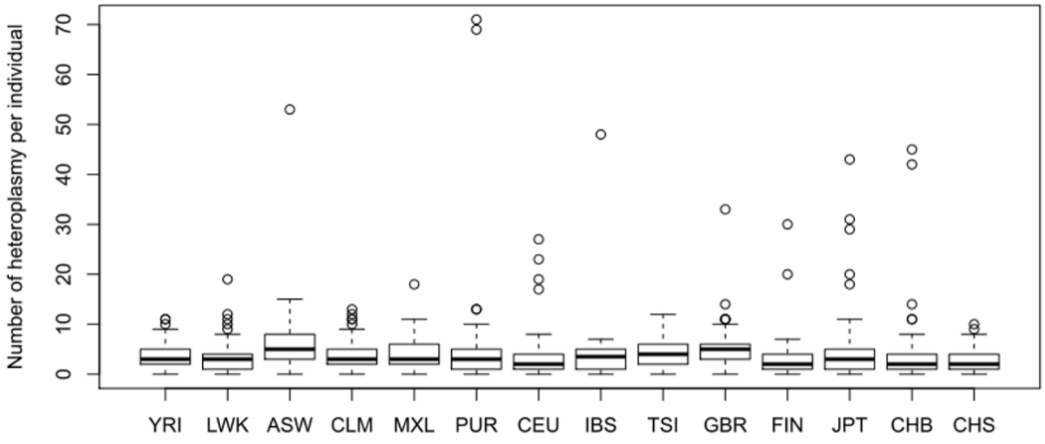
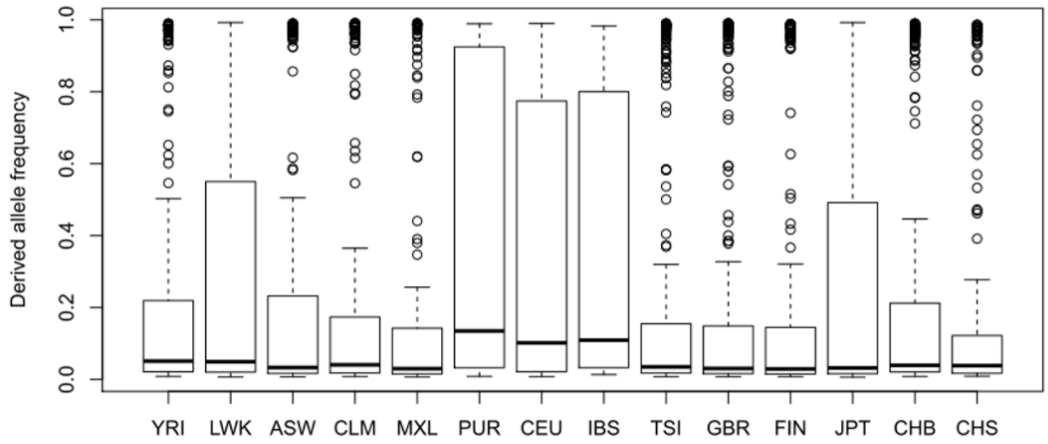
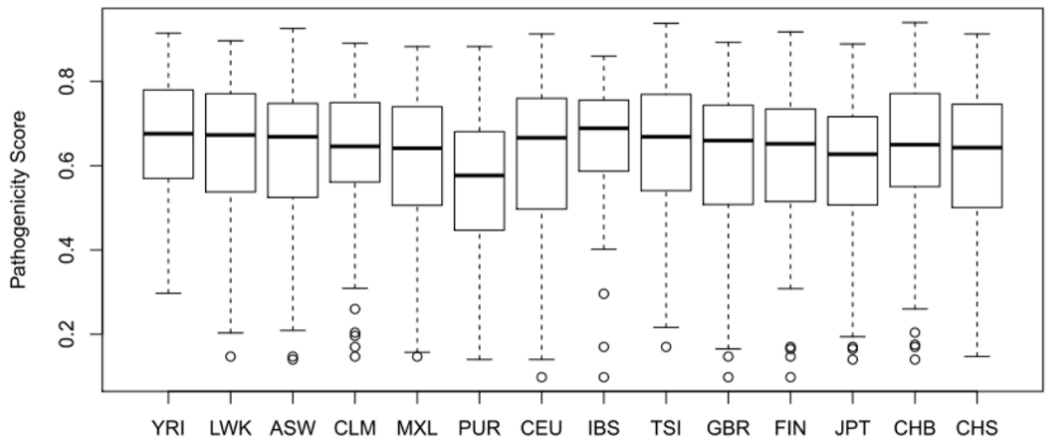


Figure 4.14. (Continued)

D



E



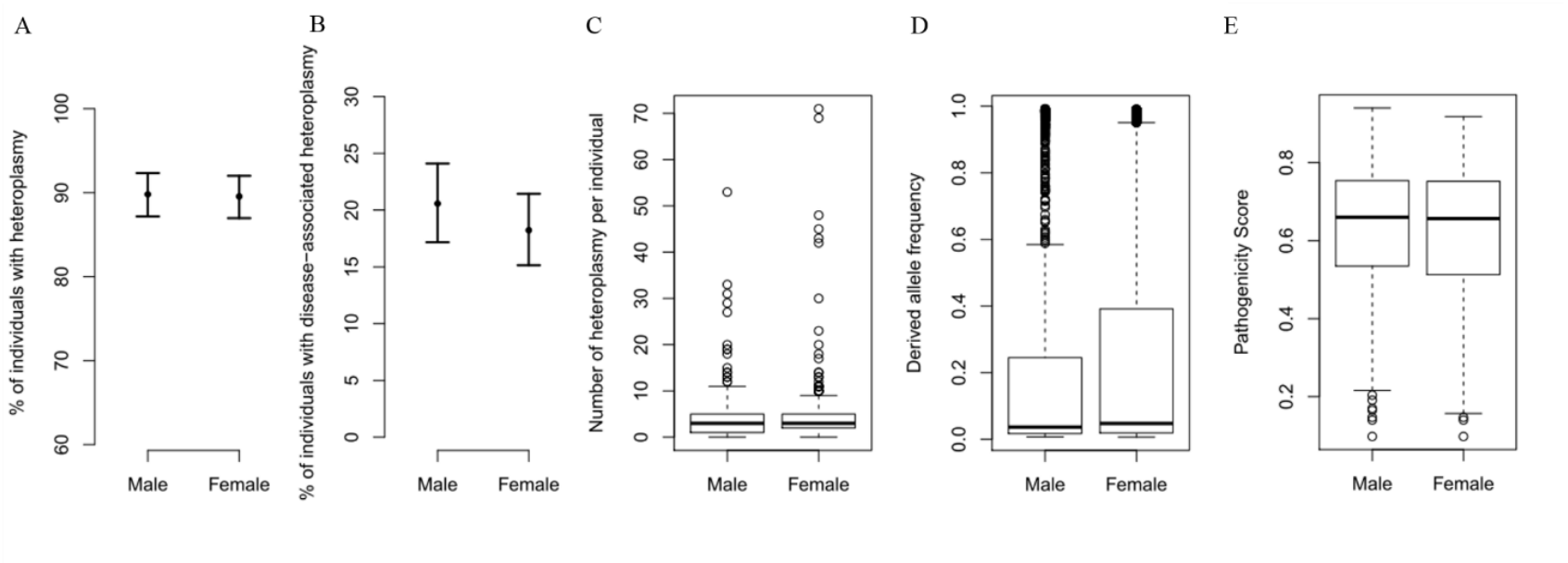


Figure 4.15. Similar heteroplasmy pattern between genders. Inter-gender comparisons of: **A.** the percentage of individuals carrying at least one heteroplasmy; **B.** the percentage of individuals carrying at least one disease-associated heteroplasmy; **C.** the number of heteroplasmy per individual; **D.** derived allele frequency; **E.** pathogenicity score of non-synonymous heteroplasmies. The error bars in A and B represent 95% CI from 10^5 bootstraps of individuals. For A and B, pairwise comparisons were performed with permutation test. For C. D and E, pairwise comparison were performed with Wilcoxon rank-sum test. Bonferroni corrections were performed with 92 tests including the inter-population comparisons. No significance were found after Bonferroni correction.

The prevalence of mitochondrial heteroplasmy at genome-wide scale has been explored in a few studies with smaller sample size and shallower sequencing depth. From the 1000 Genomes Pilot Project, 163 individuals were sequenced to 37.7~3535X coverage and 45% were observed to possess heteroplasmic sites with MAF mostly larger than 10% (12). Another study sequenced 114 individuals with ILLUMINA to a mean coverage of 67 X and 17 individuals to a mean coverage of 211 X. Among these 131 individuals 24.43% were detected to possess heteroplasmy with MAF larger than 10% (13). Moreover, a study used 454 GS FLX system and sequenced 40 Hapmap individuals to a mean coverage of 120 X and 65% individuals were found to have heteroplasmies with MAF higher than 9% (14). With a MAF cutoff of 10%, the prevalence of heteroplasmy is 44.42% in our dataset (Figure 4.4.), which is within the range of previous estimates and very close to the estimate from the 1000 genome pilot project (12). Our study benefits from higher coverage and is able to detect heteroplasmy with MAF as low as 1%. With a much larger sample size, we estimate that the prevalence of heteroplasmy in the healthy population is at least 90%. Since the majority of heteroplasmy is present at very low frequency (Figure 4.3. B), it is very likely that heteroplasmy is universal to all healthy individuals. Results from a recent study conducted on a small sample support this idea (17).

The high pathogenic potential of mitochondrial heteroplasmy is consistently demonstrated with experimentally observed disease-associated mutations, computationally predicted functional effect and the presence of weak negative selection. Firstly, experimentally reported diseases-associated mtDNA mutations are overrepresented in both polymorphic and heteroplasmic sites (Figure 4.7. A). This

pattern has been previously observed in a study with a much smaller sample size (13). It suggests that heteroplasmic and polymorphic variants are either only mildly deleterious or not yet effectively removed by purifying selection. Since polymorphic variants have gone through generations of purifying selection, their overrepresentation in disease-associated sites is likely resulted from their mild deleterious effect. In contrast, as heteroplasmic variants have a much shorter time frame for natural selection, they are likely subject to weaker purifying selection and have higher pathogenic potential. However, the overrepresentation of polymorphism and heteroplasmy in disease-associated sites may also reflect the research bias towards using known polymorphic sites in disease studies. Secondly, we artificially created all possible variants in the mitochondrial genome and computationally predicted their pathogenic effects, which serve as a pathogenicity benchmark before being subject to purifying selection. In comparison to this theoretical expectation, heteroplasmy has slightly lower pathogenicity while polymorphism has much lower effect (Figure 4.7. B, C, and D). This is consistent with the fact that polymorphism has been subject to generations of purifying selection and only variants with mild deleterious effect could survive. It also suggests that while purifying selection also acts on heteroplasmy, its strength may be weak and therefore the pathogenic effect of heteroplasmy is very close to the theoretical expectation without purifying selection. Lastly, we observed convincing signals of purifying selection on heteroplasmy and demonstrated its weaker strength than that on polymorphism, further supporting the high pathogenic potential of heteroplasmy.

The prevalence of pathogenic heteroplasmic mtDNA mutations in the general

population due to inefficient purifying selection has important clinical implication. Although only about 1 in 5000 people suffers from mitochondrial diseases (24), the incidence of pathogenic mtDNA mutations could be much higher because of the mitochondrial threshold effect which masks the deleterious effect of low-frequency pathogenic mutations. A study of ten common pathogenic mtDNA mutations revealed an incidence of at least 1 in 200 subjects (23). For these 10 mutations, the prevalence of heteroplasmy in our samples is 1 in 155 (95% CI: 83-556). When we included all identified disease-associated mtDNA mutations (7), the incidence of pathogenic heteroplasmies is 19.35%, or 1 in 5 individuals (95% CI: 4.62-5.87). Given the likely underestimation of disease-mtDNA mutation association and the observed prevalence of heteroplasmic mtDNA mutations with high predicted pathogenic scores in this study, the real frequency of pathogenic mitochondrial heteroplasmy could be much higher than this estimation.

Multiple underlying mechanisms have been proposed to modulate the expansion of deleterious mtDNA mutations at the cellular level. According to computational modeling of the relaxed replication of mtDNA in both dividing and non-dividing cells, even with random genetic drift alone the typical lifespan of an individual is more than enough for low-frequency heteroplasmy to reach high frequency or even homoplasmy in a small population of cells (34-36). On average it only takes approximately 70 generations of cell divisions to reach homoplasmy from a new mutation. That is only about 25 years for epithelial cells, which experience three cell turnovers per year (34). In post-mitotic tissues, such as skeletal muscle and neurons, the mean time to homoplasmy is about 40 years (35, 36). Besides random genetic drift during

intracellular mitochondrial turnover and cell divisions (34, 35, 37), natural selection with replicative or survival advantage has also been proposed to either accelerate or decelerate the spread of pathogenic mutations (38-40). Extensive experimental observations have recorded abundant clonally expanded mtDNA mutations in human tissues, especially in aged individuals (40, 41). More importantly, both computational modeling and experimental evidence support that mutation accumulated with age results mostly from the clonal expansion of mutations that existed early in life, rather than *de novo* mutations later in life (35, 37, 41, 42). All individuals included in the 1000 Genomes Project were healthy at the time of sample collection (25). The prevalence of pathogenic mitochondrial heteroplasmy in healthy individuals observed in this study raises the concern that they could expand to high frequency in a fraction of cells later in life, exceed the critical phenotypic threshold and lead to age-related diseases. Future studies are needed to unravel the mechanisms of clonal expansion of pathogenic heteroplasmy, to elucidate the roles of mitochondrial heteroplasmy in complex disorders, and to develop effective strategies in managing these mutations in order to prevent the progression into disease.

4.6. Materials and Methods

Reference genomes and annotations. The reference sequence for the human nuclear genome was GRCh37/hg19, as downloaded from the 1000 Genomes Project data server (http://www.1000genomes.org/). The revised Cambridge Reference Sequence (rCRS) and gene annotations for the human mitochondrial genome were downloaded from NCBI with accession number NC_012920. So were the reference mitochondrial genomes and annotations for *Pan troglodytes*, and *Pongo abelii*. Annotation of synonymous and non-synonymous changes for rCRS, and the secondary structure of tRNA and rRNA was retrieved from a previous study (45). The secondary structure of tRNA and rRNA were computed with the mfold program (46). Relative Mutation Rate (RMR) for each site was inferred as the absolute frequency of occurrence of the mutation in a phylogenetic tree constructed with 2196 global human samples (26).

Sequencing data. Sequencing reads mapped to the mitochondrial genome in the 1000 Genomes Project phase 1 data were downloaded from the 1000 genomes data server. Our analysis focused on 1085 unrelated individuals from 14 populations, which were sequenced using either ILLUMINA or SOLID platforms. There were 9 individuals sequenced by two methods (ILLUMINA and LS454). These individuals were used to confirm the reliability of our computational pipeline with ILLUMINA data. See Table 4.2. for more detailed information.

Table 4.2. Sequencing data from 1000 Genome Project

Population	# by ILLUMINA	# by SOLID	Total
ASW	50	11	61
CEU	81	0	81
CHB	81	16	97
CHS	92	8	100
CLM	50	10	60
FIN	75	18	93
GBR	70	19	89
IBS	6	8	14
JPT	78	11	89
LWK	82	14	96
MXL	52	12	64
PUR	52	3	55
TSI	98	0	98
YRI	76	12	88

ASW: Americans of African Ancestry in SW USA; CEU: Utah Residents (CEPH) with Northern and Western European ancestry; CHB: Han Chinese in Bejing, China; CHS: Southern Han Chinese; CLM: Colombians from Medellin, Colombia; FIN: Finnish in Finland; GBR: British in England and Scotland; IBS: Iberian population in Spain; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MXL: Mexican Ancestry from Los Angeles USA; PUR: Puerto Ricans from Puerto Rico; TSI: Toscani in Italia; YRI: Yoruba in Ibadan, Nigera.

Definition of ancestral alleles. A previously described method was used to define ancestral human mtDNA alleles with high confidence (47). First, LASTZ (48) was used to align the mitochondrial genomes of *Homo sapiens*, *Pan troglodytes*, and *Pongo abelii*. Furthermore, to take advantage of the better conservativeness of protein sequences than DNA sequences, we aligned the coding region based on MUSCLE alignments of protein sequences (49). Only alleles that were consistent in both *Pan troglodytes* and *Pongo abelii*, and also present in *Homo sapiens* were considered as the ancestral alleles.

Computational pipeline for calling heteroplasmy and polymorphism. Sequencing reads retrieved from the 1000 genome data sever were re-mapped to the combined human genome, both nuclear and mitochondrial genomes, using GSNAP (43). Following previous practice (32), we counted unknown characters (N) as mismatches (--query-unk-mismatch=1) and only retained sequences that mapping uniquely to the genome (-n 1 -Q). Another important parameter for mapping is the maximum number of mismatches allowed (-m). By default, the parameter is ((readlength+2)/15 - 2), corresponding to 5 mismatches for read length of 100bp. In our analysis, using the default parameters resulted in unsatisfactory coverage, especially for non-European individuals. This is due to the fact that mitochondrial DNA is much more divergent than nuclear DNA (45), and the reference mitochondrial DNA is from an individual of European origin (50). To accommodate this fact, we adjusted the parameter to allow 7% mismatches (-m 0.07), corresponding to 7 mismatches for a read length of 100 bp. To confirm that our observed patterns are not artifacts of mis-mapping, we applied both the default and the adjusted parameters. Both parameters yielded similar patterns

of heteroplasmy. Only results using a 7% mismatch threshold are presented.

After the GSNAP reads mapping, we recorded only reads that are uniquely mapped to the mitochondrial genome in order to minimize the complications of nuclear mitochondrial sequences (NumtS) (33). We further filtered the data and defined "usable sites" based on the following three quality control criteria: 1) only bases with Phred quality score >= 20 were used; 2) only sites with 10X coverage of qualified bases on both positive and negative strands were used; 3) only sites that satisfy criteria 1) and 2) in more than 95% individuals were used in analysis of heteroplasmy and polymorphism. A candidate heteroplasmic site was defined with the following two criteria: 1) the raw frequency for the minor allele is no less than 1% on both strands; 2) all alleles have support from at least 2 reads on each strand.

For each candidate heteroplasmic site, we further applied a maximum likelihood (ML) method to accurately estimate the frequency of the major allele while taking into account sequencing error (32, 44). For example, for all bases mapped to the positive strand of a locus, l bases are the major alleles and k bases are the minor alleles. Each base has respective sequencing quality, corresponding to the probability of sequencing error ε. The underlying parameter of interest is the frequency of the major allele f. The likelihood function could be written as follows:

$$L(f) = \prod_{j=1}^{l}\left[(1-f)\varepsilon_j + f\left(1-\varepsilon_j\right)\right]\prod_{j=1}^{k}\left[(1-f)(1-\varepsilon_j) + f\varepsilon_j\right]$$

We estimated f under two models: heteroplasmy (m_1) and homoplasmy (m_0). And a

log-likelihood ratio (LLR) was calculated as $log\left(\frac{L(\hat{f}_{m_1})}{L(\hat{f}_{m_0})}\right)$. A high-confidence heteroplasmy was defined as candidate heteroplasmy with LLR no less than 5 (32). With all these criteria (See Table 4.1. and Figure 4.13.), a total of 4342 heteroplasmies were defined. Among them, 153 have a minor allele frequency estimated by the ML method to be smaller than 1%, even though we required that the raw frequency for the minor allele is no less than 1% on both strands.

After detecting heteroplasmy, consensus sequences were assembled for each individual and compared among all individuals to identify polymorphic sites. Only "usable sites" satisfying the above-mentioned criteria were considered. For each individual, a consensus sequence was assembled using the alleles present at homoplasmic sites, and the major alleles at heteroplasmic sites. Sites were classified as polymorphic if there was more than one allele present in the consensus sequences of the population.

To confirm the reliability of our computational pipeline in defining heteroplasmy, we took advantage of the 9 individuals sequenced by both ILLUMINA and LS454. LS454 data were directly retrieved from the 1000 genome data sever and processed as followed: 1) Only loci defined as heteroplasmy in ILLUMINA data were examined; 2) Only reads with mapping quality no less than 20 and bases with sequencing quality no less than 20 were used; 3) Assuming a biallelic state, only the two most common alleles were retained; 4) The frequency of the heteroplasmic alleles were estimated with the ML method described above. Only heteroplasmy with the same alleles as identified by ILLUMINA was considered as confirmed.

The measure of pathogenicity. The pathogenicity scores for all possible non-synonymous changes were retrieved from a previous study (28). All possible non-synonymous changes were inferred based on the rCRS sequence and the pathogenicity of a non-synonymous change was predicted with the MutPred algorithm (27). A higher pathogenicity score indicates a higher likelihood that the non-synonymous change is pathogenic. Three types of attributes were utilized by MutPred in classifying amino acid variations: 1) attributes based on predicted protein structure and dynamics including secondary structure, solvent accessibility, transmembrane helices, coiled-coil structure, stability, B-factor, and intrinsic disorder; 2) attributes based on predicted functional properties such as DNA-binding residuals, catalytic residues, calmodulin-binding targets, and sites of phosphorylation, methylation, ubiquitination and glycosylation; 3) attributes based on amino acid sequence and evolutionary information, including sequence conservativeness, SIFT score, Pfam profile score, and transition frequencies. The software is trained with a random forest classification model to discriminate between disease-associated amino acid substitution from the Human Gene Mutation Database and putatively neutral polymorphisms from Swiss-Prot (27, 28).

The pathogenic effect of all possible non-synonymous changes were also predicted by PolyPhen-2 (29, 30). PolyPhen-2 combines sequence- and structure-based attributes and predicts the effect of missense mutation with a naive Bayesian classifier. The default HumDiv-trained predictor was used in this study. The pathogenicity predicted by MutPred and Polyphen is highly consistent (Fig. S8).

The pathogenic effect of tRNA mutations were downloaded from a previous publication (31). A tRNA mutation was deemed deleterious by a computational method taking into account the following attributes: 1) evolutionary conservation; 2) disruption of Watson-Crick pairing; 3) the tendency of co-evolution by complementary mutation in the stem.

Disease association information was obtained from MITOMAP (7).

4.7. Acknowledgements

We thank Mr. Paul Billing-Ross, Drs. Andrew Clark, Jason Locasale, Patrick Stover and Lin Xu for their discussions and comments on the manuscript. This work was supported by various funds from Cornell University, ILSI (International Life Sciences Institute) future leader award, NSF grant MCB-1243588 and NIH 1R01AI085286 to Dr. Zhenglong Gu.. Kaixiong Ye is a CVG (Center for Vertebrate Genomics) Scholar at Cornell University.

4.8. References

1. Chinnery PF, Hudson G (2013) Mitochondrial genetics. *Br Med Bull* 106: 135-159.

2. Taylor RW, Turnbull DM (2005) Mitochondrial DNA mutations in human disease. *Nat Rev Genet* 6(5): 389-402.

3. Wallace DC (2010) Mitochondrial DNA mutations in disease and aging. *Environ Mol Mutagen* 51(5): 440-450.

4. Schon EA, DiMauro S, Hirano M (2012) Human mitochondrial DNA: roles of inherited and somatic mutations. *Nat Rev Genet* 13(12): 878-890.

5. Sharpley MS, *et al.* (2012) Heteroplasmy of mouse mtDNA is genetically unstable and results in altered behavior and cognition. *Cell* 151(2): 333-343.

6. Keogh M, Chinnery PF (2013) Hereditary mtDNA heteroplasmy: a baseline for aging? *Cell Metab* 18(4): 463-464.

7. Ruiz-Pesini E, *et al.* (2007) An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res* 35(Database issue): D823-D828.

8. Rossignol R, *et al.* (2003) Mitochondrial threshold effects. *Biochem J* 370(Pt 3): 751-762.

9. Calloway CD, Reynolds RL, Herrin GJ, Anderson WW (2000) The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and increases with age. *Am J Hum Genet* 66(4): 1384-1397.

10. de Camargo MA, *et al.* (2011) No relationship found between point heteroplasmy in mitochondrial DNA control region and age range, sex and haplogroup in human hairs. *Mol Biol Rep* 38(2): 1219-1223.

11. Irwin JA, *et al.* (2009) Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. *J Mol Evol* 68(5): 516-527.

12. Abecasis GR, *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319): 1061-1073.

13. Li M, *et al.* (2010) Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *Am J Hum Genet* 87(2): 237-249.

14. Sosa MX, *et al.* (2012) Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. *PLoS Comput Biol* 8(10):

e1002737.

15. Li M, Stoneking M (2012) A new approach for detecting low-level mutations in next-generation sequence data. *Genome Biol* 13(5): R34.

16. Goto H, *et al.* (2011) Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome Biol* 12(6): R59.

17. Payne BA, *et al.* (2013) Universal heteroplasmy of human mitochondrial DNA. *Hum Mol Genet* 22(2): 384-390.

18. Sondheimer N, *et al.* (2011) Neutral mitochondrial heteroplasmy and the influence of aging. *Hum Mol Genet* 20(8): 1653-1659.

19. Ross JM, *et al.* (2013) Germline mitochondrial DNA mutations aggravate ageing and can impair brain development. *Nature* 501(7467): 412-415.

20. Kennedy SR, Salk JJ, Schmitt MW, Loeb LA (2013) Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. *PLoS Genet* 9(9): e1003794.

21. He Y, *et al.* (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* 464(7288): 610-614.

22. Larman TC, *et al.* (2012) Spectrum of somatic mitochondrial mutations in five cancers. *Proc Natl Acad Sci U S A* 109(35): 14087-14091.

23. Elliott HR, *et al.* (2008) Pathogenic mitochondrial DNA mutations are common in the general population. *Am J Hum Genet* 83(2): 254-260.

24. Schaefer AM, *et al.* (2008) Prevalence of mitochondrial DNA disease in adults. *Ann Neurol* 63(1): 35-39.

25. Abecasis GR, *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422): 56-65.

26. Soares P, *et al.* (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet* 84(6): 740-759.

27. Li B, *et al.* (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 25(21): 2744-2750.

28. Pereira L, *et al.* (2011) Comparing phylogeny and the predicted pathogenicity of protein variations reveals equal purifying selection across the global human mtDNA diversity. *Am J Hum Genet* 88(4): 433-439.

29. Adzhubei IA, *et al.* (2010) A method and server for predicting damaging

missense mutations. *Nat Methods* 7(4): 248-249.

30. Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* Chapter 7: t7-t20.

31. Kondrashov FA (2005) Prediction of pathogenic mutations in mitochondrially encoded human tRNAs. *Hum Mol Genet* 14(16): 2415-2419.

32. Picardi E, Pesole G (2012) Mitochondrial genomes gleaned from human whole-exome sequencing. *Nat Methods* 9(6): 523-524.

33. Simone D, *et al.* (2011) The reference human nuclear mitochondrial sequences compilation validated and implemented on the UCSC genome browser. *BMC Genomics* 12: 517.

34. Coller HA, *et al.* (2001) High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. *Nat Genet* 28(2): 147-150.

35. Elson JL, Samuels DC, Turnbull DM, Chinnery PF (2001) Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age. *Am J Hum Genet* 68(3): 802-806.

36. Chinnery PF, Samuels DC (1999) Relaxed replication of mtDNA: A model with implications for the expression of disease. *Am J Hum Genet* 64(4): 1158-1165.

37. Payne BA, *et al.* (2011) Mitochondrial aging is accelerated by anti-retroviral therapy through the clonal expansion of mtDNA mutations. *Nat Genet* 43(8): 806-810.

38. Diaz F, *et al.* (2002) Human mitochondrial DNA with large deletions repopulates organelles faster than full-length genomes under relaxed copy number control. *Nucleic Acids Res* 30(21): 4626-4633.

39. Fukui H, Moraes CT (2009) Mechanisms of formation and accumulation of mitochondrial DNA deletions in aging neurons. *Hum Mol Genet* 18(6): 1028-1036.

40. Nekhaeva E, *et al.* (2002) Clonally expanded mtDNA point mutations are abundant in individual cells of human tissues. *Proc Natl Acad Sci U S A* 99(8): 5521-5526.

41. Kraytsberg Y, *et al.* (2006) Mitochondrial DNA deletions are abundant and cause functional impairment in aged human substantia nigra neurons. *Nat Genet* 38(5): 518-520.

42. Khrapko K (2011) The timing of mitochondrial DNA mutations in aging. *Nat*

Genet 43(8): 726-727.

43. Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26(7): 873-881.

44. Chepelev I (2012) Detection of RNA editing events in human cells using high-throughput sequencing. *Methods Mol Biol* 815: 91-102.

45. Pereira L, *et al.* (2009) The diversity present in 5140 human mitochondrial genomes. *Am J Hum Genet* 84(5): 628-640.

46. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31(13): 3406-3415.

47. Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* 39(10): 1251-1255.

48. Harris RS (2007) Improved pairwise alignment of genomic DNA. (The Pennsylvania State University).

49. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5): 1792-1797.

50 Andrews RM, *et al.* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 23(2): 147.

Chapter 5 – Mitochondrial DNA heteroplasmy could be reliably detected with massively parallel sequencing technologies[5]

5.1. Introduction

In their comment on our publication (1), Just *et al.* pointed out that sample contamination is likely to explain the large number of heteroplasmies identified in some individuals from the 1000 Genomes Project. They suggested that this technical artifact may have invalidated our original conclusions and further questioned the reliability of using massively parallel sequencing (MPS) technologies in detecting low-frequency mitochondrial DNA (mtDNA) heteroplasmies. Recognizing the potential complication from sample contamination in the 1000 Genomes Project, which to our knowledge was never discussed in literature, we systematically evaluated its presence and impact on our previous analysis. We found that sample contamination only affects a small fraction of individuals and does not change our original conclusions.

5.2. Results

5.2.1. Observed heteroplasmy numbers rule out the prevalence of contamination

Firstly, our observed number of heteroplasmy per individual does not support the prevalence of significant contamination. If contamination is common, the number of heteroplasmy per individual is expected to approximate the number of mtDNA difference between two randomly chosen individuals. However, the average number of mtDNA differences from two randomly chosen individuals in the 1000 Genomes

[5] Published on *Proceedings of the National Academy of Sciences*. See Appendix A for inclusion authorization.

Project is 37, with a range of 20~51 for within population comparisons (Figure 5.1. A). These are much higher than per individual number of heteroplasmies. Also, we did not observe elevated per individual number of heteroplasmy in the African populations that have the highest inter-individual mtDNA differences.

5.2.2. Haplogroup analysis suggests limited contamination

Secondly, only a small fraction of individuals have suggestive evidence of contamination. For each individual, we constructed two consensus sequences, which cover the major and minor alleles at heteroplasmic sites respectively, and we defined haplogroup for either sequence based on PhyloTree (2). Although the presence of two haplogroups is not necessarily a result of sample mixture (*e.g.* one or more mutations create new haplogroup or erase the defining alleles from the original haplogroup), to be overly conservative, we considered all individuals with a secondary haplogroup to be possibly contaminated. Overall, only a small fraction of individuals (63 out of 1,085, or 5.8%) are impacted.

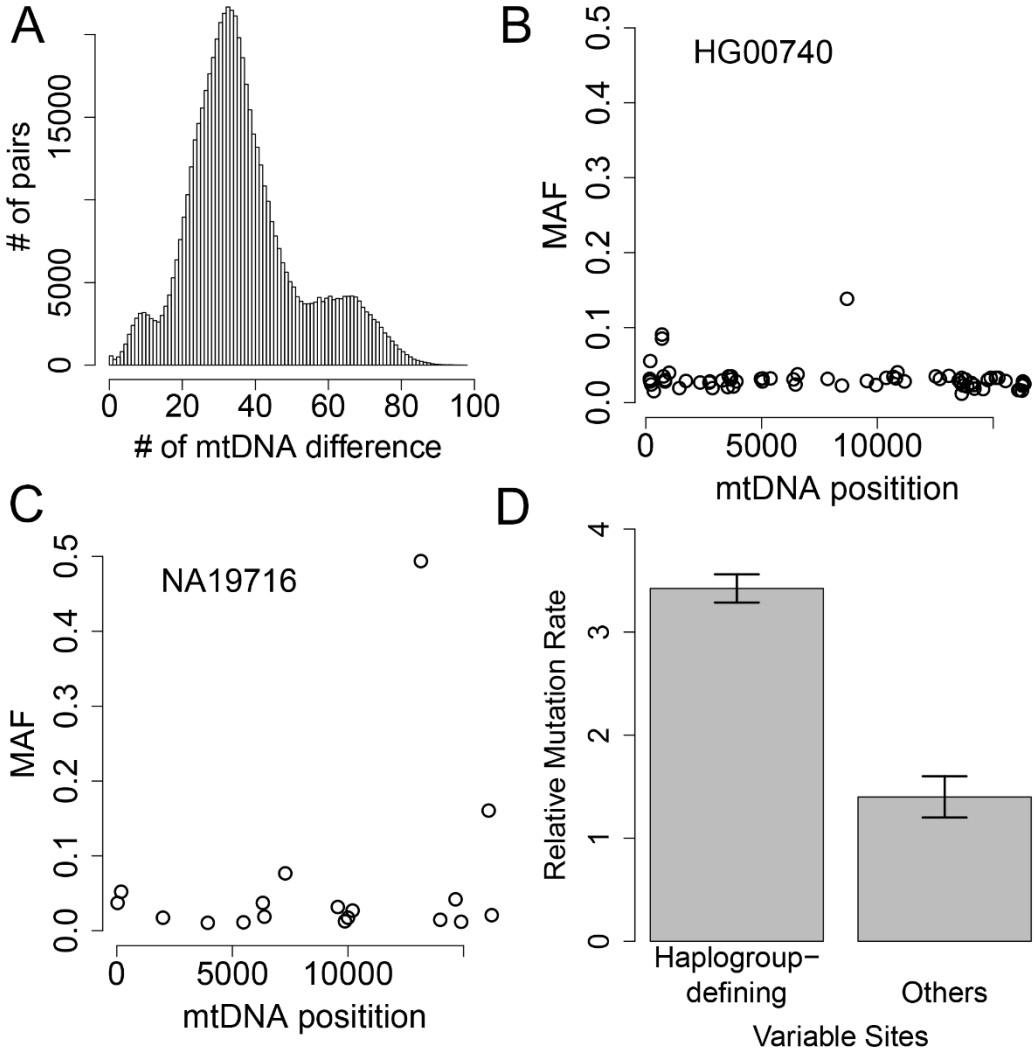


Figure 5.1. The presence of sample contamination in the 1000 Genome Project is limited. **A)** The histogram for the pairwise mtDNA difference among all 1,085 individuals. **B)** The minor allele frequency (MAF) for heteroplasmic sites identified in individual HG00740. **C)** The minor allele frequency for heteroplasmic sites identified in individual NA19716. **D)** The mean relative mutation rate for the haplogroup-defining and other variable sites. Error bar represents 1 standard error.

5.2.3. Contamination of individuals with the same haplogroup is unlikely

Thirdly, contamination of individuals with the same haplogroup is unlikely. For 1,022 remaining individuals after removal of possibly contaminated samples, it can be argued that contamination from individuals with the same haplogroup contribute to heteroplasmy by private mutations that could not be disentangled from the haplogroup analysis. However, among all within-population pairwise comparisons, there are only 2.2% individual pairs (1,019 out of 45,289) that have the same mtDNA haplogroups. This percentage likely overestimates the severity of this issue since any other sources of contamination are unlikely to have a more similar mtDNA profile to the sequenced sample than that of a random individual from the same population sample in the 1000 Genomes Project. The chance of contamination multiplying by the chance of being the same haplogroup yields extremely low probability.

5.2.4. Potentially contaminated individuals also have evidence of heteroplasmy

Certain individuals do have suggestive evidence of sample contamination, as pointed out by Just *et al*. From the haplogroup analysis, 63 individuals have a secondary haplogroup, including all 15 individuals with more than 20 heteroplasmies identified in our original study. Some of these individuals exhibit similar minor allele frequency (MAF) among potentially contaminated heteroplasmic sites, which should approximate the fraction of sample contamination (*e.g.* Figure 5.1. B). However, analyzing most of these individuals still supports clear evidence of heteroplasmic sites, as their MAFs deviate from the supposedly contaminated ones (*e.g.* Figure 5.1. C). Additionally, the enrichment of heteroplasmy on haplogroup diagnostic sites is not necessarily an indication of contamination because those sites have much higher

mutation rates than other locations (Figure 5.1. D).

5.3. Conclusions

The original conclusions remain unchanged even after we excluded all 63 individuals with a secondary haplogroup, mainly as 910 out of 1,022 individuals carry heteroplasmy, yielding a prevalence estimate of 89.04% (as compared to 89.68% we previously reported). Furthermore, we still observed significant positive correlation between the relative mutation rate and the incidence of heteroplasmy, indicating that in sites with higher mutation rate heteroplasmy is more likely to happen independently in different individuals, even after applying a MAF cutoff of 15% ($R^2 = 0.3$, $p < 2.2e-16$). We do not understand why Just *et al.* restricted their analysis to coding region heteroplasmies for their parallel analysis. Just *et al.* further questioned the reliability of using MPS in detecting low frequency heteroplasmy (lower than 5-10%) because of false positives due to chemistry, template or bioinformatic method limitations. We agree that these are all valid concerns. Hence, in our original publication, we confirmed the reliability of MPS results using nine individuals sequenced by both Illumina and LS454. The 22 heteroplasmies identified by Illumina data, including those with very low MAF, were all observed (100%) in the LS454 data with comparable heteroplasmic frequency. We believe MPS technology provides great opportunity for mtDNA heteroplasmy study, though it does require careful quality control to fend off errors from experiment, sequencing and mapping, as also demonstrated in multiple other studies (3-6).

5.4. References

1. Ye K, Lu J, Ma F, Keinan A & Gu Z (2014) Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proc Natl Acad Sci U S A* 111(29): 10654-10659.

2. van Oven M & Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* 30(2): E386-94.

3. Diroma MA, *et al* (2014) Extraction and annotation of human mitochondrial genomes from 1000 genomes whole exome sequencing data. *BMC Genomics* 15 Suppl 3: S2-2164-15-S3-S2. Epub 2014 May 6.

4. Li M & Stoneking M (2012) A new approach for detecting low-level mutations in next-generation sequence data. *Genome Biol* 13(5): R34-2012-13-5-r34.

5. Goto H, *et al* (2011) Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome Biol* 12(6): R59-2011-12-6-r59. Epub 2011 Jun 23.

6. Gardner K, Payne BA, Horvath R & Chinnery PF (2014) Use of stereotypical mutational motifs to define resolution limits for the ultra-deep resequencing of mitochondrial DNA. *Eur J Hum Genet*

AFTERWORD

The overreaching aim of my graduate research was to introduce innovative population genomics approaches into the field of human nutrition and metabolism, with the hope of contributing to the overall scientific endeavor of elucidating the genetic basis of human health and disease. My research revealed both genome-wide patterns and specific cases. The highlights, potential applications and possible follow-ups of my graduate research are discussed below.

Project 1: *Human Expression QTLs are Enriched in Signals of Environmental Adaptation*

This project was a large-scale computational analysis and was among the first few studies to reveal the importance of regulatory variants, approximated by expression QTLs, during human evolution (1-3). Very interestingly, a study by Fraser, published almost at the same time as my project, utilized the same environmental correlation dataset included in my project (4) and also examined the relative importance of regulatory and non-synonymous coding variants during human evolution (2). This study revealed that eQTLs are 10-fold more likely than non-synonymous variants to overlap with signals of local adaptation (2). Although my research analysis revealed slightly stronger enrichment signal for eQTLs than for non-synonymous variants, the overall importance in environmental adaptation is comparable between the two types of variants. The difference between the Fraser study and my study may lie in the statistical methods used and it seems that the methods utilized by Fraser are more powerful in evaluating enrichment of eQTLs in adaptive signals.

My study also identified a list of pathways, whose regulatory variants were especially important in environmental adaptation. Promising future projects could focus on confirming and elucidating the adaptive roles of regulatory variants in these pathways. Specific hypothesis could be generated based on the known function of the pathway and its relationship with the environmental factors. Candidate causal and adaptive variants could be prioritized based on their signals of environmental correlation. The adaptive role of candidate eQTLs could be firstly tested with their regulatory effects on gene expression. Promising candidates could be further examined to identify the underlying regulatory motif. Functional experiments could be designed to test if the variants could cause phenotypic changes related to the environment.

Overall, my study, along with other similar studies, emphasizes the necessity to draw more attention on regulatory variants. My study also provides a list of hypotheses and candidate adaptive variants for follow-up case studies.

Project 2: *Natural Selection on HFE in Asian Populations Contributes to Enhanced Non-heme Iron Absorption*

My second project was a case study focusing on the gene *HFE* and its regulatory adaptation to plant-based diet in Asian populations. My evolutionary analyses confirmed the presence of adaptive signals and identified a candidate Asian-common haplotype. Gene expression analysis revealed lower *HFE* expression in carriers of the Asian-common haplotype, suggesting lower expression level of hepcidin and consequently higher absorption level of non-heme iron. By recruiting 57 Asian women volunteer, we were also to observe a consistent trend of higher non-heme iron

absorption in homozygous carriers of the Asian-common haplotype than non-carriers.

Because of the limited sample size, statistical significance was not achieved in this study. The Gu research group, collaborating with the O'Brien lab, is currently planning to expend this project and to recruit a much larger group of volunteers (~600). Once the significant association is established, fine mapping could be performed to further locate the underlying causal variant and to elucidate the acting mechanism.

This project provides a very good example of local dietary adaption. Moreover, the identified adaptive variant may contribute to ethnic difference in iron status and absorption. Further studies are needed to evaluate the need of population-specific dietary recommendation for iron.

Project 3: *Extensive Pathogenicity of Mitochondrial Heteroplasmy in Healthy Human Individuals*

This was a large-scale computational analysis and was the first study to evaluate the prevalence and pathogenicity of mitochondrial DNA (mtDNA) heteroplasmy in the healthy population using next-generation deep sequencing data. My study revealed that mtDNA heteroplasmy is present in almost all healthy human individuals and most of these mtDNA mutations are highly pathogenic, possibly contributing to age-related diseases. Very excitingly, my observations were quickly confirmed in a few studies conducted at almost the same time (5, 6). Specifically, the Diroma *et al.* study utilized the Whole Exome Sequencing data from the 1000 Genomes Project and confirmed the

prevalence of mtDNA heteroplasmy in this group of healthy individuals (6).

The prevalence of mtDNA heteroplasmic mutations and their high pathogenic potential open up a new research field to investigate the role of mtDNA heteroplasmy in aging and age-related diseases. Taking advantage of the currently available next-generation sequencing data in various diseases, including cancers, the potential roles of mtDNA mutations in disease progression could be tested. Additionally, mtDNA heteroplasmy presents a valuable tool to detect sample contamination, assisting accurate identification of genotypes in both mitochondrial and nuclear genomes (7, 8).

.

In addition to its scientific significance, the Ph.D. training in Cornell University also equipped me with the knowledge and skills that are indispensable for future human genomics research as an independent researcher. I am looking forward to the exciting projects coming up.

References

1. Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET & Pritchard JK (2009) Gene expression levels are a target of recent natural selection in the human genome. *Mol Biol Evol* 26(3): 649-658.

2. Fraser HB (2013) Gene expression drives local adaptation in humans. *Genome Res* 23(7): 1089-1096.

3. Enard D, Messer PW & Petrov DA (2014) Genome-wide signals of positive selection in human evolution. *Genome Res* 24(6): 885-895.

4. Hancock AM, *et al* (2011) Adaptations to climate-mediated selective pressures in humans. *PLoS Genet* 7(4): e1001375.

5. Rebolledo-Jaramillo B, *et al* (2014) Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. *Proc Natl Acad Sci U S A* 111(43): 15474-15479.

6. Diroma MA, *et al* (2014) Extraction and annotation of human mitochondrial genomes from 1000 genomes whole exome sequencing data. *BMC Genomics* 15 Suppl 3: S2-2164-15-S3-S2. Epub 2014 May 6.

7. Ye K, Lu J, Ma F, Keinan A & Gu Z (2014) Reply to just et al.: Mitochondrial DNA heteroplasmy could be reliably detected with massively parallel sequencing technologies. *Proc Natl Acad Sci U S A* 111(43): E4548-50.

8. Just R, Irwin J & Parson W (2014) Questioning the prevalence and reliability of human mitochondrial DNA heteroplasmy from massively parallel sequencing data. *Submitted to Proc Natl Acad Sci USA*

APPENDIX A: Publication Inclusion Authorizations

Recent Advances in Understanding the Role of Nutrition in Human Genome Evolution

A version of this manuscript was accepted for publication by the journal *Advances in Nutrition* in November 2011 with the following authorship: Kaixiong Ye and Zhenglong Gu.

The *Advances in Nutrition* Authors' Statement and Copyright Release Form authorizes the inclusion of the manuscript in this dissertation in the following section:

Rights of Authors
Effective upon acceptance for publication, ASN hereby licenses the following nonexclusive rights back to authors:
a. Patent and trademark rights to any process or procedure described in the article
b. The right to photocopy or make single electronic copies of the article for their own personal use, including for their own classroom use, or for the personal use of colleagues, provided the copies are not offered for sale and are not distributed in a systematic way outside of their employing institution (e.g. via an e-mail list or public file server). Posting of the article on a secure network (not accessible to the public) within the author's institution is permitted.
c. The right, subsequent to publication, to use the article or any part thereof free of charge in a printed compilation of works of their own, such as collected writings, theses, or lecture notes; to reuse original figures and tables in future works; to present data from the article at a meeting or conference; to include the article in their thesis or dissertation.
d. Authors may post a link on a personal website that directs readers to the article on *Advances in Nutrition* website (http://advances.nutrition.org/); full text of the final, published article cannot be posted on personal or institutional websites or repositories that are accessible to the public.

E-mail correspondence with the editor of this journal shown below also confirms authorization of the inclusion of the manuscripts in this dissertation:

Cornell

Kaixiong Ye <ky279@cornell.edu>

FW: Request written permission to include my publication in my PhD's thesis

1 message

Sarah McCormack <SMcCormack@nutrition.org> Thu, Dec 11, 2014 at 5:28 PM
To: "ky279@cornell.edu" <ky279@cornell.edu>

Dear Dr. Kaixiong Ye,

Authors of content published in *Advances in Nutrition* (*AN*) are automatically licensed back the right to republish that article in their thesis. These and other author rights are posted at http://www.nutrition.org/publications/guidelines-and-policies/permissions/#rights. Thank you for publishing in *AN*, and I wish you the best of luck with your thesis.

Sarah L. McCormack

Editorial Manager

American Society for Nutrition

9650 Rockville Pike

Bethesda, MD 20814

P: 301-634-7279/F: 301-634-7892

From: Advances [mailto:advances.djs@sheridan.com]
Sent: Thursday, December 11, 2014 10:20 AM
To: Darren Early
Cc: Karine Zbiegniewicz; Katie Dunn
Subject: FW: Request written permission to include my publication in my PhD's thesis

From: Kaixiong Ye [mailto:ky279@cornell.edu]
Sent: Wednesday, December 10, 2014 10:21 PM

Human Expression QTLs are Enriched in Signals of Environmental Adaptation

A version of this manuscript was accepted for publication by the journal *Genome Biology and Evolution* in August 2013 with the following authorship: Kaixiong Ye, Jian Lu, Srilakshmi Madhura Raj, and Zhenglong Gu. E-mail correspondence with the editor of this journal shown below authorizes the inclusion of the manuscript in this dissertation:

Cornell

Kaixiong Ye <ky279@cornell.edu>

Request written permission to include my publication in my PhD's thesis

JOURNALS PERMISSIONS <Journals.Permissions@oup.com> Thu, Dec 11, 2014 at 4:37 AM
To: Kaixiong Ye <ky279@cornell.edu>

Dear Kaixiong,

Re: Kaixiong Ye, Jian Lu, Srilakshmi Madhura Raj, and Zhenglong Gu. Human Expression QTLs Are Enriched in Signals of Environmental Adaptation *Genome Biol Evol (2013) Vol. 5 1689-1701*

Thank you for your request. As part of your copyright agreement with Oxford University Press you have retained the right, after publication, to include the article in full or in part in a thesis or dissertation, provided that this is not published commercially. Please be advised that in terms of electronic versions of your thesis, you may upload a PDF of the version of record to institutional and/or centrally organized repositories, upon publication in the journal.

When uploading the version of record to a repository, authors should include a credit line and a link to the article on the OUP website. This will guarantee that the definitive version is readily available to those accessing your article from public repositories, and means that your article is more likely to be cited correctly.

For full details of the self-archiving policy for this journal please follow this link:

http://www.oxfordjournals.org/en/access-purchase/rights-and-permissions/self-archiving-policyc.html

Kind regards,

Guffi

Guffi Chohdri (Ms)
Rights Assistant

Academic Rights & Journals

Tel: +44 (0)1865 354454

Email: guffi.chohdri@oup.com

Extensive Pathogenicity of Mitochondrial Heteroplasmy in Healthy Human Individuals

and

Mitochondrial DNA heteroplasmy could be reliably detected with massively parallel sequencing technologies

These two manuscripts have been accepted for publication by the journal *Proceedings of the National Academy of Sciences* respectively in July and October 2014 with the following authorship: Kaixiong Ye, Jian Lu, Fei Ma, Alon Keinan, and Zhenglong Gu. As the author of these two manuscripts, I have the rights to include the two manuscripts in my thesis as explained by the journal "Author Rights and Permissions" as follows:

"As a PNAS author, you and your employing institution or company retain extensive rights for use of your materials and intellectual property. You retain these rights and permissions without having to obtain explicit permission from PNAS, provided that you cite the original source:

- The right to include your article in your thesis or dissertation."

E-mail correspondence with the editor of this journal shown below also confirms authorization of the inclusion of the manuscripts in this dissertation:

Cornell

Kaixiong Ye <ky279@cornell.edu>

Request written permission to include my publications in my PhD's thesis

PNAS Permissions <PNASPermissions@nas.edu> Thu, Dec 11, 2014 at 9:33 AM
To: Kaixiong Ye <ky279@cornell.edu>

Yes, that is correct: Authors do not need to obtain permission for the following uses of material they have published in PNAS: (1) to use their original figures or tables in their future works; (2) to make copies of their papers for their classroom teaching; or (3) to include their papers as part of their dissertations (this includes both PNAS articles and replies).

Please cite each PNAS publication in full when re-using the material. Because this material published after 2008, a copyright note is not needed. Feel free to contact us with any additional questions you might have.

Best regards,

Kay McLaughlin for

Diane Sullenberger

Executive Editor

PNAS

From: Kaixiong Ye [mailto:ky279@cornell.edu]
Sent: Wednesday, December 10, 2014 9:57 PM
To: PNAS Permissions
Subject: Request written permission to include my publications in my PhD's thesis

Dear Editors,

I am writing to request your formal permission for me to include my two publications in my PhD's thesis. I am the first authors for these two publications. They are:

1) Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals

Kaixiong Ye, Jian Lu, Fei Ma, Alon Keinan, and Zhenglong Gu

http://www.pnas.org/content/111/29/10654.abstract

2) Reply to Just et al.: Mitochondrial DNA heteroplasmy could be reliably detected with massively parallel sequencing technologies

Kaixiong Ye, Jian Lu, Fei Ma, Alon Keinan, and Zhenglong Gu

http://www.pnas.org/content/111/43/E4548.extract

I understand that it is the author' right "to include your article in your thesis or dissertation" but your confirmation by replying this email will be very much appreciated. Thank you very much for your time and help!

Best wishes.

Kaixiong

--

Kaixiong (Calvin) Ye

PhD

Division of Nutritional Sciences
Cornell University
317 Savage Hall
Ithaca, NY 14853
607-793-1458

http://www.human.cornell.edu/bio.cfm?netid=ky279