Global Signatures of Selection in Humans

A dissertation presented

by

Mashaal Sohail

to

The Committee on Higher Degrees in Systems Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Systems Biology

Harvard University

Cambridge, Massachusetts

April 2018

Dissertation Advisor: Professor Shamil Sunyaev Mashaal Sohail

Global Signatures of Selection in Humans

Abstract

This thesis studies the properties and prevalence of natural selection operating on our species on a global or genome-wide scale. It is still not known how negative selection against deleterious mutations operates genome-wide, whether balancing selection is responsible for maintaining higher genetic diversity in some parts of the genome, and how prevalent adaptation on complex or polygenic traits is in humans. We now have access to population genetic data from healthy populations worldwide, methods to evaluate the functional effects and age of genetic mutations, and Genome-wide Association (GWA) studies to estimate the genetic basis of a complex trait.

Using these resources, as well as developing novel statistical methodologies, we show that 1) negative selection in humans involves synergistic epistasis, or deleterious mutations in the human genome interact globally in a manner to reinforce each other's effects, 2) genes with monoallelic expression contribute disproportionately to genetic diversity in humans, which is maintained through the evolutionary forces of balancing selection and a higher mutation and recombination rate 3) the signal for polygenic

adaptation at height-associated genetic variants in humans is confounded by ascertainment biases in the GWAS used to estimate height across populations.

We conclude that (1) helps explain how humans tolerate a high incoming rate of deleterious mutations, and why sexual reproduction may have an evolutionary advantage, (2) establishes a link between genetic and epigenetic mechanisms of maintaining and exhibiting high genetic diversity in humans, and (3) demonstrates the presence of confounders that restrict interpretation of signals of polygenic adaptation in humans found using currently existing GWA studies.

Table of Contents

Table of Figures and Tables

We are what we are, because the universe is what it is. If it acts upon us, we react upon it.

- John William Draper

To my mother and father for their unconditional love, support and guidance

Acknowledgements

I am deeply indebted to many people for standing here today. This thesis would not have been possible without my advisor Professor Shamil Sunyaev to whose mentorship I owe almost all my knowledge of genetics, but first to those who led me to him.

I can root my interests in genetics, evolution and human history to conversations that my father, Sohail Muzaffar, and grandfather, Munawar Hussain, engaged me in since I was a little girl. Their knowledge of history and philosophy and speculations about science layered the base of my interests that would sprout out in my desire to pursue a PhD in systems biology and an AM in history of science. I would like to thank my mother, Ayesha Sohail, for encouraging me from a young age to challenge social norms and pursue my educational and professional dreams. My parents have sacrificed a lot to get me to this footing in the world, and for this, I am forever indebted to them.

Shamil took my untamed curiosity and channeled it into the development of a trained eye that can decipher true findings from mere caveats. With humor and humility, he gave me feedback over the years and allowed me to learn from my own mistakes. He allowed me to explore my interests even if they trespassed his own and supported my pursuits in science communication and history of science. I am grateful to Shamil for making me into a scientist, for having a contagious enthusiasm for all things genetics that he has passed on to me, and for a wonderful 5.5 years in his lab which I consider to be the most intellectually stimulating of my life so far.

I have been lucky to have many great mentors who have given me advice, encouragement and opportunities over the years. I would like to thank Professor Andrew Murray and Professor Tim Mitchison for the empathy and personal attention with which they helped me through a hump in the first year of graduate school, Professor Alexey Kondrashov for his mentorship in evolutionary genetics and writing effectively, Professors Jeff Gore, Alkes Price and Andrew Murray for serving on my qualifying exam committee, Professors Pardis Sabeti, Leonid Mirny, John Wakeley and David Reich for their support and encouragement as members of my advisory committee over the PhD years, Professor David Reich for mentorship and access to ancient DNA during my last year, and Professor John Wakeley for his support of both my pursuits in human genetics and history of science, and for thoughtful conversations at their intersection over the years. I would like to thank Professors Peter Galison, Elizabeth Lunbeck, Janet Browne and Ahmed Ragab for their time and mentorship that has allowed me to view science and my own work in historical and social context, and to learn how to communicate effectively on these matters.

I would like to thank my lab mates over the years, especially Dr. Daniel Balick, Dr. Dana Vuzman, Dr. Ivan Adzhubey, Dr. Tobias Lenz, Dr. Jae Hoon Sul and David Radke for teaching me many tricks over the years, and for being there through the thick and thin of research. I would like to thank the systems biology program for its flexibility and rigor and Samantha Reed and Elizabeth Pomerantz for always making sure I was on track. I would like to thank my collaborators Olga Vakhrusheva, Professor Paul de Bakker, Dr. Sara Pulit, Dr. Laurent Francioli. Dr. Sung Chun, Dr.

Virginia Savova and Professor Alexander Gimelbrant for the opportunity to work together on exciting, interdisciplinary global projects.

I would like to thank my friends and colleagues who made these years not only the most stimulating but also some of the most fun! I would like to thank my friends Sebastian Akle, Hong Sun, Tami Lieberman, Zain Ali, Olivia Ho-Shing, Miriam Huntley, Noam Prywes, Coleman Connelly, Winston Anthony, Adriana Vazquez, Lana Awad, Alice Nawfal, Katrina Schoen, Hiba Awad, Henry Craig, Robert Habib, Jordan Howell, Maryam Alemi, Greg Koytiger, Brian Arnold, Laura Stone, Daniel Schultz, Ali Nomani, Shireen Hamza, Palwasha Durrani and Anum Malkani for everything, for always challenging me to do and think better, for encouraging me in moments of self-doubt, and for being there to enjoy the moments in between. Having spent the last ten years of my life far from my family, my friends have been a constant source of warmth and love when I have needed it most.

I would like to thank the dervishes of the Khaniqah Nimatullahi for always having a meal ready when I needed it, for being a home away from home, and for always providing a place of stillness outside the flux of daily life. I would like to thank Dr. Alireza Nurbakhsh for being a constant source of inspiration, love and guidance during my PhD years. Last but not least, I would like to thank my partner Daniel Quesada Lombó for his constant love, humor, encouragement, support, and relentless curiosity for and critique of my intellectual pursuits.

Chapter 0. Introduction

19th century dialogues – natural selection or adaptation?

The term "natural selection" was introduced by Charles Darwin in 1859.[1] Contemporaries of Darwin had varying reactions to the term. John William Draper, a Victorian American scientist, arguably an evolutionist earlier than Darwin[2], called the phrase a regrettable addition to conversations about the evolution or transmutation of organic beings including humans that had been persisting since at least Aristotle.[3,4] Draper preferred the term adaptation.[3] He argued that living beings evolved through adapting to varying environments over sustained periods of time.[5] We start this dissertation with this dialogue between 19th century evolutionists because it will provide important context to the questions that will be the subject of its contents.

The differences between Draper and Darwin's view of the evolutionary process in the 19th century will frame the debates that we will step into in this thesis over a century later. We will first briefly summarize the two major axes of these debates as they emerged in the 19th century. The first axis concerned the degree of emphasis on adaptation vs. on natural selection. Draper, similarly to Wallace, stated the primary importance of local environments in driving evolutionary change.[5,6] A "variety" or a "species" either adapted to their local environment or went extinct.[5,6] Darwin moved the emphasis from adaptation to natural selection, or competition between individuals and "races" to survive and reproduce.[1] Draper challenged this term "natural selection" and said of it,

> It is very unscientific, very inferior to the old expression adaptation. It implies a personification of Nature. It is anthropomorphic. But Nature never selects, never accepts or rejects, knows nothing about duties, nothing about fitness or unfitness. Nature simply obeys laws. … Natural selection is thus supposed to perpetuate an organism after adaptation to its environment has taken place. The change implied by adaptation must precede it. It should be regarded rather as a metaphorical expression than a scientific statement of an actual physical event. … Darwinism, therefore, does not touch the great question as to the manner in which variation of organisms arises. It only teaches how such variations are perpetuated.[3]

The second major axis of difference concerned the degree of determinism or chance (stochasticity) in the evolutionary process. Draper believed that the "reign of law" over the evolution of organic beings was deterministic, and that there were no accidents in evolution.[7] Darwin, in contrast, used chance as a significant discursive unit in his theory.[8]

In this thesis, we will take up questions on the nature of the evolutionary process as they concern the evolution of humans along these same axes. We now have access to genome-wide genetic data from healthy populations worldwide that allow us to address questions on the nature of natural selection or adaptation in a way not possible before. We can ask how prevalent such evolutionary forces have been in our history and in today's modern industrialized world. Through modeling approaches and computer simulations, we can also interrogate if deterministic or stochastic factors are leading to the genetic signatures we observe. Each of the chapters in this thesis addresses a subset of the above

questions, which can be traced back to 19th century dialogues about evolution. We will now develop the intellectual background for each of the chapters by reviewing the state of knowledge in the field prior to the investigations set forth in this thesis.

The genetic load paradox and the evolution of sex

First conceptualized by Haldane, the genetic or mutation load came to describe the loss in average fitness of a population due to the influx of deleterious mutations every generation.[9,10] In 1950, Muller estimated the mutation load to be roughly 20% in humans by using Danforth's method to infer mutation frequency from disease frequency and persistence, and argued for rational guidance of reproduction or a eugenics agenda.[11] He had assumed a model where every deleterious mutation contributed independently to fitness. However, we can also imagine a scenario in which deleterious mutations interact with each other in a synergistic or antagonistic fashion, that is, they are epistatic with each other.

This was exactly what was imagined by the likes of Kimura and Maruyama in 1966 and others after them.[12,13] They demonstrated that if negative selection, or natural selection against deleterious mutations was indeed epistatic, the genetic load would be dramatically reduced.[12,13] Recent estimates hold the mutate rate in humans to be 1.27×10^{-8} new mutations per individual every generation.[14] This means each new born receives ~70 new mutations, of which ~7 are likely

deleterious if ~10% of the human genome is selectively constrained.[15] With such a high mutation rate, the mutation load remains unrealistic for our species if selection proceeds without epistasis, but no global investigation of the nature of epistasis with regard to negative selection in humans has been carried out as of yet.

Beyond the mutation load paradox, learning the form of negative selection and the underlying mode of epistasis is significant for understanding evolutionary dynamics and the genotype-phenotype map generally.[16] While the form of the fitness function for negative selection remains elusive in humans, investigations in model organisms suggest a trend from antagonistic to synergistic epistasis if organisms are arranged from viruses to prokaryotes to eukaryotes.[17] A particularly interesting aspect of evolutionary dynamics that is affected by the mode of epistasis is the maintenance of sex.[18] Sexual reproduction has a two-fold cost compared to asexual reproduction in a genetic sense: this has been called the enigma of sex.[19] Sex must, thus, provide a fitness advantage over asexual reproduction to persist over time.

As long as sexual reproduction creates combinations of genetic variants that increase the average fitness of the population at equilibrium, sex will have an evolutionary advantage.[20] Deterministically, whether sexual reproducers have a fitness advantage over asexual reproducers in a given mutational regime depends on the mode of epistasis.[18] In the mutational regime for humans, sexual reproducers only have a higher equilibrium fitness than asexual reproducers if

deleterious mutations interact synergistically.[18] In contrast, if deleterious mutations interact antagonistically, asexual reproducers have a higher equilibrium fitness.[18] The deterministic hypothesis for the maintenance of sex thus holds that sex will be selected for as long deleterious mutations interact synergistically.[13] Stochastic explanations hold that sex can gain an evolutionary advantage by chance even if mutations contribute to fitness independently if population sizes are small enough.[21]

Thus, with the evolutionary implications for the genetic load and the evolution of sex, and related implications for arguments for eugenics and CRISPR-mediated interventions in humans (see more in the conclusion) in mind, we undertake an investigation into the global mode of epistasis between deleterious mutations in humans, which to our knowledge, is the first of its kind (chapter 1).

The evolutionary role of stochastic epigenetic variation

Allele specific gene expression can occur in humans in a deterministic or stochastic manner as well. A well-established example of deterministic allele specific expression genome-wide is imprinting when an allele is silenced in a parent-of-origin specific manner.[22] Another example of allele specific expression is X inactivation and only affects the X chromosome, whereby one copy of the gene is silenced stochastically.[23] In the last decade, another example of stochastic

allele specific expression has emerged which impacts autosomes as well. Known as random monoallelic expression (MAE), the choice for which allele is silenced is made randomly but is clonally stable.[24] Recent advances in transcriptome-wide analyses have surprisingly found that 10-25% of human and mouse autosomal genes exhibit monoallelic expression in multiple cell types.[24,25,26,27,28,29] MAE has been directly observed in peripheral blood and derived cell lines, as well as in human placenta[24], mouse lymphoid cells and fibroblasts[25], and mouse neuroprogenitor cells[29].

It was predicted that MAE genes would harbor lower genetic diversity as they must have undergone stronger negative selection due to recessive variation in the gene body being exposed and purged more efficiently.[30] Instead, these genes showed a marked higher genetic diversity compared to genes that are biallelically expressed (chapter 2). High genetic diversity is also a feature of balancing selection whereby multiple alleles are maintained in the gene pool of a population at frequencies higher than expected by genetic drift alone.[31]

Recent investigations have suggested that a sizable portion of genetic variation in humans may be maintained through long-term balancing selection and have catalogued likely examples of such selection genome-wide.[32,33] Population genetic methods have also become available to date allelic age using nearby genetic variation[34], and to estimate time to most recent common ancestor for genes by modeling their ancestral recombination graph (ARG).[35] These along with the availability of a high quality Neanderthal genome sequence,[36] and

catalogs of population genetic variation in global human populations[37,38,39] allow for an investigation into the evolutionary basis of random MAE genome-wide through a collaboration with Alexander Gimelbrant's lab (chapter 2).

Polygenic adaptation and the GWAS revolution

Natural selection through subtle allele frequency shifts in a large number of genetic variants involved in a trait, or polygenic adaptation, has recently invoked interest.[40] A plethora of genome wide association studies (GWAS) as well as the availability of genotype data from different populations has spurred the investigation of signatures of polygenic adaptation in complex human traits.[40,41,42,43,44] In order to view population genetic differences in complex traits, studies have compared allele frequencies between populations of variants associated with a trait by GWAS weighted by their estimated effect size on the trait.[40,41] Such methods have also theoretically or empirically modeled genetic drift to infer whether any observed differences are due to genetic drift or adaptation.[40,41]

These studies have observed significant genetic differences in height between Northern and Southern Europeans and have concluded that these differences are due to polygenic adaptation.[40,41] Moreover, ancient DNA and global populations have been used to conclude that genetic differences in height persist back in Europe up to 10,000 years ago[45] and persist today across global

populations.[43] The only other complex traits for which tentative signals of polygenic adaptation have been reported are infant head circumference, hip circumference, waist-hip ratio, type 2 diabetes, inflammatory bowel disease, schizophrenia, extraversion, subjective well-being, hippocampus volume, and putamen volume.[43,44]

However, concerns remains as to the interpretation of such signals of polygenic adaptation as well as about technical caveats related to their estimation. The major concern regards the presence of residual population structure in the GWAS used to estimate genetic traits, which may propagate into artificial differences between populations. Population structure can bias estimates of effect sizes of genetic variants, as well as bias the actual variants that are discovered in the GWAS to be associated with the trait (chapter 3). Issues of transferability of GWAS estimates from one population to estimate traits in another have also been exposed; these may be due to different LD structure, genetic epistasis and genotype x environment interactions.[46] With these in mind, we sought to investigate the role of confounders in observations of genetic differences between populations, and exposed the likely role of population structure in the observation of polygenic adaptation on height in humans (chapter 3).