ON IDENTIFYING SIGNATURES OF POSITIVE SELECTION IN HUMAN POPULATIONS

A Dissertation Presented

By

JESSICA L. CRISCI

Submitted to the Faculty of the
University of Massachusetts Graduate School of Biomedical Sciences, Worcester
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

JUNE 25, 2013

POPULATION GENETICS

ON IDENTIFYING SIGNATURES OF POSITIVE SELECTION IN HUMAN POPULATIONS

A Dissertation Presented
By

JESSICA L. CRISCI

The signatures of the Dissertation Defense Committee signify completion and approval as to the style and content of the Dissertation

Jeffrey D. Jensen, Ph.D., Thesis Advisor

Daniel Bolon, Ph.D., Member of Committee

Jeffrey Bailey, M.D., Ph.D., Member of Committee

Manuel Garber, Ph.D., Member of Committee

Hopi Hoekstra, Ph.D., Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee

Konstantin Zeldovich, Ph.D., Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the school.

Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

Bioinformatics and Computational Biology

June 25, 2013

ACKNOWLEDGEMENTS

First, I would like to acknowledge my wonderful husband Tim (because he always comes first!). He has been as loving and supportive as any wife could hope for throughout my graduate career and has encouraged and deepened my love of science and learning.

I would also like to thank my advisor, Jeff Jensen, for introducing me to population genetics, for being patient, for answering my emails in the middle of the night, and for always believing in me, even when I had doubts about myself. I owe most of my success to his brilliance in promoting and supporting our work, and for constantly insisting that I can do more than I think.

Next, to my committee members, both past and present (Konstantin Zeldovich, Dan Bolon, Jeff Bailey, Manuel Garber, Hopi Hoekstra, Zhiping Weng, Evgeny Rogaev, Shahram Akbarian), I would like to thank you for your inspirational counsel, and for helping move through the various stages of my PhD career. I (quite literally) would not be writing this dissertation at all if not for your guidance and continual support. And special thanks to my program chair, Konstantin, for tolerating me, and my questions, and always providing answers, no matter how many times I bugged you. As the first student in BCB, I'm sure I made you do a lot of extra work, for which I am most grateful and sorry, in equal parts. I wish you better luck with the next upcoming graduate.

And finally, I'd like to thank my family – especially my mother and father – for their support, for pretending to know what I'm talking about and even feigning interest

when I describe what I do, and for being proud of me – because you are the reason I made it this far, and I did this as much for you as for myself.

ABSTRACT

As sequencing technology continues to produce better quality genomes at decreasing costs, there has been a recent surge in the variety of data that we are now able to analyze. This is particularly true with regards to our understanding of the human genome – where the last decade has seen data advances in primate epigenomics, ancient hominid genomics, and a proliferation of human polymorphism data from multiple populations. In order to utilize such data however, it has become critical to develop increasingly sophisticated tools spanning both bioinformatics and statistical inference. In population genetics particularly, new statistical approaches for analyzing population data are constantly being developed – unfortunately, often without proper model testing and evaluation of type-I and type-II error. Because the common Wright-Fisher assumptions underlying such models are generally violated in natural populations, this statistical testing is critical. Thus, my dissertation has two distinct but related themes: 1) evaluating methods of statistical inference in population genetics, and 2) utilizing these methods to analyze the evolutionary history of humans and our closest relatives. The resulting collection of work has not only provided important biological insights (including some of the first strong evidence of selection on human-specific epigenetic modifications (Shulha, Crisci, Reshetov, Tushir et al. 2012, PLoS Bio), and a characterization of human-specific genetic changes distinguishing modern humans from Neanderthals (Crisci et al. 2011, GBE)), but also important insights in to the performance of population genetic methodologies which will motivate the future development of improved approaches for statistical inference (Crisci et al, in review).

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

GLOSSARY

Positive selection- an increase in the frequency of an allele due to fitness benefits

Genetic drift- the fluctuation seen in allele frequencies due to random sampling

Site frequency spectrum (SFS) - a summary of all mutations and their frequencies within a sample

Selective sweep- the fixation of a beneficial allele that results in reduced heterozygosity and high-frequency derived alleles

Derived allele- a newly acquired allele within a sample distinguished by an outgroup sequence

Linkage disequilibrium (LD) - nonrandom association of alleles due to local recombination rates

Effective population size (N_e) - the number of individuals in a population that satisfy conditions of equilibrium, e.g. randomly mating, constant size, non-overlapping generations, etc.

PREFACE

Parts of this dissertation appear in:

Crisci, J. L., Wong, A., Good, J. M., & Jensen, J. D. (2011). On Characterizing Adaptive Events Unique to Modern Humans. *Gen Biol Evol*, *3*, 791–798.

Crisci, J.L. & Jensen, J.D. (2012). Evolution of the Human Genome: Adaptive Changes. In: eLS 2012, John Wiley & Sons, Ltd: Chichester http://www.els.net/

Crisci, J. L., Poh, Y.-P., Bean, A., Simkin, A., & Jensen, J. D. (2012). Recent Progress in Polymorphism-Based Population Genetic Inference. *J Hered*, *103*(2), 287–296.

Shulha, H. P.*, Crisci, J. L.*, Reshetov, D.*, Tushir, J. S.*, Cheung, I., Bharadwaj, R., Chou, H.-J., et al. (2012). Human-specific histone methylation signatures at transcription start sites in prefrontal neurons. *PLoS Biol*, *10*(11), e1001427.

Crisci, J.L., Poh, Y-P., Mahajan, S., & Jensen, J.D. The Impact of Equilibrium Assumptions on Tests of Selection. (in progress).

*equal contribution

CHAPTER I. Introduction

Ever since Charles Darwin proposed his theory of evolution by the mechanism of natural selection (1859), there has been considerable interest in the scientific community of devising ways to measure the effects of selection in populations. The field of Population Genetics was born out of the first attempts at describing how allele frequencies change between generations and how this process is affected by selection (e.g. Haldane 1927, Fisher 1930, Wright 1931). In 1974, Maynard Smith and Haigh described the hitchhiking effect of a beneficial allele and the impact this has on genetic variation. Briefly, when a new mutation increases in a population due to the effects of positive selection it will impact the frequencies of linked neutral alleles, thereby leaving a pattern of reduced genetic variation that will decrease from the site of the selective event. The extent of this signature is determined by the strength of the selection coefficient and the local recombination rate (Kaplan et al. 1989). When genome sequencing became a reality at the turn of the 20th century, this signature, known as a selective sweep, became the foundation for many statistics that have been developed to identify selection on whole-genome and sub-genomic datasets. And now sequencing technology has advanced to a point where we can sequence ancient genomes (e.g. the Neanderthal (Green et al. 2010)) and multi-species epigenomes (Shulha, Crisci, Reshetov, Tushir et al. 2012). With these novel datasets the question remains as to whether they can be used to provide new insight into how positive selection has shaped human populations.

Historically, human evolution was studied using a phenotype-first approach - relating phenotypic differences in populations to underlying genotypes. Perhaps the best-

known example is the Duffy blood antigen, believed to be driven by selection for resistance to malaria. Individuals lacking the Duffy blood antigens are resistant to infection by *Plasmodium vivax*, and this phenotype is correlated with regions in Africa where transmission of *P. vivax* is high (Miller et al. 1975). Despite the intuitive appeal of this approach, it is focusing only upon differences in phenotypic variation between populations, which may or may not have been influenced by positive selection. In fact, perhaps the true advantage of the genomic age is the ability to take a genotype-first approach – to scan the genome for adaptive mutations, using signature patterns of positive selection, in a fashion that is blind to underlying phenotype.

But this approach has its own caveats. Firstly, different mechanisms of selection leave different signatures in the genome, and it is unclear how much each of these processes affects human populations. Also, identified selective targets do not always have an obvious phenotypic consequence or advantage. If, for example, an identified gene plays a role in many cellular processes, it is difficult to determine which of these may have been targeted, and it is dangerous to assume that the one that makes the most 'sense' from an evolutionary or biological standpoint must be the right process (Pavlidis et al. 2012). This leads to questions of whether the signals we are finding are, in fact, real indicators of selection or artifacts of neutral processes. Finally, despite a proliferation of statistics designed for scanning genomes for evidence of selection (for review see Crisci et al. 2012) there is alarmingly little overlap between such studies and methodologies (Akey 2009). This raises important questions both about the mode and tempo of human evolution, as well as the efficiency of the statistics themselves. Thus, we have only a

handful of convincing examples of adaptations in human populations, largely arrived at using a phenotype-first approach – suggesting that the full benefit of population genomics has yet to be realized.

Evidence of Adaptations in Humans

Adaptation in humans is generally presented in two forms: population-specific changes that are segregating at some frequency in the species, and species-specific changes that have arisen since the split with our closest relatives (generally the chimpanzee, but recently ancient hominids as well, see below). The latter are more likely to explain distinctive neurological traits in humans, like language, learning and memory – but few convincing examples have been found to date. Thus, most examples of human adaptation are population-specific that most likely arose in response to environmental changes as humans spread to nearly every continent in the world. Common phenotypic traits affected are disease resistance, and metabolism in response to changes in diet.

Population specific adaptations

In addition to the Duffy blood group discussed above, malaria has driven selection of other traits in populations where transmission is prevalent, including sickle cell anemia (Allison 1954). Sickle cells are caused by a variant in the human hemoglobin gene (*HbS*). Individuals who are heterozygous for the trait are more resistant to infection by *Plasmodium falciparum*, whereas those who are homozygous have higher mortality rates. The resistance this variant confers on heterozygotes explains why the trait remains at a frequency of around 10% in African populations, even when it appeared at a first glance to be deleterious. This allele has both a beneficial and a deleterious phenotype,

and is indeed one of the classic examples of balancing selection. Malaria remains today a selective pressure in many extant populations, and is thought to have driven selection on multiple variants of the hemoglobin gene that cause human blood disorders (see Kwiatkowski 2005 for review).

Many diseases act as selective pressure on the immune system, promoting the evolution of resistance mutations. For instance, the CCR5 receptor is normally expressed on the membranes of CD4 T-cells and provides entry for the HIV virus. A 32bp deletion in this gene in individuals of European descent prevents this receptor from being expressed on the membranes of CD4 T-cells, and confers resistance to HIV infection (Samson et al. 1996). This deletion is present in approximately 10% of Caucasian Europeans. The age of the variant allele has been estimated to be around 1000-2000 years (for review, see Galvani and Novembre 2005); if this age were correct, it would be unlikely for this mutation to have reached this appreciable frequency by genetic drift alone (Stephens et al. 1998). And since HIV is believed to be a modern disease in humans, it is an unlikely explanation for the observed frequency. Initially, the Bubonic Plague was named as the selective pressure on the variant allele for CCR5 because of the timing of the mutation, but studies have subsequently demonstrated that this variant does not provide resistance to plague infection (Elvin et al. 2004, Mecsas et al. 2004). A more likely culprit is small pox, since it was highly transmissible for a long period of human history – and a relative of poxvirus infects cells using the CCR5 receptor (Lalani et al. 1999).

Another instance of population-specific selection as a result of environmental pressure is adaptation to a low oxygen environment in Tibetan populations – with modern populations living at approximately 2.5 miles above sea level, and thus at a 40% oxygen deficiency. Several recent papers have compared genetic data between Tibetan and Han populations to try and elucidate changes that could allow humans to live at such altitudes (Beall et al. 2010, Yi et al. 2010, Peng et al. 2011, Xu et al. 2011). All of these studies consistently highlight one gene, *EPAS1*, as being highly differentiated in the Tibetan population. *EPAS1* is responsible for regulating factors in response to hypoxia, including erythropoiesis (Patel and Simon 2008). Additionally, Yi et al (2010) find that *EPAS1* is correlated with hemoglobin levels in the blood and could explain why Tibetans have lower levels of hemoglobin at high altitudes than lowland populations.

An example of a metabolic adaptation is the lactose tolerance phenotype. Being the only mammal to continue milk consumption after infancy, humans exhibit lactose tolerance, or lactase persistence, which results from the continual expression of lactase-phlorizin hydrolase, LPH (*LCT*) into adulthood. Normally, levels of this enzyme decrease after infancy, and adults lose their ability to digest lactose in the intestines. The lactase persistence trait is present at a frequency between 40-90% in European and African populations that raise cattle (Swallow 2003). In 2007, Tishkoff et al. (2006) demonstrated that this was indeed an example of convergent evolution – with two different alleles conferring the phenotype between populations.

Human specific adaptations

The availability of both the Neanderthal and Denisovan genomic sequences is a noteworthy milestone in the study of human evolution. These two populations are much more closely related to human than chimpanzee (see Figure 2.1), and can provide unique insight into genomic changes that occurred during early human evolution. By comparing the Neanderthal genome sequence with the genomes of 5 humans from various populations and using chimpanzee as an ancestor, Green et al. (2010) identified putatively selected regions in humans. They looked for large regions of the human genome where Neanderthal had the ancestral state at polymorphic sites in humans, with the logic being that these mutations in humans must have occurred and rose in frequency after the split between humans and Neanderthal. Crisci et al. (2011), further show that this scan is capable of detecting selection in regions that would have been missed using site frequency spectrum- and divergence based approaches – with the most interesting candidates being: *CADPS2*, mutations in which are linked with autism; *NRG3*, which is expressed in the brain and located within a susceptibility locus for schizophrenia; and *DYRK1A*, also expressed in the brain and believed to be involved in learning and memory.

One interesting discovery from sequencing the genomes of these two hominins is that both populations appear to have interbred with human populations. Green et al. show that Neanderthals contributed up to 3% of their genomes to modern day Eurasians by comparing the Neanderthal genome to modern European, Asian, and African populations, finding that Neanderthals were more genetically similar to Eurasians than to Africans, suggesting gene flow. Reich et al. (2010) performed a similar analysis with the

Denisovan genome and found that this population contributed 4-6% of its genome to modern day Melanesians.

This discovery of admixture raises some interesting questions regarding the evolutionary trajectory of humans. It is possible that since these two populations were present in Europe and Asia before modern humans, they could have acquired adaptive mutations in response to environmental and dietary changes, and passed them on to human ancestors as they began migrating out of Africa. A potential example of this is the controversial evolution of the *FOXP2* gene. This gene has apparent functions in speech and language (Fisher et al. 1998, Lai et al. 2001), and contains two SNPs initially found to be unique to humans that have been argued to be under positive selection (Enard et al. 2002a). Later Krause et al. (2007) discovered that these SNPs were also present in the human-derived state in 2 Neanderthal individuals, and suggested that there was a common mutation in the ancestor of humans and Neanderthals before the two populations split over 300 Kya. But Coop et al. (2008) argue that the selective signature in humans is much too young to have occurred in an ancestor of human and Neanderthal, and that if the sweep was that old, new mutations would have returned local diversity levels back to neutral expectations. However, if there was in fact admixture between human and Neanderthal populations approximately 50 Kya (Green et al. 2010), then this selected loci very well could have arisen and swept in humans, and then the haplotype could have been passed to Neanderthals (or vise versa).

Datasets and Methodology

With many genomes now being sequenced and the ability to process large amounts of data using high-performance computing, the time required to perform large genome scans of many individuals and compare polymorphism between populations is trivial. We now have genomic sequences of the extant great apes, including human (International Human Genome Sequencing Consortium 2001, Venter et al. 2001), chimpanzee (The Chimpanzee Sequencing and Analysis Consortium 2005), gorilla (Scally et al. 2012), orangutan (Locke et al. 2011), and most recently bonobo (Prüfer et al. 2012). Also, the draft genomes of two extinct hominins have been completed within the last few years: Neanderthal (Green et al. 2010) and an individual from Denisova cave in Siberia (Reich et al. 2010). This divergence data facilitates the discovery of human-specific adaptations – often by making simple comparisons of the rate of fixation between branches. Commonly the ratio of nonsynonymous changes (d_N) to synonymous (d_S) is used as a measure of the direction of selection across a gene; with $d_N/d_S = 1$ being consistent with neutrality, $d_N/d_S > 1$ consistent with recurrent positive selection, and $d_N/d_S < 1$ consistent with recurrent purifying selection (Nei and Gojobori 1986).

Combining this divergence-based approach with polymorphism data, the McDonald-Kretiman (MK) test performs a 2x2 contingency test between fixed vs. segregating synonymous and non-synonymous sites. Under the assumption that synonymous sites are neutral, an increased rate of fixation of nonsynonymous changes between species is generally taken as evidence of recurrent adaptive fixations (McDonald and Kreitman 1991).

There are also methods that utilize polymorphism within a single species to identify patterns of selection. Next-generation genome-sequencing technology has made it faster and more cost effective to sequence entire genomes of many individuals, leading to large-scale polymorphism datasets. Indeed, the 1000 genomes project has provided scientists with the most complete set of genome-wide SNP information in humans to date (Durbin et al. 2010). All such tests rely on the patterns of variation produced by a hitchhiking event – the process by which a new beneficial allele rises quickly within a population due to positive selection, altering the frequency of linked neutral variation (Maynard Smith and Haigh 1974, Kaplan et al. 1989; Figure 1.2). For the fixation of a single beneficial mutation – these patterns are well described, including a decrease in local heterozygosity, an excess of rare mutations around the fixation, and an excess of high frequency derived mutation and linkage disequilibrium in flanking regions owing to recombination events (Figure 1.3). These changes are captured in the site frequency spectrum and may be detected in polymorphism for approximately 0.1 4N generations (where N is the effective population size) before becoming obscured by subsequent mutation and recombination events (Przeworski 2002) – or approximately 250,000 years for humans.

Mechanisms of Human Evolution

Positive selection can leave many different signatures in the human genome depending on the targets it acts upon – and there are many different models of selection. Selection can act on a single new beneficial mutation (discussed above), also known as a "hard" or "classic" sweep. Another alternative is that selection can act on multiple copies

of a beneficial mutation or standing variation (Orr and Betancourt 2001, Hermisson and Pennings 2005). This is referred to as a "soft" sweep since the beneficial mutations are present at some intermediate frequency before they begin sweeping. There are also models for incomplete sweeps—a classic sweep that has not reached fixation—which may be detectable with haplotype patterns (Kim and Nielsen 2004, Sabeti et al. 2007). Selection can also act on polygenic (Turelli and Barton 1990, Pritchard and Pickrell 2010) or epigenetic traits (Jablonka and Lamb 1998, Feinberg and Irizarry 2010). The patterns produced under all of these models differ depending on the timing, strength, and rate of selection. This all can be very confusing when attempting to scan the genome for evidence of adaptations, and contributes to the lack of concrete examples of selection at the genomic level.

The classic sweep is, perhaps, the most commonly looked for signature of selection in humans. Numerous statistics have been developed that utilize different aspects of the classic sweep pattern to try and find evidence of selection in the human genome (for review, see Crisci et al. 2012). But the results of these scans offer minimal overlap of selective targets. This is further complicated by the fact that expected sweep patterns are difficult to distinguish from background selection, i.e. the continuous removal of neutral mutations though linkage with deleterious haplotypes. This process creates a reduced level of neutral variation as is seen with a sweep (Figure 1.3). Indeed, even though coding regions genome-wide show this pattern, it is unclear whether selective sweeps are responsible, or background selection. For example, Cai et al. (2009) show that the level of neutral polymorphism in the human genome is negatively

correlated with both functional constraint and divergence from chimpanzee – pointing out that this would be consistent with either recurrent selective sweeps or background selection. Hernandez et al. (2011) find a similar negative correlation between polymorphism and functionally conserved regions, and further add that the average reduction in diversity around human amino acid substitutions is no different from reduced diversity at synonymous substitutions, suggesting that classic sweeps could not be the cause of these amino acid substitutions.

Consider also the wait time for a beneficial mutation to occur. In order for a new beneficial mutation to fix in a population via the classic sweep model, the mutation must overcome being lost by genetic drift, and reach a high enough initial frequency for selection to act on it (Kimura 1983). Thus, the waiting time for a new beneficial mutation to arise could be very long. If the primary driver of selection in humans were environmental change, selection on standing variation would allow for adaptations to fix more readily, alleviating the issue of wait time. Since multiple haplotypes are brought to fixation under both soft and standing models, this mechanism of evolution leaves a different genomic signature than classic sweeps – increasing intermediate frequency mutations and creating distinctive haplotype blocks (Przeworski et al. 2005, Pennings and Hermisson 2006).

There is also recent and intriguing evidence that selection can shape epigenetic interactions, although the details have yet to be well resolved. For example, *PRDM9* is a zinc finger protein that influences where recombination hotspots occur during meiosis (Baudat et al. 2010). The location of these hotpots differs widely between humans and

other species, and the binding domain of *PRDM9* is diverse across humans, possibly owing to a selection mechanism (for review, see Ségurel et al. 2011). Another example is the recent discovery of species-specific methylation patterns in sperm cells between humans and chimpanzees (Molaro et al. 2011). There are also brain-specific epigenetic patterns of H3K4met3 between human, chimpanzee, and macaque, which suggests that changes in gene expression play a role in the evolution of the human brain (Shulha, Crisci, Reshetov, Tushir et al. 2012). While appealing as a potential mode of rapid adaptation in natural populations, the details of epigenetic inheritance and modeling remains as a field in need of further study, though progress is beginning to be made (Geoghegan and Spencer 2011).

The Future of Human Evolution

The role of selection on genetic variation in humans has been reconciled with many different models of selection—ranging from completely neutral (Kimura 1968, 1983) to weakly deleterious (Ohta 1973) to weakly advantageous (Gillespie 1977). Another problem often ignored is the confounding effect that demography has when estimating selection (Thornton et al. 2007). Human populations violate the equilibrium assumptions underlying most tests of selection, being a non-randomly mating population that has experienced past bottlenecks and growth, as well as subdivision and migration. All of these neutral process shape the frequency spectrum (Figure 1.3), and it is essential that next generation modeling and method development be focused around jointly estimating selection and demography, rather than simply one or the other.

The Role of My Dissertation in Advancing the Study of Selection in Humans

Given the quality and quantity of human genomic data available, it is now possible to reevaluate selection estimators on a standardized set of non-equilibrium models to see how well they perform, especially considering that human populations violate many assumptions of the equilibrium neutral model upon which these statistics were founded. In fact, multiple genomic scans for selection have been performed on the human genome to date, all using different statistical estimators, and all identifying different lists of putative targets with minimal overlap between the different methods (Akey 2009). This is likely due to three reasons. First, some of these tests were performed before publically available standardized genomic polymorphism datasets for the human genome existed (e.g. HapMap (Sabeti et al. 2007), and 1000 Genomes (1000 Genomes Project Consortium, 2010)). Thus, the difference between ascertainment methods for SNPs can lead to widely different patterns when considering the site frequency spectrum (Nielsen and Signorovitch 2003). Second, these genomic scans do not deal with the impact of demographic forces on genomic variation in the same way. This affects the power of these methods to correctly identify selection and causes different regions to be highlighted. Lastly, many of these methods are outlier-based approaches, meaning they calculate a statistic across the genome and then consider the extreme values to be signatures of selective sweeps. This can lead to an overestimation of the effects of selection, since according to this practice even a non-equilibrium neutral dataset will have outliers.

I propose to fill these gaps in scientific knowledge in three ways. First I will examine the Neanderthal genome and its application in detecting more ancient sweeps along the human branch, as performed by Green et al. (2010). The rationale behind their selective sweep scan is that polymorphic sites present at high frequencies in human populations, but which are derived with respect to Neanderthal, are characteristic of sweeps that rose to fixation shortly after the human-Neanderthal split. While this is indeed an intriguing time point for revealing loci that may have uniquely contributed to modern human genetics, it is essential to extend this analysis using additional selection estimators in line with both older and more recent time scales, in order to truly identify the utility of ancient hominin genomes in human evolution.

On a related note, I consider a second novel dataset that was created exclusively for investigating evolution of the human brain. By sequencing H3K4met3 peaks in neuronal cells across 3 species of primates, including human, this dataset can elucidate epigenetic changes that have occurred specific to human neuronal cells (Shulha, Crisci, Reshetov, Tashir et al. 2012). This particular dataset provides the first opportunity to see if epigenetic changes in humans are correlated with genetic signatures of positive selection. To find out if this the case, I examine a set of human-specific H3K4met3 peaks for traditional selective sweep signatures, and also for increased DNA substitution rates along the human branch. These two methods will uncover whether any of these peak regions have been subject to recent selective sweeps or recurrent nucleotide evolution, which will help answer the extent to which positive selection influences changes in gene expression levels.

Lastly, I think it is necessary to verify the effectiveness of traditional selection estimators, now that whole genome data is becoming widely available for many different species. As these different statistics have been developed over the years, they often were not tested properly for type-I and type-II error, and if they were the analysis did not necessarily extend to models that violate equilibrium assumptions. This last point is key, since most natural populations violate these assumptions. For example, human populations have experienced several bottlenecks in the recent past as we migrated out of Africa to populate the globe (Gutenkunst et al. 2009). This is problematic for estimating selection, since bottlenecks can have a similar effect on genetic variation as selective sweeps. Consequently, most estimators reject neutrality in favor of selection when other models exist that could explain the extreme values of these statistics (Jensen et al. 2005).

It is my hope that the work contained in this dissertation can further the field's understanding of how positive selection has shaped the human genome, and clarify the state of current methods used in detecting selection. This will hopefully shape the future development of statistics that can be used without fear of the underlying non-selective forces that inevitably influence genetic variation in all natural populations.

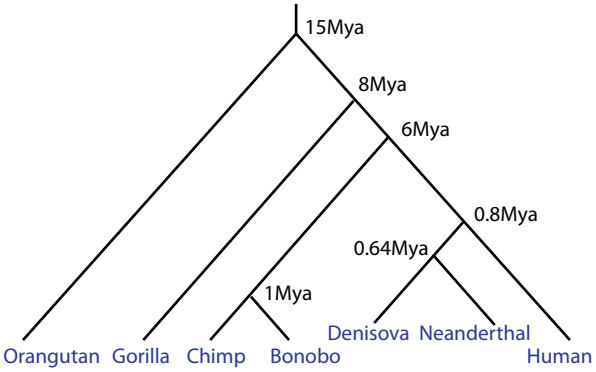


Figure 1.1. Phylogeny of the great apes and approximate divergence times. Branches are not drawn to scale. The Neanderthal and Denisova branches are intentionally truncated to indicate extinct vs. extant.

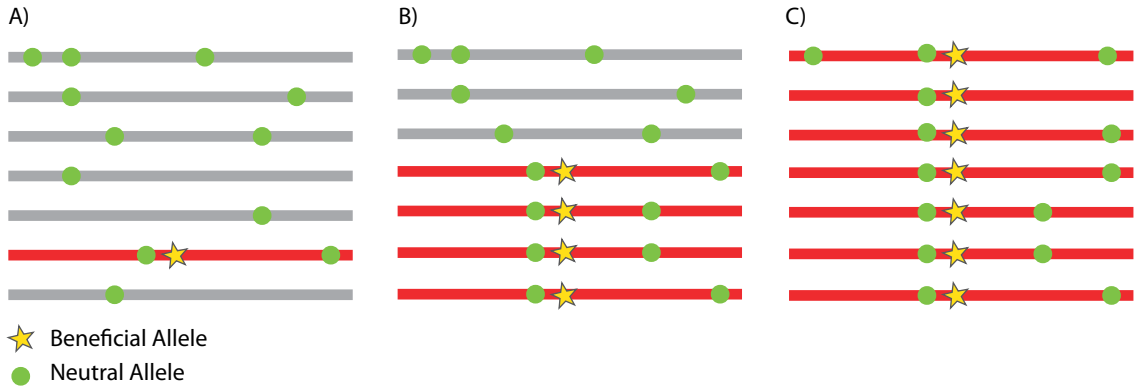


Figure 1.2. The hitchhiking effect.
Each grey or red line represents a chromosome from a single individual. A) A beneficial mutation arises in the population and is closely linked to a neutral allele. B) As the mutation rises in frequency, it brings with it linked neutral alleles. Only alleles that recombine onto the beneficial haplotype are not lost from the sample. C) After the sweep is completed the closely linked allele is fixed. Thus only high frequency alleles that have 'hitchhiked' with the beneficial mutation are visible as variation within the sample, and subsequent new mutations appear as rare variants.

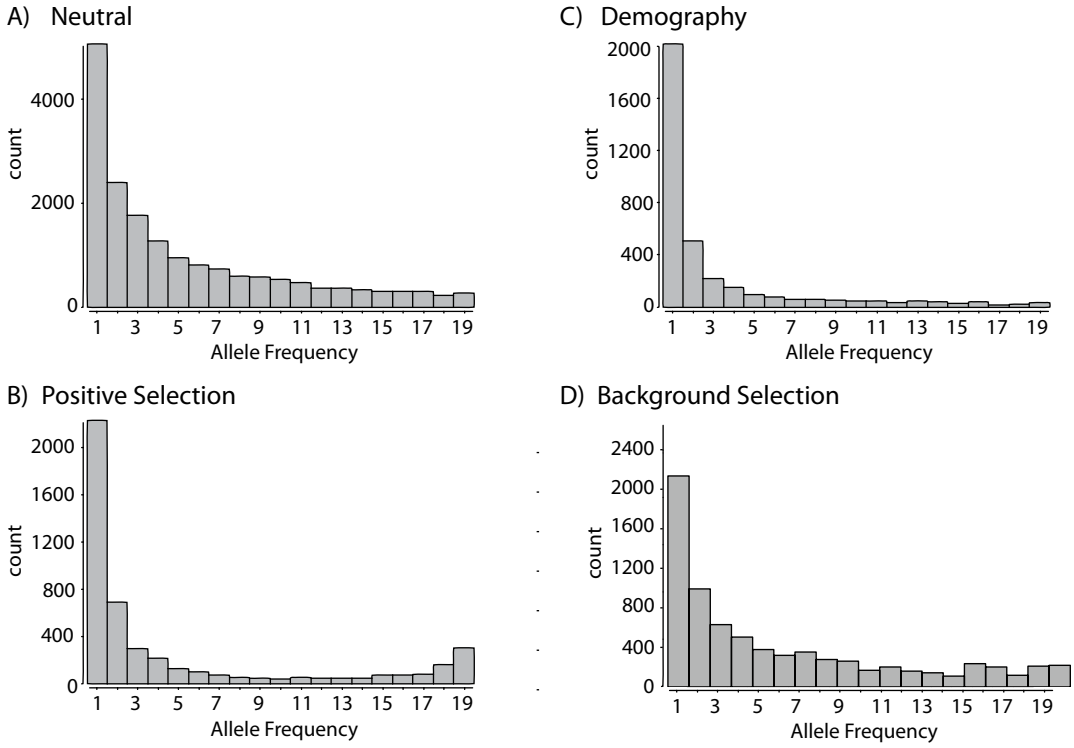


Figure 1.3. A comparison of the site frequency spectrum under equilibrium and nonequilibrium conditions.

Plots are based on simulation of a 5Kb region using either msms (A-C) or sfscode (D) with human-like parameters (effective population size of 10000, per site mutation rate of 2.35×10^{-8}, and per site recombination rate of 2.56×10^{-8}). Counts on the y-axis are the total number of mutations based on 1000 iterations. A) Equilibrium neutral population. B) 1% positive selection at a single locus. C) Nonequilibrium neutral population. Demographic parameters include an 80% reduction in population size 50 Kya, with an exponential growth of 5% for the last 1000 years. D) Background selection, with 80% of sites experiencing 1% negative selection. It is apparent from this figure that both demography and selection greatly reduce the number of mutations compared to neutrality.

CHAPTER II. On Characterizing Adaptive Events Unique to Modern Humans

Abstract

Ever since the first draft of the human genome was completed in 2001 there has been increased interest in identifying genetic changes that are uniquely human, which could account for our distinct morphological and cognitive capabilities with respect to other apes. Recently, draft sequences of two extinct hominin genomes, a Neanderthal and Denisovan, have been released. These two genomes provide a much greater resolution to identify human-specific genetic differences than the chimpanzee, our closest extant relative. The Neanderthal genome paper presented a list of regions putatively targeted by positive selection around the time of the human-Neanderthal split. We here seek to characterize the evolutionary history of these candidate regions - examining evidence for selective sweeps in modern human populations, as well as for accelerated adaptive evolution across apes. Results indicate that 3 of the top 20 candidate regions show evidence of selection in at least one modern human population ($p < 5 \times 10^{5}$). Additionally, 4 genes within the top 20 regions show accelerated amino acid substitutions across multiple apes ($p < 0.01$), suggesting importance across deeper evolutionary time. These results highlight the importance of evaluating evolutionary processes across both recent and ancient evolutionary timescales, and intriguingly suggest a list of candidate genes that may have been uniquely important around the time of the human-Neanderthal split.

Background

The identification of genomic regions that have been affected by positive selection in humans, but not in other primates, is a promising avenue for characterizing the genetic changes underlying phenotypic traits that are unique to humans. With the advent of whole-genome sequencing technology, a number of primate genomes have recently become available for such comparisons (*e.g.,* chimpanzee, The Chimpanzee Sequencing and Analysis Consortium 2005; macaque, Rhesus Macaque Genome Sequencing and Analysis Consortium 2007; orangutan, Locke et al. 2011; and gorilla, Scally et al. 2012). Additionally, two extinct hominin genomes have recently been sequenced: the Neanderthal (Green et al. 2010) and a newly discovered archaic hominin from Denisova Cave in Siberia (Reich et al. 2010). Genomic information from these extinct hominin individuals provides a unique opportunity to identify genetic changes that occurred in the evolution of modern humans (see Figure 2.1).

Green et al. produced a list of putatively swept regions in humans by aligning the human, chimpanzee, and Neanderthal genomes. They looked for spans of the genome with sites polymorphic in five modern human populations, where Neanderthal carried the ancestral allele with respect to chimpanzee. The expected number of Neanderthal derived alleles was calculated and compared to the observed number - producing a measure, *S*, which was used to quantify the absence of Neanderthal derived sites within a given region (with larger *S* corresponding to a higher confidence of a human-specific selective sweep). Because the expected number of Neanderthal derived alleles is conditioned on

the genomic average of each configuration of observed human alleles at polymorphic sites, this approach has unique power to detect older selective sweeps along the human branch. Importantly, this allows detection at time scales for which standard frequency spectrum based tests lack power (Green et al. 2010, SOM). Additionally, because the window size of variation affected by a sweep is related to s/r (the strength of selection over the recombination rate; Kaplan et al. 1989) and the transition time for a beneficial mutation is $-\log(1/2N_e)/s$ generations, they were most likely to find regions that had been affected by strong selection (*i.e.*, having fixed since the human-Neanderthal split, $\sim s > 0.001$).

In contrast, traditional genomic scans for positive selection rely on the hitchhiking pattern evident in linked neutral variation (Maynard Smith and Haigh 1974), and are limited to detecting adaptive fixations having occurred within $\sim 0.2\ 2N_e$ generations (Kim and Stephan 2002). Divergence-based methods, on the other hand, rely not on patterns in polymorphism but rather on detecting increased rates of amino acid substitution between lineages, and thus are appropriate to study recurrent selection across multiple species (*i.e.,* on a much longer evolutionary time scale) - requiring multiple beneficial fixations in order to have power.

Thus, the Green et al. approach is unique in that the timescale over which it may identify positive selection is in between purely divergence- or polymorphism-based approaches (Figure 2.1), and they provide a first glance at regions that may set humans apart from our closest evolutionary relatives. Using this method, they identified a total of

212 genomic regions, representing the top 5% of loci with signals of putative sweeps dating around the human-Neanderthal split.

As indicated by Figure 2.1, these candidate adaptive regions may be further characterized into four general categories of positive selection. They may be: 1) accelerated across apes, 2) accelerated in modern humans, 3) accelerated in the common ancestor of humans and Neanderthals, or 4) uniquely important around the time of the human-Neanderthal split. Our objective is to characterize these regions across both broad and narrow evolutionary time in order to reveal which regions may in fact have been uniquely important around the human-Neanderthal split, and to discover the extent of overlap between their method and traditional site frequency spectrum and d_N/d_S methods for detecting positive selection. We ask the question: given a list of regions that in theory represent ancient sweeps along the human linage, how many could have been detected without the use of the Neanderthal genome?

In order to distinguish among the possible alternatives we utilize two additional classes of methodology: 1) the codeml sites model and branch model (Yang 1998, Yang et al. 2000) from the software package PAML, which identifies genes that show accelerated amino acid substitution across multiple species (Yang 2007), and within a single branch, respectively, using measures of d_N/d_S and 2) SweepFinder (Nielsen et al. 2005), which identifies genetic regions that show evidence of a recent beneficial fixation within a single population using polymorphism data. This direction is similar in principle to the recent work of Cai et al. (2009) who demonstrated a relationship between high d_N and levels of polymorphism, which they interpret as evidence of recurrent positive

selection. While we are similarly comparing across multiple time-scales, our starting dataset is composed of those genes recently suggested to be important around the human-Neanderthal split (*i.e.*, as opposed to high d_N across the tree), and thus results are not directly comparable.

Our findings indicate that many of these regions would not have been detected as candidates for positive selection using traditional frequency spectrum or divergence-based approaches, and that the Neanderthal genome has indeed allowed for the identification of regions experiencing positive selection over a unique time period of the human lineage. By focusing exclusively on the putatively selected regions of the Green et al. study, we additionally parse this gene set in to those most likely to have been important in differentiating human and Neanderthal.

Evidence for selection across apes

A common approach for detecting positive selection across multiple species is to compare the ratio of the rate of non-synonymous substitutions (mutations that lead to amino acid changes; d_N) to the rate of synonymous substitutions (silent mutations; d_S), with d_N/d_S =1, <1 and >1 being consistent with neutral, purifying and positive selection, respectively. In early applications, d_N/d_S was averaged over all sites within a protein sequence and across the entire evolutionary time scale of all lineages. This application has little power to detect positive selection, because it is likely that most sites are functionally constrained ($d_N/d_S << 1$) and are primarily shaped by purifying selection. For our analysis we utilize codeml, which has a sites model allowing d_N/d_S (ω) to vary at each

site along a sequence (Yang et al. 2000). This method is still conservative in that it averages d_N and d_S over lineages at each site, but it has improved power to detect site specific positive selection in a functional protein sequence (Wong et al. 2004).

Tests of positive selection in the codeml sites model compare the fit of the data under a neutral model, to that under a model of positive selection via a likelihood ratio test. For the following analysis, three model comparisons were considered: M1a vs. M2a, M7 vs. M8, and M8a vs. M8. M1a has two subsets of sites, one where ω varies between 0 and 1 and one where ω is fixed at one; in M2a ω can be less than 1, equal to 1, or greater than 1 (Wong et al. 2004). M7 assumes a beta-distribution for ω between 0 and 1, and M8 adds an additional class of sites to M7 with ω>1 (Wong et al. 2004). In M8a this additional class is fixed at ω=1 (Swanson et al. 2003). Thus, M2a and M8 allow selection in each comparison, while M2, M7, and M8a fit the data to a neutral model. A maximum likelihood ratio is computed for each model, and the null and selection models are compared via a likelihood ratio comparison.

For our analysis we focused on the top 20 largest putative sweep regions from Green et al. (2010) and the 51 genes contained within them (Table 2.1). Orthologues were obtained in five primate species: macaque, chimpanzee, orangutan, human, and gorilla. Of the original 51 genes, 8 were noncoding RNA (MIR genes and MEG3), and thus not suitable for codeml analysis. Of the remaining 43 genes, 29 had annotated 1:1 orthologues in the above primate species in Ensembl. We did not use genes from species with more than one annotated orthologue. Multiple species alignments were constructed using the PRANK alignment algorithm (Löytynoja and Goldman 2005) and tested using

the three codeml model comparisons described above. Results are summarized in Table 2.2. Two of the 29 genes showed significant positive selection under all three comparisons: *CCDC82* and *RFX5*. Additionally, *CGN* showed significant positive selection under M1a vs. M2a and M8a vs. M8, and *THADA* was significant under M8a vs. M8. We have included this last gene in further discussions since this model comparison is the most realistic (Swanson et al. 2003).

Two of these genes are involved in human disease/immunity. *THADA*, which has been shown to be involved in beta-cell function (Simonis-Bik et al. 2010), is located close to a potential susceptibility locus of type II diabetes (Zeggini et al. 2008), and a SNP within *THADA* has been shown to be associated with type II diabetes (Schleinitz et al. 2010). *RFX5* is involved in MHC-II expression through interferon gamma (Xu et al. 2003; Garvie et al. 2008). Genes involved in immunity are among the most highly represented in scans for positive selection (Yang 2005), with several studies finding significant evidence for positive selection within the antigen recognition site of MHC-I (Hughes and Nei 1988; Yang et al. 2002) and MHC-II (Hughes and Nei 1989). The other two genes, *CCDC82*, and *CGN*, are not as well characterized, and any inference about their evolutionary significance would be purely speculative.

The codeml sites model also makes predictions regarding the most likely sites experiencing positive selection according to a Bayes empirical Bayes method (Yang et al. 2005). For each codon in a DNA sequence that is analyzed, the probability that $\omega > 1$ at that particular site is computed. A probability of greater than 0.95 was used to determine a site that showed significant positive selection. Of the four genes that were significant

for at least two tests of selection under the sites model, two such sites were identified in *CCDC82, CGN,* and *THADA*; four sites were identified in *RFX5* (Figure 2.2). In all cases, sites display accelerated rates of evolution across the species tree, but do not contain human-specific changes.

Additionally, we performed two branch tests in codeml, which specifically test for higher than expected d_N/d_S along a single branch of interest. For this analysis we tested the human branch and the branch ancestral to humans, Neanderthals, and Denisovans. This is achieved, again, by a likelihood ratio comparison between two models where a d_N/d_S ratio is assigned to each branch in the tree. Each of the models allows for two values for d_N/d_S: one for the foreground branch where positive selection is assumed (ω_1), and one for the rest of the background branches (ω_0). In the null model, ω_1 is fixed to equal 1 on the foreground branch, while ω_0 is estimated on the remaining branches. In the alternative model, ω_1 is also estimated from the data.

We found that none of the previous 29 species alignments showed significant positive selection along either the human branch or the branch ancestral to hominins *($p <$* 0.01). However, five genes did reject the null model in favor of the alternative on both branches ($p < 0.01$: *CADPS2, DYRK1A, BACH2, INPPL1, and ZFP36L2*) though $\omega_1 < 1$.

Evidence for selection in modern human populations

To detect recent selective sweeps in human populations we used ascertainment-corrected polymorphism data from Perlegen, in African-American, European-American, and Chinese populations (Williamson et al. 2007). The program SweepFinder (Nielsen et

al. 2005) was used to scan for sweeps, given the relatively large size of the genomic regions under consideration. SweepFinder computes the background site frequency spectrum (SFS) for the region in question and then identifies unusual regions relative to this background (Figure 3). A significant cutoff value is determined using neutral simulation (see Methods).

Of the top 20 putative sweep regions from Green et al., three were identified as being consistent with recent selection in modern humans (Figure 2.3). Sweep region 1 is upstream of *ZFP36L2* on chromosome 2 in the European population (Figure 2.3a). Sweep region 2 is centered around an intron of *KCNAB1* on chromosome 3 in the African population (Figure 2.3b). Finally, sweep region 3 is localized near the last exon of *DLK1* on chromosome 14 in the Chinese population (Figure 2.3c). These sweeps are distinct from those detected in the original dataset for at least two reasons. First, our sweep analysis was performed using population specific data, and thus any selective signal will be unique to a single population; whereas the Green et al. scan was based upon detecting a joint signal from all five populations considered. Second, because of the time restrictions over which a recent sweep can be detected (~100,000 years for Africans), the time scales of the two statistics are essentially non-overlapping. This scaling becomes even faster for populations of smaller effective population sizes *(i.e.,* $N_{e(Chinese)} = 510$, $N_{e(Europe)} = 1000$ (Gutenkunst et al. 2009)); thus the time to the oldest detectable sweep ~ 5100 years and ~10,000 years for the Chinese and European populations, respectively. Therefore, these results suggest recurrent selective sweeps along the human-lineage in

these regions (*i.e.,* around the human-Neanderthal split, and in modern human populations).

In an attempt to localize potential genetic targets of these peak regions, the UCSC genome browser (track SNP 130) and dbSNP were used to identify SNPs specific to the populations under consideration. Since the peak regions in chromosome 2 and 14 were less than 1Kb, an additional 2Kb of human sequence was examined on either side of the peak. One high frequency derived SNP (rs10132598) was identified in the Asian population near the significant peak of chromosome 14 (CHB+JPT= 0.83, YRI= 0.30, and CEU=0.15) according to the 1000 genomes pilot data, phase 1 (Durbin et al. 2010). This agrees well with the SweepFinder result, as the significant peak using the Perlegen dataset was specific to the Chinese population. Another SNP (rs72875566) was found near the significant peak region of chromosome 2. The significant sweep was detected in the European population, and interestingly, this SNP is at a higher frequency in individuals of European ancestry compared to Yorubans (0.85 vs. 0.61, respectively) according to the phase 1 low coverage data from the 1000 genomes project. No information on this SNP was provided for the Asian populations. This SNP is also located in a CpG island upstream of both ZFP36L2 and another predicted mRNA locus (LOC100129726, Figure 2.3) that was not in the original table in Green et al. These two genes transcribe in opposite directions and the CpG island overlaps both genes, suggesting that it may affect expression of either locus.

Discussion

By examining the candidate selection genes of Green et al. using both divergence and polymorphism data, we have parsed the list of candidate regions that may have been uniquely important in differentiating human and Neanderthal, providing an ideal list for functional validation. The extent of overlap between codeml, SweepFinder, and Green et al. is summarized in Table 2.1. Of the 20 original regions, 15 would not have been identified using the methods tested above (Table 2.1, red text). This highlights the utility of the Neanderthal genome - demonstrating power to identify regions that would have been missed by using site frequency spectrum- or d_N/d_S-based methodology alone.

The genetic functions contained within some of these novel regions are of interest in terms of human evolution. The *HoxD* gene cluster located on chromosome 2 is involved in both vertebral and limb development (for review, see Favier and Dollé 1997). Another interesting gene is *RUNX2* (*CBFA1*). This is a transcription factor involved in bone development. Mutations in *RUNX2* can lead to a skeletal disorder known as Cleidocranial dysplasia (CCD), which is characterized by short stature, underdeveloped or missing clavicles, and dental and cranial abnormalities, among other skeletal changes (Mundlos et al. 1997). Thus selection within these regions could have led to morphological differences in modern humans.

Also of note are *DYRK1A*, *NRG3*, and *CADPS2*. *DYRK1A* is located in the Down Syndrome Critical Region on chromosome 21. It is expressed during brain development, and also in the adult brain, where it is believed to be involved in learning and memory (Hämmerle et al. 2003). *NRG3* also has neurological implications. In humans, it is

expressed in the hippocampus, amygdala, and thalamus, and is believed to be a susceptibility locus for schizophrenia (Zhang et al. 1997, Wang et al. 2008). Mutations in *CADPS2* have been associated with autism (Sadakata and Furuichi 2010). Selection in these three regions during human evolution could have resulted in characteristic cognitive behavior.

The availability of extinct hominin genomic sequence, such as Neanderthal and Denisova, is an important milestone in the study of human evolution. These genomes provide much greater resolution for the identification of unique human adaptive substitutions, since they serve as a nearer outgroup than chimpanzee (Figure 2.1). Any human substitutions identified using chimpanzee may be shared among the many ancestors between human and chimpanzee, including *Australopithecus* and *Paranthropus*, whereas Neanderthal and Denisova are the two nearest known relatives of *Homo sapiens*. These two genomes also can provide a more detailed adaptive history of the human species, and, in combination with the selective scan method of Green et al., we now have power to detect adaptive fixations in deeper evolutionary time. Our results show that this method can, in fact, detect adaptive genomic regions that would have been missed using selective scans based on d_N/d_S (*i.e.*, codeml) or site frequency spectrum summary statistics (*i.e.,* SweepFinder). In their analysis, Green et al. compared their regions to two other genomic scans for selection in humans, one using an outlier approach and the other based on Tajima's *D* (Tajima 1989). They found no significant overlap between their regions and those of other studies, further suggesting power over

separate time frames. There is also no overlap between the 20 regions we examined here and the SweepFinder scan performed by Williamson et al. (2007).

It is not unexpected that the majority of genes we examined within these 20 candidate regions do not contain significant d_N/d_S. The codeml sites model requires that there be excessive d_N across all species at a particular site in order to infer positive selection, and the human branch is short relative to other apes. Thus, the non-synonymous changes are more likely to pre-date humans. Additionally, the codeml branch model averages d_N/d_S across an entire sequence, and this leads to reduced power to detect selection, as discussed above. Moreover, Green et al. identified 78 fixed non-synonymous amino acid changes in humans that were ancestral in Neanderthal, and none of the genes containing these fixed changes overlapped with the genes in the top 20 candidate regions for a selective sweep. It may well be that the target of these sweeps was not non-synonymous (*e.g.*, a synonymous or non-coding change, or that a non-synonymous change in humans was unable to be determined due to the variable depth in sequence coverage of the Neanderthal genome). In fact, 5 of the 20 candidate regions contain no annotated coding sequence (Table 2.1), and Green et al. found an additional 232 human-specific substitutions in 5'- and 3'- UTR regions, suggesting that non-coding sites may have been targeted.

Conclusion

Here we have shown that using an ancient hominin genomic sequence to scan for positive selection in humans (as performed by Green et al.) has elucidated a novel list of candidate selection regions that would not have been discovered using currently available

methods of detecting selection. Of the 15 novel regions from the Green et al. scan, 5 contained genes with interesting relations to human morphological and cognitive traits. Therefore, we conclude that using an ancient hominin genome to scan for selection in conjunction with already established methods could offer a more complete picture of how positive selection has shaped modern humans.

Methods

Multiple species alignment for codeml

Human mRNA sequences were obtained from Ensembl. Only sequences with CCDS citations were used. If there was more than one transcript, the one with the longest amino acid sequence was chosen. Macaque, chimp, gorilla, and orangutan sequences were retrieved from Ensembl using BioMart. Briefly, using the list of human gene IDs, orthologous Ensembl gene IDs for each species were obtained from the Ensembl Genes 58 human dataset using the homologs filter under Multi-species Comparisons. These IDs were then queried to get orthologous coding transcript sequences from each species using the sequences attribute. In cases where more than one transcript variant was returned, the longest was chosen. Only genes showing 1:1 homology with orthologues in all five species were used for codeml analysis. Sequences were aligned using PRANK (Löytynoja and Goldman 2005). The codon option was used, which uses the empirical codon model (ECM; Kosiol et al. 2007) to align individual codons while preserving the reading frame. The guide tree was estimated by the program and all other parameters were left as default. This method of alignment was shown by Fletcher and Yang (2010) to

be the most accurate at preserving true sequence alignment in the presence of insertions and deletion when using the PAML branch-site test.

codeml analysis

The codeml program in PAML version 4.4 (Yang 2007) was used to test for positive selection across apes (with the exception of macaque, which was included even though it is an Old World Monkey). Three different sites model tests were examined: M1a vs. M2a, M7 vs. M8, and M8 vs. M8a (see PAML documentation for parameters). A likelihood ratio test was used to determine significance. A Bonferroni corrected *p* value assuming 29 tests (0.05/29) is equal to 0.0018. We also compare with the uncorrected *p* value of 0.01 to determine significance. For both the sites and human-specific branch tests, an alignment of 5 primate species is used (human, chimpanzee, gorilla, orangutan, macaque). For the human-Neanderthal ancestral branch test an alignment of 7 species was used that included the Neanderthal and Denisovan sequences. These two sequences were excluded from sites test due to the variable coverage of both genomes, as codeml ignores sites with missing data.

Neanderthal and Denisova sequence construction

The BAM files for Neanderthal and Denisova can be found at: ftp://ftp.ebi.ac.uk/pub/databases/ensembl/neandertal and http://hgdownload.cse.ucsc.edu/downloads.html, respectively. SAMtools (Li et al. 2009) was used to retrieve the reads corresponding to each gene sequence from the Neanderthal

and Denisova BAM files using the chromosomal locations. These reads were mapped back to hg18 using Geneious version 5.3.2 (Drummond et al. 2011). A Phred scaled confidence score cutoff of 30 was applied for all sites where these sequences differed from hg18.

SweepFinder Analysis

The data used for this analysis was the same Perlegen SNP dataset as in Williamson et al 2007. The SNPs for each region were analyzed using SweepFinder (Nielsen et al. 2005), which computes the background site frequency spectrum (SFS) for a region using SNP data. It uses a likelihood framework (Kim and Stephan 2002) to compare the background SFS with that expected under a model of a selective sweep at a predetermined set of sites along the region. The number of sites is designated by the gridsize parameter, and was set to the number of nucleotides in the region. The cut off value was determined by simulating 1000 replicates in ms (Hudson 2002) under the standard neutral model for each region. The parameters for each simulated region consisted of the same SNP density (by setting the "S" parameter in ms equal to the number of SNPs from the Perlegen dataset present in the region) and gridsize as the actual region. For ms style input, SweepFinder returns the maximum LR value for each replicate. To determine significance, the top 99.995% of LR values ($p = 5 \times 10^{-5}$) were considered significant. This p value reflects a Bonferroni correction for 1000 tests.

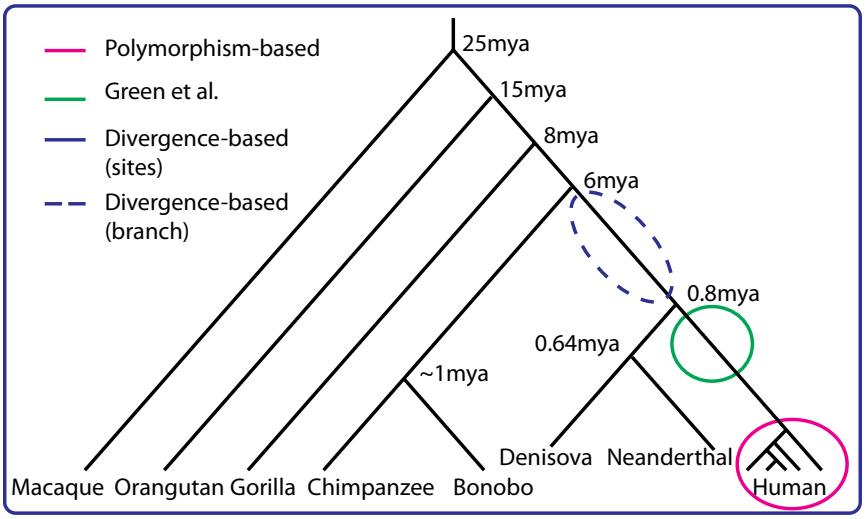


Figure 2.1 Summary of methods.
A graphical representation of the evolutionary timescale over which the methods for detecting positive selection are effective. Branch lengths are not drawn to scale. Divergence-based methods can detect positive selection across a phylogenetic tree or along a single branch; polymorphism-based methods are effective within a single population; the Green et al. method using the Neanderthal genome finds selection in humans that occurred shortly after the human-Neanderthal split.

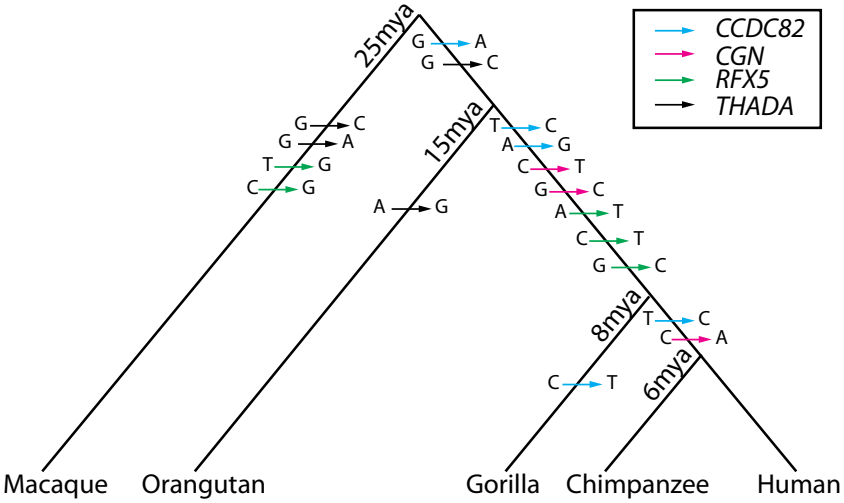


Figure 2.2. Mutations at significant sites across the primate tree.
For genes that showed significant positive selection by at least two tests in the codeml sites model, the nucleotide changes within the candidate sites for selection were mapped. In cases where there were two possible scenarios that could describe how a change originated, the simplest was assumed. Branch lengths are not drawn to scale, and the spacing and ordering of the mapped substitutions on a given branch are arbitrary.

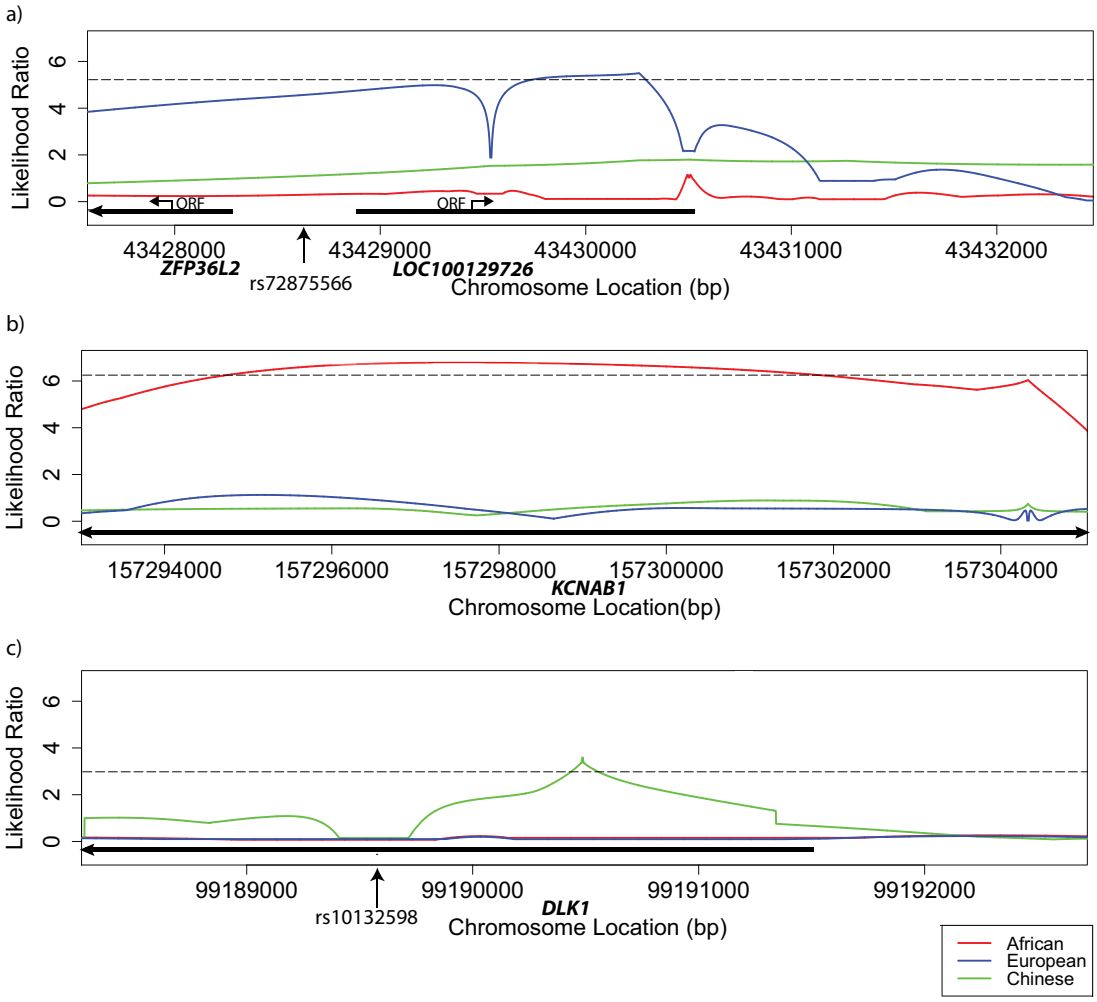


Figure 2.3. Sweep regions.
The three regions identified from the Green et al. dataset as showing evidence of a selective sweep in a modern human population using SweepFinder. The horizontal dashed line represents a Bonferroni corrected LR cutoff ($p < 5x10^{-5}$). Approximate region lengths correspond to the significant portion of the peak. Population-specific high frequency derived SNPs are marked with an arrow along the x-axis. a) A region of 531bp upstream of ZFP36L2 and *LOC100129726* in the European-American population. b) A region of ~11Kb within an intron of KCNAB1 in the African-American population. c) A region of 121bp within an intron of DLK1 in the Chinese population. For these plots, the coordinates for chromosomal location along the x-axis correspond to the hg16 genome annotation.

Table 2.1. Information on genomic regions considered and comparison of results

Region (hg18)	Width (CM)	Genes
chr2:43265008-43601389	0.5726	***ZFP36L2***; ***THADA***; ***LOC100129726***[c]
chr11:95533088-95867597	0.5538	*JRKL*; ***CCDC82***; *MAML2*
chr10:62343313-62655667	0.5167	*RHOBTB1*
chr21:37580123-37789088	0.4977	*DYRK1A*
chr10:83336607-83714543	0.4654	*NRG3*
chr14:100248177-100417724	0.4533	*MIR337; MIR665;* ***DLK1****; RTL1; MIR431; MIR493; MEG3; MIR770*
chr3:157244328-157597592	0.425	***KCNAB1***
chr11:30601000-30992792	0.3951	
chr2:176635412-176978762	0.3481	*HOXD11; HOXD8; EVX2; MTX2; HOXD1; HOXD10; HOXD13; HOXD4; HOXD12; HOXD9; MIR10B; HOXD3*
chr11:71572763-71914957	0.3402	*CLPB; FOLR1; PHOX2A; FOLR2; INPPL1*
chr7:41537742-41838097	0.3129	*INHBA*
chr10:60015775-60262822	0.3129	*BICC1*
chr6:45440283-45705503	0.3112	*RUNX2; SUPT3H*
chr1:149553200-149878507	0.3047	*SELENBP1; POGZ; MIR554;* ***RFX5****; SNX27;* ***CGN****; TUFT1; PI4KB; PSMB4*
chr7:121763417-122282663	0.2855	*RNF148; RNF133; CADPS2*
chr7:93597127-93823574	0.2769	
chr16:62369107-62675247	0.2728	
chr14:48931401-49095338	0.2582	
chr6:90762790-90903925	0.2502	*BACH2*
chr10:9650088-9786954	0.2475	

[a]The significant results using each method are either colored green (overlap between Green *et al.* and SweepFinder) or blue (overlap between Green *et al.* and codeml). Regions colored in red contain no overlap with the tested methods, and represent a novel list of genes unique to the Green *et al.* scan using Neanderthal.

[b]For codeml, genes that were significant for at least two tests of selection are underlined ($p<0.01$).

[c] *LOC100129726* was not listed in Green *et al.* Table 3.

Table 2.2. Summary of codeml results

Genes	$2\Delta\ell_{(M1a-M2a)}$	$2\Delta\ell_{(M7-M8)}$	$2\Delta\ell_{(M8a-M8)}$	ω/(Prω>1)[b]	p_{sites} ω>1[c]
BACH2	0.00	0.00	0.00		
BICC1	4.78	5.31	4.78		
CADPS2	2.28	2.50	2.28		
CCDC82	8.34**	8.35**	8.34**	5.781/0.980	0.130
CGN	6.55**	6.55	6.55**	4.186/0.952	0.376
CLPB	0.50	0.61	0.50		
DLK1	1.85	2.57	1.83		
DRYK1A	0.00	-0.18	-0.18		
EVX2	2.19	2.51	2.17		
FOLR1	0.00	0.00	0.00		
HOXD1	4.43	4.81	4.41		
HOXD4	0.20	0.47	0.20		
HOXD8	3.00	3.00	3.00		
HOXD9	0.00	0.00	0.00		
HOXD10	0.00	0.00	0.00		
INHBA	0.00	-0.32	-0.32		
INPPL1	1.77	1.94	1.76		
KCNAB1	4.72	8.63**	4.29	2.121/0.934	0.003
MAML2	0.03	0.13	0.03		
NRG3	0.00	0.00	0.00		
PHOX2A	0.00	0.00	0.00		
PI4KB	-6.26	-0.32	-1.98		
PSMB4	0.63	0.53	0.51		
RFX5	13.03**	13.05**	13.03**	7.898/0.993	0.050
SNX27	0.00	0.00	0.00		
SUPT3H	1.70	2.03	1.70		
THADA	6.35**	7.11	6.35**	3.720/0.965	0.108
TUFT1	0.14	0.19	0.14		
ZFP36L2	0.05	0.41	0.05		

** $p<0.01$

[a]Significance for each test was determined from a chi-squared distribution with degrees of freedom =1 for M8a vs. M8, and df=2 for M1a vs. M2a and M7 vs. M8.

[b]The probability that ω is greater than one at a given site in the sequence based on the BEB posterior probability For each gene showing evidence of positive selection. The highest probability observed is given with its corresponding ω value.

[c]The proportion of sites examined per sequence that fall in the category of ω being greater than one.

CHAPTER III. Human-Specific Histone Methylation Signatures at Transcription Start Sites in Prefrontal Neurons

Abstract

Cognitive abilities and disorders unique to humans are thought to result from adaptively driven changes in brain transcriptomes, but little is known about the role of cis-regulatory changes affecting transcription start sites (TSS). Here, we mapped in human, chimpanzee, and macaque prefrontal cortex the genome-wide distribution of histone H3 trimethylated at lysine 4 (H3K4me3), an epigenetic mark sharply regulated at TSS, and identified 471 sequences with human-specific enrichment or depletion. Among these were 33 loci selectively methylated in neuronal but not non-neuronal chromatin from children and adults, including TSS at *DPP10* (2q14.1), *CNTN4* and *CHL1* (3p26.3), and other neuropsychiatric susceptibility genes. Regulatory sequences at *DPP10* and additional loci carried a strong footprint of hominid adaptation, including elevated nucleotide substitution rates and regulatory motifs absent in other primates (including archaic hominins), with evidence for selective pressures during more recent evolution and adaptive fixations in modern populations. Chromosome conformation capture at two neurodevelopmental disease loci, 2q14.1 and 16p11.2, revealed higher order chromatin structures resulting in physical contact of multiple human-specific H3K4me3 peaks spaced 0.5–1 Mb apart, in conjunction with a novel *cis*-bound antisense RNA linked to Polycomb repressor proteins and downregulated *DPP10* expression. Therefore, coordinated epigenetic regulation via newly derived TSS chromatin could play an

important role in the emergence of human-specific gene expression networks in brain that contribute to cognitive functions and neurological disease susceptibility in modern day humans.

Introduction

Cognitive abilities and psychiatric diseases unique to modern humans could be based on genomic features distinguishing our brain cells, including neurons, from those of other primates. Because protein coding sequences for synaptic and other neuron-specific genes are highly conserved across the primate tree (Bayes et al 2011, King and Wilson 1975), a significant portion of hominid evolution could be due to DNA sequence changes involving regulatory and non-coding regions at the 5′ end of genes (The Chimpanzee Sequencing and Analysis Consortium 2005, McLean et al. 2011). Quantifying these differences, however, is ultimately a daunting task, considering that, for example, the chimpanzee–human genome comparison alone reveals close to 35×10^6 single bp and 5×10^6 multi-bp substitutions and insertion/deletion events (The Chimpanzee Sequencing and Analysis Consortium 2005). While a large majority of these are likely to reflect genetic drift and are deemed "non-consequential" with respect to fitness, the challenge is to identify the small subset of regulatory sequence alterations impacting brain function and behavior.

Here, we combine comparative genomics and population genetics with genome-scale comparisons for histone H3-trimethyl-lysine 4 (H3K4me3), an epigenetic mark sharply regulated at transcription start sites (TSS) and the 5′ end of transcriptional units in brain and other tissues (Zhou et al. 2011, Shilatifard 2006, Cheung et al. 2010, Shulha et al. 2011) that is stably maintained in brain specimens collected postmortem (Cheung et al. 2010, Huang et al. 2006). Our rationale to focus on TSS chromatin was also guided by

the observation that the human brain, and in particular the cerebral cortex, shows distinct changes in gene expression, in comparison to other primates (Preuss et al. 2004). While there is emerging evidence for an important role of small RNAs shaping human-specific brain transcriptomes via posttranscriptional mechanisms (Somel et al. 2011) and increased recruitment of recently evolved genes during early brain development (Zhang et al. 2011), the role of TSS and other cis-regulatory mechanisms remains unclear. Here, we report that cell type-specific epigenome mapping in prefrontal cortex (PFC, a type of higher order cortex closely associated with the evolution of the primate brain) revealed hundreds of sequences with human-specific H3K4me3 enrichment in neuronal chromatin, as compared to two other anthropoid primates, the chimpanzee and the macaque. These included multiple sites carrying a strong footprint of hominid evolution, including accelerated nucleotide substitution rates specifically in the human branch of the primate tree, regulatory motifs absent in non-human primates and archaic hominins including *Homo neanderthalensis* and *H. denisova*, and evidence for adaptive fixations in modern day humans. The findings presented here provide the first insights into human-specific modifications of the neuronal epigenome, including evidence for coordinated epigenetic regulation of sites separated by megabases of interspersed sequence, which points to a significant intersect between evolutionary changes in TSS function, species-specific chromatin landscapes, and epigenetic inheritance.

Results[1]

[1]Supplementary tables and figures for this chapter can be viewed with the original publication at http://www.plosbiology.org/

H3K4me3 Landscapes across Cell Types and Species

The present study focused on the rostral dorsolateral PFC, including cytoarchitectonic Brodmann Area BA10 and the immediately surrounding areas. These brain regions represent a higher association cortex subject to disproportionate morphological expansion during primate evolution (Semendeferi et al. 2001), and are involved in cognitive operations important for informed choice and creativity (Tsujimoto et al. 2010, 2011), among other executive functions. Given that histone methylation in neuronal and non-neuronal chromatin is differentially regulated at thousands of sites genome-wide (Cheung et al. 2010), we avoided chromatin studies in tissue homogenates because glia-to-neuron ratios are 1.4- to 2-fold higher in mature human PFC as compared to chimpanzee and macaque (Sherwood et al. 2006). Instead, we performed cell type-specific epigenome profiling for each of the three primate species, based on NeuN ("neuron nucleus") antigen-based immunotagging and fluorescence-activated sorting, followed by deep sequencing of H3K4me3-tagged neuronal nucleosomes. Prefrontal H3K4me3 epigenomes from NeuN+ nuclei of 11 humans, including seven children and four adults (Cheung et al. 2010), were compared to four chimpanzees and three macaques of mature age (Table S1). Sample-to-sample comparison, based on a subset of highly conserved Refseq TSS with one mismatch maximum/36bp, consistently revealed the highest correlations between neuronal epigenomes from the same species (Table S2). Strikingly, however, the H3K4me3 landscape in human neurons was much more similar to chimpanzee and macaque neurons, when compared to non-neuronal (NeuN−) cells (Cheung et al. 2010) from the same specimen/donor or to blood (Figure

3.1A). Therefore, PFC neuronal epigenomes, including their histone methylation landscapes at TSS, carry a species-specific signature, but show an even larger difference when compared to their surrounding glial and other NeuN− cells.

Several Hundred Loci Show Human-Specific Gain, or Loss, of Histone Methylation in PFC Neurons

To identify loci with human-specific H3K4me3 enrichment in PFC neurons, we screened 34,639 H3K4me3 peaks that were at least 500 bp long and showed a consistent >2-fold H3K4me3 increase for the 11 humans as compared to the average of the seven chimps and macaques and (ii) minimum length of 500 bp. We identified 410 peaks in the human genome (HG19) with significant enrichment compared to the two non-human primate species (with reads also mapped to HG19) after correcting for false discovery (FDR), and we call these peaks "HP" hereafter for "human-specific peaks" (Figure 3.1D; Table S3). We had previously reported that infant and child PFC neurons tend to have stronger peaks at numerous loci, compared to the adult (Cheung et al. 2010). To better age-match the human and non-human primate cohorts, we therefore repeated the analysis with our entire, recently published cohort of nine adult humans without known neurological or psychiatric disease (Cheung et al. 2010, Shulha et al. 2011). Using the same set of filter criteria (>2-fold increase in humans compared to chimpanzees and macaques), we identified 425 peaks and 296 of them overlapped with the original 410 HP (Table S3). Furthermore, 345 of the 410 peaks overlapped with the overlapped with the peaks with >1.5-fold increase for nine adult humans (compared to non-human primates; with correction for FDR) (Table S4), indicating that HPs can be detected reliably.

To obtain human depleted peaks we used a reciprocal approach where initial peaks were detected in chimpanzee and macaque. For the original cohort of 11 children and adult humans, this resulted in 61 peaks with a significant, at least 2-fold depletion in human PFC neurons (Table S5). 50 peaks defined by human-specific depletion in the mixed cohort of 11 children and adults were part of the total of 177 peaks with >1.5-fold decrease in the cohort of nine adults (compared to each of the two non-human primate species; Table S6). From this, we conclude that at least 471 loci in the genome of PFC neurons show robust human-specific changes (gain, 410; loss, 61) in histone methylation across a very wide postnatal age range.

We further explored chimpanzee-specific changes in the H3K4me3 landscape of PFC neurons by comparing human and chimpanzee peaks within the chimpanzee genome. To this end, we constructed a mono-nucleosomal DNA library from chimpanzee PFC to control for input, and mapped the neuronal H3K4me3 datasets from four chimpanzee PFC specimens, and their 11 human counterparts, to the chimpanzee genome (PT2). We identified 551 peaks in the PT2 genome that were subject to >2-fold gain and 337 peaks subject to >2-fold depletion, compared to human regardless of the H3K4me3 level in macaque (Tables S7 and S8). A substantial portion of these PT2-annotated peaks (133 and 40 peaks, respectively) with gain or loss in chimpanzee PFC neurons matched loci with the corresponding, reciprocal changes specific to human PFC neurons in HG19 (410 and 61 peaks as described above). Genetic differences among these genomes and additional, locus-specific differences in nucleosomal organization (leading to differences in background signal in the input libraries) are potential factors that would lead to only

partial matching of peaks when species-specific H3K4me3 signals are mapped within the human, or chimpanzee genome, respectively. These findings, taken together, confirm that genome sequence differences in *cis* are one important factor for the species-specific histone methylation landscapes in PFC neurons.

Human-Specific H3K4me3 Peaks in PFC Neurons Overlap with DNA Methylation Signatures in the Male Germline

Both catalytic and non-catalytic subunits of H3K4 methyltransferase complex are associated with transgenerational epigenetic inheritance in the worm, *Caenorhabditis elegans*, and other simple model organisms (Greer and Shi 2012), and furthermore, H3K4me3 and other epigenetic markings such as DNA cytosine methylation are readily detectable in non-somatic ("germline"-related) cells such as sperm, potentially passing on heritable information to human offspring (Hammond et al. 2009). Therefore, we wanted to explore whether a subset of the 410 loci with at least 2-fold H3K4me3 enrichment in human neurons are subject to species-specific epigenetic regulation in germ tissue. To this end, we screened a human and chimpanzee sperm database on DNA methylation (Molaro et al. 2011), in order to find out which, if any of the 410 sequences with human-specific H3K4me3 gain in brain overlap with a set of >70,000 sequences defined by very low, or non-detectable DNA methylation in human and chimpanzee sperm (termed (DNA) "hypomethylated regions" in Molaro et al. 2011). Of note, the genome-wide distribution of H3K4me3 and DNA cytosine methylation is mutually exclusive in germ and embryonic stem cells, and gains in DNA methylation generally are associated with loss of H3K4me3 in differentiated tissues (Isagawa et al. 2011, Yan et al. 2003).

Unsurprisingly therefore, 300/410 HP peaks in brain matched a DNA hypomethylated sequence in sperm of both species. Strikingly, however, 90/410, or approximately 22% of HP were selectively (DNA) hypomethylated in human but not in chimpanzee sperm (Table S3), a ratio that is approximately 4-fold higher than the expected 5.7% based on 10,000 simulations ($p<0.00001$; see also Text S1) (Figure 3.1B). Conversely, the portion of HP lacking DNA hypomethylation in male germ cells of either species altogether (18/410 or 4%), or with selective hypomethylation in chimpanzee sperm (2/410 or 0.5%), showed a significant, 5-fold underrepresentation in our dataset (Figure 3.1B). Thus, approximately one-quarter of the 410 loci with human-specific gain in histone methylation in PFC neurons also carry species-specific DNA methylation signatures in sperm, with extremely strong bias towards human (DNA) hypomethylated regions (22%) compared to chimpanzee-specific (DNA) hypomethylated regions (0.5%). In striking contrast, fewer than ten of the 61 loci with human-specific H3K4me3 depletion in PFC neurons showed species-specific differences in sperm DNA methylation between species (six human- and three chimpanzee-specific DNA hypomethylated regions; Table S5).

H3K4 Methylation Sites with Human-Specific Gain Physically Interact in Megabase-Scale Higher Order Chromatin Structures and Provide an Additional Layer for Transcriptional Regulation

We noticed that, at numerous chromosomal loci, HP tended to group in pairs or clusters (Table S3). There were more than 245 (163) from the total of 410 HP spaced less than 1 (or 0.5) Mb apart, which is a highly significant, 2- (or 3-) fold enrichment compared to random distribution within the total pool of 34,639 peaks (Figure 3.1C; Text

S1). Therefore, sequences with human-specific gain in H3K4me3 in PFC neurons appear to be co-regulated with neighboring sequences on the same chromosome that are decorated with the same type of histone modification. Likewise, the actual number of human-depleted peaks within one 1 Mb ($n = 6$) was higher than what is expected from random distribution ($n = 2.6$), ($p = 0.051$), albeit no firm conclusions can be drawn due to the smaller sample size ($n = 61$). This type of non-random distribution due to pairing or clustering of the majority of human-enriched sequences broadly resonates with the recently introduced concept of Mb-sized topological domains as a pervasive feature of genome organization, including increased physical interactions of sequences carrying the same set of epigenetic decorations within a domain (Dixon et al. 2012). Of note, H3K4 trimethylation of nucleosomes is linked to the RNA polymerase II transcriptional initiation complex, and sharply increased around TSS and broadly correlated with "open chromatin" and gene expression activity (Zhou et al. 2011, Shilatifard 2006). Therefore, we reasoned that a subset of human-enriched "paired" H3K4me3 peaks could engage in chromatin loopings associated with transcriptional regulation. This is a very plausible hypothesis given that promoters and other regulatory sequences involved in transcriptional regulation are often tethered together in loopings and other higher order chromatin (Schwab et al. 2011, Splinter et al. 2011).

To explore this, we screened a database obtained on chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) for RNA polymerase II, a technique designed to detect chromosomal loopings bound by the Pol II complex. Indeed, we identified at least three interactions that matched to our H3K4me3 peaks with human-

specific gain in PFC neurons (Table S9), including a loop interspersed by approximately 2.5 Mb of sequence in chromosome 16p11.2–12.2. This is a risk locus for microdeletions that are linked to a wide spectrum of neurodevelopmental disease including autism spectrum disorder (ASD), intellectual disability (ID), attention deficit hyperactivity disorder (ADHD), seizures, and schizophrenia (Kumar et al. 2008, Weiss et al. 2008, Shinawi et al. 2009, Bijlsma et al. 2009, Fernandez et al. 2010, McCarthy et al. 2009). We were able to validate this interaction by chromosome conformation capture (3C), a technique for mapping long range physical interactions between chromatin segments (Dekker 2006), in 2/2 human PFC specimens and also in a human embryonic kidney (HEK) cell line (Figure 3.2). We conclude that human-specific H3K4me3 peaks spaced as far apart as 1 Mb are potentially co-regulated and physically interact via chromatin loopings and other higher order chromatin structures.

Neuronal Antisense RNA LOC389023 Originating from a DPP10 (Chromosome 2q14) Higher Order Chromatin Structure Forms a Stem-Loop and Interacts with Transcriptional Repressors

Next, we wanted to explore whether sequences with human-specific gain in histone methylation, including those that show evidence for pairing and physical interactions, could affect the regulation of gene expression specifically in PFC neurons. To this end, we first identified which portion from the total of 410 human-specific peaks showed much higher H3K4me3 levels selectively in PFC neurons, when compared to their surrounding non-neuronal cells in the PFC. Thus, in addition to the aforementioned filter criteria (2-fold increase in human PFC neurons compared to non-human primate

PFC neurons), we searched for peaks with differential regulation among PFC neurons and non-neurons (see Text S1). We found 33 HP with selective enrichment in neuronal PFC chromatin (termed ^{neu}HP in the following) (Figure S1; Table S10). Among these were two HP spaced less than 0.5 Mb apart within the same gene, *DPP10* (chr2q14.1), encoding a dipeptidyl peptidase-related protein regulating potassium channels and neuronal excitability (Figure 3.3A–3.3B) (Maffie 2008). Interestingly, rare structural variants of DPP10 confer strong genetic susceptibility to autism, while some of the gene's more common variants contribute to a significant risk for bipolar disorder, schizophrenia, and asthma (Marshall et al. 2008, Djurovic et al. 2010, Allen et al. 2003). Histone methylation at *DPP10* was highly regulated in species- and cell type-specific manner, with both *DPP10*-1 and *DPP10*-2 peaks defined by a very strong H3K4me3 signal in human PFC neurons (Figure 3.3A), but only weak or non-detectable peaks in their surrounding NeuN− (non-neuronal) nuclei (Figure S1; Table S10) or blood-derived epigenomes (Cheung et al. 2010).

We then employed 3C assays across 1.5 Mb of the *DPP10* (chr2q14.1) in PFC of four humans. To increase the specificity in each 3C PCR assay, we positioned both the forward and reverse primer in the same orientation on the sense strand, and samples processed for 3C while omitting the critical DNA ligation step from the protocol served as negative control (Figure 3.3A–3.3B). Indeed, 3C assays on four of four human PFC specimens demonstrated direct contacts between the *DPP10*-1 and -2 peaks (Figure 3.3A). As expected for neighboring fragments (Dekker 2006), *DPP10*-1 also interacted with portions of the interspersed sequence (CR2 in Figure 3.3A). These interactions were

specific, because several other chromatin segments within the same portion of chr2q14.1 did not show longer range interactions with *DPP10*-1 (CR1, CR3 in Figure 3.3A). We further verified one of the *DPP10-1/2* physical interactions (the sequences captured by primers 6 and 17 in Figure 3.3A) in four of five brains using 3C-qPCR with a TaqMan probe positioned in fragment 6. Furthermore, *DPP10*-2 interacted with a region ("CR3" in Figure 3.3A) 400 kb further downstream positioned in close proximity to a blood-specific H3K4me3 peak. No interactions at the *DPP10* locus were observed in cultured cells derived from the H9 embryonic stem cell line (H9ESC in Figure 3.3A), suggesting that these chromatin architectures are specific for differentiated brain tissue. Of note, similar types of *DPP10* physical interactions were found in 3C assays conducted on PFC tissue of three of three macaques (Figure 3.3B). Because macaque PFC, in comparison to human, shows much weaker H3K4 methylation at these *DPP10* sequences, we conclude that the corresponding chromatin tetherings are not critically dependent on human-specific H3K4me3 dosage.

Next, we wanted to explore whether human-specific H3K4 methylation at the *DPP10* locus is associated with a corresponding change in gene expression at that locus. Notably, H3K4me3 is on a genome-wide scale broadly correlated with transcriptional activity, including negative regulation of RNA expression by generating very short (~50–200 nt) promoter-associated RNAs. These short transcripts originate at sites of H4K4me3-tagged nucleosomes and act as cis-repressors in conjunction with polycomb and other chromatin remodeling complexes (Kanhere et al. 2010, Shi et al. 2006). Therefore, transcriptional activities due to the emergence of novel H3K4me3 markings in

human PFC is likely to be complex, with unique functional implications specific to each genomic locus. To explore the transcriptome at the *DPP10* locus in an unbiased manner, we performed RNA-seq on a separate cohort of three adult human PFC (not part of the aforementioned ChIP-seq studies) and compared their transcriptional landscapes to similar datasets from chimpanzee and macaque (Liu et al. 2011, Brawand et al. 2011). Indeed, we found an antisense RNA, *LOC389023*, emerging from the second DPP10 peak, *DPP10*-2 (chr2q14.1) (Figures 3.3A and 3.4A). In an additional independent analyses (using a set of human postmortem brains different from the ones used for RNAseq) quantitative reverse transcriptase (RT)-PCR assays further validated the much higher expression of *DPP10* antisense transcript in human (Figure 3.4B), which occurred in conjunction with decreased expression of *DPP10* exons downstream of the DPP10-2 promoter (compared to chimp/macaque) (Figure 3.4A).

Consistent with the H3K4me3 enrichment specifically in neuronal chromatin, the cellular expression of *LOC389023* in adult PFC was confined to a subset of the neuronal layers (II–IV), but absent in neuron-poor compartments such as layer I and subcortical white matter (Figure 3.5A and unpublished data). Furthermore, the transcript was expressed in fetal and adult PFC but not in cerebellar cortex (Figure 3.5B). We noticed that *LOC389023* harbored a GC-rich stem loop motif that is known to associate with cis-regulatory mechanisms involved in transcriptional repression, including binding to TSS chromatin and components of *Polycomb 2* (PRC2) complex (Figure 3.5C) (Kanhere et al. 2010, Zhao et al. 2008). Consistent with a possible function inside the nucleus, *LOC389023* was highly enriched in nuclear RNA fractions from extracted prenatal and

normal (non-degenerative) adult human PFC, but not cerebellar cortex (Figure 3.5B). Indeed, in transiently transfected (human) SK-N-MC neuroblastoma cells, *LOC389023* showed a specific association with H3K4-trimethylated nucleosomes and SUZ12 (Figure 3.5D), a zinc finger protein and core component of PRC2 previously shown to interact with stem loop motifs similar to the one shown in Figure 3.5C (Kanhere et al. 2010). In contrast, EZH2, a (H3K27) methyltransferase and catalytic component of PRC-2, did not interact with LOC389023 (Figure 3.5D), consistent with previous reports on other RNA species carrying a similar stem loop motif (Kanhere et al. 2010). These observations, taken together, are entirely consistent with the aforementioned findings that levels of *DPP10* transcript, including exons positioned downstream of the *DPP10*-2 peak from which *LOC389023* originates, are significantly decreased in human PFC as compared to macaque and chimpanzee. Conversely, these two primates show non-detectable (RNAseq) or much lower quantitative RT-PCR (qRT-PCR) *LOC389023* levels in the PFC, as compared to human (Figure 3.4A–3.4B). Taken together then, these findings strongly suggest that *LOC389023* emerged de novo in human PFC neurons and interacts with localized chromatin templates to mediate transcriptional repression at the *DPP10* locus (Figure 3.6).

Association of Human-Specific H3K4-Methylation Sites with Disease

The aforementioned human-specific gains in histone methylation at *DPP10* and the emergence of human RNA de novo at this locus could reflect a phylogenetically driven reorganization of neuronal functions that may have contributed not only to the emergence of human-specific executive and social-emotional functions, but also for

increased susceptibility for developmental brain disease (Teffer and Semendeferi 2012). In this context, we noticed that the 33 ^{neu}HP (which are defined by two criteria which are (i) human-specific gain compared to non-human primates and (ii) high H3K4me3 in PFC neurons but not their surrounding non-neuronal cells) included multiple genes conferring susceptibility to neurological disease. Three loci, including *DPP10* on chromosome 2q14.1 and two genes in close proximity on chromosome 3p26.3, *CNTN4* and *CHL1*, both encoding cell adhesion molecules (Marshall et al. 2008, Fernandez et al. 2004, Sakuri et al. 2002, Glessner et al. 2009), confer very strong susceptibility to autism, schizophrenia, and related disease. Other disease-associated loci with human-specific gain selectively in PFC neurons include *ADCYAP1*, a schizophrenia (Hashimoto et al. 2007, Ayalew et al. 2012) and movement disorder gene (Nasir et al. 2006) that is part of a cAMP-activating pathway also implicated in posttraumatic stress (Ressler et al. 2011). *PDE4DIP* (*MYOMEGALIN*) (Figure 3.1D) encodes a centrosomal regulator of brain size and neurogenesis (Bond et al. 2006) that in some studies was 9-fold higher expressed in human as compared to chimpanzee cortex (Enard et al. 2002b, Caceres et al. 2003). *SORCS1* is implicated in beta amyloid processing and Alzheimer disease (Reitz et al. 2011, Lane et al. 2010) and attention deficit hyperactivity disorder (Lionel et al. 2011), which again are considered human-specific neurological conditions (Preuss et al. 2004). Because four of 33, or 12% of ^{neu}HP overlapped with neurodevelopmental susceptibility genes (*CNTN4, CHL1, DPP10, SORCS1*), we then checked whether the entire set of 410 human-specific peaks is enriched for genes and loci conferring genetic risk for autism, intellectual disability, and related neurological disease with onset in early childhood.

However, there was only minimal overlap with the Simons Foundation Autism Research Initiative database (SFARI) (Fischbach and Lord 2010), and Human unidentified Gene Encoded protein database (HuGE) for pervasive developmental disorder (including autism) associated polymorphism (Becker et al. 2004), and recent reference lists for mental retardation and/or autism-related genes (each of these databases five or fewer of the human-enriched peaks) (Neale et al. 2012). Likewise, there was minimal, and non-significant overlap with the set of 61 human- and 337 chimpanzee-depleted peaks, or the 551 chimpanzee-enriched in PFC neurons (five or fewer of peaks/database). None of the lists of peaks with human- or chimpanzee-specific gain or loss of H3K4me3 revealed statistical significance for any associations with the Gene Ontology (GO) database. We conclude that DNA sequences subject to differential histone methylation in human or chimpanzee PFC neurons are, as a group, not clustered together into specific cellular signaling pathways or functions. Table 3.1 presents examples of disease-associated genes associated with human-specific gain, or loss of H3K4-trimethylation.

Evolutionary Footprints at Sites Defined by Human-Specific Histone Methylation

To further confirm the role of phylogenetic factors in the emergence of human-specific H3K4me3 peaks, we focused on the set of 33 ^{neu}HP and calculated the total number of human-specific sequence alterations (HSAs), in a comparative genome analyses across five primates (*H. sapiens*, *P. troglodytes, G. gorilla*, *P. abelii, M. mulatta*). We recorded altogether 1,519 HSAs, with >90% as single nucleotide substitutions, five >100 bp INDELs, one (*Alu*) retrotransposon-like element at *TRIB3* pseudokinase consistent with a role of mobile elements in primate evolution (The

Chimpanzee Sequencing and Analysis Consortium 2005), and gain or loss of hundreds of regulatory motifs (Table S12). When compared to a group of (neuronal) H3K4me3 peaks showing minimal changes between the three primate species (Table S13), the ^{neu}HP, as a group, showed a significant, 2.5-fold increase in the number of HSA (20.08±5.52 HSAs versus 8.36±2.44 HSAs per 1-kb sequence, $p = 2.4e{-}06$, Wilcoxon rank sum test; Figure S3). The findings further confirm that genetic differences related to speciation indeed could play a major role for changes in the brain's histone methylation landscape, particularly for H3K4me3 peaks that are highly specific for human neurons (^{neu}HP). Interestingly, none of the above loci showed evidence for accelerated evolution of neighboring protein coding sequences (Table S11), reaffirming the view that protein coding sequences for synaptic and other neuron-specific genes are extremely conserved across the primate tree (Bayes et al. 2011, King and Wilson 1975).

These DNA sequence alterations at sites of neuron-restricted H3K4me3 peaks (with human-specific gain) point, at least for this subset of loci, to a strong evolutionary footprint before the split of human–chimpanzee lineage several million years ago (The Chimpanzee Sequencing and Analysis Consortium 2005). Next, we wanted to find out whether there is also evidence for more recent selective pressures at these loci. Indeed, a subset of ^{neu}HP contain *H. sapiens*-specific sequences not only absent in rodents, anthropoid primates, but even in extinct members of the genus *homo*, including *H. neanderthalensis* and *H. denisova* (Reich et al. 2010). Some of the ancestral alleles (including *MIAT, SIRPA, NRSN*) shared with archaic hominins exhibit very low frequencies at 0%–3% in all modern populations, and therefore it remains possible that

positive selection for newly derived alleles contributed to their high population frequencies in modern humans (Table S14). However, for the entire set of ^{neu}HP that are defined by high H3K4me3 levels in PFC neurons (but not non-neurons), the number of HSAs that emerged after the human lineage was split from *H. denisova* or *H. neanderthalensis* were 3.31% and 1.75%, respectively, which is approximately 2-fold lower as compared to 32 control H3K4me3 peaks with minimal differences among the three primate species (5.03% and 3.77%). The 2-fold difference in the number of *H. sapiens*-specific alleles (^{neu}HP compared to control peaks) showed a strong trend toward significant ($p = 0.067$) for the Denisova, and reached the level of significance ($p = 0.034$) for the Neanderthal genome (based on permutation test with 10,000 simulations (Pitman 1937)). Taken together, these results suggest that at least a subset of the TSS regions with H3K4me3 enrichment in human (compared to non-human primates) were exposed to evolutionary driven DNA sequence changes on a lineage of the common ancestor of *H. sapiens* and the archaic hominins, but subsequently were stabilized in more recent human evolution, after splitting from other hominins.

To provide an example on altered chromatin function due to an alteration in a regulatory DNA sequence that occurred after the human lineage split from the common ancestor with non-human primates, we focused on a change in a GATA-1 motif (A/TGATTAG) within a portion of *DPP10*-2 found in human, within an otherwise deeply conserved sequence across many mammalian lineages (Table S17). Gel shift assays demonstrate that the human-specific sequence harboring the novel GATA-1 site showed much higher affinity to HeLa nuclear protein extracts, compared to the

chimpanzee/other mammal sequence (Figure 3.4C). The emergence of a novel GATA-1 motif at *DPP10* is unlikely to reflect a systemic trend because the motif overall was lost, rather than gained in ^{neu}HP (10/355 versus 4/375, χ^2 $p = 0.053$). Therefore, evolutionary and highly specific changes in a small subset of regulatory motifs at *DPP10* and other loci could potentially result in profound changes in nuclear protein binding at TSS and other regulatory sequences, thereby affecting histone methylation and epigenetic control of gene expression in humans, compared to other mammals including monkeys and great apes. Of note, potentially important changes in chromatin structure and function due to human-specific sequence alterations at a single nucleotide within an otherwise highly conserved mammalian sequence will be difficult to "capture" by comparative genome analyses alone. For example, when the total set of 410 HP was crosschecked against a database of 202 sequences with evidence for human-specific accelerated evolution in loci that are highly conserved between rodent and primate lineages (Pollard et al. 2006), only one of 410 HP matched (Table S15).

Species-Specific Transcriptional Regulation

H3K4me3 is a transcriptional mark that on a genome-wide scale is broadly associated with RNA polymerase II occupancies and RNA expression (Guenther et al. 2005). However, it is also associated with repressive chromatin remodeling complexes and at some loci the mark is linked to short antisense RNAs originating from bidirectional promoters, in conjunction with negative regulation of the (sense) gene transcript (Kanhere et al. 2010, Shi et al. 2006). Indeed, this is what we observed for the *DPP10* locus (Figure 3.6). Therefore, a comprehensive assessment of all transcriptional

changes associated with the evolutionary alterations in H3K4me3 landscape of PFC neurons would require deep sequencing of intra- and extranuclear RNA, to ensure full capture of short RNAs and all other transcripts that lack polyadenylation and/or export into cytoplasm. While this is beyond the scope of the present study, we found several additional examples for altered RNA expression at the site of human-specific H3K4me3 change. There were four of 33 ^{neu}HP loci associated with novel RNA expression specific for human PFC, including the aforementioned *DPP10* locus. The remaining three human-specific transcripts included two additional putative non-coding RNAs, *LOC421321*(chr7p14.3) and *AX746692* (chr17p11.2). There was also a novel transcript for *ASPARATE DEHYDROGENASE ISOFORM 2* (*ASPDH*)(chr19q13.33) (Figure S2). Furthermore, a fifth ^{neu}HP, positioned within an intronic portion of the tetraspanin gene *TSPAN4* (chr11p15.5), was associated with a dramatic, human-specific decrease of local transcript, including the surrounding exons (Figure S2). Comparative analyses of prefrontal RNA-seq signals for the entire set of the 410 HP included at least 18 loci showing a highly consistent, at least 2-fold increase or decrease in RNA levels of human PFC, compared to the other two primate species (Table S18).

Expanded Evolutionary Analysis for Evidence of Positive Selection at Human-specific H3K4me3 Peaks

We then asked whether the subset of DNA sequences with species- and cell type-specific epigenetic regulation, including the ^{neu}HP peaks mentioned above carry a strong footprint of hominid evolution. Indeed, nucleotide substitution analysis revealed that both *DPP10 peaks DPP10 -1/2*, as well as *ADCYAP1*, *CHL1*, *CNTN4*, *NRSN2*, and *SIRPA*

show a significantly elevated rate, with 2- to 5-fold increase specifically in the human branch of the primate tree, when compared to four other anthropoid primate species (*Pan troglodytes, Gorilla gorilla*, *Pongo abelii, Macaca mulatta*) (Table 3.2). The finding that both *DPP10* peaks, *DPP10*-1 and -2 showed a significant, >4-fold increase in nucleotide substitution rates in the human branch of the primate tree—indicating "co-evolution" (or coordinated loss of constraint)—is very plausible given that chromatin structures surrounding these DNA sequences are in direct physical contact (discussed above), reflecting a potential functional interaction and shared regulatory mechanisms between peaks.

To further test whether or not there were recent, perhaps even ongoing selective pressures at loci defined by human-specific gain in H3K4me3 peaks of PFC neurons, we searched for overlap among the peaks in our study with hundreds of candidate regions in the human genome showing evidence of selection during the past 10–100,000 years from other studies. These loci typically extend over several kb, and were identified in several recent studies on the basis of criteria associated with a "selective sweep," which describes the elimination of genetic variation in sequences surrounding an advantageous mutation while it becomes fixed (Tang et al. 2007, Williamson et al. 2007, Kimura et al. 2007, Wang et al. 2006). However, screening of the entire set of 410 human gain and 61 human depleted H3K4me3 sequences against nine datasets for putative selection in humans (Akey 2009) revealed only five loci with evidence for recent sweeps (Table 3.3). One of these matched to the ^{neu}HP on chromosome 2q14.1, corresponding to the second *DPP10* (*DPP10*-2) peak (see above). In independent analyses, using the 1,000 genome database,

we further confirmed recent adaptive fixations around *DPP10-2* (Table S16), as well as two other loci, *POLL and TSPAN4*. While it is presently extremely difficult to determine how much of the genome has been affected by positive selection (of note, a recent metanalysis of 21 recent studies using total genomic scans for positive selection using human polymorphism data revealed unexpectedly minimal overlap between studies (Akey 2009)), we conclude that the overwhelming majority of loci associated with human-specific H3K4me3 gain or loss in PFC neurons (compared to non-human primates) indeed does not show evidence for more recent selective pressures.

Discussion

In the present study, we report that on a genome-wide scale, 471 loci show a robust, human-specific change in H3K4me3 levels at TSS and related regulatory sequences in neuronal chromatin from PFC, in comparison to the chimpanzee and macaque. Among the 410 sequences with human-specific gain in histone methylation, there was a 4-fold overrepresentation of loci subject to species-specific DNA methylation in sperm (Molaro et al. 2011). This would suggest that there is already considerable "epigenetic distance" between the germline of *H. sapiens* and non-human primates (including the great apes), which during embryonic development and tissue differentiation is then "carried over" into the brain's epigenome. The fact that many loci show species-specific epigenetic signatures both in sperm (Molaro et al. 2011) and PFC neurons (Figure 3.1B) raises questions about the role of epigenetic inheritance (Danchin et al. 2011) during hominid evolution. However, to further clarify this issue, additional

comparative analysis of epigenetic markings in brain and germline will be necessary, including histone methylation maps from oocytes, which currently do not exist. However, the majority of species-specific epigenetic decorations, including those that could be vertically transmitted through the germline, could ultimately be driven by genetic differences. On the basis of DNA methylation analyses in three-generation pedigrees, more than 92% of the differences in methylcytosine load between alleles are explained by haplotype, suggesting a dominant role of genetic variation in the establishment of epigenetic markings, as opposed to environmental influences (Gertz et al. 2011). A broad overall correlation between genetic and epigenetic differences was also reported in a recent human–chimpanzee sperm DNA methylation study (Molaro et al. 2011), and there is general consensus that the inherent mutability of methylated cytosine residues due to their spontaneous deamination to thymine is one factor contributing to sequence divergence at CpG rich promoters with differential DNA methylation between species (Molaro et al. 2011, Saxonov et al. 2006). Furthermore, human-specific sequences in the DNA binding domains of *PRDM9*, which encodes a rapidly evolving methyltransferase regulating H3K4me3 in germ cells, were recently identified as a major driver for human–chimpanzee differences in meiotic recombination and genome organization (Myers et al 2010). It will be interesting to explore whether PRDM9-dependent histone methyltransferase activity was involved in the epigenetic regulation of the human-enriched H3K4me3 peaks that were identified in the present study.

Another interesting finding that arose from the present study concerns the non-random distribution of histone methylation peaks with human-specific gain, due to a

significant, 2- to 3-fold overrepresentation of peak-pairing or -clustering on a 500 kb to 1 Mb scale. This result fits well with the emerging insights into the spatial organization of interphase chromosomes, including the "loopings," "tetherings" and "globules" that bring DNA sequences that are spatially separated on the linear genome into close physical contact with each other (Sanyal et al. 2011). Specifically, many chromosomal areas are partitioned into Mb-scale "topological domains", which are defined by robust physical interaction of intra-domain sequences carrying the same set of epigenetic decorations (Dixon et al. 2012). These mechanisms could indeed have set the stage for coordinated genetic and epigenetic changes during the course of hominid brain evolution. The *DPP10* (2q14.1) neurodevelopmental susceptibility locus provides a particularly illustrative example: here, two H3K4me3 peak sequences with strong human-specific gain were separated by hundreds of kilobases of interspersed sequence, yet showed a strikingly similar, 4-fold acceleration of nucleotide substitution rates specifically in the human branch of the primate tree. Importantly, the two H3K4me3 peaks, *DPP10-1* and -*2*, as shown here, are bundled together in a loop or other types of higher order chromatin. Therefore, our findings lead to a complex picture of the human-specific shapings of the neuronal epigenome, including a mutual interrelation of DNA sequence alterations and epigenetic adaptations involving histone methylation and higher order chromatin structures. The confluence of these factors could then, in a subset of PFC neurons (Figure 3.5A), result in the expression of a novel antisense RNA, which associates with transcriptional repressors to regulate the target transcript in cis, *DPP10* (Figures 3.5D and (Shilatifard 2006)).

While the present study identified a few loci, including the aforementioned *DPP10* (chromosome 2q14.1), in which DNA sequences associated with a human-specific gain in neuronal histone methylation showed signs for positive selection in the human population, it must be emphasized that the overwhelming majority of sites with human-specific H3K4me3 changes did not show evidence for recent adaptive fixations in the surrounding DNA. Therefore, and perhaps not unsurprisingly, neuronal histone methylation mapping in human, chimpanzee, and macaque primarily reveals information about changes in epigenetic decoration of regulatory sequences in the hominid genome after our lineage split from the common ancestor shared with present-day non-human primates.

Moreover, according to the present study, the subset of 33 sequences with human-specific H3K4me3 gain and selective enrichment in neuronal (as opposed to non-neuronal) PFC chromatin show a significant, 3-fold increase in human-specific (DNA sequence) alterations in comparison to non-human primate genomes. This finding speaks to the importance of evolutionary changes in regulatory sequences important for neuronal functions. Strikingly, however, the same set of sequences show a significant, approximately 1.5- to 2-fold decrease in sequence alterations when compared to the two archaic hominin (*H. denisova, H. neanderthalensis*) genomes. This finding further reaffirms that sequences defined by differential epigenetic regulation in human and non-human primate brain, as a group, are unlikely to be of major importance for more recent evolution, including any (yet elusive) genetic alterations that may underlie the suspected differences in human and neanderthal brain development (Gunz et al. 2012). However,

these general conclusions by no means rule out a critical role for a subset of human-specific sequence alterations on the single nucleotide level within any of the HPs described here, including the *DPP10* locus.

Such types of single nucleotide alterations and polymorphisms may be of particular importance at the small number of loci with human-specific H3K4me3 gain that contribute to susceptibility of neurological and psychiatric disorders that are unique to human (though it should be noticed that as a group, the entire set of sequences subject to human-specific gain, or loss, of H3K4me3 are not significantly enriched for neurodevelopmental disease genes). The list would not only include the already discussed *ADYCAP1*, *CHL1*, *CNTN4*, and *DPP10*, which were among the narrow list of 33 human-specific peaks highly enriched in neuronal but not non-neuronal PFC chromatin), but also *DGCR6*, an autism and schizophrenia susceptibility gene (Liu et al. 2002, Guilmatre et al. 2009) within the DiGeorge/Velocardiofacial syndrome/22q11 risk locus, *NOTCH4* and *CACNA1C* encoding transmembrane signaling proteins linked to schizophrenia and bipolar disorder in multiple genome-wide association studies (Ikeda et al. 2011, Sklar et al. 2011), *SLC2A3* encoding a neuronal glucose transporter linked to dyslexia and attention-deficit hyperactivity disorder (Lesch et al. 2011, Roeske et al. 2011) and the neuronal migration gene *TUBB2B* that has been linked to polymicrogria and defective neurodevelopment (Jaglin et al. 2009). Furthermore, among the 61 peaks with human-specific loss of H3K4me3 is a 700-bp sequence upstream of the TSS of *FOXP2*, encoding a forkhead transcription factor essential for proper human speech and language capabilities (Vargha-Khadem et al. 2005) and that has been subject to accelerated

evolution with amino acid changes leading to partially different molecular functions in human compared to great apes (Enard et al. 2002a, Konopka et al. 2009). The homeobox gene *LMX1B* is another interesting disease-associated gene that is subject to human-specific H3K4me3 depletion (Table 3.1). While expression of many of these disease-associated genes is readily detectable even in mouse cerebral cortex (Belgard et al. 2011), the neuropsychiatric conditions associated with them lack a correlate in anthropoid primates and other animals. This could speak to the functional significance of H3K4 methylation as an additional layer for transcriptional regulation, with adaptive H3K4me3 changes at select loci and TSS potentially resulting in improved cognition while at the same time in the context of genetic or environmental risk factors contribute to neuropsychiatric disease. More generally, our findings are in line with a potential role for epigenetic (dys)regulation in the pathophysiology of a wide range of neurological and psychiatric disorders (Tsankova et al. 2007, Robison and Nesler 2011, Day and Sweatt 2012, Jakovcevski and Akbarian 2012).

Our study also faces important limitations. While we used child and adult brains for cross-species comparisons, human-specific signatures in the cortical transcriptome are thought to be even more pronounced during pre- and perinatal development (Somel et al. 2009). Therefore, younger brains could show changes at additional loci, or more pronounced alterations at the TSS of some genes identified in the present study, including the above mentioned susceptibility genes *CNTN4* and myelomegalin/*PDE4DIP*, which are expressed at very high levels in the human frontal lobe at midgestation (Lambert et al. 2011). In this context, our finding that a large majority, or 345 of 410 H3K4me3 peaks

showed a human-specific gain both in children and adults, resonates with Somel and colleagues (Somel et al. 2011) who suggested that some of the age-sensitive differences in cortical gene expression among primate species are due to trans-acting factors such as microRNAs while *cis*-regulatory changes (which were the focus of the present study) primarily affect genes that are subject to a lesser regulation by developmental processes. More broadly, our studies support the general view that transcriptional regulation of both coding and non-coding (including antisense) RNAs could play a role in the evolution of the primate brain (Babbitt et al. 2010).

Furthermore, the cell type-specific, neuronal versus non-neuronal chromatin studies as presented here provide a significant advancement over conventional approaches utilizing tissue homogenate. However, pending further technological advances, it will be interesting to explore genome organization in select subsets of nerve cells that bear particularly strong footprints of adaptation, such as the Von Economo neurons, a type of cortical projection neuron highly specific for the hominid lineage of the primate tree and other mammals with complex social and cognitive-emotional skill sets (Butti et al. 2011). Furthermore, our focus on PFC does not exclude the possibility that other cortical regions (Konopka et al. 2012), or specialized sublayers such as within the fourth layer of visual cortex that shows a complex transcriptional architecture (Bernard et al. 2012), show human-specific histone methylation gains at additional TSS that were missed by the present study.

More broadly, the approach provided here, which is region- and cell type-specific epigenome mapping in multiple primate species, highlights the potential of epigenetic markings to identify regulatory non-coding sequences with a potential role in the context of hominid brain evolution and the shaping of human-specific brain functions. Remarkably, a small subset of loci, including the aforementioned *DPP10* (chromosome 2q14.1), shows evidence for ongoing selective pressures in humans, resulting in DNA sequence alterations and the remodeling of local histone methylation landscapes, after the last common ancestor of human and non-human primates.

Materials and Methods[2]

Primate Alignments

For nucleotide sequences used in the baseml analysis (Yang 1998), peak sequences were obtained in humans using the coordinates for human-specific that were 2-fold greater than both chimp and macaque. UCSC's liftover utility was used to obtain sequences in 3 additional primate species: chimpanzee, orangutan, and macaque. These sequences were then aligned using ClustalW (Thompson et al. 2002), with default settings.

For amino acid sequences used in codeml analyses (Yang et al. 2000), gene sequences were obtained using the BioMart tool available through Ensembl. Sequences were retrieved from five primate species: human, chimpanzee, orangutan, gorilla, and macaque. The Ensembl Genes 64 database was used for all species. Briefly, human

[2]I have included the methods for my contributions in the main body here. A complete description of methods can be found in Appendix I.

sequences were obtained by setting the gene filter to use WikiGene Names, and the "Sequence" attribute set to coding sequences. The "Homologs" attribute was used to return Ensembl Gene IDs

for the orthologous sequences in the remaining 4 primate species. These IDs were then queried under each species gene dataset in the same way as humans, except the gene filter was set to Ensembl Gene IDs, instead of WikiGene Names. These sequences were then aligned using PRANK (Löytnoja and Goldman 2005) with the codon option, which uses the empirical codon model of Kosiol et al. (2007).

Nucleotide substitution rates in humans

Baseml (Yang 1998) from PAML version 4.4 (Yang 2007) was used to determine nucleotide substitution rates for the primate nucleotide sequence alignment. Two rate classes were specified for the nucleotide sequence alignments. This was accomplished by setting the clock parameter in the baseml control file equal to 2, and then numbering the foreground branch in the tree file according to the numbering scheme explained in the PAML documentation. A likelihood ratio test with X_1^2 and p value < 0.01 was used to determine significance.

Accelerated amino acid substitution rates in humans

Codeml (Yang et al. 2000) from PAML version 4.4 was used to analyze the primate amino acid sequence alignments for differences in dN/dS. For the sites model three model comparisons were used: M1a vs. M2a, M7 vs. M8, M8a vs. M8 (see PAML

documentation for parameter settings). M1a has two subsets of sites, one where ω varies between 0 and 1 and one where ω is fixed at one; in M2a ω can be less than 1, equal to 1, or greater than 1. M7 assumes a beta-distribution for ω between 0 and 1, and M8 adds an additional class of sites to M7 with ω>1 (Wong et al. 2004). In M8a this additional class is fixed at ω=1 (Swanson et al. 2003). Thus, M2a and M8 allow selection in each comparison, while M2, M7, and M8a fit the data to a neutral model. A maximum likelihood ratio is computed for each model, and the null and selection models are compared via a likelihood ratio test, with X_1^2 for M1a vs. M2a, and M8a vs. M8, and X_2^2 for M7 vs. M8. For the branch model, two rate classes were specified for the amino acid sequence alignments, with the same specifications as in the baseml branch model except that the model parameter was set to equal 2 instead of the clock. A likelihood ratio test with X_1^2 and a p value < 0.01 was used to determine significance.

SNP dataset

SNPs were obtained for 176 Yoruban individuals from the 1000 genomes May 2010 merged SNP call release (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/). Ancestral alleles were filled in using the 6 way EPO human ancestral alignment for GRCh37 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/). Sites were omitted if no ancestral allele was identified. If there were no ancestral alleles for all sites within a region, that region was omitted from further analysis.

Sweep analysis

Regions were analyzed using Kim and Stephan's clsw program (Kim and Stephan 2002). The regions were split into windows of 100bp and a likelihood ratio (LR) is calculated for each, along with an estimate of alpha (the selection coefficient) and the most likely location of the target of selection for each window. The window containing the maximum LR is considered the ultimate location of a sweep, if one has occurred . To determine this, neutral simulations were performed using ms (Hudson 2002). Theta (the mutation rate parameter) was estimated by clsw for each region, and the human-like value for rho (the recombination parameter) was taken from Nielsen et al 2009. Any regions with LR values falling within the top 5% of the LR distribution from neutral simulations were considered significant. For these significant regions, the goodness-of-fit test (GOF) from Jensen et al. 2005 was applied to distinguish between regions that rejected neutrality due to true selection, and those that rejected neutrality because of confounding demographic factors

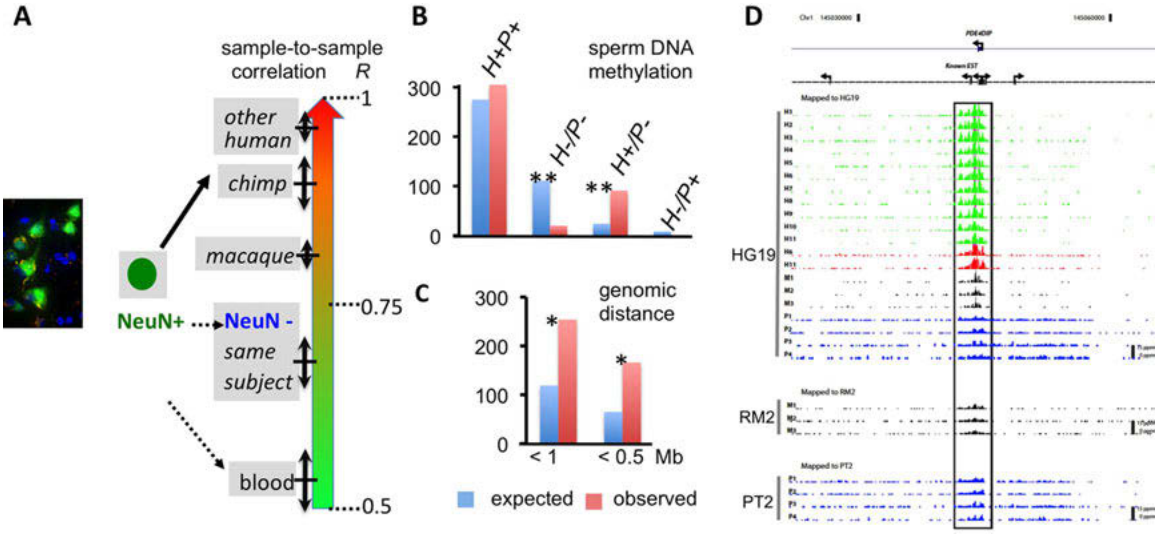


Figure 3.1. Human-specific signatures of the neuronal epigenome in PFC.
(A) Pearson correlation coefficients (*R*, mean ± standard deviation [SD]) for sample-to-sample comparison of H3K4me3 ChIP-seq normalized tag counts within Refseq promoters, revealing cell type- and species-specific signatures. (B) Expected (blue)/observed (red) counts of human-specific H3K4me3 peaks ($n = 410$) overlapping with DNA hypomethylated regions in human (H)/chimpanzee (P) sperm. Notice 4-fold enrichment for loci with human-only (H+,P−) DNA hypomethylation in dataset. (C) The actual co-localization of human-specific H3K4me3 peaks ($n = 410$) within 1- or 0.5-Mb genomic distance is 2–3-fold higher than expected (based on average distribution of entire set of 34,639 H3K4me3 peaks *(**), $p<10^{-3(-4)}$. (D) Representative example of a TSS (*PDE4DIP/Myelomegalin* ("*regulator of brain size*") with species- and cell type-specific H3K4me3 profile. Genome browser tracks showing ChIP-seq H3K4me3 signal at *PDE4DIP* (chromosome 1) locus, annotated to HG19/PT2/RM2 genomes as indicated. Green/blue/black tracks from PFC neuronal (NeuN+) nuclei of 11 humans/four chimpanzees/three macaques as indicated. Red tracks, non-neuronal (NeuN−) human PFC nuclei. Notice much stronger PDE4DIP peaks in human neurons.

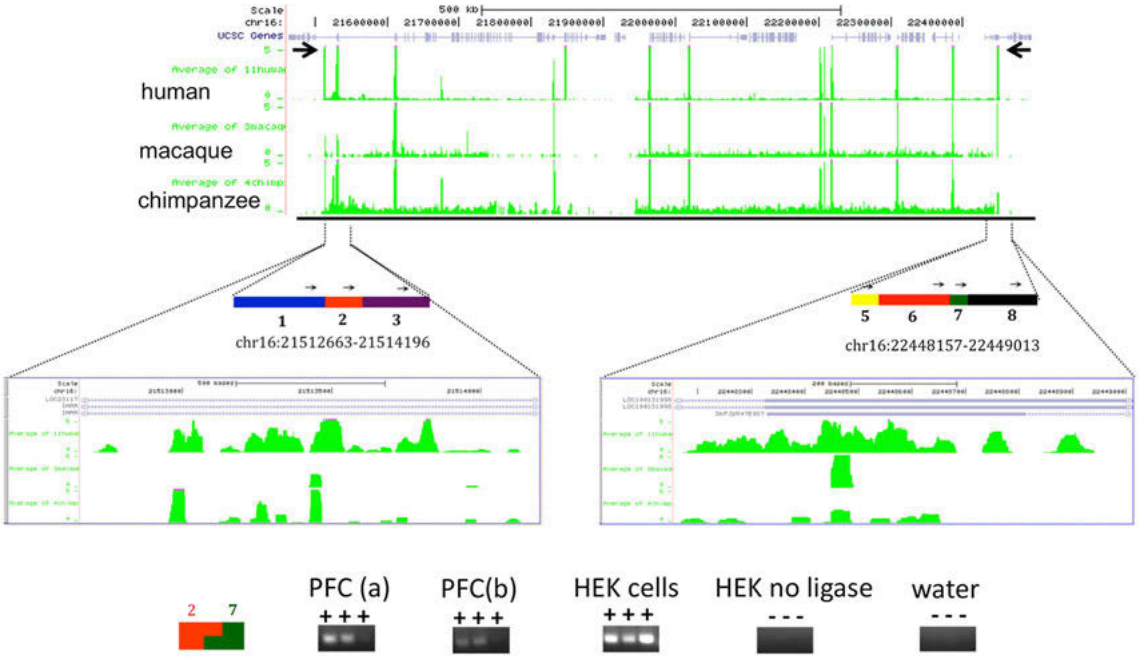


Figure 3.2. H3K4me3 landscapes and higher order chromatin at the psychiatric susceptibility locus, 16p11.2.

(Top) UCSC genome browser window track for approximately 1 Mb of human chr16: 21,462,663–22,499,013, with H3K4me3 ChIP-seq tracks from neuronal chromatin (PFC) of three primate species, as indicated. Notice human-enriched H3K4me3 peaks at chr16:21,512,663–21,514,196 and chr16:22,448,157–22,449,013 (marked by arrows) flanking numerous peaks common to all 3 species. (Bottom) Rectangles and thin arrows mark 3C HindIII restriction fragments and primers from 3C assays. Notice positive interaction of sequences captured by primers 2 and 7, agarose gels shows representative 196-bp PCR product for 3C from two PFC specimens (a,b), HEK cells, and no ligase and water controls.

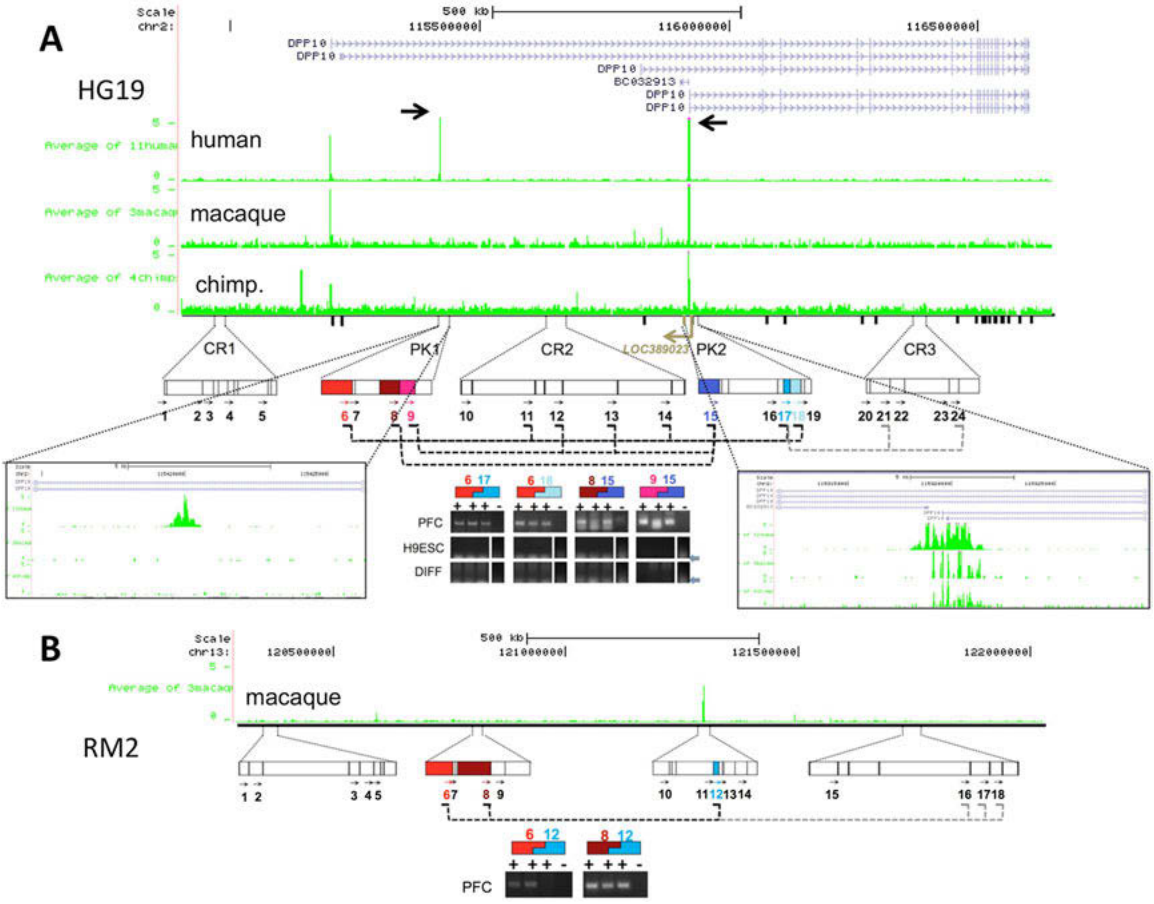


Figure 3.3. H3K4me3 landscapes and higher order chromatin at *DPP10* (2q14.1). (A) (Top) Genome browser tracks showing ChIP-seq H3K4me3 signal at *DPP10* locus annotated to HG19 and RM2 genomes. Data expressed as normalized tag densities, averaged for 11 humans, four chimpanzees, and three macaques as indicated. Human-specific peak *DPP10*-1 (1,455 bp) and *DPP10*-2 (3,808 bp) marked by arrows and shown at higher resolution in boxes, as indicated. (Bottom) Rectangles and arrows mark Hind III restriction fragments and primers from *DPP10*-1/2 (PK1, 2) and control regions (CR1-3) for 3C assays (human). Dotted lines connect primer pairs with sequence-verified product, indicating physical interaction of the corresponding fragments. Agarose gels for representative PCR products from 3C with (+) or without (−) DNA ligase (human primers 6,17: 282 bp; 6,18: 423 bp; 8,15: 160 bp; 9,15: 130 bp). (B) Rectangles and arrows mark Hind III restriction fragments and primers for corresponding DPP10 sequences in RM2, for macaque brain 3C. Macaque primers 6,12:298 bp, 8,12:154 bp. Notice positive interaction of PK1 with PK2 and neighboring CR2, but with not CR1 or CR3. Notice no signal in PFC 3C assays without DNA ligase and no signal in all 3C assays from H9 pluripotent (H9ESC) and differentiated (DIFF) cell cultures.

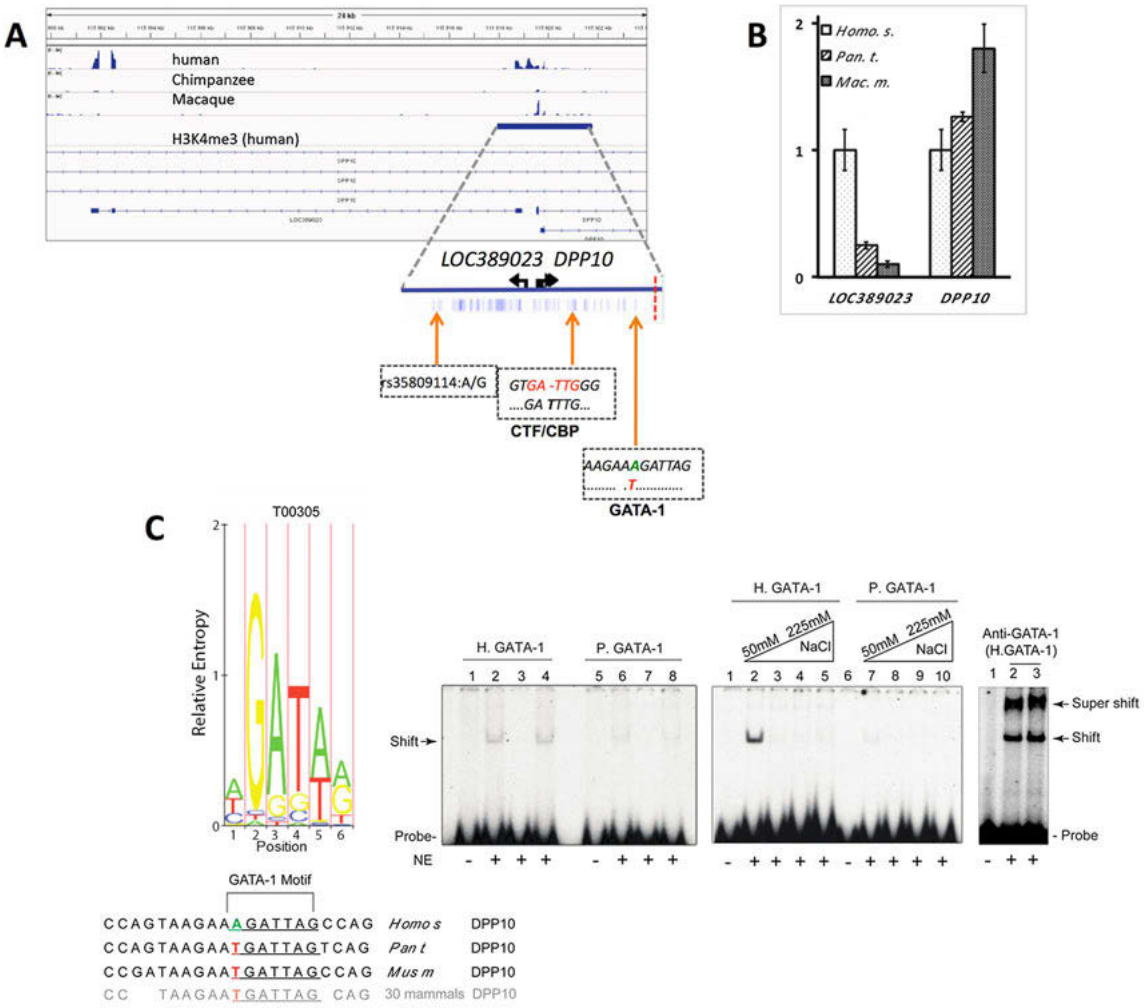


Figure 3.4. Novel transcripts and regulatory motifs at the DPP10 *locus*.
(A) (top) *DPP10* and *LOC389023*, extracted from published RNA-seq datasets from human/chimpanzee/macaque PFC. (Bottom) shows 3.8-kb *DPP10*-2 bidirectional promoter, blue tick marks for human-specific sequence divergence from five other anthropoid primates, including (from left to right) SNP rs35809114, and fixed polymorphism with novel CTF/CBP motif not found in archaic hominins (*H. denisova*, *H. neanderthalensis*) and novel GATA-1 motif within highly conserved sequence across many mammalian lineages. The vertical dotted red line marks the potential center of an adaptive fixation in modern humans (see text). (B) Bar graphs summarize qRT-PCR on PFC RNA showing much higher *LOC389023* in human, and lower expression of *DPP10* exons downstream of *DPP10*-2 peak $*^{(**)}$ p<0.05 (0.01). (C) (Left) GATA-1 consensus motifs/binding affinities (http://snpper.chip.org/mapper). (Right) HeLa nuclear extract (NE) gel shifts with ^{32}P-labeled 21 bp duplex probes for human (H) and chimpanzee (P) sequences encompassing GATA-1 motif as indicated. (Left gel) lanes (1,2,5,6) labeled probe, (3,7) cold competitor, (4,8) unrelated duplex, or increasing salt concentrations as indicated. Anti-GATA supershift assay confirms GATA-1 protein binding to probe sequence.

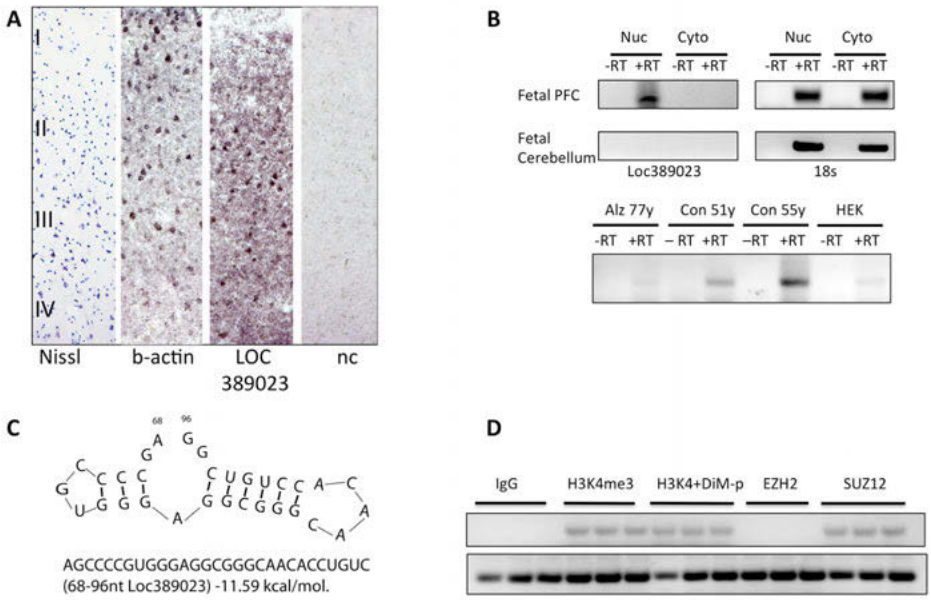


Figure 3.5. Cellular distribution and molecular affinities of human-specific RNA, *LOC389023*.

(A) Digitized images of sections from adult human PFC, stained with (left to right) Nissl, b-actin, LOC389023, and negative control (nc). Notice numerous LOC389023-expressing cells in cortical layers II–IV but not in neuron-poor layer I. (B) (Top) LOC389023, and for loading control, 18S rRNA PCR from nuclear (Nuc) and cytosolic (Cyto) RNA extracts, showing robust LOC389023 expression in nuclear fraction but not cytosolic of a prenatal (around 35 wk of gestation) PFC specimen. No LOC389023 expression was found in fetal cerebellum. (Bottom) PCR from nuclear RNA isolates of adult PFC specimens and of HEK cell line. Notice weak signal in neurodegenerative Alzheimer PFC specimen, no signal in peripheral (HEK) cells, and strong signal in PFC nuclei from normal adult controls. (C) GC rich stem loop of LOC389023 (see text). (D) RT-PCR for LOC389023 from (top) pulldowns of transfected neuroblastoma cells, (left to right) IgG, H3K4-trimethylated nucleosomal preparation co-incubated with or without dimethyl-H3K4-blocking peptide, anti-EZH2, anti-SUZ12, and (bottom) input loading control. Notice specific affinity of LOC389023 for H3K4me3 and SUZ12.

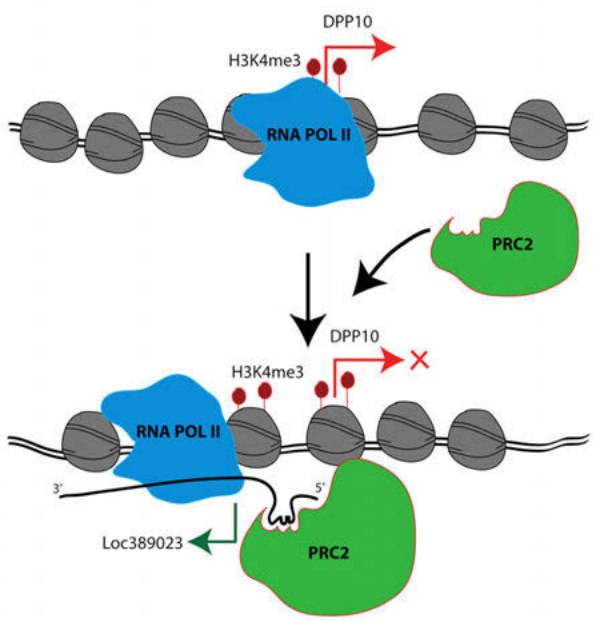


Figure 3.6. Hypothetical mechanism of action of novel human-specific RNA, LOC389023.

(Top) In non-human primate, *DPP10* transcripts are expressed by the RNA polymerase II complex from the *DPP10-2* promoter (see text) that is tagged with H3K4me3. (Bottom) In human, there is specific gain of H3K4me3 signal particularly in the 5′ portion of the DPP10-2 promoter (see text), which is associated with a novel antisense RNA, LOC389023. This RNA recruits Polycomb 2(PRC2) and other transcriptional repressors in *cis*, thereby inhibiting expression of the sense transcript, *DPP10*.

Table 3.1. Examples of disease-associated genes with human-specific gain or loss of H3K4 trimethylation in PFC neurons.

Gene; Location; HGNC	Gene	H3K4me3 Change in Human	Disease Association	Function in the Forebrain, Including Cerebral Cortex
ADCYAP1; 18p11.32; 241	adenylate cyclase activating polypeptide 1	Gain	Schizophrenia [46,47], movement disorder [48], PTSD [49]	Alternate camp signaling pathway, mediates synaptic plasticity and LTD in hippocampus [95]
CACNA1C; 12p13.33; 1390	calcium channel, voltage-dependent, L type, alpha 1C subunit	Gain	Confers genetic risk for mood, psychosis, and autism spectrum disorders [96,97]	Coupling of cell membrane depolarization to transient increase of membrane permeability for calcium [96]
CHL1; 3p26.3; *1939*	cell adhesion molecule with homology to L1CAM	Gain	Autism, schizophrenia [35,44–46]	Thalamocortical axon guidance via interaction with ephrin receptors [98,99]
CNTN4; 3p26.3; *2174*	contactin 4	Gain	Autism, intellectual disability [34,43–45]	Developmental patterning of functional odor maps in olfactory bulb, axon-associated cell adhesion molecule [34,43–45]
DGCR6; 22q11.21; 2844	DiGeorge syndrome critical region gene 6	Gain	Autism, schizophrenia [74,75]	Regulates intracellular distribution of $GABA_B$ receptor [100]
DPP10; 2q14.1; 20823	dipeptidyl-peptidase 10	Gain	Autism, mood disorder, schizophrenia, asthma [34–36]	Regulation of neuronal excitability as auxiliary subunit of potassium channels [33]
FOXP2; 7q31.1; 13875	forkhead box P2	Loss	Speech and language disorder with subtle structural and functional changes in brain circuitry [81,82]	Transcription factor regulating gene expression programs in vocal communication, including human speech and birdsong [82,101,102]
LMX1B; 9q33.3; 6654	Lim homeobox transcription factor 1, beta	Loss	ADHD and depression [103]	Key control point in gene expression programs for dopaminergic and serotonergic neurons [104,105]
NOTCH4; 6p21.3; 7884	neurogenic locus notch homolog gene 4	Gain	Schizophrenia [106,107]	Endothelial Notch 4 regulates brain vasculature [108]
PDE4DIP; 1q21.1; 15580	phoshodiesterase 4D Interacting protein	Gain	Altered phospho-diesterase signaling broadly relevant for mood and psychosis spectrum disorders [109,110]	Anchor protein for cAMP pathway in the Golgi/centrosomal complex, homologue to drosophila *centrosomin* regulating brain development and implicated in neurogenesis [50,111]
SLC2A3; 12p13.31; 11007	solute carrier family 2 (facilitated glucose transporter), member 3	Gain	Dyslexia, ADHD [78,79]	Neuronal glucose transporter, highly expressed in neuronal processes and synaptic structures and neuropil of human cerebral cortex and other brain regions [112,113]
SORCS1; 10q25.1; 16697	sortilin-related VPS10 domain containing receptor 1	Gain	ADHD [54]	In a complex with pro-NGF, involved in NGF-mediated cell signaling and neuroapoptosis [114]. Interacts, like other sortilins, with gamma-secretase implicated in Alzheimer disease [54]
TRIB3; 20p13; 16228	tribbles homolog 3 pseudo-kinase	Gain	Genetic determinant for information-processing speed in human [115] and insulin-dependent diabetes [116]	Competes in complex with ATF4 with CREB transcription factor to regulate expression of synaptosomal-associated protein 25 (SNAP-25) involved in insulin exocytosis and neurotransmission [116]
TUBB2B; 6p25.2; 30829	class IIb beta-tubulin	Gain	Cortical malformations including poly-microgyria [117], microcephaly, seizures, intellectual disability [118]	Essential for neuronal migration and other functions of the microtubuli complex [80]
ZNF423; 16p12.1; 16762	zinc-finger protein 423	Loss	16q12 microdeletion syndrome with micro-cephaly and dysmorpho-genesis of fore- and hindbrain [119,120]	C2H2-type zinc finger transcription factor that controls the switch to neuronal maturation during olfactory neurogenesis [121] and axonal projections across forebrain commissures [122].

ADHD, attention deficit hyperactivity disorder; LTD, long-term depression; NGF, nerve growth factor; PTSD, post-traumatic stress disorder.
doi:10.1371/journal.pbio.1001427.t001

Table. 3.2. Summary of Positive Selection Results

Baseml was used to estimate branch-specific nucleotide substitution rates for alignments containing human, chimpanzee, orangutan, and macaque sequences. Codeml was used to analyze site-specific amino acid substitution rates. The alignments for codeml contained sequences from 5 primate species: human, chimpanzee, gorilla, orangutan, and macaque. Only significant results with $p < 0.01$ are shown, as well as 1:1 orthologs in all species.

Coordinates	TSS	Distance to TSS	Accelerated NT Evolution[1]	Accelerated AA Evolution[2]	Selective Sweeps[3]
chr10:103328874-103331401	POLL	16572			5.16
chr11:857724-860732	TSPAN4	13279			7.49
chr18:903446-909852	ADCYAP1	0	4.90		
chr19:51016136-51018764	ASPDH	0		7.28	
chr2:115419153-115420608	DPP10	219255	4.18		
chr2:115917965-115921773	DPP10	0	4.39		5.31
chr20:1782608-1784925	SIRPA	89887	2.24		
chr20:326778-329498	NRSN2	0	1.92		
chr3:2139182-2142802	CNTN4	0	3.08		
chr3:237095-242136	CHL1	0	1.99	3.32	

[1] Nucleotide substitution rate

[2] d_N/d_S

[3] 2Ns

Table 3.3. Comparison of 410 human-specific neuronal peaks with published genomic scans for positive selection in humans

Region (hg19)	Reference(s)	Nearest TSS
Overlap with 9 scans of positive selection		
chr1:148555814-148557118	1	NBPF15
chr11:49582211-49584242	1	LOC440040
chr19:11784427-11785561	1	ZNF833P
chr2:115918003-115921686	1	DPP10
chr4:6246915-6248356	2, 3, 4	WFS1
Overlap with 202 human accelerated regions (HARs)		
chr20:61732970-61734710	5	HAR1B

1. Wang et al. 2006. Proc Natl Acad Sci U S A 103: 135-140.
2. Kimura et al. 2007. PLoS ONE 2: e286.
3. Tang et al. 2007. PLoS Biol 5: e171.
4. Williamson et al. 2007. PLoS Genet 3: e90.
5. Pollard et al. 2006. PLoS Genet 2: e168.

CHAPTER IV: The Impact of Equilibrium Assumptions on Tests of Selection

Abstract

With the increasing availability and quality of whole genome population data, various methodologies of population genetic inference are being utilized in order to identify and quantify recent population-level selective events. Though there has been a great proliferation of such methodology, the type-I and type-II error rates of many proposed statistics have not been well-described. Moreover, the performance of these statistics is often not evaluated for different biologically relevant scenarios (e.g., population size change, population structure), nor for the effect of differing data sizes (i.e., genomic vs. sub-genomic). The absence of the above information makes it difficult to evaluate newly available statistics relative to one another, and thus difficult to choose the proper toolset for a given empirical analysis. Thus, we here describe and compare the performance of four widely used tests of selection: SweepFinder, SweeD, OmegaPlus, and iHS. In order to consider the above questions, we utilize simulated data spanning a variety of selection coefficients and beneficial mutation rates. We demonstrate that the LD-based OmegaPlus performs best in terms of power to reject the neutral model under both equilibrium and non-equilibrium conditions. The results presented here ought to serve as a useful guide for future empirical studies, and provides a guide for statistical choice depending on the history of the population under consideration. Moreover, the parameter space investigated

and the Type-I and Type-II error rates calculated, represent a natural benchmark by which future statistics may be assessed.

Introduction

Population genetics seeks to characterize the forces that shape genomic variation, an endeavor that is often challenged by difficulties in unraveling the effects of selective and neutral processes. When positive selection acts on a new beneficial mutation, it will rise in frequency within a population over time, bringing nearby linked variation with it (Maynard Smith and Haigh 1974). The pattern resulting from this process is referred to as a selective sweep, and can be observed in the site frequency spectrum (SFS) and the extent of linkage disequilibrium (LD) flanking the beneficial fixation (see reviews of Nielsen (2005)). Briefly, genetic variation within a swept region is expected to be reduced, and the site frequency spectrum skewed towards an excess of both rare and high frequency derived mutations. The haplotype patterns surrounding the beneficial allele are expected to be significantly impacted (e.g., Stephan et al. 2006) as well – and it has thus been suggested that a selective sweep may be identified by a characteristic haplotype pattern in which LD is increased in regions flanking a recent beneficial fixation, but reduced across the site of fixation (Jensen et al. 2007; Pavlidis et al. 2010).

Demographic forces also affect genetic variation and haplotype structure. For instance, spontaneous changes in population size can create longer haplotypes that may strongly resemble patterns expected after a selective sweep (Pavlidis et al. 2010). Additionally, as demonstrated by Barton (1998), the expected coalescent trees

generated by a bottleneck may indeed be identical to those generated by selection, and simulation studies have demonstrated that tests of selection are prone to extremely high false positive rates under certain bottleneck models (e.g., Jensen et al. 2005; Thornton and Jensen 2007).

Numerous methods for estimating selection and demography have been developed to deal with these challenges (for review see Crisci et al. 2011). Many tests of selection have taken an outlier-based approach– thus, a statistic is computed across an entire dataset and a top fraction of values are considered selection candidates. One limitation of this approach is the assumption of an equilibrium neutral background, with deviations being interpreted as evidence of non-neutrality (rather than non-equilibrium). While it has been proposed to first fit a demographic model in order to increase power to detect selective sweeps (e.g., Williamson et al. 2005; Keightley and Eyre-Walker 2007), the demographic estimators themselves generally assume neutrality – and thus the demographic fitting may account for much of the pattern in the data owing to selection.

We here focus on identifying selection in simulated recurrent hitchhiking (RHH) and single hitchhiking (SHH) datasets using four commonly used selection estimators: SweepFinder, SweeD, OmegaPlus, and iHS (Nielsen et al. 2005, Pavlidis et al. 2013, Alachiotis et al. 2012, Voight et al. 2006). We consider equilibrium and non-equilibrium neutral and selection models. Our intent is to investigate the bias in selection estimators under non-equilibrium neutral conditions, and to characterize the demographic parameter space for which neutral and selective models may not be

differentiated. Further, given the increasing number of proposed statistics in this area, we would like to emphasize the importance of proper power testing – and we here seek to describe performance across equivalent models. We hope that the statistical testing presented here, and the simulation panel assembled, may serve as a template against which future statistics may be evaluated allowing for a direct comparison with previously proposed methodology.

For our considered models, we find that the performance of the standard implementation of SweepFinder has very few rejections of neutrality under even equilibrium models with moderately strong selection ($2Ns = 1000$). SweeD had slightly improved performance, but mainly achieved a reduced sensitivity to SNP density owing to the inclusion of monomorphic sites. OmegaPlus was found to have the most power to detect selection, but remains prone to high false-positive rates under certain neutral non-equilibrium models. Finally, while iHS performs well under equilibrium conditions, it is unable to distinguish selective effects from those of a population bottleneck. Thus, in addition to serving as a benchmark for future studies, these results highlight the need for continued methodological development in this area.

Methods

Simulation Parameters

Recurrent hitchhiking models (i.e., selective sweeps defined to occur at a specific rate) were simulated using sfs_code (Hernandez 2008), a forward simulation

program that can simulate both selection and demography simultaneously. Single hitchhiking models (i.e., a single selective sweep occurs at a specified time) were simulated using msms, which can also model both selection and demography (Ewing and Hermisson 2010). For both sets of models a single locus of 50Kb was simulated using human-like parameters for population size N=10000, mutation rate (θ=0.001/site), and recombination rate (ρ=0.001/site). For each set of parameters, 1000 simulations were performed with 40 haplotypes sampled.

Selection parameters were set as follows: for single hitchhiking events, the selected allele was located in the center of the locus with $2Ns$ = 1000, 100, and 10 for dominant alleles, and 500, 50, and 5, respectively, for heterozygous alleles. For recurrent hitchhiking, selection occurs on a new mutation with a specified probability (= 0.0002, 0.01, 0.1, or 0.25). Our models encompass equilibrium neutral, equilibrium selection, non-equilibrium neutral, and non-equilibrium selection – with bottlenecks ranging in severity from 25% to 99% size reduction and ranging in recovery time from 1000 to 4000 generations.. A complete list of the parameters of mixture models can be found in the Table legends.

Comparison of the Different Selection Statistics

We evaluate selection statistics based on either the SFS (SweepFinder, SweeD) or patterns in LD (OmegaPlus, iHS) to identify regions that contain a selective sweep. These statistics were chosen because of their widespread use in population genetics, and for their public accessibility.

SweepFinder uses information from the SFS to determine the probability of observing an allele at a given frequency and distance from a beneficial mutation (Nielsen et al. 2005, http://people.binf.ku.dk/rasmus/webpage/sf.html). This method is based on the similar framework of Kim and Stephan (2002), but the null SFS is determined from the background site frequency spectrum rather than a strictly equilibrium neutral model. This approach has been argued to make the test more robust to demographic history and variation in mutation rate. SweepFinder is designed to detect completed sweeps in both subgenomic, and genomic datasets.

SweeD is a computationally improved version of SweepFinder that is capable of analyzing much larger datasets (thousands of sequences vs. hundreds for SweepFinder) in a cluster-computing environment (Pavlidis et al. 2013, http://sco.h-its.org/exelixis/software.html). The user can also optionally specify the use of monomorphic sites (explored in Pavlidis et al. 2010), and can input parameters for an explicit demographic model to be used as the neutral SFS. SweepFinder requires a sufficiently SNP dense region in order to allow for accurate estimation, and the inclusion of a fraction of monomorphic sites evens out the SNP density as well as preserves the signature of low diversity in regions of depleted genetic variation (Pavlidis et al. 2010). Performance was evaluated with and without monomorphic sites.

OmegaPlus is a sliding-widow implementation of Kim and Nielsen's (2004) ω_{MAX} statistic that uses patterns of LD to identify selective sweeps (Alachiotis et al. 2012; http://sco.h-its.org/exelixis/software.html). It scans for windows of SNPs where

there is increased LD flanking the fixation, and reduced LD across the fixation. Like SweeD, OmegaPlus is a high performance statistic capable of analyzing very large datasets.

Finally, we evaluated iHS as a second LD-based selection estimator (Voight et al. 2006, http://coruscant.itmat.upenn.edu/software.html). This is based on the EHH statistic, which measures the decay of LD from an individual SNP (Sabeti et al. 2002). Longer haplotypes will be observed when a SNP rises faster in frequency than would be expected under neutral conditions. iHS additionally looks at the LD decay of both the derived and ancestral state of each SNP, calculates EHH for both alleles, and then integrates the area between the two curves; the notion being that this area will be larger for a selected allele vs. a neutral allele. Because of the normalization step required for raw iHS scores, a large SNP dataset is necessary.

Determining significance, and the effects of misspecification of the null

To determine the significance thresholds for SweepFinder, SweeD, and OmegaPlus, we simulated a range of neutral models in ms (Hudson 2002) using the –s option to fix the number of segregating sites. After performing each test on this neutral set of models we determined the maximum value for each of 1000 iterations and used the 95^{th} percentile as the cutoff value. The empirical models were then binned according to their average number of segregating sites and the 95^{th} percentile value was used for each bin as a cutoff for significant test values.

Next we verified that the distribution of test values in the cutoff models appropriately matched the values in the equilibrium neutral models. We observed that the distribution of values in the RHH models were a poor match for the values obtained by running SweepFinder on the neutral models simulated in ms (Figure 4.1a). However, sfscode samples 2 haplotypes from 20 individuals (producing a sample size of 40), while ms samples 40 haplotypes from a diploid population (from separate individuals). Thus, a sample size correction is necessary for proper comparison (Figure 4.1b).

Determining Threshold for Significant Sweeps in iHS

The statistic iHS computes a test score for each SNP within a locus, whereas all previously mentioned statistics compute a test value at specific points across a user-specified grid. Since iHS requires a normalization step to control for SNPs at different frequencies, we followed a slightly different procedure to determine significance values for this test. Raw iHS scores were normalized according to the method described in Voight et al. (2006). Briefly, all SNPs across each dataset were binned according to frequency. The mean and standard deviation of each bin was calculated, and these values were used to normalize raw iHS scores in the following way: for each SNP, the mean of the corresponding bin is subtracted from the raw iHS score and this result is divided by the standard deviation. This produces iHS values with a mean of approximately 0 and variance 1 for each frequency such that all SNPs can be compared directly (Voight et al. 2006).

With iHS, extreme negative values indicate a derived allele on a long haplotype, indicative of a selective sweep, and extreme positive values belong to a long ancestral haplotype. For this reason, the 1st percentiles were used to determine the significant values for the entire dataset.

Results & Discussion

SFS-Based Statistics Perform Poorly Under Recurrent Hitchhiking Models

For equilibrium neutral models our initial false positive rates for SweepFinder approached 0.30. After correcting for the sample size as described above, the false positive rates were lowered to below 0.05, which is equivalent to a p-value of 0.05 (Table 4.1). However, this correction has the unfortunate property of lowering the rejection rate of SweepFinder for equilibrium selection as well. For 2*Ns* ranging from 10 to 1000, the true positive rate for SweepFinder and SweeD is also under 0.05 (i.e., the same rejection rate as neutral models; Table 4.2). OmegaPlus is the only statistic that has power to reject neutrality as the strength of selection is increased, with a true positive rate as high as 0.44. When the probability that a new mutation is affected by selection is increased, this reduces the rejection rate of OmegaPlus, which is consistent with fewer rejections at a lower SNP density (Table 4.1), and consistent with the poor performance of this LD-based approach under recurrent hitchhiking models (Jensen et al. 2007).

Our bottleneck models consist of a severity in reduction ranging from 25% to 99%, and duration ranging from 1000 to 4000 generations. Since sfscode is a forward

simulator (and the reduction in population size begins at time 0), a longer duration is equivalent to a more recent recovery, whereas a shorter duration corresponds to an older bottleneck. In neutral bottleneck models, SweepFinder and SweeD have low power to reject for all parameter combinations (Table 4.3). OmegaPlus has a low false positive rate when the population size reduction is small, but for a 99% reduction, the rate of rejection is the same as for equilibrium selection models – suggesting an inability to distinguish these two scenarios. This is true for all duration times but is more pronounced as the recovery time decreases, with a false positive as high as 0.91 for a 99% reduction in population size that recovered only 1000 generations ago (Table 4.3). Thus, severe population size reductions can mimic this pattern of LD normally attributed to selective sweeps, consistent with previous results (Pavlidis et al. 2010).

When a bottleneck is combined with strong selection ($2Ns = 1000$), SweepFinder shows a slightly improved propensity to reject the neutral model (Table 4.4), but this is more likely due to the fact that SweepFinder has reduced sensitivity when SNP density is low, and combining strong selection with a bottleneck exaggerates this effect. OmegaPlus has a higher rejection rate at $2Ns = 100$, which is also likely due to the extreme reduction in genetic variation caused by combining strong selection with a bottleneck (Table 4.4). For non-equilibrium selection models the rate of rejection for OmegaPlus is within the same range as equilibrium selection models, which suggests that it is not capable of distinguishing selection from a bottleneck when both factors have impacted patterns of variation.

Single Hitchhiking Models

We included single hitchhiking models specifically to satisfy the sweep conditions for which SweepFinder was designed, namely that a single sweep has fixed at the time of sampling. For equilibrium selection with $2Ns = 1000$ the true positive rate for SweepFinder and SweeD is between 0.32 and 0.34 (Table 4.5), while the true positive rate for OmegaPlus is 0.46. SweepFinder's ability to reject neutrality is improved for equilibrium selection under the single hitchhiking model when selection is strong, while the performance of OmegaPlus remains constant. OmegaPlus also remains sensitive to moderate selection strengths, as the true positive rate for $2Ns = 100$ is 0.37. Thus, LD-based approaches appear to outperform SFS-based approaches in this parameter space.

Joint selection and bottleneck models follow a similar trend as previous models, with OmegaPlus being the only statistic with power to reject neutrality. The difference between the RHH and SHH joint models is that in RHH, the rejection rate is fairly uniform across all severities and recovery times. For the SHH models, a pattern similar to the neutral bottlenecks is observed, where the rejection rate is higher for more severe and recently recovered bottlenecks (Table 4.6). One reason for this uniformity when recurrent hitchhiking is combined with various bottleneck scenarios is that multiple beneficial haplotypes are amplified during the bottleneck recovery phase. The SHH models on the other hand, experience only a single selected mutation, and thus the underlying coalescent trees are primarily shaped by the

demographic history of the population. Therefore, the demographic model determines the length of the tree, and the beneficial mutation will be at varying frequencies in the population at the recovery time, depending on the demographic model.

iHS Genome-wide Approach to Detect Significant Sweeps

For iHS we initially attempted the above criteria to determine significance. However, all values were too large to afford any power for iHS to reject the neutral model. This may owe to the unique signal that iHS is trying to summarize, computing a score for each SNP instead of across an equally spaced grid. Extreme significant values are expected to occur in neutral haplotypes, but they appear more uniformly distributed than in a suspected sweep (Voight et al. 2006). This means that there is some requirement for extreme values to be clustered for a sweep, in order to distinguish a significant value left by a selective event from a random significant value. Thus, by binning by the number of segregating sites and using a neutral model to determine the cutoff, the extreme values of the neutral model may not be an accurate estimation of these clusters of SNPs left by a sweep.

For this reason we used a significance value derived from the entire dataset, following Voight et al (2006). As they point out, this method can be useful for identifying regions of interest but does not serve as a formal significance test. To define a selective sweep signal we considered the top and bottom 1% of all iHS values as our cutoffs, and then searched for instances where iHS scores greater than or equal to these values occurred consecutively at 2 or more neighboring SNPs. In

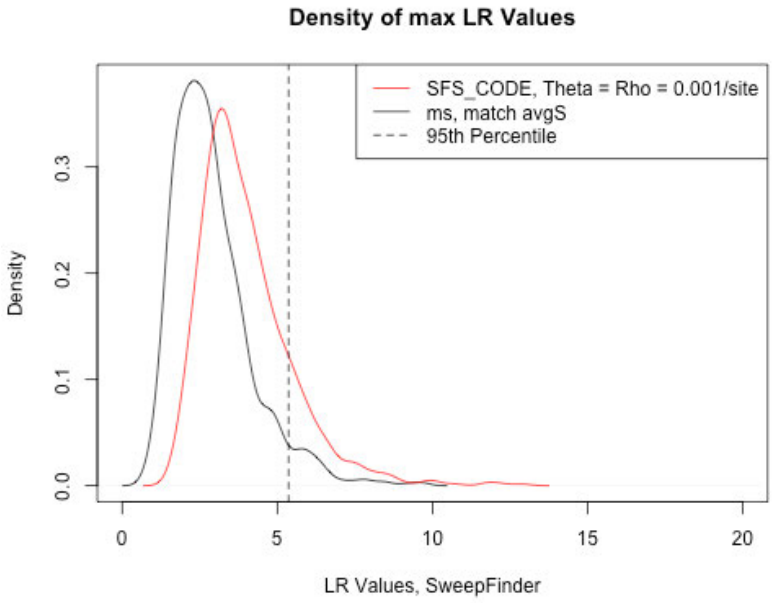
order to determine if the iHS test statistic is capable of distinguishing between a selective event and a bottleneck, we compared the fraction of sequences that contained a sweep in each model type (Figure 4.2). It is important to note, however, that iHS has a dependency on SNP dense sequences in order to be able to calculate an iHS score. For this reason, a number of replicates for 2Ns = 1000 were excluded from the recurrent hitchhiking dataset as, owing to the low SNP density, iHS was unable to calculate a value. It is also important to consider that for both SHH and RHH a majority of the sequences that contain sweep signals are from the models with weak selection (2Ns = 10) – again owing to the issue of SNP density. In fact, across both selection and neutral non-equilibrium models, SNP density is the main determinant of rejection (thus of both true and false positive rates).

Summary & Conclusions

For the models considered here, SweepFinder and iHS had the highest type II error. For both statistics, this is likely due to their dependence on SNP density, where a higher SNP density lends more power to the statistic. Thus, the lack of power under diversity reducing models (like positive selection and population bottlenecks) led to a reduced ability to reject the neutral model regardless of the presence or absence of selection. OmegaPlus showed the most sensitivity to the various model parameters, with the highest true positive rates for both RHH and SHH selection. This statistic has difficulty distinguishing selection from a severe bottleneck however, and in RHH models with joint selection and demography, the true positive rate was uniform across

all bottlenecks and within the range of true positives for equilibrium sweeps. These results emphasize the need to develop statistics that are more accurate in their identification of selective events. Many natural populations are characterized by non-equilibrium histories, and the commonly0used methods evaluated here are unable to deal with this effectively. However, these results also represent an important and well-quantified challenge to the field – and the performance of these statistics and the chosen parameter space can serve as a useful benchmark for future method development.

a)



b)

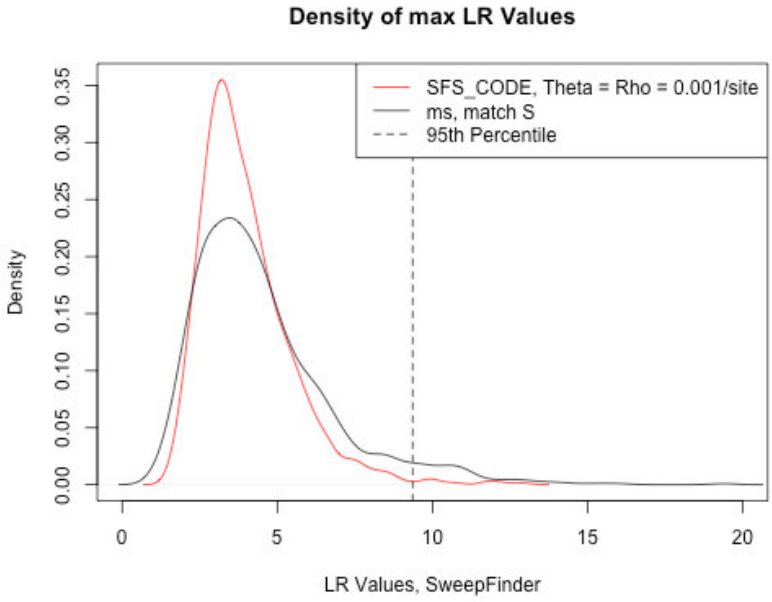


Figure 4.1. Correction for model misspecification.
Density plots for maximum likelihood ratio values for 1000 iteration of a neutral model. For sfs_code, theta = rho = 0.001 per site (red line). The same model was simulated in ms using the –s option to match the average number of segregating sites for the sfs_code model. The 95th percentile is for the ms model. A) sample size = 40, false positive rate = 0.15. B) sample size = 20, false positive rate = 0.01

Table 4.1. False Positive Rate for Equilibrium Neutral Models

	per site $\theta = \rho$									
	0.001	0.0009	0.0008	0.0007	0.0006	0.0005	0.0004	0.0003	0.0002	0.0001
SweepFinder, n = 20	0.01	0.02	0.01	0.02	0.03	0.02	0.03	0.11	0.18	0.15
SweepFinder, n = 40	0.15	0.09	0.16	0.23	0.22	0.29	0.30	0.30	0.29	0.27
SweeD	0.05	0.05	0.04	0.07	0.07	0.06	0.07	0.04	0.05	0.03
SweeD with monomorphic	0.11	0.12	0.10	0.10	0.10	0.10	0.14	0.12	0.08	0.08
OmegaPlus	0.05	0.06	0.06	0.07	0.07	0.06	0.05	0.05	0.07	0.05

Table 4.2. True Positive Rate for Equilibrium RHH Models

	2Ns = 10				2Ns = 100				2Ns = 1000			
P(sel)	0.002	0.01	0.1	0.25	0.002	0.01	0.1	0.25	0.002	0.01	0.1	0.25
SweepFinder	0.01	0.01	0.00	0.01	0.01	0.03	0.13	0.05	0.05	0.14	0.11	0.16
SweeD	0.03	0.03	0.02	0.03	0.02	0.03	0.05	0.04	0.01	0.06	0.03	0.04
SweeD with monomorphic	0.05	0.05	0.02	0.02	0.01	0.01	0.04	0.04	0.02	0.04	0.05	0.08
OmegaPlus	0.03	0.05	0.09	0.11	0.26	0.44	0.35	0.30	0.44	0.27	0.11	0.20

Table 4.3. False Positive Rate for Neutral Bottleneck Models (sfscode)

	Bottleneck Duration (generations)									
	4500					4000				
Reduction (%)	25	50	75	90	99	25	50	75	90	99
SweepFinder	0.01	0.02	0.08	0.08	0.01	0.02	0.03	0.04	0.05	0.01
SweeD	0.07	0.09	0.13	0.07	0.01	0.06	0.09	0.09	0.07	0.00
SweeD with monomorphic	0.13	0.19	0.27	0.18	0.01	0.15	0.18	0.17	0.16	0.00
OmegaPlus	0.07	0.13	0.26	0.31	0.68	0.09	0.13	0.26	0.34	0.91

	Bottleneck Duration (generations)									
	3000					1000				
Reduction (%)	25	50	75	90	99	25	50	75	90	99
SweepFinder	0.01	0.01	0.04	0.04	0.00	0.01	0.01	0.02	0.02	0.00
SweeD	0.05	0.09	0.11	0.08	0.00	0.04	0.04	0.07	0.07	0.00
SweeD with monomorphic	0.12	0.16	0.13	0.11	0.00	0.09	0.12	0.11	0.12	0.01
OmegaPlus	0.08	0.12	0.22	0.43	0.79	0.05	0.05	0.12	0.19	0.40

Table 4.4. Rejections of the Neutral Model for Joint RHH-Bottleneck Models

2Ns = 100, P(sel) = 0.01

	Bottleneck Duration (generations)														
	4000					3000					1000				
Reduction (%)	25	50	75	90	99	25	50	75	90	99	25	50	75	90	99
SweepFinder	0.05	0.04	0.02	0.04	0.01	0.05	0.05	0.03	0.03	0.05	0.05	0.05	0.04	0.04	0.04
SweeD	0.04	0.03	0.02	0.02	0.01	0.05	0.04	0.03	0.02	0.02	0.04	0.04	0.04	0.04	0.03
SweeD with monomorphic	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.02	0.01
OmegaPlus	0.48	0.46	0.40	0.42	0.84	0.47	0.44	0.42	0.47	0.59	0.46	0.49	0.46	0.43	0.47

2Ns = 1000, P(sel) = 0.01

	Bottleneck Duration (generations)														
	4000					3000					1000				
Reduction (%)	25	50	75	90	99	25	50	75	90	99	25	50	75	90	99
SweepFinder	0.15	0.14	0.15	0.15	0.17	0.15	0.17	0.15	0.18	0.16	0.16	0.17	0.16	0.14	0.17
SweeD	0.07	0.05	0.05	0.06	0.08	0.06	0.07	0.05	0.06	0.06	0.06	0.07	0.07	0.07	0.07
SweeD with monomorphic	0.05	0.04	0.04	0.05	0.06	0.04	0.06	0.04	0.05	0.04	0.04	0.05	0.06	0.05	0.05
OmegaPlus	0.26	0.23	0.16	0.27	0.28	0.18	0.27	0.26	0.26	0.26	0.28	0.24	0.26	0.28	0.27

Table 4.5. True Positive Rate for SHH Selection Models

	2Ns	10	100	1000
SweepFinder		0.05	0.14	0.33
SweeD		0.05	0.13	0.32
SweeD with monomorphic		0.12	0.15	0.34
OmegaPlus		0.07	0.37	0.46

Table 4.6. Rejections of the Neutral Model for Joint SHH-Bottleneck Models

2Ns = 100

	Bottleneck Duration (generations)											
	4500			4000			3000			1000		
Reduction (%)	50	75	90	50	75	90	50	75	90	50	75	90
SweepFinder	0.11	0.14	0.08	0.10	0.14	0.06	0.08	0.10	0.07	0.06	0.05	0.08
SweeD	0.09	0.12	0.07	0.09	0.13	0.05	0.08	0.09	0.07	0.06	0.05	0.07
SweeD with monomorphic	0.18	0.26	0.17	0.20	0.21	0.14	0.15	0.14	0.10	0.12	0.13	0.13
OmegaPlus	0.09	0.29	0.32	0.15	0.26	0.41	0.12	0.24	0.51	0.07	0.09	0.21

2Ns = 1000

	Bottleneck Duration (generations)											
	4500			4000			3000			1000		
Reduction (%)	50	75	90	50	75	90	50	75	90	50	75	90
SweepFinder	0.13	0.12	0.07	0.10	0.16	0.07	0.07	0.10	0.06	0.07	0.06	0.09
SweeD	0.12	0.11	0.07	0.09	0.13	0.06	0.07	0.09	0.05	0.06	0.05	0.08
SweeD with monomorphic	0.19	0.25	0.18	0.19	0.23	0.16	0.15	0.14	0.09	0.12	0.12	0.11
OmegaPlus	0.13	0.29	0.31	0.14	0.25	0.40	0.11	0.23	0.47	0.08	0.11	0.20

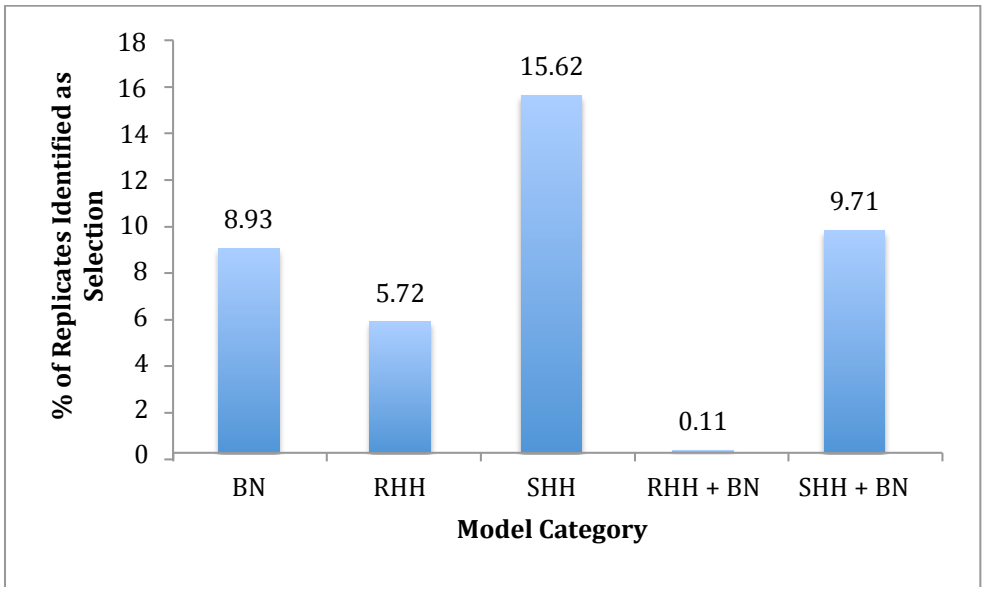


Figure 4.2. Percentage Of Sequences That Contain Selective Sweeps.
Selective sweep detection using iHS for five model categories: bottlenecks (BN), recurrent hitchhiking (RHH), single hitchhiking (SHH), and joint RHH and SHH bottleneck models. RHH models were simulated with sfcode, and SHH models were simulated with msms. These are the same models that were presented in Tables 4.1 – 4.6, and sequences with various selection and/or bottleneck parameters were pooled under each category. Percentages represent the number of replicates that were incorrectly identified as selection by iHS in each category. This plot suggests that iHS is more effective at identifying SHH events correctly, but actually many RHH replicates were eliminated due to low SNP density (see text).

CHAPTER V. Final Summary and Perspectives

The work in my dissertation is aimed at advancing the current knowledge of how positive selection has shaped human populations. To do this I investigated novel datasets that have the ability to provide a unique understanding of specific aspects of human evolution, i.e. genetic differences that set us apart from our most recent common ancestor, the Neanderthal, and unique changes in gene expression that have shaped evolution of the human brain. I also think it is important to validate the effectiveness of tests for selection that numerous scientists rely on for their ability to produce true signals of positive selection.

By checking overlap of putative sweep regions discovered with Neanderthal against methods encompassing a wider time frame, I was able to identify a set of regions that are uniquely important for differentiating modern humans – many of which contain biologically interesting genes including immune function, cognition and morphology. Obtaining the draft sequence of the Neanderthal genome was an exciting milestone for the field of human evolution, and now we have definitive evidence that using it as a tool to identify positive selection in humans can give us information about our evolutionary history that would otherwise have been missed using current methods. With improved technology to sequence high quality ancient genomes, such as the Denisovan individual, these same methods will become essential in giving us a clearer picture of the genetic differences that are unique to modern humans. This, combined with the recently released 1000 genomes dataset can provide an improved list of essential ancient sweeps – both

because the Denisovan genome is of much higher quality and coverage, and the 1000 genomes data can provide population data to compare against an ancestor, instead of the 5 single individuals that were used for the Neanderthal analysis. The authors chose 5 individuals of different ancestry in order to be able to say something about the relatedness of Neanderthal to humans from different geographical location. But, by using a population of Yorubans, for instance – which contain a higher level of genetic variation than European or Asian populations – we can identify more ancient sweeps that are important to human evolution. In time, we will be able to provide answers about our early origins that led to the intellectual differences, which set us apart from other great apes.

Perhaps the best example to date that attempts to link positive selection to evolution of human brain-specific differences is the multi-species primate epigenetic dataset presented here. Although a majority of the H3K4me3 peak sequences examined do not show direct evidence of positive selection, as no examples of strong selective sweeps could be detected, we can make some important conclusions about the evolution of human-specific epigenetic signatures involved in gene expression. It is likely that signals of positive selection are more evident in epigenetic modifications that are active during early developmental neuronal processes, since modifications differ greatly between different age groups. It is also possible that selection has influenced trans-acting factors that are responsible for H3K4me3 modification in neuronal cells. We do show that human neuronal peaks have an increased number of human-specific sequence alterations when compared to non-neuronal peaks. This, combined with the finding that several sequences show a significant increase in nucleotide substitution rates would

suggest that there is some link between genetic and epigenetic evolution within these peak regions, although positive selection has not been a major force in shaping this evolution.

As we continue our search for positive selection in human populations it is important to reevaluate the most commonly used selection estimators for robustness in a standardized manner. Many of these tests are released with performance testing for only a select number of models, and some without testing models that violate equilibrium assumptions. The last point is essential since most natural populations fall into this category. My work in this area has several important implications. First, it highlights the need for more robust selection estimators capable of distinguishing between selection and a more extreme bottleneck. Secondly, since my work was performed on simulated data where I knew the exact parameters of the demographic model, the finding that estimators do not distinguish between severe recent bottlenecks and selection points to the need for fitting a demographic model to any natural population in order to correctly identify selection.

Currently, many methods that estimate demographic parameters assume that the population is under neutral conditions. This is dangerous, as they are making use of the same information about genetic variation as selection estimators, e.g. the site frequency spectrum, and can lead to over fitting of parameters, such as a low estimation for migration rates and higher estimates in the size of population bottlenecks. Also, demographic estimators differ in the types of model they consider. For instance some estimators are more diverse in the types of parameters they fit so that you can build your

own model using migration, population size changes, and sub-divisions (Gutenkunst et al. 2009, Naduvilezhath et al. 2011) while others only consider one model with a few parameters to specify, such as 2-populaiton isolation-migration where only a split time and migration rate is considered (Nielsen and Wakeley 2001, Becquet and Przeworski 2007). All of these implementations are going to be limited by the assumptions they make about population parameters such as rate and strength of selection, recombination, and mating paradigms.

It would be nice to distill all these different methods for estimating selection and demography into one method that can jointly estimate the two processes. In order to do this it is necessary to determine what works well and what doesn't. I have started answering this question by evaluating the performance of selection estimators and trying to identify which ones will be most robust to demography. Thus far, it would seem that using LD can provide a lot more power for identifying selection correctly. The other half of this story is determining in the same standardized sort of manner which demographic estimators are most effective at correctly estimating demographic model parameters (Poh et al., in progress). This information will hopefully lead to a joint estimator that uses the most robust estimation methods to jointly estimate selection and demography in a way that will work for a large variety of datasets.

By examining the latest datasets and techniques that are being developed with respect to identifying positive selection in humans, we are gaining a better understanding of the amount of selection in the human genome, as well as the types of genetic and epigenetic changes that characterizes us as a species. The continued improvement of all

of the methods discussed in this dissertation will one day lead to a clearer picture of the processes that have shaped human evolution.

APPENDIX I. Supplementary Methods

Sample preparation (ChIP-seq and RNA-seq)

Ethics Statement: All work presented here was conducted on brain specimens collected after death. Cause of death was unrelated to the present study. All sample acquisition and processing of postmortem brain tissue was approved by the Institutional Review Boards of the participating institutions.

ChIP-seq: Procedures for extraction and sorting of NeuN+ neuronal nuclei from the cortical gray matter, and subsequent chromatin immunoprecipitation with anti-H3K4me3 antibody and ChIP-seq library preparation were recently described (Cheung et al. 2010, Connor et al. 2010, Jiang et al. 2008, Matevossian and Akbarian 2008). Cross-immunoreactivity of the anti-H3K4me3 antibody with other histone methylation forms, including mono- and di-H3K4 (H3K4me1/2) was controlled by dot blots and synthetic blocking peptides as described (Connor et al. 2010). The human ChIP-seq data sets, generated from neuronal nuclei of the pole of the frontal lobe, were published previously (Cheung et al. 2010, Shulha et al. 2012).

Postmortem brain tissue from the pole of the frontal lobe of 4 adult chimpanzees, ranging in age from 27-44 years, and of 3 adult macaque monkeys 11 years or older (**Table S1**) was processed in the same manner as previously described for the human specimens (Cheung et al. 2010, Connor et al. 2010, Jiang et al. 2008, Matevossian and Akbarian 2008). Specimens were obtained from the dorsolateral portion of the prefrontal cortex, primarily from cytoarchitectonic (Brodmann) Area 10 (BA10) and regions that border on BA10, including portions of BA9 and BA46. The quality of ChIP-seq datasets

was similar across all samples, with the total number of reads in the range of 2-10^6, of which typically 70-80% were derived from uniquely mappable sequences of the reference genome (HG19 or panTro2 or rheMac2) (**Table S1**).

RNA-seq: Three human specimens with no evidence for neurological disease or neurodegeneration were obtained from the Harvard Brain Tissue Resource Center in Belmont, MA (age 69-70, postmortem interval 15-26 hrs, all male). Rostral prefrontal cortex was processed for RNA-seq using Illumina's *mRNA-seq sample preparation kit*. Briefly, total RNAs were isolated using Trizol isolation kit and ploy-A containing RNAs were purified using poly-T oligo-attached magnetic beads. The mRNA was then fragmented into small pieces using divalent cations under elevated temperature and the cleaved RNA fragments were copied into first strand cDNA using reverse transcriptase and random primers. RNA integrity number for each sample was determined using the Agilent 2100 bioanalyzer; RIN was above 4.0 or all cases. Second strand cDNA was synthesized using DNA polymerase I and RNAseH and followed by poly "A" cloning and PCR amplification to create the final cDNA library. RNA-Seq data were generated on an Illumina Genome Analyzer IIx by single end sequencing with 35 nucleotide (nt) read length.

To further confirm species-specific differences, RNA from the frontal pole was isolated using the RNEasy Mini Kit (Qiagen, Valencia, CA). RNA concentrations were

Gene	NCBI Accession #	Forward Primer	Sequence	Reverse Primer	Sequence	Product Size (bp)
dpp10	AY172661	DPP10 Forward 98	GGGAATTGCTATTGCTCTGCTGGT	DPP10 Reverse 258	AGCCTCTGGATCGTGAAGCACAAA	160
18S	NR_003286.2	18S Forward 361	TCAACTTTCGATGGTAGTCGCCGT	18S Reverse 468	TCCTTGGATGTGGTAGCCGTTTCT	108
loc389023	NR_036580.1	loc389023 F 184	AATCCAGCCCAGATTCTCCTACCA	loc389023 R 290	TTGGGAAGGGCAGTCTGATTGAAG	107

equilibrated to 100 ng/uL, and qRT-PCR was performed using the Quantifast SYBR

Green RT-PCR kit on an AB7500 machine (Qiagen, Applied Biosystems, Carlsbad, CA) using primers shown here. Relative expression was determined using the Pfaffl method normalized to 18S and referenced to human expression. All products were sequenced for verification of specificity.

ChIP-seq analysis

Libraries were sequenced with the Illumina Genome Analyzer GAII, and images were first processed with GAPipeline (versions 1.0 and 1.4) and OLB (1.6). We performed single-end sequencing of 36bp reads. We used Bowtie (version 0.11.3) allowing up to one mismatch to map all sequence reads to the gender appropriate human genome HG19 and only retained the reads that mapped to one unique location in the genome in each sample for subsequent data analysis. Chimpanzee and macaque datasets were mapped to the appropriate genomes (rheMac2 and panTro2) and to human genome HG19 for comparison. **Table S1** online shows the sequencing statistics.

To calculate the table of Pearson correlations among H3K4me3 profiles in promoters (**Fig. 4.1A, Table S2**), the region within 2 KB of a transcriptional start site (TSS) was defined as the promoter of the TSS. If a gene has multiple TSSs, each TSS was accounted for separately. We used the RefSeq gene set from the UCSC genome browser, which contained 35,519 transcripts and 22,150 genes. ChrY was excluded from this analysis. Promoters for TSSs that were less than 2 KB apart were merged to avoid double counting. The number of tags within each promoter was tallied and divided by the size of the regions and the resulting tag densities for all annotated TSSs were used to compute Pearson correlation coefficients between each pair of samples.

In order to detect regions that were enriched in a neuron-specific manner with the H3K4me3 mark in human samples but not in chimpanzee or macaque, we first filtered a set of 34,639 peaks (without chr Y) obtained by running MACS on each of the 7 human adult samples against an input human sample (micrococcal nuclease digestion without anti-H3K4me3 antibody pull down) and taking the union of the 7 MACS outputs. The criteria for filtering are as follows: (i) the average tag density for all 11 human samples is higher than 0.01 and is more than 2 times greater than the average tag density for chimpanzee or macaque samples mapped to the human genome. (ii) The region is more than 500 bp long, and (iii) the region is detected as a peak in every human sample. We obtained 418 peaks after the filtering. To obtain human depleted peaks we used a reciprocal approach where initial peaks were detected in chimpanzee and macaque. This resulted in 63 peaks after filtering.

To evaluate significance of the 418 human-enriched and 63 human-depleted peaks, we applied Poisson statistics (the Poisson distribution has only one parameter, which is the mean called lambda). We compared human against chimp and macaque separately. For each peak, we computed the average reads in chimp and in macaque to determine the respective lambdas for the Poisson distributions. Then we computed a p-value for human vs. chimp and another p-value for human vs. macaque, using the average reads across all human samples within the peak (normalized by total reads within all annotated promoters in that sample). The Benjamini-Hochberg method was used to compute the false discovery rate (FDR) of each peak. Only the peaks with FDR<0.05 in

both human-chimp and human-macaque comparisons were kept, and there were 410 human-enriched peaks and 61 human depleted peaks with both FDRs<0.05.

Procedures similar to the ones described above were applied in order to detect chimpanzee-specific regions with significant 2-fold enrichment, or depletion, of H3K4me3.

In an additional independent analysis we probed the collection of 2,148 neuronal peaks not shared with lymphocytes under the most restrictive criteria, as described in our previous work (Cheung et al. 2010). We retained a peak if average tag density in human samples were more than 2 times greater than in both chimpanzee and macaque samples. It resulted in 33 peaks (**Table S10**), referred to hereafter and in the main manuscript as ^{neu}HP (Human-specific peak selectively enriched in neurons). To test if ^{neu}HP were identified because regions were unique to the human genome, we also mapped both macaque and chimp ChIP-seq readouts to their appropriate genomes, panTro2 and rheMac2. To allow direct comparison between peaks in HG19 and panTro2/rheMac2, we calculated tag densities normalized by sequencing depth (**Table S10**)**.**

To evaluate significance of the peaks detected we applied Poisson statistics. A set of appropriate monkey samples was used to assess background distribution. After that, a human sample with the lowest coverage at the particular peak was used to obtain the p-value.

Screenshots of ChIP-seq tracks in **Figure S1** online were normalized by the number of tags that map to 5397 orthologous promoters. Orthologous promoters were defined as +/-2kb regions around HG19 RefSeq TSSs that are uniquely lifted-over both

ways between any genomes (HG19/panTro2/rheMac2) with 95% identity. UCSC lift-over tool was used for the conversion between different genomes.

Overlap with DNA hypomethylated regions (HMR) in male germ cells

Approximately 76,000 DNA hypomethylated sequences (HMR) in human and 70,000 HMRs in chimpanzee sperm (Molaro et al. 2011) were screened for overlap with the 34,639 H3K4me3 peaks of the present study. Altogether 22,808/34,639 neuronal H3K4me3 peaks of the present study overlapped with HMRs in both human and chimpanzee sperm. A subset of 1992 H3K4me3 peak regions from PFC neurons specifically overlapped with sperm HMRs in human but not chimpanzee. Conversely, 669 human PFC neuron H3K4me3 peaks overlapped selectively with HMRs in chimpanzee but not human sperm. Next, 410 peaks out of the 25,469 peaks that overlap with sperm HMRs were picked randomly (10,000 times) and the following expected frequencies were found: human and chimpanzee sperm HMRs (H+/P+) 270/410 = 65.7%, no HMR (H-/P-), 108/410 = 26.6%, H+, P-, 24/410 = 5.7%, H-, P+, 8/410= 1.9%.

RNA-seq analysis

Using Tophat software, the first 40bp of each RNA-seq read was mapped into human, chimpanzee and macaque genomes (max. 1mismatch allowed for mapping into native genome and 2 mismatched when non-human was mapped to human. The original 418 hnp peaks were extended and clustered (united overlapped) inside of human genome. Alternatively, the 418 hnp were lifted-over to monkey, extended by 2kb and clustered. Data were expressed as (i) Chimp_HG19/Human_HG19 - sequencing depth normalized ratio; plus pseudocount, (ii) Macaque_HG19/Human_HG19 - sequencing

depth normalized ratio; plus pseudocount; (ii) Chimp_PT2/Human_HG19 - sequencing depth normalized ratio; plus pseudocount; plus normalization by region size; (iv) Macaque_RM2/Human_HG19 - sequencing depth normalized ratio; plus pseudocount; plus normalization by region size (**Table S18**).

Screenshots of RNA-seq track (**Figure S2**) were normalized by sequencing depth. Data for macaque RNA expression were downloaded as a wig file from GSE24538 (Liu et al. 2011). Data for chimpanzee RNA expression were downloaded from GSE30352 (samples "ptr br M 2,3,4") (Brawand et al. 2011). Three samples were pulled together and mapped to panTro2 genome allowing up to 2 mismatches.

Comparative analyses of human-specific alterations in Ensembl

Coordinates of human (neuron)-specific H3K4me3 peaks (referred to as hnp in the main manuscript) were converted to the Genome Reference Consortium (GCR)'s genome build Grch37 and Human Specific Alterations (HSA) were selected based on Ensembl EPO primate alignments. Altogether 1519 HSAs were identified (continuous indels in the region were considered as 1 HSA) for the subset of 33 ^{neu}HP with both species-specific and cell-type specific (present in neurons but not blood or non-neuronal brain cells.) From these, 915 were found to be conserved in primates tested (*Pan troglodytes, Gorilla gorilla, Pongo abelii, Macaca mulatta*) and 963 were located within large-scale regions with regulatory properties (Ensembl "Regulatory Build"). We downloaded *neanderthal* and *denisova* genomes from UCSC browser in bam alignment format and further checked 1519 HSAs in comparison to them.

The *neanderthal* genome (ftp://hgdownload.cse.ucsc.edu/gbdb/hg19/neandertal/seqAlis)

was in HG19 coordinates, *denisova* (ftp://hgdownload.cse.ucsc.edu/gbdb/hg18/denisova) in HG18.

The coordinates of HSAs were lifted-over with UCSC lift-over tool. The comparison to alignments was done with a set of scripts written in Perl. For denisova we found 52 (1353) HSAs that differ (similar) to human genome, for neanderthal - 20 (674) different (similar). For the remaining HSAs: 114 in case of comparison to denisova, and 674 - neanderthal we could not asses the allelic state due to incomplete coverage of archaic genomes.

Gel shift assays

Native gel electrophoretic mobility-shift assay with ^{32}P end-labeled DNA probes (1X or 50 nM) and Hela nuclear extract (20mg) for 20 minutes at 22^{o}C in binding buffer containing 10mM Tis-HCl (pH7.5), 50mM NaCl, 0.5mM DTT (1X). Probe is a 21-bp human or gorilla GATA-1 sequence duplex DNA (CCAGTAAGAA(A human/T other primate)GATTAGCCAG), non-specific probe is 5‘ ATTCGATCGGTTCGGGGCGAGC 3‘ sequence duplex DNA. To demonstrate the specificity, cold probe or non-specific duplex DNA was used at 400X/20,000nM over ^{32}P end-labeled DNA probes. All native gels were run at 150 V for 150 min in cold room in the presence of 6% glycerol, dried for 2 hours at 80^{o}C and exposed to X-ray film overnight at -70^{o}C. For binding stringency experiment same nuclear protein and probe concentrations were used with increased concentrations of sodium chloride in binding buffer (50, 100, 150 and 225 mM).

Chromosome conformation capture (3C)

To map physical interactions and loop formations between non-neighboring chromatin fragments, 1000mg of frontal pole tissue from 4 adult human specimens was used (male and female, 7 hrs autolysis interval (median), ranging in age from 30 to 70 years) in conjunction with our 3C protocol as described (Jiang et al. 2010) and 24 primers positioned on the *DPP10* (2q14.1) sense strand 5' to 3, and 8 primers for the 16p11.2 region.

Presence of physical interactions was determined by sequence-verified PCR product. Control PCRs included no input ('water') and also DNA from chromatin digested with Hind III but without the subsequent religation step ('no T4 ligase'). Additional 3C-qPCR reactions were performed using the QuantiTect Probe PCR Kit (Qiagen) and custom-made FAM-TAMRA taqman probes.

Human DPP10	chromosome conformation capture (3C) primers	
primer #	primer sequence	bp in chr2 (hg19)
1	CTCCGAGTCCCCAAACAGACTTGTTTGAAT	114960871-114960900
2	TGTAGAACAAGAGCCCAACAGTCAACATTT	114967422-114967451
3	CTCATCTCTCTGGGCTTGGTCATCTGTCTT	114969506-114969535
4	TTCAAACTGAAGCATCTCTTCTAGGGGTGC	114972863-114972892
5	TTGATTACAATCAGGCCGGATTGTCCACTC	114979317-114979346
6	TGCTGTGCTCCGAATCTGGAAGATTAATGG	115414659-115414688
7	CTGACACATTGAGCAACACCATGAGCAGAT	115415259-115415288
8	GATGGTTAGTTTTGCTCCAGTAGCCTGGAA	115423445-115423474
9	CTCTGTGGAATGTTCCCTCCTCTATTCTCT	115426353-115426382
10	CCAGTGTTCAAAAGCCTGAGACTATCCACT	115612001-115612030
11	GATTCTATTTGGCTGGGCGTTGTCTTAGGC	115623208-115623237
12	GATTTCCTCTTAACAGAGAACGTCCATCGG	115628389-115628418
13	CCCTGTGGAGCTTTTTCACCTCCATTTTAG	115637630-115637659
14	GATCCACAAATGCAGAACAACCCCCGATAT	115648085-115648114
15	CCCATCTATCAGCAATTCCCCTTACCCTCT	115911510-115911539
16	CAACTGAAATATACACATGGCACCTCGCAG	115921734-115921763
17	GACTGGGGCACTGAAAAGAAAGCTCTTCTA	115926328-115926357
18	GGGAGAGGCAACGTGATGTAAGACTGAGAA	115926666-115926695
19	TGGCAAGAGGTTACACGCAAAGCCATAATG	115929473-115929502
20	ATTGCTGCTGTTTAACGGGGACACAAATTC	116358450-116358479
21	CCTCCTGAGAGAGTAGCCCCAACTCTATTT	116361811-116361840
22	TTTCCTGGACTGATTATCAAGGAGCTGTCC	116364309-116364338
23	CTATGTCCACATGCAGAGGCAGAGGAAAAG	116372570-116372599
24	ATGACACATATCCCCCTTGGGAAATTGCTC	116374388-116374417
Human chr16	primer sequence	bp in chr16 (hg19)
1	AAATGTATCACAGTATCG	21511195-21511213
2	TCTTTGTCGGAATCCACTCGGTACACACAC	21518730-21518760
3	AGCTCAATTTAATCAACAGAACCGGGGGTT	21532483-21532514
4	TATGTGATCATCACTGCCCTACACC	22430178-22430203
5	CATCGCCACCCAGTAAACATAGTTACTGAT	22433264-22433294
6	GGGAGATATTTCTGACTTGAGACAATGCTATACTC	22446920-22446955
7	CAAGTGCTTTATTTCTTGGCTCTGGGGAGG	22450543-22450573
8	CAGAGGCTTCCTAGTTGGAAACACATTGTT	22463714-22463744
Macaque monkey	primer sequence	bp in chr13 (rm2)
M1	TCCCAACCAGTTACATCTCTGCTCAGACTC	120344263-120344292
M2	GAAGTGTTTGTATCTGTTGTCACGCAGCTC	120346568-120346597
M3	TTTTACGGGGAACAAAACTCGTCTCTCTGG	120365464-120365493
M4	GCGAAGGGTTTCTCTGGTCTACCTTTAGGT	120367862-120367891
M5	AATGAAAGTGTTGGCCAGAAAGGCCTAAGA	120369239-120369268
M6	GCTCTGCTGTGCTCCTAATCTGGAAGATTA	120806372-120806401
M7	ATATATGAGGCTGACACATTGAGCAACGCC	120806967-120806996
M8	CGGGTGGTTATTTTTGCTCCAGTAGCCTAG	120814240-120814269
M9	ATCTGTCTGTGCACACAAACAGAGAACCAG	120817252-120817281
M10	GCACCACACTAAGTGGTAGTAGGTCGGTAT	121289006-121289035
M11	AGCATCACAGTGAGGGTTTTTCAAGCACTC	121299269-121299298
M12	CTGTGTGTGCATGTGCACAAATTTACCTTG	121300506-121300535
M13	GCAAAAGAGTTATTCTTCTAGGTGCCCTTTGC	121301270-121301301
M14	GCCCAAAGCCATAATGCCTTTCTGAACATA	121306865-121306894
M15	TGTTACTATACTGCAGAGGGGAAGGACAAA	121730958-121730987
M16	GTTTACCAGGGACACCAGTTCGCATTTCTT	121756495-121756524
M17	GGGTGATGACAGTGGATTCTTAGGGTTGTC	121760129-121760158
M18	CAAGTAGCTGTCCTCTTTCATGGTCAGAGG	121762661-121762690
Human DPP10	chromosome conformation capture (3C) qPCR Probes	
probe #	FAM TAMRA probe sequence	bp in chr2 (hg19)
6	AATGGATTTTTGTCAACTTGAAACTCAAGC	115414889-115414919
8	AAAGCAGACTTGCCTATACTTGCAGTTCTG	115426412-115426442

Similar studies were conducted on prefrontal cortex from 3 adult macaques, using tissue from the right hemisphere (left hemisphere of the same animals was used for ChIP-seq). Another set of 18 macaque 3C primers was used to probe 3C in the macaque *DPP10* locus while primer pair 2/7 from human 16p11.2 was used to test the homologous sequences in the macaque. As an additional control, 3C assays were also performed with the H9 embryonic stem cell line. Embroid bodies were generated from colonies grown on feeder cells and grown in low-adherence flasks for 3 days until harvested or re-suspended in Neural Induction media and differentiated to a mixture of neural precursors and postmitotic cells, using a modified protocol (Li and Zhang 2006).

Loc389023 cloning, expression and RNA immunoprecipitation

Loc389023 expression in human brain samples and various cell lines was check either with nuclear only RNA or cytosolic enriched RNA. Two sets of primers (listed below) were used to confirm the expression. Resulted PCR products were further confirmed by sequencing. Full length Loc389023 RNA from human brain fetal nuclei was amplified using 3'end gene specific primer and cloned in pCDNA4A under CMV promoter. Loc389023 expression was verified in HEK293, Hela and SK-N-MC neural crest derived cells. RNA immuno-precipitation (RIP) was carried out as described (Zhao et al*.,* 2008). Briefly, after transfection, nuclei were isolated from SK-N-MC cells using ultracentrifugation. Nuclei pellet was resuspended in 1 ml ice cold lysis buffer (100mM KCl, 5mM MgCl2, 10mM HEPES, 0.5% NP40 along with RNAse inhibitor, lysates were mechanically lysed by passing through 27.5guage needle few times. Nuclear lysates were further diluted in 50mM Tris-HCL, 150mM NaCl and 1mM MgCl2 and pre-cleared

before incubating for 6hrs with respective antibodies (H3K4 07-736, IgG 12730: Upstate, SUZ12 3737, EZH2 4905: Cell signaling technologies). Input was saved for transfection efficiency analysis. Next day after pull down with protein G beads and several washes, RNA was isolated using Trizol reagent (Invitrogen) according to manufacturer's protocol. One step quantitative RT-PCR was performed using Quantifast SYBR kit (Qiagen). Primer sequences used for Loc389023 are 1. Left 5'*TCAACACTTGGAAGAAGGGAGCTG*3' Right 5'*GCCAGTACACCTTATTCTGACCCA*3'; 2. Left 5'*AATCCAGCCCAGATTCTCCTACCA*3' Right 5'*TTGGGAAGGGCAGTCTGATTGAAG*3'.

In Situ Hybridization

In situ hybridization 15 micron thick section from immersion-fixed human PFC specimens was performed as described previously (Mellios et al. 2008). DIG-labeled LNA oligonucleotide probes used are as follows: LOC89023 (5DigN/TTGGCTCACTCACTTACTTGCA/3Dig_N), Beta-actin (5DigN/CTCATTGTAGAAGGTGTGGTGCCA/3Dig_N) (Exiqon, Woburn, MA).

BIBLIOGRAPHY

Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19: 711–722.

Alachiotis N, Stamatakis a, Pavlidis P (2012) OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. Bioinformatics 28: 2274–2275.

Allen M, Heinzmann A, Noguchi E, Abecasis G, Broxholme J, et al. (2003) Positional cloning of a novel gene influencing asthma from chromosome 2q14. Nat Genet 35: 258–263.

Allison AC (1954) Protection afforded by sickle-cell trait against subtertian malareal infection. Brit Med J 1: 290–294.

Ayalew M, Le-Niculescu H, Levey DF, Jain N, Changala B, et al. (2012) Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. Mol Psychiatry 17: 887–905.

Babbitt CC, Fedrigo O, Pfefferle AD, Boyle AP, Horvath JE, et al. (2010) Both noncoding and protein-coding RNAs contribute to gene expression evolution in the primate brain. Genome Biol Evol 2: 67–79.

Barton NH (1998) The effect of hitch-hiking on neutral genealogies. Genet Res 72: 123–133.

Baudat F, Buard J, Grey C, Fledel-Alon A (2010) PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. Science 836.

Bayes A, van de Lagemaat LN, Collins MO, Croning MD, Whittle IR, et al. (2011) Characterization of the proteome, diseases and evolution of the human postsynaptic density. Nat Neurosci 14: 19–21.

Beall CM et al. 2010. Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. Proc Natl Acad Sci U S A. 107:11459–64.

Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. Nat Genet 36: 431–432.

Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. Gen Res 17: 1505–1519.

Belgard TG, Marques AC, Oliver PL, Abaan HO, Sirey TM, et al. (2011) A transcriptomic atlas of mouse neocortical layers. Neuron 71: 605–616.

Bernard A, Lubbers LS, Tanis KQ, Luo R, Podtelezhnikov AA, et al. (2012) Transcriptional architecture of the primate neocortex. Neuron 73: 1083–1099.

Bijlsma EK, Gijsbers AC, Schuurs-Hoeijmakers JH, van Haeringen A, Fransen van de Putte DE, et al. (2009) Extending the phenotype of recurrent rearrangements of 16p11.2: deletions in mentally retarded patients without autism and in normal individuals. Eur J Med Genet 52: 77–87.

Bond J, Woods CG (2006) Cytoskeletal genes regulating brain size. Curr Opin Cell Biol 18: 95–101.
Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, et al. (2011) The evolution of gene expression levels in mammalian organs. Nature 478: 343–348.
Butti C, Santos M, Uppal N, Hof PR (2011) Von Economo neurons: Clinical and evolutionary perspectives. Cortex E-pub ahead of print.
Caceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L, et al. (2003) Elevated gene expression levels distinguish human from non-human primate brains. Proc Natl Acad Sci U S A 100: 13030–13035.
Cai JJ, Macpherson JM, Sella G, Petrov DA (2009) Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet 5:e1000336.
Cheung I, Shulha HP, Jiang Y, Matevossian A, Wang J, et al. (2010) Developmental regulation and individual differences of neuronal H3K4me3 epigenomes in the prefrontal cortex. Proc Natl Acad Sci U S A 107: 8824–8829.
Connor C., et al. (2010) A simple method for improving the specificity of anti-methyl histone antibodies. Epigenetics 5(5): p. 392-5.
Coop G, Bullaughey K, Luca F, Przeworski Molly. (2008) The timing of selection at the human FOXP2 gene. Mol Biol Evol. 25:1257–9.
Crisci JL, Poh Y-P, Bean A, Simkin A, Jensen JD. (2012) Recent Progress in Polymorphism-Based Population Genetic Inference. J Hered. 103:287–296.
Crisci JL, Wong A, Good JM, Jensen JD. (2011) On Characterizing Adaptive Events Unique to Modern Humans. Gen Biol Evol. 3:791–798.
Danchin E, Charmantier A, Champagne FA, Mesoudi A, Pujol B, et al. (2011) Beyond DNA: integrating inclusive inheritance into an extended theory of evolution. Nat Rev Genet 12: 475–486.
Darwin CR. (1869) On the Origin of Species by Means of Natural Selection (fifth edition). John Murray, Albemarle Street, London, United Kingdom.
Day JJ, Sweatt JD (2012) Epigenetic treatments for cognitive impairments. Neuropsychopharmacology 37: 247–260.
Dekker J (2006) The three ‘C’ s of chromosome conformation capture: controls, controls, controls. Nat Methods 3: 17–21.
Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485: 376–380.
Djurovic S, Gustafsson O, Mattingsdal M, Athanasiu L, Bjella T, et al. (2010) A genome-wide association study of bipolar disorder in Norwegian individuals, followed by replication in Icelandic sample. J Affect Disord 126: 312–316.
Drummond AJ, et al. 2011. Geneious v5.4 [cited 2010 Mar 1]. Available from http://www.geneious.com/.
Durbin RM, Altshuler DL, Abecasis GR, Bentley DR, Chakravarti A, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
Elvin SJ, Williamson ED, Scott JC, Smith JN, de Lema GP, et al. (2004) Evolutionary genetics: Ambiguous role of CCR5 in Y. pestis infection. Nature 430: 417.

[a]Enard W, Przeworski M, Fisher SE, Lai CSL, Wiebe V, et al. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418: 869–872.
[b]Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, et al. (2002) Intra- and interspecific variation in primate gene expression patterns. Science 296: 340–343.
Enard W, Gehre S, Hammerschmidt K, Holter SM, Blass T, et al. (2009) A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. Cell 137: 961–971.
Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26: 2064–2065.
Favier B, Dollé P. (1997) Developmental functions of mammalian Hox genes. Mol Hum Reprod 3: 115–131.
Feinberg AP, Irizarry RA. (2010) Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc Natl Acad Sci U S A. 107 Suppl:1757–64.
Fernandez BA, Roberts W, Chung B, Weksberg R, Meyn S, et al. (2010) Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. J Med Genet 47: 195–203.
Fernandez T, Morgan T, Davis N, Klin A, Morris A, et al. (2004) Disruption of contactin 4 (CNTN4) results in developmental delay and other features of 3p deletion syndrome. Am J Hum Genet 74: 1286–1293.
Fischbach GD, Lord C (2010) The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron 68: 192–195.
Fisher RA. (1930) The Genetical Theory of Natural Selection. Oxford University Press, London, United Kingdom.
Fisher SE, Vargha-Khadem F, Watkins KE, Monaco AP, Pembrey ME, et al. (1998) Localisation of a gene implicated in a severe speech and language disorder. Nature Genetics 18: 168–170.
Fletcher W, Yang Z. (2010) The Effect of Insertions, Deletions and Alignment Errors on the Branch-Site Test of Positive Selection. Mol Biol Evol. 27:2257–67.
Galvani AP, Novembre J. (2005) The evolutionary history of the CCR5-Delta32 HIV-resistance mutation. Microbes Infect 7:302–9.
Garvie CW, Boss JM. (2008) Assembly of the RFX complex on the MHCII promoter: role of RFXAP and RFXB in relieving autoinhibition of RFX5. Biochim Biophys Acta. 1779:797–804.
Geoghegan JL, Spencer HG. (2011) Population-epigenetic models of selection. Theor Popul Biol 81:232–242.
Gertz J, Varley KE, Reddy TE, Bowling KM, Pauli F, et al. (2011) Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. PLoS Genet 7: e1002228
Gillespie J. 1977. Natural selection for variances in offspring numbers: a new evolutionary principle. Am Nat 111:1010–1014.

Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, et al. (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature 459: 569–573.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A draft sequence of the Neandertal genome. Science 328: 710–722.

Greer EL, Shi Y (2012) Histone methylation: a dynamic mark in health, disease and inheritance. Nat Rev Genet 13: 343–357.

Guenther MG, Jenner RG, Chevalier B, Nakamura T, Croce CM, et al. (2005) Global and Hox-specific roles for the MLL1 methyltransferase. Proc Natl Acad Sci U S A 102: 8603–8608.

Guilmatre A, Dubourg C, Mosca AL, Legallic S, Goldenberg A, et al. (2009) Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. Arch Gen Psychiatry 66: 947–956.

Gunz P, Neubauer S, Golovanova L, Doronichev V, Maureille B, et al. (2012) A uniquely modern human pattern of endocranial development. Insights from a new cranial reconstruction of the Neandertal newborn from Mezmaiskaya. J Hum Evol 62: 300–313.

Gutenkunst R, Hernandez RD, Williamson S, Bustamante C. D. (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5:e1000695.

Hämmerle B, Elizalde C, Galceran J, Becker W, Tejedor FJ. (2003) The MNB/DYRK1A protein kinase: neurobiological functions and Down syndrome implications. J Neural Transm Suppl. (67):129–137.

Hammoud SS, Nix DA, Zhang H, Purwar J, Carrell DT, et al. (2009) Distinctive chromatin in human sperm packages genes for embryo development. Nature 460: 473–478.

Haldane, J. (1927). A mathematical theory of natural and artificial selection, part v: Selection and mutation. PCPS-P Camb Philos S 23, 838-44.

Hashimoto R, Hashimoto H, Shintani N, Chiba S, Hattori S, et al. (2007) Pituitary adenylate cyclase-activating polypeptide is associated with schizophrenia. Mol Psychiatry 12: 1026–1032.

Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169: 2335–2352.

Hernandez, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24(23), 2786–7.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, et al. (2011) Classic selective sweeps were rare in recent human evolution. Science 331: 920–924.

Huang HS, Matevossian A, Jiang Y, Akbarian S (2006) Chromatin immunoprecipitation in postmortem brain. J Neurosci Methods 156: 284–292.

Hudson R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18:337–338.

Hughes AL, Nei M. (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci U S A. 86:958–62.

Hughes AL, Nei M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 335:167–70.

Ikeda M, Aleksic B, Kinoshita Y, Okochi T, Kawashima K, et al. (2011) Genome-wide association study of schizophrenia in a Japanese population. Biol Psychiatry 69: 472–478.

Isagawa T, Nagae G, Shiraki N, Fujita T, Sato N, et al. (2011) DNA methylation profiling of embryonic stem cell differentiation into the three germ layers. PLoS ONE 6 e26052:

Jablonka E, Lamb MJ, Avital E. (1998) "Lamarckian" mechanisms in darwinian evolution. Trends Ecol Evol 13:206–210.

Jaglin XH, Poirier K, Saillour Y, Buhler E, Tian G, et al. (2009) Mutations in the beta-tubulin gene TUBB2B result in asymmetrical polymicrogyria. Nat Genet 41: 746–752.

Jakovcevski M, Akbarian S (2012) Epigenetic mechanisms in neurological disease. Nat Med 18: 1194–1204.

Jensen JD, Kim Y. H., DuMont VB, Aquadro CF, Bustamante C. D. (2005) Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics. 170:1401–1410.

Jensen JD, Thornton K.R., Bustamante C. D., Aquadro CF. (2007) On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. Genetics. 176:2371.

Jiang, Y., et al. (2008) Isolation of neuronal chromatin from brain tissue. BMC Neurosci, 9: p. 42.

Jiang, Y., et al. (2010) Setdb1 histone methyltransferase regulates mood-related behaviors and expression of the NMDA receptor subunit NR2B. J Neurosci, 30(21): p. 7152-67.

Kanhere A, Viiri K, Araujo CC, Rasaiyaah J, Bouwman RD, et al. (2010) Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. Mol Cell 38: 675–688.

Kaplan NL, Hudson RR, Langley CH. (1989) The "hitchhiking effect" revisited. Genetics. 123:887–99.

Keightley PD, Eyre-Walker A. (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics. 177:2251–61.

Kim Y. H., Nielsen R. (2004) Linkage disequilibrium as a signature of selective sweeps. Genetics. 167:1513–1524.

Kim Y. H., Stephan W. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. 160:765–77.

Kimura M. (1968) Evolutionary rate at the molecular level. Nature. 217:624–626.

Kimura, M. (1983). Rare variant alleles in the light of the neutral theory. Mol Biol Evol, 1(1), 84–93.

Kimura R, Fujimoto A, Tokunaga K, Ohashi J (2007) A practical genome scan for population-specific strong selective sweeps that have reached fixation. PLoS One 2: e286 doi:10.1371/journal.pone.0000286.

King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188: 107–116.

Konopka G, Bomar JM, Winden K, Coppola G, Jonsson ZO, et al. (2009) Human-specific transcriptional regulation of CNS development genes by FOXP2. Nature 462: 213–217.

Konopka G, Friedrich T, Davis-Turak J, Winden K, Oldham MC, et al. (2012) Human-specific transcriptional networks in the brain. Neuron 75: 601–617.

Kosiol C, Holmes I, Goldman N. (2007) An empirical codon model for protein sequence evolution. Mol Biol Evol. 24:1464–79.

Krause J, Lalueza-Fox C, Orlando L, Enard W, Green RE, et al. (2007) The derived FOXP2 variant of modern humans was shared with Neandertals. Curr Biol 17: 1908–1912.

Kreitman M. (1996) The neutral theory is dead. Long live the neutral theory. In: BioEssays : news and reviews in molecular. Cell Dev Biol 18:678–83

Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, et al. (2008) Recurrent 16p11.2 microdeletions in autism. Hum Mol Genet 17: 628–638.

Kwiatkowski DP. (2005) How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet 77:171–92.

Lai CS, Fisher S E, Hurst J a, Vargha-Khadem F, Monaco a P. (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. Nature. 413:519–23.

Lalani AS, Masters J, Zeng W, Barrett J, Pannu R, et al. (1999) Use of Chemokine Receptors by Poxviruses Use of Chemokine Poxviruses Receptors by. Science 286: 1968–1971.

Lambert N, Lambot MA, Bilheu A, Albert V, Englert Y, et al. (2011) Genes expressed in specific areas of the human fetal cerebral cortex display distinct patterns of evolution. PLoS One 6: e17753

Lander ES, Linton LM, Birren BW, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Lane RF, Raines SM, Steele JW, Ehrlich ME, Lah JA, et al. (2010) Diabetes-associated SorCS1 regulates Alzheimer's amyloid-beta metabolism: evidence for involvement of SorL1 and the retromer complex. J Neurosci 30: 13110–13115.

Lesch KP, Selch S, Renner TJ, Jacob C, Nguyen TT, et al. (2011) Genome-wide copy number variation analysis in attention-deficit/hyperactivity disorder: association with neuropeptide Y gene dosage in an extended pedigree. Mol Psychiatry 16: 491–503.

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148: 84–98.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Li XJ, Zhang SC. (2006) In vitro differentiation of neural precursors from human embryonic stem cells. Methods Mol Biol, 331: p. 169-77.

Lionel AC, Crosbie J, Barbosa N, Goodale T, Thiruvahindrapuram B, et al. (2011) Rare copy number variation discovery and cross-disorder comparisons identify risk genes for ADHD. Sci Transl Med 3: 95ra75.

Liu H, Heath SC, Sobin C, Roos JL, Galke BL, et al. (2002) Genetic variation at the 22q11 PRODH2/DGCR6 locus presents an unusual pattern and increases susceptibility to schizophrenia. Proc Natl Acad Sci U S A 99: 3717–3722.

Liu Y, Han D, Han Y, Yan Z, Xie B, et al. (2011) Ab initio identification of transcription start sites in the Rhesus macaque genome by histone modification and RNA-Seq. Nucleic Acids Res 39: 1408–1418.

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth L V., et al. (2011) Comparative and demographic analysis of orangutan genomes. Nature 469: 529–533.

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A. 102:10557–62.

Maffie J, Rudy B (2008) Weighing the evidence for a ternary protein complex mediating A-type K+ currents in neurons. J Physiol 586: 5609–5623..

Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, et al. (2008) Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet 82: 477–488.

Matevossian A, Akbarian S. (2008) Neuronal nuclei isolation from human postmortem brain tissue. J Vis Exp 2008(20).

Maynard Smith JM, Haigh J, Smith JM. (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23–35.

McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, et al. (2009) Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet 41: 1223–1227.

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

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, et al. (2011) Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471: 216–219.

Mecsas J, Franklin G, Kuziel W, Brubaker R. (2004) Evolutionary genetics: CCR5 mutation and plague protection. Nature. 427:606.

Mellios N, et al. (2008) A set of differentially expressed miRNAs, including miR-30a-5p, act as post-transcriptional inhibitors of BDNF in prefrontal cortex. Hum Mol Genet, 17(19): p. 3030-42.

Miller LH, Mason SJ, Dvorak JA, McGinniss MH, Rothman IK. (1975) Erythrocyte receptors for (Plasmodium knowlesi) malaria: Duffy blood group determinants. Science. 189:561–562.

Molaro A, Hodges E, Fang F, Song Q, McCombie WR, et al. (2011) Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell 146: 1029–1041.

Mundlos S, Otto F, Mundlos C, Mulliken JB, Aylsworth a S, et al. (1997) Mutations involving the transcription factor CBFA1 cause cleidocranial dysplasia. Cell 89: 773–779.

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, et al. (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 327: 876–879.

Naduvilezhath L, Rose LE, Metzler D (2011) Jaatha: a fast composite-likelihood approach to estimate demographic parameters. Mol Ecol 20: 2709–2723.

Nasir J, Frima N, Pickard B, Malloy MP, Zhan L, et al. (2006) Unbalanced whole arm translocation resulting in loss of 18p in dystonia. Mov Disord 21: 859–863.

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485: 242–245.

Nei M, Gojobori T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–26.

Nielsen R, Williamson S, Kim YH, Hubisz MJ, Clark AG, Bustamante CD. (2005). Genomic scans for selective sweeps using SNP data. Gen Res 15(11), 1566–1575.

Nielsen R. (2005). Molecular signatures of natural selection. Annu Rev Genet 39:197–218.

Nielsen R, Wakeley J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158(2):885–96.

Nielsen R, Signorovitch J. (2003). Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. Theor Popul Biol 63(3), 245–255.

Pavlidis P, Jensen JD, Stephan W. (2010) Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics. 185:907–922.

Pavlidis P, Jensen JD, Stephan W, Stamatakis A. (2012). A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. Mol Biol Evol, 29(10), 3237–48.

Pavlidis P, Zivkovic D, Stamatakis A, Alachiotis N. (2013). SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. Mol Biol Evol, 1–41.

Ohta T. (1973) Slightly deleterious mutant substitutions in evolution. Nature. 246:96–97.

Orr HA, Betancourt AJ. (2001) Haldane's sieve and adaptation from the standing genetic variation. Genetics. 157:875–884.

Patel SA, Simon MC. (2008) Biology of hypoxia-inducible factor-2alpha in development and disease. Cell Death Differ 15:628–34.

Peng Y, Yang Z, Zhang H, Cui C, Qi X, et al. (2011) Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. Mol Biol Evol 28: 1075–1081.

Pennings PS, Hermisson J (2006) Soft sweeps II--molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol 23: 1076–1084.

Pitman EJG (1937) Significance tests which may be applied to samples from any population. J R Stat Soc Suppl 4: 119–130.

Pollard KS, Salama SR, King B, Kern AD, Dreszer T, et al. (2006) Forces shaping the fastest evolving regions in the human genome. PLoS Genet 2: e168

Preuss TM, Caceres M, Oldham MC, Geschwind DH (2004) Human brain evolution: insights from microarrays. Nat Rev Genet 5: 850–860.

Pritchard JK, Pickrell JK. (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20:R208–R215.

Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, et al. (2012) The bonobo genome compared with the chimpanzee and human genomes. Nature 486: 527–531.

Przeworski M. (2002) The signature of positive selection at randomly chosen loci. Genetics. 160:1179–1189.

Przeworski M., Coop G, Wall JD. (2005) The Signature of Positive Selection on Standing Genetic Variation. Evolution 59:2312–2323.

Reich D, Green RE, Kircher M, Krause J, Patterson N, et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468: 1053–1060.

Reitz C, Tokuhiro S, Clark LN, Conrad C, Vonsattel JP, et al. (2011) SORCS1 alters amyloid precursor protein processing and variants may increase Alzheimer's disease risk. Ann Neurol 69: 47–64.

Ressler KJ, Mercer KB, Bradley B, Jovanovic T, Mahan A, et al. (2011) Post-traumatic stress disorder is associated with PACAP and the PAC1 receptor. Nature 470: 492–497.

Robison AJ, Nestler EJ (2011) Transcriptional and epigenetic mechanisms of addiction. Nat Rev Neurosci 12: 623–637.

Roeske D, Ludwig KU, Neuhoff N, Becker J, Bartling J, et al. (2011) First genome-wide association scan on neurophysiological endophenotypes points to trans-regulation effects on SLC2A3 in dyslexic children. Mol Psychiatry 16: 97–107.

Sabeti PC et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature. 419:832–837.

Sabeti PC et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature. 449:913–8.

Sadakata T, Furuichi T. (2010) Ca(2þ)-dependent activator protein for secretion 2 and autistic-like phenotypes. Neurosci Res. 67: 197–202.

Sakurai K, Migita O, Toru M, Arinami T (2002) An association between a missense polymorphism in the close homologue of L1 (CHL1, CALL) gene and schizophrenia. Mol Psychiatry 7: 412–415.

Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, et al. (1996) Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 382: 722–725.

Sanyal A, Bau D, Marti-Renom MA, Dekker J (2011) Chromatin globules: a common motif of higher order chromosome structure? Curr Opin Cell Biol 23: 325–331.

Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A 103: 1412–1417.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, et al. (2012) Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169–175.

Schleinitz D, Tönjes A, Böttcher Y, Dietrich K, Enigk B, et al. (2010) Lack of significant effects of the type 2 diabetes susceptibility loci JAZF1, CDC123/CAMK1D, NOTCH2, ADAMTS9, THADA, and TSPAN8/LGR5 on diabetes and quantitative metabolic traits. Horm Metab Res 42: 14–22.

Schwab KR, Patel SR, Dressler GR (2011) Role of PTIP in class switch recombination and long-range chromatin interactions at the immunoglobulin heavy chain locus. Mol Cell Biol 31: 1503–1511.

Ségurel L, Leffler EM, Przeworski M. (2011) The Case of the Fickle Fingers: How the PRDM9 Zinc Finger Protein Specifies Meiotic Recombination Hotspots in Humans. PLoS Biol 9:e1001211.

Semendeferi K, Armstrong E, Schleicher A, Zilles K, Van Hoesen GW (2001) Prefrontal cortex in humans and apes: a comparative study of area 10. Am J Phys Anthropol 114: 224–241.

Sherwood CC, Stimpson CD, Raghanti MA, Wildman DE, Uddin M, et al. (2006) Evolution of increased glia-neuron ratios in the human frontal cortex. Proc Natl Acad Sci U S A 103: 13606–13611.

Shi X, Hong T, Walter KL, Ewalt M, Michishita E, et al. (2006) ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. Nature 442: 96–99.

Shilatifard A (2006) Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. Annu Rev Biochem 75: 243–269.

Shinawi M, Schaaf CP, Bhatt SS, Xia Z, Patel A, et al. (2009) A small recurrent deletion within 15q13.3 is associated with a range of neurodevelopmental phenotypes. Nat Genet 41: 1269–1271.

Shulha HP, Cheung I, Whittle C, Wang J, Virgil D, et al. (2011) Epigenetic signatures of autism: trimethylated H3K4 landscapes in prefrontal neurons. Arch Gen Psychiatry 69: 314–324.

Shulha HP, Crisci JL, Reshetov D, Tushir JS, Cheung I, Bharadwaj R, Chou H-J, et al. (2012). Human-specific histone methylation signatures at transcription start sites in prefrontal neurons. PLoS Biol, 10(11), e1001427.

Simonis-Bik AM et al. (2010) Gene variants in the novel type 2 diabetes loci CDC123/CAMK1D, THADA, ADAMTS9, BCL11A, and MTNR1B affect different aspects of pancreatic beta-cell function. Diabetes. 59:293–301.

Sklar P, Ripke S, Scott LJ, Andreassen OA, Cichon S, et al. (2011) Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet 43: 977–983.

Somel M, Franz H, Yan Z, Lorenc A, Guo S, et al. (2009) Transcriptional neoteny in the human brain. Proc Natl Acad Sci U S A 106: 5743–5748.

Somel M, Liu X, Tang L, Yan Z, Hu H, et al. (2011) MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates. PLoS Biol 9: e1001214

Splinter E, de Wit E, Nora EP, Klous P, van de Werken HJ, et al. (2011) The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes Dev 25: 1371–1383.

Stephan W, Song YS, Langley CH. (2006) The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics. 172:2647–63.
Swanson WJ, Nielsen R, Yang Q. (2003) Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol. 20:18–20.
Tajima F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.
Tang K, Thornton KR, Stoneking M (2007) A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol 5: e171
Teffer K, Semendeferi K (2012) Human prefrontal cortex: evolution, development, and pathology. Prog Brain Res 195: 191–218.
The Chimpanzee Sequencing And Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 437:69–87.
Thompson, JD, Gibson TJ, Higgins DG. (2002) Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics. Chapter 2: p. Unit 2 3.
Thornton Kevin R, Jensen JD. (2007) Controlling the false-positive rate in multilocus genome scans for selection. Genetics. 175:737–50.
Thornton KR, Jensen JD, Becquet C, Andolfatto P. (2007) Progress and prospects in mapping recent selection in the genome. Heredity. 98:340–348.
Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, et al. (2006) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39: 31–40.
Tsankova N, Renthal W, Kumar A, Nestler EJ (2007) Epigenetic regulation in psychiatric disorders. Nat Rev Neurosci 8: 355–367.
Tsujimoto S, Genovesio A, Wise SP (2010) Evaluating self-generated decisions in frontal pole cortex of monkeys. Nat Neurosci 13: 120–126.
Tsujimoto S, Genovesio A, Wise SP (2011) Frontal pole cortex: encoding ends at the end of the endbrain. Trends Cogn Sci 15: 169–176.
Turelli M, Barton NH. (1990) Dynamics of polygenic characters under selection. Theor Popul Biol 38:1–57.
Vargha-Khadem F, Gadian DG, Copp A, Mishkin M (2005) FOXP2 and the neuroanatomy of speech and language. Nat Rev Neurosci 6: 131–138.
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. (2006) A map of recent positive selection in the human genome. PLoS Biol. 4:e72.
Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci U S A 103: 135–140.
Wang Y-C, Chen J-Y, Chen M-L, Chen C-H, Lai I-C, et al. (2008) Neuregulin 3 genetic variations and susceptibility to schizophrenia in a Chinese population. Biol Psychiat 64: 1093–1096.
Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, et al. (2008) Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med 358: 667–675.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, et al. (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A 102: 7882–7887.
Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, et al. (2007) Localizing recent adaptive evolution in the human genome. PLoS Genet 3: e90.
Wong WSW, Yang Z, Goldman N, Nielsen R. (2004) Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites. Genetics. 168:1041–51.
Wright S (1931) Evolution in Mendelian Populations. Genetics 16: 97–159.
Xu S, Li S, Yang Y, Tan J, Lou H, et al. (2011) A genome-wide search for signals of high-altitude adaptation in Tibetans. Mol Biol Evol 28: 1003–1011.
Xu Y, Wang L, Buttice G, Sengupta PK, Smith BD. (2003) Interferon gamma repression of collagen (COL1A2) transcription is mediated by the RFX5 complex. J Biol Chem 278:49134–44.
Yan Q, Huang J, Fan T, Zhu H, Muegge K (2003) Lsh, a modulator of CpG methylation, is crucial for normal histone methylation. EMBO J 22: 5154–5162.
Yang Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol, 15(5): p. 568-73.
Yang Z (2005) The power of phylogenetic comparison in revealing protein function. Proc Natl Acad Sci U S A 102: 3179–3180.
Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.
Yang Z, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431–449.
Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22: 1107–1118.
Yang Z, Swanson WJ (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol 19: 49–57.
Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, et al. (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329: 75–78.
Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40: 638–645.
Zhang D, Sliwkowski MX, Mark M, Frantz G, Akita R, et al. (1997) Neuregulin-3 (NRG3): a novel neural tissue-enriched protein that binds and activates ErbB4. Proc Natl Acad Sci U S A 94: 9562–9567.
Zhang YE, Landback P, Vibranovski MD, Long M (2011) Accelerated recruitment of new brain development genes into the human genome. PLoS Biol 9: e1001179
Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322: 750–756.
Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet 12: 7–18.