COMPARATIVE GENOMIC AND EPIGENOMIC ANALYSES OF HUMAN AND NON-HUMAN PRIMATE EVOLUTION

A Dissertation
Presented to
The Academic Faculty

by

Ke Xu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biology

Georgia Institute of Technology
December 2013

COMPARATIVE GENOMIC AND EPIGENOMIC ANALYSES OF HUMAN AND NON-HUMAN PRIMATE EVOLUTION

Approved by:

Dr. Soojin Yi, Advisor
School of Biology
Georgia Institute of Technology

Dr. Todd Streelman
School of Biology
Georgia Institute of Technology

Dr. King Jordan
School of Biology
Georgia Institute of Technology

Dr. David Hall
Department of Genetics
University of Georgia

Dr. John McDonald
School of Biology
Georgia Institute of Technology

Date Approved: August 20, 2013

To my parents

ACKNOWLEDGEMENTS

Pursuing a PhD at Georgia Tech has been a memorable journey, during which I received overwhelming support from my advisor, committee members, family and friends. I would like to take this opportunity to thank them.

First and foremost, I would like to thank my advisor Dr. Soojin Yi for supporting me in every possible way and guiding me to a promising career. Her relentless pursuit to academic excellence has deeply influenced me and will continue to motivate me in my future endeavors. I thank my committee members Dr. King Jordan, Dr. John McDonald, Dr. Todd Streelman, and Dr. David Hall for their kind encouragements and valuable feedbacks to my research. I thank Dr. Leonid Bunimovich for supporting and encouraging me at the early stage of my PhD.

I am also grateful to my graduate school friends for sharing their knowledge and experience with me. Jianrong was extremely helpful to my programming skills in my first semester. Eddie has always been there to help me troubleshoot my problems with database. Navin was a great mentor when I was doing a summer internship in his company. Without them, graduate school to me would be less fulfilling.

Last but not least, I want to thank the School of Biology at Georgia Tech for providing me teaching assistantship opportunities. Dr. Mirjana Brockett and Dr. Linda Green are the academic professionals I worked with. I learned a lot from them. Being their teaching assistant allowed me to keep a balance between teaching and doing research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

α	Male-to-Female Mutation Rate Ratio
ANCOVA	Analysis of Covariance
BLAT	Blast Like Alignment Tool
bps	Basepairs
CEU	Utah Residents with Ancestry from Northern and Western Europe
CHBJPT	Han Chinese in Beijing, China and Japanese in Tokyo, Japan
CpG	Cytosine Immediately Followed by Guanine in 5' to 3' Direction
CpG O/E	Normalized CpG Dinucleotide Content
cpgoeDiff	CpG O/E Difference
DAF	Derived Allele Frequency
dI	Intronic Substitution Rate
dN	Nonsynonymous Substitution Rate
DNA	Deoxyribonucleic Acid
dS	Synonymous Substitution Rate
fast-X	Faster Evolutionary Rate of the X Chromosome to Autosomes
FDR	False Discovery Rate
GO	Gene Ontology
GpC	Guanine immediately followed by Cytosine in 5' to 3' direction
HAR	Human Accelerated Regions
INDELs	Insertions and Deletions
LRT	Likelihood Ratio Test
methyDiff	Methylation Level Difference
MS	Maximal Segment

N_X	Effective Population Size of the X Chromosome
N_A	Effective Population Size of Autosomes
PAML	Phylogenetic Analysis by Maximum Likelihood
RefSeq	Reference Sequence Database
RNA	Ribonucleic Acid
slow-X	Slower Evolutionary Rate of the X Chromosome to Autosomes
SVA	SINE/VNTR/Alu
TSS	Transcription Start Site
UCSC	University of California Santa Cruz
YRI	Yoruba in Ibadan, Nigeria

SUMMARY

Primates are one of the best characterized phylogenies with vast amounts of comparative data available, including genomic sequences, gene expression, and epigenetic modifications. Thus, they provide an ideal system to study sequence evolution, regulatory evolution, epigenetic evolution as well as their interplays. Comparative studies of primate genomes can also shed light on molecular basis of human-specific traits. This dissertation is mainly composed of three chapters studying human and non-human primate evolution. The first study investigated evolutionary rate difference between sex chromosome and autosomes across diverse primate species. The second study developed an unbiased approach without the need of prior information to identify genomic segments under accelerated evolution. The third study investigated interplay between genomic and epigenomic evolution of humans and chimpanzees.

Research advance 1: evolutionary rates of the X chromosome are predicted to be different from those of autosomes. A theory based on neutral mutation predicts that the X chromosome evolves slower than autosomes (slow-X evolution) because the numbers of cell division differ between spermatogenesis and oogenesis. A theory based on natural selection predicts an opposite direction (fast-X evolution) because newly arising beneficial mutations on the autosomes are usually recessive or partially recessive and not exposed to natural selection. A strong slow-X evolution is also predicted to counteract the effect of fast-X evolution. In our research, we simultaneously studied slow-X evolution, fast-X evolution as well as their interaction in a phylogeny of diverse primates. We showed that slow-X evolution exists in all the examined species, although their

degrees differ, possibly due to their different life history traits such as generation times. We showed that fast-X evolution is lineage-specific and provided evidences that fast-X evolution is more evident in species with relatively weak slow-X evolution. We discussed potential contribution of various degrees of slow-X evolution on the conflicting population genetic inferences about human demography.

Research advance 2: human-specific traits have long been considered to reside in the genome. There has been a surge of interest to identify genomic regions with accelerated evolution rate in the human genome. However, these studies either rely on *a priori* knowledge or sliding windows of arbitrary sizes. My research provided an unbiased approach based on previously developed "maximal segment" algorithm to identify genomic segments with accelerated lineage-specific substitution rate. Under this framework, we identified a large number of human genomic segments with clustered human-specific substitutions (named "maximal segments" after the algorithm). Our identified human maximal segments cover a significant amount of previously identified human accelerated regions and overlap with genes enriched in developmental processes. We demonstrated that the underlying evolutionary forces driving the maximal segments included regionally increased mutation rate, biased gene conversion and positive selection.

Research advance 3: DNA methylation is one of the most common epigenetic modifications and plays a significant role in gene regulation. How DNA methylation status varies on the evolutionary timescale is not well understood. In this study, we investigated the role of genetic changes in shaping DNA methylation divergence between humans and chimpanzees in their sperm and brain, separately. We find that for

orthologous promoter regions, CpG dinucleotide content difference is negatively correlated with DNA methylation level difference in the sperm but not in the brain, which may be explained by the fact that CpG depleting mutations better reflect germline DNA methylation levels. For the aligned sites of orthologous promoter regions, sequence divergence is positively correlated with methylation divergence for both tissues. We showed that the evolution of DNA methylation can be affected by various genetic factors including transposable element insertions, CpG depleting mutations and CpG generating mutations.

CHAPTER 1

INTRODUCTION

Primates represent one of the most diverse orders of mammals on earth. Investigating primate species is of broad interest not only because they provide an ideal model system to study evolution of closely related species but more importantly because ourselves, *Homo sapiens*, belongs to this order. For these reasons, a broad range of primate genomes are being sequenced, including Hominoids (human (International Human Genome Sequencing Consortium 2001), chimpanzee (Chimpanzee Sequencing and Analysis Consortium 2005), bonobo (Prufer et al. 2012), gorilla (Scally et al. 2012) and orangutan (Locke et al. 2011)), Old World monkeys (rhesus macaque (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007) and baboon (in process)), and New World monkeys (marmoset (in process) and squirrel monkey (in process)). Recently, scientists also began to unravel the whole genome sequences of our ancient relatives – Neanderthals (Green et al. 2010) and Denisovans (Meyer et al. 2012b).

Consequently, many researchers have been investigating genomic determinants of human-specific traits. There are generally two ways to conduct research on this aspect. One is that identifying human-specific trait first and then studying the trait-associated genes. For example, it had been known that the *FOXP2* gene is responsible for human speech and language (Lai et al. 2001) – a trait that sets human apart from the other species. A subsequent study showed that the *FOXP2* was under positive selection in humans (Enard et al. 2002). Another way to study this aspect is that identifying genomic regions with accelerated evolution in the human lineage but conserved in the other lineages first, and then studying the functions of these accelerated regions. Under this rationale, Pollard et al. identified ~200 human accelerated regions (HARs) that are evolutionarily conserved in chimpanzee, mouse, and rat (Pollard et al. 2006a). They then

demonstrated through experiments that the top one HAR is a novel RNA gene that is expressed during neocortex development (Pollard et al. 2006b).

In addition to comprehensive genomic data, abundant comparative gene expression data is also available for primates. For example, Caceres et al. measured gene expression levels of human, chimpanzee and macaque in their brains using microarray techniques and found that about 90% of the genes exhibiting expression differences between human and non-human species involved up-regulation in humans (Caceres et al. 2003). Khaitovich et al. also used microarray techniques to measure gene expression level in brain, heart, liver, kidney, and testis between human and chimpanzee (Khaitovich et al. 2005). They found that broadly expressed genes diverge less between human and chimpanzee than narrowly expressed genes both in expression level and amino acid sequences. With the advent of RNA-Seq technology, even more comprehensive comparative gene expression analyses have become possible. For example, Brawand et al. sequenced RNA from six tissues across 10 species covering a wide range of mammalian lineages and birds (Brawand et al. 2011). Their analyses of these gene expression data presented a dynamic picture of transcriptome evolution in mammals.

Epigenetics has long been of broad interest due to its associations with a variety of human diseases and influences on both DNA sequences and gene regulation (Feinberg 2007; Elango et al. 2008; Elango and Yi 2008). Recently, large-scale epigenomic studies in high resolution became possible. The next-generation sequencing technology together with the technique of bisulfite conversion of DNA sequences embarked the effort of whole-genome DNA methylation (methylomes) mapping at single-base resolution (Lister et al. 2009; Li et al. 2010; Xiang et al. 2010). Lister et al. (2009) provided the first single-base-resolution human DNA methylomes from human embryonic stem cells and fetal fibroblast, where they found widespread differences of cytosine methylation patterns between the two tissues. Later on, methylome from another human tissue - peripheral blood mononuclear cells was mapped by Li et al. (2010). In primates, comparative DNA

methylomes are available between human and chimpanzee: one from sperm (Molaro et al. 2011) and the other from brain (Zeng et al. 2012). A recent study also examined comparative histone modification mapping data between human and chimpanzee brains (Shulha et al. 2012).

The massive amount of comparative genomic, transcriptomic and epigenomic data provides us unprecedented opportunities to systematically examine molecular evolution of primates and infer genetic basis of human-specific traits. In the context of comparative analyses of primates, this dissertation investigates a series of evolutionary topics, including 1) differential evolutionary rates between sex chromosome and autosomes; 2) origin and function of clusters of human-specific single nucleotide substitutions; and 3) genetic basis of epigenetic divergence.

In chapter two of this dissertation, we take advantage of the well sequenced and annotated primate genomes to study evolutionary rate differences between the X chromosome and autosomes in a phylogeny of diverse primate species and mouse (as an outgroup). Theories predict that the X chromosome evolves at different rates from autosomes. The direction of this difference depends on the underlying evolutionary forces. A mutation based theory predicts that the X chromosome evolves slower than autosomes because of male mutation bias, or "male-driven evolution" (Li et al. 2002). In contrast, a selection based theory predicts a faster evolution of the X chromosome for the reason that X-linkage can facilitate fixation of newly arising beneficial mutations (Charlesworth et al. 1987). Although 'slow-X' evolution has been observed in many species, its degrees were found to vary between taxa (Chang et al. 1994; Makova and Li 2002; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007), which is explained by the varying ratios of the germ cell divisions in males and females between species (Borum 1961; Baker and Sum 1976; van den Hurk and Zhao 2005). Empirical studies of 'fast-X' evolution also reported mixed results. Some studies observe strong cases of fast-X while elusive in others (Bettancourt et al. 2002; Thornton et al. 2006;

Singh et al. 2008). Several factors have been proposed to account for this variation, such as total effective population size and relative effective population size of the X chromosome to autosomes (Mank et al. 2010). Intriguingly, Kirkpatrick and Hall predict that slow-X evolution makes fast-X evolution happen under a more restrictive condition (Kirkpatrick and Hall 2004). Chapter two simultaneously studies slow-X evolution, fast-X evolution, and their interactions among diverse primates. Our study reveals that the degree of slow-X evolution exhibits significant variation among the primates and this variation is in congruency with the variation of their life history traits. While slow-X evolution is observed in all the examined primates, we found that fast-X evolution is lineage-specific. We demonstrated that the degree of slow-X evolution is influential to fast-X evolution. We discussed how the varying male mutation bias between primates could contribute to the conflicting population genetic inferences about relative effective population size of the X chromosome to autosomes in human populations.

Chapter three investigates distribution of human-specific substitutions in the human genome, aiming to identify genomic regions with increased rate of human-specific substitutions and understand their origins and functions. Ever since the chimpanzee genome has been sequenced, there has been a surge of interest to study molecular uniqueness of the human genome by means of comparative genomics. Identifying genomic regions which have experienced accelerated evolution in human lineage is of special interest because these regions are likely to be under positive selection and thus be related to human-specific traits. Previous studies have yielded interesting and promising results about functionalities of these regions, however, their methodologies were either dependent upon a subset of genome (Pollard et al. 2006a; Berglund et al. 2009) or sliding windows of arbitrary sizes (Dreszer et al. 2007; Capra and Pollard 2011). In chapter three, we develop an unbiased approach to identify human genomic regions with accelerated substitution rate based on the ‘maximal segment’ algorithm (Ruzzo and Tompa 1999). This algorithm was originally designed to find contiguous subsequences

with regionally maximum scores from a sequence of score; thus, it enables us to find lineage-specific clusters of single nucleotide substitutions. By scanning human- and chimpanzee-specific substitutions through human-chimpanzee-macaque whole genome alignments, we identified a large number of human-specific clusters of single nucleotide substitutions, covering a significant amount of previously identified human accelerated regions (Pollard et al. 2006a; Berglund et al. 2009). Gene ontology analyses show that these clusters have significant overlap with genes involved in developmental processes. We also provide evidences showing that the origin of these clusters is driven by a combination of evolutionary forces, including regionally increased mutation rate, recombination associated processes, and positive selection.

Finally, chapter four investigates the role of genomic changes in shaping methylation level divergence and gene expression level divergence between humans and chimpanzees. DNA Methylation is a common epigenetic modification which usually targets cytosines followed immediately by a guanine – CpG sites. Although DNA methylation is involved in a series of regulatory activities such as suppressing gene expression (Siegfried et al. 1999) and proliferation of transposable elements (Meunier et al. 2005), reducing transcriptional noise (Huh et al. 2013), and dosage compensation (Kass et al. 1997), its pattern differs significantly even between closely related species. Recently, studies reveal that methylation patterns between humans and chimpanzees are diverged differently in different tissues – one study found that methylation levels in the brain are significantly lower in humans than in chimpanzees (Zeng et al. 2012); another study observed the opposite trend in the sperm (Molaro et al. 2011). It is well established that methylated cytosines tend to spontaneously deaminate to thymines, and thus mutations from CpG to TpG are considerably higher than other single nucleotide mutations (Fryxell and Zuckerkandl 2000; Elango et al. 2008). Recent studies suggest methylation status can be determined by nearby genetic elements (Gibbs et al. 2010; Lienert et al. 2011). In chapter four, we examine the relationships between methylation

divergence, sequence divergence, and expression divergence between humans and chimpanzees. We predict that genomic changes such as insertions/deletions and single nucleotide substitutions are related to both methylation divergence and expression divergence. The results from this study reveal inconsistent relationships between methylation divergence and CpG content divergence in different tissues – while there is a strong negative correlation between methylation level difference and CpG content difference in the sperm, this correlation is weak but positive in the brain. This can be partially explained by the fact that methylation-related mutations reflect better in germline cells. We demonstrated that genetic elements such as transposable element insertions and single nucleotide substitutions leading to CpG sites generation and depletion influence the evolution of DNA methylation of humans and chimpanzees.

In summary, the studies from this dissertation consolidated previous views about evolutionary rate difference between sex chromosome and autosomes, presented a new framework to study genomic segments under accelerated evolution, and explored various genetic factors of epigenetic evolution. My research sheds light on primate evolution on both genomic and epigenomic scale. It is a stepping stone to fully understanding molecular mechanisms of the evolution of human-specific traits.

CHAPTER 2

LINEAGE-SPECIFIC VARIATION IN SLOW- AND FAST-X EVOLUTION IN PRIMATES

Abstract

Theories predict that the evolutionary rates of X-linked regions can differ from those of autosomal regions. The male-biased mutation theory predicts a slower rate of neutral substitution on the X chromosome (slow-X evolution), as the X spends less time in male germlines, where more mutations originate per generation than in female germlines. The fast-X theory, however, predicts a faster rate of adaptive substitution on the X chromosome when newly arising beneficial mutations are, on average, partially recessive (fast-X evolution), as the X enjoys a greater efficacy of positive selection. The slow- and fast-X processes are expected to interact as the degree of male-biased mutation can in turn influence the relative rate of adaptive evolution on the X. Here we investigate lineage-specific variation in, and the interaction of, slow- and fast-X processes using genomic data from four primates. We find consistent evidence for slow-X evolution in all lineages. In contrast, evidence for fast-X evolution exists in only a subset of lineages. In particular, the marmoset lineage, which shows the strongest evidence of fast-X, exhibits the lowest male-mutation bias. We discuss the possible interaction between slow- and fast-X evolution and other factors that influence the degrees of slow- and fast-X evolution.

Introduction

Evolutionary theories provide conflicting predictions on the relative evolutionary rates of X-linked loci to those of autosomal loci. A theory based on male mutational bias, or 'male-driven evolution', predicts that the X chromosome will accumulate neutral

substitutions at a slower rate than autosomes (Miyata et al. 1987; Li et al. 2002; Ellegren 2007). Mutation rates in the male germline are generally higher than those in the female germline, owing to the greater number of replicative germ cell divisions involved in the production of sperm *versus* eggs per generation (Haldane 1947; Penrose 1955; Drost and Lee 1995). As X chromosomes spend less time in males than autosomes, the X experiences fewer mutations and hence fewer neutral substitutions. Male-biased mutation therefore predicts slow-X evolution.

In contrast, a theory based on hemizygous selection in the XY sex predicts that the X chromosome will accumulate adaptive substitutions faster than the autosomes (Charlesworth et al. 1987). In particular, provided that newly arising beneficial mutations are, on average, partially recessive ($\overline{h} < 0.5$), their probabilities of fixation are higher for X-linked loci than autosomal ones (Avery 1984; Charlesworth et al. 1987). Thus, assuming new beneficial mutations are generally partially recessive, theory predicts fast-X evolution.

Therefore, sites on the X-chromosome may evolve slower or faster than those on the autosomes, depending upon the main underlying evolutionary forces acting on them. Studies found that the strength of slow-X evolution varies among taxa (Shimmin et al. 1993; Chang et al. 1994; Bauer and Aquadro 1997; Ellegren and Fridolfsson 1997; Bohossian et al. 2000; Betancourt et al. 2002; Li et al. 2002; Makova and Li 2002; Ellegren and Fridolfsson 2003). Since slow-X evolution is a mutation-driven process, lineage effects likely reflect factors that affect lineage differences in mutation rate, such as generation time, metabolic rates and, potentially, mating system (Bartosch-Härlid et al. 2003; Blumenstiel 2007; Presgraves and Yi 2009). Likewise, empirical studies of fast-X evolution have also yielded mixed results (Betancourt et al. 2002; Thornton et al. 2006; Presgraves 2008; Singh et al. 2008; Mank et al. 2010). In particular, in mammals, evidence of fast-X is largely restricted to genes expressed in testis (Torgerson and Singh

2003; Chimpanzee Sequencing and Analysis Consortium 2005; Khaitovich et al. 2005; Baines et al. 2008).

While these studies mostly focused on testing the prediction of either of the two (slow- versus fast-X) theories, it is proposed that the mutation-based slow-X process and the selection-based fast-X process may interact. The $\bar{h} < 0.5$ condition of the fast-X theory assumed equal mutation rates in males and females ($\alpha = u_m/u_f = 1$) and assumed the effective population size of the X is ¾ that of the autosomes ($N_X/N_A = 0.75$; Charlesworth et al. 1987). The conditions for fast-X evolution depend on both. In the presence of male mutation bias, the dominance condition for fast-X evolution becomes more restrictive (Kirkpatrick and Hall 2004). If, for instance, the male-to-female mutation rate ratio (α) = 5, then fast-X occurs only when $\bar{h} < 0.3$. Male-biased mutation can thus impede fast-X evolution. When $N_X/N_A > 0.75$, the dominance condition for fast-X evolution is more permissive (Vicoso and Charlesworth 2009). If, for instance, there is greater-than-Poisson variance in male reproductive success, N_X/N_A will exceed 0.75, and fast-X evolution can occur even when new beneficial mutations are partially dominant. In XY systems, then, sexual selection on males can facilitate fast-X evolution.

In this study, we simultaneously investigate slow- and fast-X evolution as well as their interactions. To do so, we compare lineage-specific rates of substitution on the X and autosomes in four primates: human, orangutan, rhesus macaque, and marmoset. We contrast substitution data from intron and protein-coding sequences. As the evolutionary rates of introns are largely governed by the input of neutral mutations while those of nonsynonymous sites are largely governed by selection, slow- and fast-X evolution can be inferred using introns and exons, respectively. We find that all four primates show strong evidence of slow-X evolution. However, the strength of slow-X evolution varies significantly among lineages, potentially due to the variation in life history traits affecting the strength of male mutation bias. Among the primates investigated here, the marmoset

lineage exhibits the lowest male mutation bias. Interestingly, we find that evidence for fast-X evolution at nonsynonymous sites is mostly limited to marmoset.

We discuss potential interaction between the slow- and fast-X processes and other factors that influence the variation of these processes. We also demonstrate that lineage-specific variation of slow- and fast-X processes has implications for population genetic inferences about human demography.

Methods

Orthologous gene assembly

The human genome assembly is considered highly accurate or 'finished', representing approximately eightfold coverage of euchromatic regions (International Human Genome Sequencing Consortium, 2004). The genome sequences of orangutan, rhesus macaque, and marmoset are of similar coverages (6× for orangutan and marmoset, and 5× for rhesus macaque, UCSC Genome Browser). Non-human primate data are all obtained from females. Since females harbor 2 X chromosomes, the X and autosomes are sequenced to similar depths. We retrieved and assembled orthologous gene sets from four primates— human, orangutan, rhesus macaque, and marmoset — with mouse as an outgroup from the Ensembl BioMart (version: Ensembl Genes 57). For any pair of species, we chose genes with orthology type marked as 'ortholog_one2one'. For genes with multiple transcripts, the longest transcript was selected. To estimate substitution rate differences between the X chromosome and autosomes, we chose genes that have remained X-linked throughout mammalian evolution, *i.e.*, those that are X-linked in human, orangutan, rhesus macaque, marmoset and mouse. There are 303 such genes. For autosomal genes, we chose genes that are homologous to those on human chromosomes 5 and 6, which have sizes and G+C contents similar to those of the X chromosome. We further limited the analysis to genes that have remained on homologous chromosomes in orangutan and rhesus macaque; no synteny information for

chromosomes homologous to human chromosomes 5 and 6 was available for marmoset at the time of the analysis. There are 977 such genes.

Male mutation bias and protein evolutionary rates

We used intron sequences to estimate male mutation bias. We aligned the repeat-masked, concatenated intron sequences of each gene using MLAGAN v2.0 (Brudno et al. 2003). To minimize the influence of sites that may be under natural selection, we removed first introns as well as 100 bps adjacent to splice sites of the remaining introns. Introns shorter than 300 bps after this procedure were also removed. In addition, we masked hyper-mutable CpG dinucleotides (Kim et al. 2006). Lineage specific numbers of nucleotide substitutions were estimated using PAML baseml 4.2, assuming the HKY model of substitution (Yang 2007). Male-to-female mutation rate ratios (α) were then estimated following Miyata et al. (1987). Confidence intervals were obtained using the bootstrapping resampling method (Makova and Li 2002; Lu and Wu 2005; Elango et al. 2009). Estimates of the male-to-female mutation rate obtained using the Kimura's 2-parameter model are similar, demonstrating that our results are robust against different substitution models used (Supplementary Table A.1).

Coding sequences were translated to amino acid sequences first and then aligned using ProbCons 1.08 (Do et al. 2005). Genes with fewer than 100 aligned nucleotides were removed from subsequent analyses. Lineage-specific dN (number of nonsynonymous substitutions per site), dS (number of synonymous substitutions per site), and dN/dS values were estimated using PAML codeml 4.2 (Yang 2007). Again, hyper-mutable CpG dinucleotides were removed. To avoid overestimation, we removed 18 X-linked and 54 autosomal genes with dN/dS values greater than 15 in any lineage (among these genes, all but two genes had dN/dS > 100). The final data set comprised 216 X-linked genes and 702 autosomal genes. Confidence intervals were calculated by bootstrapping 1000 times. We also calculated dN/dI for genes with both intron and exon

data available following the above filtering steps. There are 151 X-linked genes and 490 autosomal genes for which dN/dI data are calculated.

Statistical tests

We tested whether evolutionary rates of protein coding sequences exhibit lineage-specific patterns of slow- and fast-X using linear models. The ANCOVA model for nonsynonymous rates is, for example, defined as

$$dN \sim \text{lineage} + \text{chromosome} + dI + \text{lineage:chromosome interaction}.$$

The lineage term accounts for the lineage effects, the chromosome term accounts for the effect of X *versus* autosomal linkage, and the dI term accounts for locus-specific variation in mutation rate. The lineage:chromosome interaction term tests the null hypothesis that the effect of X *versus* autosome is independent of lineage; rejection of the null therefore indicates lineage-specific difference in the X *versus* autosomal effects on dN. We repeated these analyses for dS and dN/dS. Because the continuous variables are not normally distributed, we used a permutation method to test for significance, with each term's significance examined by permutating the samples 100,000 times.

Results

Slow-X evolution in primates

From the four primate species, including two apes (human and orangutan), one Old World monkey (rhesus macaque), and one New World monkey (marmoset), we extracted >15 million intron sites that are likely to be evolving under little selective constraint, including >4 million from the X-chromosome (see Materials and Methods). The phylogenetic tree and lineage-specific branch lengths obtained from the autosomal intron data are consistent with previous findings (Figure 2.1). Notably, the human lineage

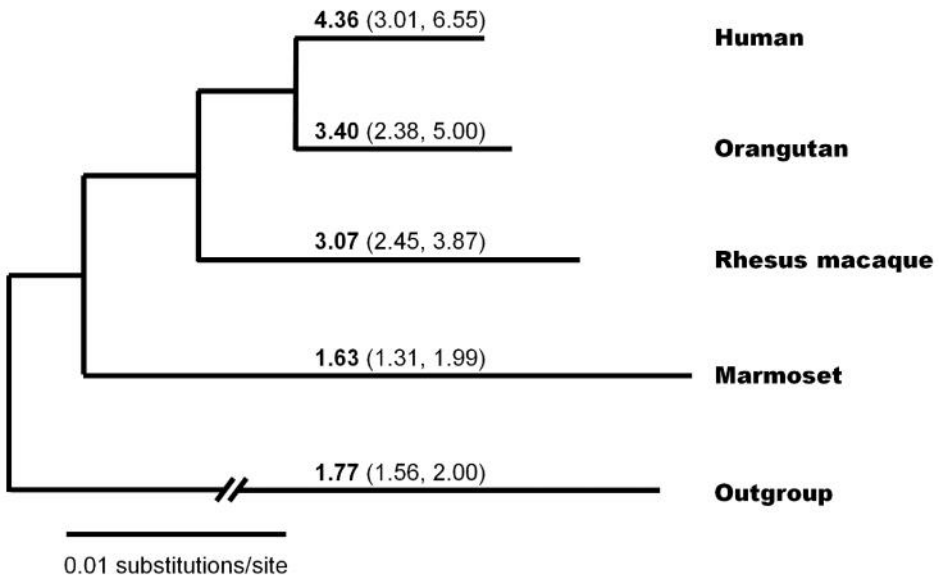

Figure 2.1. Phylogenetic tree of the four primates and the outgroup. The branch lengths estimated using the PAML follows the following tree: (((human: 0.010402, orangutan: 0.011597): 0.008192, rhesus: 0.026041): 0.010851, marmoset: 0.056405, outgroup: 0.415723). The estimates of male-to-female mutation rate ratio α are marked in bold above each branch along with the 95% confidence interval in parentheses.

has a shorter branch length than other apes (Elango et al. 2006) and Old World monkeys (Yi et al. 2002), while marmosets have a longer branch length than the catarrhines (Steiper and Young 2006).

For all four primate lineages, the estimated intron substitution rates from X-linked genes are significantly lower than those from autosomal genes (Wilcoxon test, $P < 10^{-10}$, Supplementary Table A.2), indicating slow-X evolution. We then estimated lineage-specific male-to-female mutation rate ratios (α) from the X and autosomal intron substitution rates (Miyata et al. 1987). Note that the species in this analysis are sufficiently diverged from one other that the effects of ancestral polymorphism, which can affect estimates of α, should be negligible (Makova and Li 2002). The degree of male mutation bias, measured by α, varies between different lineages, generally in accord with the previous findings. Our estimate of α for the human lineages is 4.36, which falls within the range of previously published estimates (Li et al. 2002). We report that the α from rhesus monkey is 3.07, again in accord with previous results (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007; Elango et al. 2009). We report that the male mutation bias in the orangutan lineage is 3.40. Interestingly, the marmoset lineage exhibits an extremely reduced male mutation bias, $\alpha = 1.63$. This value is significantly lower than the male mutation bias from the other three primate lineages investigated in this study (Figure 2.1). In fact, it is the lowest among the published values from primates, and similar to the estimated values from rodents. Indeed, the male mutation bias from the outgroup branch, a composite of the rodent lineage leading to the mouse and the ancestral primate lineage leading to anthropoids, is 1.77, similar to that from the marmoset.

Lineage-restricted fast-X evolution

To test for evidence of fast-X evolution, we investigated the relative evolutionary rates of protein coding sequences (Table 2.1). In human, orangutan, and rhesus macaque lineages, the X/A ratios of mean and median dN are all less than 1 (Table 2.1). Interestingly, in the marmoset lineage, the X/A ratio of mean dN is 1.25, suggesting that the nonsynonymous sites of the X chromosome evolve faster than those of autosomes. The generally lower rates of nonsynonymous substitution on the X could arise for three reasons: a greater efficacy of purifying selection on the X; a sample of X-linked genes with greater average functional constraints than the sample from chromosome 5 and chromosome 6; or a lower mutation rate at X-linked *versus* autosomal loci arising from male-biased mutation (see above).

To begin to distinguish among these possibilities, we controlled for X-autosome mutation rate differences by studying nonsynonymous rates of substitution standardized by synonymous (dN/dS) and intronic rates of substitution (dN/dI). While most branches showed X/A ratios of mean and median dN/dS and dN/dI greater than 1, the 95% confidence intervals tended to include 1. For those lineages showing some significant evidence for fast-X evolution, the signals were not consistent across measures. For instance, the rhesus macaque lineage shows suggestive evidence for fast-X evolution using mean dN/dS but not using median dN/dS, mean dN/dI, or median dN/dI. Similarly, the marmoset lineage shows suggestive evidence for fast-X evolution using mean dN/dI but not using median dN/dI, mean dN/dS, or median dN/dS. The fact that the X/A ratios of means, but not medians, tend to exceed 1 suggests that the weak signal of fast-X evolution in this analysis comes from the upper tails (high dN/dS or dN/dI) of the distributions (see below).

The X/A ratios of evolutionary rates in Table 2.1 vary considerably among lineages. We therefore used ANCOVA models with evolutionary rates (dN, dS, or dN/dS) as response variables and lineage (to account for lineage-specific evolutionary

Table 2.1. X/A ratios of mean or median dN, dS and dN/dS in four primates.

		dN	dS	dN/dS	dN/dI
Human	mean	0.80 (0.58, 1.15)	0.76 (0.68, 0.86)	1.03 (0.80, 1.30)	1.04 (0.70, 1.50)
	median	0.75 (0.61, 1.09)	0.84 (0.72, 0.94)	1.19 (0.81, 1.75)	1.06 (0.76, 1.45)
Orangutan	mean	0.80 (0.57, 1.12)	0.83 (0.69, 0.99)	0.85 (0.67, 1.06)	0.96 (0.68, 1.33)
	median	0.85 (0.62, 1.12)	0.83 (0.69, 0.98)	0.86 (0.57, 1.15)	1.04 (0.75, 1.37)
Rhesus	mean	0.89 (0.65, 1.19)	0.77 (0.64, 0.91)	1.30 (1.03, 1.60)	1.04 (0.76, 1.39)
	median	0.77 (0.58, 1.06)	0.71 (0.62, 0.83)	1.13 (0.80, 1.41)	0.94 (0.75, 1.38)
Marmoset	mean	1.25 (0.94, 1.62)	0.94 (0.83, 1.05)	1.21 (0.96, 1.50)	1.40 (1.04, 1.86)
	median	0.95 (0.76, 1.27)	0.93 (0.88, 1.05)	1.07 (0.80, 1.44)	1.04 (0.88, 1.44)

95% confidence intervals are given in parenthesis

Table 2.2. Significance of lineage, chromosome, locus specific intron divergence, and lineage-chromosome interaction on evolutionary rates.

	4 species (external branches only)		
	F-value	*Pr(>F)*	*P-value (permutation)*
Response variable: dN			
lineage	18.01	$< 10^{-10}$	$< 10^{-5}$
chromosome	2.24	NS	NS
dI	34.54	$< 10^{-8}$	$< 10^{-5}$
lineage:chromosome interaction	2.52	0.056	0.057
Response variable: dS			
lineage	14.70	$< 10^{-8}$	$< 10^{-5}$
chromosome	1.61	NS	NS
dI	76.40	$< 10^{-17}$	$< 10^{-5}$
lineage:chromosome interaction	1.14	NS	NS
Response variable: dN/dS			
lineage	7.74	$< 10^{-4}$	10^{-4}
chromosome	5.12	0.02	0.02
dI	12.39	$< 10^{-3}$	10^{-3}
lineage:chromosome interaction	*3.74*	*0.01*	*0.01*

NS: Not significant, *i.e.* $P > 0.1$

rates), chromosome (to account for X or autosomal linkage), and intron substitution rate (dI, to account for gene-specific mutation rates) as explanatory variables. We also included a lineage:chromosome interaction term to test whether relative evolutionary rates of X-linked and autosomal genes vary significantly among the 4 lineages. Since evolutionary rate data are not normally distributed, we assessed significance using permutation tests. As Table 2.2 shows, lineage has a highly significant effect on dN and dS. Notably, dN/dS, which is corrected for lineage-specific divergence, and thus supposedly measures only selective constraints, also shows a highly significant lineage effect. The effective level of functional constraints thus appears to vary among primate lineages. Neither dN nor dS show significant effects of chromosome, whereas dN/dS does. Thus, controlling for lineage and locus-specific mutation rate, dN/dS shows a significant fast-X effect. Across loci, intronic substitution rate (dI) is highly correlated with dN and dS but not with dN/dS, suggesting that the dN/dS ratio adequately standardizes for locus-specific mutation rate. Finally, the lineage:chromosome interaction term is marginally significant for dN and significant for dN/dS, but not for dS. The lineage:chromosome interaction effect implies that the magnitude of the X-autosome difference in dN/dS varies significantly among lineages.

A lineage-specific fast-X signal for genes with histories of rapid evolution

The fast-X theory is concerned with differences in the rate of adaptive evolution between the X and autosomes. We therefore tested for a signal of fast-X evolution by estimating the proportion of X-linked *versus* autosomal genes for which dN/dI >1, as expected given histories of recurrent positive selection (Table 2.3). In the marmoset lineage, the X chromosome shows a significant 2.6-fold excess of genes with dN/dI >1 relative to autosomes. In the composite outgroup lineage, the X shows a similar significant excess of genes with dN/dI >1. None of the other three primate lineages show an excess of positive selection on the X.

Table 2.3. Lineage specific signals of fast-X for genes under different dN/dI thresholds.

	X			A				
	+[a]	-[b]	%pos_sel[c]	+	-	%pos_sel	*P*[d]	X/A[e]
dN/dI>1								
Human	2	149	0.013	10	480	0.020	0.741	0.6
Orangutan	17	134	0.113	41	449	0.084	0.330	1.3
Rhesus	24	127	0.159	68	422	0.139	0.595	1.1
Marmoset	8	143	0.053	10	480	0.020	**0.047**	2.6
Outgroup	3	148	0.020	0	490	0.000	**0.013**	Inf
dN/dI>0.8								
Human	10	141	0.066	15	475	0.031	0.056	2.2
Orangutan	21	130	0.139	57	433	0.116	0.477	1.2
Rhesus	29	122	0.192	82	408	0.167	0.539	1.1
Marmoset	11	140	0.073	13	477	0.027	**0.014**	2.7
Outgroup	3	148	0.020	2	488	0.004	0.088	4.9
dN/dI>0.6								
Human	11	140	0.073	29	461	0.059	0.565	1.2
Orangutan	23	128	0.152	79	411	0.161	0.899	0.9
Rhesus	38	113	0.252	104	386	0.212	0.315	1.2
Marmoset	16	135	0.106	26	464	0.053	**0.036**	2.0
Outgroup	5	146	0.033	3	487	0.006	**0.021**	5.4
dN/dI>0.4								
Human	20	131	0.132	73	417	0.149	0.693	0.9
Orangutan	39	112	0.258	122	368	0.249	0.830	1.0
Rhesus	49	102	0.325	142	348	0.290	0.417	1.1
Marmoset	35	116	0.232	61	429	0.124	**0.003**	1.9
Outgroup	11	140	0.073	29	461	0.059	0.565	1.2
dN/dI>0.2								
Human	56	95	0.371	164	326	0.335	0.434	1.1
Orangutan	70	81	0.464	234	256	0.478	0.780	1.0
Rhesus	70	81	0.464	236	254	0.482	0.710	1.0
Marmoset	60	91	0.397	167	323	0.341	0.207	1.2
Outgroup	38	113	0.252	114	376	0.233	0.662	1.1

a. number of genes having dN/dI > threshold
b. number of genes having dN/dI < threshold
c. frequency of genes having dN/dI > threshold
d. p-value of Fisher's exact test
e. X/A ratio of number of the percentage of genes having dN/dI > threshold

The dN/dI >1 criterion for inferring positive selection is stringent. We therefore considered an arbitrary range of dN/dI values, assuming that that there is greater enrichment for positive selection as our criteria become increasingly restrictive, from dN/dI >0.2 → >1 (Table 2.3). In the marmoset lineage, the signal of fast-X evolution gets stronger as we enrich for positive selection: for dN/dI >0.8 and >1, the X shows a significant ~2.6-fold excess of rapidly evolving genes; for dN/dI >0.4 and >0.6, the signal diminishes to ~2.0-fold excess of rapidly evolving genes on the X; and for dN/dI >0.2, the signal of fast-X evolution in marmoset disappears entirely. In the composite outgroup lineage, the X shows a similar significant (or marginally significant) qualitative enrichment for rapidly evolving genes as the dN/dI criterion increases from >0.6 → >1. None of the other three primate lineages show similar enrichment of rapid evolution on the X. Finally, the distributions of dN/dI and dN/dS both differ significantly between the X and autosomes in the marmoset lineage (Wilcoxon test, P = 0.016 and 0.036, respectively), but not in the other three lineages.

Fast-X evolution confirmed at testis-expressed genes in humans

Fast-X evolution is expected to be especially strong for mutations having male-beneficial fitness effects (although the condition that $\bar{h} < 0.5$ for new favorable mutations holds; (Charlesworth et al. 1987; Vicoso and Charlesworth 2006). We therefore studied evolutionary rates at genes with testis- or sperm-specific functions in humans (Torgerson and Singh 2003; Khaitovich et al. 2005). In our data set, 5 of 151 X-linked genes and 12 of 490 autosomal genes show testis-specific expression. The X-linked testis-specific genes have significantly higher dN, dN/dS, and dN/dI than X-linked genes not expressed in testes (Supplementary Figure A.1). X-linked testis-specific genes also show greater dN, dN/dS and dN/dI than autosomal testis-specific genes, although the significance was marginal due to the small sample size (Supplementary Figure A.1). These findings are

consistent with previous ones showing that human testis-specific genes show elevated rates of substitution on the X (Torgerson and Singh 2003; Khaitovich et al. 2005).

Discussion

In this paper, we provide a simultaneous look at predictions of slow- and fast-X evolution by studying sites enriched for neutral molecular evolution in introns and sites enriched for purifying and positive natural selection in exons from four primate lineages. We find evidence for significant slow-X evolution among all four primate lineages, consistent with male mutation bias. The degree of slow-X evolution, however, varies between lineages, largely in accord with previous observations. We also find evidence consistent with fast-X evolution in marmosets and, at least for genes with male-biased expression, in humans. Below we discuss the possible causes of lineage effects on the strength of male-biased mutation and, hence, slow-X evolution. We also consider the consequences of lineage differences in slow-X evolution for fast-X evolution and, more practically, for population genetic inferences of human sex-specific demography.

Causes of lineage-specific variation of male mutation bias

Male mutation bias arises from the asymmetry in the numbers of cell divisions between male and female germlines (Haldane 1947; Miyata et al. 1987; Crow 1997; Hurst and Ellegren 1998). The magnitude of the asymmetry likely varies with life history traits across species. For example, it has been proposed that as generation time increases, the cumulative difference between the number of cell divisions in the male and female germlines will also increase, thereby increasing α (Chang et al. 1994; Li et al. 2002; Bartosch-Härlid et al. 2003; Goetting-Minesky and Makova 2006; Sayres 2011). Consistent with a generation time effect on male mutation bias, estimates of α from humans, which have relatively long generation times, converge on large values, 4-6,

whereas those from mouse, which have relatively short generation times tend to be smaller, ~2 (Li et al. 2002).

Our primate data provide further support for the generation time effect on male mutation bias. In particular, the ranks of male mutation bias observed between the primate lineages included in the current study (human > orangutan > rhesus monkey > marmoset) correspond well to the ranks of several life history traits known to co-vary with generation times. For instance, body mass, age at sexual maturity and maximum life expectancy are significantly correlated with the estimated male mutation bias (Pearson's ° = 0.97, 0.99 and 0.97, P = 0.03, 0.01 and 0.03, respectively; Figure 2.2). While suggestive, these correlations are not corrected for phylogeny. Doing so using Felsenstein's independent contrasts method (Felsenstein 1985), leaves just three phylogenetically independent points. Even though the correlations between the above life history traits and male mutation bias remain highly positive (for example, the correlation between age at sexual maturity and male mutation bias, after correcting for phylogenetic independence, is 0.96), no meaningful test of significance can be obtained from three data points. Details are presented in Supplementary Table A.3.

Other life history traits may also affect the strength of male mutation bias. The intensity of sperm competition, for instance, varies considerably with primate mating systems (Harcourt et al. 1981; Harcourt et al. 1995; Dixson and Anderson 2001). In systems for which the risk of sperm competition is high, males have evolved greater investment in sperm production— larger relative testis mass and greater numbers of sperm per ejaculate (Harcourt et al. 1981; Smith 1984; Dixson and Anderson 2001). If adaptation to sperm competition involved the evolution of more male germline cell divisions, to produce more sperm faster, then elevated male mutation bias may evolve as an incidental byproduct (Blumenstiel 2007; Presgraves and Yi 2009). Previously, we reported evidence for a positive correlation between α and relative testis mass in hominids (Presgraves and Yi 2009). The present analysis spans a wider range of taxa to

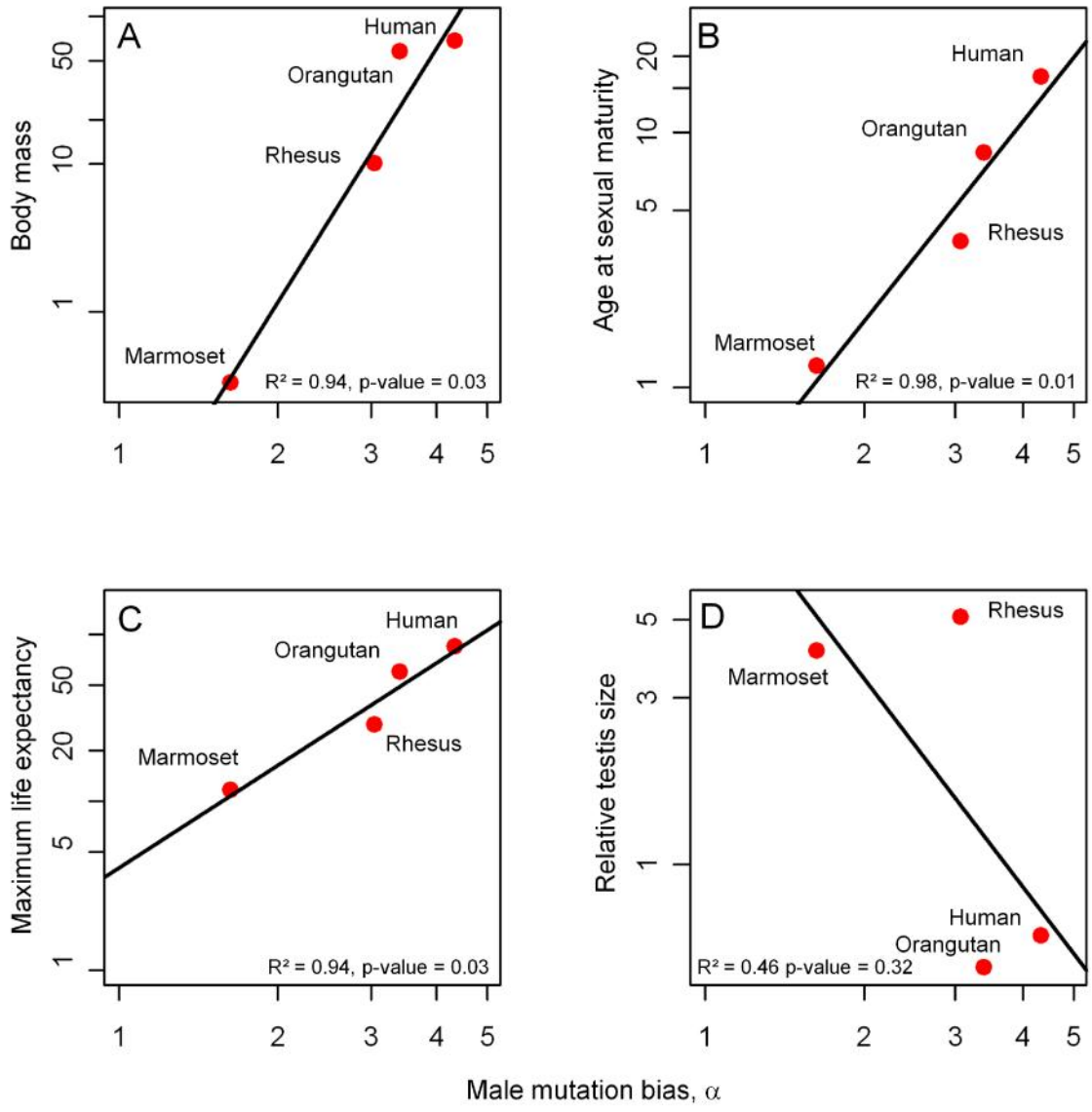

Figure 2.2. Correlations between the male mutation bias (α) and several life history traits. Male mutation bias is positively correlated with (A) body mass, (B) age at sexual maturity, (C) maximum life expectancy. However, it is not correlated with (D) relative testis mass. Values are log-transformed to improve normality.

include hominids, Old World monkeys, and New World Monkeys, but does not support a simple, general relationship between male mutation bias and indexes of sperm competition. Of the four primates studied, rhesus macaque experiences the greatest intensities of sperm competition, as evidenced by its mating system (multi-male and multi-female) and the largest relative testis mass among the species investigated (Harcourt et al. 1995, Supplementary Table A.3). In comparison, humans and orangutans exhibit generally single-male mating systems (Harcourt et al. 1995; Martin 2007). Relative testis sizes of human and orangutan are much lower than that of rhesus macaque (Supplementary Table A.2). However, male mutation bias is greater in humans and orangutans than in rhesus macaque (Figure 2.1). Marmoset, whose relative testis size is the second largest among the four primates, also exhibits little male mutation bias (Figure 2.1, Supplementary Table A.3). These observations suggest either that sperm competition has little effect on male mutation bias or that other life history traits, like generation time, are much more important. A recent study of distantly related mammals also reached a similar conclusion (Sayres et al. 2011).

Lineage-specific slow-X evolution can affect inferences about human demography

Because the X chromosome and autosomes are inherited differently with respect to sex, comparing DNA sequence polymorphism for the X and autosomes can be informative about human demographic history. In a population with equal numbers of effective males and females, the effective number of X chromosomes (N_X) should be ¾ that of autosomes (N_A). However, if the effective sex ratio deviates from 1, then the N_X/N_A ratio can deviate from ¾ (Charlesworth 2001; Vicoso and Charlesworth 2009). For instance, in a polygynous population, males have higher variance in reproductive success. Consequently, the effective number of males could be considerably lower than that of females, causing the N_X/N_A ratio to be greater than ¾. In contrast, if a population is founded by a male-biased group, the N_X/N_A ratio would be less than ¾. Two recent

studies using population genetic approaches to estimate the N_X/N_A ratio in humans reached different conclusions (Hammer et al. 2008; Keinan et al. 2009). Keinan et al. (2009) estimated that N_X/N_A is unusually low in non-African populations, and proposed a male-biased dispersal model (Keinan and Reich 2010). Hammer et al. (2008), however, estimated that N_X/N_A is approximately 1 in all populations examined, including African and non-African populations.

We hypothesize that species differences in male mutation bias may partially explain the discrepancy (see also (Bustamante and Ramachandran 2009)). N_X/N_A is inferred from the ratios of X *versus* autosomal polymorphism, corrected for X *versus* autosomal mutation rate differences:

$$\left(\frac{N_X}{N_A}\right) = \left(\frac{4N_X\mu_X}{4N_A\mu_A}\right) \times \left(\frac{\mu_A}{\mu_X}\right) = \left(\frac{\pi_X}{\pi_A}\right) \Big/ \left(\frac{\mu_X}{\mu_A}\right)$$

Lacking an experimentally defined $«_X/«_A$ from humans, the $«_X/«_A$ parameter is often inferred from human X/A ratios of divergence from an outgroup species. For instance, Keinan et al. (2009) used human divergence from macaque, whereas Hammer et al. (2008) used human divergence from orangutan. As shown above, however, the degree of male mutation bias, and consequently $«_X/«_A$, differs among these lineages. Male mutation bias is weaker in rhesus macaque than in orangutan (Figure 2.1). The $«_X/«_A$ from rhesus macaque used in Keinan et al. (2008) is ~0.875, whereas the $«_X/«_A$ from orangutan used in Hammer et al. (2008) is ~0.750, consistent with our estimates. We examined other studies that reported $«_X/«_A$ ratios, including pairwise estimates and lineage-specific estimates (Table 2.4). As expected if α for human > orangutan > rhesus macaque (Figure 2.1), the different estimates in Table 2.4 consistently show a trend that

Table 2.4. X chromosome to autosome ratios of mutation rates estimated using different outgroups. Some studies used estimates obtained from pairwise comparisons. While there exists considerable variation among studies, the rate differences between the X chromosome and autosomes are consistently larger in human-orangutan comparisons than in human-macaque comparisons. Several studies now provide lineage-specific X to autosomal ratios for human, orangutan and macaque. Again, the pattern is obvious that male mutation bias is greater in apes than in Available estimates of lineage-specific estimates are also available.

α_X/α_A in pairwise comparisons	**Human-Orangutan**	**Human-Macaque**
Hammer et al. (2008)	0.755	
Keinan et al. (2009)		0.875
(Patterson et al. 2006) 5 species comparison	0.877	0.921
Elango et al. (2009)		0.866
Rhesus macaque genome sequencing and analysis consortium (2007)		0.839
(Ebersberger et al. 2007)	0.790	0.823
Current study	0.805	0.840

Lineage-specific estimation of α_X/α_A	**Human**	**Orangutan**	**Macaque**
Patterson et al. (2006) 5 species comparison	0.785	0.892	0.941
Patterson et al. (2006) 4 species comparison	0.797	N/A	0.913
Ebersberger et al. (2007)	0.746	0.790	0.851
Current study	0.791	0.818	0.830

$«_X/«_A$ are human < orangutan < rhesus macaque. Using divergence data from macaque or orangutan to correct for the higher male-biased mutation rate in humans will thus cause underestimates of N_X/N_A. As rhesus macaque has a higher $«_X/«_A$ than orangutan, using divergence from macaque will more strongly underestimate the true N_X/N_A.

The difference in $«_X/«_A$ estimates used by Keinan et al. (2009) and Hammer et al. (2008) cannot, however, fully account for the discrepancy in N_X/N_A between the two studies. For example, if we correct nucleotide polymorphism from Keinan et al. (2009) using human-specific $«_X/«_A$ obtained from our current study, N_X/N_A of the West African population increases from 0.763 to 0.844. The estimates of N_X/N_A of Hammer et al. (2008) decreases slightly after being corrected using the human specific $«_X/«_A$, but still higher than those of Keinan et al. (2009). For example, the mean N_X/N_A of Hammer et al. (2008)'s data decreases from 1.036 to 0.989. Other factors, such as different sensitivities of analytical techniques used in the two studies (Emery et al. 2010) and the effect of natural selection on a subset of sites nearby genes (Gottipati et al. 2011) are likely to account for the remaining discrepancies.

Lineage-specific signal of fast-X evolution

To date, strong evidence for fast-X evolution resulting from beneficial substitutions has been lacking. In birds, the faster evolution of Z-linked loci is consistent with the fixation of slightly deleterious mutations due to the much smaller effective size of the Z *versus* the autosomes (Mank et al. 2009). In mammals and flies, evidence for fast-X evolution is strongest for genes with male-biased or testis-specific expression (Torgerson and Singh 2003; Khaitovich et al. 2005; Torgerson and Singh 2006; Baines et al. 2008). Over the *Drosophila* phylogeny, the distributions of dN/dS and the incidence of positively selected genes are both elevated for some lineages but not others (Singh et

al. 2008). We find similar lineage-restricted evidence for fast-X evolution over the primate phylogeny.

Marmoset is unique among the 4 primate lineages, showing several signals of fast-X evolution. First, the X/A ratio of mean dN/dI significantly exceeds 1, and the X/A ratio for mean dN/dS nearly does (Table 2.1). Second, the distributions of dN/dI and dN/dS are both significantly shifted towards higher values on the X relative to the autosomes. Third, the X shows a significant excess of rapidly evolving genes, an excess that gets stronger as the sample is progressively enriched for genes with histories of positively selected genes (Table 2.3). Among the subset of genes with dN/dI >1, GO term analyses showed no excess of genes with sperm-specific functions on the X (1 on the X, 1 on the autosomes), although these analyses are limited by the lack of experimental genome annotation data from marmoset. Nevertheless, the fact that we observe strong signals of fast-X in an unfiltered data set (with respect to male-biased expression) is highly unusual, and marks marmoset as one of the few, if any, mammalian lineages exhibiting evidence of fast-X evolution.

Why might marmoset differ from the other three primate lineages? The possibility of faster adaptive evolution on the X depends on the strength of male mutation bias (Kirkpatrick and Hall 2004) and on the ratio of effective sizes, N_X/N_A (Vicoso and Charlesworth 2009). Lineage-specific variation in either parameter could therefore give rise to lineage-restricted fast-X evolution. Kirkpatrick and Hall (2004) showed strong male mutation bias— which reduces the mutation rate on the X— impedes fast-X evolution. It is therefore interesting that the marmoset lineage, which has the lowest male mutation bias among the 4 primates investigated, has the strongest signal of fast-X evolution. Long term average N_X/N_A may vary among lineages as well, possibly with mating system. All else being equal, $N_X/N_A > ¾$ facilitates, and $N_X/N_A < ¾$ impedes, fast-X evolution (Vicoso and Charlesworth 2009). Unfortunately, we have little direct empirical knowledge of long-term N_X/N_A in primates. Nevertheless, among the four

primates we analyze, the marmoset is unique in having a polyandry-like mating system (Sussman and Garber 1987). As polyandry entails an increased variance in female reproductive success, there is some reason to expect $N_X/N_A < ¾$ in marmosets. If true, then it is surprising that marmoset provides the best signal of fast-X evolution. Given the number of species in our analysis, our conclusions on the interaction of male mutation bias and N_X/N_A on fast-X evolution must be considered tentative. Direct estimates of N_X/N_A from a range of primates along with lineage-specific estimates of slow- and fast-X evolution are needed for a larger number of primate lineages.

Conclusion

We provide a first simultaneous look at slow- and fast-X evolution as well as their interactions within a primate phylogeny. We show that slow-X evolution is universal among the four primate lineages, although its magnitude varies significantly, mostly because of the generation time effect on male mutation bias. Unlike slow-X evolution, we demonstrate that marmoset is the only species exhibiting compelling general evidence of fast-X evolution, possibly due to its weak male mutation bias compared to the other primates. Finally, we consider the possibility that the variation in the strength of male mutation bias among lineages may influence the estimates of important parameters in human demography.

CHAPTER 3

THE EVOLUTION OF LINEAGE-SPECIFIC CLUSTERS OF SINGLE NUCLEOTIDE SUBSTITUTIONS IN THE HUMAN GENOME

Abstract

Genomic regions harboring large numbers of human-specific single nucleotide substitutions are of significant interest since they are potential genomic foci underlying the evolution of human-specific traits as well as human adaptive evolution. Previous studies aimed to identify such regions either used pre-defined genomic locations such as coding sequences and conserved genomic elements or employed sliding window methods. Such approaches may miss clusters of substitutions occurring in regions other than those pre-defined locations, or not be able to distinguish human-specific clusters of substitutions from regions of generally high substitution rates. Here, we conduct a 'maximal segment' analysis to scan the whole human genome to identify clusters of human-specific substitutions that occurred since the divergence of the human and the chimpanzee genomes. This method can identify species-specific clusters of substitutions while not relying on pre-defined regions. We thus identify thousands of clusters of human-specific single nucleotide substitutions. The evolution of such clusters is driven by a combination of several different evolutionary processes including increased regional mutation rate, recombination-associated processes, and positive selection. These newly identified regions of human-specific substitution clusters include large numbers of previously identified human accelerated regions, and exhibit significant enrichments of genes involved in several developmental processes. Our study provides a useful tool to study the evolution of the human genome.

Introduction

The sequencing of the human genome and its closely related primate genomes allows us to investigate the genomic basis of molecular uniqueness in humans (Chimpanzee Sequencing and Analysis Consortium 2005; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007; Locke et al. 2011; Prufer et al. 2012). One reasonable way to investigate the genetic basis of human-specific traits is to identify genomic fragments that exhibit accelerated nucleotide substitutions confined to the human lineage. For example, the *FOXP2* gene, relevant to human speech and language, was found to be under recent positive selection in the human genome (Enard et al. 2002). Many other studies have identified genes that may have contributed to the evolution of human-specific traits (e.g., (Grossman et al. 2004; Kwiatkowski 2005; Sabeti et al. 2006; Lao et al. 2007)). Acceleration of human-specific nucleotide substitutions is also observed in non-genic regions. For example, Pollard et al. (2006a) analyzed genomic regions that are highly conserved in chimpanzee, mouse, and rat yet exhibit fast evolution in the human genome, referred to as 'human accelerated regions (HARs)'. Among the 202 HARs identified in this study, the most accelerated HAR, referred to as the *HAR1*, was found to be part of a novel RNA gene which is expressed during the development of human neocortex (Pollard et al. 2006b). In another study, a similar approach was used to identify coding exons that exhibit human-specific rate acceleration (Berglund et al. 2009). They identified 83 coding exons that show significantly accelerated nucleotide substitutions in human but are conserved in chimpanzee and macaque. Genes containing these accelerated exons are enriched in "myosin complex", "neurological system process" and "multicellular organismal process" (Berglund et al. 2009). Thus, investigating the distribution of single nucleotide substitutions specific to the human lineage has been highly useful to understanding evolutionary mechanisms that may underlie human-specific genome evolution and human adaptation.

Some studies also examined the evolutionary causes of clustered substitutions without specific *a priori* constraints on genomic regions. For example, (Dreszer et al. 2007) concluded that biased gene conversion in male germline is critical for the evolution of clustered human-specific substitutions. (Capra and Pollard 2011) examined a broader range of species and found that most species exhibit weak-to-strong substitution bias in high substitution density areas and can be well explained by GC-biased gene conversion. One potential caveat of these approaches is the fact that the substitution clusters or substitution density identified are based on arbitrary sizes of sliding windows. In addition, their definition of substitution clusters is based upon the genomic background of the same species and thus does not consider lineage specificity.

In this study, we develop a novel framework for identifying species-specific clusters of single nucleotide substitutions, independent of *a priori* knowledge on genomic regions, nor on arbitrary sliding window sizes. Specifically, we examine the human-chimpanzee-macaque whole genome alignments and identify clusters of human- and chimpanzee-specific substitutions using the 'maximal segment' algorithm (Ruzzo and Tompa 1999). This algorithm identifies the subsequences with regional maximal scores, namely, 'maximal segments'. Using this approach and employing a false-discovery rate based correction method, we identify human maximal segments that have significantly higher substitution rate than both human genomic background and their chimpanzee orthologous sequences. We find that many single nucleotide substitutions in the human lineage since the divergence from the chimpanzee lineage have occurred in close proximity to each other, or in 'clusters'. Furthermore, we show that the evolution of these clusters can be explained by a combination of several different processes, including increase of regional mutation rate, recombination-associated processes, and positive selection. These clusters provide useful tools for studying the genomic basis of human evolution.

Materials and Methods

Genome alignment

Human (hg18, 2006), chimpanzee (panTro2, 2006) and macaque (rheMac2, 2006) genome alignments were extracted and compiled from multiple alignments of 43 vertebrate genomes with human from UCSC genome browser (Kent et al. 2002). The UCSC LiftOver tool was used when converting coordinates from different versions of the human genome assembly (Kent et al. 2002). For example, the coordinates of fine scale map of human recombination from Myers et al. (2005) were converted from hg16 to hg18 to be comparable to our data (Myers et al. 2005).

Identifying maximal segments of species specific substitutions

Using the macaque genome as a reference, we identified human- and chimpanzee-specific substitutions based on parsimony. We then applied the 'maximal segment' algorithm to search for clustered human or chimpanzee-specific substitutions (Ruzzo and Tompa 1999). The maximal segment algorithm was originally designed to find the locally highest scoring contiguous subsequences from a score string. To apply the algorithm, we scored the human- and chimpanzee-specific substitutions based on the following rationale. The null hypothesis is a simple scenario in which species-specific substitutions occur randomly and the probabilities of human-specific substitution and chimpanzee-specific substitution were 50% and 50%. Alternatively, more realistic hypotheses are that one species-specific substitution happened at a probability higher than 50% and the other happened at a probability lower than 50%, depending on which species' maximal segments we were looking for. For example, to find human-specific substitution clusters, we can use the probability of human-specific substitutions in the clusters as 60% (higher than 50%) and the probability of chimpanzee-specific substitutions in the clusters as 40% (lower than 50%). Species-specific substitutions were then scored as log likelihood ratios. Specifically, human-specific substitution in the

trio alignment was scored as log(60/50), a positive score, and chimpanzee-specific substitution was scored as log(40/50), a negative score. Any other sites were scored as 0. By doing so, the trio alignment was transformed to a score string containing information on detailed regional distributions of species-specific substitutions. We then used maximal segment algorithm to find the substrings with the locally highest scores, i.e., the clusters of human-specific substitutions. The clustered chimpanzee-specific substitutions were similarly identified.

To ensure the robustness of our research, we considered the following 2 pairs of different scoring schemes: log(60/50) versus log(40/50), and log(55/50) versus log(45/50). Results from these analyses were presented in the Supplementary Table B.1 and 2, and were qualitatively similar to those presented in this manuscript. Custom perl scripts developed in Yi's laboratory were used for these processes.

To filter out non-significant maximal segments, we first applied a binomial probability criterion. Specifically, in each maximal segment with less than 10% alignment gaps, one species-specific substitution occurred with probability p and the other with probability q, where $q = 1 - p$. p and q could be estimated from the whole genome alignment. For a given maximal segment, the length of the score string was the number of random experiments (n), and the number of positive scores was the number of successful experiments (k). Then this event could be modeled under a binomial distribution because the species-specific substitution happened independently. Under this framework of a binomial distribution, we performed one-tailed tests, calculating the cumulative probability $P(X \geq k)$, which was the P-value of the maximal segment. Then a false discovery rate (FDR) approach was used to control for multiple comparisons (Benjamini and Hochberg 1995). Maximal segments with FDR $Q < 0.1$ were considered to be significant.

Genomic features

Intron sequences were used to approximate neutral evolutionary rate. First, human-chimpanzee-macaque 1:1:1 orthologs were retrieved and assembled from Ensembl Biomart (Vilella et al. 2009). Intron sequences for each gene were downloaded from UCSC genome browser (Kent et al. 2002). To reduce the influence of sites that might be under selection, we removed 100 bps on each side of splice sites. Only introns longer than 300 bps after this procedure were kept. In addition, we masked repeats and hyper-mutable CpG dinucleotides (Kim et al. 2006; Elango et al. 2008). Intron sequences were then aligned using MLAGAN v2.0 (Brudno et al. 2003). Human- and chimpanzee-specific substitution rates for each ortholog trio were estimated using PAML baseml 4.2 (Yang 2007).

Human recombination rate data were obtained from two sources: the megabase-sized deCODE map of the female and male recombination rates of the human genome (Kong et al. 2002), and the kilobase-sized map of recombination rates and hotspots of the human genome (Myers et al. 2005). For each maximal segment, we counted the numbers of weak (A,T)-to-strong (G,C), strong-to-weak, and total substitutions. The proportion of weak-to-strong or strong-to-weak substitutions was calculated as the number of weak-to-strong or strong-to-weak substitutions divided by the number of total substitutions.

Human candidate regions for recent positive selection from each population were downloaded from Voight et al. (2006), where every SNP was assigned an integrated haplotype score (iHS) for three population samples of unrelated individuals: Han Chinese from Beijing and Japanese from Tokyo (CHBJPT, 89 samples), individuals of European origin in Utah, (CEU, 60 samples), and individuals of Yoruba in Ibadan, Nigeria (YRI, 60 samples). Non-overlapping 100 kb windows of top 1% proportions of SNPs having $|iHS| > 2$ were selected as candidate selection regions (Voight et al. 2006). The chromosome ideograms of maximal segments' genomic positions were created using the web tool Idiographica Version 2.0 (Kin and Ono 2007).

Derived allele frequency analyses

Human SNP data were extracted from the 1000G low-coverage data set (The 1000 Genomes Project Consortium 2010), which included SNP information such as the position in the reference genome, ancestral alleles, derived alleles, and the count of derived alleles for the 22 autosomes from three population samples of unrelated individuals: CHBJPT (60 samples), CEU (60 samples), and YRI (59 samples). In this work, we only chose SNPs for which the ancestral allele information was available.

Putatively neutral regions were selected as the intergenic regions that did not contain any maximal segments (regions with *P*-value > 0.1). Moreover, the upstream and downstream 1000bp of such intergenic regions were removed because they may contain non-neutral regulatory elements. Derived allele frequencies were compiled for the maximal segments and compared to those of the putatively neutral regions.

Protein evolutionary rate and likelihood ratio test (LRT)

We used the alignments of 10,376 human-chimpanzee-macaque 1:1:1 orthologs from the macaque genome paper (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007). PAML codeml 4.2 was used to perform likelihood ratio tests (as in Yang (1998)) to identify genes for which the human lineage exhibited a different rate compared to the background lineages (Yang 1998, 2007). We removed genes for which dN/dS > 10 to reduce errors caused by estimation.

Testing for functional enrichment of genes

PANTHER classification system (version 8.0) was used to test for enrichment of different GO categories (Mi et al. 2013). Statistical significance was adjusted to correct for multiple testing using the Bonferroni correction.

Results and Discussion

Clusters of lineage-specific substitutions in the human genome

We identified a large number of clusters of lineage-specific substitutions from the human and chimpanzee genomes. These clusters are also referred to as 'maximal segments' in this manuscript, after the name of the algorithm we used (see Materials and Methods). Statistical significance of maximal segments is determined by a binomial test with FDR control (see Materials and Methods). Genomic positions, GC content and *P*-values of the maximal segments with FDR Q < 0.1 are available to be downloaded at http://dx.doi.org/10.1016/j.ympev.2013.06.003 (click Supplementary data 2).

Table 3.1 lists the proportions of lengths and the numbers of species-specific substitutions within maximal segments compared to genomic background under three different thresholds of statistical significance. Under the scoring scheme of human log(60/50) and chimpanzee log(40/50), the size of all maximal segments ($P < 0.05$) is 21.8% of the size of the aligned sites of human-chimpanzee-macaque whole genome alignment. In comparison, the numbers of single nucleotide substitutions within these maximal segments account for 33.1% of the human-specific substitutions (Chi-square test, $P < 1x10^{-10}$). This indicates that many single nucleotide substitutions are part of the non-random clusters identified in this study. We observe that other significance thresholds and scoring schemes consistently identify approximate 1.3- to 1.5-fold enrichment of single nucleotide substitutions in genomic regions identified as maximal segments in the human genome (Table 3.1). For example, when using a *P*-value cutoff < 0.01, 11.3% of the human genome accounts for 16.5% of human-specific substitutions (Chi-square test, $P < 1x10^{-10}$); and with FDR control at Q < 0.1, 3.5% of the human genome accounts for 4.9% of human-specific substitutions (Chi-square test, $P < 1x10^{-10}$). Our results are thus in agreement with (Schrider et al. 2011) which showed that 1% - 4% of human single nucleotide substitutions occur within short distances of each other. We note that shared polymorphisms between humans and chimpanzees may affect our ability

Table 3.1. Proportions of maximal segments length and species-specific substitutions accounted by maximal segments under two different scoring schemes and three different significance thresholds for human and chimpanzee.

	Human 60/40		Human 55/45		Chimpanzee 60/40		Chimpanzee 55/45	
	Lengths[1]	Subs[2]	Lengths[1]	Subs[2]	Lengths[1]	Subs[2]	Lengths[1]	Subs[2]
P < 0.05	21.8%	33.1%	23.8%	32.4%	18.9%	25.9%	14.8%	18.2%
P < 0.01	11.3%	16.5%	12.3%	16.1%	8.9%	11.6%	5.8%	6.9%
FDR Q < 0.1	3.5%	4.9%	2.7%	3.4%	1.1%	1.4%	0.09%	0.12%

[1]proportions of human maximal segment lengths over total lengths of three species alignments.
[2]proportions of the number of human-specific single nucleotide substitutions belonging to the maximal segments over the total number of single nucleotide substitutions in the examined alignments.

to identify maximal segments. Specifically, our approach becomes less efficient if there exist large numbers of shared polymorphisms between the two lineages. However, given that the amount of shared polymorphism between humans and chimpanzees is extremely low (Halushka et al. 1999) and likely to be localized to regions subject to trans-specific balancing selection (Leffler et al. 2013), the effect of shared polymorphism to our analysis is negligible.

Our further analyses focus on the maximal segments with FDR Q < 0.1 since they are likely to represent highly significant clusters of human-specific substitutions. Over 9000 maximal segments pass this stringent criterion and their mean length is about 5500 bps (Supplementary Table B.1). On average, there are 45 human-specific single nucleotide substitutions and 0.012 substitutions per site in a maximal segment. The number of human-specific single nucleotide substitutions in a maximal segment is negatively correlated with the *P*-value of the maximal segment (Spearman's rho = -0.25, $P < 1x10^{-10}$). Figure 3.1A illustrates the distribution of these maximal segments on each chromosome except the Y chromosome (no maximal segment is identified on the Y chromosome due to the lack of data in non-human primates). We notice that the majority of the maximal segments tend to avoid centromeres but concentrate at the telomeres. Chromosome 2 contains the least amount of maximal segments due to its limited alignment length to that of chimpanzee and macaque. Interestingly, we found the largest number of maximal segments on the X chromosome.

As expected, the total lengths of maximal segments and the lengths of alignments from each chromosome are strongly correlated (Spearman's rho = 0.97, $P < 0.001$, Figure 3.1B). Interestingly, chromosome 19 appears as an outlier according to the regression line in Figure 3.1B. To further investigate this observation, we plot the standardized residuals against the leverage of the linear regression of these two variables (Figure 3.1C) and found that the Cook's Distance of chromosome 19 is around 1, largely deviating from those of the other chromosomes, suggesting chromosome 19 is an outlier

(Chatterjee et al. 2000). In other words, chromosome 19 has fewer maximal segments than expected given its length. This finding is concordant with the observations that gene density on the chromosome 19 is higher than that on the other chromosomes and there exist a large number of evolutionarily conserved non-coding regions compared to the other vertebrates on the chromosome 19 (Grimwood et al. 2004).

Next, we examine the genomic locations of these maximal segments. We find that about 39% of the maximal segments are within the intragenic regions, about 54% are within the intergenic regions, and about 7% are across intragenic and intergenic regions (Figure 3.1A). Interestingly, the 50 maximal segments with the lowest *P*-values show a pattern deviant from this genomic background: we find that the top 50 maximal segments are enriched with those near or inside genic regions (37 overlap with intragenic regions while the expected number is 23, Fisher's exact test, $P < 0.001$). Specifically, ten of these segments are located across intragenic and intergenic regions, with three covering full gene bodies; and the other 27 segments occur completely within gene bodies (Supplementary Table B.3).

Chimpanzee maximal segments are identified by the same approach (Table 3.1 and Supplementary Table B.2). The numbers of chimpanzee specific maximal segments, identified using the same criteria as in humans, are smaller than those of human specific maximal segments, because mutation rates have increased in the chimpanzee genome compared to the human genome (Elango et al. 2006). Regardless, the chimpanzee genome exhibits approximate 1.2- to 1.4-fold enrichment of single nucleotide substitutions in the maximal segments (Chi-square test, $P < 1x10^{-10}$) (Table 3.1). Importantly, there is virtually no overlap between the maximal segments of human and chimpanzee regarding their genome alignment locations. For example, under the 60/40 scoring scheme, only 14 out of 10,811 maximal segments of human and chimpanzee overlap. Under the 55/45 scoring scheme, none of the maximal segments of human and chimpanzee overlap. This indicates that our approach can effectively identify species-

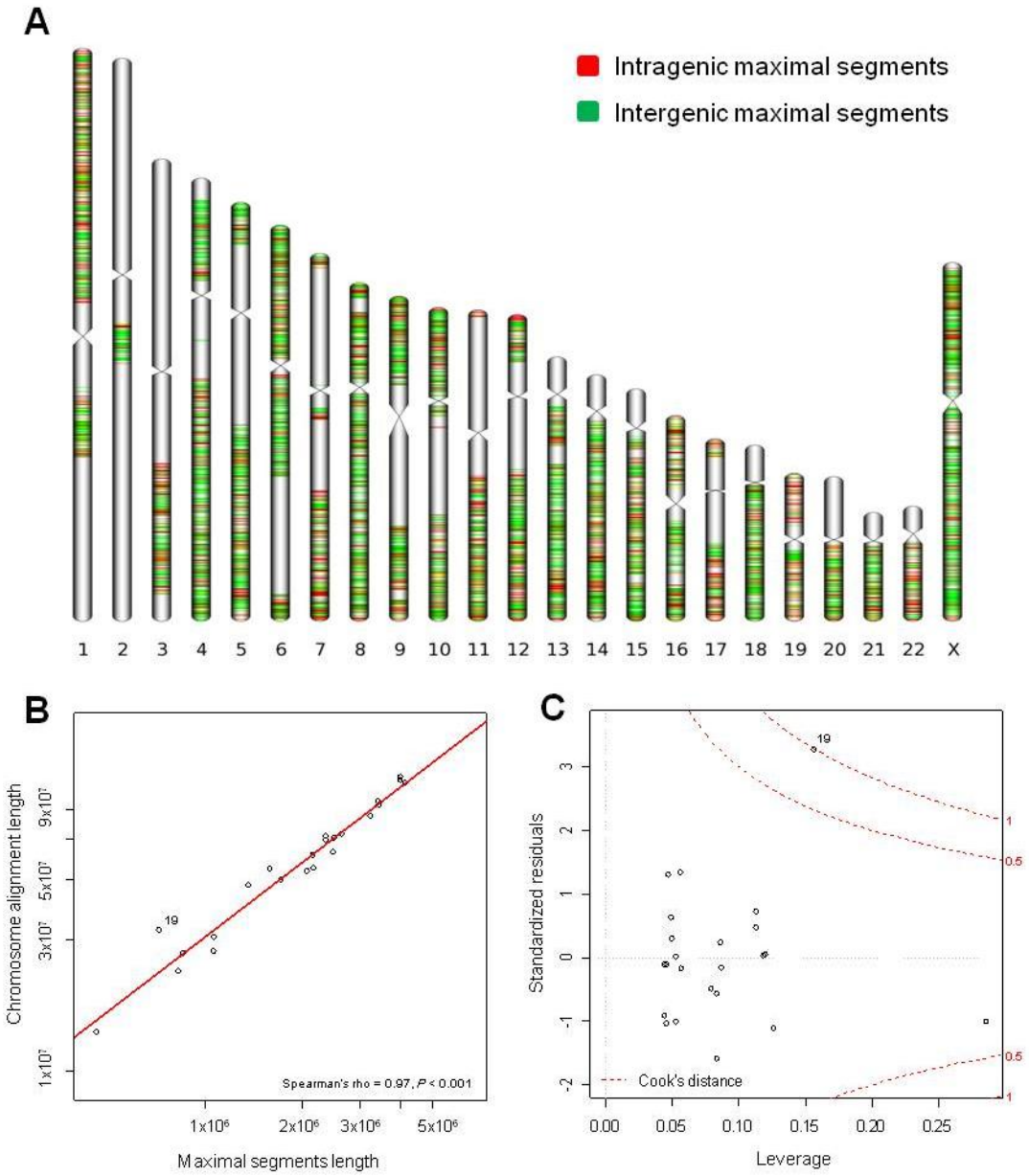

Figure 3.1. Chromosomal distribution and lengths of the identified human maximal segments. A) Chromosomal distribution of human maximal segments with FDR Q < 0.1. Intragenic maximal segments are in red and intergenic maximal segments are in green. B) Linear regression between total length of the maximal segments on each chromosome and the lengths of aligned nucleotides across three species per each chromosome. C) Plot of standardized residuals against leverages of the data points in figure B. The dotted lines represent Cook's distance 0.5 and 1.0. A data point with large Cook's distance indicates its strong influence on the regression function and in our case high likelihood of being an outlier.

specific clusters of substitutions that have occurred since the divergence of humans and chimpanzees. In the following sections, we investigate the nature of evolutionary mechanisms underlying the observed clustering of human-specific substitutions. We hypothesize the following three evolutionary mechanisms as the driving forces of the evolution of the maximal segments: 1) regionally elevated mutation rate; 2) increased recombination rate; and finally, 3) positive selection.

Maximal segments reside in the regions with elevated mutation rate

One possible mechanism by which the observed clusters of single nucleotide substitutions arise is a regional increase of neutral mutation rates, thereby resulting in several nearby substitutions. To investigate this hypothesis, we compare evolutionary rates of maximal segments to those of the genomic background. Specifically, we divide genes into two categories: genes containing at least one maximal segment, and genes containing no maximal segment at all. Genes that contain part of a maximal segment are excluded from this analysis. Intron sequences are used to approximate regional neutral evolutionary rates. For each gene, we calculate numbers of human- and chimpanzee-specific substitutions (Ki) using intron alignments of human-chimpanzee-macaque 1:1:1 orthologs (see Materials and Methods). We find that in the human genome, genes containing maximal segments exhibit significantly higher intron divergence compared to the genomic background (Figure 3.2, median Ki of genes containing maximal segment: 0.00407, median Ki of all the human genes: 0.00373, $P < 0.001$, Wilcoxon rank-sum test). Moreover, genes containing no maximal segments exhibit significantly smaller intron divergence than genomic background (Figure 3.2, median Ki of genes containing no maximal segment: 0.00368, $P < 0.05$, Wilcoxon rank-sum test). Our estimate of the genomic average substitution rate in the human lineage is lower than that from other literature because we removed hyper-mutable CpG sites (Arndt et al. 2003; Chimpanzee Sequencing and Analysis Consortium 2005). Moreover, we observe a parallel pattern in

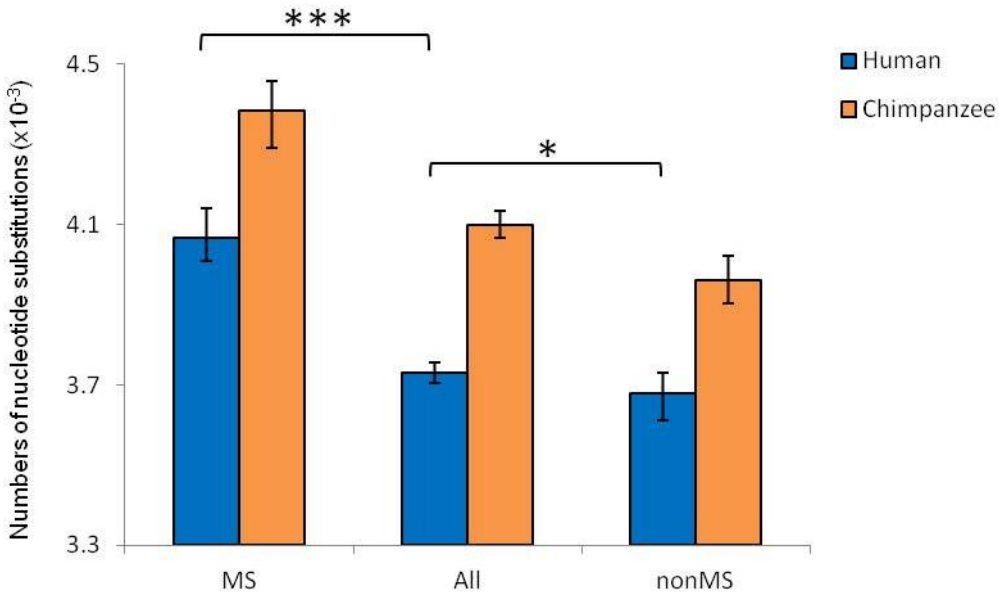

Figure 3.2. Distinctive patterns of intron divergence of human genes among three categories: genes containing maximal segments (MS), all genes (All), and genes containing no maximal segments (nonMS). For each category, median numbers of substitutions ($\times 10^{-3}$) for humans (blue bars) and for chimpanzee orthologs (orange bars) are plotted. Confidence intervals are obtained by bootstrapping 1000 times. *, $P < 0.05$; ***, $P < 0.001$, Wilcoxon rank-sum test.

chimpanzee genomic orthologs. Chimpanzee genomic regions orthologous to human genes containing maximal segments (but not harboring chimpanzee specific maximal segments) also exhibit increased intron divergence compared to other chimpanzee genes (Figure 3.2: median Ki of chimpanzee orthologs: 0.00429, median Ki of all the chimpanzee genes: 0.00410, $P < 0.001$, Wilcoxon rank-sum test). Therefore, human maximal segments tend to reside in genomic regions with regionally increased mutation rates. However, since mutation rates of these regions appear to be increased in both humans and chimpanzees, additional mechanisms are contributing to the evolution of species-specific maximal segments. In the next sections we investigate other evolutionary mechanisms.

Maximal segments occur in the regions with elevated recombination rate

Recombination may facilitate the clustering of human-specific mutations via two mechanisms: first it may increase mutation rates directly due to its mutagenicity, although the degree of this effect is debated (Perry and Ashworth 1999; Lercher and Hurst 2002; Hellmann et al. 2003; Yi et al. 2004). Second, it could increase specific subsets of mutations via biased gene conversion (Strathern et al. 1995). To test whether maximal segments tend to reside in regions of high recombination, we extract the maximal segments' recombination rates from the deCODE recombination map (Kong et al. 2002), which includes both female and male recombination rates across the whole human genome on a megabase-sized window scale. Only the autosomal recombination data are used in our study because the male recombination data are not available for the X chromosome. We find that the genomic regions where the maximal segments reside have an average female recombination rate of 1.80 cM/Mb, which is significantly higher than the female genomic average of 1.60 cM/Mb ($P < 1x10^{-10}$, Wilcoxon rank-sum test). Similarly, the average male recombination rate of the maximal segments is 1.26 cM/Mb, which is also significantly larger than the male genomic average of 0.98 cM/Mb ($P <$

$1x10^{-10}$, Wilcoxon rank-sum test). In addition, we observed a weak but significant negative correlation between the *P*-values of the maximal segments and the recombination rates (Spearman's rho = -0.03, $P < 0.01$ for both female and male). In other words, the more significant a maximal segment is, the higher its recombination rate is. Figure 3.3 illustrates this trend by dividing the maximal segments into 4 groups based on their *P*-values: 'extreme' ($P < 1x10^{-6}$, 81 cases), 'strong' ($1x10^{-6} \leq P < 1x10^{-5}$, 242 cases), 'medium' ($1x10^{-5} \leq P < 1x10^{-4}$, 1297 cases), and 'weak' ($1x10^{-4} \leq P < 1x10^{-3}$, 6841 cases). It shows that the group with the lowest *P*-values (the 'extreme' maximal segments) has significantly higher female and male recombination rates than the group with the highest *P*-values (the 'weak' maximal segments, $P < 0.05$ for female and $P < 0.01$ for male, Wilcoxon rank-sum test).

We perform similar analyses using another recombination data set which is of higher resolution (Myers et al. 2005). The results are highly similar (Supplementary Figure B.1): combining the data from autosomes and the X chromosome together, 1) the average recombination rate of the maximal segments is significantly higher than that of the genomic average (1.74 cM/Mb vs. 1.45 cM/Mb, $P < 1x10^{-10}$, Wilcoxon rank-sum test); 2) there exists a weak but significant negative correlation between the *P*-values of the maximal segments and the recombination rates (Spearman's rho = -0.02, $P < 0.05$). These results indicate the significant role of recombination to cause clustering of human-specific substitutions. We note that we do not find a significant correlation between the *P*-values of the maximal segments and their distances to the nearest recombination hotspots, which can be explained by rapid evolutionary changes of evolutionary hotspots (Ptak et al. 2005; Winckler et al. 2005; Yi and Li 2005).

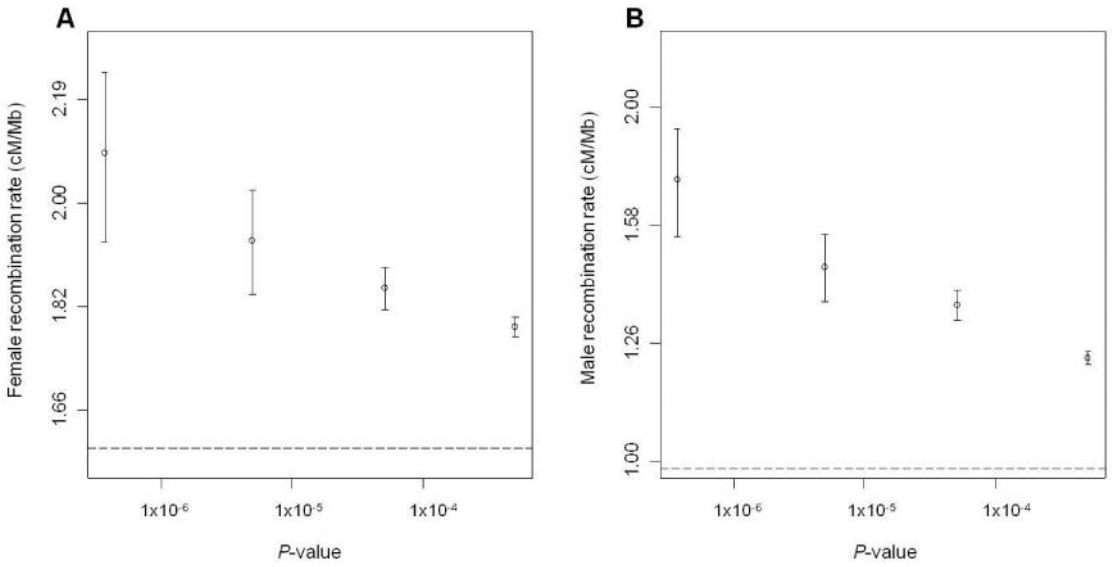

Figure 3.3. Recombination rates from four groups of maximal segments in A) female and B) male. The dashed lines represent the genomic average recombination rates in female and male.

Maximal segments show weak (A, T)-to-strong (G, C) substitution bias, suggesting biased gene conversion

In addition to exhibiting significantly higher recombination rate than genomic background, the maximal segments also exhibit higher GC contents compared to the genomic background. Compared to the genomic average GC content of 41.6% (autosomes and the X chromosome), the GC content is the highest from the 'extreme' maximal segments, followed by the group 'strong', 'medium', and 'weak' maximal segments (Figure 3.4A, Spearman's rho = -0.04, $P < 0.001$).

Given the observation that the maximal segments tend to reside in the regions with both elevated recombination rate and elevated GC content, we hypothesize that this might be due to biased gene conversion processes (a recombination-associated process which favors GC repair when there is a A:C or G:T mismatch (Chen et al. 2007)). To test this hypothesis, we calculate weak (A, T)-to-strong (C, G) and strong-to-weak substitution biases. We find that among the maximal segments, the average proportion of weak-to-strong substitutions is 44.7%, significantly higher than that of strong-to-weak substitutions of 40.6% ($P < 1x10^{-10}$, Wilcoxon rank-sum test). This finding is contradictory to the case in the whole genome, where strong-to-weak substitutions are more prevalent than weak-to-strong substitutions (43.1% and 42.2%, separately). Also, among the maximal segments there exists a weak but significant negative correlation between the *P*-values of the maximal segments and the proportion of weak-to-strong substitutions (Spearman's rho = -0.04, $P < 0.001$). In contrast, the *P*-values of the maximal segments are positively correlated with the proportions of strong-to-weak substitutions (Spearman's rho = 0.03, $P < 0.01$). Figure 3.4B shows that the weak-to-strong substitution bias is the most striking in the 'extreme' maximal segments. With the increase of the *P*-values of the maximal segments, the strength of weak-to-strong substitution bias decreases ('extreme' vs. 'strong': $P < 0.05$; 'strong' vs. 'medium': $P > 0.05$; 'medium' vs. 'weak': $P < 0.01$; Wilcoxon rank-sum test), but even in the 'weak'

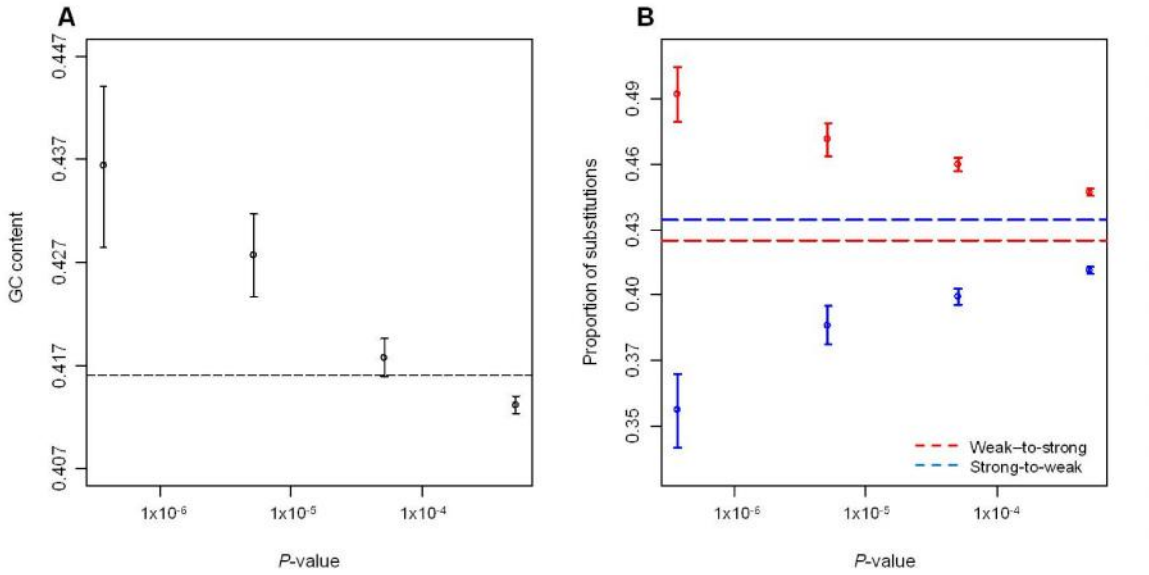

Figure 3.4. GC content and proportions of weak-to-strong and strong-to-weak substitutions in four groups of maximal segments. A) GC content in four groups of maximal segments. The dashed line is the average GC content of the human genome (autosomes and the X chromosome). B) The proportions of weak-to-strong substitutions (red) and strong-to-weak substitutions (blue) in four groups of maximal segments. The dashed red line represents the genome-wide proportion of weak-to-strong substitutions and the dashed blue line represents the genome-wide proportion of strong-to-weak substitutions.

group, this bias still exists compared to the genomic background (Figure 3.4B). These results indicate that the clustering of human-specific substitutions is partially driven by biased gene conversion.

Evidence of natural selection

In this section, we examine whether some of the observed clusters of single nucleotide substitutions evolved by positive selection favoring fixation of multiple mutations or a specific haplotype. To this end, we examine the evidence of positive selection on several different timescale along human evolution: those that occur within positively selected haplotypes (most recent positive selection events), population differences and allele frequency spectra (moderately recent positive selection), and human-specific increase of nonsynonymous substitutions over synonymous substitutions (potentially functional changes that could have originated anytime since the divergence of humans and chimpanzees) (Sabeti et al. 2006). Performing these analyses targeted for different timescale provides us an opportunity to examine the effect of natural selection distributed across different coalescent times in the human evolution.

We first test whether the identified human maximal segments are enriched in recently positively selected haplotypes. Specifically, we examine the overlap between the maximal segments and the regions identified by Voight et al. (2006) as regions under strongest recent positive selection. Voight et al. (2006) identified approximate 250 non-overlapping 100 kb windows that show the strongest signals of recent selection for each of the three human populations, namely CHBJPT, CEU, and YRI (see Materials and Methods). Our analysis does not detect any significant enrichment of the maximal segments in these regions for any of the populations (Supplementary Table B.4). For example, the total length of the strongly selected regions in CHBJPT population accounts for 0.82% of the total human genome length, and the total length of the maximal segments that fall inside the strongly selected regions accounts for 0.9% of the total

maximal segment length. Even when we narrow down the maximal segments to the ones with top 50 lowest *P*-values, there is no evidence of any increase of the strongly selected regions in the maximal segments.

In the second approach, we examine the derived allele frequency (DAF) spectra of the maximal segments. For this analysis, we divide the maximal segments into intragenic and intergenic maximal segments. We then examine the DAF spectra of these two groups of maximal segments and compare them to those from the putatively neutral regions, i.e., intergenic regions that do not overlap with any maximal segments with *P*-value < 0.1, with 1000 bps flanking these regions on both ends removed (see Materials and Methods). We use the 1000G low-coverage data set released in 2010 (The 1000 Genomes Project Consortium 2010). The results of this analysis are shown in the Supplementary Figure B.2. In all three populations, the intragenic maximal segments (grey bars) exhibit a significant excess of low-frequency derived alleles compared to the putatively neutral regions (black bars) (Chi-square test, $P < 0.001$), indicating that the intragenic maximal segments tend to be under strong purifying selection (Nielsen 2005). The intergenic maximal segments, in contrast, exhibit DAFs that are highly similar to those from the putatively neutral regions (Supplementary Figure B.2). We also examine DAFs of the most significant 50 maximal segments in inragenic and intergenic regions, separately (Figure 3.5). Interestingly, the SNPs in top 50 intragenic maximal segments show a decrease in frequency of low-frequency derived alleles and an excess of high-frequency derived alleles in all three populations (except for the slight decrease of high-frequency derived alleles in YRI), indicating that these highly significant intragenic maximal segments are under recent positive selection (Nielsen 2005) (Figure 3.5, left panels). In contrast, the top 50 intergenic maximal segments show signatures of strong purifying selection in all three populations except for YRI (Figure 3.5, right panels).

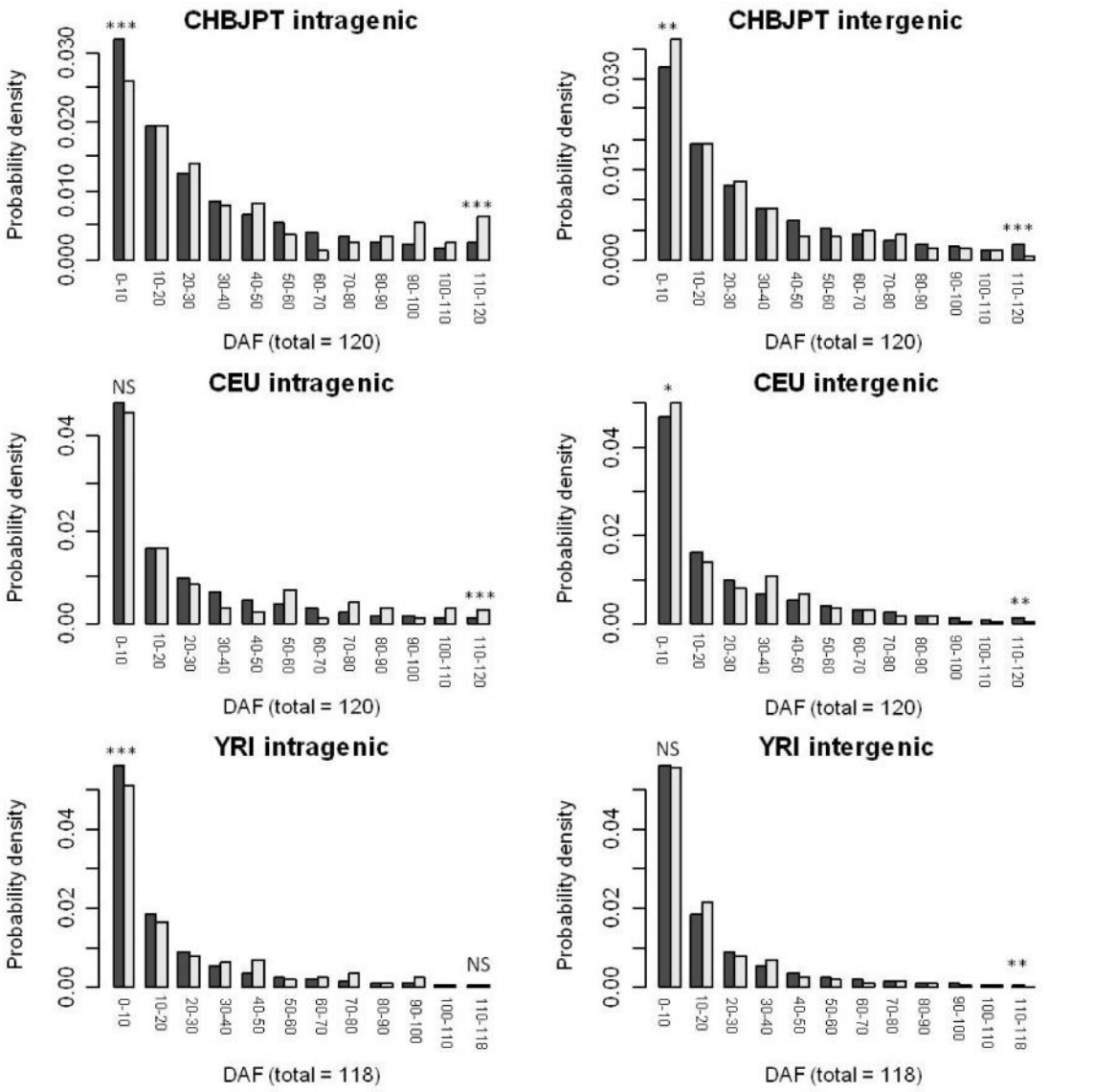

Figure 3.5. Derived allele frequency (DAF) spectra of the top 50 most significant intragenic maximal segments (left panels) and the top 50 most significant intergenic maximal segments (right panels) for three populations (CHBJPT, CEU and YRI, Materials and Methods). The grey bars represent DAFs of the maximal segments and the black bars represent DAFs of the putatively neutral regions. NS, not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Third, we investigate signatures of positive selection in coding sequences. Specifically, we use 10,376 human-chimpanzee-macaque ortholog trios (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007) and perform log-likelihood ratio tests (LRT) to identify genes that experienced accelerated protein evolution in the lineage leading to modern humans (see Materials and Methods). Genes that overlap with the maximal segments are defined by three different criteria: 1) genes with at least one exon inside a maximal segment; 2) genes with at least one intron covering a maximal segment; and 3) genes that satisfy both of the criteria. Table 3.2 lists the frequency of genes identified to have experienced accelerated protein evolution in human in each category. Genes overlapping with maximal segments exhibit overall increased frequencies of positively selected genes. For example, genes belonging to the first category (harboring at least one exon inside a maximal segment) exhibit a 1.4- fold increase compared to all the genes, while genes belonging to the second category (with at least one intron covering a maximal segment) a 1.5-fold increase. Maximal segment genes identified under the most stringent criterion (with at least one exon inside a maximal segment and at least one intron covering another maximal segment) exhibit a 2.5-fold increase compared to all the genes. These observations suggest that clusters of human-specific substitutions underwent positive selection during the history of human evolution.

We also compare our results to the results from the rhesus macaque genome analyses (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007) in which they used a slightly different approach to identify positively selected genes in the human lineage. They identified 16 positively selected genes in the human lineage, among which only two genes, leukocyte immunoglobulin-like receptor LILRB1 and hypothetical protein LOC399947, passed their stringent FDR < 0.1 criterion (Table 3.2). We find that three out of the 16 positively selected genes have at least one exon inside a maximal segment, and that LILRB1 and LOC399947 are among these three genes. In

Table 3.2. Overlap between maximal segments and accelerated sequence evolution of genes.

Genes	Genes identified to have experienced accelerated evolutionary rate by LRT	Positively selected genes identified by macaque genome paper[4]
All genes	251/9003 = 2.8%	**LILRB1**, MAGEB6, FLJ35880, ICAM1, C20orf96, **LOC399947**, TCRA, RP11-558F24.1-001, IFNA8, MRPL39, RP11-98I9.1-002, GOLGA4, HUS1B, DCXR, PRM1, TMCO5
Genes with exons in MS[1]	23/586 = 3.9%	**LILRB1**, FLJ35880, **LOC399947**
Genes with introns covering MS[2]	24/568 = 4.2%	None
AND[3]	9/127 = 7.1%	None

[1] Genes with at least one exon inside a maximal segment
[2] Genes with at least one intron covering a maximal segment
[3] Genes with at least one exon inside a maximal segment AND at least one intron covering another maximal segment
[4] Genes in bold are the genes that passed stringent FDR < 0.1 criterion (Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007)

particular, LILRB1 resides in a region covering one of the most significant (top 50) maximal segments.

Of the top 50 most significant maximal segments, three of them cover entire gene bodies (Supplementary Table B.3). Among these three maximal segments, the most significant one covers a human leukocyte antigen (HLA) gene HLA-DQB1, which is essential for human immune system yet is highly polymorphic (de Bakker et al. 2006). Its variation is related to increased risk of several autoimmune and infectious diseases (Jones et al. 2006) and has been identified to be under recent positive selection (The International HapMap Consortium 2007). The other two maximal segments cover two non-coding RNAs (KIAA1967 and FLJ14107) and a microRNA (MIR600), separately. The genomic positions of the top 50 maximal segments and their associated genes are presented in the Supplementary Figure B.3.

Maximal segments contain significant amount of previously identified human accelerated regions

Previous studies identified several genomic regions with accelerated evolutionary rate in human but conserved in the other species, for example, human accelerated regions (HARs) and human accelerated exons (HAEs) (Pollard et al. 2006a; Berglund et al. 2009). These approaches are dependent on *a priori* information, namely the identities of phylogenetically conserved genomic regions and annotated (and also phylogenetically conserved) coding exons. In comparison, our method is independent of any *a priori* information, and as a result, less restrictive. Therefore, we hypothesize that our maximal segments framework can detect many previously reported HARs and HAEs. Indeed, we find that among the 202 HARs with total length 35,099 bp, 25% of them are included in the maximal segments. Given the fact that the maximal segments are only 1.6% of the human genome, this represents a highly significant enrichment ($P < 1x10^{-10}$, Chi-square test). Similarly, among the 83 HAEs with total length 42,875 bp, 9.0% of them are

included in the maximal segments, again representing a significant enrichment ($P < 1x10^{-10}$, Chi-square test).

Genes overlapping with maximal segments are enriched in several gene ontology categories

We study functional importance of the maximal segments by analyzing genes overlapping with the maximal segments using the PANTHER classification system (Mi et al. 2013). We use the same list of genes from the third category described in Table 3.2, i.e., those containing at least one exon inside a maximal segment and at least one intron covering another maximal segment. Significant enrichments are found for several gene ontology (GO) categories, among which the top five most significant categories are strikingly representative of development-related functions: nervous system development, ectoderm development, and system development (Table 3.3). Supplementary Table B.5 lists all the significant GO categories for these genes. In addition, we also find that multiple significant GO categories in our study belong to the categories enriched by previously identified positively selected genes, including sensory perception, cell adhesion, signal transduction, transport, and developmental processes (Clark et al. 2003; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007).

Concluding Remarks

We identified a significant amount of human-specific clusters of single nucleotide substitutions using a previously developed 'maximal segment' algorithm. The advantages of using this algorithm to identify lineage specific clusters of substitutions are the facts that it does not rely on pre-defined regions or arbitrary window sizes, and relatively free from regional differences in mutation rates (as it only considers enrichment compared to orthologous regions in different species). Consequently, the resulting list of 'maximal segments' represents an unbiased, comprehensive list of genomic hotspots of

Table 3.3. Top 5 significant GO terms for genes with at least one exon inside a maximal segment and at least one intron covering another maximal segment.

	Observed number of genes	Expected number of genes	*P*-value[1]
Biological Process			
Ectoderm development	72	25.98	3.06×10^{-13}
Nervous system development	66	22.9	1.00×10^{-12}
Cell-cell adhesion	48	12.9	1.10×10^{-12}
System development	84	34.84	1.98×10^{-12}
Cell communication	128	68.91	7.17×10^{-12}
Molecular Function			
Receptor activity	58	29.33	6.25×10^{-5}
Protein kinase activity	25	8	1.09×10^{-4}
Kinase activity	28	10.51	5.00×10^{-4}
Transporter activity	34	16.05	5.43×10^{-3}
Glucosidase activity	4	0.28	2.82×10^{-2}
Cellular Component			
Extracellular matrix	21	8.9	1.09×10^{-2}
Extracellular region	21	8.92	1.11×10^{-2}
Cell junction	10	2.59	1.19×10^{-2}
Plasma membrane	10	2.77	2.02×10^{-2}
Actin cytoskeleton	18	7.44	2.16×10^{-2}

[1] *P*-values are adjusted for multiple comparisons using the Bonferroni correction

single nucleotide substitutions. Further analyses of these maximal segments reveal the importance of neighboring effects, increased regional mutation rates, recombination (in particular biased gene conversion) and positive selection at the origin of these regions.

In addition to the advantages listed above, the framework presented in this paper can be flexibly modified to suit different species divergence (by varying weights to lineage specific substitutions) and to large phylogenies. Thus we envision that similar frameworks can be applied to other phylogenies to identify species-specific clusters of single nucleotide substitutions. Nevertheless, a limiting factor of current comparative genomic approaches (including our approach) to identify lineage-specific evolutionary events is the fact that they generally compare a single genome of one species to a single species in another species, largely due to the lack of data on intra-specific genomic variation. Analyses of intragenomic variation in humans, as presented in this study, indeed elucidate the significant roles of positive selection on several different timescales of human evolution in shaping human-specific clusters of single nucleotide substitutions.

CHAPTER 4

PARALLEL EVOLUTION OF GENOMES AND EPIGENOMES BETWEEN HUMANS AND CHIMPANZEES

Abstract

DNA methylation is a crucial epigenetic modification that is phylogenetically widespread. Despite its well-appreciated functional significance, how DNA methylation patterns evolve between species is not well understood. Recently, two studies mapped genome-wide DNA methylation levels of humans and chimpanzees from sperm and prefrontal cortex, separately. Their data suggest that methylation patterns are different between species and between tissues. To understand genetic basis of the methylation divergence, in this chapter we have examined several genomic characteristics that are hypothesized to affect DNA methylation levels. Focusing on the two species' promoter regions, we find that there exists a negative correlation between their methylation level difference and their CpG content difference in the sperm but not in the brain, which is consistent with the fact that CpG contents reflect methylation patterns better in germline cells. We do not find an expected negative correlation between promoter methylation level difference and gene expression level difference in either tissue. We demonstrate that Alu repeats in promoter regions are mosaically methylated in the sperm but globally methylated in the brain. We show that regional methylation levels can be affected by genomic factors such as transposable element insertions and single nucleotide substitutions. Interestingly, CpG sites generated by single nucleotide substitutions appear to be 2-3 times more methylated than other CpG sites in both tissues of the two species.

Introduction

DNA methylation is an essential epigenetic modification that is available across a broad range of taxa. Studies have shown that DNA methylation is a crucial regulatory mechanism which is involved in various transcriptional processes such as suppressing gene expression (Kass et al. 1997; Siegfried et al. 1999), reducing transcriptional noise (Huh et al. 2013), and suppressing propagation of transposable elements (Meunier et al. 2005). Despite its functionality, DNA methylation pattern is significantly diverged between different species. For example, most vertebrate genomes' are globally methylated (~80% CpG sites are methylated), yet invertebrate eukaryotes such as *Drosophila melanogaster* and *Caenorhabditis elegans* lack DNA methylation (Tweedie et al. 1997; Suzuki and Bird 2008). Recent studies also start to reveal that even between closely related species methylation patterns are substantially diverged compared to the divergence between their genomes. Specifically, DNA methylation level in prefrontal cortex is found to be significantly lower in humans than in chimpanzees (Zeng et al. 2012) and DNA methylation level in sperm is found to be significantly higher in humans than chimpanzees (Molaro et al. 2011). Several factors including genomic composition, environmental influence and sampling effects could contribute to differential methylation patterns between closely related species.

Previous studies suggest that DNA methylation status can be determined by nearby genetic elements. For example, Lienert et al. conducted a series of experiments on a specific genomic locus in mouse stem cells demonstrating that DNA methylation of promoter sequences can be regulated by their proximal sequence elements (Lienert et al. 2011). On a broader scale, Gibbs et al. identified a large number of quantitative trait loci (QTLs) for DNA methylation and mRNA expression in four different regions of human brains (Gibbs et al. 2010). They found that DNA methylation QTLs are considerably closer to their target CpG sites compared to the distances between expression QTLs and individual transcription start sites. Given these evidences showing that some signals for

methylation status may reside in genomes, we hypothesize that we might be able to infer why and how DNA methylation levels change in evolutionary time scale from genomic sequence comparisons of human and chimpanzee. Specifically, we investigate three potential genetic determinants which may contribute to the epigenetic differences, namely, transposable elements insertions, CpG depleting mutations due to deamination, and single nucleotide substitutions generating lineage-specific CpG sites. Since relatively abundant comparative expression data sets are available between humans and chimpanzees, we also examine whether methylation levels change in promoter regions are related to gene expression levels change. So this comparison provides us a chance to examine the role of genomic changes on epigenetic divergence as well as gene expression divergence.

Our results show that between humans and chimpanzees, promoter methylation divergence is negatively correlated with CpG content divergence in the sperm but not in the brain. We find no expected negative correlation between promoter methylation divergence and gene expression divergence in neither of the tissues. We demonstrate that transposable element insertions, CpG depleting mutations and CpG generating mutations contribute to the evolution of DNA methylation in humans and chimpanzees.

Materials and methods

Sequence alignments

Whole genome alignment of human (hg19), chimpanzee (panTro2) and orangutan (ponAbe2) was extracted from multiple alignments of 45 vertebrate genomes against the human genome from UCSC Genome Browser (Meyer et al. 2012a). We first extracted human-chimpanzee-orangutan genome alignment in MAF format from the 46 species genome alignment using Kent Source Unities (Kuhn et al. 2012). After we converted the MAF format to FASTA format (Blanchette et al. 2004), we distinguished the empty parts of the alignment blocks as either incomplete sequencing or insertions/deletions using a

custom Perl script based on the annotation information in MAF alignment blocks. We then integrated this information into the FASTA alignment.

Human-chimpanzee-orangutan 1:1:1 orthologs were assembled from Ensembl Biomart (Vilella et al. 2009). 15232 trios were obtained. Promoter sequence was defined as upstream 2000 bp of transcription start site (TSS) to the end of first exon. Promoter sequence alignments were extracted from the 3 species' whole genome alignment using human promoter coordinates as reference. Because the 3 species' whole genome alignment does not include information of sequence deletions in human and sequence insertions in chimpanzee, we mapped the aligned chimpanzee and orangutan sequences back to their own genomes using BLAT to obtain the original chimpanzee and orangutan sequences (Kent 2002). For CpG O/E related analysis, we removed the orthologs if any one of the aligned length fractions of 3 species is smaller than 50%. 13512 trios were left after the filtering.

Intron sequence alignments were extracted from the 3 species' whole genome alignment using human intron coordinates as reference. Since the first intron and intron sites near the splicing sites may contain regulatory elements and not neutral (Majewski and Ott 2002), we removed the first intron and 100 bps on both ends of the rest of the introns. After this procedure, we also removed genes whose combined introns' lengths are shorter than 300 bps.

CpG O/E

CpG O/E is a normalized measure for CpG dinucleotide content. Previous studies indicate that CpG O/E is strongly negatively correlated with DNA methylation level and therefore can be used as a genomic indicator of methylation status (Weber et al. 2007; Suzuki and Bird 2008). CpG O/E is calculated as:

$$CpG\ O/E = \frac{P_{CpG}}{P_C * P_G}$$

where P_{CpG} is the frequency of CpG dinucleotide and P_C and P_G are the frequencies of cytosine and guanine.

CpG generating mutations

CpG generating site is defined as a species-specific CpG site due to a single nucleotide substitution to cytosine or guanine. For example, a human CpG generating site would be a human-specific substitution to cytosine or guanine followed by a conserved guanine site or preceded by a conserved cytosine site. Given the short evolutionary distances between human, chimpanzee and orangutan, we used parsimony method to infer species-specific substitutions. Meunier and Duret demonstrated the credibility of such method despite of the hypermutability of CpG sites (Meunier and Duret 2004). We used the same definition to identify all 16 types of dinucleotide generating sites in the human genome. Then we were able to count the number of each type of dinucleotide generating sites for all the human promoters.

Since one human-specific substitution can give rise to two types of dinucleotide generating sites, for example, ACG in human and ATG in chimp and orangutan, these two types of dinucleotide generating sites are not independent with each other. To make different dinucleotide generating sites independent with each other, we can only make two types of dinucleotide generating sites independent with each other at a time. For example, if we want to make CpG and GpC generating sites independent with each other, we will not count the dinucleotide generating sites involving triplets CGC and GCG in the target species.

CpG depleting mutations

CpG depleting mutation is defined as a mutation from CpG dinucleotide to other dinucleotide due to single nucleotide substitution. To identify such mutations, we simply reverse the process for identifying CpG generating mutations. For example, an aligned

site would be identified as a CpG depleting mutation in human if CpX or YpG is found in human and CpG is found in chimpanzee and orangutan. The 'X' could be any nucleotide other than guanine and "Y" could be any nucleotide other than cytosine. Spontaneous deamination of 5-methylcytosine to thymine, i.e., mutation from CpG to TpG, is one of the CpG depleting mutations.

DNA Methylation level

To get DNA methylation levels in promoter regions of humans and chimpanzees, we downloaded methylation information of mapped cytosines in the context of CpG dinucleotides from the brain (Zeng et al. 2012) and the sperm (Molaro et al. 2011) of humans and chimpanzees. First, we calculated fractional methylation value for each mapped cytosine site as #C/(#C + #T), where #C is the total number of cytosine reads and #T is the total number of thymine reads. Then fractional methylation values of all the mapped cytosine sites within a promoter region were averaged, which is the promoter methylation level. As methylation information of mapped cytosines from human sperm is based on genome assembly hg18, we converted the coordinates of our human promoter regions from hg19 to hg18 using UCSC LiftOver tool (Kuhn et al. 2012).

To get DNA methylation levels of aligned promoter regions, we followed the same principle. Original promoter regions were first divided into several segments, each of which must be contiguous and contains no insertions/deletions. Then we averaged the fractional methylation values of mapped cytosines on all the aligned segments of a promoter to get the methylation level of aligned promoter regions.

Data standardization

CpG O/E, methylation level, and gene expression level were standardized before calculating correlations among the differences of these variables between human and chimpanzee. We used the following equation to standardize our data:

$$z = \frac{x - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of the original values.

Identifying lineage-specific transposable element insertions

To identify human- and chimpanzee-specific transposable insertions in their promoters, we first used RepeatMasker (Smit et al. 1996-2010) to identify all the Alu and SVA elements in the orthologous promoter regions of human, chimpanzee, and orangutan. Then, we used an in-house Perl script to pick out the candidate Alu and SVA insertions, i.e., Alu and SVA elements that are identified in one species but not in the other two species. Finally, we manually checked each candidate insertion using UCSC genome browser, removing the low-quality Alus and SVAs identified by RepeatMasker as well as the partial insertions.

Results and Discussion

Methylation level difference co-varies with CpG content difference between humans and chimpanzees

With the growing interest in DNA methylation and the advent of high-throughput bisulfite sequencing technology, more and more single-base-resolution DNA methylomes are being mapped. Recently, two studies mapped DNA methylomes of humans and chimpanzees in sperm (Molaro et al. 2011) and brain (prefrontal cortex, Zeng et al. 2012), separately. Their comparative analyses of the DNA methylomes revealed surprising results: methylation divergence is much greater than genomic divergence between humans and chimpanzees in both tissues (Molaro et al. 2011; Zeng et al. 2012).

Data from Molaro et al. (2011) suggest that methylation level is generally higher in human sperm than in chimpanzee sperm, while data from Zeng et al. (2012) suggest an opposite direction in the brain (Table 4.1). CpG dinucleotide content, measured as CpG

Table 4.1. CpG Methylation level in promoter region between species and between tissues.

		Human sperm	Chimp sperm	**H/C ratio in sperm**	Human brain	Chimp brain	**H/C ratio in brain**
Whole promoter regions	Median	0.118	0.105	**1.124**	0.286	0.327	**0.875**
	Mean	0.283	0.271	**1.044**	0.392	0.417	**0.940**
Removing INDELs	Median	0.118	0.099	**1.192**	0.269	0.298	**0.903**
	Mean	0.284	0.270	**1.052**	0.386	0.404	**0.955**
Removing INDELs and lineage-specific CpGs	Median	0.113	0.095	**1.189**	0.263	0.291	**0.904**
	Mean	0.282	0.269	**1.048**	0.383	0.400	**0.958**

O/E (see Materials and Methods), has been constantly shown to be strongly correlated with methylation level and widely used as a genetic indicator of methylation status (Weber et al. 2007; Elango and Yi 2008; Sarda et al. 2012). We hypothesize that the differential methylation pattern between humans and chimpanzees could be partly explained by their CpG dinucleotide content difference. In this work, we focused on promoter regions because DNA methylation in promoter regions exhibits a bimodal distribution and is known to play a significant role in gene regulation in mammals (Kass et al. 1997; Siegfried et al. 1999; Klose and Bird 2006; Elango and Yi 2008). We plotted promoter methylation levels against CpG O/E in human sperm, chimpanzee sperm, human brain, and chimpanzee brain, separately (Figure 4.1A). As expected, strong negative correlations are observed between these two variables for all four pairs of data (Spearman's rho = -0.84 for human sperm, -0.84 for chimpanzee sperm, -0.75 for human brain, and -0.73 for chimpanzee brain, $P < 1x10^{-10}$). Next, we investigate whether the methylation level difference between human and chimpanzee co-vary with their CpG O/E difference. For both sperm and brain, we calculate the correlation between cpgoeDiff (as standardized human CpG O/E – standardized chimpanzee CpG O/E) and methyDiff (as standardized human methylation level – standardized chimpanzee methylation level). We find that for sperm, there exists a significantly negative correlation between cpgoeDiff and methyDiff (Pearson correlation coefficient in all data = -0.12, $P < 1x10^{-10}$, Figure 4.1B); yet for brain, there exists a weak but significantly positive correlation between cpgoeDiff and methyDiff (Pearson correlation coefficient in all data = 0.04, $P < 0.001$, Figure 4.1B).

Since strong correlation exists between CpG O/E and methylation level, we fit these two variables with linear regression lines for human sperm, chimpanzee sperm, human brain, and chimpanzee brain, separately (Figure 4.1A). We notice that the slopes between human sperm and chimpanzee sperm are more similar than the slopes between human brain and chimpanzee brain. To compare the regression slopes between humans

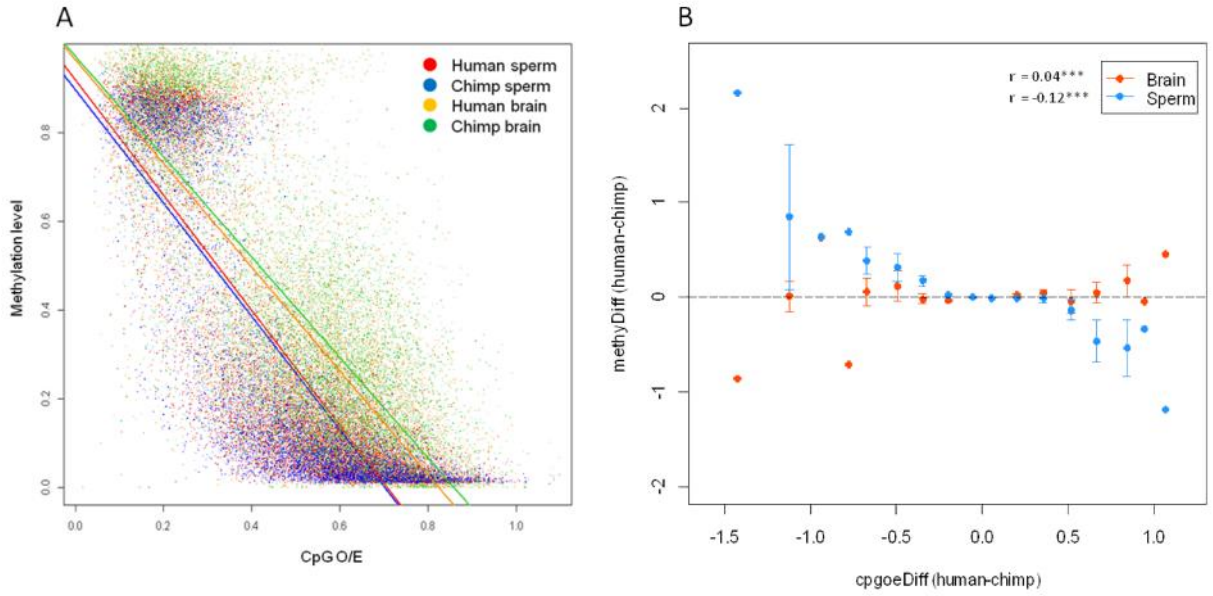

Figure 4.1. Correlations between CpG O/E and methylation level in different tissues of humans and chimpanzees. A) Correlations between CpG O/E and methylation level for human sperm, chimpanzee sperm, human brain, and chimpanzee brain, represented by different colors separately. Each of the four correlations is fitted by a linear regression line in a same color; B) correlations between CpG O/E difference and methylation level difference between humans and chimpanzees for the sperm and the brain, separately.

and chimpanzees within a same tissue, we use ANCOVA method. The results from ANCOVA confirm that the slopes between humans and chimpanzees are not significantly different in the sperm but significantly different in the brain, which is consistent with our observation that there exists a significantly negative correlation between methyDiff and cpgoeDiff in the sperm but not in the brain. The expected negative correlation between methyDiff and cpgoeDiff is only observed in the sperm could be due to two mutually non-exclusive reasons: 1) CpG O/E difference directly reflects DNA methylation difference in germlines; and/or 2) cellular heterogeneity of samples, especially the prefrontal cortex may cause bias in the estimation of methyDiff.

The publicly available comparative gene expression data provide us opportunities to examine correlations between human-chimpanzee methylation divergence and gene expression divergence. Brawand et al. (2011) thoroughly measured gene expression level of multiple species including human and chimpanzee from multiple tissues including brain and testis, allowing us to investigate the influence of methylation level difference between humans and chimpanzees on their gene expression level difference in the brain and the sperm. Since no comprehensive sperm expression data is available for humans and chimpanzees, gene expression levels from the testis are used to approximate gene expression levels of the sperm.

Consistent with previous studies showing that DNA methylation in promoters suppress gene expression (Klose and Bird 2006), we observe strong negative correlations between DNA methylation levels in promoter regions and downstream gene expression levels (Spearman's rho = -0.37 for both human brain and chimpanzee brain, -0.23 for human sperm and -0.29 for chimpanzee sperm, $P < 1x10^{-10}$). We expect that promoter methylation level change would result in gene expression level change; however, we do not find significant correlations between promoter methylation difference (as standardized human methylation level – standardized chimpanzee methylation level) and gene expression level difference (as standardized human expression level - standardized

chimpanzee expression level) in either brain or sperm. This could be due to the reason that these data sets are from different experiments using different samples.

In the following sections, we investigate what genetic factors may influence methylation level and CpG O/E in promoter region of humans and chimpanzees. We focus on two hypotheses: 1) transposable element insertions and 2) CpG generating mutations.

Alu repeats are differentially methylated in the sperm and the brain

Over half of the human genome is occupied by transposable elements, among which Alu repeats are the most abundant (International Human Genome Sequencing Consortium 2001). Previous studies using blot hybridization technique have shown that while Alus are completely methylated in human somatic tissues, some of the Alus are hypomethylated in male germ line (Hellmann-Blumberg et al. 1993; Rubin et al. 1994). We hypothesize that this might be a global pattern in both humans and chimpanzees and could contribute to the opposite correlations between CpG divergence and methylation divergence observed in the sperm and the brain.

To test this hypothesis, we extract the coordinates of all the Alu repeats within the promoter regions for both species from UCSC Table Browser (Karolchik et al. 2004). After discarding Alus shorter than 200bps, about 7,300 copies are left in both humans and chimpanzees. We then calculate average CpG methylation level for each Alu repeat. We find that for both species the methylation levels of Alus show bimodal distributions in the sperm (Figure 4.2A and C), meaning that some of the Alus are methylated but some are hypomethylated. In comparison, all the Alu repeats tend to be methylated in the brain of both species (Figure 4.2B and D). Thus, this is consistent with our hypothesis that Alus are globally differentially methylated in the sperm and the brain for both species.

Next, we examine why methylation levels of Alu repeats show bimodal distribution in the sperm. We predict that the hypomethylated Alus are relatively

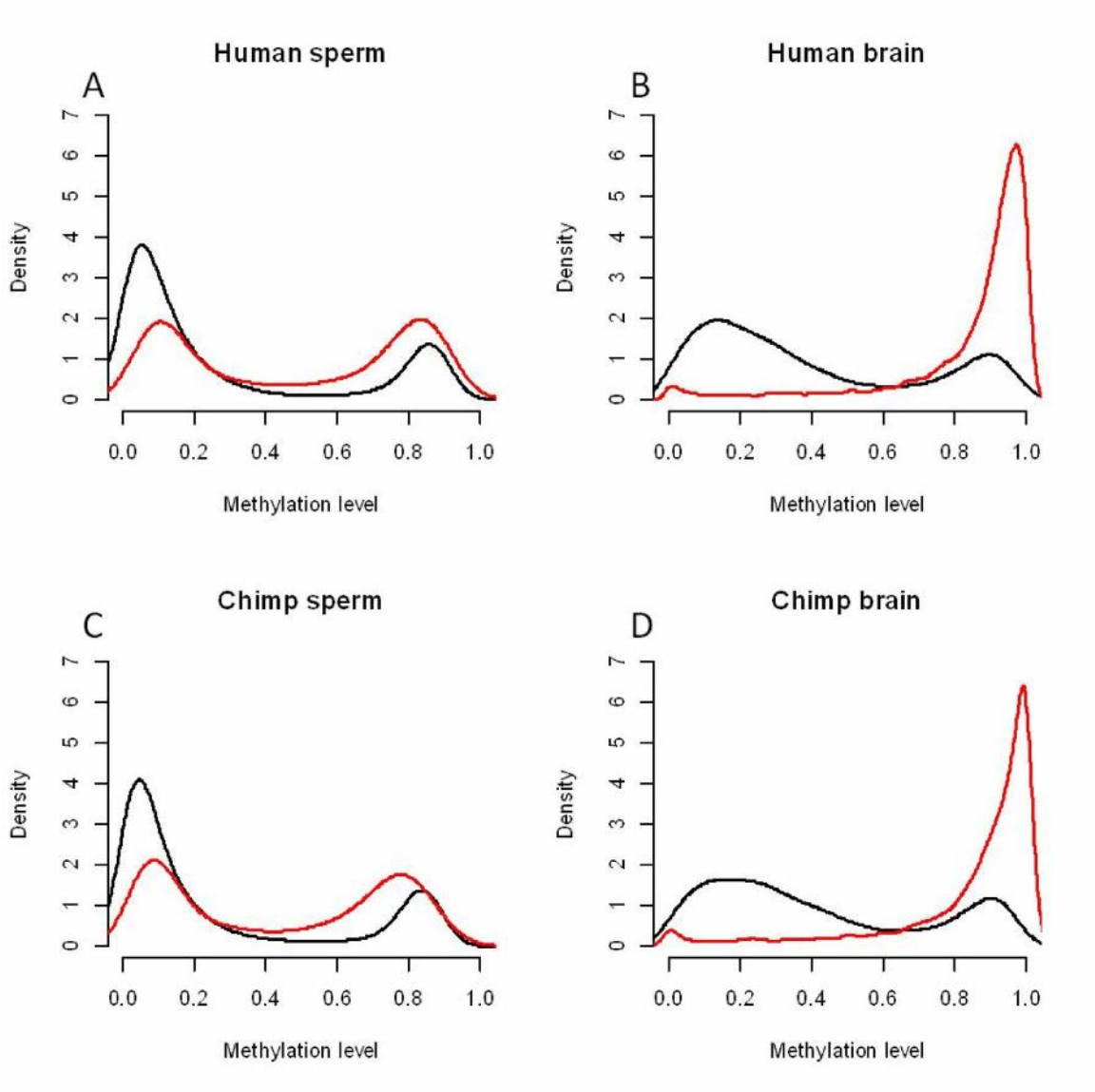

Figure 4.2. Distribution of methylation levels of Alu repeats in promoter regions. In each panel, red density curve represents distribution of methylation levels of Alu repeats in promoter regions and black density curve represents distribution of methylation levels of promoters.

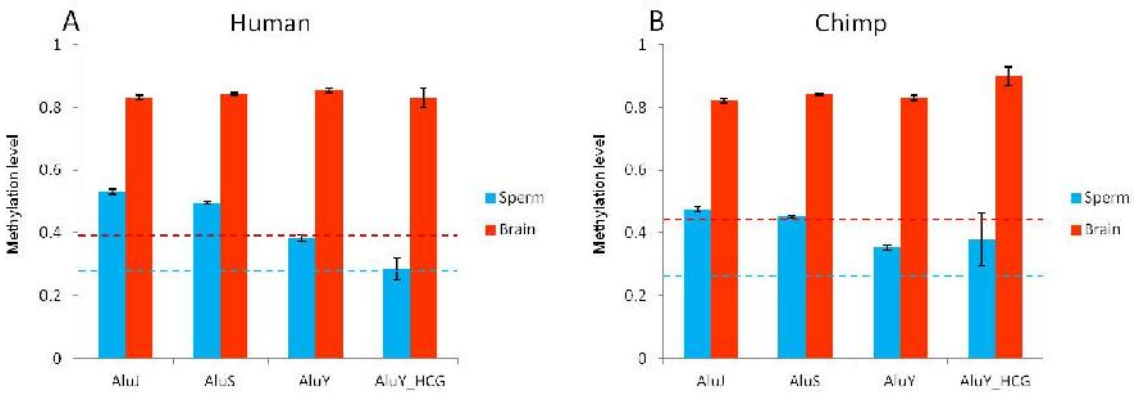

Figure 4.3. Average methylation levels of different groups of Alu repeats in promoter regions in the sperm and the brain of A) humans and B) chimpanzees. AluJ represents the oldest Alus, including AluJr, AluJr4, AluJo, and AluJb; AluS represents the second oldest Alus, including AluSz and its derived subfamilies prior to AluY; AluY represents the second youngest Alus; and AluY_HCG represents the youngest Alus, including subfamilies derived from AluY. The blue and red dotted lines represent average methylation levels of promoters in the sperm and the brain, separately.

younger than those methylated Alus in terms of their insertion time. To test this prediction, we divide the Alu repeats into four different groups: AluJ (the oldest Alus, including AluJr, AluJr4, AluJo, and AluJb), AluS (the second oldest Alus, including AluSz and its derived subfamilies prior to AluY), AluY (the second youngest Alus), and AluY_HCG (the youngest Alus, including subfamilies derived from AluY). Consistent with our prediction, we find that methylation level decreases sequentially from AluJ to AluY_HCG in humans ($P < 0.001$ for all three pairs of comparison, Wilcox Rank-sum test, Figure 4.3A). In chimpanzees, it shows the same pattern except for the group AluY_HCG ($P < 0.05$ and $P < 0.001$ for the first two pairs of comparison, Wilcox Rank-sum test, Figure 4.3B). As a control, we perform the same analyses in the brain and find that in both species methylation levels of the four different groups are all high and show no significant difference except for between AluJ and AluS (AluJ > AluS, $P < 0.01$, Wilcox Rank-sum test, Figure 4.3A and B).

Interestingly, in the sperm although the methylation levels of AluJ and AluS are almost twice higher than the average methylation level of promoters, AluY and its subfamilies have similar methylation levels as the average. In the brain, however, all groups of Alus show twice the level of promoter methylation. Since all Alu repeats are GC rich and younger Alu insertions are more likely to be lineage-specific (see section below), this differential methylation patterns between tissues may partially explain why we observed a negative correlation between CpG content difference and methylation level difference in the sperm but a weak positive correlation in the brain.

Lineage-specific transposable element insertions change promoter methylation levels and gene expression levels

Numerous insertions and deletions (INDELs) occurred in the human and chimpanzee genomes since the two species diverged from their common ancestor. Large INDELs are primarily retrotransposon insertions and have played an important role in

regulatory evolution of humans and chimpanzees (Polavarapu et al. 2011). Since some transposable elements such as Alu repeats are high in GC content and enriched in CpG dinucleotides, we hypothesize that human- and chimpanzee-specific transposable element insertions in promoters can change regional CpG content as well as DNA methylation level, and consequently, it may cause gene expression level change. To test this hypothesis, we examine evolutionarily recent transposable elements - Alus and SVAs. Alu repeats are relatively young retrotransposons that propagated in primate genomes in the past 65 million years (Batzer and Deininger 2002). SVA repeats originated after the divergence of hominoid and Old World monkeys, and are thus even younger than Alu repeats (Batzer and Deininger 2002; Wang et al. 2005). Lineage-specific Alu or SVA insertion is defined as an Alu or SVA repeat that exists in one species (human or chimpanzee) but does not exist in the syntenic regions of the other two species in the three species genome alignment (see Materials and methods for details). We identified 50 human-specific AluY insertions (AluY and its subfamilies), 7 human-specific SVA insertions, 15 chimpanzee-specific AluY insertions (AluY and its subfamilies), and 9 chimpanzee-specific SVA insertions in promoter regions. The discrepancy of the number of lineage-specific Alu insertions in promoter regions between the two species is consistent with previous observation that the human genome harbors three-fold more lineage-specific Alu repeats than the chimpanzee genome (Chimpanzee Sequencing and Analysis Consortium 2005).

For both sperm and brain, we plot methyDiff against cpgoeDiff for Alu-inserted promoters and SVA-inserted promoters separately. Figure 4.4A and B show that almost all the blue circles (human) are on the right half of the panels (average human CpG O/E = 0.60, average chimpanzee CpG O/E = 0.52, $P = 0.018$, Wilcox Rank-sum test) and almost all the red circles (chimpanzee) are on the left half of the panels (average human CpG O/E = 0.44, average chimpanzee CpG O/E = 0.52, $P = 0.074$, Wilcox Rank-sum test). This indicates that Alu-insertions tend to increase promoter CpG O/Es. For

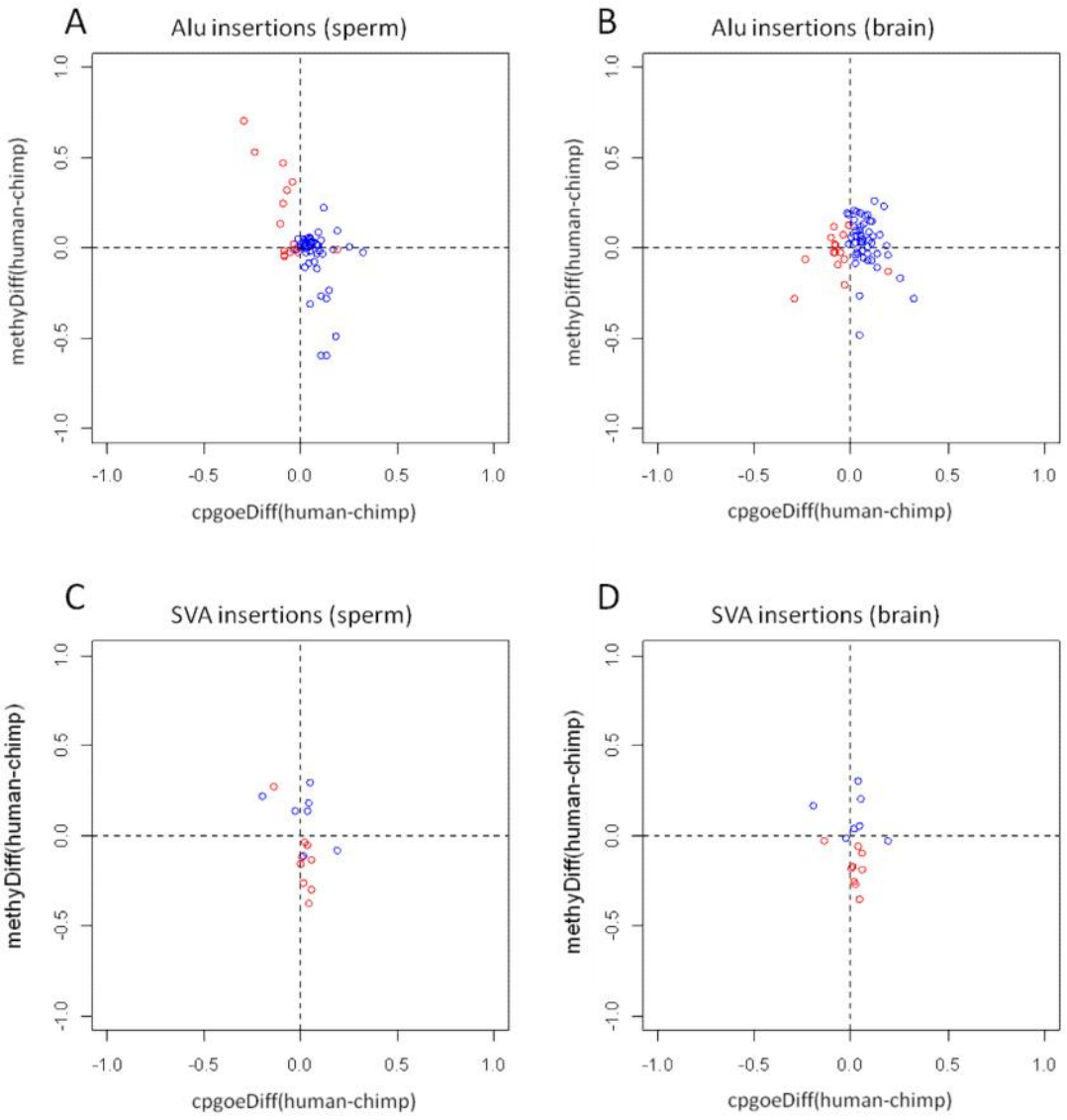

Figure 4.4. Scatter plots of human-chimpanzee CpG O/E difference and methylation difference for genes with lineage-specific Alu insertions in A) the sperm and B) the brain as wells as for genes with lineage-specific SVA insertions in C) the sperm and D) the brain. Blue circles are Alu- or SVA-inserted human promoters and red circles are Alu- or SVA-inserted chimpanzee promoters.

methylation level difference, we find that in the sperm, although the majority of the red and blue circles concentrate near the dividing line $y = 0$, the extreme positive methyDiff values tend to be enriched in red circles and the extreme negative methDiff values tend to be enriched in blue circles (Figure 4.4A). Combining the two data sets of Alu insertions in humans and chimpanzees, we see a strong negative correlation between sperm methyDiff and cpgoeDiff (Spearman's rho = -0.34, $P < 0.01$). Therefore, our data suggest that lineage-specific Alu insertions tend to decrease regional methylation level in the sperm. The brain, however, does not exhibit the same pattern (Figure 4.4B). Instead, most blue circles are in the upper panel and most red circles are in the lower panel, which means that lineage-specific Alu insertions tend to increase regional methylation level in the brain. These results are consistent with our observations from the last section.

Unlike lineage-specific Alu insertions which increase regional CpG O/E, we find that lineage-specific SVA insertions do not have strong effects on CpG O/E in either humans (average CpG O/E of SVA-inserted human promoters: 0.46 vs. average CpG O/E of chimpanzee promoters: 0.44, $P = 0.32$, Wilcox Rank-sum test) or chimpanzees (average CpG O/E of human promoters: 0.50 vs. average CpG O/E of SVA-inserted chimpanzee promoters: 0.49, $P = 0.51$, Wilcox Rank-sum test). Unlike lineage-specific Alu insertions which have different effects on regional methylation levels in different tissues, lineage-specific SVA insertions tend to increase regional methylation levels in both tissues. For example, 5 out of the 7 human-specific SVA insertions have increased sperm methylation levels in human promoters compared to chimpanzee promoters, and 8 out of the 9 chimpanzee-specific SVA insertions have increased sperm methylation levels in chimpanzee promoters compared to human promoters (Figure 4.4C). Similarly, 5 out of the 7 human-specific SVA insertions have increased brain methylation levels in human promoters compared to chimpanzee promoters, and all of the 9 chimpanzee-specific SVA insertions have increased brain methylation levels in chimpanzee promoters compared to human promoters (Figure 4.4D). Although no statistical significance is detected (except

for the effect of chimpanzee-specific SVA insertions in the brain) due to limited sample size, our results suggest that in general lineage-specific SVA insertions tend to increase regional methylation levels in both tissues of both species.

Next, we examine the influence of lineage-specific Alu and SVA insertions on gene expression levels. Since lineage-specific Alu insertions tend to lower promoter methylation levels in the sperm, we expect an increase of gene expression level. However, we observe an opposite trend: 23 out of 32 human genes (72%) with lineage-specific Alu insertions in promoters have lower expression level in the sperm than orthologous chimpanzee genes, although not significantly different from 50% (Fisher's exact test). Similarly, 7 out of 10 chimpanzee genes (70%) with lineage-specific Alu insertions in promoters have lower expression level in the sperm than orthologous human genes (also not significantly different from 50%, Fisher's exact test). For the brain, no apparent increase or decrease of expression level is detected for genes with lineage-specific Alu insertions in promoters. Since lineage-specific SVA insertions tend to increase promoter methylation levels in the sperm and brain, we expect that gene expression levels would decrease. However, we observe an opposite trend again: for 4 SVA-inserted human genes that have expression data available, 2 of them show increased expression level in the sperm and all of them show increased expression level in the brain. And for 7 SVA-inserted chimpanzee genes that have expression data available, 5 of them show increased expression levels in both the sperm and the brain. Since only limited SVA insertions are detected, no robust statistical conclusion can be drawn.

Effects of CpG depleting mutation and CpG generating mutation on methylation level

In the above two sections we have examined the effects of transposable element insertions on DNA methylation divergence between humans and chimpanzees. We hypothesize that other genetic elements such as single nucleotide substitutions can also

affect methylation divergence. To test this hypothesis, we examine correlation between sequence divergence (measured as Kimura distance) and methylation divergence (measured as absolute value of methylation level difference) of the two species after removing all the INDELs and focusing on the aligned sites only. We find that these two variables are positively correlated with each other in both tissues (Spearman's rho = 0.14 in the sperm, $P < 1x10^{-10}$ and Spearman's rho = 0.10 in the brain, $P < 1x10^{-10}$).

The fact that methylation divergence is positively correlated with sequence divergence implies that single nucleotide substitutions have strong effects on methylation levels. Since cytosines from CpG sites are the main targets of DNA methylation, we investigate the effects of CpG depleting mutations and effects of CpG generating mutations (see Materials and methods for details) on methylation levels. Since methylated CpG sites would spontaneously deaminate to TpG sites and thus be depleted, we expect a positive correlation between CpG depleting substitution rate and methylation level. Indeed, we find that partial correlations between these two variables after correcting for CpG O/E are positive in human sperm (Spearman's rho = 0.13, $P < 1x10^{-10}$), human brain (Spearman's rho = 0.11, $P < 1x10^{-10}$), chimpanzee sperm (Spearman's rho = 0.15, $P < 1x10^{-10}$), and chimpanzee brain (Spearman's rho = 0.14, $P < 1x10^{-10}$).

Generated CpG sites due to single nucleotide substitutions could be potential new targets of DNA methylation. We predict that methylation status of these new CpG sites would be different than that of other CpG sites. Partial correlation analyses reveal consistent positive correlations between methylation level and CpG generating substitution rate after correcting for CpG O/E in human sperm (Spearman's rho = 0.15, $P < 1x10^{-10}$), human brain (Spearman's rho = 0.15, $P < 1x10^{-10}$), chimpanzee sperm (Spearman's rho = 0.12, $P < 1x10^{-10}$), and chimpanzee brain (Spearman's rho = 0.15, $P < 1x10^{-10}$), suggesting that newly generated CpG sites might be more methylated than other CpG sites.

We then directly examine methylation levels of the newly generated CpG sites in human and chimpanzee promoters. We find that for both tissues, human-specific CpG sites have significantly higher methylation levels than all the human CpG sites from the aligned sites (median: 0.29 vs 0.12 in the sperm, $P < 1x10^{-10}$, Wilcox Rank-sum test; median: 0.62 vs 0.30 in the brain, $P < 1x10^{-10}$, Wilcox Rank-sum test, Figure 4.5). While chimpanzee-specific CpG sites also exhibit increased methylation levels compared to all the chimpanzee CpG sites from the aligned sites, the increase is not significant in the sperm (median: 0.23 vs 0.12 in the sperm, $P > 0.1$, Wilcox Rank-sum test; median: 0.65 vs 0.29 in the brain, $P < 1x10^{-10}$, Wilcox Rank-sum test, Figure 4.5). Interestingly, the uneven increases of methylation level of the generated CpG sites compared to the aligned CpG sites between the two species is consistent with the overall methylation level difference between the two species. For example, in the sperm, the methylation level of human-specific CpG sites is increased by 142% compared to all the human CpGs in the aligned sites, however, the increase percentage is only 92% for chimpanzee-specific CpG sites, consistent with the fact that humans have generally higher methylation level than chimpanzees in the sperm (Table 4.1). Similarly, in the brain, the methylation level of human-specific CpG sites is increased by 107% compared to all the human CpGs in the aligned sites, yet this percentage is 124% for chimpanzee-specific CpG sites, consistent with the observation that humans have generally lower methylation levels than chimpanzees in the brain (Table 4.1).

To further validate that newly generated CpG sites are more methylated than other CpG sites, we did simulation test by randomly picking the number of generated CpG sites from all the aligned CpG sites and calculating their average methylation level for 10,000 times. Our results show that for both tissues of both species, the observed average methylation level of the generated CpG sites are far higher than those from the simulation (Figure C.1), indicating that newly generated CpG sites after the divergence of humans and chimpanzees are indeed more methylated than other CpG sites. The Z-scores, which

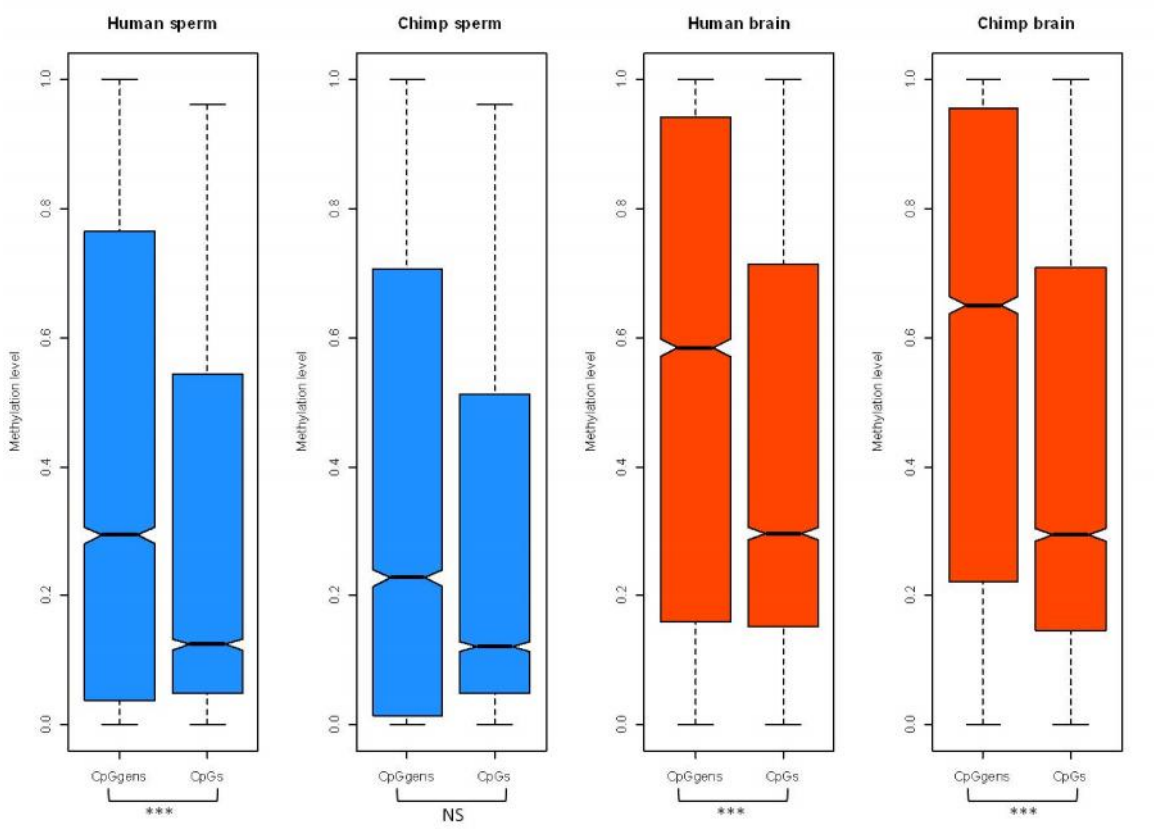

Figure 4.5. Methylation levels of newly generated CpG sites and all the aligned CpG sites in human sperm, chimpanzee sperm, human brain and chimpanzee brain. Newly generated CpG sites are noted as CpGgens and all the aligned CpG sites are noted as CpGs. ***, $P < 1x10^{-10}$, Wilcox Rank-sum test; NS, not significant, Wilcox Rank-sum test.

measure how large the observed values deviate from expectations, show consistent patterns with those from Figure 4.5. Specifically, the Z-score is lower in humans than in chimpanzees in the sperm and higher in humans than in chimpanzees in the brain (Figure C.1).

To test whether the increase of methylation level of generated CpG sites contribute to the overall methylation difference between humans and chimpanzees, we compare relative methylation level of humans to chimpanzees before and after removing lineage-specific CpG sites. Before removing lineage-specific CpG sites, i.e., considering all the CpG sites in the aligned region without INDELs, the relative methylation level of human to chimpanzee is 1.192 in the sperm and 0.903 in the brain (Table 4.1). After removing lineage-specific CpG sites, the relative methylation level of human to chimpanzee decreases to 1.189 in the sperm and increases to 0.904 in the brain (Table 4.1), but the changes are not significant ($P > 0.5$, Chi-squared test). Therefore, although the direction of the magnitude difference of methylation level elevation of generated CpG sites is consistent with the direction of overall methylation difference, this analysis indicates that generated CpG sites do not contribute significantly to the overall methylation difference between humans and chimpanzees in either tissue, which is probably because generated CpG sites are rare compared to the total number of CpG sites.

CpG generating substitutions are the most frequent among all 16 dinucleotide generating substitutions in promoter regions

As shown in the last section of the results, human-specific CpG sites are more heavily methylated than the conserved CpG sites. Therefore, we hypothesize that CpG generating mutations might be functional and under positive selection. To test this hypothesis, we compare the frequency of CpG generating substitution to the frequencies of other dinucleotide generating substitutions.

Human CpG generating mutation is defined as a human-specific substitution to cytosine or guanine followed by a conserved guanine site or preceded by a conserved cytosine site. By the same definition, we can identify the other 15 types of dinucleotide generating mutations. For all the 15,232 1:1:1 orthologs of human, chimpanzee, and orangutan, we calculate the number of all the 16 dinucleotide generating substitutions in human promoters based on the extracted sequence alignment of their promoter regions. To correct for human-specific substitution rate, we divide the number of dinucleotide generating substitution by the number of human-specific substitution for each ortholog. Because each human-specific single nucleotide substitution gives rise to two types of dinucleotide generating substitutions, for each type of dinucleotide generating substitution the average proportion of dinucleotide generating substitutions number out of human-specific substitutions number should be 1/8 if they are all neutral or equally selected. However, we find that the 16 proportions vary dramatically with 7 proportions above 1/8 and 9 proportions below 1/8 (Figure 4.6A). Moreover, we find that the proportion of CpG generating substitution is the highest among all 16 proportions of dinucleotide generating substitutions (Figure 4.6A). Since any two different dinucleotide generating substitutions may partially overlap with each other as mentioned above, and it is impossible to make all 16 types of dinucleotide generating substitutions not overlap with each other simultaneously, we make CpG and GpC generating substitutions not overlap with each other to correct for GC content (see methods). Still, we find that non-overlapping CpG generating substitutions are significantly more abundant than non-overlapping GpC generating substitutions (Figure 4.6A). The higher frequency of CpG generating substitutions than that of GpC generating substitutions means that biased gene conversion cannot fully explain the abundance of CpG generating substitutions because biased gene conversion is a neutral process which increases regional GC content and results in weak (A,T) to strong (G,C) substitutions (Chen et al. 2007). These results

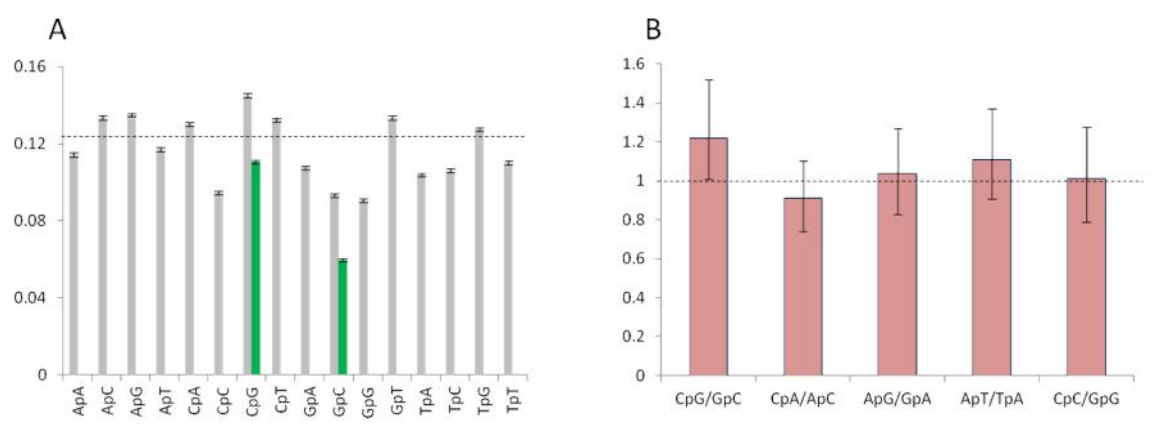

Figure 4.6. CpG generating substitution rate compared to other dinucleotide generating substitution rate. A) Proportions of 16 types of dinucleotide generating substitutions (grey bars) and non-overlapping CpG and GpC generating substitutions (green bars). The dotted line represents expected proportion 0.125; B) ω_X / ω_A ratio (see Results and Discussion for details) for 5 types of dinucleotide generating mutations. The dotted line represents expected ω_X / ω_A ratio 1.

indicate that CpG generating mutations are either more frequent for some other reasons or evolutionarily more favored than the other types of dinucleotide generating mutations.

In order to test whether CpG generating mutations in promoter region are under selection, we followed the following procedures. First, for each ortholog we calculate the ratio (α) between non-overlapping CpG and GpC generating substitution rates in human promoters (α_p) to correct for GC content and thus biased gene conversion. Then, we calculate α in the accompanying intron sequences (α_i) to correct for mutation rate. A α_p/α_i ratio (ω), which is analogous to nonsynonymous substitution rate to synonymous substitution rate ratio in coding sequences, can then be computed. Finally, we compare the ω values on the X chromosome (ω_X) and those on the autosomes (ω_A). If CpG generating mutations are under selection, we would expect (ω_X) to be greater than (ω_A) due to the fast-X effect, i.e., selection is more efficient on the X chromosome than on autosomes because of the fact that beneficial mutations on the X chromosome are exposed to natural selection on the heterogametic sex while beneficial mutations on autosomes are usually recessive or partially recessive and thus not exposed to natural selection (Charlesworth et al. 1987). After obtaining all the values mentioned above from all the orthologs, we find that the average ω_X/ω_A ratio is 1.22, significantly greater than 1 (90% confidence interval: (1.01, 1.52), bootstrapping resampling). This indicates that there indeed exists a fast-X evolution of CpG generating mutations after correction for GC content and mutation rate. As a control, we conduct the same analysis for another 4 pairs of non-overlapping dinucleotide generating mutations: CpA vs. ApC, ApG vs. GpA, ApT vs. TpA, and CpC vs. GpG. The average ω_X/ω_A ratios for these 4 pairs are 0.91, 1.04, 1.11, and 1.01, separately (Figure 4.6B). Their 90% confidence intervals (CI), however, all include 1: CI of CpA/ApC: (0.74, 1.10); CI of ApG/GpA: (0.83, 1.27); CI of ApT/TpA: (0.91, 1.37); CI of CpC/GpG: (0.79, 1.28). Thus, unlike CpG generating mutations, other dinucleotide generating mutations do not show strong signals of fast-X

evolution. Our analysis thus indicates that some of the CpG generating sites might have been driven by natural selection.

We then identify genes with the highest ω values in terms of CpG generating mutations. To do this, we only keep genes whose CpG and GpC generating mutation numbers on both promoters and introns are above 0 so that ω values can be valid. 3396 genes are left after this procedure, including 50 X-linked genes and 3346 autosomal genes. Log-transformed ω values can be approximately fitted by a normal distribution with mean value slightly below 0. We perform GO analysis for genes on the right tail of the normal distribution (quantile > 97.5%); however, no significant enrichment is found for any GO terms. A previous study shows that genes overlapping with regions of extremely high density of human-specific CpG sites are significantly enriched in neurological disease genes (Bell et al. 2012). Among these 6 neurological disease genes (ANKRD11, CHL1, EHMT1, VLDLR, DLGAP2, and DPP10), CHL1, VLDLR, and DPP10 are among our top 50 genes with the highest number of CpG generating mutations in promoters, but none of them has a ω value among top 100.

Conclusion

In this chapter, we investigated the relationships between genomic sequence divergence and DNA methylation divergence and explored genomic factors that influence the evolution of methylation in humans and chimpanzees. We find that correlation between methylation level difference and CpG content difference of the two species is negative in the sperm but positive in the brain, which is partially attributed to the differential methylation patterns of Alu repeats between the two tissues. We demonstrate that lineage-specific CpG sites are more methylated compared to all the aligned CpG sites in both tissues of both species, but the relative elevation of methylation level of the lineage-specific CpG sites has varying magnitude between species and between tissues. Alu insertions and CpG generating mutations are by no means the only genomic factors

underlying the evolution of methylation in the two species. Other factors such as flanking regions of Alu insertions, flanking regions of lineage-specific CpG sites, or specific sequence motifs may also play a significant role. Finally, we find evidence that CpG generating mutations might have been under positive selection to provide new targets for DNA methylation.

CHAPTER 5

CONCLUSIONS

This dissertation encompasses three studies of primate comparative genomics with the goal of understanding molecular basis of primate and human evolution. Chapters two and three focused on elucidating evolutionary forces underlying variation of evolutionary rates between chromosomes and between species. The contribution of genomic factors to epigenomic evolution was investigated in chapter four.

Chapter two simultaneously studied slow-X evolution, fast-X evolution and their interactions within a well established phylogeny of primates, which included two apes (human and orangutan), an Old World monkey (rhesus macaque), and a New World monkey (marmoset). Not all sequenced primate species were included in our research because ancestral polymorphism, which can bias estimation of mutation rate, has strong effects on closely related species (Makova and Li 2002). The results from this chapter revealed a consistent pattern of slow-X evolution across all the examined lineages, supporting the male-driven evolution theory. However, the degree of slow-X evolution varied significantly among the lineages. Specifically, human exhibits the greatest male-to-female mutation rate ratio, followed by orangutan, rhesus macaque, and marmoset. These ranks correlate strongly with ranks of their life history traits such as life span and age at sexual maturity, supporting the generation time effect on male mutation bias (Chang et al. 1994; Li et al. 2002; Bartosch-Härlid et al. 2003). Unlike slow-X evolution being universally observed across lineages, fast-X evolution was found to be restricted to specific lineage. At nonsynonymous sites, the evidence supporting fast-X evolution is limited to marmoset only, which happens to have the lowest male-to-female mutation rate ratio. Other indirect evidences were found to support the theory that strong male mutation bias could counteract the effect of fast-X evolution (Kirkpatrick and Hall 2004).

Finally, we discussed how lineage-specific variation of slow-X evolution might affect population genetic inferences about human demography.

The study comprising chapter three designed a novel framework based on a previously developed "maximal segment" algorithm (Ruzzo and Tompa 1999) to identify genomic segments with accelerated rates of lineage-specific substitutions without bias. Comparing to previous studies, our unbiased approach does not require any *a priori* knowledge or employing sliding windows of arbitrary sizes. We implemented this framework to identify clusters of lineage-specific substitutions (named "maximal segments" after the algorithm) by scanning score-transformed human- and chimpanzee-specific substitutions identified from human-chimpanzee-macaque whole genome alignments. The identified human-specific maximal segments have significantly higher lineage-specific substitution rate than human genomic background and chimpanzee orthologous regions, and contain significant number of human accelerated regions identified by previous studies (Pollard et al. 2006a; Berglund et al. 2009). There is virtually no overlap of genomic locations between human-specific maximal segments and chimpanzee-specific maximal segments, indicating that our framework can effectively detect lineage-specific clusters of single nucleotide substitutions. We found evidences showing that human-specific maximal segments had been driven by multiple evolutionary forces, including regionally increased mutation rate, recombination associated processes such as biased gene conversion, and positive selection. Finally, we showed that human genes overlapping with maximal segments were enriched in developmental processes.

Chapter four investigated the role of genomic changes in shaping DNA methylation divergence and gene expression divergence between humans and chimpanzees. We found that DNA methylation level difference is negatively correlated with CpG dinucleotide content difference between the two species in the sperm but not in the brain, possibly due to the reason that genetic mutations altering methylation levels

can reflect better in germline cells. We did not detect any significant correlations between methylation divergence and gene expression divergence in either sperm or brain, which might be because the data sets we used were from different experiments using different samples. We demonstrated that the evolution of DNA methylation of humans and chimpanzees can be affect by various genomic factors including transposable element insertions, CpG depleting mutations, and CpG generating mutations. Finally, we discussed that CpG generating mutations might have been evolutionarily more favored than the other types of dinucleotide generating mutations for being potential targets of DNA methylation.

These studies have inspired me and hopefully will ventilate active discussions in scientific community. With more and more whole genomes of primates, archaic human and modern human individuals being sequenced, we are in an era of being able to trace molecular evolution of primates happening in real time and ultimately understand how human beings have evolved to stand apart from other primates.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Table A.1. Comparison of alpha values obtained from the HKY method and the Kimura's 2-parameter method.

HKY

	Human	Orang	Rhesus	Marmoset	Outgroup
intron_X	0.008	0.010	0.022	0.054	0.395
intron_A	0.010	0.012	0.026	0.058	0.435
X/A	0.791	0.818	0.830	0.920	0.908
α	4.358	3.403	3.074	1.632	1.766

Kimura

	Human	Orang	Rhesus	Marmoset	Outgroup
intron_X	0.008	0.010	0.022	0.054	0.395
intron_A	0.010	0.012	0.026	0.058	0.435
X/A	0.795	0.820	0.835	0.922	0.909
α	4.209	3.339	2.963	1.606	1.752

Table A.2. Mean and median of intron substitution rates of X-linked and autosomal genes.

	Human	Orangutan	Rhesus macaque	Marmoset	Outgroup
X (mean, median)	0.008, 0.008	0.010, 0.010	0.022, 0.021	0.054, 0.051	0.395, 0.384
A (mean, median)	0.010, 0.010	0.012, 0.011	0.026, 0.026	0.058, 0.057	0.435, 0.426
X<A Significance	***	***	***	***	***

*** Wilcoxon test, $P < 10^{-10}$

Table A.3. Variation of Life History traits among the primate species examined.

	Male mutation bias, α[1]	Body mass (kg)[2]	Age at first reproduction (year)[3]	Age at sexual maturity (year)[4]	Mating system[5]	Relative testis size[7]	Maximum life expectancy (year)[4]
Human	4.36	M: 72.1 F: 62.1	19.5	F: 16.5	variable[6]	0.62	MF: 80-90
Orangutan	3.40	M: 78.5 F: 35.8	15.4	M: 9.5 F: 7	SM	0.50	MF: > 50
Rhesus	3.07	M:11.0 F: 8.8	3.0	MF: 3.5 - 4	MMMF	5.04	MF: 29
Marmoset	1.63	M: 0.317 F: 0.324	1.6	M: 1.4 F: 1	PA	4.06	MF: 11.7
Correlation with α*		0.97 (p = 0.03)	0.88 (p = 0.12)	0.99 (p = 0.01)		-0.68 (p = 0.32)	0.97 (p = 0.03)
Correlations between PICs**		0.90	0.79	0.96		-0.52	0.94

[1] Male mutation bias determined from the current study. See the main text.
[2] Body mass data from (Smith and Jungers 1997). When data from several populations were available, those from the largest population were chosen.
[3] Human data are from (Robson and Wood 2008); orangutan data is from (Wich et al. 2004); rhesus and marmoset data are from (Wootton 1987).
[4] Data from (Rowe et al. 1996). For correlation analysis, we chose the value of 59 for orangutan, which is the longest recorded life span for a Borneo Orangutan, *Pongo pygmaeus*).
[5] Mating systems are: MMMF, multi-male, multi-female; SM, single male; Mon, monogamous; PA, polyandrous. Data are obtained from (Sussman and Garber 1987; Harcourt et al. 1995).
[6] Mating system in humans varies among different societies: some human societies are monogamous while others are single male system. Other extreme cases also exist.
[7] Relative testis mass is presented as testis mass (g) divided by body mass (kg). Data from (Harcourt et al. 1995).
*Pearson correlation coefficients between each life history trait and the male mutation bias (α). Averages are used when male and female have different values. Both variables are log-transformed.
**Correlations between the phylogenetically independent contrasts (PIC, Felsenstein 1985) of each life history trait and the PIC of the male mutation bias (α). The correlations are generated by the 'Contrast' program from Phylip package (Felsenstein 1989).

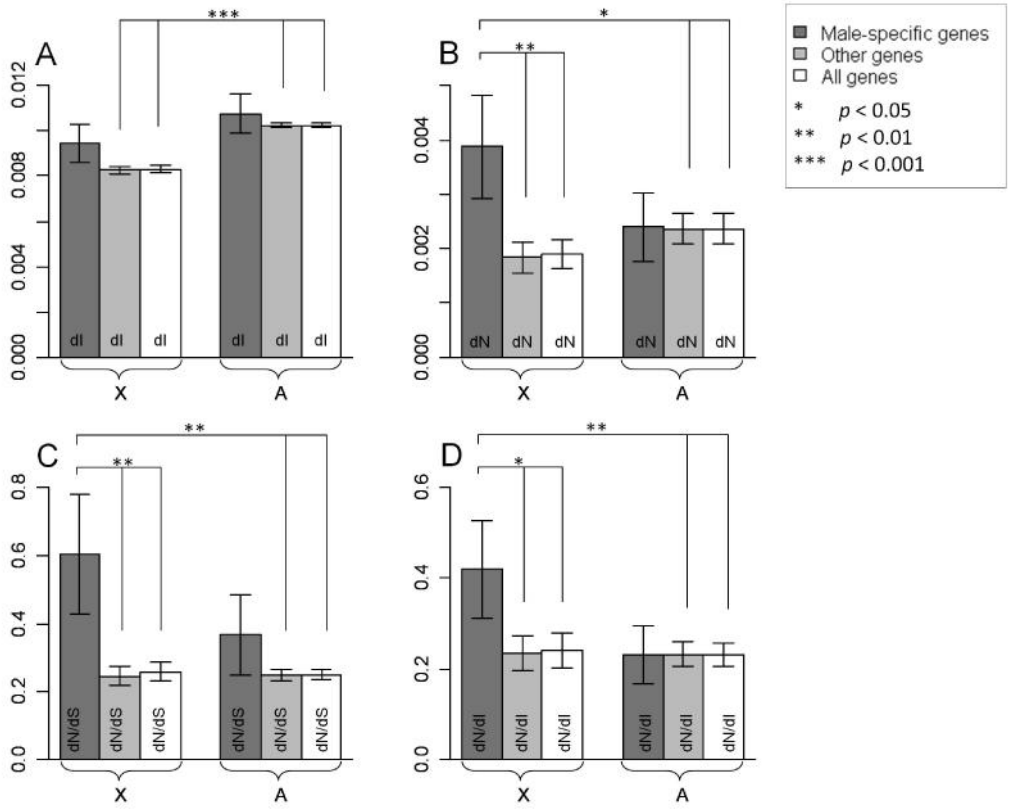

Figure A.1. Comparisons between X-linked and autosomal human genes according to their expression patterns. Significant differences, assessed using the Wilcoxon test, are marked by astericks. (A) Human X-linked introns show lower divergence than autosomal genes. (B) Among the X-linked genes, those with male-biased expression have significantly higher dN compared to other X-linked genes. (C) dN/dS and (D) dN/dI exhibit similar patterns to dN.

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

Table B.1. Statistics of the human maximal segments with FDR Q < 0.1.

	Human maximal segments (Scoring scheme: 60/40)					Human maximal segments (Scoring scheme: 55/45)				
Chr	Total number	Human-specific substitution	Proportion[a]	Total length	Proportion[b]	Total number	Human-specific substitution	Proportion[a]	Total length	Proportion[b]
chr1	687	30995	0.051	4128518	0.037	252	21249	0.035	3114417	0.028
chr2	103	4423	0.048	463079	0.033	44	2891	0.032	325482	0.023
chr3	261	10955	0.041	1363651	0.029	81	7121	0.027	1043832	0.022
chr4	717	32807	0.047	3999688	0.034	245	18680	0.027	2514155	0.021
chr5	525	25579	0.050	3249195	0.038	169	15127	0.03	2173921	0.026
chr6	608	28433	0.053	3431144	0.037	215	18261	0.034	2531389	0.027
chr7	319	13439	0.041	1590411	0.029	109	7025	0.021	924981	0.017
chr8	729	36099	0.051	4000127	0.035	203	19495	0.027	2456954	0.021
chr9	501	21713	0.056	2489218	0.039	201	16658	0.043	2146322	0.034
chr10	462	22575	0.051	2637103	0.036	180	16862	0.038	2196736	0.030
chr11	343	16903	0.053	2160771	0.039	120	12001	0.037	1782633	0.032
chr12	444	20104	0.050	2497011	0.035	179	12559	0.031	1785245	0.025
chr13	415	18896	0.046	2365762	0.034	148	11436	0.028	1578249	0.023
chr14	417	19258	0.047	2351503	0.033	134	12939	0.032	1794795	0.025
chr15	378	17852	0.049	2154092	0.035	122	12620	0.035	1716675	0.028
chr16	346	19569	0.054	2061838	0.038	113	22685	0.062	2479487	0.046
chr17	207	8852	0.048	1068843	0.035	81	6952	0.037	967173	0.031
chr18	311	14282	0.049	1715126	0.034	117	11211	0.038	1562182	0.031
chr19	186	6714	0.032	724371	0.022	64	3753	0.018	454677	0.014
chr20	189	8774	0.054	1067727	0.039	81	6471	0.04	868001	0.032
chr21	174	7998	0.046	857719	0.032	55	4714	0.027	579839	0.022
chr22	139	6832	0.048	830516	0.036	54	4976	0.035	670904	0.029
chrX	773	23043	0.055	3413980	0.035	371	16546	0.04	2675351	0.028
Sum	9234	416095	0.049	50621393	0.035	3338	282232	0.034	38343400	0.027

[a] Proportion of human-specific substitutions in the maximal segments out of all the human-specific substitutions in the chromosome.
[b] Proportion of the maximal segments lengths out of the whole chromosome length.

Table B.2. Statistics of the chimpanzee maximal segments with FDR Q < 0.1.

	Chimpanzee maximal segments (Scoring scheme: 60/40)					Chimpanzee maximal segments (Scoring scheme: 55/45)				
Chr	Total number	Chimpanzee-specific substitution	Proportion[a]	Total length	Proportion[b]	Total number	Chimpanzee-specific substitution	Proportion[a]	Total length	Proportion[b]
chr1	132	10885	0.017	1386636	0.012	10	1061	0.0016	116710	0.001
chr2a & chr2b	19	2297	0.023	250813	0.018	1	21	0.0002	474	3.41×10^{-5}
chr3	48	4040	0.014	460073	0.010	0	0	0	0	0
chr4	160	13569	0.018	1648637	0.014	4	225	0.0003	21803	0.0002
chr5	112	9439	0.018	1179365	0.014	6	360	0.0007	33314	0.0004
chr6	100	8386	0.015	1091994	0.012	6	2033	0.0036	309802	0.0033
chr7	74	4377	0.013	495616	0.009	4	303	0.0009	26718	0.0005
chr8	128	10245	0.014	1268966	0.011	7	1340	0.0018	164268	0.0014
chr9	74	7960	0.019	1010856	0.016	3	189	0.0005	26862	0.0004
chr10	89	7323	0.016	863745	0.012	4	188	0.0004	11666	0.0002
chr11	93	6588	0.019	788281	0.014	4	257	0.0008	23164	0.0004
chr12	60	4903	0.011	622362	0.009	6	751	0.0017	94424	0.0013
chr13	68	7777	0.018	1021680	0.015	5	1240	0.0028	190038	0.0027
chr14	83	5326	0.012	582545	0.008	5	163	0.0004	10926	0.0002
chr15	92	7192	0.019	846940	0.014	7	872	0.0023	108742	0.0018
chr16	66	4518	0.012	446352	0.008	4	697	0.0018	114341	0.0021
chr17	20	1430	0.007	173662	0.006	0	0	0	0	0
chr18	61	5249	0.017	669929	0.013	4	554	0.0018	67360	0.0014
chr19	14	660	0.003	46248	0.001	0	0	0	0	0
chr20	33	2638	0.015	320016	0.012	3	505	0.0029	73622	0.0027
chr21	33	4372	0.024	544645	0.020	1	20	0.0001	733	2.72×10^{-5}
chr22	18	971	0.006	88586	0.004	0	0	0	0	0
chrX	0	0	0	0	0	0	0	0	0	0
Sum	1577	130145	0.014	15807947	0.011	84	10779	0.0012	1394967	0.0009

[a] Proportion of chimpanzee-specific substitutions in the maximal segments out of all the chimpanzee-specific substitutions in the chromosome

[b] Proportion of the maximal segments lengths out of the whole chromosome length

Table B.3. Characteristics of the top 50 maximal segments with the lowest P-values and their overlapping genes.

Chr	Chr start	Chr stop	*P*-value	Length	Location	Gene symbol
chr8	4121165	4201748	1.95×10^{-14}	80584	intragenic MS	CSMD1
chr2	115634177	115641744	3.21×10^{-10}	7568	intragenic MS	DPP10
chr3	151037988	151041317	3.41×10^{-9}	3330	intragenic MS	RNF13
chr16	6922427	6923963	4.58×10^{-9}	1537	intragenic MS	RBFOX1
chr8	61275843	61299074	1.16×10^{-8}	23232	intragenic MS	CA8
chr16	7461827	7501460	2.21×10^{-8}	39634	intragenic MS	RBFOX1
chr1	3253455	3255293	2.83×10^{-8}	1839	intragenic MS	PRDM16
chr14	32087675	32114066	2.99×10^{-8}	26392	intragenic MS	AKAP6
chr6	57537396	57559032	3.41×10^{-8}	21637	intragenic MS	PRIM2
chr2	116195443	116201369	4.27×10^{-8}	5927	intragenic MS	DPP10
chr10	24829491	24854794	4.92×10^{-8}	25304	intragenic MS	KIAA1217
chr21	16583875	16614398	6.17×10^{-8}	30524	intragenic MS	LINC00478
chr1	64212709	64214871	8.41×10^{-8}	2163	intragenic MS	ROR1
chr17	2149729	2152477	8.55×10^{-8}	2749	intragenic MS	SMG6
chr10	129076214	129087446	9.43×10^{-8}	11233	intragenic MS	DOCK1
chr3	155439236	155442573	9.83×10^{-8}	3338	intragenic MS	ARHGEF26
chr21	42036467	42050128	1.06×10^{-7}	13662	intragenic MS	RIPK4
chr22	17777862	17781306	1.42×10^{-7}	3445	intragenic MS	HIRA
chr9	135648536	135651767	1.45×10^{-7}	3232	intragenic MS	VAV2
chr9	131654022	131669699	1.57×10^{-7}	15678	intragenic MS	USP20
chr6	25234488	25240090	1.61×10^{-7}	5603	intragenic MS	CMAHP
chr17	55340790	55341099	2.83×10^{-7}	310	intragenic MS	RPS6KB1
chr19	59834385	59836493	3.02×10^{-7}	2109	intragenic MS	**LILRB1**
chr18	69956943	69959120	3.47×10^{-7}	2178	intragenic MS	FBXO15
chrX	131919579	131920274	1.21×10^{-12}	696	intragenic MS	HS6ST2
chrX	76851711	76858214	3.35×10^{-8}	6504	intragenic MS	ATRX
chrX	11079257	11115925	1.77×10^{-7}	36669	intragenic MS	ARHGAP6
chr8	1632525	1644120	8.58×10^{-14}	11596	MS covering part of a gene	DLGAP2
chr15	99235780	99238597	9.07×10^{-12}	2818	MS covering part of a gene	ALDH1A3
chr8	22516149	22560460	1.13×10^{-9}	44312	MS covering part of a gene & MS covering a whole gene	BIN3, KIAA1967, FLJ14107
chr9	124909141	124915876	9.22×10^{-8}	6736	MS covering part of a gene & MS covering a whole gene	MIR600HG, MIR600
chr22	25165842	25195433	2.37×10^{-7}	29592	MS covering part of a gene	ASPHD2
chr15	60913543	60924857	2.40×10^{-7}	11315	MS covering part of a gene	TLN2
chr9	2611094	2627938	3.18×10^{-7}	16845	MS covering part of a gene	VLDLR

Table B.3 (continued)

chr22	34001672	34042313	3.85×10^{-7}	40642	MS covering part of a gene	HMGXB4
chrX	101684602	101695362	1.71×10^{-9}	10761	MS covering part of a gene	NXF4
chr6	32734579	32743824	2.20×10^{-10}	9246	MS covering a whole gene	**HLA-DQB1**
chr21	32517993	32524719	2.48×10^{-23}	6727	Intergenic MS	NA
chr5	160431655	160435334	3.88×10^{-11}	3680	intergenic MS	NA
chr8	5579749	5596163	7.55×10^{-9}	16415	intergenic MS	NA
chr9	25800500	25837809	4.71×10^{-8}	37310	intergenic MS	NA
chr11	129127082	129138061	5.80×10^{-8}	10980	intergenic MS	NA
chr19	36366360	36369954	1.00×10^{-7}	3595	intergenic MS	NA
chr10	20911380	20932842	1.36×10^{-7}	21463	intergenic MS	NA
chr10	103878676	103878957	1.41×10^{-7}	282	intergenic MS	NA
chr20	58825736	58828614	1.90×10^{-7}	2879	intergenic MS	NA
chr4	19675066	19680179	2.06×10^{-7}	5114	intergenic MS	NA
chr20	58671178	58675022	2.43×10^{-7}	3845	intergenic MS	NA
chr9	1497688	1502681	2.89×10^{-7}	4994	intergenic MS	NA
chrX	145926124	145953880	3.16×10^{-7}	27757	intergenic MS	NA

Table B.4. Summary of the maximal segments (MS) inside the genomic regions under strong recent selection.

	Genomic regions under selection[1]	MS ($P < 0.05$)	MS (FDR Q < 0.1)	Top 50 MS
	Total length of the regions / total length of the human genome	Length of MS inside the regions / total MS length	Length of MS inside the regions / total MS length	Length of MS inside the regions / top 50 MS length
CHBJPT	0.82%	0.81%	0.90%	0
CEU	0.84%	0.83%	0.69%	0
YRI	0.90%	0.95%	0.85%	2%

[1] The data are from Voight et al. (2006).

Table B.5. All significant GO terms within each ontology for genes with at least one exon inside a maximal segment and at least one intron covering another maximal segment.

	Observed number of genes	Expected number of genes	*P*-value[1]
Biological Process			
Ectoderm development	72	25.98	3.06×10^{-13}
Nervous system development	66	22.9	1.00×10^{-12}
Cell-cell adhesion	48	12.9	1.10×10^{-12}
System development	84	34.84	1.98×10^{-12}
Cell communication	128	68.91	7.17×10^{-12}
Cell adhesion	62	21.59	9.32×10^{-12}
System process	84	37.73	1.56×10^{-10}
Signal transduction	120	65.68	2.66×10^{-10}
Neurological system process	75	33.23	2.07×10^{-9}
Developmental process	95	49.02	8.43×10^{-9}
Cellular process	153	97.92	8.68×10^{-9}
Mesoderm development	61	24.8	9.84×10^{-9}
Cell motion	47	16.59	2.63×10^{-8}
Transport	88	45.66	7.98×10^{-8}
Protein modification process	49	20.7	3.54×10^{-6}
Cellular component morphogenesis	43	17.11	5.59×10^{-6}
Anatomical structure morphogenesis	43	17.11	5.59×10^{-6}
Cellular component organization	51	22.89	1.18×10^{-5}
Protein transport	56	26.38	1.20×10^{-5}
Intracellular protein transport	56	26.38	1.20×10^{-5}
Cell cycle	58	28.57	3.01×10^{-5}
Vesicle-mediated transport	44	18.97	3.62×10^{-5}
Protein amino acid phosphorylation	31	10.91	4.30×10^{-5}
Intracellular signaling cascade	51	24.68	1.15×10^{-4}
Visual perception	24	8.06	4.77×10^{-4}
Sensory perception	32	13.22	8.06×10^{-4}
Cell-cell signaling	45	22.09	8.33×10^{-4}
Transmembrane receptor protein tyrosine kinase signaling pathway	19	5.77	1.33×10^{-3}
Muscle organ development	23	8.09	1.62×10^{-3}
Endocytosis	23	8.35	2.65×10^{-3}
Female gamete generation	16	4.61	3.71×10^{-3}
Angiogenesis	18	5.75	4.52×10^{-3}
Muscle contraction	24	9.38	5.54×10^{-3}
Synaptic transmission	27	11.57	8.52×10^{-3}

Table B.5 (continued)

Cytokinesis	13	3.49	1.09×10^{-2}
Sensory perception of sound	9	1.88	2.42×10^{-2}
Homeostatic process	11	2.83	2.77×10^{-2}
JAK-STAT cascade	12	3.41	3.52×10^{-2}
Protein metabolic process	75	50.41	3.53×10^{-2}
Protein targeting	11	2.95	3.97×10^{-2}
Asymmetric protein localization	7	1.21	4.28×10^{-2}
Protein localization	7	1.21	4.28×10^{-2}
Molecular Function			
Receptor activity	58	29.33	6.25×10^{-5}
Protein kinase activity	25	8	1.09×10^{-4}
Kinase activity	28	10.51	5.00×10^{-4}
Transporter activity	34	16.05	5.43×10^{-3}
Glucosidase activity	4	0.28	2.82×10^{-2}
Transmembrane transporter activity	HYPERL	14.86	3.61×10^{-2}
Calcium ion binding	21	8.97	4.92×10^{-2}
Cellular Component			
Extracellular matrix	21	8.9	1.09×10^{-2}
Extracellular region	21	8.92	1.11×10^{-2}
Cell junction	10	2.59	1.19×10^{-2}
Plasma membrane	10	2.77	2.02×10^{-2}
Actin cytoskeleton	18	7.44	2.16×10^{-2}

[1] *P*-values are adjusted for multiple comparisons using the Bonferroni correction

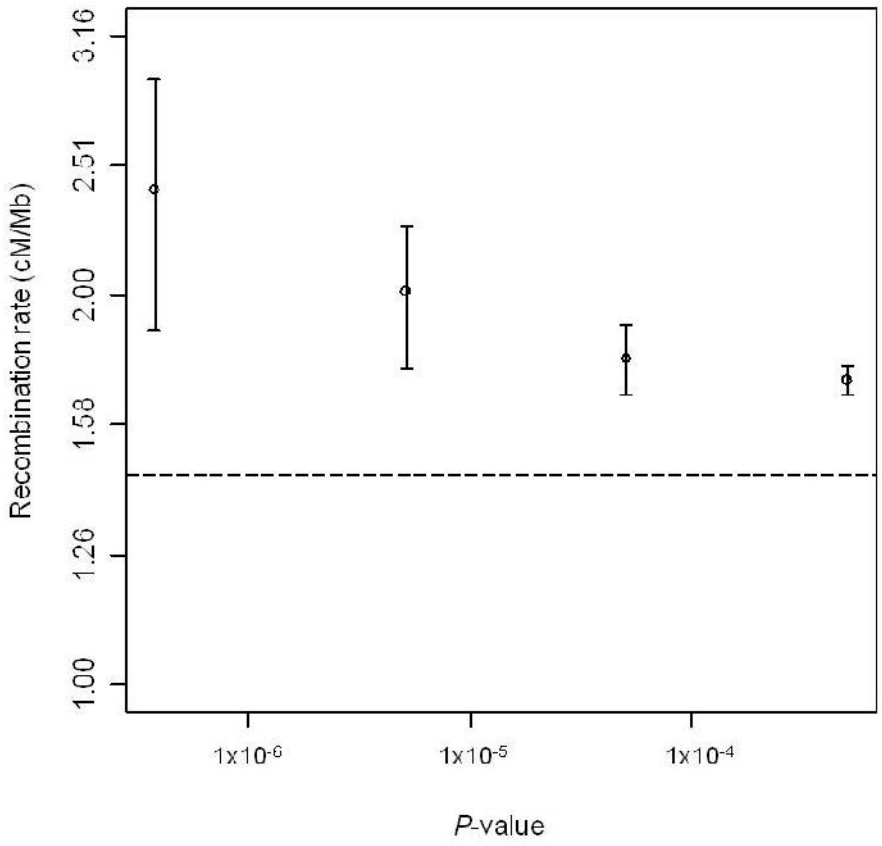

Figure B.1. Recombination rates in 4 groups ('extreme', 'strong', 'medium', and 'weak') of maximal segments using a fine scale map of recombination rates from Myers et al. (2005). The dashed line represents the genomic average recombination rate.

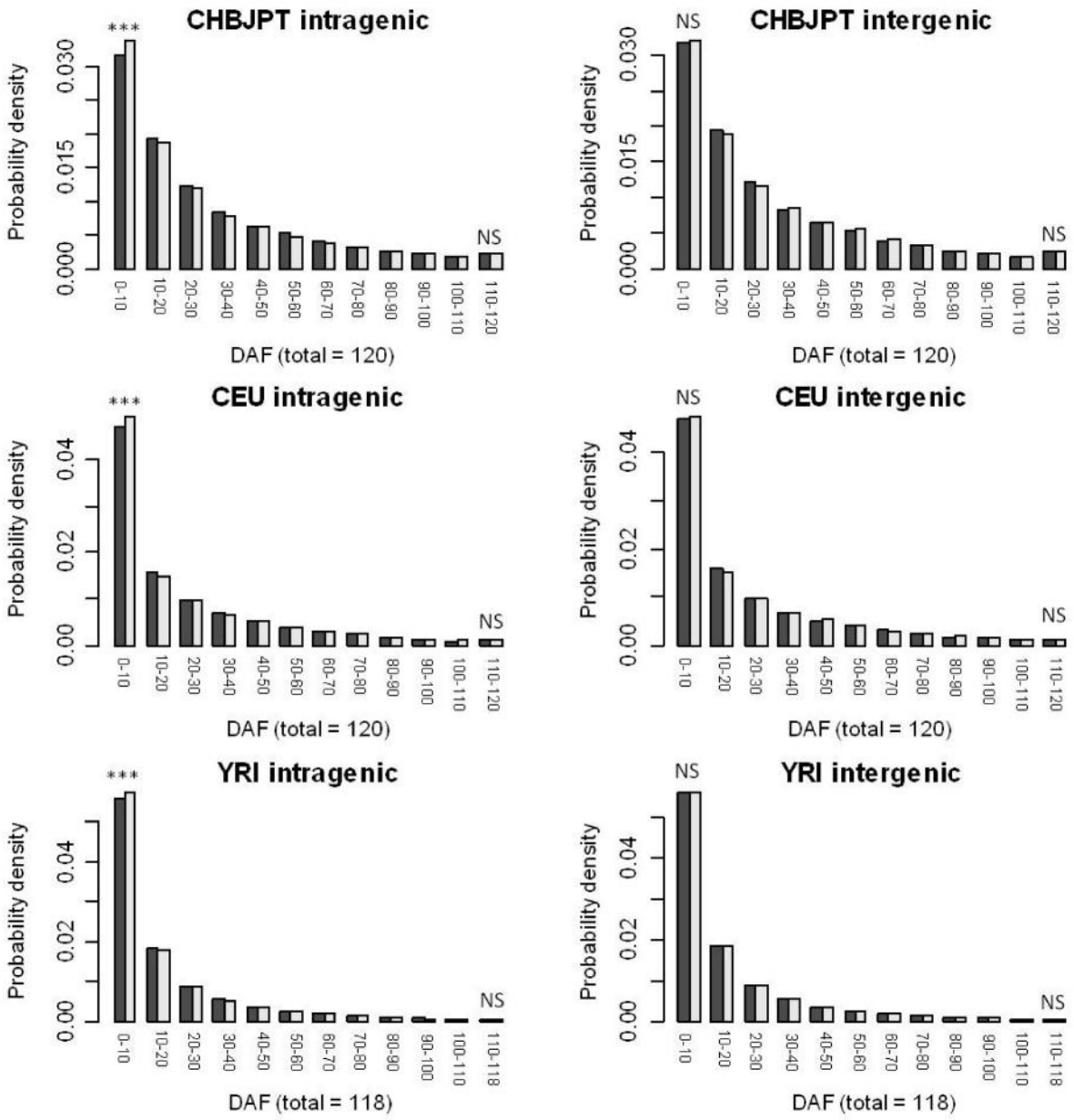

Figure B.2. Derived allele frequency (DAF) spectra of the intragenic maximal segments (left panels) and the intergenic maximal segments (right panels) for three populations. The grey bars represent the DAFs of the maximal segments and the black bars represent the DAFs of the putatively neutral regions. NS, not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

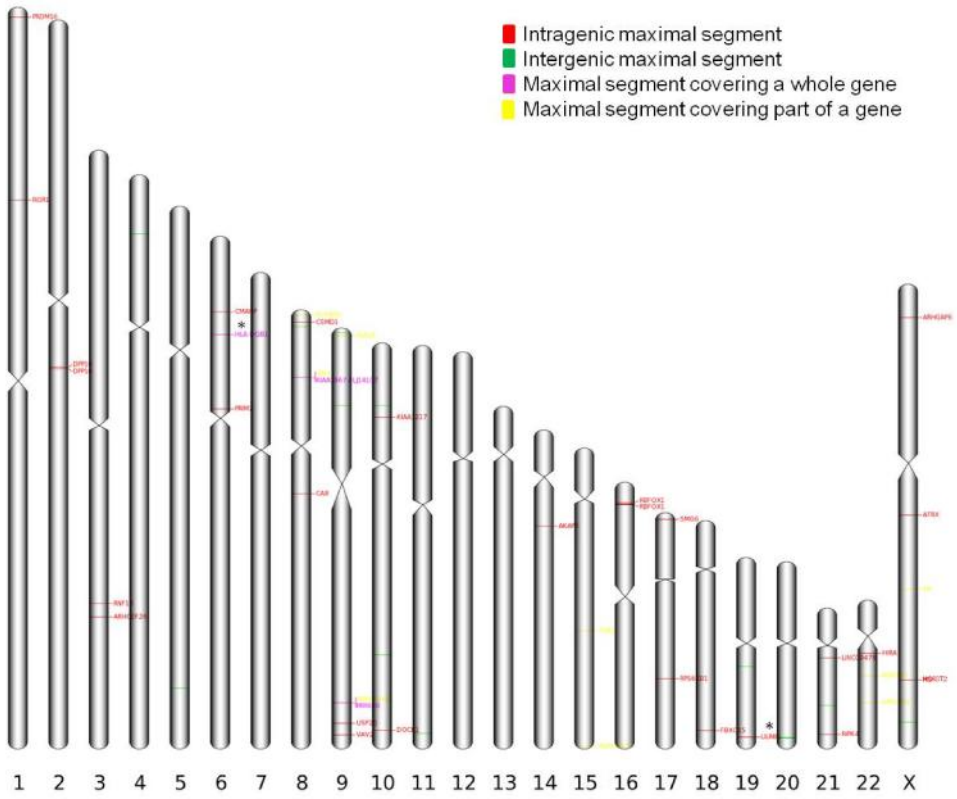

Figure B.3. Chromosomal distribution of the top 50 maximal segments with the lowest *P*-values in the human genome. Different types of genomic locations are represented in different colors: intragenic maximal segments are in red, intergenic maximal segments are in green, maximal segments covering whole genes are in pink, and maximal segments across intragenic and intergenic regions are in yellow. The gene names are annotated beside the locations. The positively selected genes HLA-DQB1 and LILRB1 are marked with stars.

APPENDIX C

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

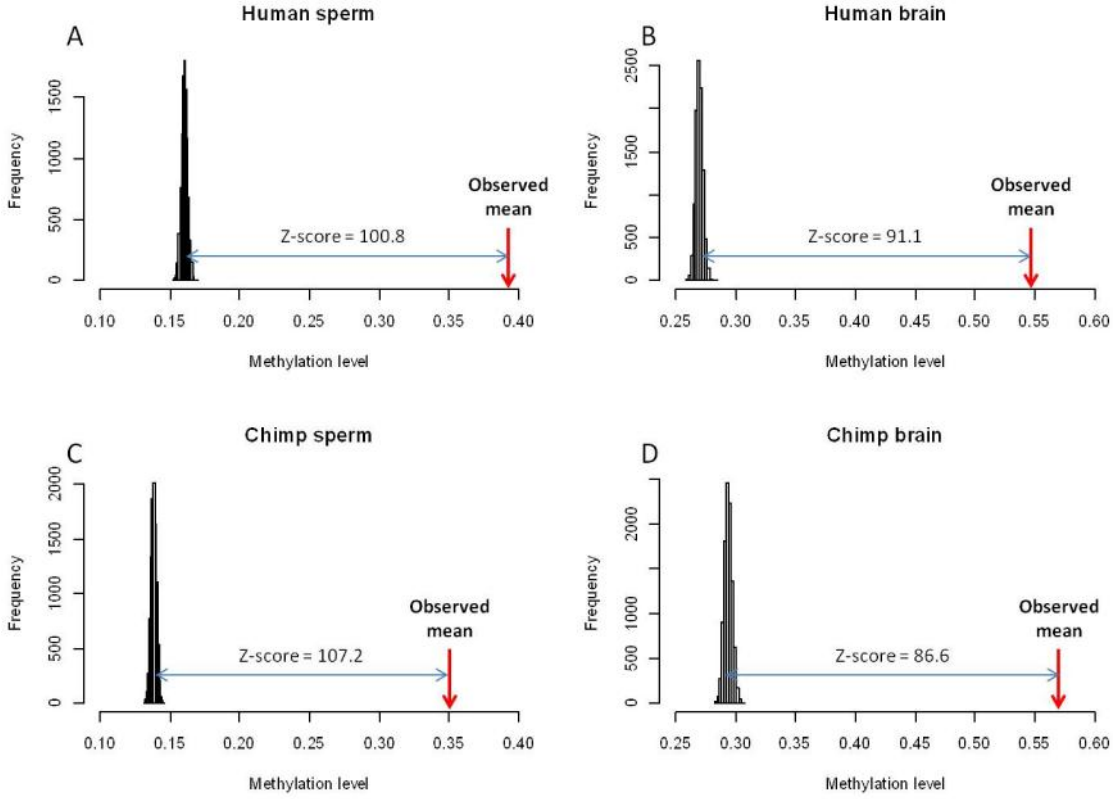

Figure C.1. Distribution of methylation levels of randomly chosen CpG sites in A) human sperm, B) human brain, C) chimpanzee sperm, and D) chimpanzee brain. The red arrows are the observed average methylation levels of newly generated CpG sites. Z-score measures how large the observed value deviates from the expected values.

REFERENCES

Arndt, P. F., D. A. Petrov, and T. Hwa. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. Mol. Biol. Evol. 20:1887-1896.

Avery, P. J. 1984. The Population-Genetics of Haplo-Diploids and X-Linked Genes. Genet. Res. 44:321-341.

Baines, J. F., S. A. Sawyer, D. L. Hartl, and J. Parsch. 2008. Effects of X-linkage and sex-biased gene expression on the rate of adaptive protein evolution in Drosophila. Mol. Biol. Evol. 25:1639-1650.

Baker, T. G., and W. Sum. 1976. Development of the ovary and oogenesis. Clin. Obsstet. Gynaecol. 3:3-26.

Bartosch-Härlid, A., S. Berlin, N. G. C. Smith, A. P. Møller, and H. Ellegren. 2003. Life history and the male mutation bias. Evolution 57:2398-2406.

Batzer, M. A., and P. L. Deininger. 2002. Alu repeats and human genomic diversity. Nat. Rev. Genet. 3:370-379.

Bauer, V. L., and C. F. Aquadro. 1997. Rates of DNA sequence evolution are not sex-biased in Drosophila melanogaster and D-simulans. Mol. Biol. Evol. 14:1252-1257.

Bell, C. G., G. A. Wilson, L. M. Butcher, C. Roos, L. Walter, and S. Beck. 2012. Human-specific CpG "beacons" identify loci associated with human-specific traits and disease. Epigenetics 7:1188-1199.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. B - Methodol. 57:289-300.

Berglund, J., K. S. Pollard, and M. T. Webster. 2009. Hotspots of biased nucleotide substitutions in human genes. PLoS Biol. 7:e26.

Betancourt, A. J., D. C. Presgraves, and W. J. Swanson. 2002. A test for faster X evolution in Drosophila. Mol. Biol. Evol. 19:1816-1819.

Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14:708-715.

Blumenstiel, J. P. 2007. Sperm competition can drive a male-biased mutation rate. J. Theor. Biol. 249:624-632.

Bohossian, H. B., H. Skaletsky, and D. C. Page. 2000. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. Nature 406:622-625.

Borum, K. 1961. Oogensis in the mouse. A study of meiotic prophase. Exp. Cell Res. 24:495-507.

Brawand, D., M. Soumillon, A. Necsulea, P. Julien, G. Csardi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, et al. 2011. The evolution of gene expression levels in mammalian organs. Nature 478:343-348.

Brudno, M., C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, E. D. Green, A. Sidow, and S. Batzoglou. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 13:721-731.

Bustamante, C. D., and S. Ramachandran. 2009. Evaluating signatures of sex-specific processes in the human genome. Nat. Genet. 41:8-10.

Caceres, M., J. Lachuer, M. A. Zapala, J. C. Redmond, L. Kudo, D. H. Geschwind, D. J. Lockhart, T. M. Preuss, and C. Barlow. 2003. Elevated gene expression levels distinguish human from non-human primate brains. Proc. Natl. Acad. Sci. USA 100:13030-13035.

Capra, J. A., and K. S. Pollard. 2011. Substitution patterns are GC-biased in divergent sequences across the metazoans. Genome Biol. Evol. 3:516-527.

Chang, B. H. J., L. C. Shimmin, S. K. Shyue, D. Hewettemmett, and W. H. Li. 1994. Weak Male-Driven Molecular Evolution in Rodents. Proc. Natl. Acad. Sci. USA 91:827-831.

Charlesworth, B. 2001. The effect of life-history and mode of inheritance on neutral genetic variability. Genet. Res. 77:153-166.

Charlesworth, B., J. A. Coyne, and N. H. Barton. 1987. The relative relative rates of evolution of sex chromosomes and autosomes. Am. Nat. 130:113-146.

Chatterjee, S., A. S. Hadi, and B. Price. 2000. Regression Analysis by Example, 3rd Edition. Wiley-Interscience, New York.

Chen, J. M., D. N. Cooper, N. Chuzhanova, C. Ferec, and G. P. Patrinos. 2007. Gene conversion: mechanisms, evolution and human disease. Nat. Rev. Genet. 8:762-775.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69-87.

Clark, A. G., S. Glanowski, R. Nielsen, P. D. Thomas, A. Kejariwal, M. A. Todd, D. M. Tanenbaum, D. Civello, F. Lu, B. Murphy, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960-1963.

Crow, J. F. 1997. The high spontaneous mutation rate: Is it a health risk? Proc. Natl. Acad. Sci. USA 94:8380-8386.

de Bakker, P. I., G. McVean, P. C. Sabeti, M. M. Miretti, T. Green, J. Marchini, X. Ke, A. J. Monsuur, P. Whittaker, M. Delgado, et al. 2006. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat. Genet. 38:1166-1172.

Dixson, A., and M. Anderson. 2001. Sexual selection and the comparative anatomy of reproduction in monkeys, apes, and human beings. Ann. Rev. Sex Res. 12:121-144.

Do, C. B., M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 15:330-340.

Dreszer, T. R., G. D. Wall, D. Haussler, and K. S. Pollard. 2007. Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. Genome Res. 17:1420-1430.

Drost, J. B., and W. R. Lee. 1995. Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among Drosophila, mouse, and human. Environ. Mol. Mutagenesis 25:48-64.

Ebersberger, I., P. Galgoczy, S. Taudien, S. Taenzer, M. Platzer, and A. von Haeseler. 2007. Mapping human genetic ancestry. Mol. Biol. Evol. 24:2266-2276.

Elango, N., S. H. Kim, E. Vigoda, and S. V. Yi. 2008. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. PLoS Comput. Biol. 4:e1000015.

Elango, N., J. Lee, Z. G. Peng, Y. H. E. Loh, and S. V. Yi. 2009. Evolutionary rate variation in Old World monkeys. Biol. Lett. 5:405-408.

Elango, N., J. W. Thomas, NISC Comparative Sequencing Program, and S. V. Yi. 2006. Variable molecular clocks in hominoids. Proc. Natl. Acad. Sci. USA 103:1370-1375.

Elango, N., and S. V. Yi. 2008. DNA methylation and structural and functional bimodality of vertebrate promoters. Mol. Biol. Evol. 25:1602-1608.

Ellegren, H. 2007. Characteristics, causes and evolutionary consequences of male-biased mutation. Proc. Biol. Sci. 274:1-10.

Ellegren, H., and A. K. Fridolfsson. 1997. Male-driven evolution of DNA sequences in birds. Nat. Genet. 17:182-184.

Ellegren, H., and A. K. Fridolfsson. 2003. Sex-specific mutation rates in salmonid fish. J. Mol. Evol. 56:458-463.

Emery, L. S., J. Felsenstein, and J. M. Akey. 2010. Estimators of the Human Effective Sex Ratio Detect Sex Biases on Different Timescales. Am. J. Hum. Genet. 87:848-856.

Enard, W., M. Przeworski, S. E. Fisher, C. S. Lai, V. Wiebe, T. Kitano, A. P. Monaco, and S. Paabo. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869-872.

Feinberg, A. P. 2007. Phenotypic plasticity and the epigenetics of human disease. Nature 447:433-440.

Felsenstein, J. 1985. Phylogenies and the Comparative Method. Am. Nat. 125:1-15.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5:164-166.

Fryxell, K. J., and E. Zuckerkandl. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol. Biol. Evol. 17:1371-1383.

Gibbs, J. R., M. P. van der Brug, D. G. Hernandez, B. J. Traynor, M. A. Nalls, S. L. Lai, S. Arepalli, A. Dillman, I. P. Rafferty, J. Troncoso, et al. 2010. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. 6:e1000952.

Goetting-Minesky, M. P., and K. D. Makova. 2006. Mammalian male mutation bias: impacts of generation time and regional variation in substitution rates. J. Mol. Evol. 63:537-544.

Gottipati, S., L. Arbiza, A. Siepel, A. G. Clark, and A. Keinan. 2011. Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. Nat. Genet. 43:741-743.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H. Fritz, et al. 2010. A draft sequence of the Neandertal genome. Science 328:710-722.

Grimwood, J., L. A. Gordon, A. Olsen, A. Terry, J. Schmutz, J. Lamerdin, U. Hellsten, D. Goodstein, O. Couronne, M. Tran-Gyamfi, et al. 2004. The DNA sequence and biology of human chromosome 19. Nature 428:529-535.

Grossman, L. I., D. E. Wildman, T. R. Schmidt, and M. Goodman. 2004. Accelerated evolution of the electron transport chain in anthropoid primates. Trends Genet. 20:578-585.

Haldane, J. B. S. 1947. The mutation rate of the gene for hemophilia, and its segregation ratios in males and females. Ann. Eugen. 13:262-272.

Halushka, M. K., J. B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. 22:239-247.

Hammer, M. F., F. L. Mendez, M. P. Cox, A. E. Woerner, and J. D. Wall. 2008. Sex-Biased Evolutionary Forces Shape Genomic Patterns of Human Diversity. PLoS Genet. 4: e1000202.

Harcourt, A. H., P. H. Harvey, S. G. Larson, and R. V. Short. 1981. Testes weight, body weight and breeding system in primates. Nature 293:55 - 57.

Harcourt, A. H., A. Purvis, and L. Liles. 1995. Sperm competition: mating system, not breeding season, affects testis size of primates. Func. Ecol. 9:468-476.

Hellmann-Blumberg, U., M. F. Hintz, J. M. Gatewood, and C. W. Schmid. 1993. Developmental differences in methylation of human Alu repeats. Mol. Cell. Biol. 13:4523-4530.

Hellmann, I., I. Ebersberger, S. E. Ptak, S. Paabo, and M. Przeworski. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. 72:1527-1535.

Huh, I., J. Zeng, T. Park, and S. V. Yi. 2013. DNA methylation and transcriptional noise. Epigenetics Chromatin 6:9.

Hurst, L. D., and H. Ellegren. 1998. Sex biases in the mutation rate. Trends Genet. 14:446-452.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Jones, E. Y., L. Fugger, J. L. Strominger, and C. Siebold. 2006. MHC class II proteins and disease: a structural perspective. Nat. Rev. Immunol. 6:271-282.

Karolchik, D., A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. 2004. The UCSC Table Browser data retrieval tool. Nucleic. Acids. Res. 32:D493-496.

Kass, S. U., N. Landsberger, and A. P. Wolffe. 1997. DNA methylation directs a time-dependent repression of transcription initiation. Curr. Biol. 7:157-165.

Keinan, A., J. C. Mullikin, N. Patterson, and D. Reich. 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat. Genet. 41:66-70.

Keinan, A., and D. Reich. 2010. Can a Sex-Biased Human Demography Account for the Reduced Effective Population Size of Chromosome X in Non-Africans? Mol. Biol. Evol. 27:2312-2321.

Kent, W. J. 2002. BLAT--the BLAST-like alignment tool. Genome Res. 12:656-664.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. 2002. The human genome browser at UCSC. Genome Res. 12:996-1006.

Khaitovich, P., I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Paabo. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science 309:1850-1854.

Kim, S. H., N. Elango, C. Warden, E. Vigoda, and S. V. Yi. 2006. Heterogeneous genomic molecular clocks in primates. PLoS Genet. 2:e163.

Kin, T., and Y. Ono. 2007. Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat. Bioinformatics 23:2945-2946.

Kirkpatrick, M., and D. W. Hall. 2004. Sexual selection and sex linkage. Evolution 58:683-691.

Klose, R. J., and A. P. Bird. 2006. Genomic DNA methylation: the mark and its mediators. Trends Biochem. Sci. 31:89-97.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, et al. 2002. A

high-resolution recombination map of the human genome. Nat. Genet. 31:241-247.

Kuhn, R. M., D. Haussler, and W. J. Kent. 2012. The UCSC genome browser and associated tools. Brief Bioinform. 14:144-161.

Kwiatkowski, D. P. 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. Am. J. Hum. Genet. 77:171-192.

Lai, C. S., S. E. Fisher, J. A. Hurst, F. Vargha-Khadem, and A. P. Monaco. 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413:519-523.

Lao, O., J. M. de Gruijter, K. van Duijn, A. Navarro, and M. Kayser. 2007. Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann. Hum. Genet. 71:354-369.

Leffler, E. M., Z. Gao, S. Pfeifer, L. Ségurel, A. Auton, O. Venn, R. Bowden, R. Bontrop, J. D. Wall, G. Sella, et al. 2013. Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. Science 339:1578-1582.

Lercher, M. J., and L. D. Hurst. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet. 18:337-340.

Li, W. H., S. J. Yi, and K. Makova. 2002. Male-driven evolution. Curr. Opin. Genet. Dev. 12:650-656.

Li, Y., J. Zhu, G. Tian, N. Li, Q. Li, M. Ye, H. Zheng, J. Yu, H. Wu, J. Sun, et al. 2010. The DNA methylome of human peripheral blood mononuclear cells. PLoS Biol. 8:e1000533.

Lienert, F., C. Wirbelauer, I. Som, A. Dean, F. Mohn, and D. Schubeler. 2011. Identification of genetic elements that autonomously determine DNA methylation states. Nat. Genet. 43:1091-1097.

Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315-322.

Locke, D. P., L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth, D. M. Muzny, S. P. Yang, Z. Wang, A. T. Chinwalla, P. Minx, et al. 2011. Comparative and demographic analysis of orang-utan genomes. Nature 469:529-533.

Majewski, J., and J. Ott. 2002. Distribution and characterization of regulatory elements in the human genome. Genome Res. 12:1827-1836.

Makova, K. D., and W. H. Li. 2002. Strong male-driven evolution of DNA sequences in humans and apes. Nature 416:624-626.

Mank, J. E., K. Nam, and H. Ellegren. 2009. Faster-Z Evolution is Predominantly Due to Genetic Drift. Mol. Biol. Evol. 27:661–670.

Mank, J. E., B. Vicoso, S. Berlin, and B. Charlesworth. 2010. Effective population size and the faster-X effect: empirical results and their interpretation. Evolution 64:663-674.

Martin, R. D. 2007. The evolution of human reproduction: A primatological perspective. Yearb. Phys. Anthropol. 50:59-84.

Meunier, J., and L. Duret. 2004. Recombination drives the evolution of GC-content in the human genome. Mol. Biol. Evol. 21:984-990.

Meunier, J., A. Khelifi, V. Navratil, and L. Duret. 2005. Homology-dependent methylation in primate repetitive DNA. Proc. Natl. Acad. Sci. USA 102:5471-5476.

Meyer, L. R., A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, et al. 2012a. The UCSC Genome Browser database: extensions and updates 2013. Nucleic. Acids. Res. 41:D64-69.

Meyer, M., M. Kircher, M. T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prufer, C. de Filippo, et al. 2012b. A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222-226.

Mi, H., A. Muruganujan, and P. D. Thomas. 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic. Acids. Res. 41:D377-386.

Miyata, T., H. Hayashida, K. Kuma, K. Mitsuyasu, and T. Yasunaga. 1987. Male-Driven Molecular Evolution - a Model and Nucleotide-Sequence Analysis. Cold Spring Harb. Symp. Quant. Biol. 52:863-867.

Molaro, A., E. Hodges, F. Fang, Q. Song, W. R. McCombie, G. J. Hannon, and A. D. Smith. 2011. Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell 146:1029-1041.

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321-324.

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

Patterson, N., D. J. Richter, S. Gnerre, E. S. Lander, and D. Reich. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441:1103-1108.

Penrose, L. S. 1955. Parental age and mutation. Lancet 269:312-313.

Perry, J., and A. Ashworth. 1999. Evolutionary rate of a gene affected by chromosomal position. Curr. Biol. 9:987-989.

Polavarapu, N., G. Arora, V. K. Mittal, and J. F. McDonald. 2011. Characterization and potential functional significance of human-chimpanzee large INDEL variation. Mob. DNA 2:13.

Pollard, K. S., S. R. Salama, B. King, A. D. Kern, T. Dreszer, S. Katzman, A. Siepel, J. S. Pedersen, G. Bejerano, R. Baertsch, et al. 2006a. Forces shaping the fastest evolving regions in the human genome. PLoS Genet. 2:e168.

Pollard, K. S., S. R. Salama, N. Lambert, M. A. Lambot, S. Coppens, J. S. Pedersen, S. Katzman, B. King, C. Onodera, A. Siepel, et al. 2006b. An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443:167-172.

Presgraves, D. C. 2008. Sex chromosomes and speciation in Drosophila. Trends Genet. 24:336-343.

Presgraves, D. C., and S. V. Yi. 2009. Doubts about complex speciation between humans and chimpanzees. Trends Ecol. Evol. 24:533-540.

Prufer, K., K. Munch, I. Hellmann, K. Akagi, J. R. Miller, B. Walenz, S. Koren, G. Sutton, C. Kodira, R. Winer, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. Nature 486:527-531.

Ptak, S. E., D. A. Hinds, K. Koehler, B. Nickel, N. Patil, D. G. Ballinger, M. Przeworski, K. A. Frazer, and S. Paabo. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. Nat. Genet. 37:429-434.

Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222-234.

Robson, S. L., and B. Wood. 2008. Hominin life history: reconstruction and evolution. J. Anat. 212:394-425.

Rowe, N., R. Mittermeier, and J. Goodall. 1996. The pictorial guide to the living primates. Pogonias Press.

Rubin, C. M., C. A. VandeVoort, R. L. Teplitz, and C. W. Schmid. 1994. Alu repeated DNAs are differentially methylated in primate germ cells. Nucleic. Acids. Res. 22:5121-5127.

Ruzzo, W. L., and M. Tompa. 1999. A linear time algorithm for finding all maximal scoring subsequences. Pp. 234-241. The Seventh International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park (CA).

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. 2006. Positive natural selection in the human lineage. Science 312:1614-1620.

Sarda, S., J. Zeng, B. G. Hunt, and S. V. Yi. 2012. The evolution of invertebrate gene body methylation. Mol. Biol. Evol. 29:1907-1916.

Sayres, M. A. W., Venditti, C., Pagel, M. and Makova, K. D. 2011. Do variations in substitution rates and male mutation bias correlate with life history traits? A study of 32 mammalian genomes. Evolution. 65:2800–2815.

Scally, A., J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, J. Herrero, A. Hobolth, T. Lappalainen, T. Mailund, T. Marques-Bonet, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483:169-175.

Schrider, D. R., J. N. Hourmozdi, and M. W. Hahn. 2011. Pervasive multinucleotide mutational events in eukaryotes. Curr. Biol. 21:1051-1054.

Shimmin, L. C., B. H. J. Chang, and W. H. Li. 1993. Male-driven evolution in DNA sequences. Nature 362:745 - 747.

Shulha, H. P., J. L. Crisci, D. Reshetov, J. S. Tushir, I. Cheung, R. Bharadwaj, H. J. Chou, I. B. Houston, C. J. Peter, A. C. Mitchell, et al. 2012. Human-specific histone methylation signatures at transcription start sites in prefrontal neurons. PLoS Biol. 10:e1001427.

Siegfried, Z., S. Eden, M. Mendelsohn, X. Feng, B. Z. Tsuberi, and H. Cedar. 1999. DNA methylation represses transcription in vivo. Nat. Genet. 22:203-206.

Singh, N. D., A. M. Larracuente, and A. G. Clark. 2008. Contrasting the efficacy of selection on the X and autosomes in Drosophila. Mol. Biol. Evol. 25:454-467.

Smit, A. F. A., R. Hubley, and P. Green. 1996-2010. RepeatMasker Open-3.0. http://www.repeatmasker.org.

Smith, R. J., and W. L. Jungers. 1997. Body mass in comparative primatology. J. Hum. Evol. 32:523-559.

Smith, R. L. 1984. Sperm competition and the evolution of animal mating systems. Academic Press, New York.

Steiper, M. E., and N. M. Young. 2006. Primate molecular divergence dates. Mol. Phylogenet. Evol. 41:384-394.

Strathern, J. N., B. K. Shafer, and C. B. McGill. 1995. DNA synthesis errors associated with double-strand-break repair. Genetics 140:965-972.

Sussman, R. W., and P. A. Garber. 1987. A new interpretation of the social organization and mating system of the Callitrichidae. Int. J. Primatol 8:73-92.

Suzuki, M. M., and A. Bird. 2008. DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev. Genet. 9:465-476.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861.

Thornton, K., D. Bachtrog, and P. Andolfatto. 2006. X chromosomes and autosomes evolve at similar rates in Drosophila: No evidence for faster-X protein evolution. Genome Res. 16:498-504.

Torgerson, D. G., and R. S. Singh. 2003. Sex-linked mammalian sperm proteins evolve faster than autosomal ones. Mol. Biol. Evol. 20:1705-1709.

Torgerson, D. G., and R. S. Singh. 2006. Enhanced adaptive evolution of sperm-expressed genes on the mammalian X chromosome. Heredity 96:39-44.

Tweedie, S., J. Charlton, V. Clark, and A. Bird. 1997. Methylation of genomes and genes at the invertebrate-vertebrate boundary. Mol. Cell Biol. 17:1469-1475.

van den Hurk, R., and J. Zhao. 2005. Formation of mammalian oocytes and their growth, differentiation and maturation within ovarian follicles. Theriogenology 63:1717-1751.

Vicoso, B., and B. Charlesworth. 2009. Effective Population Size and the Faster-X Effect: An Extended Model. Evolution 63:2413-2426.

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

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:e72.

Wang, H., J. Xing, D. Grover, D. J. Hedges, K. Han, J. A. Walker, and M. A. Batzer. 2005. SVA elements: a hominid-specific retroposon family. J. Mol. Biol. 354:994-1007.

Weber, M., I. Hellmann, M. B. Stadler, L. Ramos, S. Paabo, M. Rebhan, and D. Schubeler. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet. 39:457-466.

Wich, S. A., S. S. Utami-Atmoko, T. M. Setia, H. D. Rijksen, C. Schurmann, J. A. van Hooff, and C. P. van Schaik. 2004. Life history of wild Sumatran orangutans (Pongo abelii). J. Hum. Evol. 47:385-398.

Winckler, W., S. R. Myers, D. J. Richter, R. C. Onofrio, G. J. McDonald, R. E. Bontrop, G. A. McVean, S. B. Gabriel, D. Reich, P. Donnelly, and D. Altshuler. 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308:107-111.

Wootton, J. T. 1987. The Effects of Body-Mass, Phylogeny, Habitat, and Trophic Level on Mammalian Age at 1st Reproduction. Evolution 41:732-749.

Xiang, H., J. Zhu, Q. Chen, F. Dai, X. Li, M. Li, H. Zhang, G. Zhang, D. Li, Y. Dong, et al. 2010. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat. Biotechnol. 28:516-520.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15:568-573.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586-1591.

Yi, S., T. J. Summers, N. M. Pearson, and W. H. Li. 2004. Recombination has little effect on the rate of sequence divergence in pseudoautosomal boundary 1 among humans and great apes. Genome Res. 14:37-43.

Yi, S., D. L. Ellsworth, and W. H. Li. 2002. Slow molecular clocks in Old World monkeys, apes, and humans. Mol. Biol. Evol. 19:2191-2198.

Yi, S., and W. H. Li. 2005. Molecular evolution of recombination hotspots and highly recombining pseudoautosomal regions in hominoids. Mol. Biol. Evol. 22:1223-1230.

Zeng, J., G. Konopka, B. G. Hunt, T. M. Preuss, D. Geschwind, and S. V. Yi. 2012. Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. Am. J. Hum. Genet. 91:455-465.

PUBLICATIONS

The following publications (in chronological order) represent the work I have done during my PhD studies. Three of them are directly related to the studies from this dissertation (marked with *).

[1] *** Xu, K.**, Zeng, J., Yi, S. 2013. Parallel evolution of genomes and epigenomes between humans and chimpanzees. **(in preparation)**

[2] Jameson, N. M., Yi, S., **Xu, K.**, and Wildman, D. E. 2013. The tempo and mode of New World monkey evolution and biogeography in the context of phylogenomic analysis. ***Molecular Phylogenetics and Evolution*** **(in review)**

[3] *** Xu, K.**, Wang, J., Elango, N., and Yi, S. 2013. The evolution of lineage-specific clusters of single nucleotide substitutions in the human genome. ***Molecular Phylogenetics and Evolution*****, 69:** 276-285.

[4] Jameson, N. M., **Xu, K.**, Yi, S., and Wildman, D. E. 2012. Development and annotation of molecular markers from three New World monkey species. ***Molecular Ecology Resources*****, 12:** 950–955.

[5] *** Xu, K.**, Oh, S., Park, T., Presgraves, D. C., and Yi, S. 2012. Lineage-specific variation in slow- and fast-X evolution in primates. ***Evolution*****, 66:** 1751–1761.

[6] Park, J., **Xu, K.**, Park, T., and Yi, S. 2012. What are the determinants of gene expression levels and breadths in the human genome? ***Human Molecular Genetics*****, 21:** 46-56.

[7] **Xu, K.**, Bezakova, I., Bunimovich, L. and Yi, S. 2011. Path lengths in protein-protein interaction networks and biological complexity. ***Proteomics*****, 11:** 1857-1867.