The rhesus macaque is three times as diverse but more closely equivalent in “damaging” coding variation as compared to the human

Qiaoping Yuan[1*], Zhifeng Zhou[1*], Stephen G. Lindell[1], J. Dee Higley[2], Betsy Ferguson [3], Robert C. Thompson[4], Juan F. Lopez[5], Stephen J. Suomi[6], Basel Baghal[1], Maggie Baker[1], Deborah C. Mash[7], Christina S. Barr[1**], David Goldman[1**]

[1]Laboratory of Neurogenetics, National Institute on Alcohol Abuse and Alcoholism, NIH, Bethesda, MD 20892, USA; [2]Laboratory of Clinical and Translational Studies, NIAAA, Bethesda, MD 20892, USA; [3]Oregon National Primate Research Center, Oregon Health and Sciences University, 505 NW 185th Ave., Beaverton, OR 97006, USA; [4]Department of Psychiatry, University of Michigan, Ann Arbor, MI 48104, USA; [5]Mental Health Research Institute, University of Michigan Medical Center, 3064 NSL, 1103 East Huron Street, Ann Arbor, MI 48104, USA; [6]Laboratory of Comparative Ethology, National Institute of Child Health and Human Development, NIH, Poolesville, MD 20837, USA; [7]Department of Neurology, University of Miami School of Medicine, Miami, FL 33124, USA

Abstract - *Using a parallel next-gen sequencing and analytic pipeline, we sequenced the whole mRNA transcriptome and trimethylated histone H3-lysine 4 marked DNA regions in hippocampus from 14 humans and 14 rhesus macaques. Using this equivalent methodology and sampling space, we identified 462,802 macaque SNPs, most novel and disproportionately located in functionally important genomic regions. At least one SNP was identified in each of more than 16,000 annotated macaque genes. Comparative analyses with these SNPs equivalently identified in the two species revealed that rhesus macaque has approximately three times higher SNP density and average nucleotide diversity as compared to the human. The effective population size of the rhesus macaque is estimated to be approximately 80,000 and several times that of the human. Across five different genomic regions (intergenic, 5 Kb upstream of transcription start site, introns, untranslated, coding), intergenic regions had the highest SNP density and average nucleotide diversity and coding sequences the lowest, in both human and macaque. Although there are more coding SNPs (cSNPs) per individual in macaque than in human, the ratio of d_N/d_S in macaque is significantly lower than that in human. Furthermore, the number of predicted “damaging” nonsynonymous cSNPs in macaque is more closely equivalent to that of the human.*

Keywords: Macaque, Human, Sequencing variation, Single nucleotide diversity, SNP density, Comparative genomics

1 Introduction

Rhesus macaque (*Macaca mulatta*) monkeys and humans (*Homo sapiens*) are thought to have shared a common ancestor approximately 25 million years ago [1]. Due to their genetic, physiological and behavioral similarities with humans, and because of their hardiness, adaptability, and availability, the rhesus macaque has been widely used as a nonhuman primate model in biomedical research [2,3]. Humans presently are the most numerous and widespread of primates. Furthermore, hominid apes representing the ancestral lineage of humans were geographically widespread, their fossils having been found in both Africa and Asia. However, the human diaspora is relatively recent, with our African ancestry dating back only 80,000 to 150,000 yrs b.p [4]. Also, the number of humans worldwide numbered as low as one million as recently as 100,000 yrs ago [5], and due to limitations in dispersion and gene flow effective population sizes were probably much smaller still. Substantial evidence exists that the neutral genetic diversity of humans has been shaped, and in fact restricted, by an effective population size that until recently was less than 8,000 [6].

The geographic range of the rhesus macaque extends from Afghanistan to the East China Sea. The population presently numbers in the millions, and in its range and population size the rhesus macaque is only exceeded by the humans among primate species [7]. Fossil evidence indicates that the *Macaca* genus originated in North Africa, and dispersed to various sites in Asia at least three million years ago [8]. The rhesus macaque has adapted to a variety of natural environments, including savannah and forests, and various climatic zones. Rhesus macaques thrive in cities – where they live side by side with man. The diversity of environmental adaptations and large current and ancestral population sizes suggests that the genetic legacy of the rhesus macaque may include a higher quotient of both neutral and selectively significant genetic variation than humans. Consistent with a high degree of genetic variation, substantial morphological variation has been observed between rhesus macaques in the same populations and also between populations, with as many as 13 subspecies

[*] These authors contributed equally to this work.
[**] Correspondences: cbarr@mail.nih.gov and davidgoldman@mail.nih.gov

identified [9]. Within rhesus macaques there is some evidence for genetic distinctiveness at the molecular level, and Indian rhesus may be among the least diverse [10]. Several studies using protein polymorphisms have found higher levels of diversity in Rhesus macaques from China (where there are also more subspecies) than India, and there is some evidence for a genetic bottleneck in Indian Rhesus macaques [9]. However, substantial gene flow probably occurred later, which could refresh genetic variation. In a study of six rhesus macaque populations, including Indian, Burmese, and four Chinese populations, Indian macaques had one third to one sixth the mitochondrial DNA diversity as compared to four other populations. But the Indian macaques were approximately equal in diversity to one of the Western Chinese populations [9]. A recent study with more than 1,000 Single nucleotide polymorphisms (SNPs), which are more mutationally stable than other types of polymorphisms, revealed that Indian and Chinese rhesus macaques were nearly identical in genetic diversity [11]. Taken together, the evidence suggests that the rhesus macaque is likely to be a genetically diverse primate species but Indian macaques are if anything among the least heterogeneous populations. Genomic analysis of rhesus macaques of Indian origin would thus provide a conservative estimate of the variability of rhesus macaques.

A draft genome sequence of a single Rhesus macaque of Indian origin was completed in 2007 [3]. This draft sequence opened the opportunity to map the amount and type of macaque genomic variation. Furthermore, characterization of genetic variation in macaques would greatly improve the value of the rhesus macaque as an animal model for human biology. However, there has been no systematic genome-wide view of the genetic diversity within this species. At present, fewer than 8,000 SNPs from macaque have been recorded (dbSNP Build 131, http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi?view+summary=view+summary&build_id=131). In 2007, Malhi et al. reported about 23,000 candidate SNPs from pyrosequencing [12].

Compatible with its larger effective population size across evolutionary timeframes, the macaque appears to have higher sequence diversity than the human [3,13]. SNP density in macaques was estimated to range 1~7.8 SNPs/Kb [3,14]**.** However, the number of loci on which this conclusion is based is relatively small, and the loci were not selected in an unbiased fashion. Although >22 million human SNPs are recorded, the availability of <10,000 macaque SNPs prevents large scale sequence diversity comparison between human and macaque in different genomic regions. In this study, we used SNPs equivalently identified in 14 humans and 14 rhesus macaques by massively parallel sequencing with both H3K4me3 (trimethylated histone H3-lysine 4) ChIPseq (chromatin immunoprecipitation followed with massively parallel DNA sequencing) and RNAseq (whole transcriptome massively parallel shotgun sequencing) as sources of sequenced fragments. From more than 16,000 genes some half million macaque SNPs, most newly identified, were further analyzed and the extent of diversity was compared between humans and macaques in different genomic regions to capture effects of neutral genetic drift and selection in these two primate species. By sequencing diversity in the tissue-specific transcriptomes and histone-marked regions of the two species, we were able, without the use of DNA capture technology (that did not exist for the macaque) or whole-genome sequencing, to compare diversity in equivalent, functionally relevant genomic regions and detect effects of selection and drift on sequence substitutions in protein-coding gene regions.

2 Methods

2.1 Samples and tissues

Postmortem brain tissue (hippocampus) of 14 unrelated human (*H. sapiens*) males, age 30-50 was obtained from the University of Miami Brain Endowment Bank (Miami, FL, USA). The ethnic background of the human sample was: 6 African Americans, 8 Caucasians/Hispanics. Postmortem hippocampus of 14 rhesus macaque (*M. mulatta*) males, most unrelated, age 3.5-7, was obtained from the National Institutes of Health Animal Center in Poolesville, Maryland. Among the macaques, eleven were of Indian origin, one was of Chinese origin and two were approximately 50% Chinese/50% Indian as indicated by forensic genotyping with a panel of 96 markers optimized for macaque origin identification (Primate Genetics Program, Oregon National Primate Research Center, Table S1). The macaques at the Poolesville colony are maintained in an outbred state, with frequent introduction of new breeding stock such that their genetic diversity is expected to be equivalent to natural populations.

2.2 Construction of double-stranded cDNA libraries

Total RNA was extracted from 100 mg of hippocampus collected postmortem. Briefly, tissue samples were submerged in guanidinium thiocyanate and phenol based RNA extraction solution STAT-60 (Invitrogen, Friendswood, TX) and homogenized using a glass-Teflon homogenizer. Following mixing with chloroform and centrifugation, the aqueous phase was collected and isopropanol was added. The samples were then loaded onto RNeasy spin columns (Qiagen, Valencia, CA) for purification. To eliminate residual genomic DNA contamination, RNA samples were incubated with DNase I (Qiagen) on column at room temperature for 15 min and washed several times before collection in elution buffer. To isolate mRNA, 35 µg of total RNA was heated at 65ºC for 2 min, and then mixed with 0.5 mg of Dynabeads oligo $(dT)_{25}$ (Invitrogen) in binding buffer (20 mM Tris-HCl, pH 7.5, 1.0 M LiCl, 2 mM EDTA). After incubation at room temperature for 5 min and then washing several times, mRNA was eluted from the beads by heating at 80ºC for 2 min. The purified mRNA was fragmented to the 150 – 500 base pair range by mixing with 10 x fragmentation buffer (Ambion, Austin, TX) and heating at 70ºC for 3 min. The samples were purified with RNeasy spin column. 200 ng of fragmented mRNA was reverse-transcribed to first strand cDNA by random priming, using 3 μg of random hexamer oligos and 200 units of Superscript II reverse transcriptase (Invitrogen). The reaction was carried out at 45ºC for 1 hr in First Strand Buffer (Invitrogen) with 10 mM DTT and 0.5 mM dNTP. For second-strand cDNA synthesis, 400 units of *Escherichia Coli* DNA polymerase, 2 units of *E. Coli* RNase H, and 10 units of *E. Coli* DNA ligase was added, and the reaction was carried out at 16ºC for 2 hr in Second Strand Buffer with 0.2 mM dNTP. 20 units of T4 DNA polymerase was also added at the end of incubation for endrepair. The synthesized double-stranded cDNA library was purified with QIAquick purification kit (Qiagen).

2.3 Chromatin immunoprecipitation (ChIP)

Postmortem brain tissue (100 mg) was cut into slices less than 1 mm in thickness, and fixed in 3 ml of 1% formaldehyde/PBS solution for 10 min at room temperature to cross-link chromatin DNA and proteins. The tissue samples were then homogenized using a glass-Teflon homogenizer. Following homogenization, chromatin was isolated using the Upstate Magna ChIP G kit (Millipore, Temecula, CA). Briefly, cells were lysed in Cell Lysis Buffer in the presence of protein inhibitor cocktail. Nuclei were isolated from lysed cells by centrifugation, and re-suspended in Nuclear Lysis Buffer. The chromatin DNA was then fragmented into the 150 – 500 base-pair range by sonication using a Branson Sonifer (Branson, Danbury, Connecticut). To immunoprecipitate specific genomic regions of chromatin DNA, isolated chromatin was incubated with antibodies (Abcam, Cambridge, MA) against H3K4me3 and magnetic protein G beads (Millipore) at 4ºC for 2.5 hr. Following incubation, beads were washed with low salt, high salt, LiCl salt, and TE buffers, and reverse cross-linked by proteinase K digestion at 62ºC for 2 hr. The enriched DNA was purified after reverse cross-linking by column purification.

2.4 Sequencing with Illumina Genome Analyzer

Sample preparation and sequencing on an Illumina Genome Analyzer (Illumina, San Diego, CA) were carried out according to Illumina protocols with some modifications. Briefly, double-stranded cDNA and ChIP-enriched genomic DNA were treated with T4 DNA polymerase and Klenow fragment for end repair. The 5' ends of DNA fragments were then phosphorylated by T4 polynucleotide kinase, and an adenosine base was added to the 3' end of the fragments by Klenow (3'-5' exo$^-$). A universal adaptor was added to the both ends of the DNA fragments by A-T ligation. Following 18 cycles of PCR with Phusion DNA polymerase, the DNA library was purified on a 2% agarose gel, and fragments 170 – 350 bp in size were recovered. Approximately 10 ng of the prepared DNA was then used for cluster generation on a grafted Flow Cell, and sequenced on the Genome Analyzer for 36 cycles using the "Sequencing-by-synthesis" method.

2.5 SNP calling and sequence analyses

Sequences were called from image files with the Illumina Genome Analyzer Pipeline (GApipeline) and aligned to the corresponding reference genome (UCSC rheMac2 for macaque and UCSC hg18 for human) using Extended Eland in the GApipeline. The uniquely mapped reads were parsed with in-house Perl scripts to generate base coverage and SNP calls as described previously [15]. To reduce false positive and false negative SNP calling for low coverage sequence data, a two-step approach was used. Briefly, reads were first pooled from all samples in a species for SNP identification. At this step, no base in the uniquely mapped reads had a quality score < 8, only a single mis-match with quality score ≥ 15 was allowed in a single 36-base read, and a probable SNP had to have three independent reads representing the same alternative allele within the pooled samples. To reduce false SNP calls due to mis-mapping of cross-exon RNAseq reads, putative SNPs were filtered to remove instances in which the alternative allele was represented only by reads located one or two bases from either end of the RNAseq fragment. Candidate SNPs were then filtered at the individual sample level, where the frequency of the alternative allele in a single sample had to be the highest or second highest with a frequency ≥ 0.2. Genotypes were called for an individual sample only when sequencing coverage was ≥ 6x for the SNP site and when the allele with the lowest coverage was represented at least 3 times and heterozygotes with each allele covered by 30~70% of sequence reads. Gene structures for human were based on RefSeq Genes in UCSC hg18 and Ensembl Genes from UCSC rheMac2 were used for the macaque. PolyPhen-2 [16] was used to predict protein functional effects of nonsynonymous coding SNPs (nsSNPs). Fourteen novel macaque cSNPs were selected to be resequenced by Sanger sequencing using the BigDye Terminator Sequencing Mix (Applied Biosystems, Carlsbad, CA) and analyzed on the Applied Biosystems 3730 DNA Analyzer.

3 Results and Discussion

3.1 SNP density is three times higher in the rhesus macaque than the human

In this study diversity was determined in short sequence reads (36 bases) equivalently detected and analyzed in 14 humans and 14 rhesus macaques (Table 1) (The raw sequences generated in this study have been deposited in The Sequence Read Archive with the accession numbers of SRA028822, SRA027316, SRA029279 and SRA029275). It is important to point out that the analytical strategy of comparing diversity within the hippocampal transcriptome and in H3K4me3-marked DNA regions resulted in the analysis of equivalent regions in the macaque and in the human. There was a strong correlation between level of expression of genic associated sequences between the hippocampus of both species and in the regions strongly tagged by H3K4me3 (Fig. S1). From these equivalent genomic regions with at least 3x sequencing coverage, a total of 462,802 high quality putative SNPs (most of which were novel) were detected in the macaque, and 230,028 (most of which were known) were detected in the human. At least one SNP was identified in each of 14,675 human annotated genes and 16,797 macaque annotated genes.

Table 1. Summary of sequence coverage and putative SNPs

		Human	Rhesus
Genome size in reference assembly (Mb)		3,080	2,864
Non-gap reference genome size (Mb)		2,858	2,647
Unique coding sequence size in reference (Mb)		32.5	31.8
Sample number		14	14
Average 36-base reads per sample		17.4 x 10^6	14.4 x 10^6
Total length (Mb) of uniquely mapped reads		8,770	7,266
Mb in genome with (≥1x sequence coverage)		1,505	1,571
Mb in genome with (≥3x sequence coverage)		426	435
SNPs in dbSNP_B 131		23.7 x 10^6	7,880
SNPs in this study		230,028	462,802
Also in dbSNP_B131		206,267 (89.7%)	34 (0.0%)
Transition	AG,GA,TC,CT	155,836 (67.7%)	312,064 (67.4%)
Transversion	AC,CA,TG,GT	37,046 (16.1%)	79,061 (17.1%)
Transversion	CG,GC	25,467 (11.1%)	46,820 (10.1%)
Transversion	AT,TA	11,679 (5.1%)	24,857 (5.4%)
Genes with SNPs		14,675	16,797
Genes with SNPs in exons		11,200	12,466
SNPs located in intergenic regions		107,461 (46.7%)	269,390 (58.2%)
SNPs locate in 5Kb upstream of TSS		10,036 (4.4%)	26,303 (5.7%)
SNPs located in UTR		18,432 (8.0%)	15,455 (3.3%)
SNPs located in intron		79,875 (34.7%)	130,443 (28.2%)
SNPs located in CDS		14,224 (6.2%)	21,211 (4.6%)
Synonymous		8,329 (58.6%)	13,798 (65.1%)
Non-synonymous		5,877 (41.3%)	7,367 (34.7%)
Damaging		1,741 (29.6%)	1,525 (20.7%)
Nonsense		18 (0.1%)	46 (0.2%)

Approximately 10~25% of the putative SNPs detected in intergenic regions were found to be covered with RNAseq reads (Table S2), suggesting that significant transcription activity occurred outside of defined genic regions in both species, consistent with those reported recently [17]. Among 230,028 putative human SNPs, 90% had been recorded previously in dbSNP. This rediscovery rate is slightly higher than the 77-89% rediscovery rate for SNPs in the 1000 Genomes Project Pilot 2 deep sequencing data [18]. Also bearing on the validity of the SNP detection pipeline, the transition to transversion ratio of human and macaque SNPs was non-random. Although the random transition to transversion ratio is 1:2, this ratio is approximately 2:1 in both human and macaque. Using the same SNP calling pipeline, 22 of 26 human nsSNPs were validated by Sanger sequencing in a previous study [15]. Using Sanger sequencing, 13 of 14 novel macaque cSNPs identified in this study were also verified. Overall, the rhesus macaque had a SNP density approximately three times higher than humans (Fig.1A). Calculated across all genomic regions with at least 4x sequencing coverage in individual samples, the SNP densities for macaques and humans were 2.82 SNP/kb and 1.07 SNP/Kb, respectively (Table 2, Fig.1A).

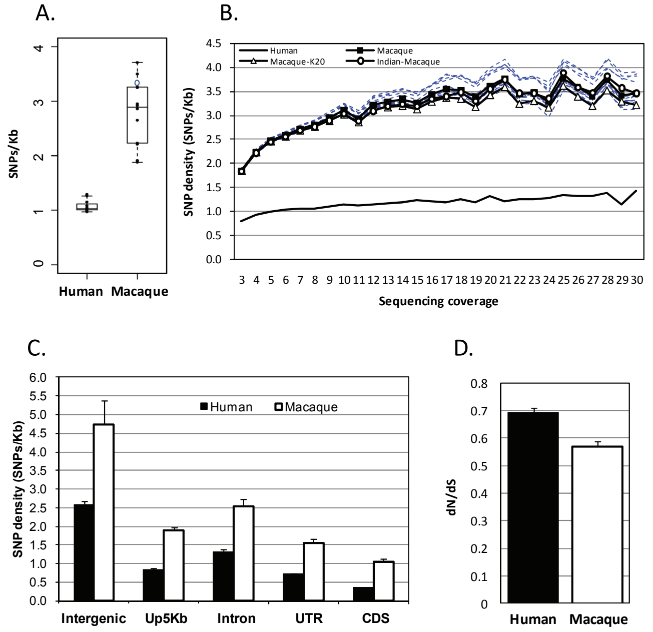

Fig. 1: Average SNP density (SNPs per 1Kb) in human and macaque. A). SNP density was calculated as the putative SNPs having different allele from reference genome divided by the unique sequenced bases in individual samples. Only bases having ≥ 4x sequence coverage were used for this calculation. Macaque sample K20, with Chinese origin, is labeled as an unfilled circle. B). Average SNP density in human and macaque was calculated for different sequencing coverage. Data from all macaque samples in solid line with filled square markers. Data with K20 omitted in solid line with unfilled triangle markers and others omitted one-by-one in dotted lines. Indian macaques only in solid line with unfilled circle markers. C). SNP density in 5 different genomic regions; D). The ratio of dN/dS for cSNPs. Error bar in C and D: standard error of mean.

Table 2. SNP density

Genome	Technology Used	SNPs/Kb
Venter	Sanger method	1.41*
Watson	454 Sequencing System (Roche)	1.46*
Chinese (YH)	Genome Analyzer (Illumina)	1.35*
African (NA18507)	Genome Analyzer (Illumina)	1.58*
African (NA18507)	SOLiD system (ABI)	1.69*
Korean (SJK)	Genome Analyzer (Illumina)	1.50*
Korean (AK1)	Genome Analyzer (Illumina)	1.51*
Proband (III-4)	SOLiD system (ABI)	1.50*
CEU,YRI	Genome Analyzer, SOLiD, 454	1.21-1.48**
Humans in this study	Genome Analyzer (Illumina)	1.07 (0.97-1.26)
Macaques in this study	Genome Analyzer (Illumina)	2.82 (1.88-3.71)

* SNP number was from Lupski et al. 2010 [19] and SNPs/Kb was calculated based on the total SNPs reported and 2.85 x 10^9 of the sequenced human genome size and 80% of accessible genome [18].

** Based on the SNPs and accessible genome from high coverage Pilot Trios data of 1000 Genomes Project [18].

Because sequencing coverage for individual samples was low for most regions, putative SNPs were called by a conservative, two-step approach as described in methods. As a result, SNP density increased in both species as sequencing coverage increased (Fig. 1B), but it can be observed that the macaque had proportionately higher SNP density at all levels of sequencing coverage (Fig. 1B). One of the macaque samples was of Chinese origin and two were approximately equally admixed between Chinese macaque and Indian macaques as described in methods. However, in the comparison between macaque and human, this Chinese macaque (K20) and the two admixed macaques did not exert a larger effect on SNP density as compared to any of the Indian macaques. This was tested by omitting individual macaques one-by-one, and also by evaluating SNP density with all three of the animals with Chinese ancestry omitted (Fig. 1B). The result is consistent with what found in a recent study where no difference was found in genetic diversity between Chinese and Indian macaques using genotype analyses with more than 1,000 SNPs [11]. As mentioned, our human sample itself included individuals of different ethnic backgrounds. Therefore, the Chinese macaque and the two admixtures were included in all analyses unless specified otherwise.

At higher coverage, SNP density approached that found by higher coverage sequencing, being 1.5 SNPs/Kb for 30x coverage across human 14 samples. A range of 3.07 ~ 3.86 x 10^6 SNPs was found in individual human genomes [19] representing approximately 1.3~1.7 SNP/Kb. Also, a SNP density of 1.2 ~ 1.5 SNPs/Kb was found in the 1000 Genomes Project Pilot 2 data for two human family trios with >40 x sequencing coverage [18]. Here, SNP densities were estimated from 14 samples in both species and with highly similar sequencing coverage, representing a methodologically equivalent view of diversity. Since intergenic and intronic regions comprise the majority of the genome in both humans and macaques, the overall SNP densities reported here are most likely underestimates because a high proportion of our data derives from coding sequences (CDS) and untranslated regions (UTR) that have the lowest SNP densities, as will be discussed below and as shown in Fig. 1C.

SNP density was compared across five different categories of genomic regions: intergenic, 5 Kb upstream of TSS (transcription start site), introns, UTR (5'- and 3'-UTRs), and CDS as annotated in refGene (human) or ENSEMBL (macaque). In all five genomic regions, macaques had significant higher SNP densities than humans (Fig. 1C). Intergenic regions had the highest SNP density and coding regions the lowest SNP density in both species (Fig. 1C). In coding regions, 76% of the cSNPs would be expected to be nsSNPs if all base substitutions were equally likely [20]. But nsSNP density was lower than synonymous cSNP density with a d_N/d_S ratio (the ratio of nonsynonymous versus synonymous substitutions, reflecting selection pressure acting on nonsynonymous sites relative to synonymous ones) in humans approximately 0.691±0.017 and d_N/d_S ratio of 0.567±0.022 in macaque (Fig. 1D). Although both adaptation and purifying selection may have occurred at numerous genes for both species, purifying selection is most likely to be predominant across the whole genome in both species as their d_N/d_S ratio values were significantly less than 1. The selection pressure on nonsynonymous substitutions may have been stronger in the macaque than in the human since the d_N/d_S ratio in macaque is significantly (t-test, p-value <0.0001) lower than human. In an equivalent genomic search space, twice as many putative SNPs were identified in macaque as compared to the human (Table 1). However, macaques only had 1.2 times as many nsSNPs, reflecting that much of the increased diversity of the macaque, even in protein-coding regions of the genome, is likely to be selectively neutral. Furthermore, the nsSNPs of macaques were less likely to be "damaging" (including "possibly damaging" and "probably damaging") as compared to the human (20% in macaque vs 30% in human), at least as predicted by PolyPhen-2 (Table 1). In line with this result, the higher d_N/d_S ratio in human may reflect a relative relaxation of purifying selection during hominoid evolution as a consequence of smaller effective population sizes or a high rate of adaptive substitution [21].

Using RNAseq and H3K4me3 ChIPseq data, a relatively high percentage of SNPs can be identified in gene coding and promoter regions, which represent functionally important domains of the genome. This could represent an advantage for certain types of gene-centric analyses. For instance, 6.2% of the human SNPs detected in this study (and 90% are previously known) were located in coding regions (cSNPs), whereas only 0.7% of the total SNPs identified in 1000 Genomes Project Pilot 2 data were cSNPs [18]. Here we sequenced only 0.426 Gb of unique human sequence at ≥3x coverage, but detected 14,224 cSNPs. This is a substantial number given that 24,192 cSNPs were detected in three Caucasian individuals with whole genome sequenced at high coverage, in the 1000 Genomes Project Pilot 2 (Fig. S2). The major limitation for SNP detection here was the proportion of genes that are not expressed in adult hippocampus or that are expressed at a low level in this tissue. The overlap of the cSNPs we detected with those reported in two individuals from the 1000 Genomes Project Pilot 2 data is consistent with the overlap that has been empirically observed between unrelated individuals (50~70% SNPs shared) on a pairwise basis (Fig. S3) [18]. Based on our sensitivity of detection of human SNPs, where 14,224 cSNPs were detected versus some 250,000 cSNPs that have been reported in NCBI (from a much larger population of subjects), we estimate that our region-focused sequencing of only 14 individuals enabled us to discover approximately 6% of the common cSNPs that are present in the rhesus macaque, although detection sensitivity was of course higher for the more abundant SNPs.

3.2 Rhesus macaques are three times as diverse as the human

The average nucleotide heterozygosity (diversity) for SNPs ($\theta_{SNP,}$ as defined by Levy et al. 2007[22]) in this study was measured as the ratio of heterozygous basepairs (both alleles with ≥3x coverage and ≥ 30% of sequence reads) divided by all basepairs sequenced at this level, within each individual. Macaque θ_{SNP} was 3 times higher than human θ_{SNP} (8.93×10^{-4} vs 3.06×10^{-4}, Fig. 2A).

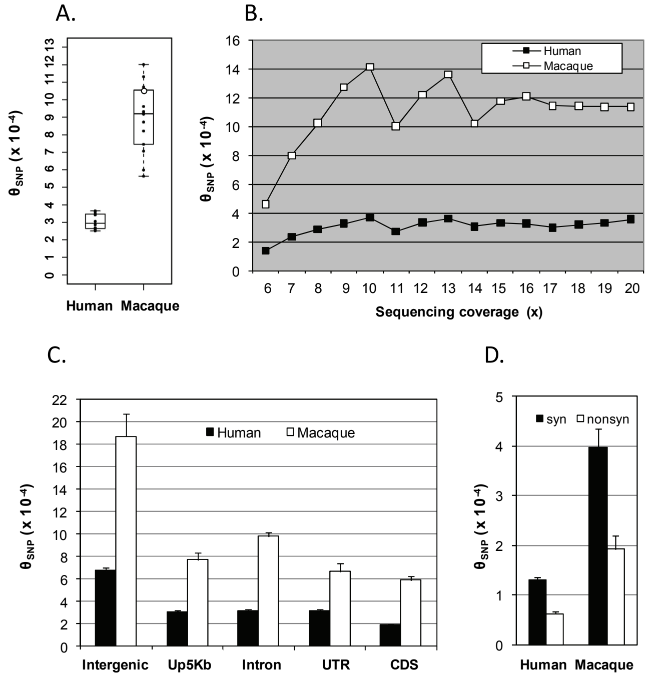

Fig. 2: Average nucleotide diversity (θ_{SNP}). A). θ_{SNP} in individual samples. Calculation was based on bases with ≥ 6x sequence coverage. Macaque sample K20, with Chinese origin, was labeled as an unfilled circle. B). Nucleotide diversity was average from all samples in each species at different sequencing coverage. C). Average nucleotide diversity in 5 different genomic regions; D). Average nucleotide diversity for synonymous cSNPs and nsSNPs. Error bar in C and D: standard error of mean.

Paralleling observations on SNP density, as sequencing coverage increases, more heterozygous basepairs are detected. With increasing sequencing coverage, θ_{SNP} increased, becoming asymptotic at about 10x coverage (Fig. 2B). At 20x sequencing coverage, θ_{SNP} was 11.4×10^{-4} in the macaque and 3.6×10^{-4} in the human (Fig.2B). Similar to what we observed for SNP density, θ_{SNP} was highest in intergenic regions and lowest in coding regions in both species and macaque had significant higher θ_{SNP} than human in all five genomic regions (Fig. 2C). Our estimated θ_{SNP} in human from all regions (3.06×10^{-4}) is lower than values that can be calculated (See supplementary Table S4) from 1000 Genomes Project Pilot 2 data (7.2×10^{-4} to 9.3×10^{-4}) and is also slightly lower than values of 5.4×10^{-4} to 8.3×10^{-4} reported previously for humans [22-26]. Our overall lower human θ_{SNP} than those reported previously was expected due to our lighter sequencing coverage and higher genic percentage of sequenced regions. However, within intergenic regions that comprise most of the genome our θ_{SNP} estimate of ~6.78×10^{-4} is actually very close to these previously reported estimates for humans based on high coverage whole genome sequencing. Therefore, the θ_{SNP} values we have computed for macaque and human appear to be robust, reflect parallel methodology and sampling and are informative for both genome-wide and regional increases in genetic diversity in the macaque compared to human.

Within coding regions it is possible to compare diversity that is more likely to be functionally significant with diversity that is more likely to be selectively neutral. In coding regions, both human and macaque had approximately 2 times more diversity for synonymous cSNPs as compared to nsSNPs (Fig. 2D), reflecting functional constraint and selection against changes in the protein sequence [25]. Concerning the possible functional significance of nsSNPs, Polyphen predicted that some 1,741 (29.6%) of the cSNPs we detected in the human and 1,525 (20.7%) of the cSNPs we detected in macaque were likely to be "damaging". The macaque cSNPs we identified include a substantial resource of putatively functional sequence variants. Supporting the functional significance of many of these SNPs, individual humans and macaques were both half as likely to be homozygous for "damaging" nsSNPs than they were to be homozygous for synonymous cSNPs and nsSNPs scored as "benign" by Polyphen (Fig. 3).

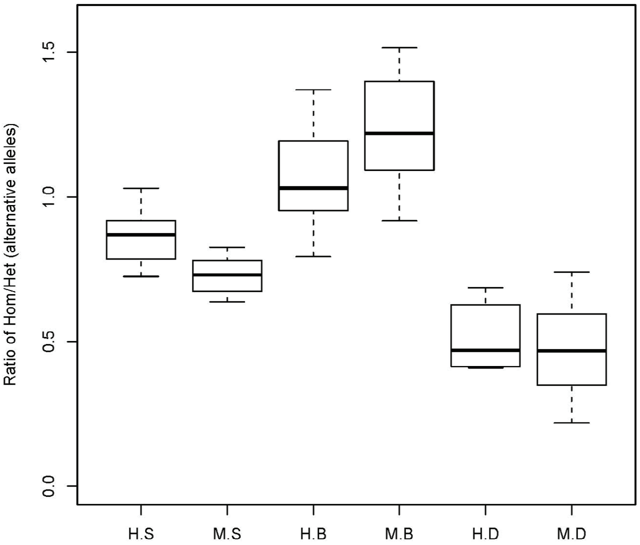

Fig. 3: The ratio of homozygous/heterozygous for the alternative alleles of cSNPs. H.S: human synonymous cSNPs; M.S: macaque synonymous cSNPs; H.B: human nsSNPs with "benign" prediction by PolyPhen; M.B: macaque nsSNPs with "benign" prediction by PolyPhen; H.D: human nsSNPs with "damaging" prediction by PolyPhen; M.D: macaque nsSNPs with "damaging" prediction by PolyPhen.

In line with the theory that most of the increased diversity of the rhesus macaque is selectively neutral in nature, the increase in macaque SNP density was not proportionately maintained from non-coding sequence to coding sequence, to nsSNPs and to putatively "damaging" nsSNPs. Instead, the macaque more closely resembled the human in its SNP density within these more functionally significant categories. Surprisingly, a different picture was observed using the diversity measure θ_{SNP} for human and macaque. By this standard, macaque was approximately three times as diverse as the human across all types of sequence categories. This could point to the maintenance of nsSNPs by balancing selection. This is an important mechanism of evolutionary adaptation in all genetically diverse species but may be operative at a larger percentage of loci in the macaque than in the human. Speculatively, although the macaque does not have proportionately more nsSNPs, those that it does have are more likely to be maintained at higher frequency by balancing selection. However, although this would explain why nsSNP density does not increase proportionately with overall SNP density and with diversity, other validating data would be required to establish this point. One indirect test would be linkage disequilibrium analysis that could detect signals of selection (selective sweeps) at genes containing nsSNPs. In fact, one use of the SNPs we have discovered would be the creation of a marker panel enabling genome wide evaluation of LD. When that is done, the results may again be surprising.

At equilibrium, LD depends on the recombination rate and effective population size. Therefore, it might be anticipated that LD blocks in the rhesus macaque will be substantially smaller than the human. Thus a macaque SNP panel effective for genome-wide use might have to be larger than human 1M panels that are now the standard. However, it is also possible that cross-population admixture has already occurred in the rhesus macaque, at least in some samples of macaques, which could have led to the presence of much larger haplotype blocks than anticipated on the basis of population size. In this same vein, cross-population comparisons of genetic variation would be valuable. The macaques analyzed here are primarily of Indian origin, but as described earlier the species is widely dispersed. In particular there is a very large population of Chinese macaques with several Chinese subspecies proposed including a subspecies representing the island of Hainan, and several unique island-based colonies including Cayo Santiago, Puerto Rico, and Morgan Island, South Carolina. The similarity of diversity of the one Chinese and Chinese/Indian admixed macaques we studied does not address whether there are significant differences at the haplotype level, and based on the analysis of these several animals we have not developed a panel of markers informative for Chinese origin. That might also require the analysis of multiple Chinese populations. Because of their population sizes and breeding structures, macaque and human founder populations, both of which are available, offer an opportunity to observe the changing impact of population dynamics on genetic diversity of different types.

There is some evidence that the mutation rate may have slowed in the hominoid ape lineage, but based on the nucleotide diversity rates we have observed we can compare the effective population sizes of rhesus macaque and human. For this purpose, we used Watterson's (1975) [27] estimator $\theta = 4Neu$ with average nucleotide diversity (θ_{SNP}) in intergenic regions (Fig. 2C) as θ because intergenic diversity is most likely to faithfully reflect neutral diversity at the whole genome level. Assuming an average mutation rate of 1×10^{-8} to 2.5×10^{-8} mutations per nucleotide site per diploid genome per generation for human [18,28-30] and an average mutation rate 5.9×10^{-9} mutations per nucleotide site per diploid genome per generation for macaque [14], the effective population size of humans is approximately 6,780-16,950 and the effective population size of the macaque is approximately 80,000. The most relevant comparison remains the diversity ratio between the human and macaque, with the macaque emerging as having an effective population size several times larger.

As mentioned, our findings on the relative diversity of Chinese and Indian macaques were limited because we studied only one individual animal of Chinese origin and two that were admixed. Furthermore, the specific geographic origin of this one Chinese macaque, and the admixture component of the two other macaques, was unknown. That could be relevant, because the mitochondrial diversity of rhesus macaques from one Western Chinese population appeared to be equivalent to Indian macaques [9], which displayed lower mitochondrial diversity than several other macaque populations. However, it should be noted that the genetic diversity of nuclear DNA is less sensitive to the effects of population bottlenecks than is the diversity of the mitochondrial genome or the haploid Y chromosome. For example, a Finnish bottleneck that left a strong imprint on Y chromosome diversity led to no reduction in autosomal diversity [31]. Recently, Kanthaswamy et al revealed that Chinese and Indian macaques appeared to have near identical genetic diversity based on genotype analysis with more than 1,000 SNPs [11]. Regardless of whether there was a population bottleneck in the rhesus macaque population of India, the Indian macaques that we studied are several times as diverse as the human. Perhaps this is due to subsequent gene flow from other populations which would have restored nuclear DNA diversity of the species on the Indian subcontinent. Considering the geographic origin of the macaques we studied, it is clear that rhesus macaque is several times as diverse compared to the human, but with indications that selection has dampened the increase in functional diversity in this species.

4 Funding

This work was supported by the Intramural Research Program of the National Institute on Alcohol Abuse and Alcoholism, National Institutes of Health.

5 References

[1] Goodman, M., Porter, C.A., Czelusniak, J., Page, S.L., Schneider, H., Shoshani, J., Gunnell, G. and Groves, C.P. (1998) Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol*, **9**, 585-598.

[2] Barr, C.S. and Goldman, D. (2006) Non-human primate models of inheritance vulnerability to alcohol use disorders. *Addict Biol*, **11**, 374-385.

[3] Rhesus Macaque Genome Sequencing and Analysis Consortium. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**, 222-234.

[4] Templeton, A. (2002) Out of Africa again and again. *Nature*, **416**, 45-51.

[5] Thomlinson, R. (1975) *Demographic Problems: Controversy over population control*. 2nd ed. Dickenson Publishing Company, Ecino, CA.

[6] Tenesa, A., Navarro, P., Hayes, B.J., Duffy, D.L., Clarke, G.M., Goddard, M.E. and Visscher, P.M. (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Res*, **17**, 520-526.

[7] Zhang, R., Zhao, T., Quan, Q. and Southwick, C.H. (1991) Distribution of macaques (*Macaca*) in China. *Acta Theriologica Sinica* **11**, 171-185.

[8] Abegg, C. and Thierry, B. (2002) Macaque evolution and dispersal in insular south-east Asia. *Biological Journal of the Linnean Society*, **75**, 555-576.

[9] Smith, D.G. and McDonough, J. (2005) Mitochondrial DNA variation in Chinese and Indian Rhesus Macaques (*Macaca mulatta*). *American Journal of Primatology*, **65**, 1-25.

[10] Ferguson, B., Street, S.L., Wright, H., Pearson, C., Jia, Y., Thompson, S.L., Allibone, P., Dubay, C.J., Spindel, E. and Norgren, R.B. (2007) Single nucleotide polymorphisms (SNPs) distinguish Indian-origin and Chinese-origin rhesus macaques (Macaca mulatta). *Bmc Genomics*, **8**, -.

[11] Kanthaswamy, S., Satkoski, J., Kou, A., Malladi, V. and Smith, D.G. (2010) Detecting signatures of inter-regional and inter-specific hybridization among the Chinese rhesus macaque specific pathogen-free (SPF) population using single nucleotide polymorphic (SNP) markers. *Journal of Medical Primatology*, **39**, 252-265.

[12] Malhi, R.S., Sickler, B., Lin, D., Satkoski, J., Tito, R.Y., George, D., Kanthaswamy, S. and Smith, D.G. (2007) MamuSNP: a resource for Rhesus Macaque (*Macaca mulatta*) genomics. *PLoS One*, **2**, e438.

[13] Magness, C.L., Fellin, P.C., Thomas, M.J., Korth, M.J., Agy, M.B., Proll, S.C., Fitzgibbon, M., Scherer, C.A., Miner, D.G., Katze, M.G. *et al.* (2005) Analysis of the *Macaca mulatta* transcriptome and the sequence divergence between *Macaca* and human. *Genome Biology*, **6**, R60.

[14] Hernandez, R.D., Hubisz, M.J., Wheeler, D.A., Smith, D.G., Ferguson, B., Rogers, J., Nazareth, L., Indap, A., Bourquin, T., McPherson, J. *et al.* (2007) Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. *Science*, **316**, 240-243.

[15] Bevilacqua, L., Doly, S., Kaprio, J., Yuan, Q., Tikkanen, R., Paunio, T., Zhou, Z., Wedenoja, J., Maroteaux, L., Diaz, S. *et al.* (2010) A population-specific HTR2B stop codon predisposes to severe impulsivity. *Nature*, **468**, 1061-1066.

[16] Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat Methods*, **7**, 248-249.

[17] Xu, A.G., He, L., Li, Z., Xu, Y., Li, M., Fu, X., Yan, Z., Yuan, Y., Menzel, C., Li, N. *et al.* (2010) Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq. *PLoS Comput Biol*, **6**, e1000843.

[18] The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.

[19] Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D.A. *et al.* (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med*, **362**, 1181-1191.

[20] Wilke, C.O. (2004) Molecular clock in neutral protein evolution. *BMC Genet*, **5**, 25.

[21] Eyre-Walker, A. and Keightley, P.D. (1999) High genomic deleterious mutation rates in hominids. *Nature*, **397**, 344-347.

[22] Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol*, **5**, e254.

[23] Bhangale, T.R., Rieder, M.J., Livingston, R.J. and Nickerson, D.A. (2005) Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum Mol Genet*, **14**, 59-69.

[24] Livingston, R.J., von Niederhausern, A., Jegga, A.G., Crawford, D.C., Carlson, C.S., Rieder, M.J., Gowrisankar, S., Aronow, B.J., Weiss, R.B. and Nickerson, D.A. (2004) Pattern of sequence variation across 213 environmental response genes. *Genome Res*, **14**, 1821-1831.

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

[26] Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet*, **22**, 231-238.

[27] Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256-276.

[28] Nachman, M.W. and Crowell, S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297-304.

[29] Kondrashov, A.S. (2003) Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Human Mutation*, **21**, 12-27.

[30] Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M. *et al.* (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, **328**, 636-639.

[31] Kittles, R.A., Perola, M., Peltonen, L., Bergen, A.W., Aragon, R.A., Virkkunen, M., Linnoila, M., Goldman, D. and Long, J.C. (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. *Am J Hum Genet*, **62**, 1171-1179.

6 Supporting Materials

Table S1. Ancestry assay for rhesus macaque samples used in this study

SampleID	%Missing Data	Inferred cluster		90% Probability Interval	
		Chinese	India	Chinese	Inida
E13	0	0.006	0.994	(0.000,0.044)	(0.956,1.000)
E20	0	0.007	0.993	(0.000,0.046)	(0.954,1.000)
E78	0	0.005	0.995	(0.000,0.034)	(0.966,1.000)
E85	0	0.524	0.476	(0.392,0.653)	(0.347,0.608)
G14	0	0.013	0.987	(0.000,0.087)	(0.913,1.000)
G36	0	0.531	0.469	(0.395,0.664)	(0.336,0.605)
K09	0	0.004	0.996	(0.000,0.027)	(0.973,1.000)
K20	0	0.999	0.001	(0.991,1.000)	(0.000,0.009)
K29	0	0.003	0.997	(0.000,0.018)	(0.982,1.000)
K41	0	0.008	0.992	(0.000,0.052)	(0.948,1.000)
K44	0	0.006	0.994	(0.000,0.043)	(0.957,1.000)
K46	0	0.005	0.995	(0.000,0.034)	(0.966,1.000)
M18	0	0.019	0.981	(0.000,0.102)	(0.898,1.000)
M40	0	0.004	0.996	(0.000,0.027)	(0.973,1.000)

Table S2. Putative SNPs covered with sequence reads from ChIPseq and/or RNAseq

	Human SNPs having reads from			Macaque SNPs having reads from		
	ChIPseq	**RNAseq**	**Both**	**ChIPseq**	**RNAseq**	**Both**
Total	130914	35615	63499	386219	23109	53474
Intergenic	80482	11594	15385	240655	9457	19278
5Kbupstream	7404	131	2501	22057	403	3843
Intron	39458	8736	31681	113994	2267	14182
Exon	3570	15154	13932	9513	10982	16171
UTR	1966	9149	7317	4249	4815	6391
CDS	1604	6005	6615	5264	6167	9780
nsSNP	814	2513	2550	2453	1864	3050
synonymous	787	3485	4057	2781	4299	6718
nonsense	3	7	8	30	4	12

Table S3. SNPs from 1000 Genomes Project Pilot 2

	CEU.trio	YRI.trio
Total	3646764	4502439
Also in dbSNP	3239544(88.8%)	3446643(76.6%)
Located in intergenic	2131596(58.4%)	2599459(57.7%)
Located in 5Kb upstream	165384(4.5%)	210006(4.7%)
Located in intron	1336273(36.6%)	1674001(37.2%)
Located in UTR	31887(0.9%)	41315(0.9%)
Located in CDS	24192(0.7%)	32244(0.7%)
nsSNP	9696(40.1%)	12853(39.9%)
Synonymous	14506(60.0%)	19412(60.2%)

Table S4. The average nucleotide heterozygosity from 1000 Genomes Project Pilot 2

Population	SampleID	θ_{SNP} (x10^{-4})*
CEU	NA12891	7.16
CEU	NA12892	7.33
YRI	NA19239	9.26
YRI	NA19238	9.06

* Calculated as the heterozygous bases divided by the sequencing accessible genome size (2.85 x 10^9 x 80%) using 1000 Genomes Project Pilot 2 data.

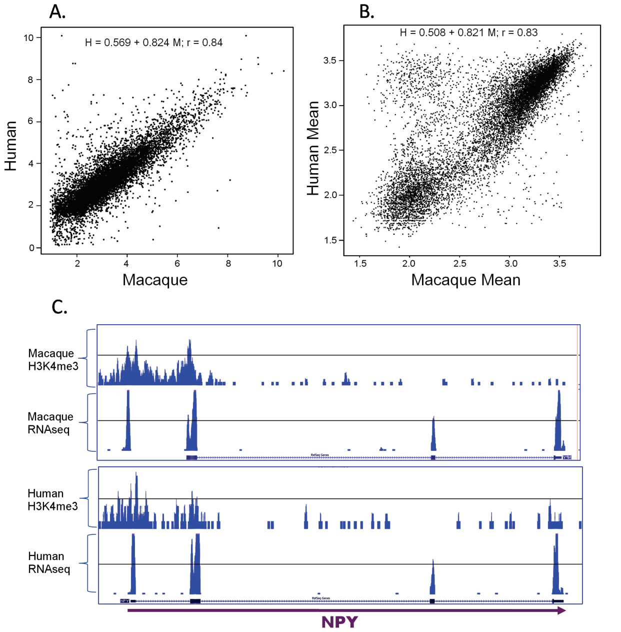

Fig. S1: A: RNAseq correlation between human vs macaque. Data points: mean of normalized gene expression level (log2). B: H3K4me3 ChIPseq correlation between human vs macaque. Data points mean of normalized area under curve (log10) of covered reads within 1Kb of TSS. C. Sequencing coverage (H3K4me3 ChIPseq and RNAseq) in NPY genic region.

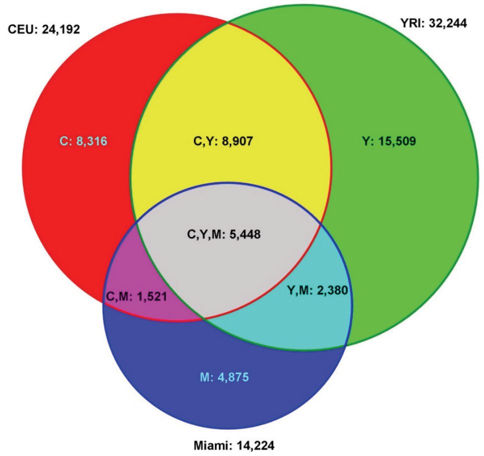

Fig. S2: Human cSNPs identified in 1000 Genomes Project Pilot 2 samples and this study. C or CEU: CEU trio from 1000 Genomes Project Pilot 2; Y or YRI: YRI trio from 1000 Genomes Project Pilot 2; M or Miami: 14 samples from Miami dataset in this study.

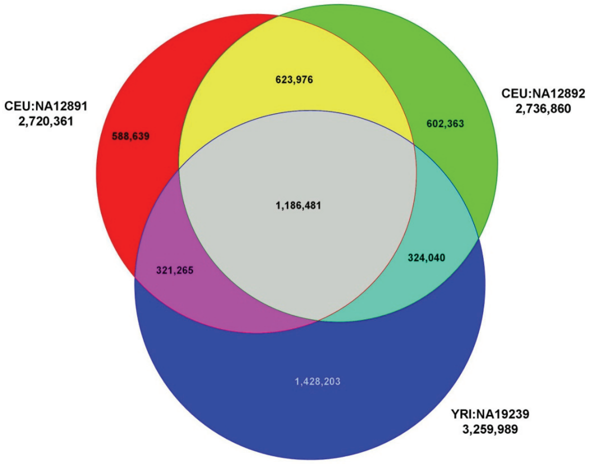

Fig. S3: SNPs shared between individuals in 1000 Genomes Project Pilot 2.