Genomics 98 (2011) 352–358

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

4040 SNPs for genomic analysis in the rhesus macaque (*Macaca mulatta*)

J. Satkoski Trask [a,b,*], W.T. Garnica [a,b], S. Kanthaswamy [a,b,c], R.S. Malhi [d,e], D.G. Smith [a,b]

[a] Department of Anthropology, 330 Young Hall, One Shields Avenue, University of California, Davis, CA, 95616, USA
[b] California National Primate Research Center, One Shields Avenue, University of California, Davis, CA, 95616, USA
[c] Department of Environmental Toxicology, 4138 Meyer Hall, One Shields Avenue, University of California, Davis, CA, 95616, USA
[d] Department of Anthropology, 109 Davenport Hall, 607 S. Matthews Avenue, University of Illinois, Urbana-Champaign, IL, 61801, USA
[e] Institute for Genomic Biology, 1206 West Gregory Drive, University of Illinois, Urbana-Champaign, IL, 61801, USA

ARTICLE INFO

Article history:
Received 26 May 2011
Accepted 17 August 2011
Available online 31 August 2011

Keywords:
Macaca mulatta
Nonhuman primate
SNP
Linkage disequilibrium

ABSTRACT

Although the rhesus macaque (*Macaca mulatta*) is commonly used for biomedical research and becoming a preferred model for translational medicine, quantification of genome-wide variation has been slow to follow the publication of the genome in 2007. Here we report the properties of 4040 single nucleotide polymorphisms discovered and validated in Chinese and Indian rhesus macaques from captive breeding colonies in the United States. Frequency-matched measures of linkage disequilibrium were much greater in the Indian sample. Although the majority of polymorphisms were shared between the two populations, rare alleles were over twice as common in the Chinese sample. Indian rhesus had higher rates of heterozygosity, as well as previously undetected substructure, potentially due to admixture from Burma in wild populations and demographic events post-captivity.

1. Introduction

Genomic characterization of the rhesus macaque (*Macaca mulatta*) has been underway since a complete genome sequence was published in 2007 [1], and is now progressing to a massively parallel scale. The National Center for Research Resources (NCRR) at the NIH currently sponsors six rhesus macaque Working Groups (WGs), one of which, the Genetics and Genomics WG, has assumed the stated goal of instituting uniform SNP-based genetic characterization protocols for parentage testing, ancestry determination and population genetic assessments [2]. In addition, several research groups are in the process of quantifying genomic variation in captive populations of rhesus macaques [3–5]. The data presented here were collected as part of a larger project to discover, quantify and validate single nucleotide polymorphisms (SNPs) in regional populations of rhesus macaques included in captive breeding populations in the United States.

The rhesus macaque is an underused model in biomedicine [6], given that most captive colonies maintain extended multi-generational pedigrees, collect detailed phenotypic data and curate extensive veterinary records, making it suitable for many studies requiring both familial data and/or quantified phenotypic or genetic variance in a complex mammalian system. Approximately six million mice are involved in biomedical research in the United States [7] while the number of nonhuman primates is less than 1/100th of that number [8]. Rhesus macaques specifically comprise even less than that proportion, although they are the most widely used non-human primate model for biomedical research.

Rhesus macaques have been shown to be an effective and adaptable research model, with applications for multi-factorial diseases, including endometriosis, type 2 diabetes and asthma [6] as well as for determining the genetic basis of behavior and gene by environment interactions [9]. Rhesus macaques have demonstrated success as a translational model, especially in testing delivery methods for gene therapies, for example, Duchenne muscular dystrophy [10], artherosclerosis [11], muscular degeneration [12] and L-Dopa-induced dyskinesia [13]. In addition, recent translational studies have demonstrated the efficacy of microbicide cells containing anti-retroviral drugs in mucosally challenged macaques both vaginally [14] and rectally [15], demonstrating the impact of macaque research on the development of cost-effective strategies to prevent HIV transmission in humans.

The most common area of research utilizing rhesus macaques is microbiology, including HIV/AIDS [16]. Rhesus macaques have been heavily used in HIV drug and vaccine initiatives, especially the STEP vaccine trial, which used a non-replicating recombinant adenovirus 5 (rAD5) vector to stimulate T-cells. Phase 2b trials sponsored by Merck and the National Institutes of Health (NIH) were terminated when it was determined that the vaccine did not provide protection against infection, nor did it reduce viral loads post-infection [17]. Haigwood [18] noted that even in the non-human primate trials, viral load was reduced by several orders of magnitude in animals infected with the chimeric human/simian virus SHIV-89.6P but not when infected with the more virulent SIV_{mac239} and the progression to human trials was overly optimistic. The difference in immune

* Corresponding author at: Department of Anthropology, 330 Young Hall, One Shields Avenue, University of California, USA. Fax: +1 530 752 8885.
E-mail address: jasatkoski@ucdavis.edu (J.S. Trask).

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.08.004

response underlines the point that successful application of animal models will only come from a more sophisticated understanding of host genetics. Unlike *Drosophila*, *Caenorhabditis elegans* or the mouse models, there is no central curated resource for genomic and phenotypic information (flybase.org, wormbase.org and www.informatics.jax.org, respectively) on the rhesus macaque model. The development of such a resource would encourage collaboration between biomedical and genetic research, allowing greater integration of phenotypic, immunological and genomic information.

In this study we focus on curating genomic information and report SNP discovery and validation in rhesus macaques. As we develop rhesus macaques as a research resource we will expand available information to include identification and quantification of copy number variants, location of population-specific genomic rearrangements, and other genome-level factors known to influence phenotype. The greater availability of these data will make rhesus macaques an even more attractive research model for genetic epidemiology, multi-factorial disease and translational medicine.

2. Methods

Our method of SNP discovery is described in detail in Malhi et al. [19]. A DNA sample from a female rhesus macaque of western Chinese origin was submitted to 454 Life Sciences (Roche Diagnostics, Branford, CT) for large-scale parallel pyrosequencing, producing a total of 339,967 reads with an average read length of 104 bp. The reads were aligned against the published rhesus genome version 1.1 [1], known to be derived from an Indian-origin animal. This alignment identified approximately 23,000 prospective polymorphisms.

Malhi et al. [19] described the discovery of approximately 23,000 candidate SNPs distributed throughout the rhesus macaque genome. Our goal was to select and validate markers distributed approximately 1 Mb apart from this pool of candidates. However, the median distance between adjacent candidate polymorphisms is only 65 kb (mean = 125 kb ± 223 kb), indicating that the majority of candidate markers were far too close together for the construction of an equidistantly spaced SNP map and thus the actual number of suitable markers for such a map was much smaller than 23,000. Accordingly, 8342 of these candidate SNPs were selected for validation by identifying the most proximal polymorphism on each chromosome and polymorphisms spaced approximately 1 Mb apart across the entire sequence. When probes for the polymorphisms were not designable on the Illumina GoldenGate™ platform, failed to amplify during the genotyping reaction or did not show any segregating polymorphisms in the genotyped individuals, the nearest verifiable polymorphism, either upstream or downstream, was included instead.

Quality-screening of the candidate SNPs is described in Satkoski et al. [4]. Polymorphic locations with pyrofragment Phred scores less than 20 and only a single overlapping fragment were discarded. For the remaining putative polymorphisms, the chromosome and nucleotide position of each fragment containing a candidate SNP within the rhesus genome was confirmed with the genome BLAST [20] function of the National Center for Biotechnology Information website (NCBI, www.ncbi.nlm.nih.gov). Fragments that were confirmed as single copy and produced a high-quality (+98%) match to the rhesus genome were selected for further analysis. Fifty-two of the 8342 candidate markers selected for validation produced no BLAST matches and 3494 produced multiple BLAST hits, suggesting that the sequence flanking the polymorphism is repetitive or exists in multiple copies, leaving 4796 SNPs for validation.

We employed Illumina (San Diego, CA) GoldenGate technology to genotype the resulting candidate markers. Of the candidates, 125 could not be incorporated into the Illumina oligo pool (OPA), resulting in 4671 markers submitted for validation. These markers were genotyped on both the BeadXPress (with one 96-plex OPA and four 384-plex OPAs) and the iScan platforms (with two 1536-plex OPAs). The individuals selected for genotyping were, to the best of our knowledge, not first or second degree relatives; sample information is shown in Table 1. These animals were either imported directly from the country of origin (VBS, TSS) or had sufficient colony documentation to support their assignment to the appropriate region (CPRC, UM, ONPRC and CNPRC). In addition, all individuals had been sequenced for a 830-bp section of the mitochondrial genome [21] and 24 nuclear microsatellite loci [22], used to confirm their assignment to a specific geographic region (partial data presented in Satkoski Trask et al. [23], additional data not shown). Approximately 5% of the markers and at least two individuals were chosen at random and duplicated across each OPA and each run as controls. Of the genotyped markers, 365 did not meet the minimum quality (GenTrain) score of 0.4 and were excluded from further analysis. An additional 266 markers exceeded the maximum 5% missing genotype criterion and were also excluded, leaving a total of 4040 validated SNPs.

Minor allele frequencies (MAF) and observed heterozygosities were calculated with PLINK 1.06 (http://pngu.mgh.harvard.edu/purcell/plink, [24]). Principal component analysis (PCA) was performed with the adegenet 1.2-8 package for R [25] to identify genetic structure within the data independent of a priori assignment to a particular geographic origin or breeding center. To identify which markers were located within genes, SNPs were localized relative to known genes using the RefSeq Genes track for the MGSC Merged 1.0/rheMac2 assembly of the rhesus macaque genome in the UCSC Genome Browser (http://genome.ucsc.edu [26,27]) through the Galaxy web interface [28,29].

Once validated, information on each polymorphism was submitted to the dbSNP online database (http://www.ncbi.nlm.nih.gov/projects/SNP). Information on each SNP described herein, including chromosome and nucleotide position, dbSNP ss#, flanking sequence and MAF values from the Indian and Chinese samples can be found at http://primate.bioinformatics.ucdavis.edu, a custom UCSC Genome Browser instance.

Linkage disequilibrium (LD) was measured as r^2, the correlation coefficient between the allele frequencies of the two markers [30], and calculated with Haploview [31]. Only pairwise LD calculations with non-zero T-int values were considered; T-int is a statistic used by the HapMap project that measures the completeness of information provided by a set of markers in a genomic region [31]. For both the Indian and Chinese samples, markers were sorted into MAF bins of 0.1, 0.2, 0.3, 0.4 and 0.5. The r^2 values were filtered to include only markers in the same MAF bin to create frequency-matched pairs [32]. By comparing the positions of the markers in the frequency-matched pairs to the MGSC Merged 1.0/rheMac2 version of the rhesus genome on the UCSC genome browser (http://genome.ucsc.edu), it was determined that all the frequency-matched pairs were located outside of known genes (pairs located within genes had been removed during frequency-matching). Hernandez et al. [33] used SNPs located in ENCODE

Table 1
Geographic origin and mtDNA haplotype of the Indian and Chinese samples. *CPRC*, Caribbean Primate Research Center; *UM*, University of Miami; *ONPRC*, Oregon National Primate Research Center; *VBS*, Valley Biosystems; *TSS*, Three Springs Scientific; *CNPRC*, California National Primate Research Center; *COV*, Covance Inc., Alice TX.

Geographic origin	mtDNA haplotype	Sample size	Source	Sex ratio
India				
Uttar Pradesh, Central	Ind1	20	CPRC	M = 24%
Kashmir	Ind1	2	UM	F = 56%
Central India	Ind2	3	ONPRC	U = 20%
China				
Guangdong	ChiE	1	VBS	
Sichuan	ChiW	11	VBS, TSS	M = 16%
Suzhou	ChiW (N = 4)	12	CNPRC	F = 48%
	ChiE (N = 8)			U = 36%
Unknown	ChiS	1	COV	

regions to estimate that LD in rhesus macaques decayed completely by 50 kb; we considered all frequency-matched pairs regardless of distance to ensure that we captured the point at which LD reached zero.

3. Results

Of the 8342 markers chosen for validation, 4040 (40%) met all the QC and genotype standards. These SNPs are distributed throughout the genome, located on all 20 autosomes and the X chromosome. We have validated one SNP approximately every 723 kb (±142 kb) on the autosomes, and one SNP approximately every 1.9 Mb on the X chromosome. No SNPs were discovered on the Y chromosome due to the lack of a published Y chromosome sequence. The number and spacing of SNPs on each chromosome is shown in Table 2. Of the 4040 validated SNPs, 80 (2%) were located in 76 different genes.

Of the 4040 markers, 2760 (68%) were polymorphic in both the Indian and Chinese samples. Thirteen percent of the markers were polymorphic only in the Chinese sample, compared to 19% of the markers polymorphic only in the Indian sample. The MAF distributions for both the Indian and Chinese samples are shown in Fig. 1. The average MAF for the Indian sample was 0.19, compared with 0.17 in the Chinese sample. This difference is statistically significant ($p<0.000001$, two-sample t-test, unequal variances). Not only are the average MAF value and proportion of population-specific SNPs higher in the Indian sample, but, as shown in Fig. 1, many more markers are of high frequency in the Indian sample with a corresponding low frequency in the Chinese sample than the reverse. Of the 3023 markers polymorphic in Indian animals, 308 (10.2%) were below 5% minor allele frequency, compared with 767 (23.0%) of the 3336 markers polymorphic in Chinese individuals. Of the markers polymorphic in both populations, only 65 (2.4%) had a minor allele frequency below 5%. Differences in observed heterozygosity are shown in Fig. 2. The average heterozygosity for the Chinese sample was 0.22, while the average heterozygosity in the Indian sample was 0.24. This difference was statistically significant ($p<0.0001$, two-sample t-test for equal variances).

The r^2 values for the two samples are shown plotted against distance in Fig. 3. Six hundred and twenty-five marker pairs met the stated criteria in the Indian sample, compared with 724 marker pairs in the Chinese sample. We fit logarithmic, exponential and linear models to the data and found that the logarithmic model had the greatest predictive power. Using the appropriate equation for each population, we calculated the slope and x intercept of the logarithmic regression line, which allowed us to estimate the value of r^2 at 10 kb ($x=10$) as well as the point at which LD dissipates completely (x intercept). For the Indian sample, r^2 at 10 kb was 0.54, compared to 0.30 for the Chinese sample. LD was predicted to reach 0 at 582.55 kb and 1.81 Mb, respectively.

The results of the PCA analyses are illustrated in Figs. 4 and 5. PC1 represents 24.7% of the total sample variance, and PC2 and 3 represent 3.2% and 2.9%, respectively. As shown in Fig. 4, the Chinese and Indian samples are sorted cleanly, with the exception of individual 22375. This individual was classified as Indian and determined to have the Ind2 mitochondrial haplotype that is relatively rare (5%) in Indian rhesus macaques yet fixed in Burmese [21] and Bangladesh [34] rhesus macaques. Given that there appears to be structure in the Indian sample along the PC_2 axis, we plotted PC_2 against PC_3 in Fig. 5. These individuals form at least three, possibly four, distinct clusters unrelated to mitochondrial haplotype or geography: one cluster consisting solely of Ind1 animals from the CPRC, one cluster containing CPRC animals and Ind2 individual 22375 and one containing both Ind1 and Ind2 animals from the other sample sources.

Table 2
The number of markers on each chromosome and the average distance between markers.

Rhesus chromosome	Number of SNPs	Mean gap (kb)	Median gap (kb)
1	344	664.57	511.01
2	315	600.81	494.68
3	288	681.46	538.90
4	258	647.70	509.50
5	280	651.23	542.89
6	272	648.67	522.97
7	273	618.72	543.87
8	214	688.63	593.58
9	201	662.12	484.88
10	138	685.66	550.93
11	207	652.38	543.18
12	154	677.20	551.21
13	214	636.28	497.95
14	151	881.63	633.93
15	158	695.24	625.23
16	101	784.60	789.43
17	152	618.62	531.11
18	81	894.22	758.78
19	53	1194.82	769.40
20	101	871.52	571.50
X	82	1862.98	1371.66
		Grand mean	Grand median
		777.10	543.87

4. Discussion

Of the SNPs chosen for validation, 42% were rejected due to their location within duplicated or repetitive regions, as reported by a BLAST search against the rhesus macaque reference genome. The human genome is composed of approximately 50% repetitive sequences [35], and the amount of repetitive DNA in the rhesus macaque is comparable [1] although the proportion of this sequence attributable to segmental duplication (2.3%) is substantially lower than the comparable values for humans and chimpanzees. Therefore, the percentage of markers rejected due to their location in a repetitive or duplicated region is reasonable. This high rate of failure for marker validation seems to be an innate quality of the primate genome [35] rather than an issue with the validation method, and should be taken into account when researchers seek to estimate the number of usable markers resulting from any SNP discovery effort.

Previously undetected flanking polymorphisms in the Indian and Chinese samples or repetitive DNA sequence not reflected in the published genome are potential explanations for the 365 markers (an additional 4% of the markers originally selected for validation) that failed to meet the GenTrain quality threshold during Illumina GoldenGate genotyping. A much smaller proportion of markers chosen for validation (1%) failed during the probe-design process. A probe design failure (as opposed to a polymorphic site that is designable but flagged with a warning) is likely due to low sequence complexity in the flanking region (Illumina Technical Note: Designing Custom GoldenGate® Assays). These failures highlight the importance of genomic completeness and adequate genome annotation during SNP map construction. The available annotation of the rhesus macaque genome has not been updated since its release in 2006 [36]. Although, as demonstrated here, it is possible to identify sequence regions inappropriate for SNP genotyping through genotype failure or low quality scores, bioinformatic screening of markers prior to wet lab validation is much faster and massively more cost effective.

The largest collection of Indian and Chinese rhesus macaque SNPs was previously published by Hernandez et al. [33]. Their sample included nine individuals from China (seven from Suzhou, one from Kunming and one from Guangdong) and thirty-eight individuals of Indian origin, the ancestry of most of which could not be attributed to a specific geographic region. Hernandez et al. [33] sequenced five ENCODE regions in 166 non-overlapping windows, resulting in the discovery of 1476 SNPs. In contrast, our study includes 50 individuals, with Chinese and Indian individuals equally represented. Simulations of LD using simulated data have suggested that sufficient sample size is an important consideration when designing a study of linkage disequilibrium, since insufficient

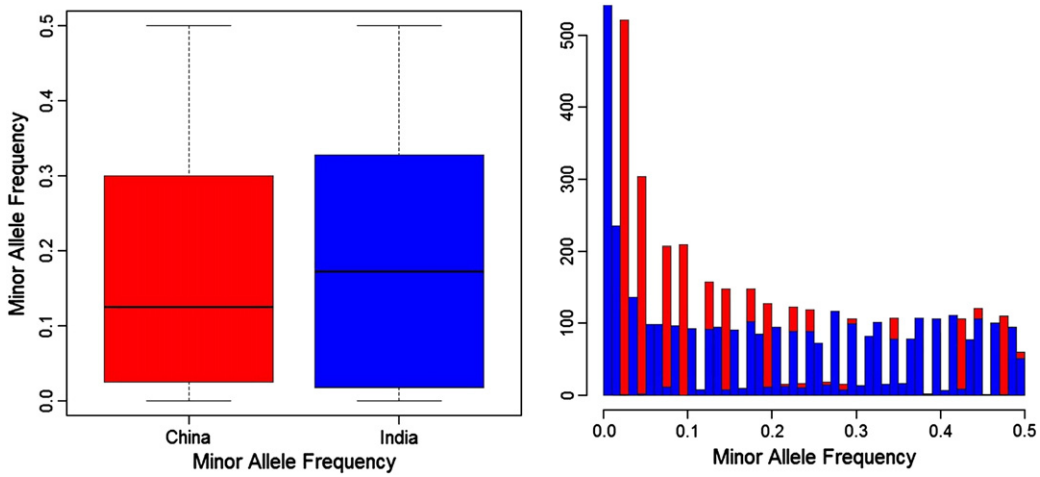

Fig. 1. Distribution of minor allele frequencies in the Chinese and Indian samples. Mean minor allele frequency is significantly different ($p<0.01$).

population representation can fail to capture recombination events and overestimate haplotype block size [37]. Fifty individuals (or 100 chromosomes) are generally considered to be the minimum viable sample size for estimating LD around common alleles, a number confirmed by studies of LD in human populations [38].

Hernandez et al. [33] did not report the individual heterozygosity or MAF values of their reported markers but noted that very few of the SNPs were shared across populations. Only 33% of their markers, compared to 68% in the present study, were shared between both populations, while 61% and 39% were found only in the Chinese and Indian samples, respectively, compared to 13% and 19% in our study. A comparison with the baboon genome (to infer the ancestral state) determined that the Chinese population contained an excess of rare markers under the assumption of constant population size, while the Indian sample contained many SNPs of intermediate and high frequency. A much greater proportion of the markers examined in our study were polymorphic in both the Indian and Chinese samples, and the proportion of markers that were population specific much smaller. This difference is probably largely due to a bias toward the discovery of high MAF, shared SNPs inherent in our discovery method [23]: the comparison of just two individuals, one Indian and one Chinese, led to preferential discovery of high frequency polymorphisms shared between the two populations. Low frequency polymorphisms had a lower probability of discovery, while discovery of SNPs polymorphic only in the Indian or Chinese populations required the sequenced individual to be a heterozygote. This phenomenon is probably exacerbated by the fact that the Indian genome used for alignment represented only a single strand and thus, could not be heterozygotic. Another factor that potentially contributes to this difference is the larger, more geographically variable composition of our Chinese sample. A third likely explanation is the difference in marker location: while the markers described in Hernandez et al.'s study were located entirely in ENCODE regions, only 2% of the markers in our study are located in coding regions and probably less subject to the action of selection on the Indian and Chinese rhesus macaque populations. One result of this study did reproduce the results of Hernandez et al., specifically, the paucity of intermediate and high-frequency alleles in the Chinese, relative to the Indian sample. As illustrated in Fig. 1, while markers with low MAF in the Indian sample also tended to have a low MAF in the Chinese sample, the converse was not true. Rare alleles (MAF equal to or less than 5%) were over twice as common in the Chinese sample, relative to the Indian sample. This pattern is consistent with a genetic bottleneck in Indian (but not Chinese) rhesus macaques that eliminated some of the rare alleles in their common ancestors, as speculated by Hernandez et al. [33].

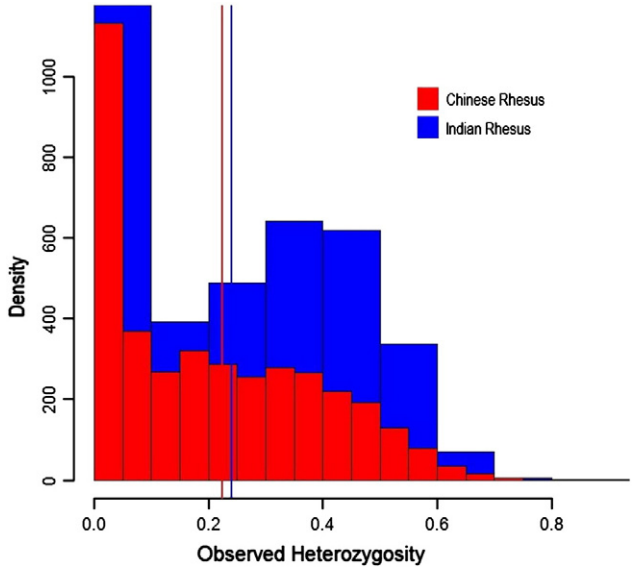

Fig. 2. Histograms of observed heterozygosity in the Chinese and Indian samples. The mean heterozygosities for the two samples are statistically significantly different ($p<0.01$).

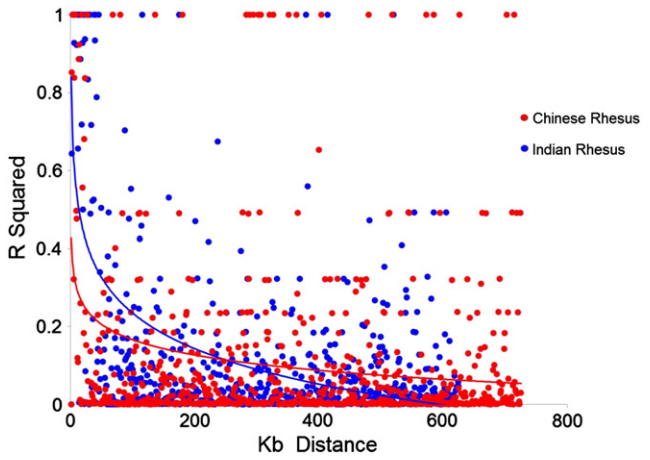

Fig. 3. The decay of linkage disequilibrium in the Indian and Chinese samples. R^2 values are calculated for frequency-matched markers located on the same chromosome. The x intercept of each line marks the complete disintegration of linkage disequilibrium.

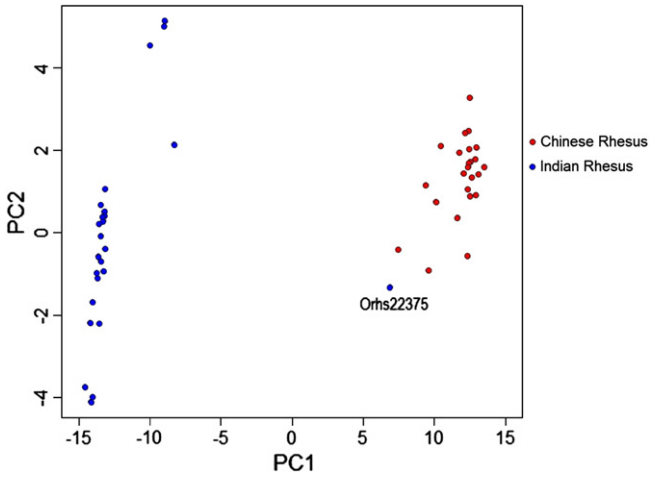

Fig. 4. Principal component analysis of Indian and Chinese rhesus macaques. Principal component (PC) 1 separates Indian from Chinese individuals with the exception of sample Orhs22375.

As shown in Fig. 3, and in agreement with Hernandez et al. [33], linkage disequilibrium dissipates much more rapidly in Indian rhesus macaques compared to their Chinese counterparts. The linkage distance in the Chinese sample was over twice that of the Indian sample, although both were much greater than predicted by Hernandez et al. [33], making it difficult to assess the impact of admixture of the Indian sample with rhesus macaques dispersing from the east carrying the Ind2 mtDNA haplogroup. The LD values reported by Hernandez et al. [33], r^2 at 10 kb of 0.15 and 0.52 for Chinese and Indian rhesus macaques, respectively, were somewhat different from those estimated in the present study (r^2 at 10 kb of 0.30 and 0.54, respectively), although more so for the Chinese animals. This difference may be due to the fact that only nine Chinese individuals were included in the former study, compared to 38 Indian individuals, raising the possibility that Chinese recombinants were identified at a lower rate. Also probably contributing to the difference is the fact that LD in the Hernandez et al. [33] study was calculated from SNPs in ENCODE regions, while the frequency-matched LD values presented here eliminated all comparisons of markers located within the same gene and contained only LD measurements from SNPs in non-coding regions.

Consistent between the two studies was the substantial difference between LD in the Indian and Chinese rhesus macaque samples. The longer linkage distance observed in Indian rhesus could potentially

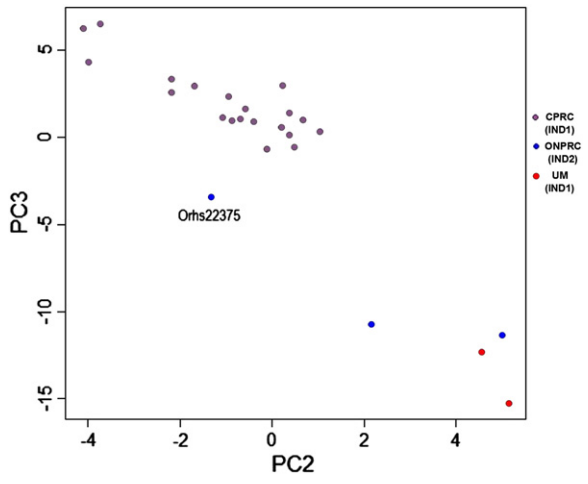

Fig. 5. Differentiation within the Indian samples shown by PC2 and PC3 in the principal component analysis.

be due to nuclear genetic admixture with Burmese rhesus, via the Bramaputra River [21]. Additionally, a selective sweep, or local reduction in genetic variation, can be caused by the rapid fixation of a beneficial mutation, resulting in high LD around the site of this mutation [39]. If the selective sweep happened quickly, local variation will diminish to zero, followed by the re-accumulation of variation through novel mutation and recombination, leading to an overabundance of rare alleles. In human populations, researchers have identified incomplete selective sweeps around the genes for lactase (*LCT*) and glucose-6-phosphate dehydrogenase (*G6PD*), displaying haplotypes that appear to be selection, but have not yet reached 100% frequency. This process produces a pattern of locally identical haplotypes segregating at high frequencies, with the other haplotypes displaying normal variability [40]. Preliminary analysis of LD in geographically and phenotypically variable rhesus macaques have identified regions potentially under selection in Indian, but not Chinese populations, which could contribute to the large difference in linkage distance [41]. Although the measurements of LD presented here and in Hernandez et al. [33] provide insight into the differing evolutionary histories of the Indian and Chinese rhesus populations, a true linkage map will not be possible until these SNPs can be genotyped in extended families [42].

Although the differentiation between the Indian and Chinese samples was quite strong and consistent with previously reported results [4,23], two results unique to the present study are the greater heterozygosity in Indian than in Chinese rhesus macaques and the presence of population structure of unknown source in the Indian sample (Figs. 4 and 5). Previous studies of microsatellite (STR) markers [43], mtDNA sequence [22] and SNPs in the 3′ ends of rhesus macaque genes [3] and ENCODE regions [33] have reported higher levels of heterozygosity in Chinese than in Indian rhesus macaques. While the majority of the Indian-origin animals sampled by Hernandez et al. [33] came from the Yerkes National Primate Research Center, most of the Indian-origin animals in the present study were selected from the Caribbean Primate Research Center (CPRC). Fig. 4 shows that the individuals form several distinct clusters along PC_2, and this continuum is also visible along PC_3, with the CPRC individuals forming two groups, exclusive of the Indian-origin animals from the University of Miami (UM) and the Oregon National Primate Research Center (ONPRC). In contrast, the individuals from the latter two centers are far less differentiated than the individuals in the UM and ONPRC colonies. While individual 22375 appeared to be of Chinese origin, it clusters with other individuals of Indian origin in Fig. 4, consistent with the hypothesis that the IND2 haplotype originated in Burma with these individuals serving as a source of novel alleles in the Indian population. We have previously reported the presence of a historic signal of Chinese admixture among individuals of haplotype IND2 from ONPRC [44] using a smaller set of 829 SNPs.

The early history of the CPRC colony is described in detail by Carpenter [45], Buettner-Janusch et al. [46] and Johnsen [47]. The population was initiated in 1938/1939 with the release of 409 individuals, 14 gibbons and three *Macaca nemestrina* on the island of Cayo Santiago, off the southeastern coast of Puerto Rico. The population in March of 1940 was approximately 350 animals and dropped to 150 prior to 1956 but grew to 791 by 1968. In 1970, when the colony on Cayo Santiago became part of the CPRC, the population was reduced to 333 animals. A genetic study of the transferrin locus by Buettner-Janusch et al. [46] conducted both before and after the 1970 population reduction found that although the allele frequencies had not changed significantly, the total number of transferrin phenotypes had fallen from 15 to 12. Olivier et al. [48] confirmed this result with low Fst values among social groups calculated from the serum protein transferrin and the isozymes carbonic anhydrase II and 6-phosphogluconate dehydrogenase. In contrast, Duggleby [49] found that the phenotypes of the red blood cell systems I, J, K, L, P and Q were significantly heterogeneous over social groups. These early protein polymorphism studies suggest, as do these

results, that although the complicated demographic history of the CPRC has not significantly impacted variation of selectively important loci, the impact on neutral loci, or loci under weak selection, could be profound. Although Indian rhesus within United States breeding colonies are generally considered to be far more genetically homogeneous than their Chinese counterparts (who are often recently imported or the offspring of imported animals), the results of our study call this into question and demonstrate cryptic population structure in Indian rhesus macaques; the importance of this structure on the phenotypic and immunological variance within the United States captive Indian-origin rhesus population is difficult to gage at this point.

5. Conclusions

Forty-two percent of SNPs selected for validation were rejected due to their location in a duplicated or repetitive region, confirming estimates that the amount of repetitive DNA in the rhesus macaque genome is comparable to that in humans. Unlike previously published reports of SNP variability in rhesus macaques [33], the majority of polymorphisms were shared between the Indian and Chinese samples. Rare alleles were over twice as common in Chinese rhesus. Linkage disequilibrium was much stronger in the Indian sample relative to the Chinese sample, potentially due to a combination of admixture with rhesus macaques from Burma and differential selection on Indian populations. Although a paucity of rare alleles in the Indian sample is consistent with the hypothesis of a bottleneck in this population, high heterozygosity and the presence of previously undetected substructure indicates that Indian-origin rhesus macaques in United States breeding centers may contain cryptic genetic variation. The results of this study underline the importance of quantifying genomic variation present in biomedical research models, especially as the rhesus macaque increases in popularity as a translational model.

Acknowledgments

The authors acknowledge funding from the following sources: NIH/NCRR R24RR025871 and NIH/NCRR R24RR005090. The authors would like to thank the staff of the University of California (UC), Davis Molecular Anthropology Laboratory for assistance with DNA extraction and preparation. We would also like to acknowledge the staff and faculty of the UC Davis Genome Center, especially Dr. Dawei Lin and the Bioinformatics core, for insight on data analysis, and Dr. Charles Nicolet and the Genome Technologies Core, for valuable discussion of SNP genotyping.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.ygeno.2011.08.004.

References

[1] R. Gibbs, R.M.G.S.a.A. Consortium, Evolutionary and biomedical insights from the rhesus macaque genome, Science 316 (2007) 222–234.
[2] S. Kanthaswamy, J.P. Capitanio, C.J. Dubay, B. Ferguson, T. Folks, J.C. Ha, C.E. Hotchkiss, Z.P. Johnson, M.G. Katze, L.S. Kean, H. Michael Kubisch, S. Lank, L.A. Lyons, G.M. Miller, J. Nylander, D.H. O'Connor, R.E. Palermo, D.G. Smith, E.J. Vallender, R.W. Wiseman, J. Rogers, Resources for genetic management and genomics research on non-human primates at the National Primate Research Centers (NPRCs), J. Med. Primatol. 38 (2009) 17–23.
[3] B. Ferguson, S.L. Street, H. Wright, C. Pearson, Y. Jia, S.L. Thompson, P. Allibone, C.J. Dubay, E. Spindel, R.B. Norgren, Single nucleotide polymorphisms (SNPs) distinguish Indian-origin and Chinese-origin rhesus macaques (*Macaca mulatta*), BMC Genomics 8 (2007) 43.
[4] J. Satkoski, R.S. Malhi, S. Kanthaswamy, R.Y. Tito, V. Malladi, D.G. Smith, Pyrosequencing as a method for SNP identification in the rhesus macaque (*Macaca mulatta*), BMC Genomics 9 (2008) 256.
[5] G.L. Fawcett, M. Raveendran, D.R. Derios, D. Chen, F. Yu, R.A. Harris, Y. Ren, D. Muzny, J.G. Reid, D.A. Wheeler, K.C. Worley, S.E. Shelton, N.H. Kalin, A. Milosavljevic, R. Gibbs, J. Rogers, Characterization of single-nucleotide variation in Indian-origin Rhesus Macaques (*Macaca mulatta*), BMC Genomics 12 (2011) 311.
[6] R.M. Hadfield, J.G. Pullen, K.F. Davies, S.E. Wolfensohn, J.W. Kemnitz, D.E. Weeks, S.T. Bennett, S.H. Kennedy, Toward developing a genome-wide microsatellite marker set for linkage analysis in the rhesus macaque (*Macaca mulatta*): identification of 76 polymorphic markers, Am. J. Primatol. 54 (2001) 223–231.
[7] L.A. Hart, A. Dassler, Mouse in Science: Why Mice?, in, UC Davis Center for Animal Alternatives.
[8] U.S.D.o. Agriculture, Annual Report Animal Usage by Fiscal Year: Fiscal Year 2009, in: A.a.P.H.I. Service (Ed.), 2011.
[9] C.S. Barr, T.K. Newman, M.L. Becker, C.C. Parker, M. Champoux, K.P. Lesch, D. Goldman, S.J. Suomi, J.D. Higley, The utility of the non-human primate model for studying gene by environment interactions in behavioral research, Genes Brain Behav. 2 (2003) 336–340.
[10] L.R. Rodino-Klapac, P.M.L. Janssen, C.L. Montgomery, B.D. Coley, L.G. Chicoine, K.R. Clark, J.R. Mendell, A translational approach for limb vascular delivery of the micro-distrophin gene without high volume or high pressure for treatment of Duchenne muscular dystrophy, J. Transl. Med. 5 (2007) 45.
[11] E.A. DiBlasio-Smith, M. Arai, E.M. Quinet, M.J. Evans, T. Kornaga, M.D. Basso, L. Chen, I. Feingold, A.R. Halpern, Q.Y. Liu, P. Nambi, D. Savio, S. Wang, W.M. Mounts, J.A. Isler, A.M. Slager, M.E. Burczynski, A.J. Dorner, E.R. LaVallie, Discovery and implementation of transcriptional biomarkers of synthetic LXR agonists in peripheral blood cells, J. Transl. Med. 6 (2008) 59.
[12] J. Kota, C.R. Handy, A.M. Haidet, C.L. Montgomery, A. Eagle, L.R. Rodino-Klapac, D. Tucker, C.J. Shilling, W.R. Therfall, C.M. Walker, S.E. Weisbrode, P.M.L. Janssen, K.R. Clark, Z. Sahenk, J.R. Mendell, B.K. Kaspar, Follistatin gene delivery enhances muscle growth and strength in nonhuman primates, Sci. Transl. Med. 1 (2009) 6ra15.
[13] M.R. Ahmed, A. Berthet, E. Bychkov, G. Porras, Q. Li, B.H. Bioulac, Y.T. Carl, B. Bloch, S. Kook, I. Aubert, S. Dovero, E. Doudnikoff, V.V. Gurevitch, E.V. Gurevitch, E. Bezard, Lentiviral overexpression of GRK6 alleviates L-Dopa-induced dyskinesia in experimental Parkinson's Disease, Sci. Transl. Med. 2 (2010) 28ra28.
[14] U.M. Parikh, C. Dobard, S. Sharma, M. Cong, H. Jia, A. Martin, C.P. Pau, D.L. Hanson, P. Guenthner, J. Smith, E. Kersh, J.C. Garcia-Lerma, F.J. Novembre, R. Otten, T. Folks, W. Heneine, Complete protection from repeated vaginal SHIV exposures in macaques by a topical gel containing tenofovir alone or with emtricitabine, J. Virol. 83 (2009) 10358–10365.
[15] J.G. Garcia-Lerma, M. Cong, J. Mitchell, A.S. Youngpairoj, Q. Zheng, S. Masiotra, A. Martin, Z. Kuklenyik, A. Holder, J. Lipscomb, C.P. Pau, J.R. Barr, D.L. Hanson, R. Otten, L. Paxton, T. Folks, W. Heneine, Intermittent prophylaxis with oral Truvada protects macaques from rectal SHIV infection, Sci. Transl. Med. 2 (2010) 14ra14.
[16] H.E. Carlsson, S.J. Schapiro, J. Hau, Use of primates in research: a global overview, Am. J. Primatol. 63 (2004) 225–237.
[17] D. Barouch, Challenges in the development of an HIV-1 vaccine, Nature 455 (2008) 613–619.
[18] N.L. Haigwood, Update on animal models for HIV research, Eur. J. Immunol. 39 (2009) 1991–2058.
[19] R.S. Malhi, B. Sickler, D. Lin, J. Satkoski, R.Y. Tito, D. George, S. Kanthaswamy, D.G. Smith, *Mamu*SNP: a resource for rhesus macaque (*Macaca mulatta*) genomics, PLoS ONE 2 (2007) e438.
[20] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[21] D.G. Smith, J. McDonough, Mitochondrial DNA variation in Chinese and Indian rhesus macaques (*Macaca mulatta*), Am. J. Primatol. 65 (2005) 1–25.
[22] D.G. Smith, D. George, S. Kanthaswamy, J. McDonough, Identification of country of origin and admixture between Indian and Chinese rhesus macaques, Int. J. Primatol. 27 (2006) 881–898.
[23] J.A. Satkoski Trask, R.S. Malhi, S. Kanthaswamy, J. Johnson, W.T. Garnica, V. Malladi, D.G. Smith, The effect of SNP discovery method and sample size on estimation of population genetic data for Chinese and Indian rhesus macaques (*Macaca mulatta*), Primates 52 (2011) 129–138.
[24] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. Bakker, M.J. Daly, P.C. Sham, PLINK: a toolset for whole-genome association and population-based linkage analysis, Am. J. Hum. Genet. 81 (2007) 559–575.
[25] T. Jombart, Adegenet: a R package for the multivariate analysis of genetic markers, Bioinformatics 24 (2008) 1403–1405.
[26] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006.
[27] P.A. Fujita, B. Rhead, A.S. Zweig, A.S. Hinrichs, D. Karolchik, M.S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho, M. Diekhans, T.R. Dreszer, B.M. Giardine, R.A. Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R.M. Kuhn, K. Learned, C.H. Li, L.R. Meyer, A. Pohl, B.J. Raney, K.R. Rosenbloom, K.E. Smith, D. Haussler, W.J. Kent, The UCSC genome browser database: update 2011, Nucleic Acids Res. 39 (2010) 1–7.
[28] D. Blankenberg, G. Von Kuster, N. Corador, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, J. Taylor, Galaxy: a web-based genome analysis tool for experimentalists, Current Protocols in Molecular Biology, 2010, pp. 1–21.
[29] J. Goecks, A. Nekrutenko, J. Taylor, T.G. Team, Galaxy: a comprehensive approach for supporting accessible, reproducible and transparent computational research in the life sciences, Genome Biol. 11 (2010) R86.
[30] P. Hedrick, S. Kumar, Mutation and linkage disequilibrium in human mtDNA, Eur. J. Hum. Genet. 9 (2001) 969–972.
[31] J.C. Barrett, B. Fry, J. Maller, M.J. Daly, Haploview: analysis and visualization of LD and haplotype maps, Bioinformatics 21 (2005) 263–264.
[32] M.A. Eberle, M. Rieder, L. Kruglyak, D. Nickerson, Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome, PLos Genet. 2 (2006) 1319.

[33] R.D. Hernandez, M. Hubisz, D.A. Wheeler, D.G. Smith, B. Ferguson, J. Rogers, L. Nazareth, A. Indap, T. Bourquin, J. McPherson, D. Muzny, R. Gibbs, Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques, Science 316 (2007) 240–243.
[34] M.K. Hasan, Mitochondrial DNA of rhesus macaques (*Macaca mulatta*) from Bangladesh, The 23rd Biannual Meeting of the International Primatological Society, Kyoto, Japan, 2010.
[35] E.S. Lander, I.H.G.S. Consortium, Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[36] A.S. Hinrichs, D. Karolchik, R. Baertsch, G.P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T.S. Furey, R.A. Harte, F. Hsu, J. Hillman-Jackson, R.M. Kuhn, J.S. Pederson, A. Pohl, J. Raney, K.R. Rosenbloom, A. Siepel, K.E. Smith, C.W. Sugnet, A. Sultan-Qurraie, D.J. Thomas, H. Trumbower, R.J. Weber, M. Weirauch, A.S. Zweig, D. Haussler, W.J. Kent, The UCSC genome browser database: update 2006, Nucleic Acids Res. 34 (2006) D590–D598.
[37] N. Wang, J.M. Akey, K. Zhang, R. Chakraborty, L. Jin, Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination and mutation, Am. J. Hum. Genet. 71 (2002) 1227–1234.
[38] D. Reich, M. Cargill, S. Bolk, J. Ireland, P.C. Sabeti, D.J. Richter, T. Lavery, R. Kouyoumjian, S.F. Farhadian, R. Ward, E.S. Lander, Linkage disequilibrium in the human genome, Nature 411 (2001) 199–204.
[39] Y. Kim, R. NieIsen, Linkage disequilibrium as a signature of selective sweeps, Genetics 167 (2004) 1513–1524.
[40] R. Nielsen, I. Hellmann, M. Hubisz, C. Bustamante, A. Clark, Recent and ongoing selection in the human genome, Nat. Rev. Genet. 8 (2007) 857–868.
[41] J.A. Satkoski Trask, W.T. Garnica, R.S. Malhi, S. Kanthaswamy, D.G. Smith, High-throughput single-nucleotide polymorphism discovery and the search for candidate genes for long-term SIVmac nonprogression in Chinese rhesus macaques (*Macaca mulatta*), J. Med. Primatol. 40 (2011) 224–232.
[42] J. Rogers, R. Garcia, W. Shelledy, J. Kaplan, A. Arya, Z. Johnson, M. Bergstrom, L. Novakowski, P. Nair, A. Vinson, D. Newman, G. Heckman, J. Cameron, An initial genetic linkage map of the rhesus macaque (*Macaca mulatta*) genome using human microsatellite loci, Genomics 87 (2006) 30–38.
[43] D.G. Smith, J. McDonough, D. George, Mitochondrial DNA variation within and among regional populations of longtail macaques (*Macaca fascicularis*) in relation to other species of the fascicularis group of macaques, Am. J. Primatol. 69 (2007) 182–198.
[44] S. Kanthaswamy, J. Satkoski, A. Kou, V. Malladi, D. Glenn Smith, 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, J. Med. Primatol. 39 (2010) 252–265.
[45] C.R. Carpenter, Breeding colonies of macaques and gibbons on Santiago Island, Puerto Rico, in: W.I.B. Beveridge (Ed.), Breeding Primates, S. Karger, Basel, 1972, pp. 76–87.
[46] J. Buettner-Janusch, G.A. Mason, V. Buettner-Janusch, D.S. Sade, Genetic studies of serum transferrins of free-ranging rhesus macaques of Cayo Santiago, *Macaca mulatta* (Zimmerman 1780), Am. J. Phys. Anthropol. 41 (1974) 217–232.
[47] D.O. Johnsen, History, in: B.T. Bennett, C.R. Abee, R. Hendrickson (Eds.), Nonhuman Primates in Biomedical Research, Academic Press, San Diego, 1995, pp. 1–12.
[48] T.J. Olivier, C. Ober, J. Buettner-Janusch, D.S. Sade, Genetic differentation among matrilines in social groups of rhesus monkeys, Behav. Ecol. Sociobiol. 8 (1981) 279–285.
[49] C. Duggleby, Blood group antigens and the population genetics of *Macaca mulatta* on Cayo Santiago: I. Genetic differentiation of social groups, Am. J. Phys. Anthropol. 48 (1978) 35–40.