ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Evans BJ, Tosi AJ, Zeng K, Dushoff J, Corvelo A, Melnick DJ. 2017 Speciation over the edge: gene flow among non-human primate species across a formidable biogeographic barrier. *R. Soc. open sci.* **4**: 170351.
http://dx.doi.org/10.1098/rsos.170351

Received: 13 April 2017
Accepted: 18 September 2017

Subject Category:
Biology (whole organism)

Subject Areas:
bioinformatics/evolution/genetics

Keywords:
mechanisms of speciation, gene flow, X chromosome, genomics, primate evolution, Wallace's Line

Author for correspondence:
Ben J. Evans
e-mail: evansb@mcmaster.ca

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3899413.

THE ROYAL SOCIETY PUBLISHING

Speciation over the edge: gene flow among non-human primate species across a formidable biogeographic barrier

Ben J. Evans[1,5], Anthony J. Tosi[2], Kai Zeng[3], Jonathan Dushoff[1], André Corvelo[4] and Don J. Melnick[5]

[1]Biology Department, Life Sciences Building Room 328, McMaster University, 1280 Main Street West, Hamilton, ON, Canada L8S4K1
[2]Anthropology Department, Kent State University, 238 Lowry Hall, Kent, OH 44242, USA
[3]Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK
[4]New York Genome Center, 101 Avenue of the Americas, New York, NY 10013, USA
[5]Department of Ecology, Evolution, and Environmental Biology, Columbia University, 10th floor Schermerhorn Extension, 119th Street and Amsterdam Avenue, New York, NY 10027, USA

BJE, 0000-0002-9512-8845

Many genera of terrestrial vertebrates diversified exclusively on one or the other side of Wallace's Line, which lies between Borneo and Sulawesi islands in Southeast Asia, and demarcates one of the sharpest biogeographic transition zones in the world. Macaque monkeys are unusual among vertebrate genera in that they are distributed on both sides of Wallace's Line, raising the question of whether dispersal across this barrier was an evolutionary one-off or a more protracted exchange—and if the latter, what were the genomic consequences. To explore the nature of speciation over the edge of this biogeographic divide, we used genomic data to test for evidence of gene flow between macaque species across Wallace's Line after macaques colonized Sulawesi. We recovered evidence of post-colonization gene flow, most prominently on the X chromosome. These results are consistent with the proposal that gene flow is a pervasive component of speciation—even when barriers to gene flow seem almost insurmountable.

1. Background

1.1. Wallace's Line and the drivers of speciation

A species is a group of reproductively compatible individuals with ancestor–descendant relationships that evolve through time [1]. Early ideas about the drivers of speciation recognized geographical isolation as an important prezygotic barrier to reproduction that contributes to this process, with this reasoning being heavily influenced by zoogeographic patterns (e.g. [2,3]). One particularly influential pattern is the sharp faunal transition that occurs between the islands of Borneo and Sulawesi, across 'Wallace's Line' [4,5]. Many vertebrate genera do not span this barrier [6] and it is generally thought that many species in this region evolved in allopatry [7–10] as a consequence of a dynamic history of connectivity or isolation of large landmasses [11]. However, our understanding of the drivers of speciation here and elsewhere is in flux, including recent reappraisals of the degree to which gene flow exists among closely related species (e.g. [12]), the evolutionary consequences of gene flow and adaptation during speciation (e.g. [13]) and the degree to which geographical isolation is associated with 'extrinsic' (abiotic) versus 'intrinsic' (ecological) barriers to reproduction [14–16]. Wallace's Line is an important evolutionary arena for studying speciation because some groups have anomalous distributions that span this barrier. These groups permit us to test whether allopatric lineages on either side of a precipitous biogeographic barrier are in fact isolated genetically and, if not, what genomic regions were affected by gene flow, and what were the adaptive implications.

1.2. Macaque monkeys have an anomalous distribution across Wallace's Line

Macaque monkeys (*Macaca*) have the largest distribution of all non-human primate genera [17] and are among the most diverse [18]. Although they originated in Africa [19], almost one-third of macaque species occur just east of Wallace's Line on the island of Sulawesi, which at less than 200 000 km^2 comprises less than 4% of the geographical distribution of macaques [17]. The Sulawesi macaques are endemic to this island, allopatrically distributed and differentiated in behaviour [20,21], cranial morphology [22], pelage and other aspects of external morphology [23], and genetic variation [24,25]. Molecular studies support monophyly of the Sulawesi macaques and a sister relationship to the pigtail macaque, although phylogenetic relationships among the Sulawesi macaques remain poorly resolved (e.g. [25–28]). That macaques dispersed at least once across Wallace's Line opens the possibility that it happened multiple times, and that variation among genomic regions could exist in the levels of gene flow across this biogeographic divide. To explore this possibility, we used restriction site-associated DNA (RADseq) data to characterize phylogenetic relationships and molecular variation among macaques in this geographical region, and we then used whole genome-sequence (WGS) data from one individual from each of three species to test for evidence of gene flow across Wallace's Line.

2. Material and methods

2.1. Molecular data

To perform tests for gene flow discussed below, it was necessary to establish a phylogenetic framework for our samples. We therefore generated and analysed RADseq [29] data from genetic samples of 40 individuals from 10 macaque species including all Sulawesi macaques and *Macaca nemestrina* and *M. siberu* from several sites in the Sunda Region (Borneo, Peninsular Malaysia, Sumatra and the Mentawai Islands). Genetic samples used in this study were collected from captive individuals as previously described [25,30] and the geographical origins of these samples are depicted in figure 1. Additional details on these samples and on macaque taxonomy are presented in the electronic supplementary material.

Two RADseq libraries were prepared by Floragenex (http://www.floragenex.com/). The first library was previously reported and constructed from nine *M. tonkeana* samples [31], and was sequenced using one Illumina HiSeq 2500 lane and 100 base pair (bp) paired-end reads. The second library has not previously been reported, included 31 new samples and was sequenced on one Illumina HiSeq 2500 lane with single-end 100 bp reads. Because reverse reads were available for only nine individuals, we restricted our analysis to the forward reads from these RADseq data.

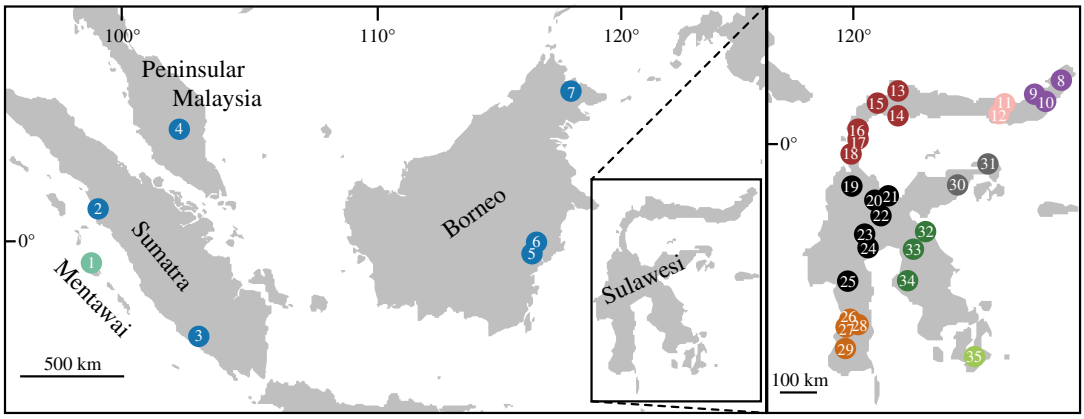

Figure 1. Thirty-five geographical origins of the 40 genetic samples analysed in this study. Numbered localities correspond to the approximate geographical origins of samples as follows, with asterisks denoting samples for which precise provenance is unknown: (1) *M. siberu*, (2) Ngasang, (3) Kedurang, (4) Malay*, (5) PM665, (6) PM664*, (7) Sukai, Gumgum, (8) PF660, (9) PF1003, (10) PF1001*, (11) PM1000, (12) PF654, (13) PF648, (14) PF651, (15) PM645, (16) PM639, (17) PF644, (18) PF643, (19) PF515, (20) PM565, PM566, PM567, (21) PM561, (22) PM582, (23) PM584, (24) PM592, (25) PM602, (26) PM613, (27) PM618, (28) PM614, PF615, PM616, (29) PF713, (30) PF549, (31) PM545, (32) PM571, (33) PM596, (34) PF625, (35) PF707. Dots are coloured by species as detailed in figures 2 and 3.

In addition to RADseq, WGS was performed on three of these samples with a specific aim of testing for evidence of gene flow across Wallace's Line. WGS data were collected for one male *M. nemestrina* (PM664), one male *M. tonkeana* (PM592) and one female *M. nigra* (PF660) using the Illumina HiSeqX platform with paired-end 150 bp reads.

2.2. Read filtering and quality control

RADseq data were de-multiplexed, trimmed and filtered using the process_radtags program of STACKS v. 1.21 [32,33]. Reads were initially truncated to 75 bp; miscalled barcodes that differed by up to three mutations from only one barcode were rescued, and those with uncalled bases or bases with an average Phred-scaled quality score lower than 10 were removed. All reads were then filtered again with TRIMMOMATIC v. 0.36, removing overrepresented sequences that were identified using FASTQC [34] and requiring retained sequences to have a minimum length of 36 bp and an average Phred-scaled quality score of at least 15 in a sliding window of 4 bp. After filtering repetitive regions (described below), the number of mapped RADseq reads per individual ranged from a low of 256 400 (for *M. maura* individual PM613) to a high of 6 346 894 (for *M. tonkeana* individual PM592) with an average and standard error of 2 016 557 and 301 929 reads per individual, respectively.

2.3. Genotyping and data filtering

For both RADseq and WGS data, the 'MEM' algorithm of BWA v. 0.7.8 [35] was used to map reads from each individual to a rhesus reference genome (rhemac2) which was downloaded from the University of California Santa Cruz Genome Browser (https://genome.ucsc.edu/). For the RADseq data, coverage of mapped reads ranged from a minimum of 7X (for *M. nigrescens* individual PF654) to a maximum of 83X (for *M. tonkeana* individual PM603), with an average and standard error of 36X and 4X, respectively. For the WGS data, coverage was greater than 40X for each of the three samples.

The Genome Analysis Toolkit (GATK) v. 3.6 [36] was used to perform genotyping and filtering as recommended by the 'Best Practices' pipeline [37,38]. Example command lines for initial generation of the WGS genotypes are provided in the electronic supplementary material. This included realignment of insertion/deletion (indel) polymorphisms with the RealignerTargetCreator and IndelRealigner functions. Genotyping was performed with the HaplotypeCaller and GenotypeGVCFs functions using, respectively, the `EMIT_ALL_CONFIDENT_SITES` and `includeNonVariantSites` functions of these commands. After this, the VariantFiltration and SelectVariants functions of GATK and a perl script were used to identify and remove positions that spanned an indel plus a buffer of 3 bp in both directions, repetitive regions identified in the reference genome by RepeatMasker [39], and individual genotypes that had coverage of less than 5X.

The WGS data were handled somewhat differently from the RADseq data in order to accommodate the different nature of these data (paired-end rather than single-end, shotgun sequencing rather than RADseq). Instead of trimming the WGS data with TRIMMOMATIC, we relied on the BWA MEM algorithm to softclip adapter sequences and used the BaseRecalibration function of GATK to recalibrate quality scores, excluding from the error model variant positions that were pre-called using Haplotypecaller. We did not perform base recalibration on the RADseq data because we performed stricter quality filtering on those data than the WGS data prior to mapping. Additionally, we did not perform de-duplication on the RADseq data because most of these data were single-end reads; for the WGS data de-duplication was performed with the MarkDuplicates function of PICARD (http://broadinstitute.github.io/picard).

Because our RADseq and WGS data included a mixture of male and female individuals, a haploid genotype was inferred for all sites on the X chromosome based on the allele with the highest depth of coverage (hereafter we refer to this as 'X chromosome genotyping by depth of coverage'). For heterozygous sites, the single nucleotide polymorphism (SNP) with the highest coverage was used; if two SNPs had equal coverage, one was randomly selected. We also performed other approaches to filter and genotype the X chromosomes of males and females, which are discussed in the Results and discussion section, with similar results. For all analyses, we assumed that sites that mapped to the rhesus X chromosome also are on the X chromosome of the other macaque species we examined, and the same for the autosomes.

2.4. Phylogenetic and principal components analysis of restriction site-associated DNA data

Phylogenetic analysis of the RADseq data was performed using IQTREE v. 1.5.0a [40] on the concatenated RADseq data from the autosomes and also a separate analysis of the RADseq X chromosome data. For both datasets, outgroup sequences were included from a human and an anubis baboon (genome assemblies hg19 and papAnu2, with the genome alignment to the rhesus macaque obtained from the University of California Santa Cruz Genome Browser or generated using LASTZ [41], respectively, as described in Evans *et al.* [31]). For the analysis of autosomal DNA, IQTREE selected the general time reversible model with Γ distributed rate heterogeneity based on the Bayesian information criterion (BIC). For the analysis of X chromosome DNA, IQTREE selected the TVM model of evolution with a proportion of invariant sites and a Γ distributed rate heterogeneity. For both analyses, node support was evaluated using the ultrafast bootstrap approach as implemented by IQTREE. Because this analysis involves concatenated data, divergence times may not appropriately accommodate incomplete lineage sorting (ILS), and this aspect of the analysis of autosomal data is intended for qualitative rather than quantitative purposes. The phylogenetic analysis of the X (featured in the electronic supplementary material) carries the additional caveat that it was performed on a subset of the intra-individual molecular polymorphism in each female, using only the variant at each polymorphic site with the highest depth of coverage, as described above.

The program MCMCTREE, which is part of the PAML v. 4.8 software suite [42], was then used to convert the maximum-likelihood (ML) topologies obtained from IQTREE to a chronogram. The independent evolutionary rates model was used, and the analysis included only those data that had no heterozygous sites or missing data. For the autosomal and X chromosome analyses, the HKY85+Γ model of nucleotide substitution was deployed—of the models implemented by MCMCTREE, this was the most similar to the model selected by the BIC in IQTREE. The alpha parameter for the Γ distribution was set to the ML estimate of 0.2176 which was recovered from IQTREE for the autosomes and 1.106 for the X. For calibration of both analyses, the 95% confidence interval (95% CI) for the age of the Old World monkeys and apes (Catarrhini) was set to a lower and upper bound of 28 and 36 million years ago (Ma) and the divergence time of baboons and macaques was set to a lower and upper bound of 10 and 13 million years ago (Ma), following Finstermeier *et al.* [43]. To expedite the MCMCTREE analysis of the autosomal DNA only, the 'cleandata' option was used to exclude sites with ambiguous or missing data.

To further visualize genetic relationships among these data, we performed a principal components analysis (PCA) on the filtered autosomal RADseq genotypes using the program SNPRELATE [44]. The SNPs were re-coded based on the dosage of the reference allele for all variant sites ('method =copy.num.of.ref') and were pruned to include only SNPs from sites with no missing data and that had a linkage disequilibrium threshold of 0.2 or less based on the composite measure of linkage disequilibrium [45] within a genomic window of size 500 000 bp. We performed a PCA on the full RADseq autosomal dataset and also on a reduced RADseq autosomal dataset including only the Sulawesi macaques. We also performed analyses of polymorphism on the X and autosomes as described and presented in the electronic supplementary material.

2.5. Gene flow analysis of whole genome-sequence data and divergence

If no gene flow occurred after macaques colonized Sulawesi from Borneo, phylogenetic analyses presented below indicate that most genomic regions of the Sulawesi macaques would be expected to be monophyletic with respect to the pigtail macaque *M. nemestrina*, with the exception of regions with ILS. ILS is expected to cause some genomic regions in a Sulawesi macaque to be more closely related to *M. nemestrina* than to other Sulawesi macaques (i.e. paraphyly of genetic variation in Sulawesi macaques), even though Sulawesi macaques are monophyletic over most of their genome. If gene flow occurred among macaques on either side of Wallace's Line after macaques reached Sulawesi, there might be an excess of derived mutations (based on comparison to an outgroup genome) that are shared by a pigtail macaque from eastern Borneo and a macaque from western Sulawesi (e.g. *M. tonkeana*) when compared with derived sites that are shared between a pigtail macaque and a macaque from eastern Sulawesi (e.g. *M. nigra*). This expectation forms the basis of the D-statistic, which is also known as the ABBA-BABA test [46–50].

We hypothesized that if gene flow did occur between macaques on either side of Wallace's line, then it would more likely be between the pigtail macaque (*M. nemestrina*) on Borneo and the tonkean macaque (*M. tonkeana*) on west central Sulawesi than between the pigtail macaque and the Celebes crested macaque (*M. nigra*) from the northeast tip of Sulawesi, because the first species pair are geographically closer to each other than the second species pair. To test this, we calculated the D-statistic and also a modification of the admixture fraction f proposed by Green *et al.* [46], called f_{DM}, which was calculated as described on page 8 of the electronic supplementary material of Malinsky *et al.* [51]. f_{DM} is distributed on the interval $[-1, 1]$ and, under the null hypothesis of no introgression after colonization of Sulawesi, this statistic should be symmetrically distributed around zero. If the relative rate of gene flow is higher between *M. nemestrina* and *M. tonkeana* than between *M. nemestrina* and *M. nigra*, then f_{DM} will be greater than zero. Alternatively, if the opposite is true, f_{DM} will be less than zero. The null hypothesis that f_{DM} is equal to zero was evaluated using the weighted block jackknife approach [46] with f_{DM} values in non-overlapping 5 000 000 bp genomic regions that were weighted by the sum of the numbers of ABBA and BABA sites ([46] and defined below) in each window. For the autosomes, the gene flow statistics were calculated from heterozygous and homozygous genotypes. For the X chromosome, these statistics were calculated from variants from each individual with the highest depth of coverage as described above. We also explored other genotyping approaches for the X chromosome WGS data that are detailed below. Justification for genotyping the X chromosome by depth, at least in males, was based in part on the identification of pseudoheterozygous genotypes in the non-pseudoautosomal region of the X when diploid genotypes were inferred for this region. To illustrate the number of positions with shared and unshared heterozygous genotypes across individuals in the non-pseudoautosomal region of the X, a Euler diagram was generated using the R package 'eulerr' [52]. For all of these analyses, sites with missing genotypes were excluded.

Divergence was calculated as the average per-site per cent nucleotide difference between both alleles carried by two individuals. Except where stated in the electronic supplementary material, tables, divergence is presented without correction for multiple substitutions.

2.6. Testing for contamination

As a rough assessment of potential sample contamination by human DNA, which could potentially affect our inferences of gene flow, 1% of the WGS read data from each sample were randomly selected and classified against an index containing 12 full primate genomes (*Microcebus murinus*, *Chlorocebus sabaeus*, *Nomascus leucogenys*, *Callithrix jacchus*, *M. fascicularis*, *M. mulatta*, *Papio anubis*, *Gorilla gorilla gorilla*, *Pan paniscus*, *P. troglodytes*, *Pongo abelii*, *Homo sapiens*), using TAXMAPS [53]. With an aim of minimizing false positives due to (i) sequence similarity between macaques and these other species and (ii) the more complete sequencing coverage of the human reference genome, we used a strict paired-end classification mode that requires both mates to be mapped, and opted to compute the 'lowest common ancestor' between the independent classification of each mate in the pair.

3. Results and discussion

3.1. Evolutionary patterns among Southeast Asian macaque monkeys

Despite a large amount of missing data (electronic supplementary material), the ML topology recovered from autosomal RADseq provided strong statistical support for relationships among Southeast Asian

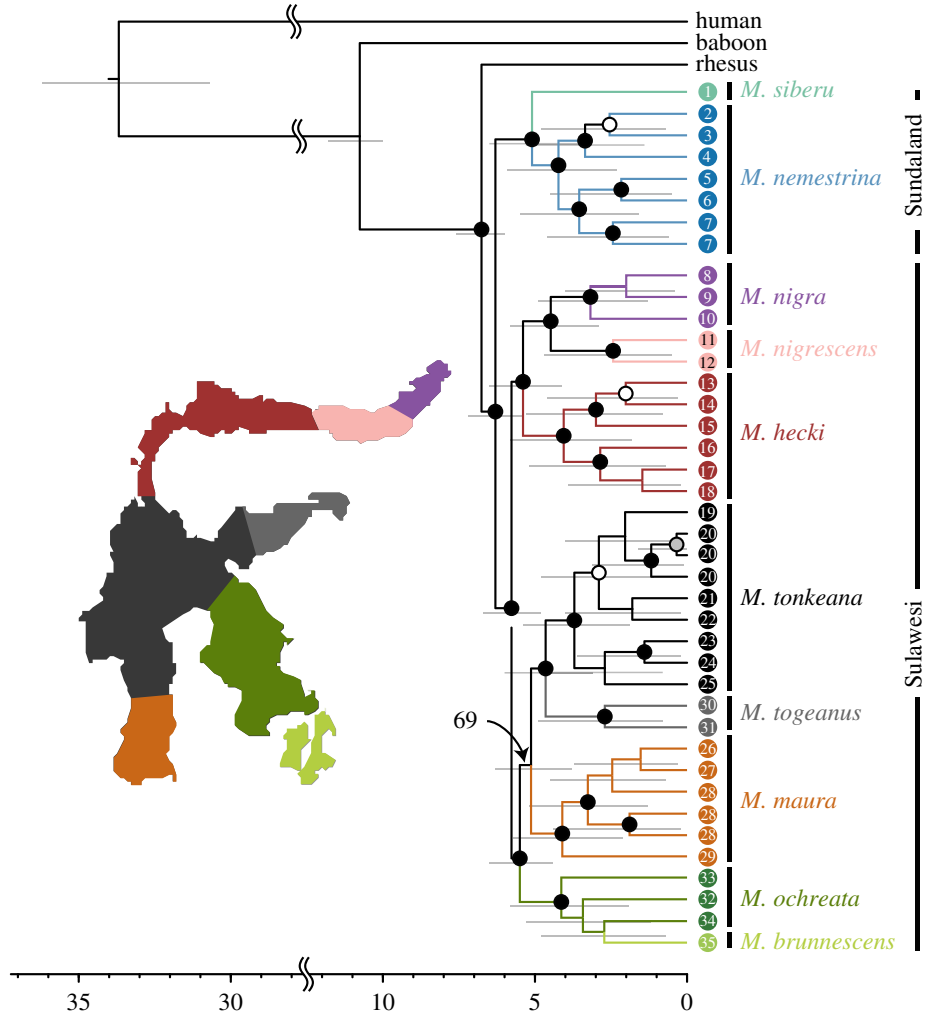

Figure 2. Time-calibrated phylogeny (chronogram) recovered from analysis of autosomal RADseq data. Scale indicates divergence in million years ago (Ma). Black, grey, and white dots over nodes reflect ultrafast bootstrap values from IQTREE that are greater than 99, 95 and 90, respectively. Grey bars near each node indicate the 95% CI for divergence times recovered from MCMCTREE. An inset indicates ranges of Sulawesi macaques and low bootstrap support for one node is indicated with an arrow. Tips are numbered according to their geographical localities depicted in figure 1.

macaques that clustered by species and geographical region, and with unprecedented statistical support for relationships among Sulawesi macaques (figure 2). Well-supported relationships include strong support for monophyly of *M. nemestrina* + *M. siberu*, and for monophyly of the Sulawesi macaques. Within the Sulawesi macaques, two clades correspond, respectively, to species in the north peninsula and species in the rest of Sulawesi. Within the strongly supported clade that includes macaques from the rest of Sulawesi, there is poor resolution among three well-supported clades which include (i) *M. maura*, which occupies the southwestern peninsula, (ii) (*M. ochreata* + *M. brunnescens*), which occupy the southeast peninsula and surrounding islands, and (iii) (*M. tonkeana* + *M. togeanus*), which occur in central/central eastern Sulawesi.

Geographical structure of phylogenetic relationships is observed within species as well. For example, samples from the southern and eastern portions of the range of *M. hecki* each form a clade, and samples from southern and northern portions of the range of *M. tonkeana* also each form a clade. The only species that was not monophyletic in this analysis is *M. ochreata*, the southernmost sample from which forms a weakly supported clade with the *M. brunnescens* sample. Divergence estimates point to a similar timing of divergence of extant Sulawesi macaques from each other, and of *M. siberu* from *M. nemestrina*, both of which occurred approximately 5–6 Ma (figure 2). This corresponds with the earliest fossil evidence of Asian macaques approximately 5.5 Ma [54].

Phylogenetic analysis of the X chromosome RADseq data also recovered strong phylogenetic support for monophyly of the Sulawesi macaques with respect to *M. nemestrina* and for monophyly

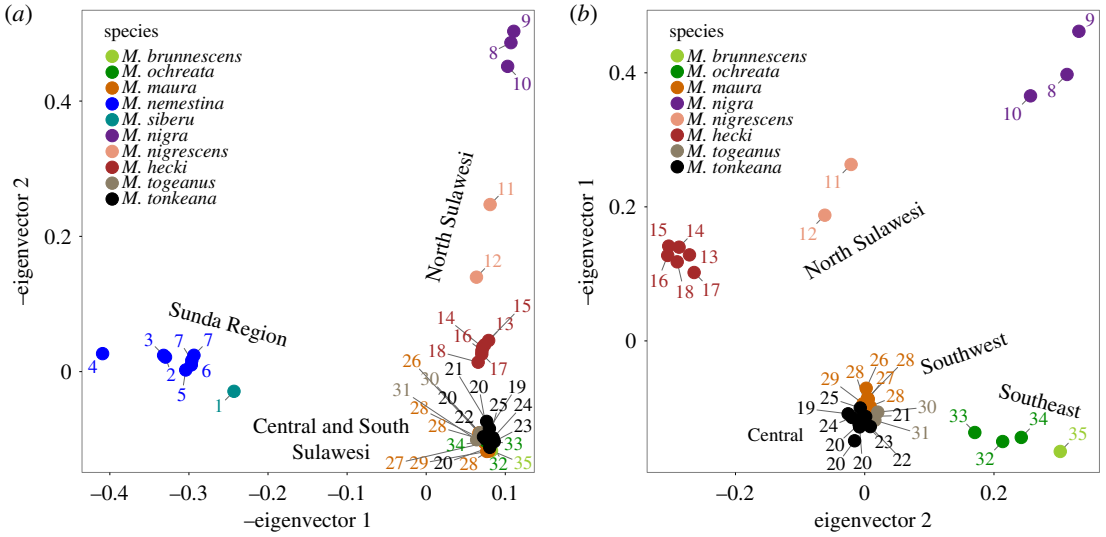

Figure 3. PCA analysis for all RADseq data (*a*) and RADseq data from only Sulawesi (*b*). Individual samples are numbered according to their geographical origins depicted in figure 1.

of the Sulawesi macaques of the northern peninsula with respect to the rest of Sulawesi (electronic supplementary material, figure S1). Some differences in phylogenetic relationships were inferred in this analysis compared to the autosomes, but none were well supported. Polymorphism on the X chromosome was lower than expected in four species with population sampling, but did not depart from expectations after allowing for a dynamic demography and selection on GC content using an ML model (model described in electronic supplementary material, tables S3–S11 and figure S2, and in [31]).

Similarly to the phylogenetic analysis, PCA of the autosomal RADseq data clustered samples by species and geographical origin (figure 3). The first PCA considered 2845 variable positions that had no missing data and low or no linkage disequilibrium among the 40 samples. The first eigenvector, which accounted for 7.45% of the variation in the data, separated variation in macaques from the Sunda Region and Sulawesi. The second eigenvector, which accounted for 4.85% of the variation in the data, separated variation in individuals of the northern peninsula of Sulawesi from the remainder of Sulawesi.

When autosomal RADseq data from only Sulawesi samples were analysed by a second PCA, 2969 variable positions had no missing data and low or non-existent level of linkage disequilibrium among the 32 samples (figure 3). The first eigenvector accounted for 6.88% and separated macaques of the northern peninsula from the remainder of Sulawesi, and also within the northern peninsula. The second eigenvector accounted for 5.86% of the variation and further separated geographical variation in macaques within each of these regions of Sulawesi.

Overall, these analyses provide strong support for monophyly of the Sulawesi macaques across the autosomes and X chromosomes, a strong correspondence between geography and macaque phylogenetic relationships on Sulawesi, and identify a non-significant dearth of molecular variation on the X chromosome.

3.2. Gene flow across Wallace's Line, especially on the X

Using WGS data from three individuals and a rhesus macaque genome sequence as an outgroup, a significant excess of shared derived sites was observed between *M. nemestrina* from Borneo and *M. tonkeana* compared to between *M. nemestrina* and *M. nigra* on the X chromosome, but not on the autosomes. This excess of shared derived sites spans Wallace's Line, even though most variation throughout the genome strongly supports monophyly of the Sulawesi macaques in the autosomes and the X (electronic supplementary material, figure 2).

On the X, the f_{DM} statistic was positive, which is consistent with gene flow between *M. nemestrina* and *M. tonkeana*; the magnitude and confidence intervals varied depending on the filters applied, and some combinations of filters and outgroups did not recover f_{DM} statistics that departed significantly from zero. However, for reasons discussed below, we view some of these analyses to be conservative. After removing repetitive regions and genotyping the X chromosome of males and females based on depth,

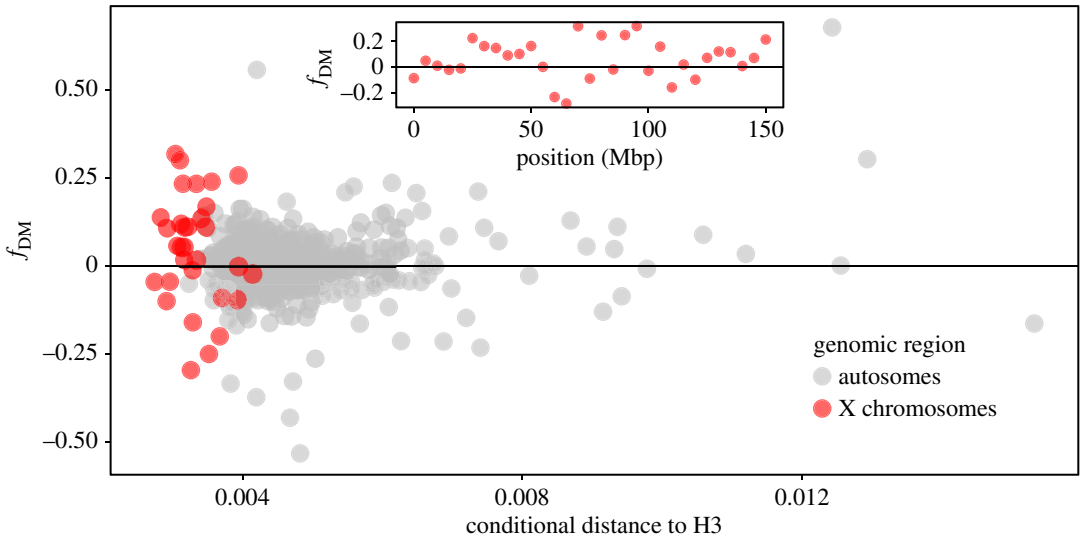

Figure 4. Conditional genetic distance to the *M. nemestrina* genome (H3) as a function of f_{DM} statistic for autosomes (grey, calculated including sites with heterozygous and homozygous genotypes) and the X chromosome (red, based on genotyping by depth after excluding positions with heterozyous diploid genotypes in males). Statistics are based on WGS data divided into non-overlapping windows of the reference genome spanning five million base pairs. Positive values of f_{DM} indicate an excess of derived sites (relative to the rhesus macaque) that are shared between the *M. tonkeana* (H2) genome and H3; negative values indicate an excess of derived sites that are shared between the *M. nigra* (H1) genome and H3. Genetic distances are 'conditional' in the sense that the uncorrected per cent of divergent sites between H2 and H3 or between H1 and H3 is plotted depending on whether f_{DM} is positive or negative, respectively, for each genomic window. Inset depicts f_{DM} in 5 Mbp windows on the X chromosome.

the f_{DM} statistic was 0.3049 (95% CI: 0.1675–0.4423). Because these values have only one nucleotide called per site per individual, the values of Patterson's *D* are the same as f_{DM}. In these X chromosome data, there were 40 304 BBAA sites, 3781 ABBA sites and 2014 BABA sites, which, respectively, refer to sites with shared derived nucleotides in *M. tonkeana* and *M. nigra* ('BBAA' sites), *M. tonkeana* and *M. nemestrina* ('ABBA' sites), or *M. nigra* and *M. nemestrina* ('BABA' sites). When autosomal sites were considered, there was also an excess of shared derived sites between *M. nemestrina* from Borneo and *M. tonkeana*, but this was not significantly greater than zero. The autosomal f_{DM} statistic was 0.0042 (95% CI: −0.0065–0.0148) and Patterson's *D* was 0.0027 (95% CI: −0.0044–0.0098). In the autosomal data, there were 1 203 957 BBAA sites, 145 277 ABBA sites and 144 491 BABA sites.

One concern in this analysis is that the *M. nigra* individual was a female, whereas the *M. nemestrina* and *M. tonkeana* individuals were male, raising the possibility that there was some sort of systematic bias in genotyping the X chromosome of male and female individuals. For this reason, we explored the effect of deleting sites for which a heterozygous genotype was recovered in the diploid genotype of males prior to genotyping sites on chromosome X based on the highest depth of coverage. After excluding these sites plus a 3 bp buffer, the f_{DM} statistic was lower but still significantly greater than zero: 0.0752 (95% CI: 0.0210–0.1293). This analysis had 39 825, 1938 and 1667 BBAA, ABBA and BABA sites, respectively. Significantly higher f_{DM} statistics were also recovered when diploid rather than haploid genotypes were used for the female in each of these analyses (electronic supplementary material, table S2). When baboons were used as an outgroup instead of the rhesus macaque and/or when the coverage cut-off was increased from 5X to 12X, a significantly higher f_{DM} statistic was recovered when all sites were considered, but not after excluding sites in which males had heterozygous diploid genotypes (electronic supplementary material, table S2). When we additionally excluded sites with heterozygous diploid genotypes in the female, a conservative measure based on the analysis of heterozygous genotypes on the X chromosome discussed below, the f_{DM} statistic was still positive in most of the analyses, although the 95% CIs overlapped zero (electronic supplementary material, table S2).

In figure 4, we present the results from chromosome X that were recovered after genotyping all individuals by depth after excluding positions with heterozyous diploid genotypes in males and using the rhesus macaque as an outgroup. The rhesus macaque sequence is generally a better outgroup for this analysis than the baboon sequence because it is more closely related to the ingroup taxa and therefore

has fewer lineage-specific mutations. Additionally, the genome sequence is more complete for the rhesus macaque, so more data are considered when this species is set as an outgroup.

To further explore possible sex-specific genotype biases, we used information from [55] to infer that the boundary of the rhesus macaque pseudoautosomal region is at approximately position 403 495 of the chromosome X sequence in the rhemac2 rhesus macaque genome assembly. Of the pseudodiploid genotypes that were inferred from the male X outside of the pseudoautosomal region, a very small proportion of sites (0.046% or 0.048% for *M. tonkeana* and *M. nemestrina*, respectively) of the genotypes were heterozygous, and most of these sites were heterozygous in both males (electronic supplementary material, figure S3). In this same region of the X, twice as many genotypes (0.096%) of the female *M. nigra* were heterozygous, and most of these sites were not heterozygous in either male (electronic supplementary material, figure S3). These results are consistent with the proposal that many of the pseudoheterozygous genotypes in males arose as a consequence of mismapped reads from the Y chromosome, or other forms of genotyping error. By contrast, most of the heterozygous sites in the female appear to be real, although some of these also appear to be due to genotyping error (e.g. heterozygous positions on the non-pseudoautosomal region of the female chromosome X that are also heterozygous in both males). A summary of heterozygous non-pseudoautosomal region sites that had heterozygous genotypes in each individual is presented in electronic supplementary material, figure S3. Overall then, this analysis argues that most of the heterozygous genotype inferences in the non-pseudoautosomal region of the female X chromosome have a biological basis, as opposed to being genotype errors. Clearly, a useful direction for further study would be to analyse samples from only one sex in order to minimize, or at least homogenize, genotyping error across all samples in the analysis.

Another alternative explanation for a significant departure of ABBA-BABA statistics from zero is that there were different rates of evolution in each species on Sulawesi. For example, if *M. nigra* were to evolve more quickly than *M. tonkeana*, we would expect positions with a shared derived nucleotide in both Sulawesi species and also the pigtailed macaque (i.e. 'BBBA' sites) to evolve more frequently into CBBA sites than to BCBA sites, where 'C' refers to a derived site that evolved from 'B', which itself is a derived site compared to the rhesus outgroup sequence (which is designated 'A'). To the extent that 'C' is a reversion to an ancestral 'A' nucleotide (which is expected in about one out of every three mutations), this could increase the number of apparent ABBA sites compared to BABA sites, and thus elevate the ABBA-BABA statistics (f_{DM}, D) above zero without gene flow.

To explore this possibility, we quantified the number of CBBA and BCBA sites in the data with the expectation that they should be equivalent if there was no substantial difference in the rate of evolution between *M. nigra* and *M. tonkeana*. On the X, there were slightly more BCBA than CBBA sites (138 and 135, respectively), and there were 360 'CCBA' sites (i.e. sites with three segregating nucleotides in which one had the highest frequency in both Sulawesi macaques, another had the highest frequency in the pigtailed macaque and a third was in the rhesus genome). This was also the case after removing sites in which males had heterozygous diploid genotypes (this analysis recovered 93 BCBA sites, 85 CBBA sites and 106 CCBA sites). On the autosomes, there also were slightly more BCBA than CBBA sites (5265 and 5212, respectively), and there were 7636 CCBA sites. These observations suggest that the rates of evolution in *M. tonkeana* and *M. nigra* were very similar; in fact, that there were slightly more BCBA than CBBA sites on the X and the autosomes make the ABBA-BABA statistics conservative with respect to the inference of gene flow between *M. tonkeana* and *M nemestrina*.

We also explored the possibility that variation in the level of contamination by human DNA could somehow influence the results of the ABBA-BABA statistics. Analysis with TAXMAPS suggested that the level of contamination of human DNA was very low (0.0061%, 0.0055% and 0.0065% of the reads from the *M. nemestrina*, *M. tonkeana* and *M. nigra* individuals, respectively). These proportions are not substantially different among the samples, and not substantially different in the male samples compared to the female. We suspect these percentages are overestimates caused by the more complete genome sequence of humans compared to the other species in this analysis.

The estimated divergence time between the Sulawesi macaques and *M. nemestrina* is 6 Ma (figure 2). Whether these genomic patterns are most consistent with a pulse of gene flow soon after Sulawesi was colonized by macaques, ongoing gene flow or some other scenario remains beyond the scope of this study. However, some general insights may be gleaned by examining divergence in putatively recently exchanged genomic regions. Recently, exchanged genomic regions are expected to have low divergence between the species that exchanged them, whereas anciently exchanged regions or regions with incomplete lineage sorting (ILS) are expected to have high divergence because more time has elapsed since the genomic region was shared. Interestingly, several genomic regions with high f_{DM} values also have relatively low divergence between *M. nemestrina* and *M. tonkeana* (figure 4). This again suggests

against the relatively high f_{DM} values of these regions being due to ILS alone, and argues for further study of the nature and timing of possible gene flow across Wallace's Line using WGS data (analyses of the RADseq data, not shown, provided insufficient statistical power for this analysis of gene flow). Taken together, these results open the possibility that gene flow occurred between macaques on either side of Wallace's Line after the initial colonization of Sulawesi, and in a way that more profoundly affected variation on the X chromosome than the autosomes. A previous study did not recover evidence for gene flow across the Makassar Strait [26], although data analysed in that study (Sanger re-sequencing from a few dozen genic regions) would be unlikely to detect low levels of gene flow. Of note is that Evans *et al*. [26] detected paraphyletic molecular variation in the X-linked gene *TBL1X* in Sulawesi macaques, which could be a consequence either of ILS or gene flow.

Evidence of gene flow has been recovered from all major primate groups, including other papionins (e.g. [56]). Thus, the most striking aspects of this result are not that gene flow may have occurred between two primate species, but instead that (i) it may have occurred across such a precipitous biogeographic barrier, and (ii) its effect on the X chromosome may be more substantial than on the autosomes. There are several non-mutually exclusive scenarios that could explain this pattern. One possibility is that a low level of gene flow between *M. nemestrina* and *M. tonkeana* resulted in a small amount of shared variation in the autosomes and the X, and that this was followed by genetic drift or natural selection that increased the frequencies of some transferred regions on the X to a greater degree than on the autosomes. Indeed, strong effects of natural selection on the X have been reported in several other primates, including humans (e.g. [57–59]). A higher level of gene flow across the Makassar Strait on the X than the autosomes could also result if it was mediated mostly by female migration. But this would also be surprising because female papionin monkeys are generally philopatric [60] and this scenario is the opposite of expectations of a 'large X effect' in speciation [61], which predicts a lower level of gene flow between species on the X compared to the autosomes. Potentially relevant to these findings is the occurrence of interspecies hybridization between all parapatric species of Sulawesi macaque [24,62–66]. If gene flow across hybrid zones on Sulawesi were mediated mostly by male migration, then molecular varation introduced onto Sulawesi across Wallace's Line would be expected to become more widely distributed (and thus less detectable by ABBA-BABA statistics) in the autosomes than on the X chromosome.

It is also possible that the significant departure of ABBA-BABA statistics from zero may in fact be due to factors other than gene flow. Although we did not find evidence that variation in the rate of evolution could account for this pattern, it is conceivable that population structure, for example due to isolation by distance, of the X chromosome existed in the ancestor of (*M. nemestrina* + the Sulawesi macaques) prior to dispersal of a subpopulation from Borneo to Sulawesi. If this were the case, the three whole-genome sequences we considered may not have captured a representative sample of molecular variation in each species, which could result in a misleading signal of gene flow. As stated above, it is also possible that natural selection or genetic drift could alter allele frequencies in such a way as to deliver a significant ABBA-BABA statistic. These possibilities could be explored with additional data from other individuals, and possibly allow genomic patterns that stem from ILS to be teased apart from those that stem from gene flow across Wallace's Line.

4. Conclusion

In nature, speciation plays out in several dimensions: the geographical context ranges from allopatry to sympatry, gene flow varies from absent to extensive, and differentiation can be driven mostly by genetic drift or more prominently sculpted by natural selection. At first glance, speciation on either side Wallace's Line appears to have unfolded largely with no gene flow across this barrier. However, our analyses of Southeast Asian macaque monkeys raise doubts about this assertion—at least for macaque monkeys—and provide several insights into diversification in this region, and to the process of speciation in general. Strong geographical structure of molecular variation in macaque RADseq data supports an important role of geography in regional faunal evolution. Thus, while Sulawesi Island may have been an archipelago in the past [11,67], the dispersal route of macaques among these palaeo-islands matches the modern geography of Sulawesi. We also identified a genomic signature of gene flow across Wallace's Line, with the most pronounced signal on the X chromosome. This finding opens the possibility that gene flow can occur across formidable biogeographic barriers, and that in such cases it may vary in magnitude among genomic regions. Other explanations—such as demography, unsampled molecular variation and natural selection—are also plausible, and warrant further testing with additional samples. To the extent that gene flow across Wallace's Line can be confirmed, this would contribute to examples of gene flow

between species pairs in the genus *Macaca* (e.g. [24]), in other papionin genera (e.g. [56]) and in primate genera that are more closely related to humans (e.g. [68]).

Ethics. Research and collection permits for this study were provided by the Indonesian Institute of Sciences/Lembaga Ilmu Pengetahuan Indonesia (LIPI). Genetic samples for this project were obtained using methods approved by the Institutional Animal Care and Use Committee (IUCAC) at Columbia University. No special research ethics approval was required for this research.

Data accessibility. All RADseq and WGS data and draft genome assemblies mapped to a rhesus macaque reference genome are deposited in the National Center for Biotechnology Information Short Read Archive accession numbers SRP041222, PRJNA398316 and PRJNA398316. Phylogenetic data, including trees and alignments, genotype files and scripts are deposited in Dryad as http://dx.doi.org/10.5061/dryad.3j218 [69]. Sampling locations are depicted in figure 1.

Authors' contributions. This study was designed by B.J.E. Samples were collected by B.J.E. and D.J.M., and library preparation and sequencing costs were contributed by B.J.E. and A.J.T. Analyses were performed by B.J.E., K.Z., J.D. and A.C. The paper was written by B.J.E. and all the authors provided comments.

Competing interests. We have no competing interests.

Funding. This work was supported by a grant to B.J.E. from the Natural Sciences and Engineering Research Council of Canada (RGPIN/283102-2012 and RGPIN-2017-05770) and Kent State University.

Acknowledgements. We thank Brian Golding for access to computing facilities and sharcnet (www.sharcnet.ca) and Janet Kelso for helpful discussion about bioinformatics. We thank Brian Charlesworth and two anonymous reviewers for helpful discussion and comments on a previous version of this manuscript. We thank the New York Genome Center for project support provided by Catherine Reeves and Bridget Riley-Gillis, and Joseph Solomon for laboratory assistance.

References

1. de Queiroz K. 2007 Species concepts and species delimitation. *Syst. Biol.* **56**, 879–886. (doi:10.1080/10635150701701083)
2. Mayr E. 1963 *Animal species and evolution*. Cambridge, MA: Harvard University Press.
3. Dobzhansky T *et al.* 1970 *Genetics of the evolutionary process, vol. 139*. New York, NY: Columbia University Press.
4. Wallace AR. 1863 On the physical geography of the Malay Archipelago. *J. R. Geol. Soc. Lond.* **33**, 217–234. (doi:10.2307/1798448)
5. Mayr E. 1944 Wallace's Line in the light of recent zoogeographic studies. *Q. Rev. Biol.* **19**, 1–14. (doi:10.1086/394684)
6. Cranbrook EO. 1981 The vertebrate faunas. In *Wallace's Line and plate tectonics* (ed. TC Whitmore), pp. 57–69. Oxford, UK: Oxford University Press.
7. Keast A. 2001 The vertebrate fauna of the Wallacean island interchange zone: the basis of inbalance and impoverishment. In *Faunal and floral migrations and evolution in SE Asia-Australasia* (eds I Metcalfe, JMB Smith, M Morwood, I Davidson), pp. 287–310. Rotterdam, The Netherlands: Balkema.
8. Musser G. 1987 The mammals of Sulawesi. In *Biogeographical evolution of the Malay Archipelago* (ed TC Whitmore), pp. 73–93. Oxford, UK: Clarendon Press.
9. Merker S, Driller C, Perwitasari-Farajallah D, Pamungkas J, Zischler H. 2009 Elucidating geological and biological processes underlying the diversification of Sulawesi tarsiers. *Proc. Natl. Acad. Sci.* **106**, 8459–8464. (doi:10.1073/pnas.0900319106)
10. Stelbrink B, Albrecht C, Hall R, von Rintelen T. 2012 The biogeography of Sulawesi revisited: Is there evidence for a vicariant origin of taxa on Wallace's anomalous island? *Evolution* **66**, 2252–2271. (doi:10.1111/j.1558-5646.2012.01588.x)
11. Hall R. 2009 Southeast Asia's changing palaeogeography. *Blumea-Biodiversity, Evolution and Biogeography of Plants* **54**, 148–161. (doi:10.3767/000651909X475941)
12. Kuhlwilm M *et al.* 2016 Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**, 429–433. (doi:10.1038/nature16544)
13. Lamichhaney S *et al.* 2015 Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**, 371–375. (doi:10.1038/nature14181)
14. Harrison RG. 2012 The language of speciation. *Evolution* **66**, 3643–3657. (doi:10.1111/j.1558-5646.2012.01785.x)
15. Feder JL, Egan SP, Nosil P. 2012 The genomics of speciation-with-gene-flow. *Trends. Genet.* **28**, 342–350. (doi:10.1016/j.tig.2012.03.009)
16. Endler JA. 1982 Problems in distinguishing historical from ecological factors in biogeography. *Am. Zool.* **22**, 441–452. (doi:10.1093/icb/22.2.441)
17. Fa JE. 1989 The genus *Macaca*: a review of taxonomy and evolution. *Mammal Rev.* **19**, 45–81. (doi:10.1111/j.1365-2907.1989.tb00401.x)
18. Roos C, Zinner D. 2015 Diversity and evolutionary history of macaques with special focus on *Macaca mulatta* and *Macaca fascicularis*. In *The Nonhuman Primate in Nonclinical Drug Development and Safety Assessment* (eds J Blümel, S Korte, E Schenck, G Weinbauer), pp. 3–16. Amsterdam, The Netherlands: Elsevier.
19. Stewart C-B, Disotell TR. 1998 Primate evolution–in and out of Africa. *Curr. Biol.* **8**, R582–R588. (doi:10.1016/S0960-9822(07)00367-3)
20. Thierry B, Bynum E, Baker S, Kinnaird MF, Matsumura S, Muroyama Y, O'Brien TG, Petit O, Watanabe K. 2000 The social repertoire of Sulawesi macaques. *Primate Res.* **16**, 203–226. (doi:10.2354/psj.16.203)
21. Riley EP. 2010 The endemic seven: four decades of research on the Sulawesi macaques. *Evol. Anthropol: Issues, News, and Reviews* **19**, 22–36. (doi:10.1002/evan.20246)
22. Albrecht GH. 1977 The craniofacial morphology of the Sulawesi macaques: multivariate approaches to biological problems. *Contrib. Primatol.* **13**, I–VIII.
23. Fooden J. 1969 *Taxonomy and evolution of the monkeys of Celebes*. Basel: Karger.
24. Evans B, Supriatna J, Melnick D. 2001 Hybridization and population genetics of two macaque species in Sulawesi, Indonesia. *Evolution* **55**, 1686–1702. (doi:10.1111/j.0014-3820.2001.tb00688.x)
25. Evans BJ, Supriatna J, Andayani N, Melnick DJ. 2003 Diversification of Sulawesi macaque monkeys: decoupled evolution of mitochondrial and autosomal DNA. *Evolution* **57**, 1931–1946. (doi:10.1111/j.0014-3820.2003.tb00599.x)
26. Evans BJ, Pin L, Melnick DJ, Wright SI. 2010 Sex-linked inheritance in macaque monkeys: implications for effective population size and dispersal to Sulawesi. *Genetics* **185**, 923–937. (doi:10.1534/genetics.110.116228)
27. Ghenu A-H, Bolker BM, Melnick DJ, Evans BJ. 2016 Multicopy gene family evolution on primate Y chromosomes. *BMC. Genomics.* **17**, 156. (doi:10.1186/s12864-015-2187-8)
28. Tosi AJ, Morales JC, Melnick DJ. 2003 Paternal, maternal, and biparental molecular markers provide unique windows onto the evolutionary history of macaque monkeys. *Evolution* **57**, 1419–1435. (doi:10.1111/j.0014-3820.2003.tb00349.x)
29. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008 Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS. ONE* **3**, e3376. (doi:10.1371/journal.pone.0003376)
30. Rosenblum LL, Supriatna J, Melnick DJ. 1997 Phylogeographic analysis of pigtail macaque populations (*Macaca nemestrina*) inferred from mitochondrial DNA. *Am. J. Phys. Anthropol.* **104**, 35–45. (doi:10.1002/(SICI)1096-8644(199709)104:1<35::AID-AJPA3>3.0.CO;2-C)

31. Evans BJ, Zeng K, Esselstyn JA, Charlesworth B, Melnick DJ. 2014 Reduced representation genome sequencing suggests low diversity on the sex chromosomes of tonkean macaque monkeys. *Mol. Biol. Evol.* **31**, 2425–2440. (doi:10.1093/molbev/msu197)
32. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013 Stacks: An analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–3140. (doi:10.1111/mec.12354)
33. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. 2011 Stacks: building and genotyping loci *de novo* from short-read sequences. *G3 (Bethesda)* **1**, 171–182. (doi:10.1534/g3.111.000240)
34. Andrews S. 2010 Fastqc: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed 8 October 2017).
35. Li H, Durbin R. 2010 Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595. (doi:10.1093/bioinformatics/btp698)
36. McKenna A *et al.* 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303. (doi:10.1101/gr.107524.110)
37. DePristo MA *et al.* 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498. (doi:10.1038/ng.806)
38. Van der Auwera GA *et al.* 2013 From fastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **11**, 1–11. (doi:10.1002/0471250953.bi1110s43)
39. Smit A, Hubley R, Green P. 1996 Repeat feature annotation. *RepeatMasker Open* **3**, 1996–2004. See http://www.repeatmasker.org.
40. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015 IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274. (doi:10.1093/molbev/msu300)
41. Harris RS. 2007 Improved pairwise alignment of genomic DNA. Doctoral dissertation, Pennsylvania State University University Park, PA, USA: ProQuest.
42. Yang Z. 2007 Paml 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591. (doi:10.1093/molbev/msm088)
43. Finstermeier K, Zinner D, Brameier M, Meyer M, Kreuz E, Hofreiter M, Roos C. 2013 A mitogenomic phylogeny of living primates. *PLoS. ONE.* **8**, e69504. (doi:10.1371/journal.pone.0069504)
44. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012 A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328. (doi:10.1093/bioinformatics/bts606)
45. Weir BS. 1979 Inferences about linkage disequilibrium. *Biometrics* **35**, 235–254. (doi:10.2307/2529947)
46. Green RE *et al.* 2010 A draft sequence of the Neandertal genome. *Science* **328**, 710–722. (doi:10.1126/science.1188021)
47. Martin SH, Davey JW, Jiggins CD. 2015 Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**, 244–257. (doi:10.1093/molbev/msu269)
48. Durand EY, Patterson N, Reich D, Slatkin M. 2011 Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252. (doi:10.1093/molbev/msr048)
49. Reich D *et al.* 2010 Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060. (doi:10.1038/nature09710)
50. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012 Ancient admixture in human history. *Genetics* **192**, 1065–1093. (doi:10.1534/genetics.112.145037)
51. Malinsky M *et al.* 2015 Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science* **350**, 1493–1498. (doi:10.1126/science.aac9927)
52. Larsson J. 2016 *eulerr*: area-proportional Euler diagrams, R package version 1.0.0.
53. Corvelo A, Clarke WE, Robine N, Zody MC. 2017 Taxmaps-ultra-comprehensive and highly accurate taxonomic classification of short-read data in reasonable time, *bioRxiv*, 134023. (doi:10.1101/134023)
54. Delson E. 1996 The oldest monkeys in Asia. In *International Symposium: Evolution of Asian Primates*, vol. 40, Freude and Kyoto University Primate Research Institute.
55. Hughes JF *et al.* 2012 Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**, 82–86. (doi:10.1038/nature10843)
56. Wall JD, Schlebusch SA, Alberts SC, Cox LA, Snyder-Mackler N, Nevonen KA, Carbone L, Tung J. 2016 Genomewide ancestry and divergence patterns from low-coverage sequencing data reveal a complex history of admixture in wild baboons. *Mol. Ecol.* **25**, 3469–3483. (doi:10.1111/mec.13684)
57. Veeramah KR, Gutenkunst RN, Woerner AE, Watkins JC, Hammer MF. 2014 Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans, *Mol. Biol. Evol.* **31**, 2267–2282. (doi:10.1093/molbev/msu166)
58. Dutheil JY, Munch K, Nam K, Mailund T, Schierup MH. 2015 Strong selective sweeps on the X chromosome in the human-chimpanzee ancestor explain its low divergence. *PLoS Genet.* **11**, e1005451. (doi:10.1371/journal.pgen.1005451)
59. Nam K *et al.* 2015 Extreme selective sweeps independently targeted the X chromosomes of the great apes. *Proc. Natl. Acad. Sci.* **112**, 6413–6418. (doi:10.1073/pnas.1419306112)
60. Silk JB. 1987 Social behavior in evolutionary perspective. In *Primate societies* (eds BB Smuts, DL Cheney, RM Seyfarth, RW Wrangham, TT Struhsaker), pp. 318–329. Chicago, IL: University of Chicago Press.
61. Coyne JA, Orr HA. 1989 Two rules of speciation. In *Speciation and its consequences* (eds D Otte, J Endler), pp. 180–207. Sunderland, MA: Sinauer Associates.
62. Watanabe K, Lapasere H, Tantu R. 1991 External characteristics and associated developmental changes in two species of Sulawesi macaques, *Macaca tonkeana* and *M. hecki*, with special reference to hybrids and the borderland between the species. *Primates* **32**, 61–76. (doi:10.1007/BF02381601)
63. Watanabe K, Matsumura S, Watanabe T, Hamada Y. 1991 Distribution and possible intergradation between *Macaca tonkeana* and *M. ochreata* at the borderland of the species in Sulawesi. *Primates* **32**, 385–389. (doi:10.1007/BF02382680)
64. Bynum E, Bynum D, Supriatna J. 1997 Confirmation and location of the hybrid zone between wild populations of *Macaca tonkeana* and *Macaca hecki* in central Sulawesi, Indonesia. *Am. J. Primatol.* **43**, 181–209. (doi:10.1002/(SICI)1098-2345(1997)43:3<181::AID-AJP1>3.0.CO;2-T)
65. Ciani AC, Stanyon R, Scheffrahn W, Sampurno B. 1989 Evidence of gene flow between Sulawesi macaques. *Am. J. Primatol.* **17**, 257–270. (doi:10.1002/ajp.1350170402)
66. Watanabe K, Matsumura S. 1991 The borderlands and possible hybrids between three species of macaques, *M. nigra*, *M. nigrescens*, and *M. hecki*, in the northern peninsula of Sulawesi. *Primates* **32**, 365–370. (doi:10.1007/BF02382677)
67. Hall R. 1998 The plate tectonics of Cenozoic SE Asia and the distribution of land and sea. In *Biogeography and geological evolution of SE Asia* (eds R Hall, JD Holloway), pp. 99–131. Leiden, The Netherlands: Backhuys Publishers.
68. De Manuel M *et al*. 2016 Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**, 477–481. (doi:10.1126/science.aag2602)
69. Evans BJ, Tosi AJ, Zeng K, Dushoff J, Corvelo A, Melnick DJ. 2017 Data from: Speciation over the edge: gene flow among non-human primate species across a formidable biogeographic barrier. Dryad Digital Repository. (doi:10.5061/dryad.3j218)

Supplementary Information

Taxonomy of silenus group macaques

Macaques are divided into four species groups based on morphology [1, 2] and these groups each correspond to distinct phylogenetic clades [3]. The silenus group includes the liontail macaque *M. silenus*, which occurs in southwest India, and several species from Southeast Asia, which are the focus of this study, including the pigtail macaque (*M. nemestrina*) and the Sulawesi macaques: *M. tonkeana*, *M. maura*, *M. ochreata*, *M. brunnescens*, *M. hecki*, *M. nigrescens*, *M. nigra* [4, 5]. Within *M. tonkeana*, individuals in the west and east are significantly differentiated from each other [6] and the latter population is also known as *M. togeanus* [7]. *Macaca nemestrina*, as recognized by Fooden (1975), has been divided into several species including representatives from the Sunda Region (Sumatra, Borneo, Peninsular Malaysia – *M. nemestrina*), from the Mentawai Islands (*M. siberu* and *M. pagensis*; [8]), and from the northern portion of their range (*M. leonina*) [9]. Macaques, along with tarsiers and humans, are the only primates that have dispersed across Wallace's Line.

Genetic samples

RADseq data was collected from 40 samples in total, including representatives of all species from Sulawesi: *M. brunnescens* (1 female), *M. hecki* (4 females, 2 males), *M. maura* (2 females, 4 males), *M. nigra* (2 females, 1 male), *M. nigrescens* (1 female, 1 male), *M. ochreata* (1 female, 2 males), *M. tonkeana* (1 female, 8 males), and

M. togeanus (1 female, 1 male), seven *M. nemestrina* individuals, including four from Borneo (1 female, 3 males), two from Sumatra (both female), and one from Peninsular Malaysia (a female), and one female *M. siberu* from Siberut Island in the Mentawai Archipelago.

Data

The RADseq dataset had a substantial amount of missing data in an alignment of all samples; for autosomal DNA there was an average of 52.2% missing data per taxon (range 42.0% – 78.1%) and for the X there was an average of 42.8% (range 24.1% – 84.1%). Three individuals had over 70% missing data in the X and in the autosomal data: *M. maura* PM613, *M. nigrescens* PF654, and *M. ochreata* PM596. A summary of sequence data statistics for each sample is provided in Supplementary Table S1.

Genotyping

A general description of our bioinformatics pipeline is presented in the Methods. To supplement this, we include below examples of the commandlines used for generating the initial genotype files for the WGS data.

```
#ALIGN:
{bwa} mem -M -t 10 {reference} {R1_fastq} {R2_fastq} | {samtools} view
    -Shb -o {outputDir}/{lane}.mem.bam -

#SORTBAM:
```

```
{java} -Djava.io.tmpdir={javaTmpDir} -Xmx24576m -jar
    {picard}/AddOrReplaceReadGroups.jar MAX_RECORDS_IN_RAM=2000000
    CREATE_INDEX=true SORT_ORDER=coordinate VALIDATION_STRINGENCY=SILENT
    I={outputDir}/{lane}.mem.bam O={lane_bam} RGID={lane} RGLB={sample}
    RGSM={sample} RGPU="Unknown" RGCN={center} RGDS="RefVersion:rheMac2"
    RGPL="illumina"

#MERGE:
{java} -Djava.io.tmpdir={javaTmpDir} -Xmx24576m -jar
    {picard}/MergeSamFiles.jar USE_THREADING=true
    MAX_RECORDS_IN_RAM=2000000 CREATE_INDEX=true
    VALIDATION_STRINGENCY=SILENT INPUT={lane_bam_merge_string}
    OUTPUT={outputDir}/{sample}.merged.bam
{java} -Djava.io.tmpdir={javaTmpDir} -XX:ParallelGCThreads=5 -Xmx24576m
    -jar {picard}/MarkDuplicates.jar MAX_RECORDS_IN_RAM=2000000
    VALIDATION_STRINGENCY=SILENT CREATE_INDEX=true
    M={outputDir}/{sample}.dedup.metrics I={outputDir}/{sample}.merged.bam
    O={outputDir}/{sample}.dedup.bam

#REALIGN_CREATOR:
{java} -Djava.io.tmpdir={javaTmpDir} -Xmx24576m -jar {gatk} -T
    RealignerTargetCreator --interval_padding 200 -rf BadCigar -nt 4 -R
    {reference} -I {outputDir}/{sample}.dedup.bam -o
    {outputDir}/{sample}.forRealigner.intervals
```

```
799
800 #REALIGN:
801 {java} -Djava.io.tmpdir={javaTmpDir} -Xmx24576m -jar {gatk} -T
802     IndelRealigner -dcov 1000 -rf BadCigar --consensusDeterminationModel
803     USE_READS -R {reference} -targetIntervals
804     {outputDir}/{sample}.forRealigner.intervals -I
805     {outputDir}/{sample}.dedup.bam -o {outputDir}/{sample}.realigned.bam
806
807
808 #BQSR:
809 {java} -Djava.io.tmpdir={javaTmpDir} -Xmx24576m -jar {gatk} -T
810     BaseRecalibrator --interval_padding 200 -rf BadCigar
811     --downsample_to_fraction 0.1 -nct 4 -R {reference} -I
812     {outputDir}/{sample}.realigned.bam -o {outputDir}/{sample}.recal.grp
813     -knownSites {bqsr_sites_file}
814 {java} -Djava.io.tmpdir={javaTmpDir} -Xmx24576m -jar {gatk} -T PrintReads
815     -nct 4 -rf BadCigar --disable_indel_quals --emit_original_quals -R
816     {reference} -I {outputDir}/{sample}.realigned.bam -o
817     {outputDir}/{sample}.final.bam -BQSR {outputDir}/{sample}.recal.grp
818
819
820 #HAPLOTYPECALLER:
821 {java} -XX:ParallelGCThreads=2 -Djava.io.tmpdir={javaTmpDir} -Xmx65536M
822     -jar {gatk} -T HaplotypeCaller --genotyping_mode DISCOVERY -A
```

```
    AlleleBalanceBySample -A DepthPerAlleleBySample -A DepthPerSampleHC -A
    InbreedingCoeff -A MappingQualityZeroBySample -A StrandBiasBySample -A
    Coverage -A FisherStrand -A HaplotypeScore -A
    MappingQualityRankSumTest -A MappingQualityZero -A QualByDepth -A
    RMSMappingQuality -A ReadPosRankSumTest -A VariantType -A
    StrandOddsRatio --emitRefConfidence GVCF -l INFO -rf BadCigar -R
    {reference} -nct 1 -I {outputDir}/{sample}.final.bam -o
    {outputDir}/{sample}_{intervalName}.g.vcf -L {intervalValue}

#COMPRESS/INDEX:
{bgzip} -f {outputDir}/{sample}_{intervalName}.g.vcf
{tabix} -f -p vcf {outputDir}/{sample}_{intervalName}.g.vcf.gz

#MERGE/EMIT:
{java} -XX:ParallelGCThreads=2 -Djava.io.tmpdir={javaTmpDir} -Xmx65536M
    -jar {gatk} org.broadinstitute.gatk.tools.CatVariants -R {reference}
    --assumeSorted {vcf_list} -o {outputDir}/{sample}.g.vcf.gz
{java} -XX:ParallelGCThreads=2 -Djava.io.tmpdir={javaTmpDir} -Xmx65536M
    -jar {gatk} -T GenotypeGVCFs -R {reference} --variant
    {outputDir}/{sample}.g.vcf.gz -o {outputDir}/{sample}.vcf

#COMPRESS/INDEX:
{bgzip} -f {outputDir}/{sample}.vcf
{tabix} -f -p vcf {outputDir}/{sample}.vcf.gz
```

Alignments for phylogenetic analysis

The alignment of autosomal RADseq data, including gapped positions, was 10,639,115 bp (out of a total genome size of about 3 Gigabases), and 91.6% of these positions were invariant. For the X, the alignment including gapped positions was 217,762 bp, and 92.7% of these were invariant.

Divergence

Based on the WGS data, genome-wide divergence per site between either of the Sulawesi species and the pigtailed macaque was 0.5% for autosomal DNA and 0.3% for the X. Divergence between *M. tonkeana* and *M. nigra* was 0.3% for the autosomes and 0.2% on the X. Similar to the results from analysis of the X/A polymorphism ratio discussed below, a higher level of divergence on the autosomes compared to the X is consistent with faster male evolution.

The X/A polymorphism ratio

In a constant sized population with equal variance in reproductive success between the sexes, the null expectation for the relative level of polymorphism on the X and autosomes is 0.75 [10]. We used two approaches to test this null hypothesis, with our analyses focused on each of the four species or populations for which we had data from at least four individuals in the RADseq data. This included *M. tonkeana* (9 individuals), *M. maura* (5 or 6 individuals depending on the analysis; see below),

M. hecki (6 individuals), and *M. nemestrina* from Borneo (4 individuals). We excluded data from the east population of *M. tonkeana* (i.e., *M. togeanus*) and the *M. nemestrina* populations from Sumatra and Peninsular Malaysia to reduce the effects of population subdivision on the results.

As described in Evans et al. 2014, the first method standardized the ratio of pairwise nucleotide diversity per site (π) on the X over that of the autosomes using the Jukes-Cantor corrected (1969) divergence from baboons for the X and autosomes, respectively, and included a correction for ancestral polymorphism as detailed in Charlesworth and Charlesworth (2010). Because this method required no missing data within a species in order for a site to be considered, we excluded *M. maura* individual PM613 due to low coverage. The second method estimated the X/A polymorphism ratio (λ) using a model of evolution that included the possibility of a dynamic demographic history and natural selection on GC content (and/or GC-biased gene conversion), as described elsewhere [11, 14–17]. Thus, $N_{eX}=\lambda N_{eA}$, where N_{eX} and N_{eA} are the effective population sizes of the X and autosomes, respectively, and λ is not influenced by differences between these genomic regions in mutation rate. This model allowed for missing data, so no individuals were excluded within each of these species or populations. The models to which the data from each species were fitted have several parameters, and include a 'full' model in which all parameters are estimated independently, and several nested models in which one or more parameters are fixed to some constant, or set to be equivalent to one another. The full model included two time intervals with different effective population sizes with instantaneous change between the ancestral and recent population size occurring τ

generations ago, and the ratio of the current to ancestral population size being equal to ρ. To account for the possibility that natural selection acted on GC content, or the equivalent genomic effects of GC-biased gene conversion, we fitted the data to a model of evolution between two types of nucleotides: those with a weak bond (adenosines and thymines) and those with a strong bond (guanines and cytosines). The polymorphism data were recoded to include only variable sites in which a guanine (G) or a cytosine (C) nucleotide was segregating with an adenine (A) or a thymine (T). In these models, the parameters θ_{01} and θ_{10} refer to the mutation parameters from G or C nucleotides to A or T nucleotides, and the reverse, respectively, as detailed in Evans et al. 2014. The parameter γ reflects whether GC-biased gene conversion (gBGC) or natural selection on GC content favors an increase in GC content (a positive parameter value) or a decrease in GC content (a negative parameter value). In the full model, γ is estimated separately for the autosomes (γ_A) and the X (γ_X), and in some of the nested models γ_A and γ_X are set to be equivalent and/or equal to zero, which corresponds to no gBGC or neutrality of GC content. The polymorphism data were also fitted to an equilibrium model in which population size is constant and for which there is no X/A polymorphism ratio (λ) parameter. More detailed information and the statistical rationale for these models are available elsewhere [11, 14–17]. If the equilibrium model was poorly supported, weighted parameter estimates were then calculated across all models using AIC weights, as described by Wagenmakers and Farrell 2004.

Low molecular polymorphism on the X can be explained by demography and natural selection

Fig. S2 depicts the X/A polymorphism ratio calculated from standardized π using RADseq data from four species, after separating the polymorphism data into categories based on genomic position relative to annotated genes in the rhesus genome. Additional polymorphism statistics are presented in Supplementary Tables S3 – S6. As expected, diversity and divergence was similar in *M. tonkeana* to that previously reported based on an expanded dataset that included paired-end sequences [11], even though there were differences in the bioinformatic analyses of each study.

In the four species in which we assessed population genetic variation, *M. nemestrina* from Borneo was the most polymorphic. The 95% CIs for π and θ_{W} overlapped for the three Sulawesi species with population genetic data from at least 4 individuals (*M. tonkeana*, *M. maura*, and *M. hecki*). In genomic regions far from genes, which presumably are least affected by natural selection, the X/A polymorphism ratio was lower than expectations (Fig. S2). However, in these genomic regions Tajima's D of autosomal DNA was significantly negative (Tables S3 – S6), indicating an excess of low frequency polymorphisms. This could stem from population expansion, although in *M. maura*, the 95% CI for this parameter was near zero suggesting population size of that species may have varied less than that of the others.

That Tajima's D is significantly different from zero provides circumstantial evidence for a dynamic demography in at least some of these species, and changes in population size are known to influence the X/A polymorphism ratio [19]. Addition-

ally, other factors such as gBGC or natural selection on GC content have the potential to affect the X/A polymorphism ratio (e.g., Evans et al. 2014). For these reasons, we fitted the polymorphism data from genomic regions far (>51,000 bp) from genes to several models of evolution to polymorphism data from the X and autosome for each of the four populations or species for which we had data from >3 individuals. For the three species where the equilibrium model (with no change in population size over time) was not supported, the weighted average of parameter estimates over all models based on AIC weights are presented in Table S11. Parameter estimates for each model for each species or population are presented in Supplemental Tables S7 – S9.

M. maura was unusual among the 4 populations we tested because the equilibrium model was provided a relatively good fit to the data. This is illustrated by the AIC weight for the equilibrium model (∞) in Table S9 that is over twice as high as that of any other model. The other three populations/species each had evidence for a dynamic demography and zero weight for the equilibrium model.

For all species, gBGC and/or selection favoring increased GC content is supported because the the maximum likelihood estimates and/or model averages for the gBGC parameters on the autosomes and the X, γ_A and γ_X respectively, are greater than zero. This indicates that gBGC and/or selection for GC content favor an increase in GC content (Table S11). Moreover, all models that set γ_A or γ_X equal to zero had low AIC weights (Tables S7 – S9). If the strength of gBGC and/or selection favoring increased GC is similar on the X and autosomes, we expect $\gamma_X = \lambda\gamma_A$; for all four species/populations the AIC weights of these models were high compared

to the other models that relaxed this constraint, suggesting that the forces driving GC-biased molecular evolution in these genomic regions were indeed similar. This latter finding was also recovered previously for *M. tonkeana* [11] using an expanded dataset from that species that also analyzed paired end sequences from RADseq.

In this analysis, variable positions are re-coded into two states, A_0 and A_1, where A_0 refers to G or C nucleotides and A_1 refers to A or T nucleotides (see Methods). Additionally, $\theta_{01} = 4N_e\mu_{01}$, where μ_{01} represents the mutation rate from $A_0 \rightarrow A_1$, and $\theta_{10} = 4N_e\mu_{10}$, where μ_{10} represents the mutation rate from $A_1 \rightarrow A_0$. In all species, $\theta_{01} > \theta_{10}$ for the X and for the autosomes (Table S11). This suggests that in each of these different species there are more variable positions in which a G or a C is the major (more common) allele than where an A or a T is the major allele. Also of interest is the observation that in each species, $\theta_{ijA}\ \lambda/\theta_{ijX} > 1$, where ij is 01 or 10 and A and X refer to the autosomes and X respectively. This indicates that the mutation rate is higher in the autosomes than in the X, which is suggestive of male driven evolution – a result that is also suggested by pairwise divergence between the three species for which we performed WGS as described below.

Thus, while we recovered significantly lower polymorphism on the X than expected based on π in four macaque species, in each one this could be accounted for by an evolutionary model that includes a dynamic demography and selection on gBGC/natural selection on GC content. One factor not incorporated in these analyses and that is beyond the scope of this study is the possibility that hybridization among species via male dispersal could influence the X/A ratio. As discussed above, most papionin monkeys have female philopatry and hybridization has been detected

between all parapatric species pairs on Sulawesi. If hybridization were mostly mediated by male migration, diversity in the autosomes would be increased to a greater extent than the X. This possibility could explain the lower (but not significantly so) levels of diversity on the X. Added to this, other factors such as natural selection in males on deleterious recessive mutations, could also decrease diversity on the X.

References

[1] Fooden, J. Provisional classification and key to living species of macaques (Primates: *Macaca*). *Folia primatologica*, **25**, (1976) 225–236.

[2] Delson, E. Fossil macaques, phyletic relationships and a scenario of deployment. *The Macaques: Studies in ecology, behavior and evolution*, (1980) 10–30.

[3] Tosi, A. J., Morales, J. C., and Melnick, D. J. Paternal, maternal, and biparental molecular markers provide unique windows onto the evolutionary history of macaque monkeys. *Evolution*, **57**, (2003) 1419–1435.

[4] Fooden, J. Taxonomy and evolution of liontail and pigtail macaques (Primates : Cercopithecidae). *Fieldiana Zoology*, **67**, (1975) 1–169.

[5] Fooden, J. *Taxonomy and evolution of the monkeys of Celebes* Karger Basel, 1969.

[6] Evans, B., Supriatna, J., and Melnick, D. Hybridization and population genetics of two macaque species in Sulawesi, Indonesia. *Evolution*, **55**, (2001) 1686–1702.

[7] Froehlich, J. W. and Supriatna, J. Secondary intergradation between *Macaca maurus* and *M. tonkeana* in South Sulawesi, and the species status of *M. togeanus*. *Evolution and ecology of macaque societies*, (1996) 43–70.

[8] Roos, C., Ziegler, T., Hodges, J. K., Zischler, H., and Abegg, C. Molecular phylogeny of Mentawai macaques: taxonomic and biogeographic implications. *Molecular Phylogenetics and Evolution*, **29**, (2003) 139–150.

[9] Groves, C. P. *Primate taxonomy* Smithsonian Books, 2001.

[10] Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, **10**, (2009) 195–205.

[11] Evans, B. J., Zeng, K., Esselstyn, J. A., Charlesworth, B., and Melnick, D. J. Reduced representation genome sequencing suggests low diversity on the sex chromosomes of tonkean macaque monkeys. *Molecular Biology and Evolution*, **31**, (Sept. 2014) 2425–2440.

[12] Jukes, T. H. and Cantor, C. R. Evolution of protein molecules. *Mammalian Protein Metabolism*, **3**, (1969) 132.

[13] Charlesworth, B. and Charlesworth, D. *Elements of Evolutionary Genetics* Roberts Publishers, 2010.

[14] Haddrill, P. R., Zeng, K., and Charlesworth, B. Determinants of synonymous and nonsynonymous variability in three species of Drosophila. *Molecular Biology and Evolution*, **28**, (2011) 1731–1743.

[15] Zeng, K. and Charlesworth, B. Studying patterns of recent evolution at synonymous sites and intronic sites in Drosophila melanogaster. *Journal of Molecular Evolution*, **70**, (2010) 116–128.

[16] Zeng, K. and Charlesworth, B. The joint effects of background selection and genetic recombination on local gene genealogies. *Genetics*, **189**, (2011) 251–266.

[17] Zeng, K. and Charlesworth, B. Estimating selection intensity on synonymous codon usage in a nonequilibrium population. *Genetics*, **183**, (2009) 651–662.

[18] Wagenmakers, E.-J. and Farrell, S. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, **11**, (2004) 192–196.

[19] Pool, J. E. and Nielsen, R. Population size changes reshape genomic patterns of diversity. *Evolution*, **61**, (2007) 3001–3006.

Supplementary Tables and Figures

Table S1: Information on sequence data analyzed in this study including the species (Species), sample identification number (SampleID), sex (Sex), number of reads before and after trimming (Untrimmed and Trimmed, respectively), read length after trimming if performed or before trimming for the HiSeqX data (Readlength), GC content (GC), and the number of mapped reads (mapped). For the HiSeqX data, trimming was not performed (np). Mapped reads were computed for RADseq and HiSeqX data by the flagstat command of SAMTOOLS and GATK, respectively.

Species	SampleID	Sex	Untrimmed	Trimmed	Readlength	GC	mapped
RADseq							
M. nemestrina	Gumgum	F	3609945	3503070	36-75	52	2413976
M. nemestrina	Kedurang	F	2811747	2686390	36-75	52	1837733
M. nemestrina	Malay	F	2888661	2820801	36-75	52	1926671
M. nemestrina	Ngasang	F	1863406	1819857	36-75	52	1250488
M. nemestrina	PM664	M	4794910	4720761	36-75	52	3242719
M. nemestrina	PM665	M	8785118	2447219	36-75	53	1666291
M. nemestrina	Sukai	M	2561836	2505653	36-75	53	1777480
M. siberu	pagensis	F	2374748	2315204	36-75	53	1602553
M. nigra	PF1001	F	1103174	1726287	36-75	53	1193708
M. nigra	PF660	F	1767769	1080813	36-75	53	744724
M. nigra	PM1003	M	2709319	2637075	36-75	53	1847731
M. nigrescens	PF654	F	3259548	911105	36-75	53	472852
M. nigrescens	PM1000	M	2860128	2778827	36-75	53	1935144
M. hecki	PF643	F	5770315	5628031	36-75	53	3888077
M. hecki	PF644	F	9630884	9306553	36-75	54	6627357
M. hecki	PF648	F	4852016	4768222	36-75	53	3322875
M. hecki	PF651	F	5038203	5035540	36-75	53	3509593
M. hecki	PM639	M	5269803	5154271	36-75	53	3551944
M. hecki	PM645	M	5855161	5746693	36-75	53	3974092
M. maura	PF615	F	5625181	5532103	36-75	53	3815996
M. maura	PF713	F	10698886	10557589	36-75	54	7524686
M. maura	PM613	M	935933	921611	36-75	53	637370
M. maura	PM614	M	4160191	4058140	36-75	53	2843009
M. maura	PM616	M	6667369	6526349	36-75	53	4530947
M. maura	PM618	M	4285281	4160860	36-75	54	2969422
M. tonkeana	PF515	F	11982906	11702493	36-75	53	8148982
M. tonkeana	PM561	M	11624883	11309233	36-75	53	7841954
M. tonkeana	PM565	M	12132625	11852221	36-75	53	8241475
M. tonkeana	PM566	M	9921261	9688908	36-75	53	6712122
M. tonkeana	PM567	M	11892242	11561026	36-75	53	8033420
M. tonkeana	PM582	M	12296341	11967241	36-75	54	8439642
M. tonkeana	PM584	M	12812865	12442854	36-75	53	8654276
M. tonkeana	PM592	M	13843609	13368052	36-75	54	9412709
M. tonkeana	PM602	M	13737591	13373987	36-75	54	9521922

Continued on next page

Continued from previous page

Species	SampleID	Sex	Untrimmed	Trimmed	Readlength	GC	mapped
M. togeanus	PF549	F	1096486	1070693	36-75	53	741166
M. togeanus	PM545	M	1783072	1762114	36-75	53	1239608
M. ochreata	PF625	F	2164485	2128202	36-75	53	1477886
M. ochreata	PM571	M	1687457	1667269	36-75	53	1169513
M. ochreata	PM596	M	2414864	2124502	36-75	54	685724
M. brunnescens	PF707	F	2277761	2217043	36-75	53	1567209
WGS							
M. nemestrina	PM664	M	949729200	np	151	41	927910917
M. nigra	PF660	F	916229358	np	151	41	894519506
M. tonkeana	PM592	M	913316498	np	151	41	892349062

Table S2: Gene flow statistics for the X chromosome calculated from data after different filtering steps, and using different outgroup taxa. The first panel indicates the f_{DM} statistic with confidence intervals in parentheses. The second panel indicates the number of ABBA, BABA, and BBAA sites in each analysis, with the definition of these site patterns provided in the main text. Each panel is divided into rows based on whether sites with heterozygous diploid genotypes were included and called as a haploid genotype based on depth (included), excluded if either or both males had a heterozygous diploid genotype (no male hets), or excluded if any individual had a heterozygous diploid genotype (no hets). For each of these filters, we calculated gene flow statistics after removing sites where any genotype had <5X or <12X coverage (5X or 12X respectively), using a rhesus or baboon outgroup (rhesus or baboon, respectively), and/or using a haploid genotype for all individuals or diploid genotype for the female all (all haploid or female diploid, respectively). In the top panel, asterisks indicate f_{DM} statistics whose confidence intervals are higher than zero, which is consistent with gene flow between *M. nemestrina* and *M. tonkeana*. In general, less data are considered by the analyses on the left and bottom of each panel.

	5X, rhesus, all haploid	5X, rhesus, female diploid	5X, baboon, all haploid	12X, rhesus, all haploid	12X, baboon, all haploid
f_{DM}					
included	0.30492 (0.16752 - 0.44232)*	0.38206 (0.28071 - 0.48342)*	0.16879 (0.07368 - 0.26390)*	0.31821 (0.17475 - 0.46167)*	0.16973 (0.07373 - 0.26573)*
no male hets	0.07517 (0.02101 - 0.12933)*	0.19022 (0.12465 - 0.25579)*	0.03392 (-0.01003 - 0.07788)	0.04351 (-0.00781 - 0.09483)	0.03408 (-0.00984 - 0.07801)
no hets	0.04690 (-0.01601 - 0.10980)	0.04726 (-0.01560 - 0.11012)	-0.01926 (-0.07222 - 0.03369)	0.04733 (-0.00489 - 0.09954)	0.03851 (-0.00474 - 0.08176)
ABBA;BABA;BBAA					
included	3781; 2014; 40304	4156.5; 2401.5; 40312	3123; 2221; 30999	3457; 1788; 36736	2898; 2057; 28566
no male hets	1938; 1667; 39825	2104.5; 1725; 39722.5	1905; 1780; 30349	1523; 1396; 35709	1608; 1502; 27426
no hets	1518; 1382; 38538	1518; 1381; 38538	1451; 1508; 29363	1527; 1389; 35705	1618; 1498; 27424

Table S3: Borneo *M. nemestrina* polymorphism ($n = 1$ female, 3 males). Data are divided into three categories based on their position relative to genes, including positions spanning genes ± 1000 bp in both directions (plusminus), positions 1000–51000 bp from genes (1000to51000), and positions >51000 bp from genes (51000plus). Statistics include the number of sites genotyped (Sites), the number of RAD tags (RADtags), the number of segregating sites (S), Watterson's θ (θ_{W}), pairwise nucleotide diversity (π), divergence from humans with Jukes-Cantor correction and correction for ancestral polymorphism (divergence), Tajima's D (TajD), and the number of singleton sites over the number of segregating sites (S_{e}/S). 95% confidence intervals are in parentheses.

Statistic	plusminus	1000to50000	51000plus
aDNA			
*S*ites	424889	756435	2471619
RADtags	4614	8242	27524
S	2862 (2759 - 2966)	5311 (5162 - 5463)	19342 (19048 - 19627)
θ_{W}	0.00260 (0.00250 - 0.00269)	0.00271 (0.00263 - 0.00279)	0.00302 (0.00297 - 0.00306)
π	0.00242 (0.00233 - 0.00251)	0.00254 (0.00246 - 0.00262)	0.00281 (0.00277 - 0.00285)
divergence	0.06065 (0.05989 - 0.06141)	0.06363 (0.06303 - 0.06421)	0.06637 (0.06601 - 0.06668)
π/divergence	0.040 (0.03829 - 0.04159)	0.040 (0.03869 - 0.04121)	0.042 (0.04165 - 0.04304)
TajD	-0.376 (-0.441 - -0.313)	-0.339 (-0.384 - -0.294)	-0.381 (-0.406 - -0.355)
$S\mathrm{e}/S$	0.507 (0.489 - 0.526)	0.502 (0.488 - 0.516)	0.519 (0.512 - 0.526)
xDNA			
*S*ites	4803	10354	33621
RADtags	68	139	455
S	11 (5 - 18)	21 (13 - 31)	78 (61 - 96)
θ_{W}	0.00125 (0.00057 - 0.00204)	0.00111 (0.00068 - 0.00163)	0.00127 (0.00099 - 0.00156)
π	0.00118 (0.00056 - 0.00194)	0.00105 (0.00063 - 0.00153)	0.00124 (0.00097 - 0.00152)
divergence	0.05906 (0.05166 - 0.06691)	0.05166 (0.04712 - 0.05635)	0.06343 (0.06071 - 0.06622)
π/divergence	0.020 (0.00924 - 0.03310)	0.020 (0.01208 - 0.03062)	0.020 (0.01507 - 0.02428)
TajD	-0.558 (-0.847 - 0.083)	-0.557 (-0.857 - -0.117)	-0.216 (-0.489 - 0.082)
$S\mathrm{e}/S$	0.909 (0.727 - 1.000)	0.905 (0.762 - 1.000)	0.795 (0.705 - 0.872)

Table S4: *M. tonkeana* polymorphism (n = 1 female, 8 males). Labels follow Table S3.

Statistic	plusminus	1000to50000	51000plus
aDNA			
*S*ites	620178	1076052	3606494
RADtags	5312	9329	31803
S	4650 (4520 - 4783)	8385 (8205 - 8572)	31344 (31000 - 31691)
θ_W	0.00218 (0.00212 - 0.00224)	0.00227 (0.00222 - 0.00232)	0.00253 (0.00250 - 0.00255)
π	0.00164 (0.00158 - 0.00169)	0.00172 (0.00167 - 0.00176)	0.00190 (0.00188 - 0.00193)
divergence	0.06114 (0.06050 - 0.06175)	0.06425 (0.06376 - 0.06474)	0.06672 (0.06646 - 0.06699)
π/divergence	0.027 (0.02582 - 0.02779)	0.027 (0.02602 - 0.02748)	0.029 (0.02813 - 0.02893)
TajD	-1.069 (-1.125 - -1.009)	-1.034 (-1.076 - -0.990)	-1.056 (-1.079 - -1.034)
*S*e/*S*	0.512 (0.498 - 0.526)	0.492 (0.481 - 0.503)	0.502 (0.496 - 0.507)
xDNA			
*S*ites	14076	24685	95369
RADtags	127	224	892
S	41 (29 - 54)	44 (31 - 57)	253 (225 - 285)
θ_W	0.00107 (0.00076 - 0.00141)	0.00066 (0.00046 - 0.00085)	0.00098 (0.00087 - 0.00110)
π	0.00094 (0.00066 - 0.00126)	0.00058 (0.00040 - 0.00076)	0.00085 (0.00074 - 0.00096)
divergence	0.05448 (0.05059 - 0.05896)	0.05228 (0.04931 - 0.05522)	0.05786 (0.05633 - 0.05961)
π/divergence	0.017 (0.01197 - 0.02328)	0.011 (0.00757 - 0.01474)	0.015 (0.01283 - 0.01663)
TajD	-0.626 (-1.131 - -0.087)	-0.565 (-1.116 - -0.004)	-0.691 (-0.895 - -0.484)
*S*e/*S*	0.585 (0.439 - 0.732)	0.591 (0.455 - 0.727)	0.565 (0.506 - 0.628)

Table S5: *M. maura* polymorphism ($n = 2$ females, 3 males). Labels follow Table S1.

Statistic	plusminus	1000to50000	51000plus
aDNA			
*S*ites	570287	1007276	3332818
RADtags	5198	9220	31199
S	2530 (2431 - 2643)	4469 (4345 - 4598)	16465 (16219 - 16718)
θ_W	0.00157 (0.00151 - 0.00164)	0.00157 (0.00152 - 0.00161)	0.00175 (0.00172 - 0.00177)
π	0.00156 (0.00149 - 0.00163)	0.00150 (0.00146 - 0.00155)	0.00169 (0.00166 - 0.00172)
divergence	0.06197 (0.06130 - 0.06267)	0.06494 (0.06440 - 0.06549)	0.06756 (0.06726 - 0.06784)
π/divergence	0.025 (0.02404 - 0.02635)	0.023 (0.02234 - 0.02385)	0.025 (0.02460 - 0.02546)
TajD	-0.032 (-0.111 - 0.043)	-0.214 (-0.270 - -0.157)	-0.161 (-0.192 - -0.127)
*S*e/*S*	0.397 (0.379 - 0.417)	0.442 (0.427 - 0.457)	0.423 (0.415 - 0.431)
xDNA			
*S*ites	9447	18642	65077
RADtags	101	198	760
S	15 (8 - 23)	30 (20 - 42)	101 (81 - 121)
θ_W	0.00076 (0.00041 - 0.00117)	0.00077 (0.00051 - 0.00108)	0.00074 (0.00060 - 0.00089)
π	0.00070 (0.00036 - 0.00108)	0.00076 (0.00049 - 0.00109)	0.00073 (0.00058 - 0.00087)
divergence	0.05305 (0.04828 - 0.05875)	0.05086 (0.04763 - 0.05426)	0.05987 (0.05803 - 0.06186)
π/divergence	0.013 (0.00677 - 0.02132)	0.015 (0.00971 - 0.02162)	0.012 (0.00974 - 0.01463)
TajD	-0.609 (-1.200 - 0.132)	-0.104 (-0.674 - 0.449)	-0.201 (-0.475 - 0.103)
*S*e/*S*	0.800 (0.600 - 1.000)	0.633 (0.467 - 0.800)	0.663 (0.574 - 0.752)

Table S6: *M. hecki* polymorphism ($n = 4$ females, 2 males). Labels follow Table S1.

Statistic	plusminus	1000to50000	51000plus
aDNA			
*S*ites	571933	999066	3375989
RADtags	5187	9182	31368
S	2957 (2851 - 3072)	5264 (5118 - 5395)	20390 (20110 - 20658)
θ_W	0.00171 (0.00165 - 0.00178)	0.00174 (0.00170 - 0.00179)	0.00200 (0.00197 - 0.00203)
π	0.00145 (0.00139 - 0.00151)	0.00149 (0.00144 - 0.00153)	0.00175 (0.00172 - 0.00178)
divergence	0.06186 (0.06119 - 0.06253)	0.06493 (0.06441 - 0.06545)	0.06735 (0.06705 - 0.06765)
π/divergence	0.023 (0.02240 - 0.02440)	0.023 (0.02221 - 0.02364)	0.026 (0.02554 - 0.02640)
TajD	-0.736 (-0.807 - -0.670)	-0.688 (-0.745 - -0.636)	-0.592 (-0.619 - -0.564)
Se/S	0.524 (0.506 - 0.542)	0.508 (0.495 - 0.521)	0.489 (0.482 - 0.496)
xDNA			
*S*ites	11500	21152	83301
RADtags	115	211	870
S	13 (7 - 21)	28 (18 - 38)	144 (121 - 168)
θ_W	0.00050 (0.00027 - 0.00080)	0.00058 (0.00037 - 0.00079)	0.00076 (0.00064 - 0.00088)
π	0.00048 (0.00023 - 0.00076)	0.00054 (0.00035 - 0.00076)	0.00071 (0.00060 - 0.00083)
divergence	0.05360 (0.04946 - 0.05778)	0.05168 (0.04858 - 0.05493)	0.05892 (0.05717 - 0.06061)
π/divergence	0.009 (0.00430 - 0.01429)	0.010 (0.00666 - 0.01511)	0.012 (0.01012 - 0.01412)
TajD	-0.244 (-1.069 - 0.668)	-0.410 (-0.953 - 0.243)	-0.404 (-0.640 - -0.130)
Se/S	0.615 (0.385 - 0.846)	0.643 (0.464 - 0.821)	0.653 (0.569 - 0.729)

Table S7: Parameter estmates from model fitting for *M. nemestrina* from Borneo ($n = 4$) polymorphism data. Models indicated with abbreviations and symbols that are defined in the Methods and discussed in detail in Evans et al. 2014. $\bar{x}$ indicates weighted average of parameter values.

model	θ_{01X}	θ_{10X}	γ_X	θ_{01A}	θ_{10A}	γ_A	λ	ρ_1	τ_1	$\ln L$	δAIC	wAIC
One Epoch												
∞	0.00195	0.00058	1.13186	0.00439	0.00114	1.44463	-	-	-	-3050429.949	357.59	0.000
Two Epochs												
full	0.00137	0.00049	0.93517	0.00253	0.00078	1.26642	2.313	2.359	0.652	-3050249.900	3.49	0.068
$\gamma_A = 0$	0.00125	0.00044	0.95318	0.00004	0.00004	0(fixed)	95.451	100.000	0.898	-3051716.354	2934.40	0.000
$\gamma_X = 0.75\gamma_A$	0.00106	0.00037	$0.75\gamma_A$	0.00254	0.00079	1.26558	0.75(fixed)	2.348	0.645	-3050250.156	0.00	0.386
$\gamma_X = 0$	0.00089	0.00080	0(fixed)	0.00253	0.00078	1.26648	2.405	2.360	0.653	-3050252.877	7.44	0.009
$\gamma_X = \lambda\gamma_A$	0.00115	0.00036	$\lambda\gamma_A$	0.00254	0.00078	1.26540	0.838	2.350	0.646	-3050250.108	1.90	0.149
$\lambda = 0.75$	0.00110	0.00036	1.02612	0.00254	0.00078	1.26545	0.75(fixed)	2.349	0.646	-3050250.139	1.96	0.145
$\theta_{01A} = \theta_{10A}$	0.00128	0.00046	0.91700	θ_{10A}	0.00144	0.08210	1.632	2.788	0.436	-3051539.332	2580.35	0.000
$\theta_{01X} = \lambda\theta_{10A}$	$\lambda\theta_{01A}$	0.00030	1.11559	0.00254	0.00078	1.26551	0.400	2.349	0.646	-3050250.332	2.35	0.119
$\theta_{01X} = \theta_{10X}$	θ_{10X}	0.00084	-0.10042	0.00253	0.00078	1.26641	2.356	2.359	0.652	-3050253.556	8.80	0.005
$\theta_{10X} = \lambda\theta_{10A}$	0.00099	$\lambda\theta_{10A}$	1.06700	0.00254	0.00078	1.26566	0.394	2.349	0.646	-3050250.336	2.36	0.119
$\bar{x}$	0.00108	0.00037	0.99594	0.00254	0.00078	1.26560	0.808	2.349	0.646			

Table S8: Parameter estimates from model fitting of *M. tonkeana* data ($n = 9$). Labels follow Table S7.

model	θ_{01X}	θ_{10X}	γ_X	θ_{01A}	θ_{10A}	γ_A	λ	ρ_1	τ_1	$\ln L$	δAIC	wAIC
One Epoch												
∞	0.00139	0.00055	0.83380	0.00357	0.00105	1.28046	-	-	-	-4171971.390	4026.82	0.000
Two Epochs												
full	0.00072	0.00032	0.72472	0.00201	0.00077	1.02231	0.501	3.932	0.134	-4169956.509	3.06	0.075
$\gamma_A = 0$	0.00066	0.00029	0.73049	0.00117	0.00123	0(fixed)	0.524	4.565	0.140	-4171922.702	3933.45	0.000
$\gamma_X = 0.75\gamma_A$	0.00084	0.00035	$0.75\gamma_A$	0.00201	0.00077	1.02205	0.75(fixed)	3.925	0.135	-4169956.977	0.00	0.345
$\gamma_X = 0$	0.00052	0.00047	0(fixed)	0.00201	0.00077	1.02223	0.517	3.934	0.134	-4169963.181	14.41	0.000
$\gamma_X = \lambda\gamma_A$	0.00075	0.00036	$\lambda\gamma_A$	0.00201	0.00077	1.02267	0.630	3.928	0.135	-4169956.696	1.44	0.168
$\lambda = 0.75$	0.00081	0.00036	0.69784	0.00201	0.00077	1.02243	0.75(fixed)	3.925	0.135	-4169956.913	1.87	0.135
$\theta_{01A} = \theta_{10A}$	0.00066	0.00029	0.73020	θ_{10A}	0.00121	0.04892	0.510	4.544	0.136	-4171741.885	3571.81	0.000
$\theta_{01X} = \lambda\theta_{10A}$	$\lambda\theta_{01A}$	0.00027	0.84651	0.00202	0.00077	1.02178	0.339	3.940	0.133	-4169957.061	2.17	0.117
$\theta_{01X} = \theta_{10X}$	θ_{10X}	0.00048	-0.10748	0.00201	0.00077	1.02219	0.497	3.936	0.134	-4169965.323	18.69	0.000
$\theta_{10X} = \lambda\theta_{10A}$	0.00067	$\lambda\theta_{10A}$	0.69695	0.00202	0.00077	1.02264	0.390	3.937	0.133	-4169956.751	1.55	0.159
$\bar{x}$	0.00077	0.00033	0.73158	0.00201	0.00077	1.02229	0.606	3.930	0.134			

Table S9: Parameter estimates from model fitting of *M. maura* ($n = 6$) data. Labels follow Table S7.

model	θ_{01X}	θ_{10X}	γ_X	θ_{01A}	θ_{10A}	γ_A	λ	ρ_1	τ_1	$\ln L$	δAIC	wAIC
One Epoch												
∞	0.00104	0.00039	0.87006	0.00233	0.00089	1.05198	-	-	-	-3034844.889	0.00	0.477
Two Epochs												
full	0.00097	0.00038	0.83765	0.00203	0.00078	1.04255	14.460	1.157	5.000	-3034844.676	5.57	0.029
$\gamma_A = 0$	0.00029	0.00008	1.16941	0.00023	0.00025	0(fixed)	2.352	6.956	2.249	-3035631.009	1576.24	0.000
$\gamma_X = 0.75\gamma_A$	0.00100	0.00041	$0.75\gamma_A$	0.00234	0.00089	1.05618	0.75(fixed)	0.010	0.003	-3034844.769	1.76	0.198
$\gamma_X = 0$	0.00066	0.00060	0(fixed)	0.00201	0.00078	1.04193	14.455	1.168	5.000	-3034847.515	9.25	0.005
$\gamma_X = \lambda\gamma_A$	0.00104	0.00039	$\lambda\gamma_A$	0.00234	0.00089	1.05588	0.832	0.010	0.004	-3034844.741	3.70	0.075
$\lambda = 0.75$	0.00102	0.00039	0.86792	0.00229	0.00088	1.04768	0.75(fixed)	1.022	1.223	-3034844.839	3.90	0.068
$\theta_{01A} = \theta_{10A}$	0.00090	0.00037	0.79876	θ_{10A}	0.00118	0.08340	3.370	1.416	1.095	-3035506.559	1327.34	0.000
$\theta_{01X} = \lambda\theta_{10A}$	$\lambda\theta_{01A}$	0.00039	0.87835	0.00234	0.00089	1.05552	0.446	0.010	0.003	-3034844.767	3.76	0.073
$\theta_{01X} = \theta_{10X}$	θ_{10X}	0.00063	-0.10469	0.00202	0.00078	1.04236	14.816	1.161	5.000	-3034848.278	10.78	0.002
$\theta_{10X} = \lambda\theta_{10A}$	0.00104	$\lambda\theta_{10A}$	0.87110	0.00234	0.00089	1.05555	0.443	0.010	0.003	-3034844.767	3.76	0.073
$\bar{x}$	0.00103	0.00040	0.84863	0.00232	0.00089	1.05298	0.851	0.116	0.266			

Table S10: Parameter estimates from model fitting of *M. hecki* data ($n = 6$). Labels follow Table S7.

model	θ_{01X}	θ_{10X}	γ_X	θ_{01A}	θ_{10A}	γ_A	λ	ρ_1	τ_1	$\ln L$	δAIC	wAIC
One Epoch												
∞	0.00098	0.00045	0.68700	0.00294	0.00089	1.28375	-	-	-	-3073983.524	744.36	0.000
Two Epochs												
full	0.00067	0.00034	0.57742	0.00213	0.00078	1.09147	0.502	2.526	0.132	-3073609.850	3.01	0.063
$\gamma_A = 0$	0.00064	0.00032	0.57000	0.00121	0.00131	0(fixed)	0.614	2.890	0.155	-3074877.126	2535.57	0.000
$\gamma_X = 0.75\gamma_A$	0.00081	0.00032	$0.75\gamma_A$	0.00213	0.00078	1.09019	0.75(fixed)	2.517	0.134	-3073610.343	0.00	0.287
$\gamma_X = 0$	0.00051	0.00046	0(fixed)	0.00213	0.00078	1.09144	0.502	2.528	0.132	-3073612.043	5.40	0.019
$\gamma_X = \lambda\gamma_A$	0.00068	0.00035	$\lambda\gamma_A$	0.00213	0.00078	1.09148	0.521	2.525	0.132	-3073609.852	1.02	0.172
$\lambda = 0.75$	0.00073	0.00037	0.57960	0.00213	0.00078	1.09133	0.75(fixed)	2.517	0.134	-3073609.974	1.26	0.153
$\theta_{01A} = \theta_{10A}$	0.00064	0.00032	0.56993	θ_{10A}	0.00127	0.08219	0.595	2.895	0.150	-3074690.970	2163.25	0.000
$\theta_{01X} = \lambda\theta_{10A}$	$\lambda\theta_{01A}$	0.00029	0.64843	0.00214	0.00078	1.09148	0.287	2.542	0.129	-3073610.134	1.58	0.130
$\theta_{01X} = \theta_{10X}$	θ_{10X}	0.00048	-0.10659	0.00213	0.00078	1.09144	0.495	2.528	0.132	-3073612.931	7.18	0.008
$\theta_{10X} = \lambda\theta_{10A}$	0.00064	$\lambda\theta_{10A}$	0.56548	0.00213	0.00078	1.09166	0.422	2.531	0.131	-3073609.877	1.07	0.168
$\bar{x}$	0.00070	0.00034	0.63586	0.00213	0.00078	1.09112	0.573	2.525	0.133			

Table S11: Weighted average of parameter estimates over all models for *M. nemestrina* from Borneo, *M. hecki*, and *M. tonkeana*. Model averages are not reported for *M. maura* because the equilibrium model was not rejected and the λ parameter therefore could not be estimated (Supplementary Table S9). As described in the Supplement, $\theta_{01\mathrm{X}}$ is the polymorphism parameter for sites on the X where the derived mutation is an A or T, $\theta_{10\mathrm{X}}$ is the polymorphism parameter for sites on the X where the derived mutation is an G or C, γ_{X} is the selection parameter for GC content on the X. $\theta_{01\mathrm{A}}$, $\theta_{10\mathrm{A}}$, and γ_{A} are the corresponding parameters for the autosomes. λ, ρ_1, and τ_1 refer respectively to the X/A polymorphism ratio, the ratio of the current to ancestral population size, and number of generations before the present that the population size changed

Species	$\theta_{01\mathrm{X}}$	$\theta_{10\mathrm{X}}$	γ_{X}	$\theta_{01\mathrm{A}}$	$\theta_{10\mathrm{A}}$	γ_{A}	λ	ρ_1	τ_1
M. nemestrina Borneo	0.00108	0.00037	0.99594	0.00254	0.00078	1.26560	0.808	2.349	0.646
M. hecki	0.00070	0.00034	0.63586	0.00213	0.00078	1.09112	0.573	2.525	0.133
M. tonkeana	0.00077	0.00033	0.73158	0.00201	0.00077	1.02229	0.606	3.930	0.134

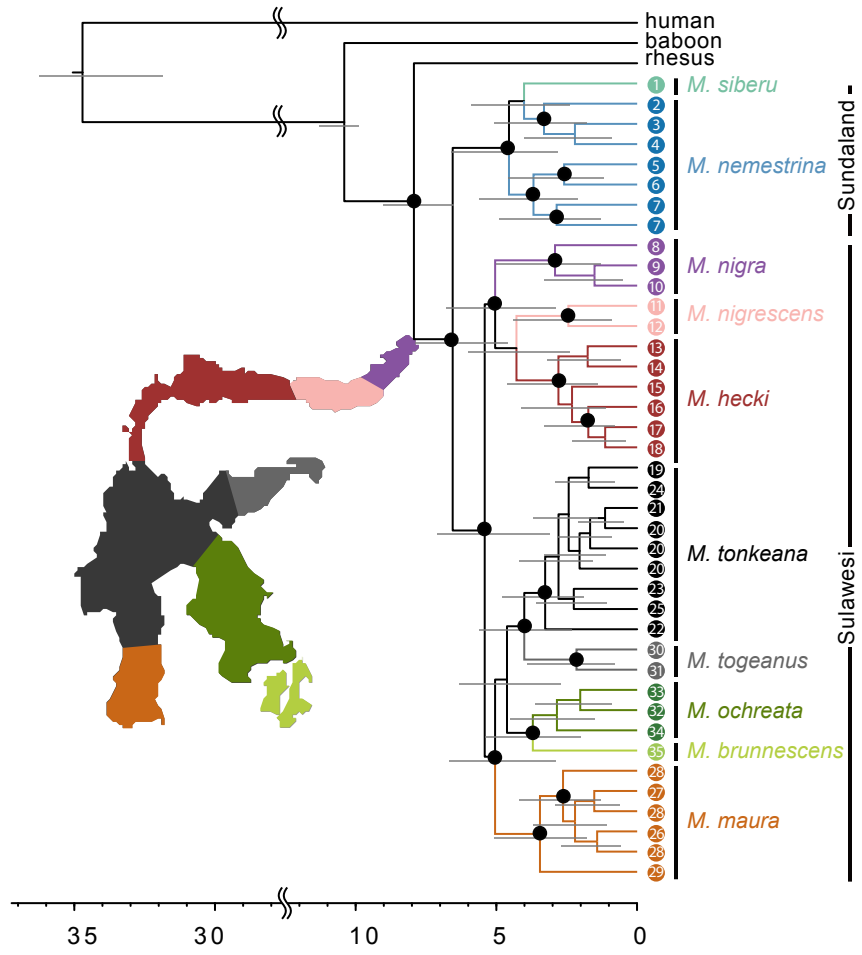

Figure S1: Chronogram estimated from RADseq data from the X chromosome. Labels follow Fig 2.

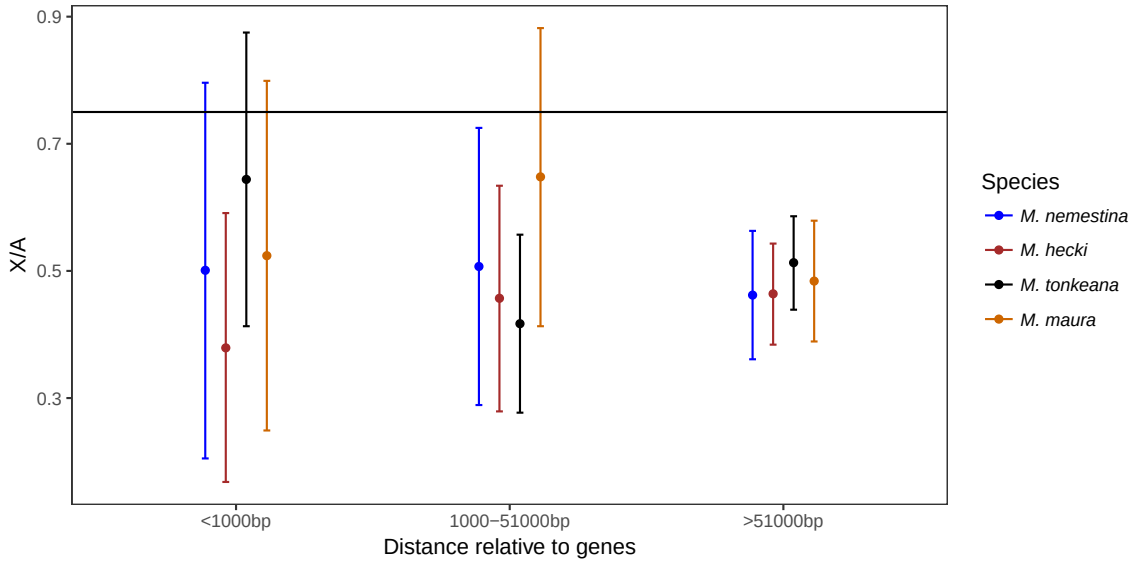

Figure S2: X/A polymorphism ratios (X/A) based on RADseq data for four species. Ratios were calculated in three genomic categories: (1) exonic and intronic sequences and flanking regions less than 1000 bp from genes (<1000bp), (2) nongenic regions that are between 1,000 and 51,0000 bp from genes (1000–51000bp), and (3) nongenic regions that are greater than 51,000 bp from genes (>51000bp). Bars indicate 95% confidence intervals and accommodate uncertainty in polymorphism and divergence. A horizontal line indicates the 0.75 expectation.

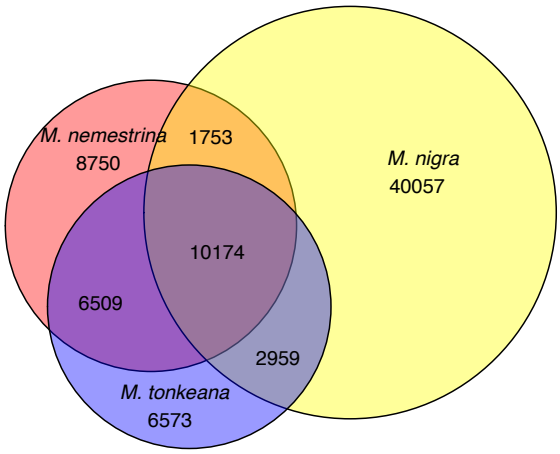

Figure S3: A Euler diagram of the number of sites in the non-pseudoautosomal region of the X chromosome for which a heterozygous genotype was inferred for one or more of the three individudals used in the analysis of gene flow across Wallace's Line. Numbers in each region of the chart refer to the number of shared or unshared heterozygous genotypes, and is based on ~57.5 million genotyped sites in each individual after discarding repetitive regions. Most of the (pseudo-) heterozygous genotypes in each male were shared with the other, whereas most of the heterozygous genotypes in the female were not shared with either male. This analysis is consistent with the proposal that most of the heterozygous genotypes in the female are real.