Supporting Information

Natural selection for the Duffy-null allele in the recently admixed people of Madagascar

Jason A. Hodgson[a,b], Joseph K. Pickrell[c,d], Laurel N. Pearson[a], Ellen E. Quillen[a,e], António Prista[f], Jorge Rocha[g,h], Himla Soodyall[i], Mark D. Shriver[a], and George H. Perry[a,j]

[a]Department of Anthropology, The Pennsylvania State University, University Park, PA 16802, USA, [b]Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot, Berkshire SL5 7PY, UK, [c]New York Genome Center, New York, NY 10013, USA, [d]Department of Biological Sciences, Columbia University, New York, NY 10027, USA, [e]Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX 78245, USA, [f]Faculdade de Educação Física e Desporto, Universidade Pedagógica, Moçambique, [g]Centro de Investigação em Biodiversidade e Recursos Genéticos da Universidade do Porto (CIBIO), Vairão, Portugal, [h]Departamento de Biologia, Faculdade de Ciências da Universidade do Porto, Porto, Portugal, [i]Human Genomic Diversity and Disease Research Unit, Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of Witwatersrand and the National Health Laboratory Service, Johannesburg 2000, South Africa, [j]Department of Biology, Penn State University, University Park, PA 16802, USA

1. SUPPLEMENTARY MATERIAL AND METHODS

Samples. The samples used in this study consist of 19 Malagasy, 16 Bantu speakers from South Africa, 20 Burunge from East Africa, 20 Mozambican, 42 Europeans, 11 Mala and 11 Brahmin from India, and 20 Chinese or Japanese. 10k SNP data were collected using methods described in ref. [1]. The Burunge, Europeans, Mala, Brahmin, Chinese and Japanese, and Mozambican samples have been described previously [2-4]. The European, and Chinese and Japanese samples are from the Coriell "Caucasian" and "East Asian" collections. The Malagasy and South African Bantu samples are newly described here. The Malagasy sample consists of Merina individuals collected from the Central Highlands of Madagascar. The sample of South African Bantu speakers was collected in Johannesburg, South Africa. Both the Malagasy and South African samples were collected with approval from the Human Research Ethics Committee of the University of Witwatersrand.

Duffy-null allele genotyping. We genotyped the C -> T SNP that determines the Duffy positive/null genotype (rs2814778) using Sanger sequencing. A 464bp segment of the *DARC* promoter region containing rs2814778 was amplified with the PCR (electronic supplementary material, figure S2a) using the following primers (both 5'-3'): forward ACTTTCTGGTCCCCACCTTT, and reverse ACAAGAGGGAGCTAGGAGGC. Sequencing was performed on the positive strand using an internal primer TAGTCCCAACCAGCCAAATC on an Applied Biosystems 3730 capillary sequencer. Electropherograms were visualized with Geneious. Homozygotes were resolved as a single peak of normal magnitude, while heterozygotes were resolved as double peaks at half the normal magnitude (electronic supplementary material, figure S2b).

Quantification of Malagasy admixture. An identity by state-multidimensional scaling (IBS-MDS) analysis was performed with PLINK v1.07 [5]. First, an IBS matrix was calculated using autosomal SNPs, excluding SNPs with MAF < 0.01, SNPS with > 20% missing data, and individuals with >50% missing data. The resulting IBS matrix was then used as input for the MDS analysis. Results were plotted using the ggplot2 library [6] for R v2.15.1 [7]. The same excluded dataset used to calculate the IBS matrix was also used in the ADMIXTURE analysis. To determine the number of ancestral populations, K, that best describes the data, we calculated cross-validation error for K=2 through K=6 (electronic supplementary material, figure S1), and chose K that minimized the error (K = 3), using the ADMIXTURE cross-validation

procedure. Statistical analyses and plots of estimated ancestry proportions were performed with R v2.15.1 [7].

Expected allele frequency in the Malagasy. Expected allele frequencies were calculated according to the following formula:

$E(p_d) = \alpha(p_1) + (1 - \alpha)(p_2)$

where $E(p_d)$ is the expected frequency of allele p in the admixed daughter population, p_1 is the frequency of the allele in the parent population 1, p_2 is the frequency of the allele in the parent population 2, and α is the admixture proportion contributed by parent population 1.

Likelihood ratio test to compare a history of Duffy-null allele frequency stasis to a history of change. Because we observed wide variance in sub-Saharan African ancestry proportions among our Merina population sample (figure 1b), we tested whether the observed Duffy-null genotypes were different than expected, given the corresponding individual estimates of sub-Saharan African ancestry (and assuming Duffy-null allele frequencies of 0.0 and 1.0 in the founding Austronesian and mainland African populations, respectively).

Specifically, we used a likelihood ratio test to determine whether the observed Duffy genotypes (SNP rs2814778), and the corresponding estimates of sub-Saharan African ancestry for 18 Malagasy individuals were more consistent with the null hypothesis of no allele frequency change since admixture began, to the alternative hypothesis that Duffy-null allele frequency has changed significantly. We assume no genotyping error and no error in the ancestry estimates. Let $\vec{g} = [g_1, g_2, ...g_i]$ be a vector containing the numbers of Duffy-null alleles in each individual *i* and $\vec{f} = [f_1, f_2, ...f_i]$ be the corresponding vector of African ancestry proportions. If the Duffy-null allele was fixed in the African population that colonized Madagascar and absent from the Austronesian population that colonized Madagascar (as well as from any Eurasian populations that also contributed to Malagasy ancestry), then:

$$P(g_i \mid f_i) \sim Bin(2, f_i)$$

We then define a simple model that allows the Duffy-null allele frequency to have changed since admixture. Let δ be a parameter controlling the amount of change. We define the following likelihood function:

$$L(\vec{g} \mid \vec{f}) = \sum_{i=1}^{18} Bin(g_i \mid 2, f_i + \delta(1 - f_i)),$$

where *Bin(x| n, p)* represents the binomial density function with parameters *n* and *p* evaluated at *x*.Using this likelihood function, we used a likelihood ratio test to compare the null hypothesis where $\delta = 0$, to the alternative hypothesis where $\delta \neq 0$.

This likelihood ratio test rejected the null hypothesis that the frequency of the Duffy-null allele has not changed since the time of admixture (*P* = 0.004), consistent with the notion of a recent Duffy-null allele frequency increase in the Malagasy.

Computer simulations. A program, *fastadsim.pl*, was written to simulate admixture between two parent populations. The program simulates genetic drift for a single allele through time, assuming i) instantaneous admixture between two parent populations, ii) diploidy, iii) random mating, iv) no mutation, and the following user-defined parameters: a) allele frequency of parent population 1, b) allele frequency of parent population 2, c) admixture proportion, d) starting effective population size, e) per generation population growth, f) the number of generations of drift, g) per generation migration rate from population 1, h) per generation migration rate from population 2, and i) the number of simulation replicates to perform.

An alternative test of the null hypothesis that the observed Duffy-null allele frequency change can be explained by genetic drift alone could be to compare the observed allele frequency shift to a null distribution derived from putatively neutrally evolving SNPs in the Malagasy with similar frequency difference between the mainland sub-Saharan African and East Asian source populations. However, the extreme Duffy-null allele frequency difference observed between sub-Saharan Africans and the rest of the world is practically unknown elsewhere in the genome. For example, the 1000 Genomes Project found only four fixed differences between West Africans and Western Europeans, including the Duffy-null allele, and only 72 between West Africans and East Asians [8], none of which are included in the Affymetrix 10k SNP panel that we used for this study. Of the 11,552 SNPs genotyped for this study, only 2 had frequency differences greater than 0.9, and 37 greater than 0.8, between Mozambicans and East Asians. Also, SNPs with such extreme frequency differences between populations may have resulted

from natural selection in at least one of the two populations [9, 10], making such SNPs inappropriate from which to derive a neutral null distribution. Consequently, we used the neutral evolution computer simulation approach described above to explore the amount of genetic drift that can be expected for the Duffy-null allele in Madagascar under various conservative demographic scenarios.

Expected allelic diversity lost to drift. We used *fastadsim.pl* to estimate the amount of genetic diversity lost to drift that we would expect to observe for a given demographic scenario. We performed 200 simulations under the specified demographic parameters for alleles with starting frequencies of 0.001 – 0.500 in increments of 0.001 (100,000 simulations), recording the final allele frequency of each simulation. Then to account for sampling bias in our observed SNP data (i.e., low frequency polymorphisms will often be sampled to be monomorphic), for each simulation we drew a sample of 2N alleles with probability equal to the simulated final allele frequency, where N is the observed sample size. We then recorded the proportion of simulations in which one of the two alleles was sampled to be lost due to drift (fixation).

Estimation of selection coefficient. We estimated the coefficient of selection, s, assuming $w_{(+/+)} = 1$, $w_{(+/-)} = 1$, and $w_{(-/-)} = 1 + s$, where $w_{(+/+)}$ is the fitness of Duffy-positive phenotype homozygotes, $w_{(+/-)}$ is the fitness of Duffy-positive heterozygotes, and $w_{(-/-)}$ is the fitness of Duffy-negative phenotype homozygotes. We used the following equation to calculate the frequency of the Duffy-null allele in the next generation (q') given the frequency of Duff-null in the current generation (q) and the coefficient of selection (s):

$$q' = ((1-q)q + q^2 + sq^2) / (1 + sq^2)$$

This equation was used recursively to find values of s that resulted in the observed change in Duffy-null allele frequency given g generations. A perl script was used to automate this process. This equation was also used recursively to estimate the number of generations required for a Duffy-null selective sweep.

2. SUPPLEMENTARY TABLE

Table S1. Estimated selection coefficients.

Population	Duffy-null freq change	# gens	s
Merina	0.48 → 0.72*	33	0.052
Merina	0.48 → 0.72*	43	0.040
Merina	0.48 → 0.72*	77	0.022
Merina	0.48 → 0.78**	33	0.066
Merina	0.48 → 0.78**	43	0.050
Merina	0.48 → 0.78**	77	0.028
South coast	0.67 → 0.92**	33	0.065
South coast	0.67 → 0.92**	43	0.049
South coast	0.67 → 0.92**	77	0.027

Estimated selection coefficients assuming selection for the Duffy-null phenotype (e.g. w+/+ = 1, w+/- = 1, w-/- = 1 + s), and various assumptions about the number of generations of selection. *Duffy-null allele frequency in our sample of 18 Merina individuals. **Duffy-null allele frequency from Highland (N=69) and south coast (N=86) Malagasy [11].

3. SUPPLEMENTARY FIGURES

Figure S1. Plot of ADMIXTURE cross validation error from K=2 through K=6. We chose K=3 to analyze the SNP data, as the value that minimizes the error.

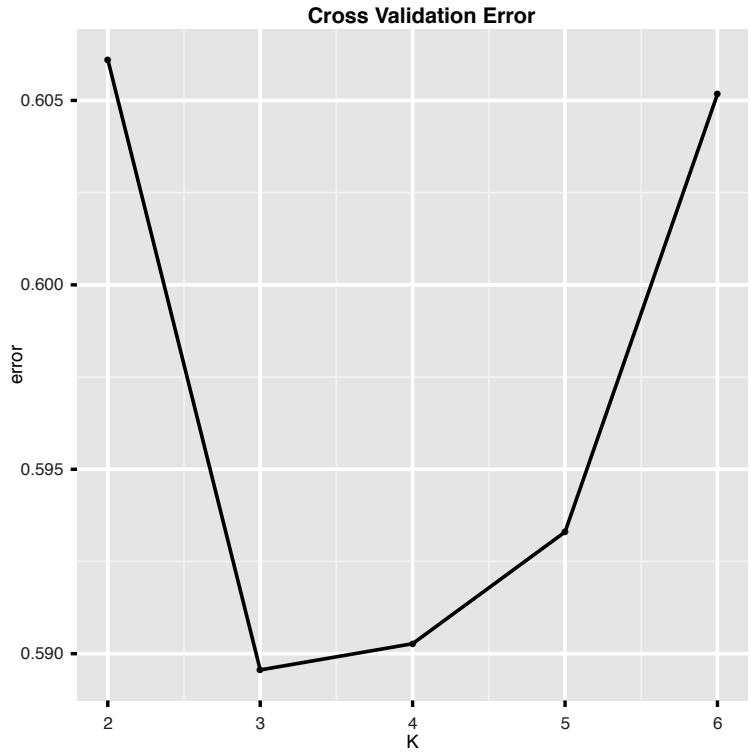


Figure S2. Duffy-positive/ null genotyping. a) The Duffy-positive/ null phenotype is determined by SNP rs2814778 located in the promoter region of the Duffy Antigen Receptor for Chemokines (*DARC*) gene, just upstream of the 5' UTR. We successfully genotyped this SNP in 18 of 19 Malagasy, by amplifying a 464 bp. region around the SNP and Sanger sequencing the PCR product. b) Genotypes were determined by examining the electropherograms with large single peaks recorded as homozygotes, and small double peaks recorded as heterozygotes.

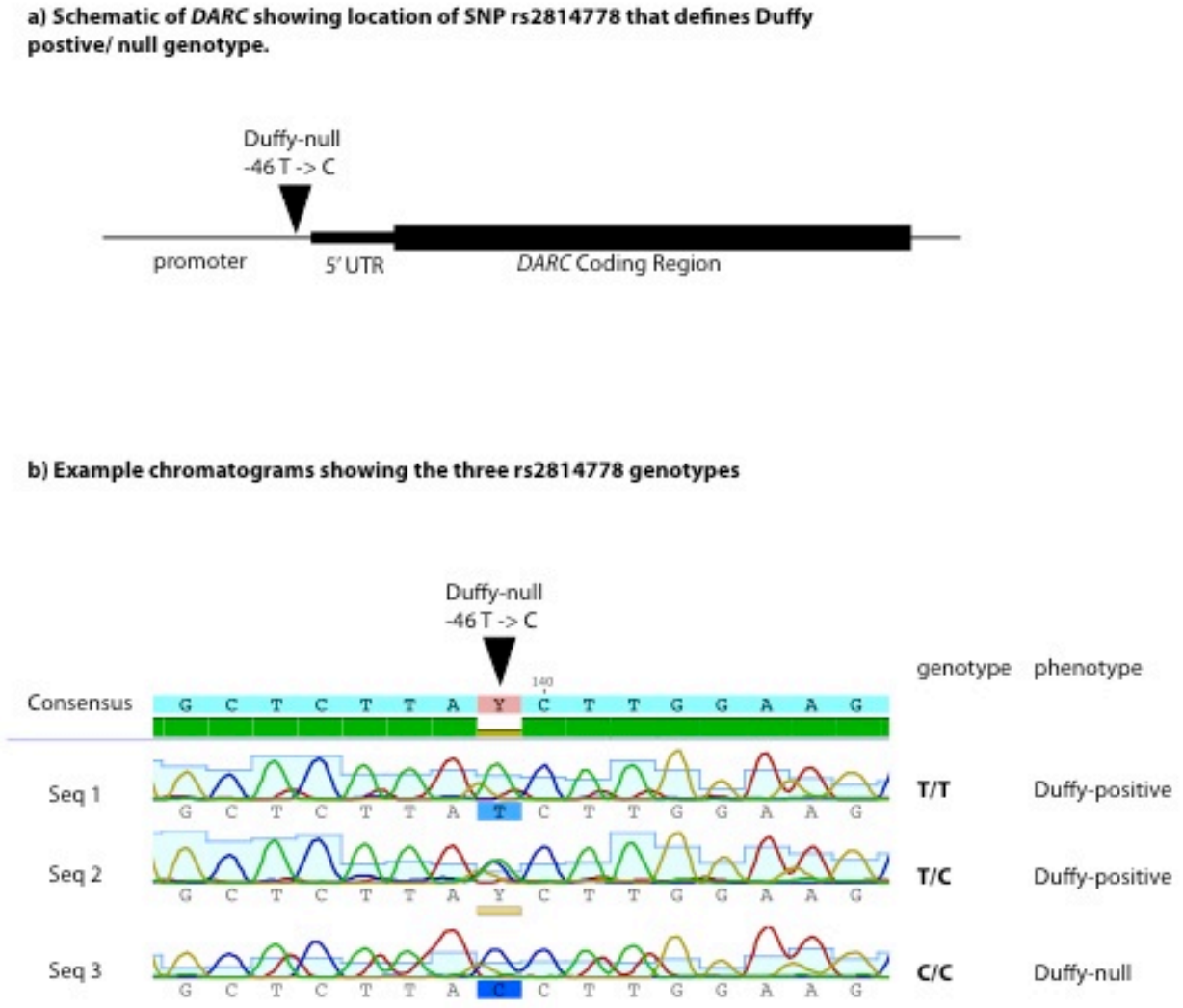


Figure S3. Simulated frequencies of the Duffy-null allele under genetic drift with constant migration from sub-Saharan Africa. For each demographic scenario – comprised of variable initial effective population sizes (N_e = 100, 200, or 500), number of generations (33, 43, 77; equivalent to ~1,000, 1,300, and 2,300 years respectively), number of migrants per generation from sub-Saharan Africa (1, 2, or 4 when possible to result in a final average allele frequency of ~0.48), and 2% population growth per generation – forward-evolution simulations were performed starting with the indicated initial admixture proportion (α_0). Initial admixture proportions were chosen such that the final admixture proportion (α_x) would be ~0.48 given the migration rate. The final allele frequency was recorded for each of 10,000 simulations for each demographic scenario. The bars show the 50%, 75%, 95%, and 99% distributions of the final simulated allele frequencies for each demographic scenario. The solid vertical line indicates the observed frequency of 0.78 for the Duffy-null allele in a sample of Highland Merina and Bezanozano Malagasy (N=69) [11], while the dashed line indicates the frequency of the Duffy-null allele in our sample of 18 Merina. The percentage of simulations with an allele frequency greater than or equal to 0.72 and 0.78 are shown in the right hand column (>= 5% shown in bold).

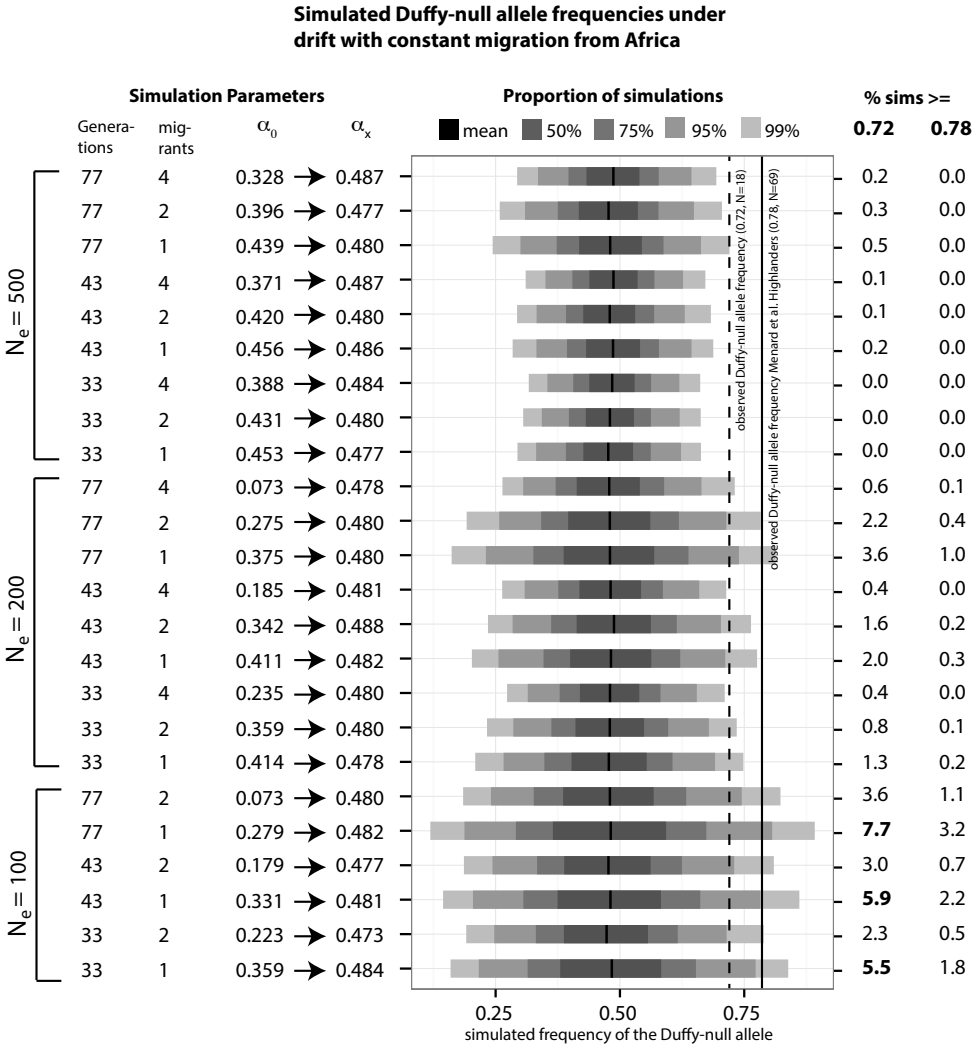


Figure S4. Simulated frequencies of the Duffy-null allele under genetic drift with constant migration from Eurasia. For each demographic scenario – comprised of variable initial effective population sizes (N_e = 100, 200, or 500), number of generations (33, 43, 77; equivalent to ~1,000, 1,300, and 2,300 years respectively), number of migrants per generation from Eurasia (1, 2, or 4 when possible to result in a final average allele frequency of ~0.48), and 2% population growth per generation – forward-evolution simulations were performed starting with the indicated initial admixture proportion (α_0). Initial admixture proportions were chosen such that the final admixture proportion (α_x) would be ~0.48 given the migration rate. The final allele frequency was recorded for each of 10,000 simulations for each demographic scenario. The bars show the 50%, 75%, 95%, and 99% distributions of the final simulated allele frequencies for each demographic scenario. The solid vertical line indicates the observed frequency of 0.78 for the Duffy-null allele in a sample of Highland Merina and Bezanozano Malagasy (N=69) [11], while the dashed line indicates the frequency of the Duffy-null allele in our sample of 18 Merina. The percentage of simulations with an allele frequency greater than or equal to 0.72 and 0.78 are shown in the right hand column (>= 5% shown in bold).

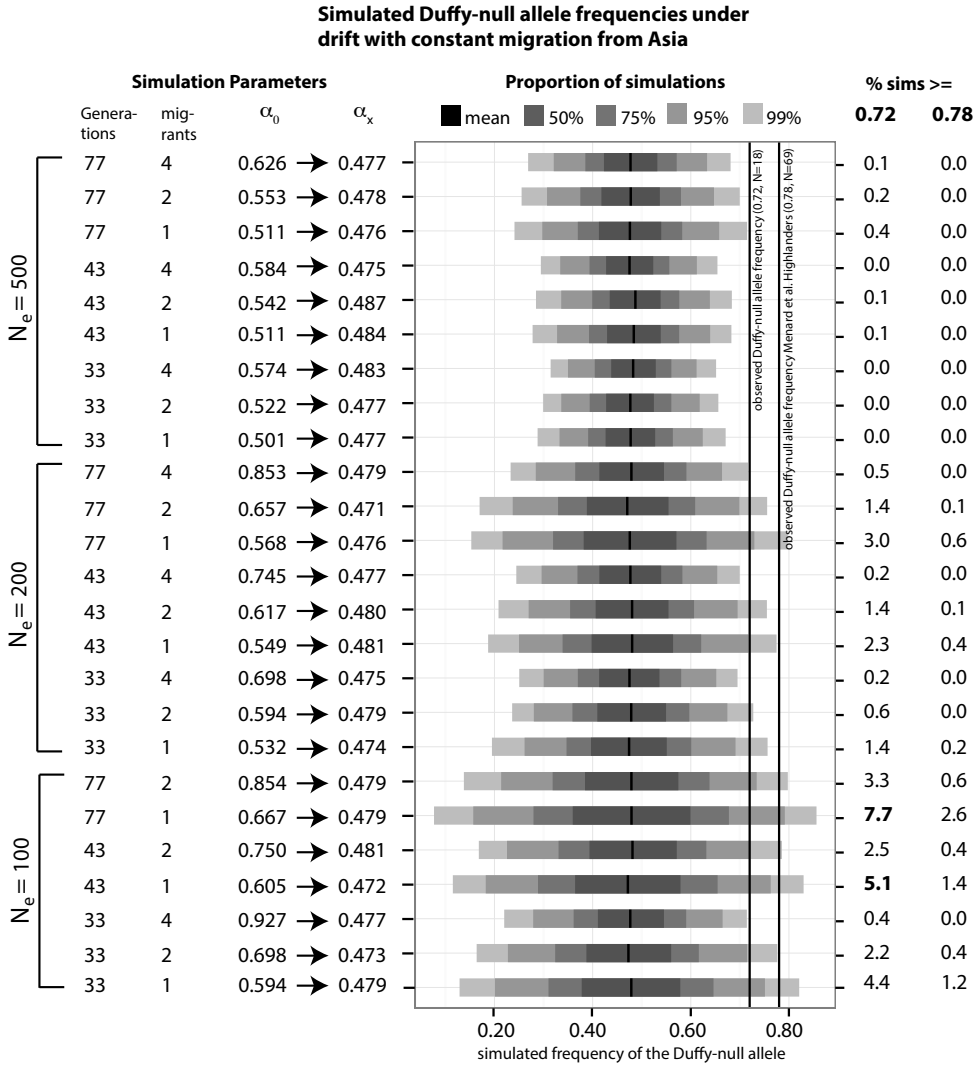


Figure S5. Simulated frequencies of the Duffy-null allele under genetic drift in southern coastal populations with 67% sub-Saharan African ancestry. For each demographic scenario - comprised of variable initial effective population sizes (Ne = 100, 200, or 500), number of generations (33, 43, 77; equivalent to ~1,000, 1,300, or 2,300 years, respectively), and levels of population growth (2%, 5% or 10%) - forward-evolution simulations were performed from a starting allele frequency of 0.67 based on the estimate of African ancestry in southern coastal Malagasy groups [12]. The final allele frequency was recorded for each of 10,000 simulations run for each demographic scenario. The bars show the 50%, 75%, 95%, and 99% distributions of the final simulated allele frequencies for each demographic scenario. The solid vertical line indicates the observed Duffy-null frequency of 0.92 from the southern coastal location of Faragangana (N=86) [11]. The percentage of simulations with a final Duffy-null allele frequency greater than or equal to 0.92 are shown in the right hand column (>= 5% shown in bold).

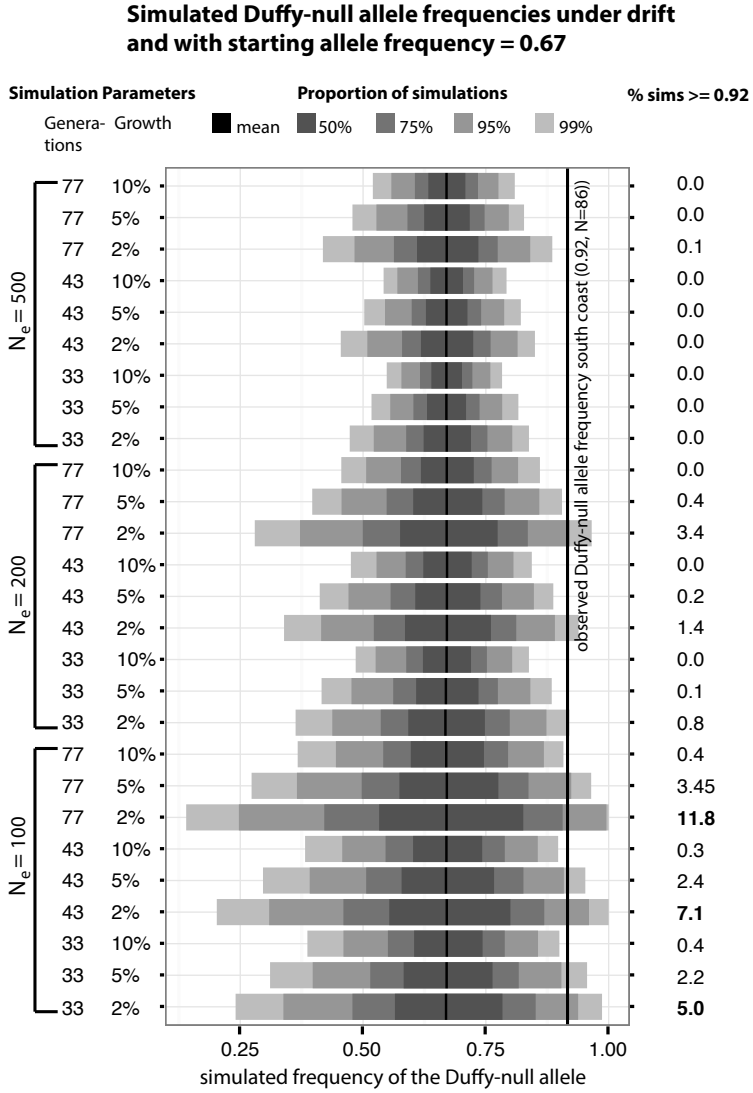


Figure S6. Theoretical versus observed levels of allelic fixation in the Merina with 48% sub-Saharan African ancestry, and with a Duffy-null allele frequency of 0.72. Proportions of fixed SNP loci in the Merina from simulated (dashed line) and observed (solid line) data, binned by predicted allele frequencies at the initial time of admixture, and observed allele frequencies in present-day sub-Saharan African and East Asian population samples, given the least extreme scenario given 48% sub-Saharan admixture and a Duffy-null allele frequency of 0.72 (see Fig. 2). The observed data are based on the Affymetrix 10k SNP data from our population sample of 19 Merina individuals. In each case the observed allelic fixation is less than that expected under the type of extreme demographic conditions necessary for such a large shift in the Duffy-null allele frequency in the absence of positive selection.

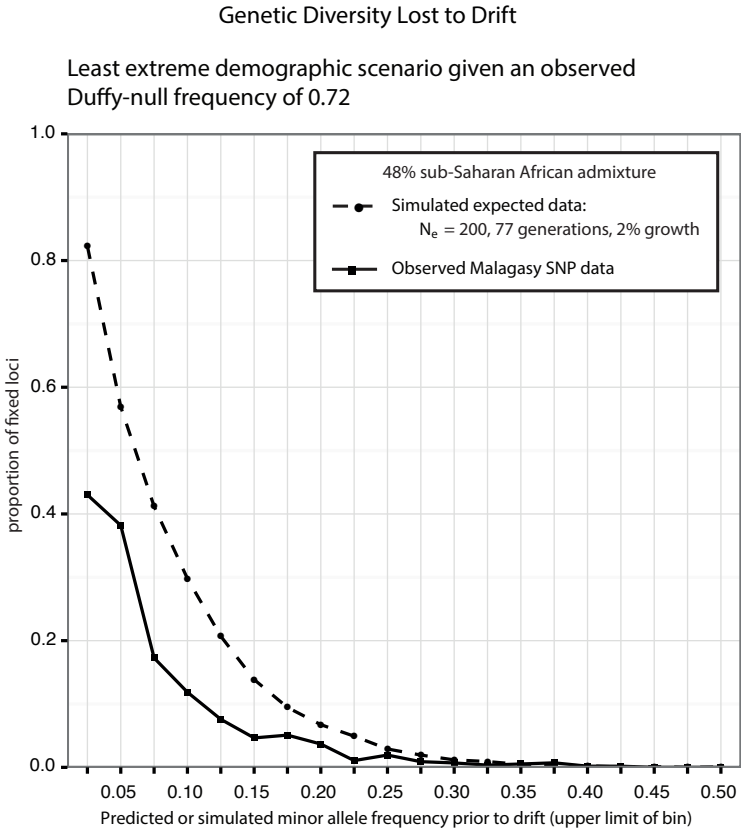
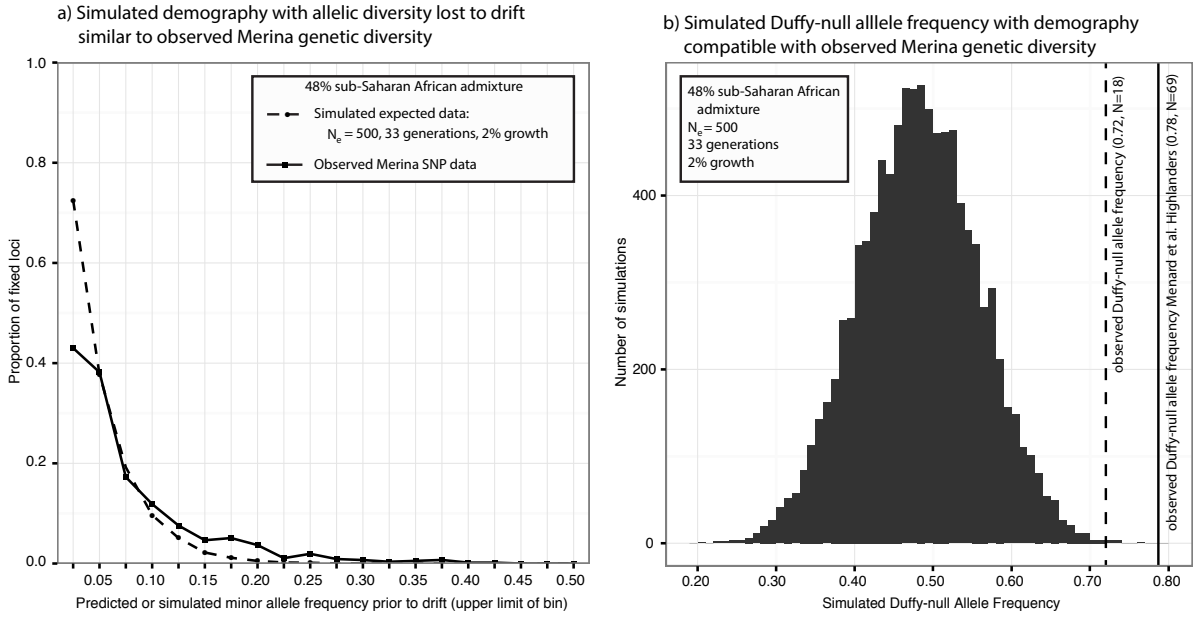


Figure S7. Simulated expected distribution of genetic drift for the Duffy-null allele for a demography consistent with observed levels of allelic fixation in the Merina. a) Expected level of allelic fixation given N_e = 500, 33 generations, and 2% growth; a demography we found to produce similar allelic fixation to the observed Merina SNP data. b) Distribution of 10,000 genetic drift simulations for the Duffy-null allele given the demography this close fit demography. None of the 10,000 simulations finished with a Duffy-null allele frequency ≥ 0.78. Only seven of 10,000 simulations finished with a Duffy-null allele frequency ≥ 0.72.



4. REFERENCES

1. Shriver M.D., Mei R., Parra E.J., Sonpar V., Halder I., Tishkoff S.A., Schurr T.G., Zhadanov S.I., Osipova L.P., Brutsaert T.D., et al. 2005 Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. *Hum Genomics* **2**(2), 81-89.

2. Norton H.L., Kittles R.A., Parra E., McKeigue P., Mao X., Cheng K., Canfield V.A., Bradley D.G., McEvoy B., Shriver M.D. 2007 Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Mol Biol Evol* **24**(3), 710-722.

3. Bauchet M., McEvoy B., Pearson L.N., Quillen E.E., Sarkisian T., Hovhannesyan K., Deka R., Bradley D.G., Shriver M.D. 2007 Measuring European population stratification with microarray genotype data. *Am J Hum Genet* **80**(5), 948-956.

4. Alves I., Coelho M., Gignoux C., Damasceno A., Prista A., Rocha J. 2011 Genetic Homogeneity Across Bantu-Speaking Groups from Mozambique and Angola Challenges Early Split Scenarios between East and West Bantu Populations. *Human Biology* **83**(1), 13-38.

5. Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J., et al. 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**(3), 559-575.

6. Wickham H. 2009 ggplot2: elegant graphics for data analysis. (New York, NY, USA, Springer.

7. R_Development_Core_Team. 2011 R: A language and environment for statistical computing. (Vienna, Austria, R Foundation for Statistical Computing.

8. Durbin R.M., Abecasis G.R., Altshuler D.L., Auton A., Brooks L.D., Gibbs R.A., Hurles M.E., McVean G.A. 2010 A map of human genome variation from population-scale sequencing. *Nature* **467**(7319), 1061-1073.

9. Barreiro L.B., Laval G., Quach H., Patin E., Quintana-Murci L. 2008 Natural selection has driven population differentiation in modern humans. *Nat Genet* **40**(3), 340-345.

10. Excoffier L., Hofer T., Foll M. 2009 Detecting loci under selection in a hierarchically structured population. *Heredity* **103**(4), 285-298.

11. Menard D., Barnadas C., Bouchier C., Henry-Halldin C., Gray L.R., Ratsimbasoa A., Thonier V., Carod J.F., Domarle O., Colin Y., et al. 2010 Plasmodium vivax clinical malaria is commonly observed in Duffy-negative Malagasy people. *Proc Natl Acad Sci U S A* **107**(13), 5967-5971.

12. Pierron D., Razafindrazaka H., Pagani L., Ricaut F.X., Antao T., Capredon M., Sambo C., Radimilahy C., Rakotoarisoa J.A., Blench R.M., et al. 2014 Genome-wide evidence of Austronesian-Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar. *Proc Natl Acad Sci U S A*.