Supporting information for *Demographic history and rare allele sharing among human populations*.

Simon Gravel, Brenna M. Henn, Ryan N. Gutenkunst, Amit R. Indap, Gabor T. Marth, Andrew G. Clark, The 1000 Genomes consortium, Carlos D. Bustamante

I. ESTIMATING PARAMETER CONFIDENCE INTERVALS VIA BOOTSTRAP

In order to obtain confidence intervals on the inferred demographic parameters, we used a bootstrap approach. Since linkage effectively reduces the number of independent measurements, we re-sampled genomic regions rather than individual SNPs. More specifically, we considered regions with a coding sequence (CDS) annotation in Gencode [1] version 3b, and merged regions within 50kbp of each other. We were then left with 9328 distinct and well-separated regions. We re-sampled these with replacement, forming a set of overlapping genomic regions. For each of these bootstrapped sets, we extracted SNPs generated from both the high- and low-coverage data. We obtained inferred values for both the error rates due to low coverage and the resulting inferred demographic parameters.

II. SAMPLE SELECTION

Short read sequencing results in high site-to-site and sample-to-sample variability in coverage beyond the independent draw, Poisson sampling [Marth *et al.*, in submission]. Since we are interested in obtaining high quality reads from the capture data, we discarded individuals for which the mean over exons of the median coverage per exon was below 15. We also discarded individuals with unusually high discrepancy with validation or HapMap data. As a result, we removed samples [NA18524, NA18529, NA18530, NA18531, NA18543, NA18557, NA18599, NA18615, NA18620, NA18627, NA18628, NA18635, NA18642, NA18749, NA18773] (15 samples) from CHB, [NA06985, NA06994, NA11840, NA11995, NA12004, NA12006, NA12156, NA12414, NA12763, NA12815, NA12829, NA12842, NA12872, NA12873, NA12874, NA12889] (16 samples) from CEU, [NA18948, NA18957, NA18961, NA18964, NA18969, NA18979, NA18981, NA18985, NA18988, NA18989, NA18994, NA19006, NA19054, NA19062, NA19065, NA19068, NA19077, NA19090, NA19554, NA19559, NA19562] (21 samples) from JPT and [NA18489, NA18499, NA18504, NA18516, NA18522, NA18865, NA18870, NA18871, NA18917, NA19102, NA19116, NA19137, NA19141, NA19181, NA19201, NA19207, NA19210, NA19220] (18 samples) from YRI, for a total of 70 samples in the 4 populations sequenced in the low-coverage pilot. For exon-only analysis, we also removed, for the same reasons, samples [NA20511, NA20512, NA20518, NA20519, NA20524, NA20527, NA20528, NA20529, NA20530, NA20540, NA20587, NA20763] (12 samples) from TSI [NA17966, NA18109, NA18112, NA18147, NA18670] (5 samples) from CHD, and [NA19046, NA19307, NA19310, NA19312, NA19317, NA19321, NA19441, NA19453, NA19456] (9 samples) from LWK.

III. ERROR CORRECTION

The matrix $\mathbf{A}$ describing the error model can be inverted to give a correction model

$$S^0_{ijk} = \sum_{i'j'k'} \left(\mathbf{A}^{-1}\right)_{ijk;i'j'k'} S_{i'j'k'},$$

with

$$\left(\mathbf{A}^{-1}\right)_{ijk;i'j'k'} = \left(A^1\right)^{-1}_{ii'} \left(A^2\right)^{-1}_{jj'} \left(A^3\right)^{-1}_{kk'} \qquad \text{(S1)}$$

and

$$\left(A^p\right)^{-1}_{ii'} = \begin{cases} \frac{1}{1-\epsilon^p_{i'}} & \text{if } i = i' \\ \frac{-\epsilon^p_{i'}}{1-\epsilon^p_{i'}} & \text{if } i = 0, i' \neq 0 \\ 0 & \text{otherwise.} \end{cases} \qquad \text{(S2)}$$

Note $\left(A^p_{ii'}\right)^{-1} = 0$ for $i' = 0, i \neq 0$, implying that it is not necessary to know the number of observed invariant sites to estimate the corrected SFS for variable sites. The same holds true for $\mathbf{A}$, obtaining S^0_{ijk} for $(i,j,k) \neq (0,0,0)$ does not require the knowledge of $S_{0,0,0}$.

In practice, this correction model should be used with caution; in some cases, estimated entries in the corrected SFS result from a difference of two large numbers, and can even be negative. The Poisson random field approximation, used in the maximum likelihood calculation, would not hold for such cases. However, the corrected model is a sensible approach for calculating basic population demographic estimates (such as F_{ST}) that are not based on likelihood estimates.

Comparison of exon data, low-coverage data, and this error-correction model are shown on Figures S1, S2, S3.

IV. COMPARISON BETWEEN MODEL AND DATA FOR MARGINAL FREQUENCY SPECTRA.

We compare in Figure S5 the model predicted and marginal site frequency spectra for the CEU, CHB+JPT, and YRI panels.

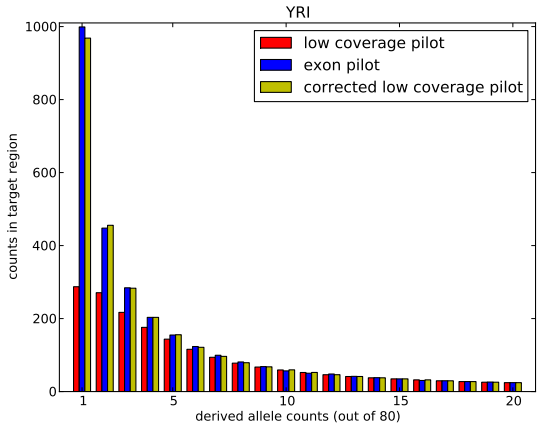

FIG. S1. Site frequency spectra for the YRI panel for sites with variant calls in the exon pilot. Shown are the exon pilot data, the low-coverage data, and a corrected SFS using the exponentially decaying error model, for derived allele frequency below 25%. All spectra are polarized using the March 2006 assembly of the chimp genome (PanTro2). The corrected SFS captures the bulk of the differences between the exon and low coverage data across the population frequencies.

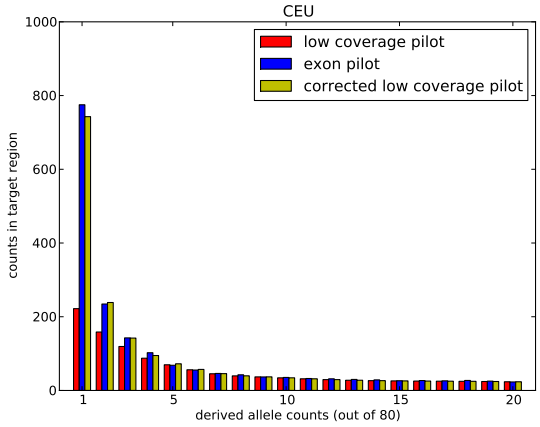

FIG. S2. Site frequency spectra for the CEU panel for sites with variant calls in the exon pilot. Shown are the exon pilot data, the low-coverage data, and a corrected SFS using the exponentially decaying error model, for derived allele frequency below 25%. All spectra are polarized using the March 2006 assembly of the chimp genome (PanTro2). The corrected SFS captures the bulk of the differences between the exon and low coverage data across the population frequencies.

V. JACKKNIFE EXPANSION

The Burnham-Overton jacknife was introduced to predict the number of unobserved objects (typically, animals) based on a finite number of mark-and-recapture field trips. The basic assumption of this model is a postulated relation between the total number of objects $V(\infty)$, the number $V(n)$ of objects observed after n field trips, and n itself :

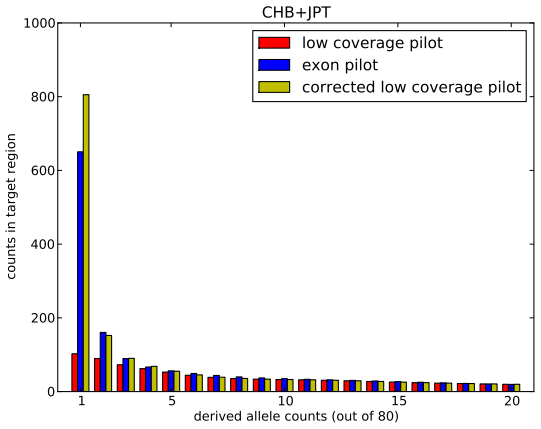

FIG. S3. Site frequency spectra for the CHB and JPT panels for sites with variant calls in the exon pilot. Shown are the exon pilot data, the low-coverage data, and a corrected SFS using the exponentially decaying error model, for derived allele frequency below 25%. All spectra are polarized using the March 2006 assembly of the chimp genome (PanTro2). The corrected SFS captures the bulk of the differences between the exon and low coverage data across the population frequencies.

$$\hat{V}(\infty) = V(n) + \sum_{i=1}^{p} \frac{a_i^p}{n^i}$$

.

The Burnham-Overton jackknife estimator has been shown to perform well in a variety of situations [3, 4], even though it has been shown that, since an arbitrary number of very rarely observed objects might eventually be discovered, no point estimator can be generally unbiased in such infinite extrapolation [5, 6].

The genetics context requires a more modest extrapolation to a finite number of observations. In this case, the potential problem of an infinite number of rarely observed objects is eliminated: based on the random sampling assumption, we know that the rate of new discoveries per sample cannot increase with the number of samples. The number of discoveries can therefore be bounded with good confidence above and below. This does not guarantee the existence of an unbiased point estimator, but supports the idea that a jackknife estimator for finite extrapolation has the potential to be more reliable than one for infinite extrapolation.

With this in mind, we write the expansion

$$\hat{V}_n(N) = V(n) + \sum_{i=1}^{p} a_i^p \Delta^i(N, n), \tag{S3}$$

where $\Delta(N, n) = \sum_{j=n}^{N-1} 1/j$ for a fixed jackknife order p. This assumption provides the expected behavior as $n \to N$, and is exact for $p \geq 1$ in the case of a constant-size, neutrally evolving population. The parameters a_k^p are calculated by equating estimators $\hat{V}_n(N) = \hat{V}_{n-1}(N) = \cdots = \hat{V}_{n-p}(N)$,

and solving the resulting p linear equations. Explicit expressions are given below.

Figure S6 shows that, based on subsampled simulated data from our demographic model, our jackknife estimator could predict with good accuracy the number of segregating sites in the full data.

A. Parameters of the jackknife expansion

The jackknife expansion parameters for the number of undiscovered variants in a sample of size N, based on the site frequency spectrum $\phi(i)$ from a sample of size n, are easily derived using symbolic software using $V_n(N) = V(n) + \sum_{i=1}^{p} a_i^p \Delta^i(N,n)$, together with the equality $\hat{V}_n(N) = \hat{V}_{n-1}(N) = \cdots = \hat{V}_{n-p}(N)$, and

$$\begin{aligned}\sum_{j=1}^{k} \frac{\binom{k}{j}}{\binom{N}{j}} \Phi(j) =& V(n) - V(n-k) \\ =& \sum_{j=1}^{p} a_j^p \left(\Delta^j(N,n-k) - \Delta^j(N,n)\right).\end{aligned} \tag{S4}$$

The resulting equations for the number of missed variants m_p are bulky, but are simple rational expressions of n, N, and the $\phi(i)$. The predicted number m_p of missing variants at second and third order are

$$\begin{aligned}m_2 =& \left(\frac{(-1+5n-6n^2+2n^3)\Delta(N,n) + (-1+n)(2-3n+n^2)\Delta(N,n)^2}{n(3-5n+2n^2)}\right)\phi(1) \\ &+ \left(\frac{-2(-2+n)^2\Delta(N,n) - 2(-2+n)(2-3n+n^2)\Delta(N,n)^2}{n(3-5n+2n^2)}\right)\phi(2)\end{aligned} \tag{S5}$$

at second order, and

$$\begin{aligned}m_3 =& \frac{\left(127-501n+885n^2-837n^3+432n^4-114n^5+12n^6\right)\Delta(N,n)}{n\left(-165+521n-637n^2+377n^3-108n^4+12n^5\right)}\phi(1) \\ &+ \frac{(-2+n)\Delta(N,n)^2\left(-5-51n+81n^2-39n^3+6n^4+2(-2+n)^2\left(3-4n+n^2\right)\Delta(N,n)\right)}{n\left(165-356n+281n^2-96n^3+12n^4\right)}\phi(1) \\ &+ \frac{2\left(248-402n+61n^2+249n^3-195n^4+57n^5-6n^6\right)\Delta(N,n)}{(-1+n)^2 n\left(165-356n+281n^2-96n^3+12n^4\right)}\phi(2) \\ &- \frac{2\left(-560+1130n-731n^2+55n^3+127n^4-51n^5+6n^6\right)\Delta(N,n)^2}{n\left(-165+521n-637n^2+377n^3-108n^4+12n^5\right)}\phi(2) \\ &- \frac{4(-2+n)^2\left(-39+52n-22n^2+3n^3\right)\Delta(N,n)^3}{n\left(165-356n+281n^2-96n^3+12n^4\right)}\phi(2) \\ &+ \frac{6(-3+n)^3\Delta(N,n)\left(-3+2n+\left(5-8n+3n^2\right)\Delta(N,n)\right)}{(-1+n)^2 n\left(-55+82n-39n^2+6n^3\right)}\phi(3) \\ &+ \frac{6(-3+n)^3(-2+n)\Delta(N,n)^3}{n\left(-55+82n-39n^2+6n^3\right)}\phi(3)\end{aligned} \tag{S6}$$

at third order.

A limit of interest occurs when $n >> 1$, when the estimates simplify to

$$\begin{aligned}m_1 =& \Delta(N,n)\phi(1) \\ m_2 =& \left(\Delta(N,n) + \frac{\Delta^2(N,n)}{2}\right)\phi(1) - \Delta^2(N,n)\phi(2) \\ m_3 =& \left(\Delta(N,n) + \frac{\Delta^2(N,n)}{2} + \frac{\Delta^3(N,n)}{6}\right)\phi(1) \\ &- \left(\Delta^3(N,n) + \Delta^2(N,n)\right)\phi(2) + \Delta^3(N,n)\phi(3).\end{aligned} \tag{S7}$$

Note that the number of terms of the site frequency spectrum that are used in the estimate is equal to the order of the jackknife expansion, and the magnitude of the coefficient of each SFS term increases with the perturbation order, in a way that may diverge for sufficiently large Δ. As a result, for large values of Δ, the jackknife expansion becomes unstable and sensitive to noise: the choice of the jackknife expansion order is therefore always a compromise.

B. Jackknife estimates for all exon pilot populations

In figure S7, we show the jackknife projected number of sites to be discovered in all 7 populations from the exon pilot project. In Figure S6, we compare the results of the jacknife estimator applied to the expected SFS resulting from sampling 80 individuals from each of the three panels in our analysis.

VI. LIKELIHOOD PROFILES AND BOOTSTRAP DISTRIBUTION

Figure S8 provides a comparison of bootstrap estimates to likelihood profiling. We provide the results in genetic units, as likelihood profiling requires holding parameters constant while optimizing other parameters, a task much simplified if carried in the units used in the simulation. Given N_a and generation time g (see Methods), the transformation from genetic to physical units is

$$\begin{aligned} T_{EuAs} &= 2N_a g \tau_{EuAs} \\ T_B &= 2N_a g \left(\tau_{EuAs} + \tau_B\right) \\ T_{Af} &= 2N_a g \left(\tau_{EuAs} + \tau_B + \tau_A f\right) \\ N_{Af} &= N_a \nu_{Af} \\ N_B &= N_a \nu_B \\ N_{EU0} &= N_a \nu_{EU0} \\ N_{AS0} &= N_a \nu_{AS0} \\ r_{AS} &= \left(\nu_{AS}/\nu_{AS0}\right)^{g/T_{EuAs}} - 1 \\ r_{EU} &= \left(\nu_{EU}/\nu_{EU0}\right)^{g/T_{EuAs}} - 1 \\ m_i &= M_i/2N_a. \end{aligned} \tag{S8}$$

As can be seen in Figure S8, likelihood profiles are in qualitative agreement with the bootstrap estimates. Bootstrap 95% confidence interval correspond roughly to the range 10-20 log-likelihood units (in base e). Note that the profiles display composite likelihoods since linkage between neighboring sites in the Poisson Random Field approach creates correlations between neighboring sites. As a result, we expect that likelihood-based confidence intervals provide an underestimate of the true variance, better captured by the bootstrap analysis.

[1] Harrow, J et al. (2006) Gencode: producing a reference annotation for encode. Genome Biol **7**, S4.1–9.

[2] Hernandez, R. D, Williamson, S. H, & Bustamante, C. D. (2007) Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol **24**, 1792–1800.

[3] Burnham, K & Overton, W. (1978) Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika **65**, 625–633.

[4] Burnham, K & Overton, W. (1979) Robust estimation of population size when capture probabilities vary among animals. Ecology **60**, 927–936.

[5] Link, WA (2003) Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities Biometrics **59**, 1123-1130.

[6] Link, WA (2006) Rejoinder to: "On identifiability in capture-recapture models". Biometrics **62**, 936-939.

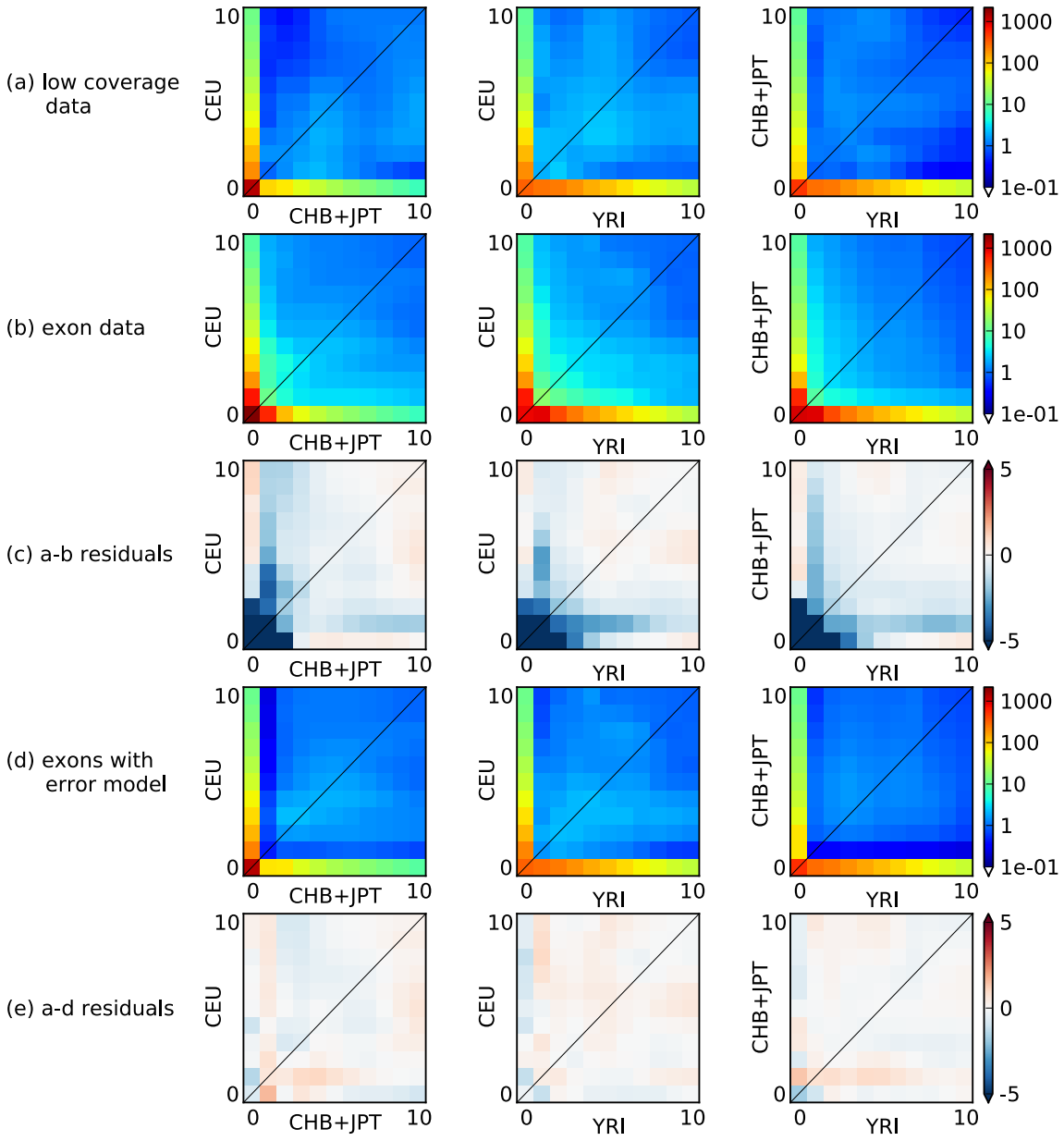

FIG. S4. Two-dimensional marginal frequency spectra for (a) the low-coverage pilot, (b) the exon pilot, and (d) the exon pilot once the error model has been applied. (c) shows the Anscombe residuals between the two pilots, whereas (e) shows the residuals after the error model has been applied to the exon pilot. All spectra are polarized using the March 2006 assembly of the chimp genome (PanTro2).

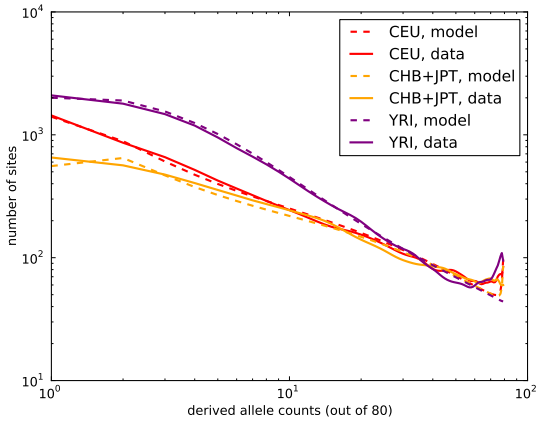

FIG. S5. Comparison of model predictions, based on our maximum-likelihood parameters, and data from 4-fold synonymous sites polarized using chimp as an outgroup. The discrepancy for very common variants is due to cases where the chimp allele is not the ancestral allele [2].

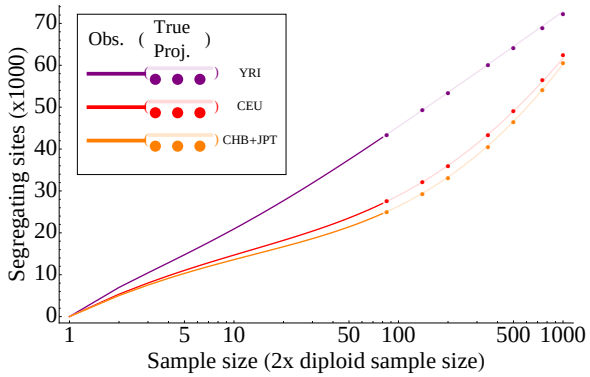

FIG. S6. To control for bias of the jackknife estimator, we used our demographic model to generate "True" SFS for 1000 chromosomes in our 3 population panels. We then calculated average jackknife projections based on "observed" subsamples of 80 chromosomes for each population. In the absence of sampling, we find little bias for the first decade of extrapolation.

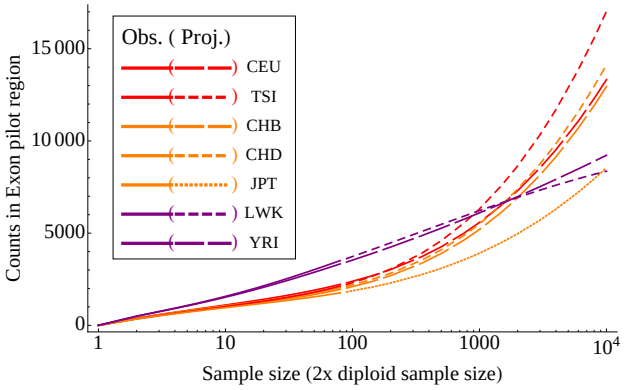

FIG. S7. Observations and third-order jackknife predictions of the number of variants to be discovered in the exon capture panels as sample size is increased, based on 80 chromosomes per population.

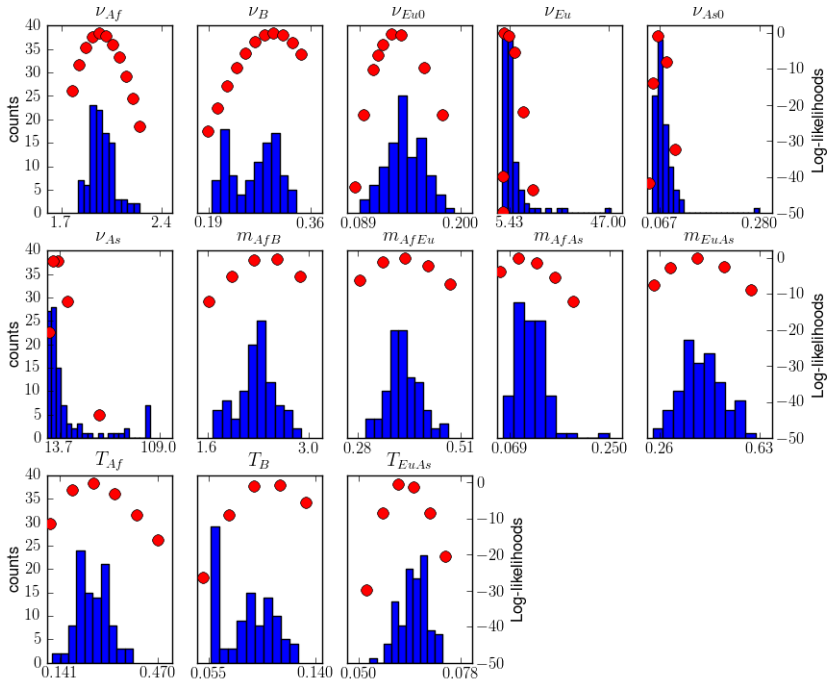

FIG. S8. Bootstrap and likelihood profiles for the Out-Of-Africa model for parameters expressed in genetic units. Likelihood profiles are obtained in the usual way by fixing one parameter and optimizing the likelihoods over the other parameters.