Inference of Human Population History From Whole Genome Sequence of A Single Individual (Supplementary Information)

Heng Li and Richard Durbin

May 4, 2011

1 Coalescent simulation

Coalescent simulation was done by msHOT (Hudson, 2002; Hellenthal and Stephens, 2007). Ancestral population sizes used in simulation are shown in the main paper. The scaled mutation and recombination rates were set to those inferred from YH. One hundred 30Mbp diploid sequences were simulated with the same parameters. We call this simulation the *standard simulation* as it follows a standard coalescent-with-recombination process (Hudson, 1983; Griffiths and Marjoram, 1996). The *ms* command is:

```
ms 2 100 -t 81960 -r 13560 30000000 -eN 0.01 0.05 -eN 0.0375 0.5 -eN 1.25 1
```

1.1 Alternative simulation

To verify PSMC, we also tried five alternative demographic history (namely, sim-1, sim-2, sim-3, sim-YH, sim-split and sim-split2, respectively):

```
ms 2 100 -t 30000 -r 6000 30000000 -eN 0.01 0.1 -eN 0.06 1 -eN 0.2 0.5 -eN 1 1 -eN 2 2
ms 2 100 -t 3000 -r 600 30000000 -eN 0.1 5 -eN 0.6 20 -eN 2 5 -eN 10 10 -eN 20 5
ms 2 100 -t 60000 -r 12000 30000000 -eN 0.01 0.05 -eN 0.0150 0.5 -eN 0.05 0.25 -eN 0.5 0.5
ms 2 100 -t 65130.39 -r 10973.82 30000000 -eN 0.0055 0.0832 -eN 0.0089 0.0489 \
         -eN 0.0130 0.0607 -eN 0.0177 0.1072 -eN 0.0233 0.2093 -eN 0.0299 0.3630 \
         -eN 0.0375 0.5041 -eN 0.0465 0.5870 -eN 0.0571 0.6343 -eN 0.0695 0.6138 \
         -eN 0.0840 0.5292 -eN 0.1010 0.4409 -eN 0.1210 0.3749 -eN 0.1444 0.3313 \
         -eN 0.1718 0.3066 -eN 0.2040 0.2952 -eN 0.2418 0.2915 -eN 0.2860 0.2950 \
         -eN 0.3379 0.3103 -eN 0.3988 0.3458 -eN 0.4701 0.4109 -eN 0.5538 0.5048 \
         -eN 0.6520 0.5996 -eN 0.7671 0.6440 -eN 0.9020 0.6178 -eN 1.0603 0.5345 \
         -eN 1.4635 1.7931
ms 2 100 -t 10000 -r 2000 10000000 -I 2 1 1 -n 1 1 -n 2 1 -ej 0.06 2 1 -n 0.06 1
ms 2 100 -t 10000 -r 2000 10000000 -T -l -I 2 1 1 -eM 0 4 -eN 0 1 -en 0.01 1 0.1 \
         -eM 0.06 0 -ej 0.06 2 1 -eN 0.06 1 -eN 0.2 0.5 -eN 1 1 -eN 2 2
```

where sim-1 (Figure S1a) represents a history similar to the PSMC estimate for non-African populations, sim-2 (Figure S1b) evaluates if the large recent population size is an innate defect of our PSMC model; sim-3 (Figure S1c) shows the accuracy given a very sharp bottleneck; sim-YH (Figure S1d) checks if PSMC may recover the history estimiated by itself (the blue line in the figure is the PSMC estimate for CHN.A); sim-split (Figure S1e) shows the PSMC estimate given two constant-sized populations split at 60kya; sim-split2 (Figure S1f) simulates a little more plausible history between African and non-African populations with one population going through a severe bottleneck but the other not.

To quantify the accuracy of the PSMC estimate, we define the following metric:

$$d(t_0,t_1)=\frac{1}{\log t_1-\log t_0}\int_{t_0}^{t_1}\frac{|N_0(t)-N_1(t)|}{N_0(t)+N_1(t)}\frac{dt}{t}$$

which is the averge fraction difference in the logarithm scale in the interval $[t_0,t_1)$. Due to the way d is defined, it is not affected by the scaling of t and N. We computed $d(10\text{kya},2\text{Mya})$ for

simulations without admixtures. The following table gives the result:

Simulation	Figure	d(10kya, 2Mya)
plain	Figure 2a	0.13
sim-1	Figure S1a	0.10
sim-2	Figure S1b	0.10
sim-3	Figure S1c	0.12
sim-YH	Figure S1d	0.06

In general, PSMC does well in recovering the history, although it may smooth out steep changes in the population size. In addition, the simulation implies that PSMC is reasonably good at estimating the time when two populations split (Figure S1d). After this split point, there are no coalescences, which PSMC reflects as an infinite population size.

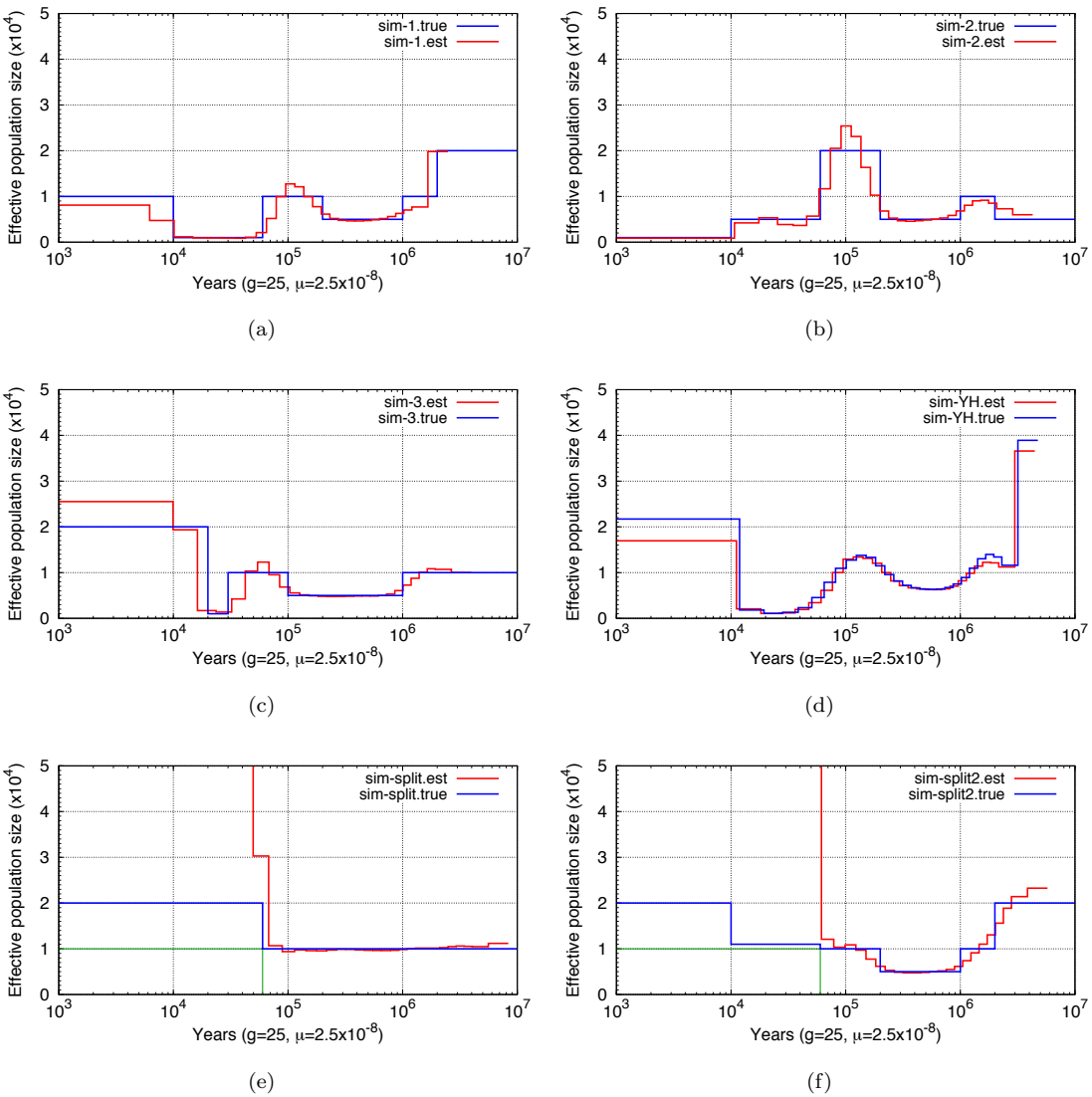
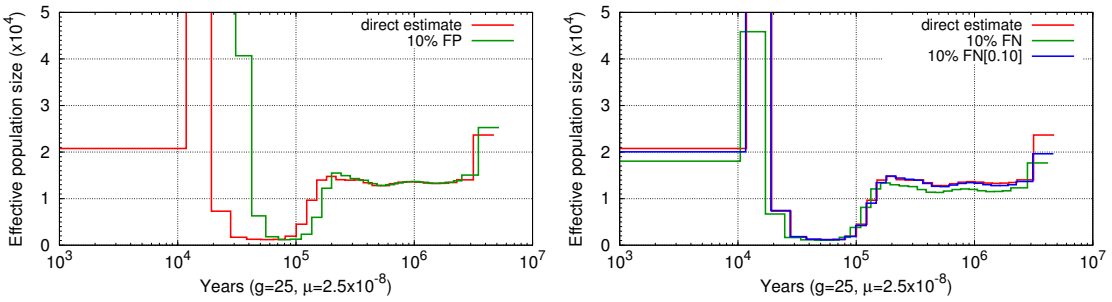


Figure S1: PSMC estimate on simulated data as in Section 1.1.

1.2 Simulation with uniform SNP ascertainment errors



(a) (b)

Figure S2: Effect of SNP ascertainment errors. (a) Uniform false heterozygotes. (b) Uniform missing heterozygotes.

The SNP ascertainment may have errors. To see how PSMC performs on data with false SNPs, we randomly added 10% more heterozygotes to the sequence generated by the standard simulation and then run PSMC to estimate the history. We see that adding uniform false heterozygotes (FP) increases the TMRCA in all time frames and pushes the estimate away from the origin along the X axis (Figure S2a). Adding FP also breaks long recent segments and decreases the number of recent recombinations, which has a similar effect to increased recent population sizes.

Insufficient read depth may lead to missing heterozygotes in an essentially uniform manner. To simulate this scenario, we uniformly changed 10% of heterozygotes to homozygotes on the sequence generated from the standard simulation. We see that the estimate given missing heterozygotes (FN) is shifted toward the origin (the green curve in Figure S2b). Fortunately, FN is largely equivalent to lower neutral mutation rate μ. We can correct for FN by reducing μ (the blue curve).

1.3 Simulation with long hypermutated regions

Balancing selection or false heterozygotes caused by segmental duplications may lead to excessively long segments with high heterozygosity. To see the effect of these segments, we simulated three 30Mbp diploid sequences with mutation rate 10 times higher than that used in the standard simulation. The *ms* command is:

```
ms 2 3 -t 819600 -r 13560 30000000 -T -eN 0.01 0.05 -eN 0.0375 0.5 -eN 1.25 1
```

We mixed the three sequences with the 100 sequences generated from the standard simulation and run PSMC to infer population sizes. Figure S3 indicates that these hypermutated segments may lead to excessively large ancient population sizes, but the rest of curve is essentially unaffected.

1.4 Simulation with variable mutation rates

1.4.1 Calculating regional human-macaque mutation rate

We downloaded the 4-way EPO alignment between human, chimpanzee, orangutan and macaque from Ensembl FTP[1] (v50) and converted the EMF format to Multiple Alignment Format (MAF) with emf2maf.pl[2] available from Ensembl's EURS repository. We excluded alignments containing paralogous regions and removed the columns containing gaps or ambiguous bases in either human or macaque. 2,156,898,990 columns remained on autosomes with 130,132,715 substitutions. We calculated the mean divergence in each 20Kbp sliding window with a step 100bp. Windows with less than 10,000 columns in the EPO alignment were dropped. 24,102,686 windows were left after

[1] ftp://ftp.ensembl.org/pub/release-50/emf/ensembl-compara/epo_4_catarrhini/
[2] http://cvs.sanger.ac.uk/cgi-bin/viewcvs.cgi/*checkout*/ensembl-compara/scripts/dumps/emf2maf.pl

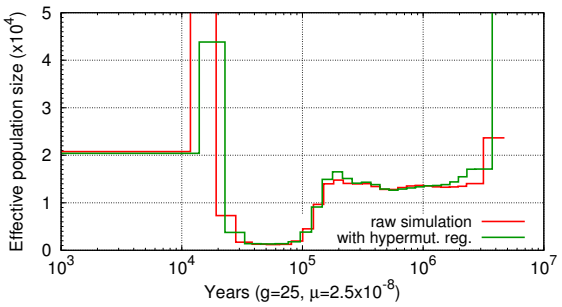


Figure S3: Effect of long hypermutated regions.

this process. The mean substitution rate of these windows is 6.03% with a standard deviation 0.76%. Figure S4 shows the substitution rate as a function of coordinate on chromosome 6.

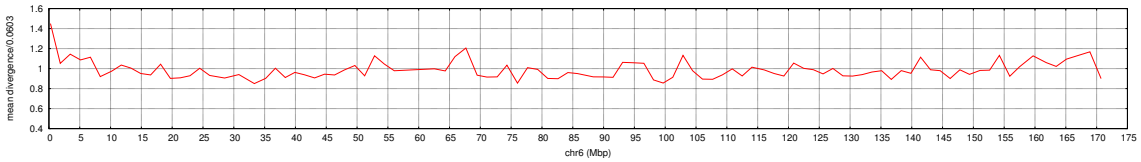


Figure S4: Substitution rate as a function of coordinate on human chromosome 6. The curve was smoothed with the Bezier method provided by gnuplot.

1.4.2 Simulating sequences under variable mutation rate

After calculating the regional mutation rate, we got at each position i on the human genome a pair of values $(a_i, r_i) \in \{0, 1\} \times \Re^+$, where r_i is the average mutation rate in the 20Kbp window between $[i - 10000, i + 10000]$ divided by the overall substitution rate 0.0603, and a_i indicates whether such a window is dropped in calculation as is described in the previous section.

In simulation, we randomly extracted 30Mbp region from the human genome starting from i and generated a 30Mbp diploid sequence as follows. Given the k-th position on the simulated sequence, we attached an ambiguous base 'N' if a_{i+k} equals 0; if a_{i+k} equals 1, we attached a heterozygote with a probability $e^{-\theta t_k r_{i+k}}$ or a homozygote otherwise. Here θ is the scaled mutation rate used in simulation and t_k is the TMRCA at position k according to the local coalescent tree given by *ms* in the standard simulation.

1.5 Simulation with recombination hotspots

To simulate recombination hotspots, we obtained the hotspot map from HapMap Release 21. We generated three hundred 10Mbp sequences with the hotspots on each sequence drawn from a randomly selected 10Mbp region from the hotspot map. We assumed the recombination rate in each hotspot is ten times higher than the non-hotspot regions.

Note that our PSMC model essentially uses the distribution of heterozygosity to infer parameters, while local variation in recombination rate has little effect on this distribution. Therefore our model is robust to hotspots as is shown in Figure 2 in the main text.

1.6 Simulation with structured population

1.6.1 Structured population and effective population size

Assume in a constant-sized population the ancestral population split into two equal-sized subpopulations at time t which joined back at time s ($s < t$). The probability that two lineages at time s

coalesce before t is $(1-e^{-2(t-s)})/2$, which is the probability that the two lineages are chosen from the same subpopulation and then coalesce before t. In contrast, in a constant-sized population without structure, the probability of two lineages at s coalescing before t is $1-e^{-(t-s)}$. We have

$$\left[1-e^{-(t-s)}\right]-\frac{1}{2}\left[1-e^{-2(t-s)}\right]=\frac{1}{2}\left[1-e^{-(t-s)}\right]^2>0$$

which means coalescences occur less frequently in the structured population, and the longer the split the more significant the effect. As having fewer coalescences is equivalent to a larger effective population size, the effective population size in the structured population between time s and t is larger than the sum of the sizes of the two sub-populations.

1.6.2 Effect of structured population

We took the PSMC estimate on the YRI autosomes, changed the original estimate by removing the hump in the size history between 30 and 600kya, and added population split and admixture in a simulation to study the effect of structured population. We assumed at 250kya the ancestral population split into two subpopulations with equal sizes which admixed to one at 60kya. The sum of the sizes of the subpopulations remain the same across 60-250kya. The *ms* command-line for this simulation is:

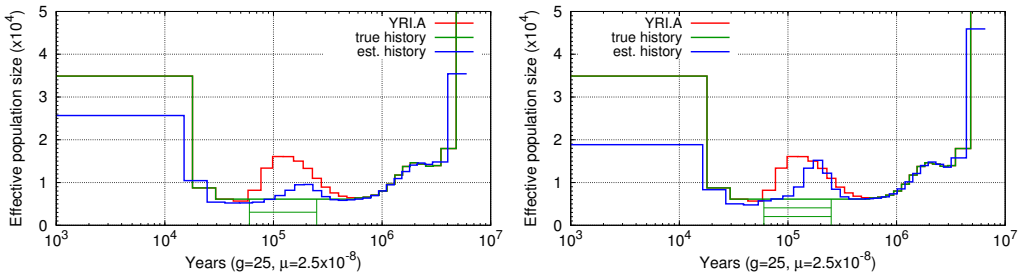
```
ms 2 100 -t 104693 -r 13862 30000000 -T -eN 0.0052 0.2504 \
  -eN 0.0084 0.1751 -es 0.0172 1 0.5 -en 0.0172 1 0.08755 \
  -en 0.0172 2 0.08755 -ej 0.0716 2 1 -eN 0.0716 0.1833 \
  -eN 0.1922 0.1885 -eN 0.2277 0.2022 -eN 0.2694 0.2295 \
  -eN 0.3183 0.2754 -eN 0.3756 0.3367 -eN 0.4428 0.3939 \
  -eN 0.5216 0.4190 -eN 0.6141 0.4104 -eN 0.7225 0.3954 \
  -eN 0.8496 0.3998 -eN 0.9987 0.5144 -eN 1.3785 1.8311
```

Figure S5a shows that during the period of population split, PSMC predicts a larger effective population size than the sum of sizes of sub-populations. The effect is stronger if the population split into three smaller subpopulations (Figure S5b). This is expected in theory given the discussion in the previous section. The *ms* command line used in Figure S5b is:

```
ms 2 100 -t 104693 -r 13862 30000000 -T -eN 0.0052 0.2504 \
  -eN 0.0084 0.1751 -es 0.0172 1 0.33333 -es 0.0172 2 0.5 -en 0.0172 1 0.08755 \
  -en 0.0172 2 0.08755 -ej 0.0716 3 2 -ej 0.0716 2 1 -eN 0.0716 0.1833 \
  -eN 0.1922 0.1885 -eN 0.2277 0.2022 -eN 0.2694 0.2295 \
  -eN 0.3183 0.2754 -eN 0.3756 0.3367 -eN 0.4428 0.3939 \
  -eN 0.5216 0.4190 -eN 0.6141 0.4104 -eN 0.7225 0.3954 \
  -eN 0.8496 0.3998 -eN 0.9987 0.5144 -eN 1.3785 1.8311
```

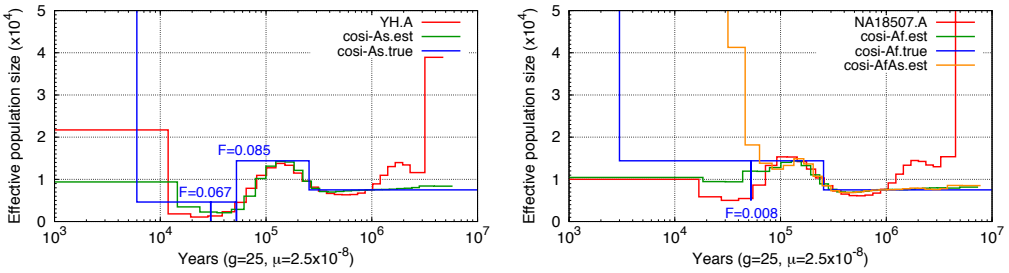
1.7 Simulation from Schaffner *et al.* (2005)

We estimated the population history on diploid sequences simulated from the best fit model by Schaffner et al. (2005) which considers variable recombination rates, recombination hotspots, migration and gene conversion. PSMC still works reasonably well (Figure S6). However, the reconstructions from simulated data are not able to reproduce the mild bottleneck 20–60kya that we observe in he African (YRI) data, or very large ancestral population size beyond 1Mya as is seen in real data. In addition, even given the migrations between Asian and African populations in the simulation, PSMC predicts a sudden rise in population size right at the split of African and Asian populations around 52kya (Figure S6b, orange line), which is different from the estimate from the NA18507-CHN pseudo-diploid X chromosome comparison, showing that the extended genetic exchange we observe between Africa and Asia in the real data is not produced by the method if the real final split time is earlier.



(a) (b)

Figure S5: Effect of population split and admixture. (a) population split to two equal-sized subpopulations at 250kya which admixed at 60kya. (b) population split to three equal-sized subpopulations. In both panels, the red curves are the original PSMC estimate on YRI autosomes. The green curves show the ancestral population size used in *ms* simulation, or the sum of subpopulation sizes during the population split. The thin green lines indicate the period of population split and the size of each subpopulation. The blue curves give the PSMC estimates on the simulated sequences.



(a) (b)

Figure S6: PSMC estimate on data simulated from the best fit model by Schaffner et al. (2005). (a) Asian diploid genome. (b) African diploid genome and African-Asian hybrid genome. In both figures, vertical short blue lines indicate the position of bottlenecks in simulation with inbreeding coefficient (F) labeled nearby.

2 PSMC inference for data sets

2.1 PSMC inference for human individuals

We have also applied the PSMC method on the two trios sequenced by the 1000 Genomes Project, as well as the COLO-829-BL genome (European ancestry; Pleasance et al., 2010) and the NA18506 and NA18508 genomes (SRA009225 and SRA009347, respectively). Figure S7a and S7c indicates that the PSMC results on automsomes are highly consistent except for the very recent history, demonstrating the power of using whole-genome data. Estimates on X chromosomes (Figure S7b and S7d) are noisier, but estimates from similar ancestry still well agree with each other.

2.2 PSMC inference for orangutan individuals

We downloaded from the Short Read Archive the short read sequences for three orangutan individuals (Locke et al., 2011), two Borneans (KB5404 and KB4204) and one Sumatran (KB5883), and processed the data in the same way as we processed the human sequences (i.e. alignment with BWA and consensus calling with SAMtools). As KB4204/Bornean2 and KB5883/Sumatran are both males, we are able to construct a pseudo-diploid X chromosome to investigate the divergence time between the two Orangutan species.

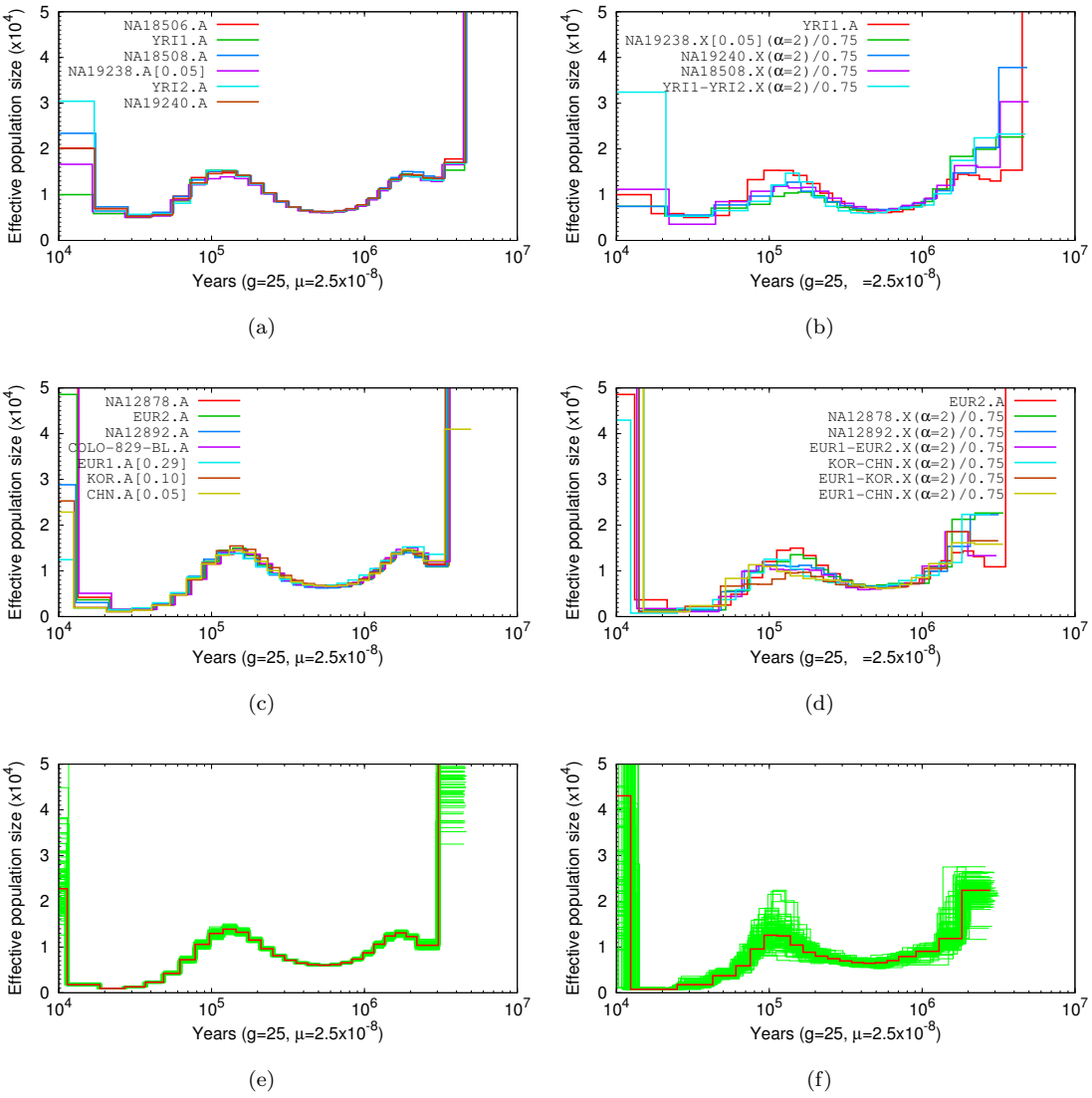


Figure S7: (a) PSMC estimate for Yoruban autosomes. (b) Estimate for Yoruban X chromosome. (c) Estimate for non-African autosomes. (d) Estimate for non-African X chromosomes (e) Block bootstrapping for Korean autosomes (KOR.A). Thin green lines represent 100 rounds of resampling. (f) Block bootstrapping for Korean-Chinese combined pseudo-diploid X chromosome (KOR-CHN.X).

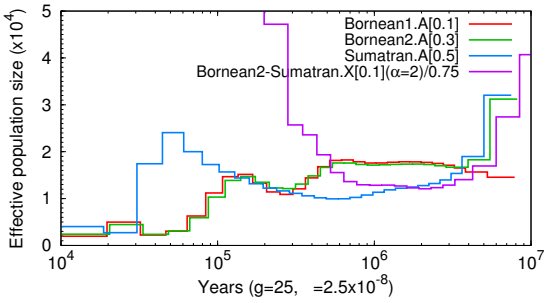


Figure S8: PSMC inference for three individuals from two orangutan species, Bornean and Sumatran.

Figure S8 shows the PSMC estimate of the population size history of the two species. Notably, although KB5404/Bornean1 has much more data, the inferred history is nearly identical to that of KB4204/Bornean2 when the false negative rate is considered, which again reveals the robustness of PSMC. From the figure, Sumatran (the blue curve) may appear to deviate from Bornean (the red and green curves) a few million years ago, but the relatively small population size inferred from the pseudo-diploid X chromosome (the purple curve) between 500kya and 5Mya indicates that the ancestral populations of Bornean and Sumatran largely remained as one population in this period. The effective population size inferred from the X chromosome started to increase several hundred thousand years ago and rapidly went to infinity at 300kya or so, which should imply that the process of speciation of the two orangutan species may last hundreds of thousands of years, and the final gene flow may occur around 300kya, consistent with the estimate of 334 ± 145kya by Locke et al. (2011) using the coal-HMM method (Hobolth et al., 2007).

3 Other potential artifacts in PSMC estimate

3.1 Rescaling to real time

Species	Substitutions (%)	Divergence time (Mya)	mutation per-year ($\times 10^{-9}$)
Chimpanzee	1.31	7	0.94
Orangutan	3.28	18	0.91
Macaque	6.03	25	1.21

Table S1: Mutation rate per site per year between human and other primates. Pairwise substitution rates are estimated based on the Ensembl 4-way EPO whole-genome alignment. Human-chimp divergence time is taken from Patterson et al. (2006), human-orangutan from Satta et al. (2004), and human-macaque from Rhesus Macaque Genome Sequencing and Analysis Consortium (2007).

The TMRCA estimated by the PSMC model is in the units of mutation per site. To rescale TMRCA in the units of years, we need to know the mutation rate per site per year, which can be estimated by using closely related species. Table S1 implies that in primates, the mutation rate is broadly around 10^{-9} per site per year, the rate we used in rescaling the PSMC estimate (we assumed a 2.5×10^{-8} mution rate per site per generation and a 25-year generation time, which is translated to a 1.0×10^{-9} mutation rate per site per year).

However, recent direct measurement using whole genome sequences in pedigrees suggest that in the individuals examined the mutation rate per site per generation approaches 10^{-8} (Roach et al., 2010; 1000 Genomes Project Consortium, 2010), twice smaller than the rate we use. Nonetheless, what matters for population genetic based methods such as PSMC is the time average. A comparatively small fraction of higher mutation rates could change this average significantly. We therefore feel that although direct measurements are clearly valuable, there are not enough yet to change the mutation rates used in population genetic based analyses.

3.2 Inaccuracy in scaled recombination rate

PSMC is not good at inferring recombination events that result in small changes in TMRCA (Figure S9). Thus it may systematically underestimate the recombination rate. In the standard simulation, the ratio of scaled mutation rate and recombination rate (θ/ρ) is set to 5, but the PSMC estimates this ratio as 8.63 ± 0.10, significantly deviating from 5. However, as seen in Figure S1, this underestimate of the scaled recombination rate does not seem to affect the accuracy of the PSMC estimate.

On real data, PSMC estimates that there are 2.3×10^5 recombination events from the Chinese autosomes, which amounts to 12kbp unrecombined blocks in average. The average length of unrecombined blocks drops to 10kbp for a Yoruban genome due to the larger effective population size. However, due to the underestimated recombination rate by PSMC, the true length of unrecombined blocks should be smaller.

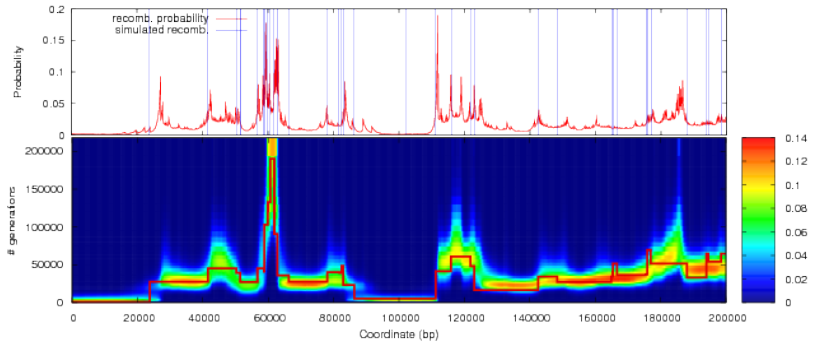
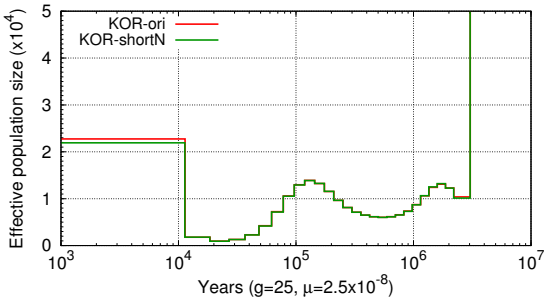


Figure S9: Estimating recombination events. PSMC tends to miss recombination events that lead to small changes in TMRCA.

3.3 Effect of ambiguous bases

Most of the sequences in the telomere and centromere regions are not present in the human reference genome or highly repetitive. They appeared as 'N's in the input of PSMC. To see whether long stretches of 'N's may affect the PSMC estimate, we remove contiguous N longer than 100Kbp and split the input sequence there. The resulting estimate is almost identical to that without this preprocessing (Figure S10).



(a)

Figure S10: Effect of ambiguous bases.

3.4 Effect of the ratio of male-to-female mutation rate

There are debates about the value of the ratio of male-to-female mutation rates α (Ebersberger et al., 2002; Makova and Li, 2002; Taylor et al., 2006; Burgess and Yang, 2008), which may affect time scaling of the estimate on X chromosomes, given that:

$$\mu_X = \mu_A \cdot \frac{2(2+\alpha)}{3(1+\alpha)}$$

Nonetheless, for α ranged between 2 and 5, μ_X only varies slightly from 2.22×10^{-8} to 1.94×10^{-8} if we assume $\mu_A = 2.5 \times 10^{-8}$, and may not impact our conclusion (Figure S11).

3.5 Convergence of Baum-Welch iteration

We apply 20 rounds of Baum-Welch iterations from the constant-sized history. Figure S12a shows that the bottleneck between 11–50kya is obvious even after a single round of iteration. Estimate between 50kya–2Mya takes more iterations to get stabilized, but the estimate is not changed much

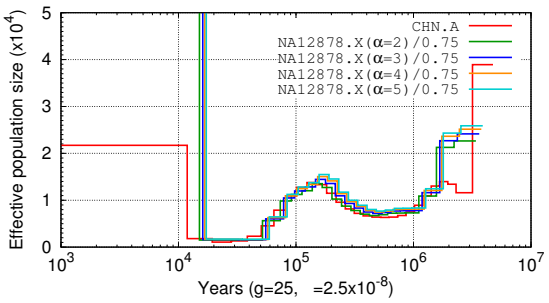
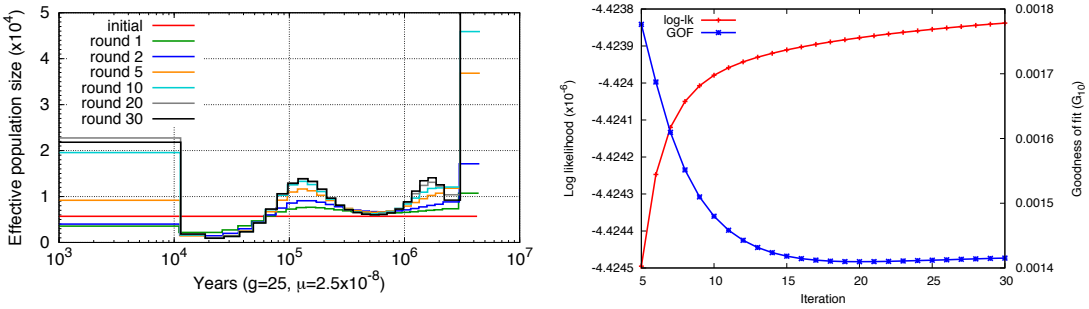


Figure S11: Effect of the ratio of male-to-female mutation rate (α).

after the 10th iteration. In addition, although the likelihood of the data continues to increase after 30 iterations, the goodness of fit (GOF) statistics G_{10} gets worse after 20 (Figure S12b), which implies that applying more iterations may lead to overfitting.



(a) (b)

Figure S12: Convergence and goodness of fit (GOF) of PSMC estimate. (a) PSMC estimate given different rounds of iterations. (b) Log likelihood (log-LK) and GOF as a function of iterations, starting from constant-sized history.

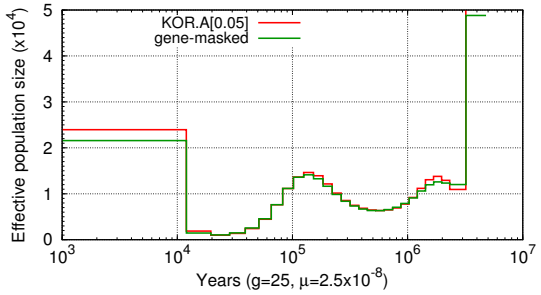


Figure S13: PSMC estimate on KOR.A with exons and their 10kbp (5kbp on each side of an exon) flanking regions masked out. Estimate across the entire autosomes is assumed to have about 5% lower mutation rate than in non-coding regions.

3.6 Effect of coding regions

In comparison to non-coding regions, coding regions are expected to be subjected to purifying selection, which may further affect the heterozygosity of the flanking regions. To see if selection

around coding regions may change the PSMC estimate, we acquired the GenCode annotation of human genes and masked out all exons plus their 10kbp (5kbp on each side of an exon) flanking regions. About 1.0Gbp sequences were masked consequently. The observed heterozygosity in the regions left after masking is 5% higher than the genome average. Figure S13 implies that if we correct for this difference in mutation rate, the PSMC estimates from all autosomes and from non-coding regions are nearly identical.

4 Comparison to previous studies

4.1 Inbreeding coefficient

The probability that two lineages coalesce in the time interval $[s, t)$ is:

$$F(s,t) = 1 - e^{-\int_s^t \frac{du}{\lambda(u)}}$$

where $\lambda(t) = N(t)/N_0$ is the relative population size and the time scale is the number of generations divided by $2N_0$ (Griffiths and Tavare, 1994). $F(s,t)$ is called the *inbreeding coefficient* or the *intensity* between time s and t.

4.2 Other studies on population divergence and the size history

Table S2 shows the divergence time between African and non-African populations, the timing and strength of the bottleneck, estimated in other studies. All the estimates are based on nuclear DNA data.

It is clear that the divergence time between African and non-African populations varies greatly between studies. This is mostly caused by the assumptions of the basic demographic models. Garrigan et al. (2007), Fagundes et al. (2007) and Cox et al. (2008) assume population is reduced immediately following the divergence of African and non-African populations. Forced by the timing of the population reduction, these models are unlikely to predict deep divergence time. The very recent divergence time by Cox et al. (2008) may also be related to the use of X chromosome data exclusively. On the other extreme end, Gutenkunst et al. (2009) infer a very deep divergence time, but they also infer a large migration rate (2.5×10^{-4}) between 140kya and 21.2kya, which compensates for the early population split. Another possible cause of the difference is the different data ascertainment procedures (e.g. autosomal data vs. X-linked data, noncoding only vs. full gene sequences, and genotyping vs. resequencing).

Most of the studies in Table S2 broadly agree on the timing and the strength of the reduction/bottleneck, even though some (Marth et al., 2004; Keinan et al., 2007; Wall et al., 2009) model this by a piecewise constant history, some (Adams and Hudson, 2004; Garrigan et al., 2007; Cox et al., 2008) by reduction followed by exponential growth and some (Schaffner et al., 2005; Gutenkunst et al., 2009) by a mixture of reduction and bottlenecks. The differences between these results can be contributed to the assumptions made in the demographic models; the difference in ascertainment procedures may also be a major cause given that similar methods/models may lead to different conclusions (e.g. Marth et al. (2004) vs. Keinan et al. (2007)). Our PSMC model does not explicitly infer the bottleneck. Nonetheless, when we generated an AFS from our CHN.A estimate and fit the AFS with a 3-epoch model (Marth et al., 2004), we got a bottleneck between 12–40kya with an inbreeding coefficient 0.36, consistent with other studies. However, when we applied the same method to YRI.A, we are unable to recover the mild bottleneck in YRI, which may indicate that the lack of African bottleneck in AFS-based studies might be due to the lack of power in fitting the AFS.

As to the comparison with individual studies, the PSMC estimates on YRI.A and CHN.A well agree with Schaffner et al. (2005) in that the PSMC estimates on data simulated from the best-fit model are very similar to the estimates on real data (Figure S6). Nonetheless, the recent genetic exchanges inferred from YRI-CHN.X disagree. This is possibly because Schaffner et al. (2005) assume a uniform migration rate after the out-of-African event, while our model implies a large migration rate beyond 20kya but a low rate afterwards, more similar to the model by Gutenkunst et al. (2009).

	Method [1]	T_d [2] (kya)	T_b [3]	T_e [4] (kya)	F [5]	Data	Comment
Marth et al. (2004)	AFS	N/A	76	64	0.095	42×33k sites	
Adams and Hudson (2004)	coal-sim	N/A	30	N/A	N/A	Seattle SNPs	Size reduction followed by exponential growth.
Schaffner et al. (2005)	coal-sim	52.5	52.5/30	N/A	0.085/0.067	3988 sites, genotyping	Two sharp bottlenecks mixed with reduced population size (Figure S6)
Plagnol and Wall (2006)	coal-sim	130	60	N/A	0.20/0.24	34 × 3.2Mb, EGP	0.20 for model without introgression; 0.24 with introgression; F calculated from the *ms* command line
Liu et al. (2006)		56.1	N/A	N/A	N/A		
Garrigan et al. (2007)	coal-sim	39.5	39.5	N/A	N/A	431 × 16kb from M,X,Y	Quoting the divergence time between Dogon and Mongolian. Reduction followed by exponential growth.
Fagundes et al. (2007)	coal-sim	51.1	51.1	N/A	N/A	50 × 25kb from auto.	Reduction followed by exponential growth.
Keinan et al. (2007)	AFS	N/A	23 ± 2	N/A	0.18 ± 0.01	HapMap and Perlegen	One-bottleneck model. Bottleneck at 32 ± 3 for European.
Cox et al. (2008)	coal-sim	27.7	N/A	N/A	N/A	100 × 100kb from X	Quoting the divergence time between Han and Biaka. Divergence time estimed on chrX with migration modeled.
Wall et al. (2009)	coal-sim	80	30+1	30	0.27/0.33	58 × 5.3Mb, EGP	0.27 without introgression; 0.33 with introgression; bottleneck duration fixed at 1kya; Eu-Af diverged at 120kya.
Gutenkunst et al. (2009)	diffusion	140	140/21.2	N/A	N/A	EGP, non-coding only	Two bottlenecks at the Af–Eu/As divergence and at the Eu–As divergence.

Table S2: Inferred out-of-Africa event in current literature. Underlined numbers are fixed (not inferred) in the model. [1] 'coal-sim' denotes the category of methods that use a coalescent simulator to fit the observed data; 'AFS' denotes methods that explicitly compute the likelihood of an allele frequency spectrum given a piecewise constant population size history; 'diffusion' denotes the diffusion approximation on the joint allele frequency spectrum. [2] The divergence time of African and non-African populations. If the time is in unit of generations in the literature, it is scaled to years under the assumption of 10^{-9} mutation per site per year (equivalent to $\mu = 2.5 \times 10^{-8}$ and $g = 25$). [3] Time of the start of the bottleneck or population size reduction. [4] Time of the end of the bottleneck. [5] Inbreeding coefficient. F can only be calculated when the model explicitly models the bottleneck.

In contrast to the variability of the studies using nuclear DNA, studies using mitochondrial DNA (mtDNA) (Ingman et al., 2000; Macaulay et al., 2005) are highly consistent, probably because of the similarity in the demographic model and data processing as well. These studies found that the M and N mtDNA haplogroups coalesced around 60kya. However, timing with mtDNA using the current method gives the genetic divergence which predated the population divergence. In addition, it is possible to infer the population size history from mtDNA (Atkinson et al., 2008), but the variance is large as mtDNA is short.

4.3 Comparison to the palaeoanthropological evidence

Fossil evidence supports that anatomically modern humans had migrated to Europe during 41–46kya (Mellars, 2006b), to Malaysia by around 45kya (Barker et al., 2002) and to Australia by at least 45kya (Stringer, 2002; Mellars, 2006a). Although anatomically human fossils identified at Skhul and Qafzeh in Israel were dated back to 100–135kya (Vanhaereny et al., 2006), this migration is believed to be early unsuccessful dispersal by some researchers (Mellars, 2006a).

On the other hand, several studies using nuclear DNA placed the East Asian-European divergence around 17–25kya (Keinan et al., 2007; Garrigan et al., 2007; Gutenkunst et al., 2009). Our PSMC estimate from the combined Venter and YH X chromosomes is also very recent (Figure S7d). This leads to the apparent inconsistency with the fossil evidence that anatomically modern human have spread across the continent by at least 40kya. One of the possible explanations is that during the Last Glacial Maximum at about 20kya, the non-African populations retreated southward (Forster, 2004), and gene flows may have occurred between the different populations again. Under this hypothesis, the recent gene flow between YRI.X and KOR.X would be reasonable, although autosomal data from more populations are needed to further confirm the existence of the recent gene flow.

References

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–73.

Adams, A. M. and Hudson, R. R. (2004). Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics*, 168:1699–1712.

Atkinson, Q. D., Gray, R. D., and Drummond, A. J. (2008). mtdna variation predicts population size in humans and reveals a major southern asian chapter in human prehistory. *Mol Biol Evol*, 25:468–474.

Barker, G. et al. (2002). Prehistoric foragers and farmers in south-east asia: renewed investigations at niah cave, sarawak. *Proceedings of the Prehistoric Society*, pages 147–164.

Burgess, R. and Yang, Z. (2008). Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol Biol Evol*, 25:1979–1994.

Cox, M. P., Woerner, A. E., Wall, J. D., and Hammer, M. F. (2008). Intergenic dna sequences from the human x chromosome reveal high rates of global gene flow. *BMC Genet*, 9:76.

Ebersberger, I., Metzler, D., Schwarz, C., and Paabo, S. (2002). Genomewide comparison of dna sequences between humans and chimpanzees. *Am J Hum Genet*, 70:1490–1497.

Fagundes, N. J. R. et al. (2007). Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci*, 104:17614–17619.

Forster, P. (2004). Ice ages and the mitochondrial dna chronology of human dispersals: a review. *Philos Trans R Soc Lond B Biol Sci*, 359(1442):255–64; discussion 264.

Garrigan, D., Kingan, S. B., Pilkington, M. M., Wilder, J. A., Cox, M. P., Soodyall, H., Strassmann, B., Destro-Bisol, G., de Knijff, P., Novelletto, A., Friedlaender, J., and Hammer, M. F. (2007). Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, x and y chromosome resequencing data. *Genetics*, 177(4):2195–207.

Griffiths, R. C. and Marjoram, P. (1996). An ancestral recombination graph. In Donnelly, P. and Tavare, S., editors, *IMA volume on mathematical population genetics*, pages 257–270. Berlin: Springer.

Griffiths, R. C. and Tavare, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci*, 344:403–410.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet*, 5(10):e1000695.

Hellenthal, G. and Stephens, M. (2007). mshot: modifying hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics*, 23:520–521.

Hobolth, A., Christensen, O. F., Mailund, T., and Schierup, M. H. (2007). Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS Genet*, 3:e7.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*, 23:183–201.

Hudson, R. R. (2002). Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338.

Ingman, M., Kaessmann, H., Pääbo, S., and Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature*, 408(6813):708–13.

Keinan, A., Mullikin, J. C., Patterson, N., and Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in east asians than in europeans. *Nat Genet*, 39:1251–1255.

Liu, H., Prugnolle, F., Manica, A., and Balloux, F. (2006). A geographically explicit genetic model of worldwide human-settlement history. *Am J Hum Genet*, 79(2):230–7.

Locke, D. P. et al. (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, 469(7331):529–33.

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., Blackburn, J., Semino, O., Scozzari, R., Cruciani, F., Taha, A., Shaari, N. K., Raja, J. M., Ismail, P., Zainuddin, Z., Goodwin, W., Bulbeck, D., Bandelt, H.-J., Oppenheimer, S., Torroni, A., and Richards, M. (2005). Single, rapid coastal settlement of asia revealed by analysis of complete mitochondrial genomes. *Science*, 308(5724):1034–1036.

Makova, K. D. and Li, W.-H. (2002). Strong male-driven evolution of dna sequences in humans and apes. *Nature*, 416(6881):624–626.

Marth, G. T., Czabarka, E., Murvai, J., and Sherry, S. T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166:351–372.

Mellars, P. (2006a). Going east: new genetic and archaeological perspectives on the modern human colonization of eurasia. *Science*, 313(5788):796–800.

Mellars, P. (2006b). A new radiocarbon revolution and the dispersal of modern humans in eurasia. *Nature*, 439(7079):931–5.

Patterson, N. et al. (2006). Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441:1103–8.

Plagnol, V. and Wall, J. D. (2006). Possible ancestral structure in human populations. *PLoS Genet*, 2:e105.

Pleasance, E. D. et al. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463:191–6.

Rhesus Macaque Genome Sequencing and Analysis Consortium (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, 316:222–34.

Roach, J. C. et al. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328:636–9.

Satta, Y. et al. (2004). Ancestral population sizes and species divergence times in the primate lineage on the basis of intron and bac end sequences. *J Mol Evol*, 59:478–87.

Schaffner, S. F. et al. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res*, 15:1576–1583.

Stringer, C. (2002). Modern human origins: progress and prospects. *Philos Trans R Soc Lond B Biol Sci*, 357(1420):563–579.

Taylor, J., Tyekucheva, S., Zody, M., Chiaromonte, F., and Makova, K. D. (2006). Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Mol Biol Evol*, 23:565–573.

Vanhaereny, M., d'Errico, F., Stringer, C., James, S. L., Todd, J. A., and Mienis, H. K. (2006). Middle paleolithic shell beads in israel and algeria. *Science*, 312(5781):1785–8.

Wall, J. D., Lohmueller, K. E., and Plagnol, V. (2009). Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol Biol Evol*, 26(8):1823–1827.

The Pairwise Sequentially Markovian Coalescent Model

Heng Li and Richard Durbin

24 January 2008

This document gives the mathematical basis for the PSMC, including all necessary theorems and equations, with discussion. Lemma 1 and 2 give two general facts which will be used later. Theorem 1 proves several central results of the continuous-time PSMC model. This theorem establishes the foundation of the whole PSMC theory. Corollary 4 and Remark 2 show how to calculate or approximate various probabilities when time is discretized. Remark 4 presents the construction of HMM, and Remark 6 and 8 explain several catches in implementation. Remarks 9-11 show methods on estimating the variance and testing the goodness of fit (GOF).

1 PSMC: The Pairwise Sequentially Markovian Model

1.1 General Fomulae

This section presents two lemmas for general functions. Lemma 1 will be used to prove the normalization of the conditioned transition probability in the PSMC continuous-time Markov chain; Lemma 2 will be used to derive the stationary distribution of coalescent time.

Lemma 1. *Given*

$$f(t|s) = h(t)\int_0^{\min\{s,t\}} \frac{g(u)}{\int_0^s g(w)\,dw}\cdot e^{-\int_u^t h(v)\,dv}\,du \tag{1}$$

where $g(t)$ and $h(t)$ are any functions that can be integrated on $[0,\infty)$, the following equation always stands:

$$\int_0^\infty f(t|s)\,dt = 1$$

Proof. Let:

$$t = \phi(\tilde{t})$$

and

$$\tilde{g}(\tilde{u}) = \frac{g(\phi(\tilde{u}))}{h(\phi(\tilde{u}))}$$

where $\phi(\tilde{t})$ satisfies $\phi(0) = 0$ and

$$\phi'(\tilde{u})\cdot h(\phi(\tilde{u})) = 1$$

The integral becomes:

$$f(t|s)\,dt = \frac{f(\phi(\tilde{t})|\phi(\tilde{s}))}{h(\phi(\tilde{t}))}\,d\tilde{t} = \frac{\int_0^{\min\{\tilde{s},\tilde{t}\}}\tilde{g}(\tilde{u})e^{-(\tilde{t}-\tilde{u})}d\tilde{u}}{\int_0^{\tilde{s}}\tilde{g}(\tilde{u})d\tilde{u}}\,d\tilde{t}$$

If we note that for any $g(t)$ that can be integrated:

$$\begin{aligned}
&\int_0^\infty e^{-v}dv\int_0^{\min\{v,t\}} g(u)e^u\,du\\
=\;&\int_0^t g(u)e^u du\left(\int_u^t e^{-v}dv + \int_t^\infty e^{-v}dv\right)\\
=\;&\int_0^t g(u)\,du
\end{aligned}$$

always stands, we get:

$$\int_0^\infty f(t|s)\,dt = \int_0^\infty \frac{\int_0^{\min\{\tilde{s},\tilde{t}\}} \tilde{g}(\tilde{u})e^{-(\tilde{t}-\tilde{u})}d\tilde{u}}{\int_0^{\tilde{s}} \tilde{g}(\tilde{u})d\tilde{u}}\,d\tilde{t} = 1$$

□

Lemma 2 (Stationary distribution)**.** *Let:*

$$\pi(t) = \frac{h(t)}{C}e^{-\int_0^t h(v)dv}\int_0^t g(u)\,du \tag{2}$$

where C is a scaling constant:

$$C = \int_0^\infty g(u)e^{-\int_0^u h(v)dv}\,du \tag{3}$$

The following equations always stand:

$$\int_0^\infty f(t|s)\pi(s)\,ds = \pi(t)$$

$$\int_0^\infty \pi(t)\,dt = 1$$

Proof.

$$\begin{aligned}
&\int_0^\infty f(t|s)\pi(s)\,ds \\
= \;& \frac{h(t)}{C}\int_0^\infty \frac{ds}{\int_0^s g(w)\,dw}\cdot\left[h(s)e^{-\int_0^s h(v)dv}\int_0^s g(w)\,dw\right]\int_0^{\min\{s,t\}} g(u)\,e^{-\int_u^t h(v)dv}\,du \\
= \;& \frac{h(t)}{C}\int_0^\infty h(s)e^{-\int_0^s h(v)dv}ds\int_0^{\min\{s,t\}} g(u)\,e^{-\int_u^t h(v)dv}\,du \\
= \;& \frac{h(t)}{C}\int_0^t g(u)\,e^{-\int_u^t h(v)dv}\,du\int_u^\infty e^{-\int_0^s h(v)dv}h(s)\,ds \\
= \;& \frac{h(t)}{C}\int_0^t g(u)\,e^{-\int_u^t h(v)dv}\,du\int_u^\infty d\left[-e^{-\int_0^s h(v)\,dv}\right] \\
= \;& \frac{h(t)}{C}\int_0^t g(u)\,e^{-\int_u^t h(v)dv}\,e^{-\int_0^u h(v)dv}\,du \\
= \;& \frac{h(t)}{C}e^{-\int_0^t h(v)dv}\int_0^t g(u)\,du
\end{aligned}$$

i.e.:

$$\int_0^\infty f(t|s)\pi(s)\,ds = \pi(t)$$

Then $\pi(t)$ is the density of the stationary distribution. Furthermore, as we require that

$$\begin{aligned}
1 &= \int_0^\infty \pi(t)\,dt \\
&= \int_0^\infty \frac{h(t)}{C}e^{-\int_0^t h(v)dv}\,dt\int_0^t g(u)\,du \\
&= \int_0^\infty g(u)\,du\int_u^\infty \frac{h(t)}{C}e^{-\int_0^t h(v)dv}\,dt \\
&= \frac{1}{C}\int_0^\infty g(u)\,du\int_u^\infty d\left[-e^{-\int_0^t h(v)\,dv}\right] \\
&= \frac{1}{C}\int_0^\infty g(u)e^{-\int_0^u h(v)dv}\,du
\end{aligned}$$

the constant C can thus be calculated. □

1.2 List of Symbols

Symbol	Type	Meaning
a, b	discrete	Coordinate on the sequence
t, s, Δ	continuous	Coalescent time
T_a	continuous, r.v.	Coalescent time at a
R_a	binary, r.v.	Recombination or not between a and $a+1$
X_a	binary, r.v.	Mutation or not at a
N, N_0	continuous	Population size
λ, λ_0	continuous	Relative population size
θ, θ_0	continuous	Per-site mutation rate
ρ, ρ_0	continuous	Per-site recombination rate
p, q	function	transition probability
i, j, k, l	discrete	State of the HMM
u, v, w	continuous	Coalescent time (in integration)
C, C_π, C_σ	continous	Scaling constant

1.3 PSMC Model

In this section, Theorem 1 estabilishes the foundation of the PSMC continuous-time Markov chain. It gives the equations of transition, and the stationary distribution. The following corollaries show how to approximate the constants in the Theorem when the scaled mutation and recombination rates are small.

The discrete-time Markov chain, which will be presented in the next section, is derived from the continuous-time Markov chain by integrating probability desities in time intervals.

Theorem 1 (PSMC)**.** *Let the population size be:*

$$N(t) = N_0\lambda(t)$$

where t equals the number of generations divided by $2N_0$. The scaled mutation rate and recombination rate per nucleotide are θ and ρ. respectively. Given two haplotypes, let T_a be the coalescent time at position $a \in [1, L]$, and define:

$$R_a = \begin{cases} 1 & \text{a recombination happens between } a \text{ and } a+1 \\ 0 & \text{otherwise} \end{cases}$$

$$\Lambda(t) = \int_0^t \lambda(u)\, du$$

$$C_\pi = \int_0^\infty e^{-\int_0^u \frac{dv}{\lambda(v)}}\, du \tag{4}$$

$$C_\sigma = \int_0^\infty \frac{\pi(t)}{1-e^{-\rho t}} dt \tag{5}$$

According to the SMC (Sequentially Markov Coalescent) model (McVean and Cardin, 2005; Marjoram and Wall, 2006), the following equations stand:

$$q(t|s)\, dt = \Pr\{T_{a+1} = t | T_a = s, R_a = 1\} = \frac{dt}{\lambda(t)} \int_0^{\min\{s,t\}} \frac{1}{s} \cdot e^{-\int_u^t \frac{dv}{\lambda(v)}}\, du \tag{6}$$

$$\pi(t) = \Pr\{T_{a+1} = t | R_a = 1\} = \frac{t}{C_\pi \lambda(t)} e^{-\int_0^t \frac{dv}{\lambda(v)}} \tag{7}$$

$$p(t|s) = \Pr\{T_{a+1} = t | T_a = s\} = (1-e^{-\rho s}) q(t|s) + e^{-\rho s}\delta(t-s) \tag{8}$$

$$\sigma(t) = \Pr\{T_a = t\} = \frac{\pi(t)}{C_\sigma(1-e^{-\rho t})} \tag{9}$$

$$\Pr\{R_a = 1\} = \frac{1}{C_\sigma} \tag{10}$$

Furthermore,

$$\int_0^\infty q(t|s)\pi(s)\,ds = \pi(t) \tag{11}$$

$$\int_0^\infty p(t|s)\sigma(s)\,ds = \sigma(t) \tag{12}$$

and

$$\int_0^\infty q(t|s)\,dt = \int_0^\infty p(t|s)\,dt = \int_0^\infty \pi(t)\,dt = \int_0^\infty \sigma(t)\,dt = 1 \tag{13}$$

Proof. Equation 6 is the root of all the other equations.

1. When a recombination happens, the probability that it happens in $[u, u+du)$ is:

   $$P_1(u|s)\,du = \frac{1}{s}\,du$$

   At time u, two alleles coalesce at $[t, t+dt)$ is (Hein et al., 2005; Griffiths and Tavare, 1994):

   $$P_2(t|u)\,dt = \frac{1}{\lambda(t)} \exp\left\{ -\int_u^t \frac{dv}{\lambda(v)} \right\} dt$$

   When we know s and t, $u \in [0, \min\{s,t\})$. Then:

   $$q(t|s) = \int_0^{\min\{s,t\}} P_2(t|u) \cdot P_1(u|s)\,du = \frac{1}{\lambda(t)} \int_0^{\min\{s,t\}} \frac{1}{s} \cdot e^{-\int_u^t \frac{dv}{\lambda(v)}}\,du$$

   This proves Equation 6.

2. In Lemma 1 and Lemma 2, let $g(u) = 1$ and $h(u) = 1/\lambda(u)$. We have:

   $$\int_0^\infty q(t|s)\,dt = 1$$

   $$\int_0^\infty q(t|s)\pi(s)\,ds = \pi(t)$$

   This proves Equation 7 and 11.

3. Equation 8 comes *naturally*, and

   $$\int_0^\infty p(t|s) = (1 - e^{-\rho s}) \int_0^\infty q(t|s)\,dt + e^{-\rho s} = 1$$

   $$\begin{aligned} \int_0^\infty p(t|s)\sigma(s)\,ds &= \frac{1}{C_a} \int_0^\infty (1 - e^{-\rho s}) \frac{q(t|s)\pi(s)}{1 - e^{-\rho s}}\,ds + \frac{e^{-\rho t}}{C_\sigma(1 - e^{-\rho t})}\pi(t) \\ &= \frac{\pi(t)}{C_\sigma(1 - e^{-\rho t})} \\ &= \sigma(t) \end{aligned}$$

   This proves Equation 9 and 12.

4. Given coalescent time $T_a = t$, the probability that a recombination happens between a and $a+1$ is:

   $$\Pr\{R_a = 1 | T_a = t\} = 1 - e^{-\rho t}$$

   Then

   $$\Pr\{R_a = 1\} = \int_0^\infty (1 - e^{-\rho t})\sigma(t)\,dt = \frac{1}{C_\sigma}$$

   This proves Equation 10.

□

Corollary 1 (Approximating C_σ)**.** *When ρ_0 is sufficiently small:*

$$C_\sigma = \frac{1}{C_\pi \rho} + \frac{1}{2} + o(\rho) \tag{14}$$

Proof.

$$\begin{aligned} C_\sigma &= \int_0^\infty \frac{t}{C_\pi \lambda(t)[1-e^{-\rho t}]} e^{-\int_0^t \frac{dv}{\lambda(v)}}\, dt \\ &= \frac{1}{C_\pi \rho} \int_0^\infty \Big[1 + \frac{\rho t}{2} + o(\rho^2)\Big] \frac{1}{\lambda(t)} e^{-\int_0^t \frac{dv}{\lambda(v)}}\, dt \\ &= \frac{1}{C_\pi \rho} \int_0^\infty \frac{1}{\lambda(t)} e^{-\int_0^t \frac{dv}{\lambda(v)}}\, dt + \frac{1}{2} \int_0^\infty \pi(t)\, dt + o(\rho) \\ &= \frac{1}{C_\pi \rho} + \frac{1}{2} + o(\rho) \end{aligned}$$

□

Corollary 2 (Rate of pairwise difference)**.** *When both θ_0 and ρ_0 are sufficiently small:*

$$\Pr\{X_a = 1\} = C_\pi \theta \cdot \big[1 + o(\rho + \theta)\big] \tag{15}$$

Proof.

$$\begin{aligned} \Pr\{X_a = 1\} &= \int_0^\infty \Pr\{X_a = 1 | T_a = t\} \Pr\{T_a = t\}\, dt \\ &= \int_0^\infty (1 - e^{-\theta t}) \sigma(t)\, dt \\ &= \frac{1}{C_\sigma} \int \frac{1 - e^{-\theta t}}{1 - e^{-\rho t}} \pi(t)\, dt \\ &= \frac{1}{C_\sigma} \int \frac{\theta + o(\theta^2)}{\rho + o(\rho^2)} \pi(t)\, dt \\ &= \frac{\theta}{C_\sigma \rho} \int \Big[1 + o(\rho + \theta)\Big] \pi(t)\, dt \\ &= C_\pi \theta \cdot \big[1 + o(\rho + \theta)\big] \end{aligned}$$

□

Corollary 3 (First-order approximation)**.** *Under the first-order approximation with respect to θ and ρ, the following equations stand:*

$$\Pr\{R_a = 1\} = \frac{1}{C_\pi} = C_\pi \rho$$

$$\Pr\{X_a = 1\} = C_\pi \theta$$

$$\sigma(t) = \frac{1}{\lambda(t)} e^{-\int_0^t \frac{dv}{\lambda(v)}}$$

$$\int_0^t \sigma(u)\, du = 1 - e^{-\int_0^t \frac{dv}{\lambda(v)}}$$

Remark 1 (Distribution of segment lengths). Let L_{a+1} be the length of the segment following a recombination occurring at a. Conditional on the recombination, L_{a+1} follows a exponential distribution (more precisely, a geometric distribution in fact):

$$\Pr\{L_{a+1} = l | R_a = 1, T_{a+1} = t\}\, dl = \rho t e^{-\rho t l}\, dl$$

Then,

$$\Pr\{L=l\}\,dl = dl\int_0^\infty \frac{\rho t^2 e^{-\rho t l}}{C_\pi \lambda(t)} e^{-\int_0^t \frac{dv}{\lambda(v)}}\,dt = dl\,\frac{\rho}{C_\pi}\int_0^\infty e^{-\int_0^t \frac{dv}{\lambda(v)}}\,d\Big(t^2 e^{-\rho t l}\Big)$$

The mean segment length is thus

$$\begin{aligned}
&\int_0^\infty \rho t l e^{-\rho t l}\,dl \int_0^\infty \frac{t}{C_\pi\lambda(t)} e^{-\int_0^t \frac{dv}{\lambda(v)}}\,dt \\
= \;& \frac{1}{C_\pi \rho}\int_0^\infty \frac{1}{\lambda(t)} e^{-\int_0^t \frac{dv}{\lambda(v)}}\,dt \\
= \;& \frac{1}{C_\pi\rho} \approx C_\sigma
\end{aligned}$$

1.4 Discrete-Time PSMC Model

This section presents the discrete-time PSMC Markov Chain, its transition probabilities between time intervals and the stationary distribution. The proof of Corolary 4 is given in the Appendix.

Corollary 4 (Discrete-time PSMC)**.** *Let*

$$0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = \infty$$

Assume in each time interval $[t_k, t_{k+1})$ *function* $\lambda(t)$ *is a constant* λ_k*. Define:*

$$\pi_k = \int_{t_k}^{t_{k+1}} \pi(t)\,dt$$

$$\sigma_k = \int_{t_k}^{t_{k+1}} \sigma(t)\,dt$$

$$q_{kl} = \frac{1}{\pi_k}\int_{t_k}^{t_{k+1}} ds \int_{t_l}^{t_{l+1}} q(t|s)\pi(s)\,dt$$

$$p_{kl} = \frac{1}{\sigma_k}\int_{t_k}^{t_{k+1}} ds \int_{t_l}^{t_{l+1}} p(t|s)\sigma(s)\,dt$$

Then:

$$\pi_k = \frac{1}{C_\pi}\left[(\alpha_k - \alpha_{k+1})\Big(\sum_{i=0}^{k-1}\tau_i + \lambda_k\Big) - \alpha_{k+1}\tau_k\right]$$

$$\sigma_k = \frac{1}{C_\sigma}\left[\frac{1}{C_\pi\rho}(\alpha_k - \alpha_{k+1}) + \frac{\pi_k}{2} + o(\rho)\right] \tag{16}$$

Furthermore, for $l < k$*:*

$$q_{kl} = \frac{\alpha_k - \alpha_{k+1}}{C_\pi \pi_k}\left[(\alpha_l - \alpha_{l+1})\Big(\beta_l - \frac{\lambda_l}{\alpha_l}\Big) + (t_{l+1} - t_l)\right]$$

for $l = k$*:*

$$q_{kl} = \frac{1}{C_\pi \pi_k}\left[(\alpha_k - \alpha_{k+1})^2\Big(\beta_k - \frac{\lambda_k}{\alpha_k}\Big) + 2\lambda_k(\alpha_k - \alpha_{k+1}) - 2\alpha_{k+1}(t_{k+1} - t_k)\right]$$

and for $l > k$*:*

$$q_{kl} = \frac{\alpha_l - \alpha_{l+1}}{C_\pi \pi_k}\left[(\alpha_k - \alpha_{k+1})\Big(\beta_k - \frac{\lambda_k}{\alpha_k}\Big) + (t_{k+1} - t_k)\right]$$

and

$$p_{kl} = \frac{\pi_k}{C_\sigma \sigma_k} q_{kl} + \delta_{kl}\Big(1 - \frac{\pi_k}{C_\sigma \sigma_k}\Big) \tag{17}$$

where:

$$\tau_k = t_{k+1} - t_k$$

$$\alpha_k = \exp\left(- \sum_{i=0}^{k-1} \frac{t_{i+1} - t_i}{\lambda_i} \right)$$

$$\beta_k = \sum_{i=0}^{k-1} \lambda_i \Big(\frac{1}{\alpha_{i+1}} - \frac{1}{\alpha_i} \Big)$$

$$C_\pi = \sum_{k=0}^{n} \lambda_k (\alpha_k - \alpha_{k+1})$$

Remark 2 (Mutation probability). The average mutation probability in an interval $[t_k, t_{k+1})$ cannot be analytically calculated. But we can seek another way. From Equation 17, a recombination occurs in $[t_k, t_{k+1})$ as if it occurs at time $-\log\Big[1 - \pi_k/(C_\sigma \sigma_k)\Big]/\rho$. If we assume mutation also exactly occurs at this time point, the probability of a mutation is:

$$e_k(1) = \exp\left[-\frac{\theta}{\rho} \log\Big(1 - \frac{\pi_k}{C_\sigma \sigma_k}\Big) \right] = \Big(1 - \frac{\pi_k}{C_\sigma \sigma_k}\Big)^{\theta/\rho} \tag{18}$$

Remark 3 (Determining N_0). If we know μ, the neutral mutation rate, $N_0 = \theta/4\mu$. On autosomes, μ is typically 2.5×10^{-8} (Nachman and Crowell, 2000). Note that Equation 15 agrees with Marth et al. (2004) in case of two haplotypes.

2 PSMC Hidden Markov Model

2.1 The basic of HMM

Remark 4 (PSMC-HMM). We denote a hidden state in the HMM by k, which means a coalescence between the two haplotypes at this point in the sequence lies in the time interval $[t_k, t_{k+1})$. A mutation is emmitted with a probability $e_k(1)$ (Equation 18) and the transition probability is p_{kl} (Equation 17). The stationary distribution of the hidden states is $\{\sigma_k\}$ (Equation 16). All these parameters can be analytically approximated with a precision of order-two Taylor expansion when ρ_0 is sufficiently small.

Remark 5 (PSMC-PHMM). It is possible to use PSMC to approximately model two diploid sequences. In this case, a hidden state is (k, l) and the transition probability of $(k, l) \to (k', l')$ is $p_{kk'} p_{ll'}$. The stationary distribution of hidden states is $\{\sigma_k \sigma_{k'}\}$.

Remark 6 (Missing data). Missing data can be easily incorporated into an HMM. When there is no observation at a, $e_k(x_a) = 1$ for all k.

Remark 7 (Choosing time intervals). We choose a set of $\{t_i\}_{i=0\ldots n}$ that are approximately evenly distributed in the log space, but because we require $t_0 = 0$, the intervals will not strictly evenly distributed. In practice, we set

$$t_i = 0.1(e^{\frac{i}{n}\log(1+10T_{max})} - 1)$$

where $T_{max} = t_n$ is chosen such that no more than a few percent of coalescences occur beyond T_{max}.

Remark 8 (Reducing parameters). In principle, we can estimate all the $n + 1$ values of λ_k with EM. However, at both small and large t, the expected number of segments is very small. Separate estimates of λ_k in these intervals will lead to overfitting due to insufficient data. An effective way to tell whether overfitting occurs is to check $C_\sigma \pi_k$, the expected number of segments in the interval

$[t_k, t_{k+1})$. If this number is small (less than 20, for instance), the λ_k estimated from EM cannot be trusted due to statistical fluctuations. In this case, we should use fewer free parameters by using the same λ spanning several adjacent intervals. This will lower the resolution but will yield much better estimation.

2.2 Assessing variance and fitness

Remark 9 (Bootstrapping). The variance can be estimated by boostrapping. We split the input diploid sequence into L'-long non-overlapping segments and randomly resample the segments with replacement to generate a new diploid sequence of the same length as the original one. Parameters are then estimated from the new sequence. We repeat this process B times and regard the variance of the B resampled estimates as the variance of the estimate on the original sequence. Typically, we take $L' = 30,000,000$ and $B = 100$.

Remark 10 (Measuring GOF with σ_k). On one hand, from Equation 16 we can calculate σ_k from free parameters of the model without looking at the data. On the other hand, from forward-backward algorithm we can estimate the posterior expectation of the occurences $\hat{c}_k$ of state k:

$$\hat{c}_k = \frac{1}{P(D)} \sum_i f_k(i) b_k(i)$$

Normalizing $\hat{c}_k$ gives:

$$\hat{\sigma}_k = \frac{\sum_i f_k(i) b_k(i)}{\sum_{k,i} f_k(i) b_k(i)}$$

where $f_k(i)$ is the forward function of sequence position i and state k and $b_k(i)$ is the backward function. If the model fits the data, we would expect to see $\{\hat{\sigma}_k\}$ is identical to $\{\sigma_k\}$. And therefore the relative entroy $D(\sigma||\hat{\sigma})$ would be an indicator of GOF:

$$G^\sigma = \sum_k \sigma_k \log \frac{\sigma_k}{\hat{\sigma}_k}$$

Remark 11 (Measuring GOF with l-long subsequences). MacKay Altman (2004) pointed out we can test GOF by comparing the distribution of l-long subsequences from direct calculation with the observed distribution. Given an integer $n \in [0, 2^l - 1]$, let $\{p_n\}$ be the theoretical distribution of the binary sequence represented by n and let $\{\hat{p}_n\}$ be the one directly counted from the observed sequence. The relative entroy between them is:

$$G_l = \sum_{n=0}^{2^l-1} p_n \log \frac{p_n}{\hat{p}_n}$$

which measures GOF. Typically, l is ranged from 10 to 20.

A Proof of Corollary 4

Proof.

$$C_\pi = \int_0^\infty e^{-\int_0^u \frac{dv}{\lambda(v)}} du = \sum_{k=0}^n \alpha_k \int_{t_k}^{t_{k+1}} e^{-\frac{u-t_k}{\lambda_k}} du = \sum_{k=0}^n \lambda_k(\alpha_k - \alpha_{k+1})$$

$$
\begin{aligned}
\pi_k &= \int_{t_k}^{t_{k+1}} \pi(t)\,dt \\
&= \frac{1}{C_\pi}\int_{t_k}^{t_{k+1}} \frac{dt}{\lambda(t)} e^{-\int_0^t \frac{dv}{\lambda(v)}} \cdot t \\
&= \frac{1}{C_\pi \lambda_k}\int_{t_k}^{t_{k+1}} \alpha_k e^{-\frac{t-t_k}{\lambda_k}} \left[\sum_{i=0}^{k-1}\tau_i + (t-t_k)\right] dt \\
&= \frac{\alpha_k}{C_\pi \lambda_k}\left[\sum_{i=0}^{k-1}\tau_i \int_{t_k}^{t_{k+1}} e^{-\frac{t-t_k}{\lambda_k}}\,dt + \int_{t_k}^{t_{k+1}} (t-t_k) e^{-\frac{t-t_k}{\lambda_k}}\,dt\right] \\
&= \frac{\alpha_k}{C_\pi \lambda_k}\left[\lambda_k \sum_{i=0}^{k-1}\tau_i\Big(1-e^{-\frac{\tau_k}{\lambda_k}}\Big) + \lambda_k^2 \int_0^{\frac{\tau_k}{\lambda_k}} u e^{-u} du\right] \\
&= \frac{\alpha_k}{C_\pi}\left[\sum_{i=0}^{k-1}\tau_i\Big(1-e^{-\frac{\tau_k}{\lambda_k}}\Big) + \Big(\lambda_k - \lambda_k e^{-\frac{\tau_k}{\lambda_k}} - \tau_k e^{-\frac{\tau_k}{\lambda_k}}\Big)\right] \\
&= \frac{1}{C_\pi}\left[(\alpha_k - \alpha_{k+1})\Big(\sum_{i=0}^{k-1}\tau_i + \lambda_k\Big) - \alpha_{k+1}\tau_k\right]
\end{aligned}
$$

$$
\begin{aligned}
\sigma_k &= \int_{t_k}^{t_{k+1}} \sigma(t)\,dt \\
&= \int_{t_k}^{t_{k+1}} \frac{1}{C_\sigma(1-e^{-\rho t})} \cdot \frac{t}{C_\pi \lambda(t)} e^{\int_0^t \frac{dv}{\lambda(v)}}\,dt \\
&= \frac{1}{C_\sigma C_\pi \rho}\int_{t_k}^{t_{k+1}} \Big[1 + \frac{1}{2}\rho t + o(\rho^2)\Big]\frac{1}{\lambda(t)} e^{\int_0^t \frac{dv}{\lambda(v)}}\,dt \\
&= \frac{1}{C_\sigma C_\pi \rho}\left[\int_{t_k}^{t_{k+1}} \frac{1}{\lambda(t)} e^{\int_0^t \frac{dv}{\lambda(v)}}\,dt + \frac{1}{2}C_\pi \rho \int_{t_k}^{t_{k+1}} \pi(t)\,dt + o(\rho^2)\right] \\
&= \frac{1}{C_\sigma}\left[\frac{1}{C_\pi \rho}(\alpha_k - \alpha_{k+1}) + \frac{\pi_k}{2} + o(\rho)\right]
\end{aligned}
$$

$$
\begin{aligned}
q_{kl} &= \frac{1}{\pi_k}\int_{t_k}^{t_{k+1}} ds \int_{t_l}^{t_{l+1}} q(t|s)\pi(s)\,dt \\
&= \frac{1}{C_\pi \pi_k}\int_{t_k}^{t_{k+1}} \frac{1}{\lambda(s)} e^{-\int_0^s \frac{dv}{\lambda(v)}}\,ds \int_{t_l}^{t_{l+1}} \frac{dt}{\lambda(t)} \int_0^{\min\{s,t\}} e^{-\int_u^t \frac{dw}{\lambda(w)}}\,du \\
&= \frac{\alpha_k}{C_\pi \pi_k \lambda_k \lambda_l}\int_{t_k}^{t_{k+1}} e^{-\frac{s-t_k}{\lambda_k}}\,ds \int_{t_l}^{t_{l+1}} dt \int_0^{\min\{s,t\}} e^{-\int_u^t \frac{dw}{\lambda(w)}}\,du
\end{aligned}
$$

$l < k$:

$$
\begin{aligned}
q_{kl} &= \frac{\alpha_k - \alpha_{k+1}}{C_\pi \pi_k \lambda_l}\int_{t_l}^{t_{l+1}} dt \left(e^{-\frac{t-t_l}{\lambda_l}} \sum_{i=0}^{l-1}\frac{\alpha_l}{\alpha_i}\int_{t_i}^{t_{i+1}} e^{\frac{u-t_i}{\lambda_i}}\,du + \int_{t_l}^{t} e^{-\frac{t-u}{\lambda_l}}\,du\right) \\
&= \frac{\alpha_k - \alpha_{k+1}}{C_\pi \pi_k \lambda_l}\int_{t_l}^{t_{l+1}} dt \left[\alpha_l e^{-\frac{t-t_l}{\lambda_l}} \sum_{i=0}^{l-1}\lambda_i\Big(\frac{1}{\alpha_{i+1}} - \frac{1}{\alpha_i}\Big) + \lambda_l\Big(1 - e^{-\frac{t-t_l}{\lambda_l}}\Big)\right] \\
&= \frac{\alpha_k - \alpha_{k+1}}{C_\pi \pi_k}\left[\Big(1 - \frac{\alpha_{l+1}}{\alpha_l}\Big)\beta_l \alpha_l + (t_{l+1} - t_l) - \lambda_l\Big(1 - \frac{\alpha_{l+1}}{\alpha_l}\Big)\right] \\
&= \frac{\alpha_k - \alpha_{k+1}}{C_\pi \pi_k}\left[(\alpha_l - \alpha_{l+1})\Big(\beta_l - \frac{\lambda_l}{\alpha_l}\Big) + (t_{l+1} - t_l)\right]
\end{aligned}
$$

$l > k$:

$$\begin{aligned} q_{kl} &= \frac{\alpha_k}{C_\pi \pi_k \lambda_k} \int_{t_k}^{t_{k+1}} e^{-\frac{s-t_k}{\lambda_k}}\, ds \cdot (\alpha_l - \alpha_{l+1}) \left[\beta_k + \frac{\lambda_k^2}{\alpha_k} \Big(e^{\frac{s-t_k}{\lambda_k}} - 1 \Big) \right] \\ &= \frac{\alpha_k(\alpha_l - \alpha_{l+1})}{C_\pi \pi_k \lambda_k} \int_{t_k}^{t_{k+1}} e^{-\frac{s-t_k}{\lambda_k}}\, ds \cdot \left[\Big(\beta_k - \frac{\lambda_k}{\alpha_k} \Big) + \frac{\lambda_k}{\alpha_k} e^{\frac{s-t_k}{\lambda_k}} \right] \\ &= \frac{\alpha_l - \alpha_{l+1}}{C_\pi \pi_k} \left[(\alpha_k - \alpha_{k+1}) \Big(\beta_k - \frac{\lambda_k}{\alpha_k} \Big) + (t_{k+1} - t_k) \right] \end{aligned}$$

$l = k$:

$$\begin{aligned} q_{kl} &= \frac{\alpha_k}{C_\pi \pi_k \lambda_k} \int_{t_k}^{t_{k+1}} e^{-\frac{s-t_k}{\lambda_k}}\, ds \cdot \left[\beta_k(\alpha_k - \alpha_{k+1}) + (s - t_k) - \frac{\lambda_k \alpha_{k+1}}{\alpha_k} \Big(e^{\frac{s-t_k}{\lambda_k}} - 1 \Big) \right] \\ &= \frac{\alpha_k}{C_\pi \pi_k} \int_0^{\tau_k} e^{-u}\, du \cdot \left[\beta_k(\alpha_k - \alpha_{k+1}) + \frac{\lambda_k \alpha_{k+1}}{\alpha_k} + \lambda_k u - \frac{\lambda_k \alpha_{k+1}}{\alpha_k} e^u \right] \\ &= \frac{1}{C_\pi \pi_k} \left[\beta_k(\alpha_k - \alpha_{k+1})^2 + \frac{\lambda_k}{\alpha_k}(\alpha_k - \alpha_{k+1})(\alpha_k + \alpha_{k+1}) - 2\tau_k \alpha_{k+1} \right] \\ &= \frac{1}{C_\pi \pi_k} \left[(\alpha_k - \alpha_{k+1})^2 \Big(\beta_k - \frac{\lambda_k}{\alpha_k} \Big) + 2\lambda_k(\alpha_k - \alpha_{k+1}) - 2\alpha_{k+1}(t_{k+1} - t_k) \right] \end{aligned}$$

$$\begin{aligned} p_{kl} &= \frac{1}{\sigma_k} \int_{t_k}^{t_{k+1}} ds \int_{t_l}^{t_{l+1}} p(t|s)\sigma(s)\, dt \\ &= \frac{1}{\sigma_k} \int_{t_k}^{t_{k+1}} \frac{\pi(s)\, ds}{C_\sigma(1 - e^{-\rho s})} \left[\int_{t_l}^{t_{l+1}} (1 - e^{-\rho s}) q(t|s)\, dt + \delta_{kl} e^{-\rho s} \right] \\ &= \frac{1}{C_\sigma \sigma_k} \int_{t_k}^{t_{k+1}} \pi(s)\, ds \int_{t_l}^{t_{l+1}} q(t|s)\, dt + \frac{\delta_{kl}}{C_\sigma \sigma_k} \int_{t_k}^{t_{k+1}} \Big(\frac{1}{1 - e^{-\rho s}} - 1 \Big) \pi(s)\, ds \\ &= \frac{\pi_k q_{kl}}{C_\sigma \sigma_k} + \frac{\delta_{kl}}{C_\sigma \sigma_k} \left[\int_{t_k}^{t_{k+1}} \frac{\pi(s)\, ds}{1 - e^{-\rho s}} - \pi_k \right] \\ &= \frac{\pi_k q_{kl}}{C_\sigma \sigma_k} + \delta_{kl} \Big(1 - \frac{\pi_k}{C_\sigma \sigma_k} \Big) \end{aligned}$$

□

References

Griffiths, R. C. and Tavare, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci*, 344(1310):403–410.

Hein, J., Schierup, M. H., and Wiuf, C. (2005). *Gene genealogies, variation and evolution*. Oxford University Press.

MacKay Altman, R. (2004). Assessing the goodness-of-fit of hidden markov models. *Biometrics*, 60(2):444–450.

Marjoram, P. and Wall, J. D. (2006). Fast "coalescent" simulation. *BMC Genet*, 7:16.

Marth, G. T., Czabarka, E., Murvai, J., and Sherry, S. T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166(1):351–372.

McVean, G. A. T. and Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*, 360(1459):1387–1393.

Nachman, M. W. and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304.