**Estimating time to the most recent common ancestor (TMRCA): comparison and application of eight methods --- Supplementary**

Contents

[I Simulation Analysis 2](#_Toc419969114)

[1.1 Command line input to simulate data 2](#_Toc419969115)

[1.2 Command line to apply eight methods in simulation 2](#_Toc419969116)

[II TMRCA of Southeast Asian Malays and South Asian Indians 4](#_Toc419969117)

[2.1 Data 4](#_Toc419969118)

[2.2 Methods of Estimating Divergence Time of Malay and Indian Population 5](#_Toc419969119)

[III Supplementary Table 7](#_Toc419969120)

[IV Methodology Review 9](#_Toc419969121)

[4.1 Statistical Estimators of TMRCA 9](#_Toc419969122)

[4.2 MCMC methods 10](#_Toc419969123)

[4.3 HMM Methods 11](#_Toc419969124)

[4.4 Differential Approximation Methods 14](#_Toc419969125)

[REFERENCE 14](#_Toc419969126)

# Simulation Analysis

## Command line input to simulate data

We use ms program and seq-gen program to generate sequence data with population parameters mu=1e-9, g=25, L=1e7, Ne\_ref=1e4, r=5e-9 (rho=4Ne\_ref\*r\*L). The simulation data we used can be downloaded from <http://www.statgen.nus.edu.sg/~TMRCA/>. The command line we used are:

#### Simple Isolation Model

 for i in {1..10}; do ms 1001 1 -t 10000 -r 2000 10000000 -I 3 500 500 1 -n 1 1 -n 2 1 -n 3 1 -ej 0.02 2 1 -en 0.02 1 1 -ej 4.1 3 1 -en 4.1 1 1.0 -p 7 > ms\_Simple\_Isolation\_$i.txt ; done &

#### Isolation Migration Model

 for i in {1..10}; do ms 1001 1 -t 10000 -r 2000 10000000 -I 3 500 500 1 -n 1 1 -n 2 1 -n 3 1 -em 0.0 1 2 4 -em 0.0 2 1 4 -ej 0.02 2 1 -eM 0.02 0.0 -en 0.02 1 1.0 -ej 4.1 3 1 -en 4.1 1 1.0 -p 7 > ms\_Isolation\_Migration\_$i.txt ; done &

#### Bottleneck - Non-bottleneck Model

 for i in {1..10}; do ms 1001 1 -t 10000 -r 2000 10000000 -I 3 500 500 1 -n 1 1 -n 2 0.5 -n 3 0.5 -eg 0.0 1 100.0 -eg 0.023 1 -45.0 -ej 0.06 2 1 -eN 0.06 0.5 -ej 4.1 3 1 -en 4.1 1 0.5 -p 7 > ms\_Bottleneck\_NonBottleneck\_$i.txt ; done &

#### Bottleneck Bottleneck Model

 for i in {1..10}; do ms 1001 1 -t 10000 -r 2000 10000000 -I 3 500 500 1 -n 1 1 -n 2 1 -n 3 0.5 -eg 0.0 1 58.0 -eg 0.0 2 58.0 -ej 0.04 2 1 -eG 0.04 0.0 -ej 0.04 2 1 -eN 0.04 0.5 -ej 4.1 3 1 -en 4.1 1 0.5 -p 7 > ms\_Bottleneck\_Bottleneck\_$i.txt ; done &

## Command line to apply eight methods in simulation

### Hayes’ method (T-LD)& McEvoy’s method (T-FST)

We wrote some scripts to calculate TMRCA according to Hayes’ method and McEvoy’s method, T-LD-FST.bash, which can be downloaded from <http://www.statgen.nus.edu.sg/~TMRCA/>.

### DADI

We wrote python scripts for analyses of DADI:

DADI-Simulation.py and DADI\_Simulation\_demographic\_models.py.

The scripts can be downloaded from <http://www.statgen.nus.edu.sg/~TMRCA/>.

The command line we used for analyses are:

#### Simple Isolation Model (default setting)

python DADI-Simulation.py input Isolation-Migration True False > output

#### Isolation Migration Model (prior information)

python DADI-Simulation.py input Isolation-Migration True True > migration.output

#### Bottleneck - Non-bottleneck Model (prior information)

python DADI-Simulation.py input Bottleneck-NonBottleneck True False > Bottleneck-NonBottleneck.output

#### Bottleneck Bottleneck Model (prior information)

python DADI-Simulation.py input Twopop-Bottleneck-Model True False > Bottleneck-Bottleneck.output

### MIMAR

The command line we used for analyses are:

#### Simple Isolation Model

mimar 300000 100000 30 -lf input -u 2.5e-8 -t u 0.0001 0.002 -n u 0.0001 0.002 -N u 0.0001 0.002 -ej u 500 3000 -v 3e-4 3e-4 300 3e-4 0.25 0.25 -i 10 -r 2e-4 -o exsoutput > output

#### Isolation Migration Model

mimar 300000 100000 30 -lf input -u 2.5e-8 -t u 0.0001 0.002 -n u 0.0001 0.002 -N u 0.0001 0.002 -ej u 500 3000 -v 3e-4 3e-4 300 3e-4 0.25 0.25 -i 10 -r 2e-4 –o exsoutput > output

mimar 300000 100000 30 -lf input -u 2.5e-8 -t u 0.0001 0.002 -n u 0.0001 0.002 -N u 0.0001 0.002 -ej u 500 3000 -M l -5 3 -v 3e-4 3e-4 300 3e-4 0.25 0.25 -i 10 -r 2e-4 -o migration.exsoutput > migration.output

#### Bottleneck - Non-bottleneck Model

mimar 300000 100000 30 -lf input -u 2.5e-8 -t u 0.0001 0.002 -n u 0.0001 0.002 -N u 0.0001 0.002 -ej u 1000 5000 -v 3e-4 3e-4 300 3e-4 0.25 0.25 -i 10 -r 2e-4 -o exsoutput > output

mimar 300000 100000 30 -lf input -u 2.5e-8 -t u 0.0001 0.002 -n u 0.0001 0.002 -N u 0.0001 0.002 -eg 0.0 1 100.0 -ej u 1000 5000 -eg 0.0 1 100.0 -eg 0.38 1 -45.0 -eg 1.0 1 0.0 -v 3e-4 3e-4 300 3e-4 0.25 0.25 -i 10 -r 2e-4 -o BN.exsoutput > BN.output

#### Bottleneck\_Bottleneck

mimar 300000 100000 30 -lf input -u 2.5e-8 -t u 0.0001 0.002 -n u 0.0001 0.002 -N u 0.0001 0.002 -ej u 500 3000 -v 3e-4 3e-4 300 3e-4 0.25 0.25 -i 10 -r 2e-4 -o exsoutput > output

mimar 300000 100000 30 -lf input -u 2.5e-8 -t u 0.0001 0.002 -n u 0.0001 0.002 -N u 0.0001 0.002 -ej u 500 3000 -en 0.0 1 1.0 -en 0.0 2 1.0 -eg 0.0 1 58.0 -eg 0.0 2 58.0 -eg 1.0 1 0.0 -eg 1.0 2 0.0 -en 1.0 1 0.5 -en 1.0 2 0.5 -v 3e-4 3e-4 300 3e-4 0.25 0.25 -i 10 -r 2e-4 -o BB.exsoutput > BB.output

### GPho-CS

The parameters are set as follows:

|  |  |  |
| --- | --- | --- |
| Parameter | Simulation Study | Real Data Analysis |
| Burn-in iterations | 100 000 | 100 000 |
| Sample iterations | 200 000 | 300 000 |
| Sample-skip | 10 | 10 |
| prior for all parameters  |   |   |
| prior for all *m* parameters |   |   |
| Prior for divergence time  |   |   |
| Prior for divergence time between human and outgroup |  |   |
| Search finetune automatically | True | True |

### CoalHMM

All the analyses are applied with the command below and the mean of 1000 MCMC samples are used as the parameter estimation.

python isolation-model-mcmc.py -o output --logfile mcmc.log --mc3 -n 1000 --theta 0.001 --rho 0.4 input

### PSMC

All the analyses are applied with the command below and the parameters obtained in the 20th iteration is used as the estimation.

psmc -N25 -t15 -r5 -p "4+25\*2+4+6" -o output.psmc input.psmcfa

### MSMC

All the analyses are applied with the command below and the parameters obtained in the 20th iteration is used as the estimation.

msmc\_linux\_64bit --fixedRecombination --skipAmbiguous -P 0,0,1,1 –o output input

# TMRCA of Southeast Asian Malays and South Asian Indians

## Data

### Malay Sample From Singapore Genome Variation Project (SGVP):

SSM.chr1.2012\_05.bgl.phased.vcf.gz ~ SSM.chr22.2012\_05.bgl.phased.vcf.gz and SS6003427.bam

### Indian Sample Singapore Genome Variation Project (SGVP):

chr1.phasing.vcf.gz ~ chr22.phasing.vcf.gz and SS6003427.bam

### Human Genome Reference

Human reference file is downloaded from 1000 Genomes Phase III FTP: human\_g1k\_v37.fasta.gz

### panTro2 reference

chr1.hg19.panTro2.synNet.axt.gz ~ chr22.hg19.panTro2.synNet.axt.gz are download from UCSC: <http://hgdownload.cse.ucsc.edu/goldenpath/hg19/vsPanTro2/syntenicNet/>.

## Methods of Estimating Divergence Time of Malay and Indian Population

### Hayes’ method, McEvoy’s method

We used phased variants data from SGVP for Hayes’ method, McEvoy’s method. The scripts used to calculate T-LD and T-FST are the same as the scripts used in simulation.

### DADI

We used phased variants data from SGVP for DADI. We wrote python scripts for analyses of DADI:

DADI-Malay-Indian.py and demographic\_models.py.

The scripts can be downloaded from <http://www.statgen.nus.edu.sg/~TMRCA/>.

The command line we used for analyses are:

python DADI-Malay-Indian.py chr{$CHR}.input Isolation-Migration True False > DADI-Malay-Indian-chr{$CHR}-Isolation.output

python DADI-Malay-Indian.py chr{$CHR}.input Twopop-Bottleneck-Model True False > DADI-Malay-Indian-chr{$CHR}-Twopop-Bottleneck-Model\_nomigration.output

### MIMAR

We used phased variants data from SGVP for MIMAR. The command line we used to run MIMAR is:

mimar 300000 100000 30 -lf chr{$CHR}.input -u 2.5e-8 -t u 0.0001 0.002 -n u 0.0001 0.002 -N u 0.0001 0.002 -ej u 500 3000 -v 3e-4 3e-4 300 3e-4 0.25 0.25 -i 10 -o MIMAR-Malay-Indian-chr{$CHR}.exsoutput > MIMAR-Malay-Indian-chr{$CHR}.output

### CoalHMM

We used phased variants data from SGVP for CoalHMM and we use human reference panel to reconstruct the haploid sequence. The command line we used to run CoalHMM is:

isolation-model-mcmc.py -o CoalHMM-Malay-Indian-chr{$CHR}.mcmc --logfile CoalHMM-Malay-Indian-chr{$CHR}.mcmc.log --mc3 --split 0.00001 --theta 0.001 chr{$CHR}.input;

### GPho-CS

The neutral sites list as well as filter bed files are downloaded from <http://compgen.bscb.cornell.edu/GPhoCS/data.php>.

We used samtools 1.0 to pileup the raw reads and call genotypes by bcftools. The command is:

samtools mpileup -C 50 -u -r <chr> -f <ref.fa> <bam> | bcftools –cgI

We used shapeit program to phase the variants with command shown below. For those positions that are not present in the 1000GP reference panel, we randomly assign the haplotypes to two haploid sequences.

shapeit -V tmp.vcf -M genetic\_map\_chr${CHR}\_combined\_b37.txt --input-ref 1000GP\_Phase3\_chr$CHR.hap.gz 1000GP\_Phase3\_chr$CHR.legend.gz 1000GP\_Phase3.sample -O shapeit.out --exclude-snp alignments.snp.strand.exclude --no-mcmc

shapeit -convert --input-haps shapeit.out --output-vcf phased.vcf

As suggested by GPhoCS developer, we filtered out simple repeats, recent transposable elements, indels, sites with effective coverage < 5, regions now showing conserved synteny in human/chimpanzee alignments, recent segmental duplications, CpGs and sites likely to be under selection such as exons of protein-coding genes, noncoding RNAs, and conserved noncoding elements.

The parameters used for GPho-CS is given in Section 1.2.4.

### PSMC

We used samtools 1.0 to pileup the raw reads and call genotypes by bcftools.

samtools mpileup -C50 -uf ref.fa aln.bam | bcftools view -c - | vcfutils.pl vcf2fq -d 10 -D 60 | gzip > diploid.fq.gz

We phased the heterozygotes with the same fashion as GPho-CS and obtain one haploid sequence. The pseudo-diploid sequence is constructed from two haploid sequences which each from one population.

### MSMC

We follow the pipeline of MSMC to call variants and phase the data according to 1000 Genome Project Phase III as reference.

# Supplementary Table

Table 1. Mean error rate (MER) and corresponding 95% confidence interval.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Simulation** | **T\_LD** | **T\_FST** | **MIMAR** | **MIMAR-prior** | **DADI** | **DADI-prior** | **CoalHMM** | **Gpho-CS** | **PSMC 1e6** | **PSMC 1e5** | **PSMC 5e4** | **MSMC** |
| **mean\_se\_SI** | -8.9% | -0.5% | 15.7% | NA | -1.2% | NA | -33.8% | -15.5% | 0.6% | 29.4% | 40.1% | -9.5% |
| **CI.dn\_SI** | -13.3% | -5.5% | 9.3% | NA | -4.2% | NA | -65.3% | -55.1% | -16.0% | -7.7% | 4.7% | -36.7% |
| **CI.up\_SI** | -4.6% | 4.4% | 22.1% | NA | 1.8% | NA | -2.4% | 24.0% | 17.2% | 66.5% | 75.5% | 17.6% |
| **mean\_se\_IM** | -24.0% | -9.7% | 0.2% | 44.3% | -17.0% | -0.1% | 13.6% | -52.1% | 37.5% | 57.9% | 64.5% | -16.4% |
| **CI.dn\_IM** | -28.2% | -13.4% | -3.5% | 41.7% | -18.5% | -5.7% | -19.2% | -89.9% | 5.8% | 17.8% | 26.6% | -24.6% |
| **CI.up\_IM** | -19.8% | -5.9% | 3.9% | 47.0% | -15.5% | 5.6% | 46.4% | -14.3% | 69.2% | 98.0% | 102.5% | -8.1% |
| **mean\_se\_BN** | -13.0% | -46.3% | -15.7% | -19.6% | 23.8% | 14.0% | -5.9% | 4.3% | -35.3% | -23.2% | -16.9% | -10.9% |
| **CI.dn\_BN** | -36.5% | -48.9% | -25.1% | -22.7% | -15.9% | 5.5% | -12.7% | -5.8% | -46.2% | -35.7% | -28.4% | -25.7% |
| **CI.up\_BN** | 10.5% | -43.6% | -6.4% | -16.5% | 63.5% | 22.5% | 0.9% | 14.5% | -24.4% | -10.8% | -5.4% | 4.0% |
| **mean\_se\_BB** | -12.0% | -14.0% | -7.9% | 1.1% | 110.0% | 50.0% | 2.5% | 23.3% | -25.6% | -10.3% | -7.9% | -2.4% |
| **CI.dn\_BB** | -24.7% | -18.6% | -11.6% | -3.1% | 98.6% | 44.0% | -12.2% | 2.5% | -37.8% | -25.6% | -20.7% | -11.1% |
| **CI.up\_BB** | 0.7% | -9.3% | -4.3% | 5.2% | 121.4% | 56.1% | 17.3% | 44.0% | -13.4% | 5.0% | 4.8% | 6.4% |

Note: SI, IM, BN, BB represents (i) simple isolation model, (ii) isolation migration model, (iii) bottleneck-nonbottleneck model and (iv) bottleneck-bottleneck model, respectively. CI.dn and CI.up represents the upper and lower boundary of 95% confidence interval. –prior represents the results when proper demographic model is used in scenario (ii)-(iv). We includes the results of PSMC with three different thresholds, 1e6, 1e5 and 5e4.

Table 2. TMRCA of Southeast Asian Malays and South Asian Indians and corresponding 95% confidence interval estimated by eight methods.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TMRCA of Southeast Asian Malays and South Asian Indians** | **T\_LD** | **T\_FST** | **MIMAR** | **DADI.SI** | **DADI.BB** | **CoalHMM** | **Gpho-CS** | **PSMC 1e6** | **PSMC 1e5** | **PSMC 5e4** | **MSMC** |
| **mean** | 24173 | 59429 | 32535 | 54358 | 44512 | 17546 | 6594 | 20591 | 20715 | 36824 | 27508 |
| **CI.dn** | 23136 | 56242 | 31304 | 50205 | 20767 | 15156 | 5652 | 19963 | 20011 | 31726 | 25220 |
| **CI.up** | 25211 | 62615 | 33766 | 58511 | 68256 | 19936 | 7537 | 21218 | 21419 | 41922 | 29796 |

Note: CI.dn and CI.up represents the upper and lower boundary of 95% confidence interval. We includes the results of DADI with two demographic model, simple isolation model and bottleneck-bottleneck model in column DADI.SI and DADI.BB respectively. We includes the results of PSMC with three different thresholds, 1e6, 1e5 and 5e4.

# Methodology Review

## Statistical Estimators of TMRCA

### Hayes’ method (T-LD)

*Idea of T-LD*

Hayes uses the decline in correlation of LD between two offspring populations with increasing genetic distance to estimate their divergence time. LD structure should be the same in offspring populations right after their divergence from ancestral population. Hence the correlation between LD of two daughter populations should be 1.0 right after divergence, and decays with time due to recombination. If LD is measured by correlation coefficient , Hill and Robertson showed that decays in a manner that after generations over genetic distances of , assuming constant population size and natural population (finite unselected random mating population) [1]. Therefore, follows a linear function of distances () and divergence time can be estimated by ().

*Methodology*

1. Extract the sites that are segregating (polymorphic) in all populations under the study. Estimate LD by correlation coefficient , in each population separately for each pair of SNPs of genetic distance between 0.005cM and 0.1cM and adjust by (1/n) to account for sample size.
2. are binned into 19 categories with equal length of genetic distance and incremental upper boundaries from 0.01cM to 0.1cM. For each LD bins, estimate the correlation of LD () between two populations of interest.
3. Regress onto genetic distance and obtain divergence time .

### McEvoy’s method (T-FST)

*Idea of T-FST method*

Under neutral evolutionary theory, the population genetic differentiation sources from gene drift and can be estimated by . The extent of gene drift depends on effective population size (*)* and population divergence time by [2], thus is an estimator of . However, changed dramatically in human history. According to Hill and Robertson, LD between markers far apart reflects recent and the LD between markers closed together reflect ancient . Sved and Nei 1987 both reported that is approximately true for generations ago, where is the square of genetic correlation coefficient and is the genetic distance [2-4]. Therefore, the effective population size can be estimated by LD structure and population divergence time can be thus derived.

*Methodology*

1. E*xtract the sites that are segregating (polymorphic) in all populations under the study. Compute the average of the SNP-wise .*
2. *Estimate LD by the square of correlation coefficient in each population separately for each pair of SNPs of* *genetic distance from 0.005cM to 0.1cM. values are adjusted for each population by () to account for experimental sample size. Similar to Hayes’ method, are binned into 19 categories with equal length of genetic distance and incremental upper boundaries from 0.01cM to 0.1cM. Effective population size is computed by for each bin and a single point estimation takes the average of the 19 values for each population separately.*
3. *Harmonic means of the effective population sizes of the two populations of interest () is computed and the population divergence time is estimated as .*

## MCMC methods

### MIMAR

MIMAR is a multilocus model for estimating population parameters under isolation-migration model allowing recombination [5]. The parameters of interest include three population mutation rate (), migration rate ( assuming symmetric migration or () assuming asymmetric migration), divergence time () and recombination rate (). Here the population mutation rate , where represents the mutation rate per site per generation. Hence with a fixed , and are equivalent. MIMAR explores the posterior probabilities of and infers the parameters through MCMC.

The data that MIMAR utilizes are the segregating sites summaries for multiple independent neutral loci, each of which has length of hundreds of base pairs. The segregating sites summaries used by MIMAR are , where are numbers of polymorphisms unique to the samples from populations 1 and 2 respectively; is the number of shared SNP between the two samples; is the number of fixed variants in either sample. In order to derive the likelihood, MIMAR assumes loci are independent to each other, thus the likelihood is the product of the likelihood of each locus. Although it is hard to obtain the analytical formula for the full likelihood at one locus ( , where is the observed data at a single locus), can be expressed by a Bayesian framework:

Given an ARG and parameter , likelihood can be either derived from coalescent theory or estimated by the traditional substitution Markov model. The conditional probability can be evaluated using coalescent theory. ARGs {} are sampled for each locus, then can be approximated by . So the overall posterior distribution is in form of

where represent *Y* loci, represents the data at locus . represents ARGs for locus . The prior distributions for are uniform with provided or default boundary.

MIMAR designs an MCMC procedure to sample the parameters and infer the population divergence time by the expected value of its stationary distribution. Instead of directly sample from posterior distribution MIMAR samples from and can be approximated by the sampling values of from this chain at stationary. If now at (), propose a move to () according to the transition kernels (normal with mean) and . With these settings, the Metropolis-Hasting acceptance probability is:

### GPho-CS



Figure 1. Illustrate the model of GPho-CS. There are eight lineages, which two from population A, four from population B and two from population C. The genealogy is compatible with a known phylogeny tree with two migration bands. The scaled population mutation rates for population *A,B,C* and ancestral population AB and ABC are *θA, θB, θC, θAB* and *θABC* respectively. This figure is adopted from a similar figure in the supplementary material of [6].

GPho-CS is a Bayesian MCMC method which utilizes sequence alignments at many neutral loci to explore the posterior distribution of population sizes and population divergence times with a known phylogeny of multi-populations [6]. GPho-CS assumes no intralocus recombination and allows multiple migration bands. In our application, we assume two populations and isolation-migration model.

Consider a known population phylogeny (tree) . For each population , population mutation rate and population divergence time are the parameters of interest (Figure 1). Input observations are haploid (or diploid) sequence alignments at multiple loci {} (represents locus ). GPho-CS uses MCMC to sample parameters according to their joint posterior density , which consists of two main components: (a) the computation of the data density function and (b) the update scheme for ({},{},{},{}), where represents the genealogy for loci and represents the migration rate of migration band .

With several independent assumptions, the data density function is expressed by:

where the prior and are Gamma distribution specified by user; the second product only includes ancient populations; is computed based on coalescent theory and is computed using traditional Markov model of nucleotide substitution [7]. Migration event that lineages migrate from target population to source population in a coalescent process is modelled by a Poisson process with rate . GPho-CS designs a series of Metropolis-Hastings procedure to update the layers of ‘latent’ variables one by one.

## HMM Methods

### CoalHMM

CoalHMM is a hidden Markov model that utilizes a pair of whole-genome haploid alignments which each from one population to estimate population parameters under isolation model [8]. It assumes that the process is Markovian along the alignments and only considers the genealogies of pairs of adjacent nucleotides. CoalHMM uses a discrete state Markov model to depict the coalescent time along the sequences (coalescent HMM model), and uses continuous time finite state Markov models (CTMC) to depict the ancestry of two adjacent nucleotides back in time. The CTMC helps compute the transition probability of the coalescent HMM model.

*CTMC*

CoalHMM developed two CTMCs to model the ancestry of two adjacent nucleotides: one-sequence system and two-sequence system. Back in time, when two populations are isolated, the process of adjacent nucleotides on each alignment are modelled by one-sequence system separately. When two populations coalesce, the process is modelled by a two-sequence system. The hidden states of the one-sequence system and the two-sequence system are shown in Table 3 and Table 4 respectively.

Table 3. The hidden states of two adjacent nucleotides in one sequence system. Linked edge means the two nucleotides are on the same sequence.

|  |  |  |
| --- | --- | --- |
| Index | 1 | 2 |
| State |  |  |

Table 4. The hidden states of two adjacent nucleotides in two sequences system. Open circle means the two sequences found MRCA at the locus, whereas filled circle means MRCA is not found yet. Linked edge means the two nucleotides are on the same sequence. {ΩB, ΩL, ΩR, ΩE} represent the state sets of non-coalescence on both nucleotides, coalescence at left nucleotide, coalescence at right nucleotide, coalescence at both nucleotides, respectively.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Set |  |  |  |  |
| Index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| State |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

The CTMCs have transition rate matrix and and transition matrix and for one-sequence system and two-sequence system respectively.



****

where is the coalescence rate, is the scaled recombination rate, and is the reference effective population size.

*Coalescent HMM Model*

CoalHMM developed a discrete state Markov model to depict the coalescent time along the alignments. The hidden states are discretized coalescent time intervals with break points and , where represents the divergence time. State represents the coalescence occurs in . The distribution of the CTMC states when entering the HMM at state is given by for . The transition probability from state to state is then given by .

When ,

When i<j,

When i>j,

The transition probability calculated in this way are exact the probability according to coalescent theory with recombination.

Emission probabilities are the probabilities that a given pair of nucleotides differs in a given time, which is computed by Jukes-Cantor substitution models. In the discrete model, mid-point of corresponding time interval is used.

There are two ways to estimate parameters: (a) maximum likelihood parameters optimized by a modified Newton-Raphson algorithm and derivatives are computed numerically; (b) MCMC. In our applications, we used MCMC to estimate the parameters.

### PSMC



Figure 2. PSMC uses a hidden Markov model to infer the historical population size based on the basis of the local density of heterozygotes. The hidden states are discretized TMRCA and the transitions are ancestral recombination events. Homozygotes and heterozygotes are colored in red and blue respectively. The figure is adopted from a similar figure in [9].

PSMC is a Markovian method to infer the piece-wise constant ancestral effective population size[9]. The SMC, developed by McVean in 2005, makes the simulation of genomic size sequence possible and likelihood inference tractable. PSMC is a case of two chromosomes SMC model. If PSMC is applied to a pseudo-diploid sequence which each haploid sequence from one population, the divergence time could be qualitatively inferred by the time when combined population’s size increase to infinity.

PSMC utilizes a diploid consensus sequence and assumes that the process is Markovian along the sequence (Figure 2). The hidden states are discretized coalescent time intervals . Within each coalescent time interval, PSMC has a free parameter of effective population size. The transition probability and emission probability of continuous-state HMM are given below and those of discrete-state HMM are computed by taking the integral on the intervals. The maximum likelihood parameters are obtained through Viterbi Learning EM algorithm.

The transition probability is derived from SMC model and given by:

where is the scaled recombination rate, is the Dirac delta function and is the transition probability conditional on there being a recombination event, where is the relative population size at state .

The emission probability follows an exponential distribution of rate (scaled mutation rate): , , where 1 means heterozygote and 0 means homozygote.

### MSMC

MSMC is a multi-sequence extension of PSMC that also infers the piece-wise constant ancestral effective population size [10]. The hidden state of MSMC is the first coalescence represented by a triplet , where is the first coalescence time and and are the label of the two lineages with regard to the first coalescence. Coalescence time is also discretized into intervals with boundary . Suppose there are haploid sequences, then MSMC has hidden states. The transition probability and emission probability are derived under the SMC’ framework and parameters are optimized by Baum-Welch algorithm [11].

When MSMC is applied to two populations, three coalescence rates, and , are used, where and represent the within population coalescence rates for population 1 and population 2 and represents the cross population coalescence rate. MSMC defines cross-coalescence rate = as a measure of relative gene exchange rate between two populations. Population divergence process is shown by cross-coalescence rate decreasing from around one to close to zero.

## Differential Approximation Methods

### DADI

DADI is a diffusion approximation approach which utilizes multi-population allele frequency spectrum (AFS) to infer population evaluation parameters under a particular demographic model [12]. The basic idea is: firstly solve a diffusion equation of AFS, then calculate the expected AFS and compare it with observed AFS, and iterate above steps to find the optimal parameters which maximize the likelihood.

Given a number of sequences from populations, which sequences from population , AFS is defined as a dimensional matrix with each entry, (), counting the biallelic polymorphic sites that the number of derived allele occurrence is in population [13]. Let be the process of the density of derived mutations having relative allele frequency () at a forward time . Under Wright-Fisher model, follows the diffusion equation:

where represents the relative population size of population , represents the scaled fitness coefficient of variants in population , represents the scaled migration rate from population to population . Boundary conditions are no-flux except where all population frequencies are 0 or 1. Complex demographic structure can be modelled by altering the parameters or dimensionality of . The diffusion process can be solved through a finite different method and the expected AFS can be subsequently computed by:

DADI assumes the entries of AFS to be independent Poisson variables of mean . Hence the likelihood of parameter can be derived as below and maximum likelihood parameters can subsequently obtained:

# REFERENCE

1. Hill, W.G. and A. Robertson, *Linkage disequilibrium in finite populations.* Theor Appl Genet, 1968. **38**(6): p. 226-31.

2. Nei, M., *Molecular evolutionary genetics*. 1987, New York: Columbia University Press. x, 512 p.

3. Sved, J.A., *Linkage disequilibrium and homozygosity of chromosome segments in finite populations.* Theor Popul Biol, 1971. **2**(2): p. 125-41.

4. Hayes, B.J., et al., *Novel multilocus measure of linkage disequilibrium to estimate past effective population size.* Genome Res, 2003. **13**(4): p. 635-43.

5. Becquet, C. and M. Przeworski, *A new approach to estimate parameters of speciation models with application to apes.* Genome Res, 2007. **17**(10): p. 1505-19.

6. Gronau, I., et al., *Bayesian inference of ancient human demography from individual genome sequences.* Nat Genet, 2011. **43**(10): p. 1031-4.

7. Felsenstein, J., *Evolutionary trees from DNA sequences: a maximum likelihood approach.* J Mol Evol, 1981. **17**(6): p. 368-76.

8. Mailund, T., et al., *Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model.* PLoS Genet, 2011. **7**(3): p. e1001319.

9. Li, H. and R. Durbin, *Inference of human population history from individual whole-genome sequences.* Nature, 2011. **475**(7357): p. 493-6.

10. Schiffels, S. and R. Durbin, *Inferring human population size and separation history from multiple genome sequences.* Nat Genet, 2014. **46**(8): p. 919-25.

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

12. Marth, G.T., et al., *The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations.* Genetics, 2004. **166**(1): p. 351-72.

13. Keinan, A., et al., *Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans.* Nat Genet, 2007. **39**(10): p. 1251-5.