

**STATISTICAL CHALLENGES IN NEXT GENERATION  
POPULATION GENOMICS STUDY**

**ZHOU JIN**

**(BSc. Peking University)**

**A THESIS SUBMITTED**

**FOR THE DEGREE OF DOCTOR OF PHILOSOPHY**

**DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY**

**NATIONAL UNIVERSITY OF SINGAPORE**

**2015**

## DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Candidate: Jin Zhou

Signed: *Zhou Jin*

Date: 26/07/2015

## **ACKNOWLEDGEMENTS**

This thesis and all the work over the last 4 years would not have been possible without the love and support of everyone who has stood behind me all the way.

I would like to express my sincere thanks to A/P TEO Yik Ying for his supervision and suggestions throughout these 4 years. YY brought me to a new interesting world of bioinformatics. YY was very generous with his time and knowledge and assisted me in each step towards the completion of the thesis.

I would like to acknowledge the invaluable help rendered by Rick, Wang Xu, Xuanyao, Woei Yuh and Eryu for their accompaniment for the four years. I realized the importance of moral support when working together. I would also like to thank the IT experts Taylor, Anthony and Bowen. These guys have never turned me away when I have problems with work.

I would like to thank NUS Department of Statistics and Applied Probability for financial support for my four years' study. I am most indebted to Genome Institute of Singapore for providing me resources of genotyping data. Especially I would like to thank Dr. Chiea-Chuen Khor and Dr. Erwin Tantoso's support for my research.

Finally I will like to thank the support from my family, for always believing in me, and for the endless encouragement that has kept me going.

## CONTENTS

Contents .....	ii
LIST OF TABLES .....	vi
LIST OF FIGURES .....	vii
Chapter 1. Introduction.....	1
1.1 High Throughput Genetics Era .....	1
1.2 Genotype Calling.....	3
1.3 Linkage Disequilibrium.....	5
1.4 Rare Variants.....	7
1.5 Population Structure and Migration .....	8
1.6 Population Divergence Time Estimation .....	10
1.7 Concepts in Population Genetics.....	13
1.8 Description of the Thesis.....	16
Chapter 2. Large Scale Genotype Calling .....	18
2.1 Background .....	18
2.2 Chip Design and Data .....	20
2.3 Existing Genotype Calling Algorithms .....	23
2.4 Method .....	38
2.5 Application to Data from Exome Microarray & Method Comparison.....	47
2.6 Discussion .....	54

2.7	Supplementary Information.....	57
Chapter 3.	Statistical Evaluation of TMRCA algorithms.....	64
3.1	Modern Methods of Estimating TMRCA .....	64
3.2	Theories in Population Genetics .....	64
3.3	Methods for Estimating TMRCA.....	74
3.4	Simulation and Real Data Application.....	88
3.5	DISCUSSION .....	102
Chapter 4.	Conclusion .....	106
4.1	Future work for iCall.....	108
4.2	Future Work for TMRCA .....	109
REFERENCE.....		110

## **SUMMARY**

After a decade of international whole genome endeavors, common patterns of genetic variations of the human genome are catalogued through microarrays genotyping. With the advent of next generation sequencing platforms, an overwhelming majority of novel variants identified by whole genome sequencing are of lower frequencies, encouraging the array-based follow-up studies emphasizing on low frequency and rare variants. Besides, next generation sequencing facilitates population genetics research by providing fine-scale haplotype sequence of individual genome that contains all forms of polymorphisms, linkage disequilibrium information and patterns of genetic variations.

The first study in this thesis investigates the microarray genotype calling issue for low frequency and rare variants. Existing genotype calling algorithms are developed mainly for common SNPs and present many problems for rare variants. In this thesis, we design and introduce a new method, iCall, for a robust genotyping of common, low-frequency and rare SNPs, and we show that iCall outperforms existing genotype calling algorithms.

The second study in this thesis continues the theme of investigating the impact that sequencing technologies bring to genetics research. Specifically we evaluate existing methods for estimating the divergence time of closely related populations. This considers and compares genetic data obtained from genotyping and sequencing, and evaluate the relative performance of the different methods in terms of their robustness and accuracy through a series of

simulations under different demographic scenarios, followed by estimating the population divergence time between Southeast Asian Malays and South Asian Indians.

## LIST OF TABLES

Table 1. Comparison of iCall against optiCall, Illuminus, GenCall and GenoSNP at 16 428 SNPs at different sample sizes for calling, where genotypes from whole-genome sequencing of 81 samples are used as benchmark. This figure has been adapted from Table 1 in Zhou et al. (2014) Bioinformatics Vol. 30 no. 12 [62].	51
Table 2. Comparison of iCall+zCall, GenCall+zCall and optiCall+zCall at 16 428 SNPs at different sample sizes for calling, where genotypes from whole-genome sequencing of 81 samples are used as benchmark. This figure has been adapted from Table 2 in Zhou et al. (2014) Bioinformatics Vol. 30 no. 12 [62].	52
Table 3. The resources of whole genome sequencing data. 45 samples came from SSMP and 36 samples came fromSSIP.	59
Table 4. Comparison of TMRCA methods.	63
Table 5. The hidden states of two adjacent nucleotides in one sequence system. Linked edge means the two nucleotides are on the same sequence. This table has been adapted from a similar figure in reference [88].	82
Table 6. 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. $\{\Omega_B, \Omega_L, \Omega_R, \Omega_E\}$ represent the state sets of non-coalescence on both nucleotides, coalescence at left nucleotide, coalescence at right nucleotide, coalescence at both nucleotides, respectively. This table has been adapted from a similar figure in reference [88].	82



## LIST OF FIGURES

Figure 1. Illustration of a biallelic SNP. ....	3
Figure 2. (a) Isolation model (b) Isolation migration model .....	11
Figure 3. ‘Out-of-Africa’ global migration history. Figure has been adapted from a similar figure in reference [49]......	12
Figure 4. Illustration of the Wright Fisher model. Consider a diploid population consisting of $N_0$ diploidy individuals, in total $2N_0$ haploid copies of genes. All haploid copies of gene in generation $t$ are drawn randomly from all copies of gene in generation $t-1$ . ....	15
Figure 5. Illustration of the design of Illumina microarray. A target sequence is bound to the matched oligonucleotide bead and its fluorescent end fluoresces red/green light to signal the hybridization of corresponding allele. ....	20
Figure 6. Illustration of hybridization intensity profiles for three different SNPs at both the allelic intensities axes and the transformed contrast scale axes. The three SNPs correspond to (i) a common SNP with $MAF \geq 5\%$ (panels A and B); (ii) a polymorphic SNP with $MAF < 5\%$ (panels C and D); (iii) a common SNP with shifted intensity clusters (panels E and F). In each panel, the assigned genotypes are colored accordingly as AA (red), AB (green), and BB (blue). This figure has been adapted from Figure 1 in Zhou et al. (2014) Bioinformatics Vol. 30 no. 12 [62]. ....	22
Figure 7. Illustration of the erroneous calling made by Illuminus. Illuminus is not robust for low frequency and rare variants since it tends to cluster intensities in more clusters than observed. ....	31
Figure 8. Illustration of the erroneous calling of GenoSNP. The three panels in the first row illustrate the calling for single sample across multiple SNPs. The three panels in the second row illustrate the calling for single SNP across multiple samples. ....	34
Figure 9. Illustration of erroneous calling made by optiCall. ....	36
Figure 10. Illustration of the algorithm of zCall. Two intensity thresholds $t_x$ and $t_y$ are used to cluster the intensities into genotype classes. ....	37
Figure 11. Histograms of the absolute value of the contrast coordinates for 12 370 samples at three SNPs with different MAFs, corresponding to a (A) common SNP ( $MAF \geq 5\%$ ); (B) low-frequency or rare SNP ( $0\% < MAF < 5\%$ ); and (C) monomorphic SNP ( $MAF = 0\%$ ). This figure has been adapted from Figure 2 in Zhou et al. (2014) Bioinformatics Vol. 30 no. 12 [62]. ....	39
Figure 12. The first penalty term penalizes on small distances between the heterozygous cluster and the two homozygous clusters. ....	43
Figure 13. An example of a genealogical tree of a sample of 6 genes. The column on the right shows the equivalent relations of the genealogy. ....	65

Figure 14. Jukes-Cantor model. The four nucleotides substitute in a Markovian manner.....	68
Figure 15. An example of a genealogy of 3 genes. Vertices 1-3 represent present genes, vertices 4 represent an ancestral gene and vertex 0 represents their MRCA. Edges v1-4 represent the length of time past for a coalescence event.....	69
Figure 16. Illustration of ARG of a sample of three genes. (A) Two recombination occurred at $t_2$ and $t_4$ . These recombination separate the gene into three segments ( $b_1, b_2$ ), ( $b_2, b_3$ ) and ( $b_3, b_4$ ) (ancestral material colored in yellow, green and blue, respectively; non-ancestral material colored in grey). (B) The corresponding genealogies of ( $b_1, b_2$ ), ( $b_2, b_3$ ) and ( $b_3, b_4$ ) are colored in yellow, green and blue respectively.....	70
Figure 17. Illustration of the sequential Markov coalescent model with 5 genes. The cross-mark indicates the point of recombination, which is uniformly distributed on the genealogy. The branch above the recombination point is removed, resulting in a floating branch which coalesces with existing lineages at the rate proportional to the number of lineages present. The figure has been adapted from a similar figure in reference [81]. .....	72
Figure 18. Illustration of the SMC' model with 5 genes. The cross-mark indicates the point of recombination, which is uniformly distributed on the tree. The floating branch coalesce with existing lineages at the rate proportional to the number of lineages present before erasing the branch above the recombination point. (a) represents the situation that the floating branch coalesces with branch other than its ancestral branch; (b) represents the situation that the floating branch coalesces with its own ancestral branch (in this case, the recombination event will not change LD pattern).....	74
Figure 19. Illustration of the model of GPho-CS. There are eight lineages, with 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 $\theta_A, \theta_B, \theta_C, \theta_{AB}$ and $\theta_{ABC}$ respectively. This figure has been adapted from a similar figure in reference [87].....	80
Figure 20. 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 TMRCAs and the transitions are ancestral recombination events. Homozygotes and heterozygotes are colored in red and blue respectively. The figure has been adapted from a similar figure in reference [89]. .....	84
Figure 21. Illustration of the four demographic scenarios considered in our simulation study. An ancestral population diverged into two populations (population_1 and population_2) at time $T_{split}$ . $N_1, N_2$ and $N_a$ are the effective population size of population_1, population_2 and the ancestral population, respectively. (i) simple-isolation-model: ancestral population split into two populations at 20Kya. (ii) isolation-migration-model: a symmetric	

migration rate is added after the split. (iii) bottleneck-nonbottleneck-model: ancestral population split into two populations at 60Kya after which population\_2 has constant effective population size and population\_1 experienced a bottleneck. (iv) bottleneck-bottleneck-model: ancestral population split into two populations at 40Kya, after which both population\_1 and population\_2 have population size declined instantly and afterwards increased exponentially.....89

Figure 22. Mean error rate and 95% confidence interval are obtained from 10 iterations. Except MIMAR-prior and DADI-prior, the estimations are obtained with simple isolation model. MIMAR-prior and DADI-prior show the results obtained with prior knowledge of the demographic model for scenario (ii), (iii) and (iv).....98

Figure 23. Illustrate the point estimation and corresponding 95% confidence interval of TMRCA for Southeast Asian Malays and South Asian Indians by the eight methods. DADI.SI and DADI.BB show the estimates of DADI with isolation model and bottleneck-bottleneck-model respectively..... 101

Figure 24. Illustrate the estimation of TMRCA by (A) PSMC and (B) MSMC on whole-genome sequencing data for the 22 autosomal chromosomes from Southeast Asian Malays and South Asian Indians. Both the effective population size (panel A) and the cross-coalescence rate (panel B) are modelled as step functions. The divergence time for the two populations is defined for (A) PSMC as the time when the effective population size increases to infinity, which in practice is implemented as a threshold such as 100,000 in our study; (B) MSMC as the most recent time when the cross-coalescence rate decreases below an arbitrarily selected threshold, which in our study the threshold is selected as 0.5..... 103

## **CHAPTER 1. INTRODUCTION**

### **1.1 High Throughput Genetics Era**

In modern genetics, the invention of cloning and sequencing technologies utilizing recombinant DNA has enabled us to understand and study the nature of genetic information directly [1-3]. The molecular basis for genes is deoxyribonucleic acid (DNA) that consists of nucleotide sequences for known or unknown cellular functions or processes. The nucleotide sequences are read and translated by cells to produce amino acid sequences which in turn fold into proteins. The genomes of any two individuals are about 99.9% identical remaining 0.1% DNA sequence variation is largely attributed to: (i) single nucleotide polymorphisms (SNPs), also called markers, referring to single base changes in the human genome sequence [4]; and (ii) structural variants comprising of genomic alterations such as copy number polymorphisms, insertions, deletions and duplications [5]. SNPs are notably the most common genetic variation [6]. A large majority of the SNPs has a minimal impact on biological system, whereas a few SNPs can be functional, causing changes in amino acids, mRNA transcriptions and translations [7]. The human genome contains millions of SNPs and all of which can potentially contribute to cell function [8]. High throughput techniques leverage automation to quickly assay the human gene that encompass from the target regions to the whole-genome. It makes the unfinished genomic sequence data rapidly available to the researches.

The field of human genetics has developed rapidly in the past decade. It is highly encouraged by the development of the genomic mapping technology,

from genome-wide linkage mapping to low and high throughput single nucleotide polymorphism genotyping, and the most recent high throughput genome sequencing [9]. With the advent of cost-efficient technology, it is now feasible to generate and analyze terabytes of genetic data to investigate gene association with complex diseases, the biological processes of DNA inheritance and evolutionary histories of human populations.

The first two generations of linkage maps were restriction fragment length polymorphisms (RFLPs) and microsatellites, and both covered only hundreds of polymorphic markers on each platform. The third generation linkage map, SNP genotyping (also called microarray genotyping), emerged in the mid 1990s and was first used to study genetic variation in 2000 [10]. Microarrays developed rapidly during the course of the last decade, covering from tens of thousands to several million polymorphism markers of, which marks the epoch of high throughput genetics.

The International HapMap Project (HapMap) is a multi-country effort, launched in 2002, with aims to catalogue common patterns of genetic variations through microarray genotyping. In 2008, the HapMap project catalog contained 3.5 million common SNPs across 11 populations around the globe. It investigated the linkage disequilibrium (LD) structure of human genome, guided the design of genetic studies and was the key resource for researchers to find genetic variants affecting health as well as investigate population diversity and population structure.

The design of microarrays has depended on existing information about genetic variants in the human genome, and this has resulted in a greater propensity to include genetic variants that are more likely to be polymorphic across multiple populations than genetic variants that are of lower frequencies or rarer. As a result, the coverage of genotyping microarrays is skewed in favor of common variants. The advent of next-generation sequencing (NGS) brought a more comprehensive discovery of low-frequency and rare variants. NGS sharply reduced the cost of sequencing and has enabled rapid sequencing of large stretches of DNA base pairs spanning entire genomes.

The 1000 Genome Project (1KGP) is an international genetic research effort, launched in 2008, aiming to establish the most detailed catalogue of human genetic variation. More specifically, 2500 individuals from populations of Asian, European, African, and American ancestry will be sequenced and information on variants with frequencies down to 1% can be gathered. To date, 1KGP has provided a deep characterization of human genomic variations, which brings an unprecedented opportunity to study population evolution.

## 1.2 Genotype Calling

TTGCAGTGCAAG<sup>A</sup><sub>C</sub>ACAGTAAGCTCA

**Figure 1. Illustration of a biallelic SNP.**

Genotype calling is the process of determining the genotype of an individual at each SNP. Most typical SNPs are biallelic, with two possible alleles segregating in a population (Figure 1). Mendelian inheritance states that every

individual contains a pair of alleles for each particular trait (assuming diploidy). Hence if we let  $A$  and  $B$  represent generically the two possible alleles, a biallelic SNP has three possible genotypes:  $AA, AB, BB$ .

High throughput genotyping of millions of genetic variants can be achieved by using pre-designed high density oligonucleotide microarray chips. On each chip, there are hundreds of thousands of probes of defined sequences so that many SNPs could be interrogated simultaneously. Matched probes as well as mismatched probes are included in the chip, both of which have the potential to hybridize to target DNA. To reduce the effect of erroneous hybridization, several redundant probes are used to interrogate each SNP. The genotype can be determined by comparing the differential amount of hybridization of the target DNA to each of these redundant probes [11].

There are two major producers of oligonucleotide microarray chips, Affymetrix Inc. and Illumina Inc. Affymetrix introduced its microarrays, including GeneChip Mapping 10K Array, Mapping 100K Array, Mapping 500K Array, Human SNP Array and Genome-wide Human SNP Array, between 2004 and 2009. The chip's feature has improved over time, from only containing 10,000 markers to more than 1.8 million markers [12-16]. Each array consists of millions of 25 base pair oligonucleotide probes, which emits fluorescence at the fluorescent end when they bind to the target sequences. Each SNP is interrogated by five probe quartets, each of which consists of four pairs of perfect match and mismatch probes. Genotype can be called according to the pixel intensity of fluorescence for each SNP [13].

Illumina introduced its first microarray, the Human-1 Genotyping BeadChip, in 2005, followed by the HumanHap family and the Omni family, increasing their total dataset from 100,000 markers to 5 million markers. The latest microarrays, the Omni family, are characterized as high throughput genotyping arrays which provide access to newly discovered SNPs with lower frequencies [17]. Genomic markers are interrogated and detected through the process: (i) 50-mer probes hybridize to the loci of interest; (ii) marker specificity is conferred by enzymatic single-base extension to incorporate a labeled nucleotide; (iii) dual-color fluorescent measures the intensities of two alleles [18].

The platforms offered by these companies differ in terms of array fabrication, probe design, sample preparation and hybridization protocol [19]. Hence, genotype calling algorithms are usually developed for specific platforms. These calling procedures are usually automated due to the massive scale of the genotyping and erroneous calls are possible. These erroneous calls have the potential to cause confounding in downstream studies [20]. Thus the development of accurate calling algorithms is an important topic of research.

### **1.3 Linkage Disequilibrium**

Microarrays are predesigned with tagging SNPs according to existing genome annotations which are mostly common variants (minor allele frequency (MAF)  $\geq 5\%$ ). Although genotyping microarrays contain an increasing number of markers, it is at present too expensive to directly interrogate all common



variants in the human genome to catalogue common patterns of genetic variations.

Linkage disequilibrium (LD) refers to the non-random association between neighboring alleles resulting from coinheritance of genetic SNPs [21].

Assuming all the alleles descended from a single ancestral chromosome, new alleles differing from the ancestral chromosome arise from historical mutations [22]. Collections of specific alleles orderly arranged on a chromosome that are likely to be inherited together are called haplotypes [23]. New haplotypes are generated by mutations or recombination and the coinheritance of the haplotypes reflect LD structure. LD extends the promise of being able to survey the genome by choosing a minimal number of markers for each LD block as proxies.

The extent and strength of LD is affected by genetic factors such as mutation, recombination and selection, as well as human demographic factors such as population structure and migration [24]. Hence research on LD is very important in understanding population evolutionary history [25, 26].

There are many measures formulated to assess the strength of LD. The genetic correlation coefficient  $r$ , the square of genetic correlation coefficient  $r^2$  and Lewontin's  $D'$  are commonly used [27, 28]. Consider the haplotypes for two biallelic loci, with allele  $A$  and  $a$  at one locus and allele  $B$  and  $b$  at the other locus. Let  $p_A, p_a, p_B, p_b$  denote the four allele frequencies, and  $p_{AB}, p_{Ab}, p_{aB}, p_{ab}$  represent the four haplotype frequencies.

The genetic correlation coefficient,  $r$ , is defined as:

$$r = \frac{p_{AB} - p_A p_B}{\sqrt{p_A p_B p_a p_b}}$$

Lewontin's  $D'$  is defined as:

$$D' = \begin{cases} \frac{p_{AB} - p_A p_B}{\min(p_A p_b, p_a p_B)} & \text{if } p_{AB} - p_A p_B > 0 \\ \frac{p_{AB} - p_A p_B}{\min(p_A p_B, p_a p_b)} & \text{if } p_{AB} - p_A p_B < 0 \end{cases}$$

## 1.4 Rare Variants

Microarray genotyping, facilitated by LD, successfully captures more than 90% of genetic variation and was used to establish the linkage map of the human genome. However, microarray genotyping prioritizes common variants and is ineffective in discovering low frequency ( $1\% \leq \text{MAF} < 5\%$ ) and rare ( $\text{MAF} < 1\%$ ) variants. Because of next generation sequencing, it has become possible to directly sequence and formulate accurate haplotype information of human genome. A majority of rare variants is discovered by a variety of international sequencing endeavors.

Rare variants show a systematically different and typically stronger population stratification than common variants, especially as rare variants are found to be more geographically localized and tend to be population specific. Studies show that rare variants can reveal fine-scale population substructures beyond those inferred by common variants [29], demonstrating that rare variants can contain more information about recent human population evolution than common variants.

## 1.5 Population Structure and Migration

In our context, population structure (also called population stratification) refers to the systematic genetic variation in allele frequencies between populations. Human populations have gone through a complex migration and settlement process, and diverged into sub-populations with possible mating preference and restrictions such as geography, environment or social interaction. As a result of genetic drift or divergent natural selection, populations become genetically differentiated over time and the amount of genetic differentiation is related to the historical evolution process [30]. Studies show that genetically related populations are more likely to cluster geographically. Correlations in genotype data cluster well for continents of origin including Eurasia, east Asia and Africa [31, 32].

### 1.5.1 F-statistics ( $F_{ST}$ )

Quantifying patterns of human genetic variation across global populations is important in understanding the population structure. The F-statistics, first introduced by Wright in 1921, describe the partitioning of genetic diversity within and among populations and are the most widely used metrics to quantify and detect population structure [33]. The three interrelated parameters introduced are  $F_{IT}$ ,  $F_{ST}$ ,  $F_{IS}$ , representing the correlation of genes within individual relative to in the combined population, of different individuals in the same population relative to in the combined population, and within individuals relative to within the population it belongs to, respectively. They follow the relation  $(1 - F_{IT}) = (1 - F_{ST})(1 - F_{IS})$ [34].

$F_{ST}$  is directly related to the variance of allele frequency among different populations [35]. A large  $F_{ST}$  value indicates significant differences in allele frequencies among populations and small  $F_{ST}$  value indicates similarity in allele frequencies among populations. At one locus, if natural selection favors one allele in one population, its  $F_{ST}$  value tends to be larger than other loci without selection. If natural selection favors one allele in both populations, its  $F_{ST}$  value tends to be smaller than that of loci with pure genetic drift. Many estimations of  $F_{ST}$  have been developed with different assumptions about sample sizes or the number of populations. The most widely used estimation was introduced by Weir and Cockerham in 1984, which we used to estimate  $F_{ST}$  in this thesis.

Consider allele A. Let  $a, b, c$  represent the variance of the frequency of allele A between populations, between individuals within populations, and between gametes within individuals, respectively, and  $p$  represent the frequency of allele A in the ancestral population. The expectations of  $a, b, c$  take the forms of [36]:

$$E(a) = p(1 - p)F_{ST}$$

$$E(b) = p(1 - p)(F_{IT} - F_{ST})$$

$$E(c) = p(1 - p)(1 - F_{IT})$$

Then estimates of the three F-statistics are given by:

$$1 - \widehat{F}_{IT} = \frac{c}{a + b + c}$$

$$\widehat{F}_{ST} = \frac{a}{a + b + c}$$

$$1 - \widehat{F}_{IS} = \frac{c}{b + c}$$

$a$ ,  $b$  and  $c$  can be estimated from a weighted analysis of variance [37]:

$$a = \frac{\bar{n}}{n_c} \left\{ s^2 - \frac{1}{\bar{n} - 1} \left[ \bar{p}(1 - \bar{p}) - \frac{r - 1}{r} s^2 - \frac{1}{4} \bar{h} \right] \right\}$$

$$b = \frac{\bar{n}}{\bar{n} - 1} \left[ \bar{p}(1 - \bar{p}) - \frac{r - 1}{r} s^2 - \frac{2\bar{n} - 1}{4\bar{n}} \bar{h} \right]$$

$$c = \frac{1}{2} \bar{h}$$

where

$\tilde{p}_i$  is the frequency of allele A in the sample of size  $n_i$  from population  $i$  ( $i = 1, 2, \dots, r$ ),

$\tilde{h}_i$  is the proportion of individuals heterozygous for allele A in population  $i$  ( $i = 1, 2, \dots, r$ ),

$\bar{n} = \sum_i \frac{n_i}{r}$ , the average sample size,

$n_c = (r\bar{n} - \sum_i \frac{n_i^2}{r\bar{n}}) / (r - 1) = \bar{n}(1 - C^2/r)$ , with  $C^2$  denotes the squared coefficient of variation of sample sizes,

$\bar{p} = \sum_i n_i \tilde{p}_i / r\bar{n}$ , the average sample frequency of allele A,

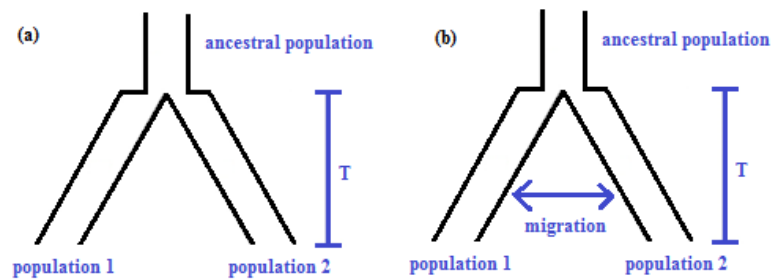
$s^2 = \sum_i n_i (\tilde{p}_i - \bar{p})^2 / (r - 1)\bar{n}$ , the sample variance of allele A frequencies over populations,

$\bar{h} = \sum_i n_i \tilde{h}_i / r\bar{n}$ , the average heterozygote frequency for allele A.

## 1.6 Population Divergence Time Estimation

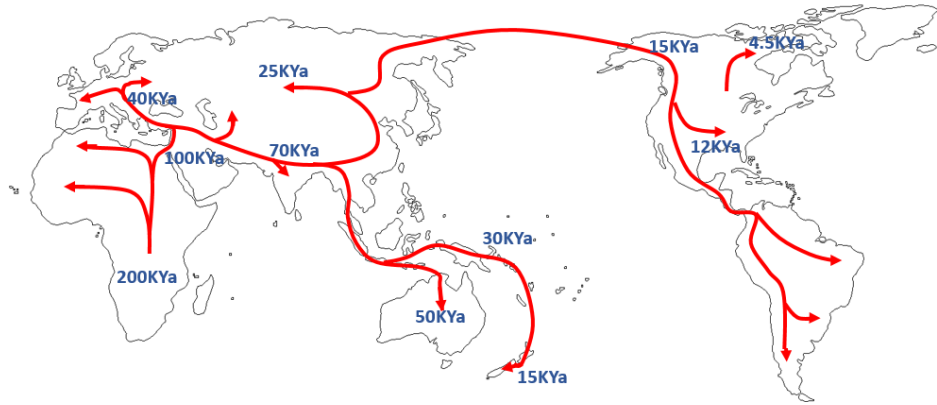
Population divergence is the process in which populations of the same ancestry accumulate genetic mutations independently over a period of time, producing sufficient genetic distinction between these populations as a result of an extended period of reproductive isolation. The population structure can be summarized by a tree showing the genetic distances between populations

and the history of population divergence [38]. The simplest population structure is isolation model (Figure 2 (a)), in which two random mating populations 1 and 2 with constant effective population sizes diverged  $T$  generations ago from an ancestral population. The isolation migration model (Figure 2 (b)) adds migration to the two populations after the original divergence. Although the assumptions of random mating and constant population size are not fully realizable, these two models are widely used by researchers to study and simulate human evolution.



**Figure 2. (a) Isolation model (b) Isolation migration model**

The inference of the divergence time between populations has been of fundamental interest in the study of population evolution. While there is a consensus around the origin and proliferation of modern humans in Africa, dated respectively at about 200,000 and 100,000 years ago, there have been several conflicting theories on the exact nature of modern human dispersal across the globe [39, 40]. The availability of genome-wide data by technologies ranging from genotyping to next-generation sequencing provides the unprecedented opportunity to study the anthropology and migration of modern humans that shaped the existing global distribution of human populations, in an evolutionary process driven by demographic changes, genetic drift and natural selection [41, 42]. Already, valuable insights have been derived from deep genetic surveys of populations in Africa [43, 44], Asia [45, 46], Europe [47], and the Americas [48].



**Figure 3. ‘Out-of-Africa’ global migration history. Figure has been adapted from a similar figure in reference [49].**

An early notable study on population divergence time was the ‘Out of Africa’ theory [40] (Figure 3). Early studies estimated population divergence time mainly based on molecular clock theory and mtDNA or Y chromosomes [40, 50]. However, the assumption of molecular clock theory that mutations in a particular genetic system occur at a deterministic and steady rate is criticized for failing to take account of the stochasticity of gene drift. Although mtDNA and Y chromosome are convenient to use, they account for only a minority of the heritable sequences and contain much less evolutionary information owing to ancestral recombination events on autosomal chromosomes.

Modern population genetics theories made improvement in considering the stochasticity and complex gene forces. Coalescent theory, developed independently by several researchers [21, 35, 51, 52] and formalized by John Kingman in 1982, is the most important theory now (see Section 3.2.1). It provides a stochastic model to trace all alleles in a sequence shared by all members of a population backward in time to a single ancestral copy. Based on coalescent theory, many methods have been developed to estimate

TMRCAs (see Section 1.7.7), a reasonable surrogate of the population divergence time.

## **1.7 Concepts in Population Genetics**

### **1.7.1 Mutation**

A mutation refers to a change in the nucleotide sequence in the genome, which can be a single base substitution, insertion or deletion. Mutation provides a continual source of genetic variation to a population that is passed on to subsequent generations. Many of these mutations are likely removed through the process of negative selection, while the remainder can accumulate to a high frequency in the population over time [41]. The rate of mutation is low, and independent estimates have suggested that the mutation rate in autosomal chromosomes is between  $1.0 \times 10^{-8}$  to  $2.5 \times 10^{-8}$  mutations per site per generation [53].

### **1.7.2 Recombination and Genetic Distance**

Recombination (also called crossover) refers to the genetic events that two chromosomes of a homologous pair exchange their genetic material and produce recombinant chromosomes during the formation of a gamete in meiosis. The expected number of crossovers between the loci per meiosis is used to measure the genetic distance of the loci. The unit of genetic distance is the Morgan (M) (or centiMorgan (cM)), referring to the distance within which an average of one crossover occurs for every meiosis (or every 100 meioses).

### **1.7.3 Random Mating and Hardy-Weinberg Equilibrium**



Random mating assumes that there are no mating preferences or restrictions such as environment or social interaction, and every individual has the same chance to mate with every other individual in the population. When an infinite large random mating population is free from other evolutionary forces, it is in Hardy-Weinberg Equilibrium (HWE), which states that the allele and genotype frequencies in a population will remain constant from generation to generation. Assuming in a population the frequency of allele A and allele B is  $p$  and  $(1 - p)$ , the frequencies of three genotypes in the population are:

$$p_{AA} = p^2, p_{AB} = 2p(1 - p), p_{BB} = (1 - p)^2.$$

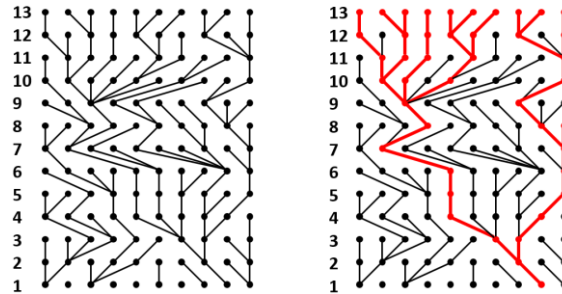
#### 1.7.4 Natural Selection, Neutrality and Genetic Drift

Natural selection is a key mechanism of evolution that favors or induces survival and perpetuation of one kind of biological traits over others. The key concept in natural selection is fitness, which describes the ability to both survive and reproduce.

In 1960s, Motoo Kimura introduced the neutral theory of molecular evolution, using diffusion equations to calculate the distribution of the allele frequencies. Neutral theory claims that most polymorphisms do not influence the fitness of an individual and are not subjected to selection. The main force that changes allele frequencies is genetic drift [54] that the allele frequency of a new allele introduced by mutation is a stochastic process and can rise and spread in a population or get lost due to the random sampling of organisms. When the population size is large, the allele frequency will not fluctuate dramatically and will remain stable. When the population size is small, gene

drift will lead to the allele frequency changing rapidly and will cause some alleles to become fixed and some alleles lost in the population.

### 1.7.5 Wright Fisher Model



**Figure 4. Illustration of the Wright Fisher model.** Consider a diploid population consisting of  $N_0$  diploidy individuals, in total  $2N_0$  haploid copies of genes. All haploid copies of gene in generation  $t$  are drawn randomly from all copies of gene in generation  $t-1$ .

The Wright Fisher Model is a genetic drift model for a single locus with assumptions of finite population size, discrete and non-overlapping generations, random mating, equal sex ratio and equal fitness for all individuals [42]. Successive generations are produced by multinomial sampling from previous generation so that all individuals have an equal probability to be picked as a parent. Assume a diploidy population of size  $N_0$ .

Here are two straightforward conclusions:

- a. Considering two lineages, the number of generations until two lineages have a common parent follows a geometric distribution of rate  $\frac{1}{2N_0}$ .
- b. For a sample of size  $n$ , there are  $C_2^n$  possible coalescent pairs. Let  $W_n$  be the number of generations until the first coalescence, then  $W_n \sim \text{Geometric}(C_2^n/2N_0)$  or  $W_n \sim \text{Exponential}(C_2^n/2N_0)$  when  $n \ll N_0$ .

### 1.7.6 Effective Population Size

The Wright-Fisher model assumes that all individuals in a population have an equal chance of breeding. The effective population size refers to ‘the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift’, which should be smaller than the census population size [55]. Effective population size is an important parameter in population genetic studies and most of the estimates of human population size are in order of  $10^4$ .

### 1.7.7 TMRCA

In the Wright-Fisher model as well as coalescent theory, all gene samples are ultimately inherited from a single ancestral copy called the most recent common ancestor (MRCA). The time to the most common ancestor (TMRCA) refers to the time that has elapsed since the MRCA of a set of gene copies lived. For example, in the Wright-Fisher model, let  $W_n$  be the time until the first coalescence occurs in a sample of size  $n$ . Thus  $TMRCA = \sum_{k=2}^n W_k$ . TMRCA can be estimated by statistical estimators based on DNA data and established mutation rates as practiced in genetic genealogy. The TMRCA has been commonly used as a reasonable surrogate for the population divergence time.

## 1.8 Description of the Thesis

High throughput genetics allows array-based and sequencing-based population genetic research to proliferate and extends the catalog of genetic variation to the whole allele frequency spectrum. This chapter has provided an introduction to some key concepts of population genetics. Subsequent chapters

will explore two issues raised in the high throughput genetic era related to population genetics.

Through next generation sequencing, a large collection of rare variants have been discovered, resulting in new microarray designs that have been customized to interrogate genetic variants of lower frequencies. Chapter 2 discusses the problem of large-scale genotype calling for rare variants and provides a brief review of existing methods. Subsequently we introduce a novel genotype calling algorithm and compare it against existing methods.

Population divergence time is an important parameter in understanding population structure and evolution. Chapter 3 briefly introduces some key theories of population genetics, followed by an in-depth review of different methods for estimating population divergence time. Subsequently we perform a formal statistical evaluation study on the existing methods using a systematic simulation and apply the methods to sequencing data of Southeast Asian Malays and South Asian Indians.

The last chapter discusses the main conclusions of our work and also some aspects that could be explored in future work.

## **CHAPTER 2. LARGE SCALE GENOTYPE CALLING**

### **2.1 Background**

Early generations of genotyping microarrays prioritized tagging SNPs identified from the International HapMap Project [4] that are selected on their ability to provide adequate coverage of the human genome in the HapMap populations. Over the last decade, the International HapMap Project has provided a useful and functional haplotype map of the human genome, which facilitated many types of genetic studies such as population evolution study, association studies and pharmacogenomics.

In Phase 3 of the HapMap, over 1.8 million SNPs were genotyped in 1,184 reference individuals from 11 global populations [6]. The SNPs in the HapMap database were selected to preferentially include common variants and included only a small subset of low-frequency variants, as only 10-13% had  $MAF < 5\%$ . In addition, 100-kb regions of 692 individuals were sequenced in HapMap phase III, in which 42-66% of the segregating sites have  $MAF < 5\%$  [6], showing that a substantial proportion of variants on human DNA have lower frequencies.

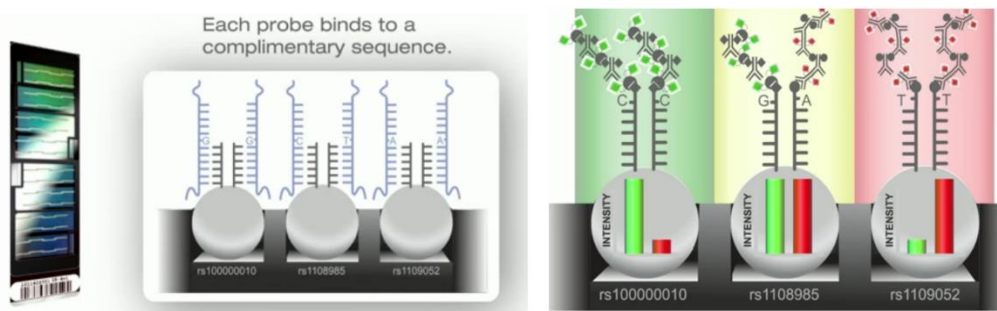
Next-generation genotyping microarrays have been designed with insights from 1KGP and whole genome and exome-sequencing studies to increase genome coverage and to include low-frequency and rare variants that are often ancestry specific [56]. Such microarrays help to provide a more comprehensive understanding of the variations in human populations and to

explore the role of rare variants [57]. For example, the exome microarray was specifically designed to contain mostly variants with  $MAF < 1\%$ .

Determining the genotypes of these low-frequency and rare variants from hybridization intensities is challenging as there is less support to locate the presence of the minor alleles when the allele counts are low. Existing genotype calling algorithms are mainly designed for calling common variants and are notorious for failing to generate accurate calls for low-frequency and rare variants. Therefore, there is a need for a robust genotype calling algorithm that is capable of accurately determining the genotypes for both common and rare variants.

In the following sections, we will discuss the design of the Illumina microarrays and provide a review of the existing genotype calling algorithms developed for Illumina microarray. Subsequently we will propose our new method, which is benchmarked against four of the most commonly used single-stage algorithms as well as different iterations of two-stage calling with zCall.

## 2.2 Chip Design and Data



**Figure 5. Illustration of the design of Illumina microarray. A target sequence is bound to the matched oligonucleotide bead and its fluorescent end fluoresces red/green light to signal the hybridization of corresponding allele.**

The Illumina Infinium SNP genotyping array consists of hundreds of thousands of beads that are clustered into sets called ‘beadpools’. Each beadpool consists of beads that are manufactured at the same time and physically located at similar positions on the microarray [58]. Each bead is covered with hundreds of thousands of copies of specific oligonucleotide that act as the capture sequences in one of assays [59]. The oligonucleotide sequence has a fluorescent end which fluoresces when the sequence binds to the appropriate target sequence. The degree of fluorescence yields a pixel intensity measuring the degree of hybridization to each of the alleles. Two color single base extension (SBE) chemistry [60] is used on each bead which enables it to assay two alleles (Figure 5).

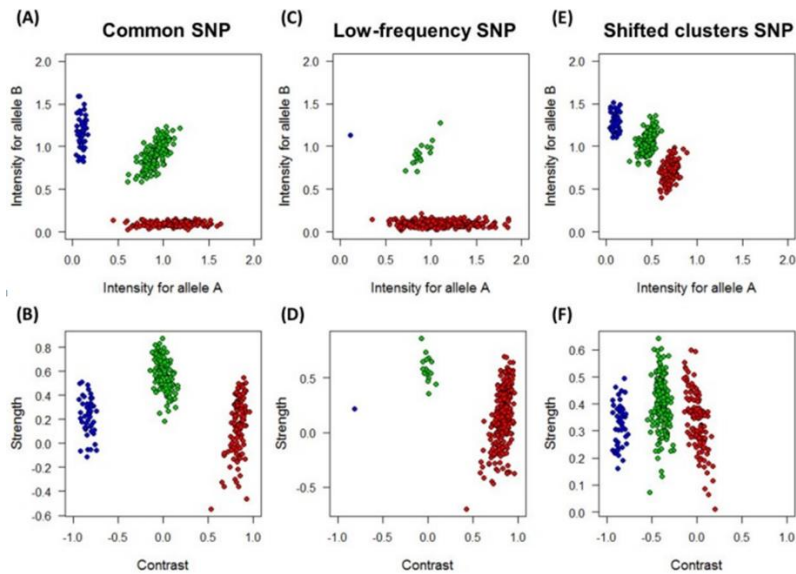
A chip is assayed for each individual and the genotypes are determined using genotype calling procedures based on the observed fluorescent intensities of every SNP on the chip. Because the vast majority of SNPs are biallelic [61], this process has predominantly been applied to probes that query two possible allelic outcomes at a genomic variant (generically defined as allele A and allele B). Translating both sets of allelic hybridization intensities thus allows

discrete decisions to be made with respect to whether the genotype of a sample at a particular SNP is  $AA$ ,  $AB$  or  $BB$ .

Different platforms rely on different technologies and produce raw allele intensities in different dimensions. For example, the Illumina array has on average twenty beads per SNP which generate twenty pairs of hybridization intensities, which before genotype calling, these twenty pairs of intensities are averaged to produce one pair of summary intensity for each SNP [58]. In contrast, early designs of the Affymetrix microarrays generate four-dimensional data at each SNP, namely the intensities of perfect match of allele  $A$ , mismatch of allele  $A$ , perfect match of allele  $B$  and mismatch of allele  $B$ . These four dimensional data are typically reduced to two dimensions with platform specific dimension reduction methods.

In general, microarrays will eventually give a pair of summarized allele-specific intensities  $(x, y)$  for each sample at each SNP, corresponding to allele  $A$  and allele  $B$  respectively. A number of genotype algorithms have been established to process the intensities into genotype calls. Although they model the data differently, in principle, individuals with high  $x$  and low  $y$  are asserted to be genotype  $AA$ , whereas the opposite to be genotype  $BB$ . Individuals with moderate  $x$  and  $y$  are asserted to be genotype  $AB$  (Figure 6A).





**Figure 6. Illustration of hybridization intensity profiles for three different SNPs at both the allelic intensities axes and the transformed contrast scale axes. The three SNPs correspond to (i) a common SNP with  $MAF \geq 5\%$  (panels A and B); (ii) a polymorphic SNP with  $MAF < 5\%$  (panels C and D); (iii) a common SNP with shifted intensity clusters (panels E and F). In each panel, the assigned genotypes are colored accordingly as AA (red), AB (green), and BB (blue). This figure has been adapted from Figure 1 in Zhou et al. (2014) *Bioinformatics* Vol. 30 no. 12 [62].**

Genotype calling algorithms perform the calling based on hybridization intensities either in the original coordinates or translate the intensities into other coordinates. The commonly used transformations are contrast-strength coordinates and log scale coordinates:

$$\text{Contrast-strength coordinates: } \text{contrast} = \frac{x-y}{x+y} ; \text{ strength} = \log(x + y)$$

$$\text{Log scale coordinates: } x' = \log(x + 1) ; y' = \log(y + 1)$$

In practice, the genotypes for the bulk of the SNPs can be accurately determined with straightforward rules that partition the distinctively hybridization intensities. However, SNPs with lower minor allele frequencies or with shifted intensities will not conform to these simple rules and they usually require more sophisticated statistical strategies to accurately call genotypes. The lower allele frequency spectrum of the majority of these SNPs

presents a significantly different challenge where only a small fraction of the samples is heterozygous and there is usually no homozygous cluster for the minor allele (Figure 6C-D). This can thwart algorithms that perform multi-sample calling as these algorithms often set out to locate three genotype clusters. Shifts in the positions of the genotype clusters due to intrinsic hybridization chemistry for a fraction of the SNPs can compound the problem of multi-sample genotype calling (Figure 6E-F). When the emphasis switches to low-frequency spectrum of the genetic variants, an accurate and robust genotype calling algorithm becomes important in new generation genetic study.

### **2.3 Existing Genotype Calling Algorithms**

Statistical algorithms have automated the process of calling genotypes in large-scale microarrays genotyping in which up to five million variants can be assayed simultaneously. Early methods call genotypes based on intensities of multiple redundant probes at a single SNP of a single sample. Newer methods often use the standardized two-dimensional intensities and improve their calling accuracy by incorporating these two types of information:

- a. Information from multiple SNPs for each individual
- b. Information from multiple individuals at the same SNP

Consequently, existing algorithms can be broadly classified into four categories: single-sample single-SNP calling algorithms; multi-sample single-SNP calling algorithms; single-sample multi-SNP calling algorithms and multi-sample multi-SNP calling algorithms.

### *Single-sample single-SNP calling algorithm*

The earliest type of calling algorithms incorporates the raw intensities from different probes and chips at the same SNP for a single person, and aims to reduce the probe and chip effects to the lowest level and model the background noise effectively to reduce false calls.

The Dynamic Modeling (DM) algorithm [63] performs genotype calling based on Affymetrix four-dimensional raw data from a single chip. It assumes two underlying normal distribution (representing foreground and background distribution) for the intensities of every probe quartet for each SNP. The genotype for each probe quartet is selected based on a probe-level log likelihood and the final genotype for each SNP is subsequently determined by a non-parametric test that compares the p-values of four genotype models (AA, AB, BB, NULL).

GEL [64] uses information on multiple chips. It utilizes DM calls as preliminary genotype calls to obtain an empirical distribution of the transformed two-dimensional intensities of each genotype. A genotype of each SNP can be subsequently assigned by Bayes rule.

RLMM [65] is also a multi-chip model. In contrast to DM and GEL, RLMM is a supervised learning algorithm. It fits a linear model for each allele of each SNP to extract the chip effect and the probe effect, and then it derives a discriminant function based on the Mahalanobis distance with parameters trained by well-defined genotype groups and assigns genotype calls to new

data. RLMM has been criticized for its dependence on the availability of training data, which is not often available or appropriate.

BRLMM [66] is similar to RLMM except its use of DM calls to initialize the algorithm and an additional Bayesian step which introduces a prior distribution to each of the parameters, thus removing the need by RLMM for prior training data to initialize the parameters. However, the reliance on DM calls has been found to introduce serious errors and systematic biases in the recalibration process.

DM, GEL, RLMM and BRLMM were developed for the Affymetrix platform. They were early methods and exhibited poorer performance than methods developed subsequently. We will thus not discuss their methodologies.

*Multiple-sample single-SNP calling algorithm (population based)*

Population strategies jointly consider the intensity measurements at each SNP across multiple samples in a cluster analysis framework to learn about genotype cluster characteristics before making the calls. It has been shown that pooling information across multiple individuals can improve the calling quality. In addition, population based methods could effectively genotype for thousands of individuals simultaneously.

GenCall is in the proprietary software of BeadStudio and GenomeStudio.

According to one of Illumina's technical reports, GenCall analyses DNA from a population of several individuals by a set of multiplexed arrays. A custom

clustering algorithm that incorporates several biological heuristics models the behavior of each locus. In cases where fewer than three clusters are observed, locations and shapes of the missing clusters are estimated using neural networks. The genotype call is asserted as the one having the best performance based on a Bayesian procedure [67].

Illuminus [68] is an unsupervised clustering method based on a mixture model of t-distributions which is fitted to the strength and contrast of each SNP through an Expectation Maximization (EM) algorithm [4]. Genotype calls are assigned by choosing the class that has the highest posterior probability and the probability serves as a call confidence measure. Illuminus is also developed for Illumina platform. A limitation of both GenCall and Illuminus is the higher error rates in genotype calling when the minor allele frequency is low.

#### *Single-sample multiple-SNP calling algorithm (SNP based)*

With the rapid development of microarray technology, increasing number of SNPs can be genotyped in one assay. It is cheaper and easier to assay thousands of SNPs of a single person than thousands of individuals at a single SNP. SNP-based strategy has become another choice of the genotype calling strategy.

GenoSNP [58] is a SNP-based strategy which clusters the intensities for all SNPs within a single individual by fitting a Bayesian hierarchical model to the logarithm transformed intensity ( $\log_2(x + 1), \log_2(y + 1)$ ) through the EM

algorithm. The genotype is asserted as the one that has the maximum posterior probability.

SNP-based methods would be desirable if each probe on a genotyping array had a similar response characteristic regardless of which genomic region was being queried. However, it has been questioned whether the SNPs have similar patterns and the within-cluster variation is less than that between clusters [58]. Compared with population-based algorithms, GenoSNP has more SNPs that fail the test for Hardy-Weinberg Equilibrium, which indicates the violation of the assumption that the behavior of all SNPs is similarly across the whole genome.

#### *Multiple-sample multiple-SNP calling algorithm*

This type of algorithm attempts to jointly evaluate both the population-based and SNP-based information to reduce the false calling rate.

The Modified Mixture Model (M3) [69] defines a two-stage calling procedure. In the first stage, it utilizes a typical population-based Bayesian model to call genotype preliminarily. Then it defines an average posterior rate (APR) based on the posterior probability obtained in the first stage to measure the calling quality. Those SNPs with low MAF and poor APR are chosen to be recalled in the second stage, in a manner very similar to GenoSNP, to perform calling across multiple SNPs. A reference SNP is used for each poorly called SNP to assist the recalling process in the second stage. The reference SNP has good

SNP quality, good clustering properties and similar pattern with the testing SNP.

MAMS [70] was designed for the Affymetrix platform. First, it utilizes a typical Bayesian model to call genotypes of single-array multi-SNP data (SAMS call). Second, it applies agglomerative hierarchical clustering to call genotypes of multi-array single-SNP data (MASS call). Finally, it calculates the silhouette width for both SAMS call and MASS call and asserts the final genotype by comparing their silhouette width scores.

A genotype calling algorithm, optiCall [67], is specifically developed for the Illumina platform. It uses multi-SNP multi-sample data to construct a prior distribution and call genotypes within each SNP with a Bayesian hierarchical model. Subsequently, it performs a chi-square HWE test and applies Illuminus to reassign the genotypes for SNPs that are not in HWE. The optiCall algorithm improved the calling accuracy on low frequency and rare SNPs to some degree.

Another Illumina specific platform to improve the calling accuracy on rare SNPs is zCall [71]. It uses two intensity thresholds to separate data of each SNP into genotype AA, AB, BB and NULL. Based on the genotypes obtained from a default genotype caller, zCall applies a linear regression analysis to the mean intensities of genotype AA and BB as well as standard deviations of intensities of genotype AA and BB for common SNPs. Rare SNPs are

genotyped according to the major allele threshold determined by the default call and the minor allele threshold determined by the regression model.

We will provide a more detailed review of Illuminus, GenoSNP, optiCall and zCall in the following sections. Since GenCall is a proprietary method of Illumina, M3 is only available in Matlab and MAMS is designed for the Affymetrix platform, we will not review their methodologies in details. Subsequently, we will introduce our novel method, iCall, and compare it with GenCall, Illuminus, optiCall, GenoSNP and zCall.

### 2.3.1 Illuminus (single-SNP multiple-sample calling algorithm)

The Illuminus algorithm uses the normalized hybridization intensities for the respective two alleles at each SNP that are generated from the proprietary software GenomeStudio as the input, and transforms the intensity signals into contrast-strength scale. Contrast and strength of sample  $j$  at SNP  $l$  is denoted as  $(c_{jl}, s_{jl})$ . Because Illuminus is a population-based method, for the sake of brevity, I leave out the SNP label  $l$  in the following text in this section.

Illuminus fits a three-component bivariate mixture model for  $\mathbf{X}_j = (c_j, s_j)$  using multivariate truncated  $t$  distributions, where the three components correspond to the genotype classes of  $AA$ ,  $AB$  and  $BB$ . Let  $f(\mathbf{x}; \mathbf{M}, \mathbf{\Sigma}, \nu)$  represent the probability density at  $\mathbf{x}$  of a  $t$  distribution with location parameter  $\mathbf{M}$ , variance-covariance matrix  $\mathbf{\Sigma}$  and degree of freedom  $\nu$ . The density for  $\mathbf{X}_j$  can be written as



$$F(\mathbf{X}_j) = \sum_{k=1}^3 \lambda_k \phi_k(\mathbf{X}_j; \mathbf{M}_k, \boldsymbol{\Sigma}_k, \nu_k)$$

where  $(\lambda_1, \lambda_2, \lambda_3)$  are mixture proportions following HWE and

$$\phi_1(\mathbf{X}_j; \mathbf{M}_1, \boldsymbol{\Sigma}_1, \nu_1) = \frac{f(\mathbf{X}_j; \mathbf{M}_1, \boldsymbol{\Sigma}_1, \nu_1)}{1 - \int_{-\infty}^{-1} f(\mathbf{X}_j; \mathbf{M}_1, \boldsymbol{\Sigma}_1, \nu_1) dc}$$

$$\phi_2(\mathbf{X}_j; \mathbf{M}_2, \boldsymbol{\Sigma}_2, \nu_2) = \frac{f(\mathbf{X}_j; \mathbf{M}_2, \boldsymbol{\Sigma}_2, \nu_2)}{\int_{-1}^1 f(\mathbf{X}_j; \mathbf{M}_2, \boldsymbol{\Sigma}_2, \nu_2) dc}$$

$$\phi_3(\mathbf{X}_j; \mathbf{M}_3, \boldsymbol{\Sigma}_3, \nu_3) = \frac{f(\mathbf{X}_j; \mathbf{M}_3, \boldsymbol{\Sigma}_3, \nu_3)}{1 - \int_1^{\infty} f(\mathbf{X}_j; \mathbf{M}_3, \boldsymbol{\Sigma}_3, \nu_3) dc}$$

Illuminus introduced a fourth bivariate Gaussian component with zero mean and large variances as a background distribution of intensity serving for outliers whose intensity profile is not clear to be classified as any of the three genotype clusters.

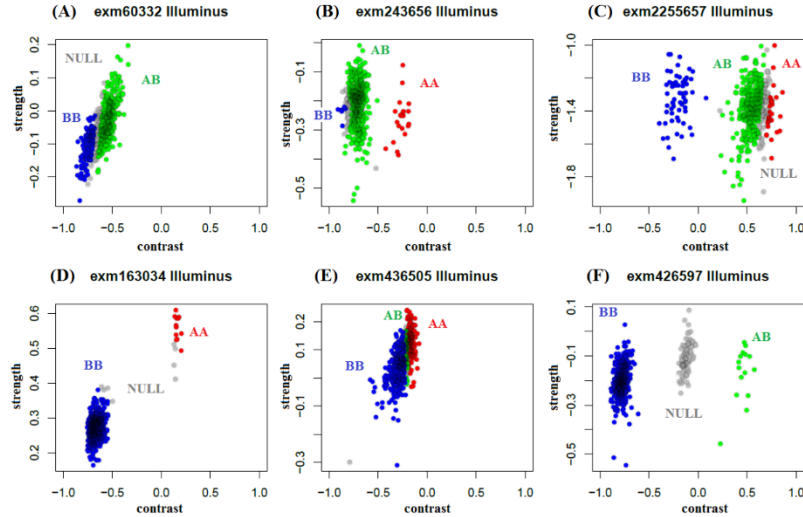
$$\phi_4(\mathbf{X}_j; \mathbf{M}_4, \boldsymbol{\Sigma}_4) = N(\mathbf{X}_j; \mathbf{M}_4, \boldsymbol{\Sigma}_4)$$

where  $\mathbf{M}_4 = (0, 0)$ ,  $\boldsymbol{\Sigma}_4 = \begin{bmatrix} 100000 & 0 \\ 0 & 100000 \end{bmatrix}$ .

The parameters  $(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3, \lambda_1, \lambda_2, \lambda_3)$  are calibrated by the EM algorithm and the genotype is assigned if its posterior probability exceeds 0.95. A good set of initial starts plays an important role in the EM algorithm. Poor initial starts may result in poor classification. To deal with this problem, Illuminus provides five guided starts, from which the initialization of the location parameters is chosen:

$$\begin{bmatrix} -0.9 & 0 & 0.9 \\ -0.9 & -0.5 & 0.9 \\ -0.9 & 0.5 & 0.9 \\ -0.9 & 0.5(\max(c) + \min(c)) & 0.9\max(c) \\ 0.9\min(c) & 0.5(\max(c) + \min(c)) & 0.9 \end{bmatrix}$$

where  $c$  denotes contrasts at the specific SNP. Selecting best initialization starts makes Illuminus more robust to intensity location shifts.



**Figure 7. Illustration of the erroneous calling made by Illuminus. Illuminus is not robust for low frequency and rare variants since it tends to cluster intensities in more clusters than observed.**

Illuminus performs well for SNPs with  $MAF > 1\%$ . However, it presents many problems when dealing with rare variants (Figure 7). One problem originates from the dynamic temping to choose the initial starts with highest likelihood function. Consequently, it has a preference of classifying samples into three clusters with small variance to achieve high likelihood score although sometimes the data present fewer clusters (Figure 7A-C, E). Consider a rare variant ( $MAF < 1\%$ ), the sample sizes and intensity patterns of the three genotype clusters differ significantly. When the population size is small (hundreds), it is possible to observe only two or only one genotype cluster at a rare SNP. Another problem originates from the set of the five guide starts,

which fails to capture some patterns of intensity profiles, and results in erroneous calls (Figure 7 D, F).

### 2.3.2 GenoSNP (multiple-SNP single-sample calling algorithm)

GenoSNP uses the raw hybridization intensities that are generated from the proprietary software GenomeStudio as the input and transforms the intensities into logarithm coordinates. The transformed intensity of sample  $j$  at SNP  $l$  is  $(\log_2(x_{jl} + 1), \log_2(y_{jl} + 1))$ . Since GenoSNP calls genotype across SNPs sample by sample, for the sake of brevity, I will leave out the sample label  $j$  in the following text in this session.

Let  $\mathbf{X}_l = (\log_2(x_l + 1), \log_2(y_l + 1))$  be the pair of transformed intensities for the  $l$ th SNP;  $g_l \in \{1,2,3,4\}$  represent the genotype {AA, AB, BB, NULL} at SNP  $l$ . GenoSNP uses a four-component Gaussian mixture model to fit the data and calls are obtained by finding the genotype with the maximum probability.

$$p(g_l; \boldsymbol{\theta}) = \prod_{k=1}^4 \lambda_k^{I(g_l=k)}$$

$$p(\mathbf{X}_l; g_l, u_l, \boldsymbol{\theta}) = \prod_{k=1}^4 N(\mathbf{X}_l; \mathbf{M}_k, u_l \boldsymbol{\Lambda}_k)^{I(g_l=k)}$$

where  $\{\lambda_k, k = 1,2,3,4\}$  represent the mixture proportions;  $\{\mathbf{M}_k, k = 1,2,3,4\}$  and  $\{\boldsymbol{\Lambda}_k, k = 1,2,3,4\}$  represent the means and scale correlation matrix of Gaussian distributions;  $u_l$  represents a scaling parameter at SNP  $l$ . The parameters are further modelled by a hierarchical model given as follows.

$$p(\boldsymbol{\lambda}|\boldsymbol{\kappa}) \propto \prod_{k=1}^4 \lambda_k^{\kappa_0-1}$$

$$p(u_l; g_l, \boldsymbol{\theta}) = \prod_{k=1}^4 G(u_l; \nu_k/2, \nu_k/2)^{I(g_l=k)}$$

$$p(\mathbf{M}_k, \boldsymbol{\Lambda}_k) = N(\mathbf{M}_k; \mathbf{m}_0, \eta_0 \boldsymbol{\Lambda}_k) W(\boldsymbol{\Lambda}_k | \gamma_0, \mathbf{S}_0)$$

where  $\{\lambda_k\}$  follows a Dirichlet distribution,  $G(x; \alpha, \beta)$  represents the pdf of a gamma distribution with shape parameter  $\alpha$  and scale parameter  $\beta$ ;

$N(\mathbf{x}; \mathbf{M}, \boldsymbol{\Lambda})$  represents the pdf of a bi-variate Gaussian distribution with mean

$\mathbf{M}$  and covariance matrix  $\boldsymbol{\Lambda}$ ;  $W(\boldsymbol{\Lambda}; \gamma, \mathbf{S})$  represents the pdf of a Wishart

distribution with degree of freedom  $\gamma$  and scale matrix  $\mathbf{S}$ ;  $\nu_k$  are fixed at 4;

$\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \mathbf{M}, \boldsymbol{\Lambda}\}$ ; hyper-parameters  $\kappa_0=1.1, \eta_0=1, \gamma_0=1, \mathbf{S}_0 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$ ,

$\mathbf{m}_0=[(9,6), (8,8), (6,9), (6,6)]$  for  $k=1,2,3,4$  respectively.

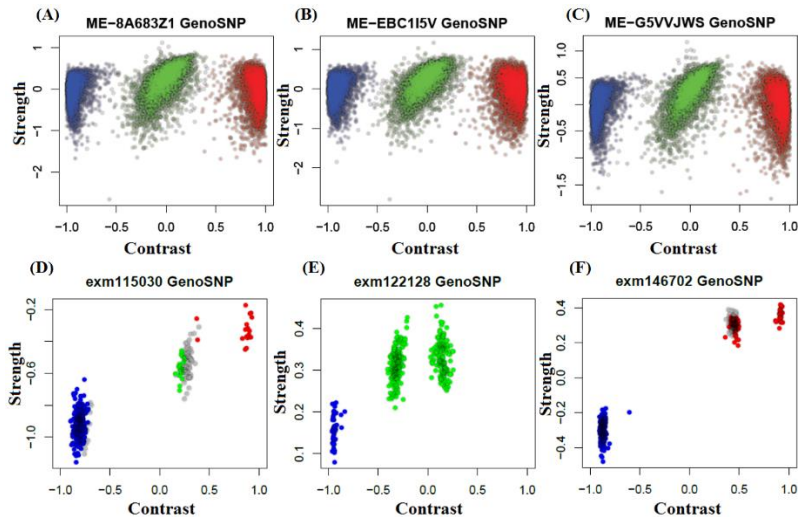
Instead of implementing a standard EM algorithm, GenoSNP uses a

Variational Bayes EM algorithm (VB-EM) to perform the optimization, which

was proved to be more robust than standard EM algorithm, and the final

genotypes are asserted to be the genotype that maximizes the variational

approximation probability.



**Figure 8. Illustration of the erroneous calling of GenoSNP. The three panels in the first row illustrate the calling for single sample across multiple SNPs. The three panels in the second row illustrate the calling for single SNP across multiple samples.**

GenoSNP has been criticized for its assumption that SNPs across the genome have similar intensity patterns. It performs poorly when the intensity clouds deviate from their expected locations. As shown in Figure 8, the intensity profile across multiple SNPs is similar and stable for different individuals (Figure 8A-C). The clustering based on cross SNPs intensity is generally correct. However, many erroneous calls are made when there is location shift (Figure 8D-F).

### 2.3.3 optiCall (multiple-SNP multiple-sample calling algorithm)

A Bayesian hierarchical model to the normalized hybridization intensities is fit using optiCall. A prior distribution is fitted to cross-sample cross-SNP intensities and genotypes are called within each SNP as which has the highest posterior probability.

*STEP 1: Create across sample cross SNP prior*

Let  $\mathbf{X}_{jl} = (x_{jl}, y_{jl})$  represent the normalized intensities of sample  $j$  at SNP  $l$ .

The optiCall algorithm takes a random subset  $S$  of intensity values from the dataset, which contains intensities across SNPs and across samples. Similar to Illuminus, a four-component Student's  $t$  mixture model is fitted to the subset of intensities. Let  $\mathbf{X}_{jl}$  represent a sample in  $S$  and  $g_{jl} \in \{1,2,3,4\}$  represent its genotype. The joint pdf of  $(\mathbf{X}_{jl}, g_{jl})$  is given by

$$F(\mathbf{X}_{jl}, g_{jl}; \mathbf{M}_k, \mathbf{\Sigma}_k, \nu_k) = \prod_{k=1}^4 [\lambda_k f_k(\mathbf{X}_{jl}; \mathbf{M}_k, \mathbf{\Sigma}_k, \nu_k)]^{I(g_{jl}=k)}$$

where  $f(\mathbf{x}; \mathbf{M}, \mathbf{\Sigma}, \nu)$  denotes the density function for data  $\mathbf{x}$  at a Student's  $t$  distribution with location parameter  $\mathbf{M}$ , variance-covariance matrix  $\mathbf{\Sigma}$  and degree of freedom  $\nu$ .  $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$  are the mixture proportion of the four classes that need not follow HWE. Degree of freedom  $\nu$  for all classes is set to 1. The parameters for NULL class are fixed to  $\mathbf{M}_4 = (0, 0), \mathbf{\Sigma}_4 = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}$ . The EM algorithm is applied to fit the model and infer the parameters  $\{\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3, \lambda_1, \lambda_2, \lambda_3\}$ .

*STEP 2: Genotype calls across samples with prior information across SNPs*

The optiCall algorithm clusters intensities with another four-component mixture Student's  $t$  model for each SNP separately. The parameters of the  $t$ -distributions have Normal-inverse-Wishart prior determined in STEP 1.

$$F(\mathbf{X}_j, g_j; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k, \nu_k) = \prod_{k=1}^4 [\pi_k f_k(\mathbf{X}_j; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k, \nu_k)]^{I(g_j=k)}$$

$$p(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) \propto N\left(\boldsymbol{\mu}_k; \boldsymbol{\alpha}_k, \frac{\boldsymbol{\sigma}_k}{\beta_k}\right) W(\boldsymbol{\sigma}_k^{-1}; \gamma_k, \mathbf{S}_k), k = 1, 2, 3$$

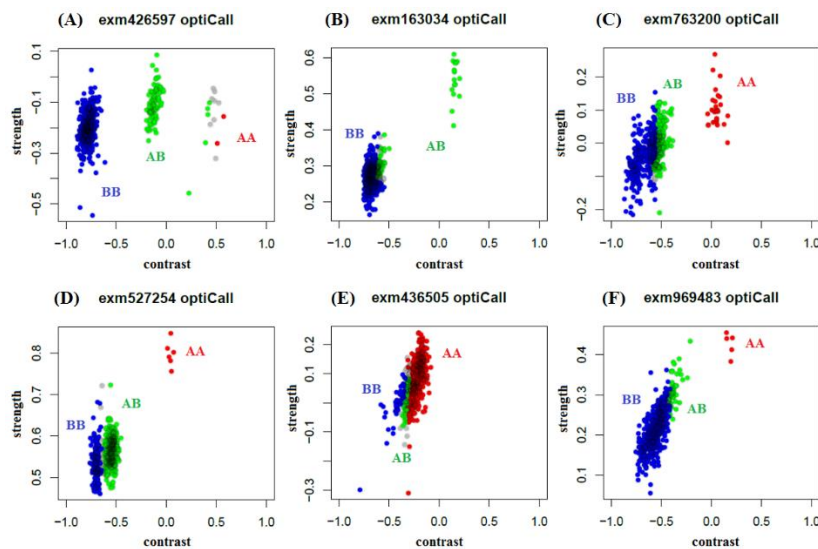
$$\boldsymbol{\mu}_4 = (0, 0), \mathbf{\Sigma}_4 = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}$$

where  $v_1 = v_3 = 1, v_2 = 1.3$ .  $\alpha_k$  are set to  $M_k$  obtained in STEP 1,  $S_k$  are set to be the inverse of  $\Sigma_k$  obtained in STEP 1,  $\beta_k = 1$  and  $\gamma_k = 100, k = 1,2,3$ .

The EM algorithm is applied to calibrate the parameters and genotype with maximum posterior probability will be assigned if its posterior is above 0.9.

### STEP 3: Rescue

The optiCall method uses the p-value of HWE chi-square test as a measure of clustering quality and uses Illuminus algorithm to reclassify SNPs that have poor clustering qualities.

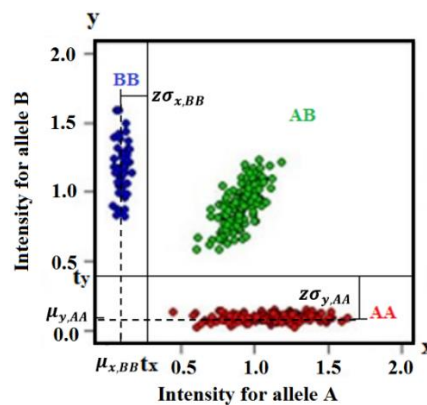


**Figure 9. Illustration of erroneous calling made by optiCall.**

optiCall achieves higher robustness for rare variants than Illuminus since its prior distributions effectively locate each genotype, especially the minor allele. Nevertheless, it performs poorly for the SNPs whose location of intensity clouds shifts (Figure 9). The prior distribution obtained by multi-SNP multi-sample intensity may not be appropriate for all SNPs. Heterozygotes' mean often has a larger variability and heterozygotes' covariance matrix is often larger than that of homozygotes. This characteristic may strongly affect

the calling procedure and lead to erroneous calls (Figure 9A-B). The creators of optiCall are aware that the prior distribution would cause problem for those SNPs with location shift, hence use Illuminus as a rescue process. However, this will reintroduce the shortcoming of Illuminus – its low accuracy for rare variants – for those SNPs being rescued (Figure 9C-F).

#### 2.3.4 zCall (multiple-SNP multiple-sample calling algorithm)



**Figure 10. Illustration of the algorithm of zCall. Two intensity thresholds  $t_x$  and  $t_y$  are used to cluster the intensities into genotype classes.**

In recognition of the challenges associated with calling the genotypes for rare SNPs, zCall was introduced to post-process the genotype calls from a default calling algorithm such as GenCall [71]. This relied on calibrating the positions of the other two genotype clusters on the basis of the dominant homozygous cluster to improve the accuracy and call rate (Figure 10). The input of zCall is the Illumina normalized intensity  $\mathbf{X}_{jl} = (x_{jl}, y_{jl})$  and the genotype calls made by a default population-based genotype caller  $\{g_{jl}\}$ .

##### *STEP 1: linear regression model*

zCall picks out all SNPs with  $\text{MAF} \geq 5\%$  based on the default call, and then uses linear regression to analyze the relation between mean value of intensities



of allele A and allele B of the homozygotes BB and AA respectively and the relation between standard deviation of intensities of allele A and allele B of homozygotes BB and AA respectively ( $\mu_{Y,AA} \sim \mu_{X,BB}$ ;  $\sigma_{Y,AA} \sim \sigma_{X,BB}$ ;  $\mu_{X,BB} \sim \mu_{Y,AA}$ ;  $\sigma_{X,BB} \sim \sigma_{Y,AA}$ ).

*STEP 2: recall rare variants*

At a rare variant, the mean and standard deviation of major homozygotes is well defined by the default call. The mean and standard deviation of minor homozygotes can be determined by the linear model obtained from STEP 1. The genotype clusters can be determined by a vertical ( $x = t_x$ ) and horizontal ( $y = t_y$ ) line, where  $t_x = \mu_{x,BB} + 7 \cdot \sigma_{x,BB}$ ;  $t_y = \mu_{y,AA} + 7 \cdot \sigma_{y,AA}$ .

**2.4 Method**

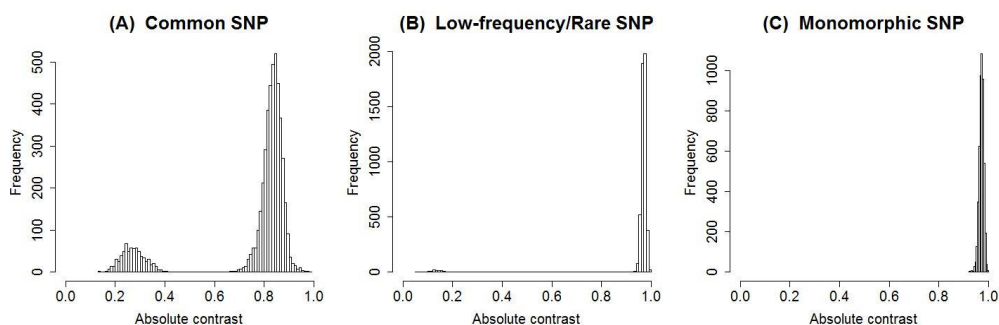
We introduce a new genotype calling strategy for Illumina arrays, iCall, which performs multi-sample calling at a single SNP to improve accuracy across the full allele frequency spectrum. This algorithm adopts the classical three-component student's  $t$ -mixture model framework that Illuminus adopts, but focuses on deriving appropriate penalties to find the best seeding parameters to initialize the EM procedure to recognize the variety of situations where calling becomes difficult, such as when (i) the MAF is low; (ii) the total number of samples for joint calling is small; or (iii) the hybridization intensities deviate substantially from usual. iCall is implemented in C++ for use on Linux operating systems and is available for download at

<http://www.statgen.nus.edu.sg/~software/icall.html>

Similar to Illuminus and GenCall, iCall is a population-based algorithm. iCall also uses the normalized hybridization intensities for the respective two alleles at each SNP that is generated from the proprietary software GenomeStudio as the input. We generally define the two alleles as A and B, and let  $(x_j, y_j)$  denote the normalized intensities for sample  $j$  at a specific SNP. The iCall algorithm transforms the normalized intensities to the contrast-strength coordinate system  $(c_j, s_j)$ .

#### 2.4.1 Identifying the parameters to initialize calling

The performance of the genotype calling can depend crucially on the set of initial calls used to seed the algorithm, especially if the mathematical framework for initializing the calls is similar to the framework for subsequent calling. For instance, if the initial set of calls already assumes the presence of only one genotype cluster, subsequent iterations of a calling algorithm will usually remain within the same domain space unless the empirical data provides a strong motivation to introduce additional genotype clusters.



**Figure 11. Histograms of the absolute value of the contrast coordinates for 12 370 samples at three SNPs with different MAFs, corresponding to a (A) common SNP ( $MAF \geq 5\%$ ); (B) low-frequency or rare SNP ( $0\% < MAF < 5\%$ ); and (C) monomorphic SNP ( $MAF = 0\%$ ). This figure has been adapted from Figure 2 in Zhou et al. (2014) *Bioinformatics* Vol. 30 no. 12 [62].**

iCall adopts a framework to generate the initial set of calls, by considering the information presented by the absolute contrast measurements (or  $|c_j|$ ). When

considered across multiple samples, the density profile of the absolute contrast can inform the potential locations of each genotype clusters (Figure 11). A common SNP will usually yield a density profile with two distinct peaks (around 0 and 1 for  $|c_j|$  respectively), while a rare or low-frequency SNP will give a profile with a small peak near 0 and a significantly larger peak around 1, and a monomorphic SNP will yield only one peak around 1. To model this, we consider two scenarios: (i) the first assumes a normal distribution for  $|c_j| \sim Normal(\mu, \sigma^2)$ , and this aims to capture the situation when the SNP is monomorphic; (ii) the second aims to identify the situation for a non-monomorphic SNP and assumes a two-component normal mixture model for  $|c_j|$  such that  $|c_j| \sim p \cdot Normal(\mu_1, \sigma_1^2) + (1 - p) \cdot Normal(\mu_2, \sigma_2^2)$  with  $0 \leq \mu_1 < \mu_2 \leq 1$  and all the parameters  $(p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$  are estimated from the data within an EM algorithm framework. The first scenario is actually a special case of the second scenario where  $p = \mu_1 = \sigma_1^2 = 0$ .

In order to identify which scenario is more appropriate for the observed data, we defined a penalized log likelihood score. We assign a penalty term on small values of  $(\mu_2 - \mu_1)$  for both scenarios (in scenario,  $\mu_1 = 0, \mu_2 = \mu$ ). Without this penalty term, the two-component mixture model will always yield a higher log-likelihood due to the better fit of the data into two normal distributions with smaller variances.

The penalized log-likelihood functions are calculated as

$$\sum_j \log \left( \phi(|c_j|; \mu, \sigma^2) \right) + n \cdot S(\mu), \quad \text{for scenario 1}$$

$$\sum_j \left\{ \log \left( \phi \left( |c_j|_{j \in \text{class}_1}; \mu_1, \sigma_1^2 \right) \right) + \log \left( \phi \left( |c_j|_{j \in \text{class}_2}; \mu_2, \sigma_2^2 \right) \right) \right\} + n \cdot$$

$S(\mu_2 - \mu_1)$ , for scenario 2

where the penalty term  $S(x) = \log \left( \frac{\psi(x | \text{meanlog}=0.4, \text{variancelog}=0.4)}{\int_0^1 \psi(y | \text{meanlog}=0.4, \text{variancelog}=0.4) dy} \right)$ ,  $n$

represents the number of samples used for the joint calling,  $\psi(\cdot)$  is the pdf of a lognormal distribution with mean and variance of the distribution on the log scale equal to meanlog and variancelog, and  $\phi(\cdot)$  is the density function of a normal distribution. The intuition here is when the values for  $\mu_1$  and  $\mu_2$  are not significantly different, the calling algorithm prefers to combine the two components instead of forcing the presence of two clusters.

The scenario with the higher log-likelihood is chosen to generate eight sets of location parameters to initialize the genotype calling in a three-component univariate Gaussian mixture model for  $c_j$ , where the eight sets are

$$\begin{bmatrix} -\mu & 0 & \mu \\ -1 & 0 & \mu \\ -\mu & 0 & 1 \\ -\mu & \mu & 1 \\ -1 & -\mu & \mu \\ -1 & -0.8 & \mu \\ -\mu & 0.8 & 1 \\ t_1 & \frac{t_1+t_2}{2} & t_2 \end{bmatrix}$$

if scenario 1 yields the higher log-likelihood, or

$$\begin{bmatrix} -\mu_2 & 0 & \mu_2 \\ -1 & 0 & \mu_2 \\ -\mu_2 & 0 & 1 \\ -\mu_2 & -\mu_1 & \mu_2 \\ -\mu_2 & \mu_1 & \mu_2 \\ -\mu_2 & -\mu_1 & \mu_1 \\ -\mu_1 & \mu_1 & \mu_2 \\ t_1 & \frac{t_1+t_2}{2} & t_2 \end{bmatrix}$$

if scenario 2 yields the higher log-likelihood, and  $t_1$  and  $t_2$  are chosen from the trimmed empirical distribution of  $c_j$  as

$$t_1 = \frac{Q_c^{0.999} + Q_c^{0.001}}{2} - 0.95 \times \frac{Q_c^{0.999} - Q_c^{0.001}}{2}$$

and

$$t_2 = \frac{Q_c^{0.999} + Q_c^{0.001}}{2} + 0.95 \times \frac{Q_c^{0.999} - Q_c^{0.001}}{2}$$

where  $Q_c^x$  denotes the 100x quantile value of the distribution of the empirical contrast values. This allows the initialization parameters to be guided by the observed contrast values, which is particularly useful in the situation where the intensities for the genotype clusters are shifted significantly.

#### 2.4.2 Initializing the genotype calling

iCall uses the same calling structure as Illuminus where the latent genotype variable  $g_j \in \{1,2,3,4\}$ . The density function of  $\mathbf{X}_j = (c_j, s_j)$ , under the three-component bivariate truncated  $t$  mixture model, is given by

$$F(\mathbf{X}_j) = \sum_{k=1}^3 \lambda_k \psi_k(\mathbf{X}_j | \mathbf{M}_k, \Sigma_k, \nu_k)$$

where

$$\psi_1(\mathbf{X}_j | \mathbf{M}_1, \Sigma_1, \nu_1) = \frac{f(\mathbf{X}_j | \mathbf{M}_1, \Sigma_1, \nu_1)}{1 - \int_{-\infty}^{-1} f(\mathbf{X}_j | \mathbf{M}_1, \Sigma_1, \nu_1) dc}$$

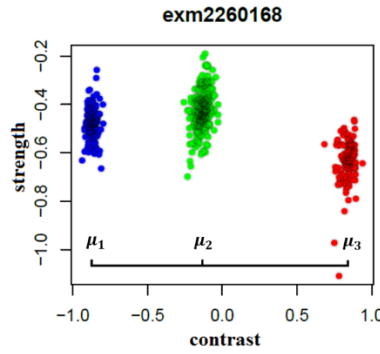
$$\psi_2(\mathbf{X}_j | \mathbf{M}_2, \Sigma_2, \nu_2) = \frac{f(\mathbf{X}_j | \mathbf{M}_2, \Sigma_2, \nu_2)}{\int_{-1}^1 f(\mathbf{X}_j | \mathbf{M}_2, \Sigma_2, \nu_2) dc}$$

$$\psi_3(\mathbf{X}_j | \mathbf{M}_3, \Sigma_3, \nu_3) = \frac{f(\mathbf{X}_j | \mathbf{M}_3, \Sigma_3, \nu_3)}{1 - \int_1^{\infty} f(\mathbf{X}_j | \mathbf{M}_3, \Sigma_3, \nu_3) dc}$$

with  $f(\mathbf{X}_j | \mathbf{M}_k, \Sigma_k, \nu_k)$  representing the density function of a bivariate  $t$  distribution with location parameter  $\mathbf{M}_k$ , variance-covariance matrix  $\Sigma_k$  at  $\nu_k$

degrees of freedom, and  $\{\lambda_1, \lambda_2, \lambda_3\}$  representing the proportion of each genotype that follows the HWE.

Genotype variable  $\{g_j\}$  is assigned to the one whose posterior probability is higher than a threshold (default threshold is 0.8 in iCall). The parameters  $\{\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3, \lambda_1, \lambda_2, \lambda_3\}$  and the latent genotype  $\{g_j\}$  are updated by the EM algorithm.



**Figure 12. The first penalty term penalizes on small distances between the heterozygous cluster and the two homozygous clusters.**

To initializing this EM procedure, iCall uses a three-component univariate truncated Gaussian mixture model with equal weights for the contrast measurements to generate the first iteration of latent genotype  $\{g_j^{(1)}\}$ . The joint density probability of  $(c_j, g_j)$  is given by

$$F(c_j, g_j | \mu^{(1)}, \mu^{(2)}, \mu^{(3)}, \sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)}) = \sum_{k=1}^3 [h_k(c_j | \mu^{(k)}, \sigma^{(k)})]^{I(g_j=k)}$$

where

$$h_1(c_j | \mu^{(1)}, \sigma^{(1)}) = \frac{\phi(c_j; \mu^{(1)}, \sigma^{(1)})}{1 - \int_{-\infty}^{-1} \phi(y; \mu^{(1)}, \sigma^{(1)}) dy}$$

$$h_2(c_j | \mu^{(2)}, \sigma^{(2)}) = \frac{\phi(c_j; \mu^{(2)}, \sigma^{(2)})}{\int_{-1}^1 \phi(y; \mu^{(2)}, \sigma^{(2)}) dy}$$

$$h_3(c_j | \mu^{(3)}, \sigma^{(3)}) = \frac{\phi(c_j; \mu^{(3)}, \sigma^{(3)})}{1 - \int_1^{\infty} \phi(y; \mu^{(3)}, \sigma^{(3)}) dy}$$

with  $\phi(\cdot)$  representing the density function of a univariate normal distribution. Each of the eight sets of location parameters is used as  $(\mu^{(1)}, \mu^{(2)}, \mu^{(3)})$ . The same standard deviation of 0.1 is assumed for the three genotype classes in the first three guided starts, and 0.05 ( $Q_c^{0.999} - Q_c^{0.001}$ ) for the three genotype classes in the other five guided starts. Moreover, two penalized log-likelihood functions were calculated to select two sets of parameters among the eight, which take the form of

$$\begin{aligned} \text{like}_1 &= \sum_{j=1}^n \log \left( F(c_j, g_j | \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3) \right) + n \cdot \frac{S(\mu_2 - \mu_1) + S(\mu_3 - \mu_2)}{2} \\ \text{like}_2 &= \sum_{j=1}^n \log \left( F(c_j, g_j | \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3) \right) + n \cdot \frac{S(\mu_2 - \mu_1) + S(\mu_3 - \mu_2)}{2} - \\ &10 \times \sum_{i=1}^3 \frac{(O_i - E_i)^2}{E_i} \end{aligned}$$

where  $S(\cdot)$  is the same as in Section 2.4.1. Note that  $\{\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3\}$  are empirically updated according to  $\{g_j^{(1)}\}$ , which are often different from  $\{\mu^{(1)}, \mu^{(2)}, \mu^{(3)}, \sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)}\}$ .

The intuition behind the two penalty terms is: the first term,  $n \cdot$

$$\frac{S(\mu_2 - \mu_1) + S(\mu_3 - \mu_2)}{2},$$

penalizes on small distances between the heterozygous cluster and the two homozygous clusters, while the second term,

$$-10 \sum_{i=1}^3 \frac{(O_i - E_i)^2}{E_i},$$

penalizes on genotype call configuration at a SNP that deviates further from the state of HWE. Of the eight guided starts, we identify the two guided starts ( $\text{seed}_1, \text{seed}_2$ ) that yield the highest  $\text{like}_1$  and  $\text{like}_2$  respectively, and these two guided starts are subsequently used to seed the genotype calling. Note that the two sets of seeding start may be identical if the

same guided start yields the highest penalized log-likelihoods in both calculations.

### 2.4.3 Genotype calling

Each of the two sets of seeding starts is used to initialize the three-component bivariate truncated  $t$  mixture model that Illuminus adopts. The calling algorithm adopts an EM framework to yield two sets of genotype call configurations, each initialized from one of the two seeding starts.

The EM procedure used is described as follows. In the M-step, the means, variance-covariance matrixes and mixture proportions are updated by maximizing the log-likelihood function conditional on the assigned genotypes. In the E-step, we do not calculate the Q function (the expected value with regard to the conditional distribution of latent genotypes given the intensities and the parameters). Instead, we assign a genotype to each sample as which has posterior probability exceeds a threshold, 0.8.

M-step:

$$\mathbf{M}_k = \left( \frac{\sum_j c_j \cdot I(g_j = k)}{\sum_j I(g_j = k)}, \frac{\sum_j s_j \cdot I(g_j = k)}{\sum_j I(g_j = k)} \right) \quad k = 1,2,3$$

$$\mathbf{\Sigma}_k = \begin{bmatrix} \text{var}(c_{g_j=k}) & \text{cov}(c_{g_j=k}, s_{g_j=k}) \\ \text{cov}(c_{g_j=k}, s_{g_j=k}) & \text{var}(s_{g_j=k}) \end{bmatrix} \quad k = 1,2,3$$

$$p_A = \frac{\sum_j 2 \times I(g_j = 1) + I(g_j = 2)}{\sum_j 2 \times I(g_j \neq 4)}$$

$$p_B = \frac{\sum_j 2 \times I(g_j = 3) + I(g_j = 2)}{\sum_j 2 \times I(g_j \neq 4)}$$

$$(\lambda_1, \lambda_2, \lambda_3) = (p_A^2, 2p_A p_B, p_B^2)$$



E-step:

$$g_j = \operatorname{argmax}_k \{\lambda_k \psi_k(\mathbf{X}_j | \mathbf{M}_k, \Sigma_k, \nu_k), k = 1, 2, 3\}$$

The default is to accept the genotype call configuration initialized with seed<sub>1</sub>, except when the evidence against HWE is more significant in the configuration generated by seed<sub>1</sub> than the configuration generated by seed<sub>2</sub> and likelihood in the configuration generated by seed<sub>1</sub> is smaller than which is generated by seed<sub>2</sub>, in which case the genotype calls generated with seed<sub>2</sub> are accepted as the final calls. This minimizes the inadvertent miscalling that happens due to shifts in genotype clouds resulting in genotype calls that tend to deviate from HWE.

#### 2.4.4 Chromosomes X, Y and mitochondria

For calling genotypes at SNPs on the mitochondria and the non-pseudo-autosomal regions of the sex chromosomes, the genotype calling additionally require information on the gender of each sample which determines the direction of hybridization inactivation. For SNPs on chromosome X, genotypes for females are determined in the same fashion as autosomal SNPs, while the genotypes for males will only be called as either AA or BB. For SNPs on chromosome Y, NULL calls will be produced for females and the calling only considers the intensity data for male samples and similarly yields genotype calls of either AA or BB. The situation is reversed for SNPs on the mitochondria, where NULL calls will be produced for males and the calling only considers the intensity data for female samples and produces calls of either AA or BB.

## 2.5 Application to Data from Exome Microarray & Method Comparison

The performance of iCall was compared against four single-stage genotype calling algorithms: GenCall, optiCall, Illuminus and GenoSNP. Intensity data were available for 12,370 samples that have been genotyped on the Illumina exome chip, of which 348 samples came from the Singapore Integrative Omics Project (iOmics) and 12,022 samples came from multiple complex disease studies that have been genotyped at a single facility at the Genome Institute of Singapore.

To compare the performance of different genotype calling algorithms, we need to derive a set of gold standard calls that we subsequently assumed to be perfect for benchmarking the genotype calls made by different algorithms.

Whole genome sequencing is a completely different method for calling genotypes, which is regarded as a good comparison. Of the 348 iOmics samples, 81 samples have been additionally whole-genome sequenced to a target coverage of 30-fold as part of the Singapore Sequencing Studies (<http://www.statgen.nus.edu.sg/>), and the sequence calls after quality checks for these samples were regarded as the gold standard calls that were subsequently used to benchmark the performance of the different methods.

A total of 16,428 SNPs were present on the exome chip that overlapped with the polymorphic variants identified from the high-coverage sequencing. These SNPs were classified as common ( $MAF \geq 5\%$ , 13,542 SNPs), low frequency ( $1\% \leq MAF \leq 5\%$ , 1,356 SNPs) and rare ( $MAF \leq 1\%$ , 1,530 SNPs) according to the GenCall genotypes for all 12,370 samples.

In order to evaluate how the number of samples available for joint calling impact the algorithms, we thinned the dataset into four smaller sets with 500, 1000, 3000 and 5000 samples, which always included the 81 samples with gold standard calls. Genotypes are generated by running iCall, optiCall, Illuminus and GenoSNP on the datasets of different size and accuracies are assessed only based on the gold standard subset. Because of the resource limitation, GenCall genotypes were available for the 348 iOmics samples and 12,022 samples independently. Therefore, in comparison, we only provide accuracy of GenCall with sample size of 348. GenoSNP is a single sample caller where the performance is not affected by the size of the available samples

The performance of iCall, optiCall, Illuminus and GenoSNP is evaluated using five metrics: (i) call rate, defined as the percentage of valid genotype calls that are not assigned as NULL; (ii) concordance, defined as the percentage of valid genotype calls that are identical to the gold standard calls; (iii) overall concordance, defined as the percentage of genotype calls out of all possible calls that are identical to the gold standard calls, and is calculated as the product of the call rate and the concordance; (iv) minor allele concordance for rare and low-frequency SNPs, defined as the percentage of the heterozygous and minor allele-homozygous calls that are identical to the gold standard calls out of the total number of such calls made for these SNPs; and (v) missed minor allele call rate, defined as the percentage of the heterozygous and minor allele homozygous calls that are not identified out of the total number of available minor allele calls in the gold standard. The last two metrics

effectively evaluate the true positive and false negative rates for making a genotype call involving at least one minor allele at a low-frequency or rare SNP. The calculations of all five metrics are made using only the 81 samples for which there are gold standard calls available. In the gold standard subset, there are 1,222,885 valid calls (not NULL call) in total, where 14,063 minor allele calls belong to the 1,356 low-frequency SNPs and 6,371 minor allele calls belong to the 1,530 rare SNPs.

On the basis of call rates and concordance with the gold standard calls, iCall yielded the highest overall concordance rate and call rate regardless of the sample size (Table 1). We observed that GenCall yielded the highest concordance rate but tend to be more conservative at making calls, but still managed to deliver an overall concordance rate that was consistently higher than the performance by optiCall. The performance of Illuminus and GenoSNP were comparatively less satisfactory, with GenoSNP yielding an overall concordance rate that was below 97%.

When evaluating the ability to correctly call genotypes carrying at least one copy of the minor allele that is present in the dataset at a frequency <5%, iCall consistently yields the highest accuracy and the lowest missed allele calls compared with GenCall, optiCall, Illuminus and GenoSNP at low-frequency SNPs (Table 1). For example, iCall achieved a minor allele concordance rate of 97.140% and 97.168% at the sample sizes of 500 and 12,370 respectively, compare with compared with optiCall at 96.932 and 97.033%, respectively, and the next-best performing algorithm (GenCall) at 97.083% at the sample

size of 348 (Table 1). At rare SNPs, iCall similarly delivered the highest minor allele concordance rates across all sample sizes considered (at least 97.435%, with all other methods delivering concordance <97%). This suggests that whenever iCall made a call involving a minor allele, it was more likely to be correct than existing algorithms.

**Table 1. Comparison of iCall against optiCall, Illuminus, GenCall and GenoSNP at 16 428 SNPs at different sample sizes for calling, where genotypes from whole-genome sequencing of 81 samples are used as benchmark. This figure has been adapted from Table 1 in Zhou et al. (2014) Bioinformatics Vol. 30 no. 12 [62].**

Sample size	Call rate (%)	Concordance	Overall concordance (%)	Low-frequency SNPs				Rare SNPs			
				Correct minor allele calls	Minor allele calls	Minor allele concordance rate (%)	Missed minor allele call rate (%)	Correct minor allele calls	Minor allele calls	Minor allele concordance rate (%)	Missed minor allele call rate (%)
<b>iCall</b>											
500	99.993	97.683	97.676	13 653	14 055	97.140	2.915	6161	6316	97.546	3.296
1000	99.990	97.683	97.673	13 653	14 055	97.140	2.915	6153	6315	97.435	3.422
3000	99.990	97.685	97.675	13 657	14 054	97.175	2.887	6163	6316	97.578	3.265
5000	99.988	97.685	97.673	13 656	14 055	97.161	2.894	6163	6316	97.578	3.265
12 370	99.986	97.686	97.672	13 658	14 056	97.168	2.880	6162	6317	97.546	3.280
<b>GenCall</b>											
348	99.983	97.688	97.671	13 647	14 057	97.083	2.958	6113	6315	96.801	4.050
<b>optiCall</b>											
500	99.987	97.667	97.654	13 617	14 048	96.932	3.171	6160	6378	96.582	3.312
1000	99.988	97.665	97.653	13 623	14 078	96.768	3.129	6175	6448	95.766	3.076
3000	99.985	97.675	97.660	13 621	14 044	96.988	3.143	6179	6463	95.606	3.014
5000	99.987	97.681	97.668	13 625	14 048	96.989	3.115	6180	6418	96.292	2.998
12 370	99.990	97.681	97.662	13 637	14 054	97.033	3.029	6182	6565	94.166	2.967
<b>Illuminus</b>											
500	99.805	97.652	97.462	13 036	14 203	91.783	7.303	5043	6436	78.356	20.844
1000	99.834	97.650	97.488	13 254	14 067	94.221	5.753	5120	6587	77.729	19.636
3000	99.873	97.645	97.521	13 496	14 052	96.043	4.032	5187	6700	77.418	18.584
5000	99.862	97.651	97.516	13 500	14 048	96.099	4.003	5035	6696	75.194	20.970
12 370	99.848	97.661	97.513	13 563	14 059	96.472	3.555	4772	6542	72.944	25.098
<b>GenoSNP</b>											
Single SNP	99.607	96.734	96.354	13 349	15 267	87.437	5.077	6059	7583	79.902	4.897

Among the 16 428 SNPs considered, 13 542 are common SNPs, 1356 are low-frequency SNPs and 1530 are rare SNPs. Within the gold standard, there are 1 222 885 valid genotype calls in total, which include 14 063 minor allele calls at low-frequency SNPs and 6371 minor allele calls at rare SNPs.

**Table 2. Comparison of iCall+zCall, GenCall+zCall and optiCall+zCall at 16 428 SNPs at different sample sizes for calling, where genotypes from whole-genome sequencing of 81 samples are used as benchmark. This figure has been adapted from Table 2 in Zhou et al. (2014) Bioinformatics Vol. 30 no. 12 [62].**

Sample size	Call rate (%)	Concordance	Overall concordance (%)	Low-frequency SNPs				Rare SNPs			
				Correct minor allele calls	Minor allele calls	Minor allele concordance rate (%)	Missed minor allele call rate (%)	Correct minor allele calls	Minor allele calls	Minor allele concordance rate (%)	Missed minor allele call rate (%)
<b>iCall+zCall</b>											
500	99.999	97.682	97.681	13 653	14 054	97.147	2.915	6167	6315	97.656	3.202
1000	99.998	97.683	97.681	13 654	14 055	97.147	2.908	6167	6315	97.656	3.202
3000	99.998	97.684	97.682	13 660	14 054	97.197	2.866	6167	6315	97.656	3.202
5000	99.998	97.684	97.682	13 658	14 055	97.175	2.880	6167	6315	97.656	3.202
12 370	99.998	97.685	97.683	13 661	14 056	97.190	2.859	6167	6315	97.656	3.202
<b>GenCall+zCall</b>											
500	99.998	97.685	97.684	13 654	14 054	97.154	2.908	6163	6314	97.608	3.265
1000	99.998	97.685	97.683	13 654	14 054	97.154	2.908	6162	6313	97.608	3.280
3000	99.998	97.685	97.683	13 654	14 054	97.154	2.908	6161	6312	97.608	3.296
5000	99.998	97.685	97.683	13 654	14 054	97.154	2.908	6161	6312	97.608	3.296
12 370	99.998	97.685	97.683	13 654	14 054	97.154	2.908	6161	6312	97.608	3.296
<b>optiCall+zCall</b>											
500	99.999	97.660	97.659	13 628	14 040	97.066	3.093	6170	6369	96.875	3.155
1000	99.999	97.659	97.658	13 629	14 065	96.900	3.086	6186	6438	96.086	2.904
3000	99.999	97.672	97.671	13 626	14 041	97.044	3.107	6189	6414	96.492	2.857
5000	99.999	97.678	97.676	13 628	14 045	97.031	3.093	6190	6377	97.068	2.841
12 370	99.998	97.670	97.668	13 641	14 052	97.075	3.001	6189	6521	94.909	2.857

Among the 16 428 SNPs, 13 542 are common SNPs, 1356 are low-frequency SNPs and 1530 are rare SNPs. Within the gold standard, there are 1 222 885 valid genotype calls in total, 14 063 minor allele calls at low-frequency SNPs and 6371 minor allele calls at rare SNPs.

However, a high minor allele concordance can be achieved by a conservative algorithm that only calls the easy-to-call minor allele genotypes but misses out on most of the genuine minor allele calls. We additionally evaluated the extent that each caller is missing genuine minor allele calls. For low-frequency SNPs, iCall consistently exhibited the lowest missed minor allele call rate (with a maximum of 2.915%), compared with 2.958% for GenCall and 3.029% for optiCall with 12 370 samples. However, for rare SNPs, iCall was more conservative and made less minor allele genotype calls than optiCall, especially when the sample size is large (missed minor allele call rate of 3.280 and 2.967% for iCall and optiCall, respectively) although the genotype calls by iCall are much more likely to be correct (concordance of 97.546% by iCall versus 94.166% by optiCall). As the number of samples available for joint calling increases, optiCall appears to be more liberal at making minor allele calls, whereas iCall appears to be stable. GenCall, Illuminus and GenoSNP consistently performed poorly when measured with these two minor allele metrics.

zCall is a post-processing caller that uses intensities and genotypes generated from a standalone caller as input data. We also compare the performance of GenCall+zCall, optiCall+zCall and iCall+zCall (Table 2). The results show that zCall improves the genotype calls generated from all the three callers. However, zCall improves GenCall in a higher degree than iCall in calling minor alleles, with the greatest degree of improvement observed for GenCall genotypes. GenCall+zCall is slightly better than iCall+zCall with marginally higher overall concordance rate (GenCall+zCall at 97.681% and 97.684% with



sample sizes of 500 and 12,370 respectively, against iCall+zCall's 97.681% and 97.683%) and concordance rate (GenCall+zCall's 97.685% against iCall+zCall's 97.682%-97.685%). But iCall+zCall is slightly better than GenCall+zcall in calling minor alleles at both low-frequency and rare SNPs. optiCall+zCall exhibited the same characteristics as optiCall, where it is more aggressive in calling minor allele genotypes but at the expense of making more erroneous calls.

## **2.6 Discussion**

We have introduced iCall, a method for calling genotypes that yields comparatively better performance than existing genotype calling algorithms, particularly in accurately calling the genotypes involving minor alleles at low-frequency or rare SNPs. One important aspect of genotype calling is that determining the genotypes accurately is straightforward for the majority of the SNPs, but there are SNPs where the MAF is considerably lower or when the hybridization profiles differ from the usual that require more robust considerations to accurately determine the genotypes. Our method improves on the framework of Illuminus by using a series of penalty functions to identify the optimum parameters to seed the EM model. The availability of a large dataset that has been genotyped on the exome chip meant that we could evaluate the performance of existing algorithms across different sample sizes.

We have benchmarked the genotype calls obtained from different methods against a set of gold standard calls that was derived from deep sequencing. As a stand-alone caller, iCall performs the best in terms of delivering the most

accurate genotype calls while minimizing the number of missed calls, particularly for genotypes involving minor alleles at low-frequency and rare SNPs. The better performance at low-frequency and rare SNPs was similarly observed when iCall was incorporated as part of a two-stage calling process with zCall.

We have compared iCall against existing methods using two additional metrics that specifically focused on the ability to call the genotypes that involved at least one minor allele at rare and low-frequency SNPs. This is in line with the intended purpose of the exome microarray for finding low-frequency or rare SNPs that are associated with phenotypes. Measuring how accurately and sensitively a calling algorithm can call a heterozygous or minor allele-homozygous genotype is thus more important. After all, an algorithm that erroneously calls a rare SNP as major-allele monomorphic will have attained a concordance of at least 98%. In quantifying the association evidence at rare or low-frequency SNPs, it is common to pool allele counts across similar SNPs in a genomic region to assess allelic burden [72-75]. Erroneously calling the presence of a minor allele genotype, or the failure to call a minor allele genotype when it exists, can thus directly impact the power and false-positive rate of the association analyses.

One challenge with assessing the quality of rare variant genotyping calls is the lack of a gold standard reference for the true genotypes of a given SNP in a given sample [57]. The three ways available for assessing rare variant genotypes calls are: (i) to calculate transmission rates and Mendelian

inheritance errors from a large dataset of related individuals; however, data for this approach are usually not available. (ii) An alternative approach is to simulate rare variants by sampling from common, high-quality, and well-accepted SNP genotype data and comparing the new call rate and genotype information with the original genotype calls; however, this sampling usually assumes Hardy-Weinberg Equilibrium and the intensities of those high-quality calls often have clear classification profile and less noises. These properties are sometimes not true for real data. (iii) Use different platforms to genotype the samples and obtained the consistent set of calls to be the gold standard. Extracting consistent genotype calls from different platforms could eliminate much of the non-biological noises introduced in the genotyping procedure caused by different microarray designs or other technical sources of variation. In our study, we used the third approach. More careful assessment could be done by including approach (i) and approach (ii) to evaluate the performance of iCall in future work.

Automated algorithms for calling genotypes have contributed to the success of large-scale genomic studies, and this is likely to continue with the continuous introduction of next-generation genotyping microarrays designed with knowledge gained from large-scale sequencing studies, querying up to 5 million SNPs across the genome or variants found specifically in the exons. Although these technologies provide the opportunity to investigate new hypotheses on the evolution of the human genome and the genetic etiology of diseases and traits, this can only happen if the content in the human genome can be accurately determined. We have introduced a calling algorithm that

provides a better accuracy in calling genotypes for rare and low-frequency SNPs, and consistently performs well at common SNPs.

## **2.7 Supplementary Information**

### **2.7.1 Intensity Data**

iCall, optiCall, Illuminus, GenCall use the normalized hybridization intensities for the respective two alleles at each SNP that is generated from the proprietary software GenomeStudio as the input, whereas GenoSNP uses the raw intensities as input instead of normalization hybridization intensities.

12022 samples from multiple complex disease studies that are being carried out at the Genome Institute of Singapore:

GSGT Version        1.9.4

Processing Date        6/26/2013 8:01 AM

Content        Exome\_Asian\_30K\_ExomePlus\_15031624\_B.bpm

348 samples from Singapore Integrative Omics Project:

GSGT Version    1.9.4

Processing Date 9/11/2013 3:57 PM

Content        Exome\_Asian\_30K\_ExomePlus\_15031624\_B.bpm

### **2.7.2 Genotype Calling Algorithm Implementation**

#### **2.7.2.1 GenCall**

12022 samples from multiple complex disease studies that are being carried out at the Genome Institute of Singapore:

GSGT Version 1.9.4

Processing Date 6/26/2013 8:01 AM

Content Exome\_Asian\_30K\_ExomePlus\_15031624\_B.bpm

348 samples from Singapore Integrative Omics Project:

GSGT Version 1.9.4

Processing Date 9/16/2013 12:39 PM

Content Exome\_Asian\_30K\_ExomePlus\_15031624\_B.bpm

#### 2.7.2.2 optiCall

The version of optiCall used is: tss101-opticall-76a3850f251a (updated in 2012-10-15). We ran optiCall with default parameters using command:

```
./opticall -in sample_intensity.txt -out sample_opticall
```

#### 2.7.2.3 Illuminus

The Illuminus program was obtained by requesting from the author on date 2012-10-17. We ran Illuminus with default parameters using command:

```
./Illuminus -i sample_intensity.txt -o sample_Illuminus -c
```

#### 2.7.2.4 GenoSNP

The version of GenoSNP used is: GenoSNP\_Exe\_v1.3. We ran GenoSNP with default parameters using command:

```
./GenoSNP -snps snpfile.txt -samples samplefile.txt -calls calls.txt
```

#### 2.7.2.5 iCall

The iCall program we used can be downloaded at

<http://www.statgen.nus.edu.sg/~software/icall.html>. We ran iCall with default parameters using command:

```
./iCall -i sample_intensity.txt -o sample_iCall -c
```

#### 2.7.2.6 zCall

The version of zCall used is: zCall\_Version3.3\_GenomeStudio. We ran zCall with default parameters. We applied zCall to intensity and calls generated by default caller GenCall, optiCall, and iCall and ran zCall respectively, with command:

```
STEP1: Python findMeanSD.py -R zCall_input.txt > my.mean.sd.txt
```

```
STEP2: Rscript findBetas.r my.mean.sd.txt my.betas.txt 1
```

```
STEP3: python findThresholds.py -B my.betas.txt -R zCall_input.txt -Z 7 -I  
0.2 > my.output.threshold.txt
```

```
STEP4: python zCall.py -R zcall_input.txt -T my.output.threshold.txt -O  
my.output.root.for.tped_tfam
```

#### 2.7.3 Whole Genome Sequencing Genotyping

Whole-genome sequencing genotype were obtained on Illumina Hiseq 2000 platform at a deep coverage of 30-fold. Among the 81 sequencing samples, 45 samples came from Singapore Sequencing Malay Project (SSMP), 36 samples came from Singapore Sequencing Indian Project (SSIP) (Table 3).

**Table 3. The resources of whole genome sequencing data. 45 samples came from SSMP and 36 samples came from SSIP.**

Sample.ID	Ethni c	Sample.ID	Ethni c	Sample.ID	Ethni c
-----------	------------	-----------	------------	-----------	------------

ME-1QK9PV6	Mala y	ME-G5VVJWS	Mala y	ME- WA1WVCC	India n
ME-2DY341S	Mala y	ME-GCXTDKT	Mala y	ME-WSU2EF6	India n
ME-2LK97T1	Mala y	ME-GF98MZ5	Mala y	ME-WXX4H6D	India n
ME-34PV687	India n	ME-GTJBBGA	India n	ME-X915CPN	India n
ME-5JUYP3D	Mala y	ME-HCL8VLX	Mala y	ME- XHNW5GG	India n
ME-5SMRK9M	India n	ME-HR7SP2B	Mala y	ME- Y5YHAWK	Mala y
ME-69QXS73	Mala y	ME-IL2JEBQ	India n	ME-YIQ4TX1	India n
ME-6D8MVP3	Mala y	ME-IR7CT2Z	Mala y	ME-YWBM8JR	India n
ME-6E5M6ZY	Mala y	ME-ISBZ269	Mala y	ME-ZR962SS	Mala y
ME-6PTNSHM	Mala y	ME-JA7MIHW	Mala y	ME-ZS5EI5A	Mala y
ME-7G29KYE	India n	ME-JLH2YFN	Mala y	ME-1W363FP	Mala y
ME-7R5VPTN	India n	ME-K62U8HX	India n	ME-3S3MIXQ	Mala y
ME-88VI9T6	Mala y	ME-LFNREY2	India n	ME-6BZLEI6	Mala y
ME-8BNI435	Mala y	ME-LIBZ5CL	India n	ME- AWKAXEF	Mala y
ME-8N76Z3J	India n	ME-LL1YTLG	Mala y	MECA0710190 2	India n
ME-8YNDYMF	Mala y	ME-LVW6UAR	Mala y	MECA0803221 4	Mala y
ME-9CFKAVL	Mala y	ME- LWMDWTB	India n	ME- DCHWYRB	Mala y
ME-9JZJEXG	India n	ME-MFW9KEJ	Mala y	ME-FGKB1M2	India n
ME-A4D116T	India n	ME-NHIM15G	Mala y	ME-HBA2T3S	Mala y
ME-A5F6GSV	India n	ME-NU63IA8	India n	ME-LJ3Y9SP	India n
ME-AS9L56T	India n	ME-Q9LIAKM	India n	ME-M5E8X6U	Mala y
ME-B8U42YS	Mala y	ME-R2ZU9N9	India n	ME- MHH1MLY	Mala y
ME-BXH58LW	India n	ME-SKZLAZ8	India n	ME-SWPVF5V	Mala y
ME-DRHR42A	Mala y	ME-T4SIGJB	India n	ME-VZU46ZV	India n
ME-EG31367	India n	ME-T5AKQPW	India n	ME-WI8KCKR	India n
ME- EMM3MTN	Mala y	ME-USH3597	Mala y	ME-XC14WVV	Mala y

ME-F45PX59	India n	ME-VKP2PDU	Mala y	ME-YLF7IFK	Mala y
------------	------------	------------	-----------	------------	-----------

### 2.7.3.1 Singapore Sequencing Malay SNP discovery and quality control

The SNP discovery and quality control are performed in two ways: (i) single-sample SNP calling using CASAVA with the small variant caller module; (ii) multi-sample SNP calling using SAMTOOLS 0.1.17 [76]. The final set of SNPs that are used in our study only included those that have been discovered by both CASAVA and SAMTOOLS.

#### *CASAVA – Quality Control*

- Remove the candidate SNPs that possesses a  $Q(\text{snp}) < 20$ .
- Remove the candidate SNPs that possesses call depth greater than 3 times the mean sequencing depth of the chromosome.
- Removed all heterozygous SNPs for SNPs discovered in chromosome Y and mitochondria.

#### *SAMTOOLS – Quality Control*

- Remove the candidate SNPs that possesses variant quality  $\leq 3$
- Remove the candidate SNPs that possesses read depth smaller than 3 or higher than maximum read depth (mean read depth +  $3 \times$  standard deviation of read depth).
- Remove the candidate SNPs within 10bp of a gap.
- Removed all heterozygous SNPs for SNPs discovered in chromosome Y and mitochondria.



### 2.7.3.2 Singapore Sequencing Indian SNP discovery and quality control

The SNP discovery and quality control are performed in two ways: (i) single-sample SNP calling using CASAVA with the small variant caller module; (ii) multi-sample SNP calling using GATK 2.1.8 [77]. The final set of SNPs that are used in our study only included those that have been discovered by both CASAVA and GATK.

#### *CASAVA – Quality Control*

- Remove the candidate SNPs that possesses a  $Q(\text{snp}) < 20$ .
- Remove the candidate SNPs that possesses call depth greater than 3 times the mean sequencing depth of the chromosome.
- Removed all heterozygous SNPs for SNPs discovered in chromosome Y and mitochondria.

#### *GATK – Quality Control*

- Firstly, use GATK to realign the bam file, remove the duplicates and recalibrate the bases.
- SNPs are called by recalibrating variants with SNP call annotations (QD, HaplotypeScore, MQRankSum, ReadPosRankSum, FS, MQ, InbreedingCoeff, and DP) and removing the candidate SNPs whose variant quality score is below 99.0.

**Table 4. Comparison of TMRCA methods**

Name	Methodology	Input Classification	Input Information	TMRCA	Demographic model
T-LD	Statistical estimator from LD	Summary statistics of alignments scalable to genomic size	LD of variants with genetic distance within 0.005cM - 0.1cM	Point estimation	Isolation migration model
T-FST	Statistical estimator from FST and LD	Summary statistics of alignments scalable to genomic size	LD of variants with genetic distance within 0.005cM - 0.1cM SNP-wise FST of the two populations.	Point estimation	Isolation migration model
MIMAR	MCMC	Summary statistics at multiple neutral loci of size ~1000bp	The summary statistics at each locus: the numbers of polymorphisms unique to the samples from populations 1 and 2; the number of shared alleles between the two samples and the number of fixed alleles in either sample.	Posterior distribution mean±sd	Isolation migration model or more complex model specified by user
GPho	MCMC	Full data at multiple neutral loci of size ~1000bp	Each locus provides several samples of diploid or haploid sequences of multiple populations Out-group sequence can be used for mutation rate calibration.	Posterior distribution mean±sd	Phylogeny tree given by user Constant population size to be estimated
DADI	Diffusion Approximation	Summary statistics of alignments scalable to genomic size	Allele Frequency Spectrum of multiple populations. Out-group information can be used for polarization.	Point estimation	Demographic function specify by user with sets of parameters to be estimated.
CoalHMM	HMM - MCMC	Full data of alignments scalable to genomic size	Two genomic size haploid sequences: one from population_1 and the other one from population_2.	Posterior distribution mean±sd	Isolation model
PSMC	HMM – Maximize Likelihood Estimation	Full data of alignments scalable to genomic size	Pseudo-diploid sequences constructed from two genomic size haploid sequences: one from population_1 and the other one from population_2	Qualitative estimation. PSMC provides an estimation of historical population size as a step function of time. The time when population size tends to infinity is the divergence time.	A step function with boundaries of the intervals specified by users and function values to be estimated.
MSMC	HMM – Maximize Likelihood Estimation	Full data of alignments scalable to genomic size	Small samples of genomic size phased sequences from two populations. Normally equal numbers of sequences in each of the two populations (2-4 haploid sequences for each population).	Qualitative estimation. MSMC provides a metric, relative cross coalescence rate, to measures the gene exchange between two populations. It is a step function of time having value in [0,1]. It shows the dynamic process of relative gene flow changes between two populations, indicating the process of population divergence.	A step function with boundaries of the intervals specified by users and function values to be estimated.

## **CHAPTER 3. STATISTICAL EVALUATION OF TMRCA**

### **ALGORITHMS**

#### **3.1 Modern Methods of Estimating TMRCA**

The inference of the divergence time between populations has been of fundamental interest in the study of population evolution. There is broad consensus for the “Out-of-Africa” hypothesis which states that modern human arose about 200,000 years ago in Africa and spread throughout the continents around 100,000 years ago. This was followed by several waves of major population dispersals across the globe, although the exact nature of the population divergence remains debatable. Existing methods to estimate population divergence time differ in their methodological frameworks and demographic assumptions, and require different types of genetic data as input. These fundamental differences often result in the methods producing inconsistent estimates of the population divergence time, further confounding attempts to robustly uncover the history of human migration, especially when most population genetic studies do not employ multiple methods to estimate the time to the most recent common ancestor (TMRCA). Therefore, a systematic evaluation of the existing methods is needed to provide guidance for researchers who attempt to investigate population divergence time.

#### **3.2 Theories in Population Genetics**

##### **3.2.1 Coalescent Theory**

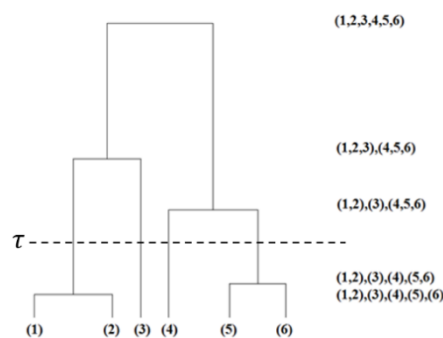
In this session, we will discuss the main ideas of coalescent theory briefly in the absence of selection, complex population structure and unfixed population size.

### 3.2.1.1 Poisson Process

The Poisson process is a fundamental theory in coalescent theory. A Poisson process is a stochastic process counting events that occur independently and randomly in the time. In a Poisson process with rate  $\lambda$ , the probability of an event happens in the time interval  $(t, t + \delta t)$  is  $\lambda\delta t$  and the time before the first event or between two adjacent events follows an exponential distribution with mean  $1/\lambda$ .

Consider two Poisson processes, process A and process B of rate  $a$  and  $b$  respectively. The combined process (defining an event as either A or B) is a Poisson process of rate  $(a + b)$ . Moreover, given an event occurs, the probability that the event is A equals  $a/(a + b)$  and the probability that the event is B equals  $b/(a + b)$ .

### 3.2.1.2 Coalescent Process



**Figure 13. An example of a genealogical tree of a sample of 6 genes. The column on the right shows the equivalent relations of the genealogy.**

Consider a sample of  $n$  genes taken at the present time. Let  $\tau$  represent a time moving backward before the samples are taken. Genes are defined in the same equivalent class at time  $\tau$  if they have a common ancestor. For example in Figure 13, there are six genes sampled at time 0. At time  $\tau$ , gene 1 and gene 2

have a common ancestor; gene 5 and gene 6 have another common ancestor.

Hence, there are four equivalent classes at time  $\tau$ . The equivalence relation

can be described by

$\{(1), (2), (3), (4), (5), (6)\}$  at time 0

$\{(1,2), (3), (4), (5,6)\}$  at time  $\tau$

If amalgamating two equivalence classes in an equivalent relation can produce

a new equivalence relation (e.g. in Figure 13, amalgamating (4) and (5,6)

results in equivalent relation  $\{(1,2), (3), (4,5,6)\}$ ), this amalgamating is called

a coalescence and the process of successive amalgamations is called a

coalescence process [42]. Let  $\phi_1 = \{(1,2,3, \dots, n)\}$  and  $\phi_n =$

$\{(1), (2), (3), \dots, (n)\}$  be two equivalence relations. Kingman (1982) uses a

stochastic model to describe the coalescence process moving from  $\phi_n$  to  $\phi_1$ .

The coalescence process can be typically illustrated by a genealogical tree.

Let  $\zeta$  represent an equivalence relation and  $\eta$  represent any possible

equivalence relation that can be obtained after a coalescence occurs on  $\zeta$ .

Suppose there are  $k$  equivalent classes in  $\zeta$ , and thus there are  $C_2^k$  possible

coalescences. According to Kingman's coalescent theory, the time to the next

coalescence event is an exponential distribution of mean  $1/C_2^k$ . Let  $T_k$  be the

time to the next coalescence event when there are  $k$  equivalent classes present

in the coalescent process and  $T_{MRCA}$  be the coalescent time to the most recent

common ancestor (MRCA) of  $n$  genes. Then  $T_{MRCA} = \sum_{k=2}^n T_k$ .

### 3.2.1.3 Relation between Coalescent Theory and Wright Fisher Model

The equivalent class in coalescent theory is equivalent to a lineage in the Wright Fisher model (see Section 1.7.5). On one hand, let  $T_k$  be the time to the next coalescence event when there are  $k$  equivalent classes present in the coalescent process. Then  $T_k \sim \text{Exponential}(C_2^k)$ . On the other hand, consider  $k$  lineages in a diploid population of size  $N_0$  ( $k \ll N_0$ ) and let  $W_k$  represent the number of generations until any two of the  $k$  lineages have a common ancestor. Then  $W_k \sim \text{Exponential}(C_2^k/2N_0)$ . Therefore, the relations between time in generations in the Wright-Fisher model and coalescence time is

$$W_k = T_k \times 2N_0$$

### 3.2.1.4 Coalescent theory with Recombination

When recombination is included in the coalescent process, recombination events can be modelled by a Poisson process with rate  $R/2$  ( $R = 4N_0r$  and  $r$  is the recombination rate). Define an event to be either a recombination or a coalescence and suppose that there are  $k$  equivalent classes at time  $\tau$  in the process. The probability that an event occurs in  $(\tau, \tau + \delta\tau)$  is  $\frac{1}{2}k(k-1)\delta\tau + \frac{1}{2}kR\delta\tau = \frac{1}{2}k(k+R-1)\delta\tau$ , and the probability that the event occurring is a recombination is  $R/(k-1+R)$ .

### 3.2.1.5 Scaling with $N_0^{ref}$

If we want to use the reference population size  $N_0^{ref}$  as the scaling factor,

define the relative population size  $\lambda = \frac{N_0^{ref}}{N_0}$  ( $\lambda$  could be a function of time  $t$ ).

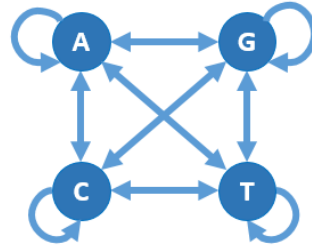
$T_k$  can be modelled by a Poisson process of rate  $\lambda C_2^k$  (also denoted as

coalescence rate  $C = \lambda C_2^k$ ). Considering recombination ( $R = 4N_0^{ref}r$ ),  $T_k$  can

be modelled by a Poisson process of rate  $\lambda C_2^k + \frac{1}{2}kR$ . The coalescent time in generations is  $2N_0^{ref} \times$  coalescence time.

### 3.2.2 Full Data Likelihood Calculation based on Substitution Markov model

#### 3.2.2.1 Jukes-Cantor Model



**Figure 14. Jukes-Cantor model. The four nucleotides substitute in a Markovian manner.**

In order to make any statistical inference about the genealogy, we need to compute the likelihood of a genealogy given the observed DNA sequences.

The DNA substitution model first introduced by Jukes and Cantor in 1969 [78] is a one parameter continuous time Markov model (Figure 14), which assumes that transition rates between any two nucleotides are equal and that all sites in the sequence are independent. Hence the likelihood is the product of the probabilities taken across sites and the main problem is to calculate the likelihood at one site.

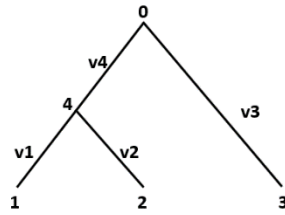
Assuming the continuous time Markov model has four states {A, T, C, G} and the transition rate matrix  $Q$  has the form as follows.

$$Q = \begin{bmatrix} -\frac{3u}{4} & \frac{u}{4} & \frac{u}{4} & \frac{u}{4} \\ \frac{u}{4} & -\frac{3u}{4} & \frac{u}{4} & \frac{u}{4} \\ \frac{u}{4} & \frac{u}{4} & -\frac{3u}{4} & \frac{u}{4} \\ \frac{u}{4} & \frac{u}{4} & \frac{u}{4} & -\frac{3u}{4} \end{bmatrix}$$

So the transition matrix  $P(t) = \exp(tQ)$  has the form as follows.

$$P(t) = \begin{bmatrix} 1 - 3a_t & a_t & a_t & a_t \\ a_t & 1 - 3a_t & a_t & a_t \\ a_t & a_t & 1 - 3a_t & a_t \\ a_t & a_t & a_t & 1 - 3a_t \end{bmatrix}, \text{ where } a_t = \frac{1 - \exp(-4t)}{4}$$

and where  $P_{ij}(t)$ ,  $i, j \in \{A, T, C, G\}$  represents the probability that a lineage which is initially in state  $i$  will be in state  $j$  after  $t$  units of time. The stationary distribution is defined by  $\pi = (0.25, 0.25, 0.25, 0.25)$ .



**Figure 15. An example of a genealogy of 3 genes. Vertices 1-3 represent present genes, vertices 4 represent an ancestral gene and vertex 0 represents their MRCA. Edges v1-4 represent the length of time past for a coalescence event.**

Given a genealogy at one site, the probability of obtaining a give set of alleles at the tips can be computed as the product of the base substitution probabilities of all lineages. For example in Figure 15, there are three gene samples at one locus. The likelihood is

$$L = \sum_{s_0} \sum_{s_4} \pi_{s_0} P_{s_0 s_4}(v_4) P_{s_4 s_1}(v_1) P_{s_4 s_2}(v_2) P_{s_0 s_3}(v_3)$$

where  $s_i$  represents the state (base) at point  $i$ ;  $v_i$  represents the waiting time for a coalescence;  $\pi_{s_0}$  represents the prior probability of the states at point 0.

### 3.2.2.2 Felsenstein's Estimation Method

In 1981, Felsenstein extended Jukes-Cantor Model and proposed a four-parameter continuous time Markov model [79]. In Felsenstein's model, the stationary distribution  $\pi$  could be any probability and the transition probability  $P(t)$  can be subsequently derived from  $\pi$  by the relation:

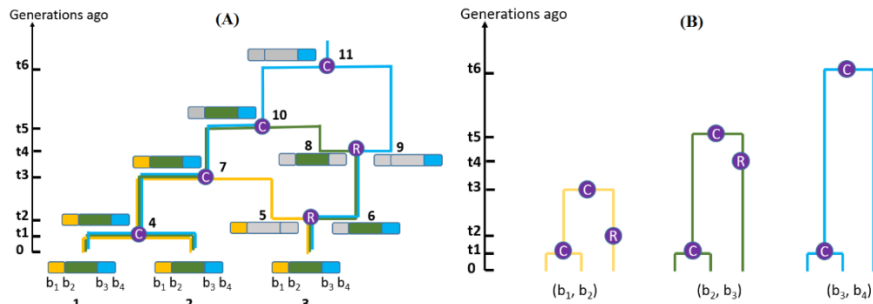


$$P_{ij}(t) = e^{-ut}I_{ij} + (1 - e^{-ut})\pi_j$$

where  $I_{ij} = 1$  if  $i = j$ .

Felsenstein suggested the stationary distribution  $\pi$  to be the base composition proportions in the samples under the study.

### 3.2.3 Ancestral Recombination Graph



**Figure 16. Illustration of ARG of a sample of three genes. (A) Two recombination occurred at  $t_2$  and  $t_4$ . These recombination separate the gene into three segments  $(b_1, b_2)$ ,  $(b_2, b_3)$  and  $(b_3, b_4)$  (ancestral material colored in yellow, green and blue, respectively; non-ancestral material colored in grey). (B) The corresponding genealogies of  $(b_1, b_2)$ ,  $(b_2, b_3)$  and  $(b_3, b_4)$  are colored in yellow, green and blue respectively.**

Recombination plays a critical role in reproduction. Recent studies on haplotype patterns and LD structures show that recombination harbors a considerable amount of information about recent population history [80]. In the scenario assuming no recombination, each sequence has a single ancestor in its parent generation. Thus all sequences ultimately have a single common ancestor and the inheritance relationships of them could be represented by a genealogical tree. In the scenario with recombination, sequences would be broken up by recombination events into segments that have different genealogies. An ancestral recombination graph (ARG) has been proposed to depict the coalescent process integrating a series of coalescent and recombination events.

In an ARG, assuming at a time point of the coalescent process, there is a set of  $k$  lineages, the  $i$ th lineage of which contains ancestral gene material at  $m_i$  ordered non-overlapping intervals on a continuous unit interval  $\mathbf{x}_i = \{(x_{i1}, y_{i1}), (x_{i2}, y_{i2}), \dots, (x_{im_i}, y_{im_i})\}$ . The coalescence process can be modelled by a Poisson process with coalescent rate  $\lambda_C$  and recombination rate  $\lambda_R$ , where  $\lambda_C = \sum_{i \neq j} I_{ij}$ ,  $\lambda_R = R/2 \sum_i (y_{im_i} - x_{i1})$ , and  $R = 4N_e r$ .

If a coalescence occurs, the resulting lineage contains the union of the ancestral material intervals of the two coalescing lineages (e.g. Figure 16A at time  $t_1$  and  $t_3$ ). If a recombination occurs, one lineage splits into two at a splitting point uniformly distributed in the interval  $(x_{i1}, y_{im_i})$  (e.g. Figure 16A at time  $t_2$  and  $t_4$ ). If one interval is represented by only one lineage, this interval has already found its MRCA and is removed from the process (e.g. Figure 16A at time  $t_3$  and  $t_5$ , interval  $(b_1, b_2)$  and interval  $(b_2, b_3)$  are removed respectively). The coalescence process of a sample of size  $n$  starts at  $k = n$ ,  $m_i = 1$ ,  $x_{i1} = 0$ ,  $y_{i1} = 1$  for  $i = 1, \dots, n$  and ends when all the samples (0, 1) have identified the MRCA. Hence, with recombination, samples could have different genealogies and a different MRCA (Figure 16B).

Since ARG is not a tree, MCMC method suffers from a huge computational burden because of the exponentially increasing number of lineages. Hence a type of MCMC method uses multiple loci models assuming recombination within loci and free-recombination and an independent probabilistic property between loci, and explores the posterior distribution by simulating ARG for each locus independently. The size of each locus is often between several

hundred to thousand base pairs and it is challenging to scale to the whole genome.

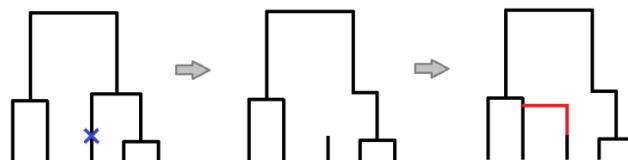
### 3.2.4 Sequential Markov Coalescent Model (SMC)

Simulating historical sequences through ARG to infer the genealogical history is severely restricted by the computational burden as the length of the region increases. To overcome this difficulty, McVean developed a Markovian model, the sequential Markov coalescent model (SMC), which considerably simplifies the model and makes the simulation of genomic size sequences possible and likelihood inference tractable [81].

In the standard coalescent model, the coalescence rate  $\lambda_c = \sum_{i \neq j} I_{ij}$ , where  $I_{ij} = 1$  for all  $i$  and  $j$  if  $i \neq j$ . In McVean's SMC,  $I_{ij} = 1$  if  $i \neq j$  and the  $i$ th lineage and the  $j$ th lineage share common ancestral material. This modification largely reduces the space complexity of ARG, and more importantly, provides the process Markovian properties.

*The algorithm of sequential Markov coalescent model (SMC)*

Assuming a continuous ancestral gene material  $(0, 1)$ , the sequential Markov coalescent algorithm is described below (Figure 17).

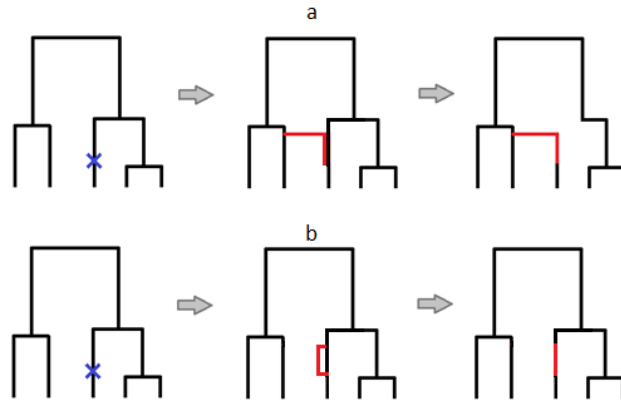


**Figure 17. Illustration of the sequential Markov coalescent model with 5 genes. The cross-mark indicates the point of recombination, which is uniformly distributed on the genealogy. The branch above the recombination point is removed, resulting in a floating branch which coalesces with existing lineages at**

**the rate proportional to the number of lineages present. The figure has been adapted from a similar figure in reference [81].**

- a. Simulate a standard coalescent history at point  $x = 0$ . The resulting genealogical tree has a total branch length of  $T(x)$ .
- b. The distance of the first recombination from  $x$  (to the right) is exponentially distributed with rate  $RT(x)/2$ . If the point of the recombination event is less than 1, the recombination breaks up the gene. The left emerging region follows the old genealogy, and the right emerging follows a new genealogy sampled as follows: The time point when the recombination event occurs is drawn uniformly from the old genealogy and the older portion of the selected branch is erased, resulting in a floating lineage. The floating lineage coalesces with the remaining genealogy at rate proportional to the number of lineages present and forms a new genealogy. Update  $x$  to the left end of the right emerging region and calculate tree length  $T(x)$  for the new genealogy.
- c. Repeat process b until the position of next recombination event occurs beyond 1.

*SMC'*



**Figure 18. Illustration of the SMC' model with 5 genes. The cross-mark indicates the point of recombination, which is uniformly distributed on the tree. The floating branch coalesce with existing lineages at the rate proportional to the number of lineages present before erasing the branch above the recombination point. (a) represents the situation that the floating branch coalesces with branch other than its ancestral branch; (b) represents the situation that the floating branch coalesces with its own ancestral branch (in this case, the recombination event will not change LD pattern).**

Marjoram modified SMC and proposed SMC' in 2006 [82]. SMC' did a slight modification that the older portion of the branch where the recombination occurs is deleted after the new line is added (Figure 18). In this way, it allows coalescence between two lineages resulting from a recombination (Figure 18b).

### 3.3 Methods for Estimating TMRCA

Many of the existing population genetics inference and methodologies have been built on the foundation of the coalescent theory [21, 83, 84], although these can be generally classified according to the type of genetic data used as input and the assumptions about population demography (Table 4). For example, one class of methods for estimating the time to the most recent common ancestor (TMRCA) considers multiple neutral loci each of around 1,000 bases only in multiple populations, such as MIMAR[85, 86] and GPhoCS [87]. Another class of methods infers the TMRCA from full chromosomal information, such as CoalHMM [88], PSMC [89] and MSMC [90]. The third

class of methods essentially infers the TMRCA on the extent of linkage disequilibrium (LD), population diversity measured by the  $F_{ST}$  parameter and population allele frequency, such as the approaches by Hayes and colleagues (abbreviated subsequently as T-LD) [26, 91], by McEvoy and colleagues (abbreviated subsequently as T-FST) [24], and DADI [92]. These methods differ by the type of input data required (sequence-level information or summary statistics), and by the assumption around the presence of genetic recombination during migration [92].

These different methods can also be classified by the statistical framework used in the design of the methods. Notably, MIMAR and GPho-CS are Markov chain Monte Carlo (MCMC)-based methods which implement an MCMC algorithm to sample the posterior distribution of the TMRCA parameter, and possess the advantage of incorporating greater complexity in the model to allow for recombination and gene flows through migration. However, such methods are typically computationally expensive and scaling up to allow whole-genome sequences to be considered as input remains intractable. Conversely, methods such as CoalHMM, PSMC and MSMC adopt a hidden Markov model (HMM) framework which assumes a Markovian behavior when considering recombination events. This reduces the computing burden and has been extended to allow the whole genomic sequence to be analyzed. T-LD and T-FST derive the TMRCA by computing statistics measuring the extent of LD or  $F_{ST}$ , while DADI infers the TMRCA between two populations from a diffusion approximation of the allele frequency spectrum.

In the following sections, we review the methodologies of eight existing methods used to estimate TMRCA (T-LD, T-FST, MIMAR, GPho-CS, DADI, CoalHMM, PSMC and MSMC).

### 3.3.1 Statistical Estimators of TMRCA

We reviewed two methods for estimating population divergence time using genotyping data. T-LD is a statistical estimator of TMRCA based on LD structure and T-FST is another statistical estimator based on LD and FST information.

#### 3.3.1.1 T-LD

##### *Conceptual framework 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 ( $r_{pop}$ ) should be 1.0 right after divergence, and decays with time due to recombination. If LD is measured by correlation coefficient  $r$ , Hill and Robertson showed that  $r_{pop}$  decays in a manner that after  $T$  generations over genetic distances ( $c$ ) of  $r_{pop} = e^{-2cT}$ , assuming constant population size and finite unselected random mating population [93].

##### *Methodology*

- a. Extract the sites that are segregating (polymorphic) in all populations under the study. Estimate LD by correlation coefficient  $r$  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.
- b.  $r$  values 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 ( $r_{pop}$ ) between two populations of interest.
- c. Regress  $r_{pop}$  onto genetic distance to obtain the divergence time  $T$ .

### 3.3.1.2 T-FST

#### *Conceptual framework of T-FST*

Under neutral evolutionary theory, the population genetic differentiation sources from gene drift and can be estimated by  $F_{ST}$ . The extent of gene drift depends on effective population size ( $N_0$ ) and the population divergence time ( $T$ ) such that  $F_{ST} \approx T/(2N_0)$  [94]. According to Hill and Robertson, LD between markers far apart reflects recent  $N_0$  and the LD between markers closed together reflects ancient  $N_0$ . Sved and Nei (1987) both reported that  $E(r^2) \approx 1/(2 + 4N_0c)$  is approximately true for  $N_0 \frac{1}{2c}$  generations ago, where  $r^2$  is the square of genetic correlation coefficient and  $c$  is the genetic distance [91, 94, 95]. Therefore, the effective population size can be estimated by LD structure and population divergence time can be thus derived.



## Methodology

- a. Extract the sites that are segregating (polymorphic) in all populations under the study. Compute the average of the SNP-wise  $F_{ST}$ .
- b. Estimate LD by the square of correlation coefficient  $r^2$  in each population separately for each pair of SNPs of genetic distance from 0.005cM to 0.1cM. Adjust  $r^2$  values for each population by  $(1/n)$  to account for experimental sample size. Similar to T-LD,  $r^2$  are binned into 19 categories. The effective population size is computed by  $\left[\frac{1}{E(r^2)} - 2\right] / 4c$  for each bin and a single point estimation takes the average of the 19 values for each population separately.
- c. The harmonic mean of the effective population sizes of the two populations of interest ( $N_0$ ) is computed and the population divergence time is estimated as  $T \approx 2N_0F_{ST}$ .

### 3.3.2 MCMC methods

Many Bayesian methods have been established to estimate the evolutionary parameters including effective population size and population divergence time through simulating genealogies or ARGs based on coalescence theory and inferring parameters from their posterior distributions.

#### 3.3.2.1 MIMAR

MIMAR is a multilocus model for estimating population parameters under isolation-migration model allowing recombination [85]. The parameters of interest include three population mutation rate, ( $\theta_A, \theta_1$  and  $\theta_2$ ), one or two migration rates ( $m$  assuming symmetric migration or  $(m_1, m_2)$  assuming

asymmetric migration), a divergence time ( $T$ ) and a recombination rate ( $R$ ), where  $\theta = 4N_0u$  and  $u$  represents the mutation rate per site per generation. Hence  $N_0$  and  $\theta$  are equivalent with a fixed  $u$ . MIMAR explores the posterior probabilities of  $\Theta = \{\theta_1, \theta_2, \theta_A, m, T, R\}$  and infers the parameters through MCMC.

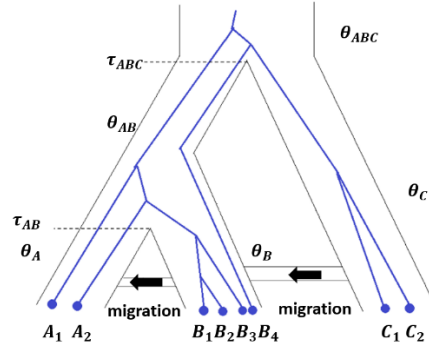
The data that MIMAR utilizes are the segregating sites summaries for multiple independent neutral loci, each of which has a length of hundreds of base pairs. The segregating sites summaries used by MIMAR are  $\mathbf{X} = (S_1, S_2, S_s, S_f)$ , where  $S_1$  and  $S_2$  are numbers of polymorphisms unique to the samples from populations 1 and 2 respectively;  $S_s$  is the number of shared SNP between the two samples; and  $S_f$  is the number of fixed variants in either sample. MIMAR assumes loci are independent to each other, and 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 ( $P(\Theta|\mathbf{X})$ , where  $\mathbf{X}$  is the observed data at a single locus),  $P(\Theta|\mathbf{X})$  can be expressed by a Bayesian framework.

$$P(\Theta|\mathbf{X}) \propto f(\Theta) \int_G f(\mathbf{X}|G, \Theta) f(G|\Theta) dG$$

Given an ARG  $G$  and parameter  $\Theta$ , the likelihood  $f(\mathbf{X}|G, \Theta)$  can be either derived from coalescent theory or estimated from a traditional substitution model. The conditional probability  $f(G|\Theta)$  can be evaluated using coalescent theory. The prior distributions  $f(\Theta)$  for  $\theta_1, \theta_2, \theta_A, T$  and  $\log(m)$  are uniform with provided or default boundaries. MIMAR designs a Metropolis-Hasting

MCMC procedure to sample the parameters and infers the population divergence time using the expected value of its stationary distribution.

### 3.3.2.2 GPho-CS



**Figure 19. Illustration of the model of GPho-CS. There are eight lineages, with 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  $\theta_A$ ,  $\theta_B$ ,  $\theta_C$ ,  $\theta_{AB}$  and  $\theta_{ABC}$  respectively. This figure has been adapted from a similar figure in reference [87].**

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 multiple populations. GPho-CS assumes no intralocus recombination and allows multiple migration bands. In our application, we assume two populations and an isolation-migration model.

Consider a known population phylogeny (tree)  $T$ . For each population  $p$ , the population mutation rate  $\theta_p$  and population divergence time  $\tau_p$  are the parameters of interest (Figure 19). Input observations are haploid (or diploid) sequence alignments at multiple loci  $\{X_i\}$  ( $i$  represents locus  $i$ ). GPho-CS uses MCMC to sample parameters according to their joint posterior density, using two main components: (a) the computation of the data density function

$P(\{\mathbf{X}_i\}, \{G_i\}, \{\theta_p\}, \{\tau_p\}, \{m_b\} | T)$  and (b) the update scheme for  $(\{\theta_p\}, \{\tau_p\}, \{G_i\}, \{m_b\})$ , where  $G_i$  represents the genealogy for locus  $i$  and  $m_b$  represents the migration rate of migration band  $b$ .

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

$$P(\{\mathbf{X}_i\}, \{G_i\}, \{\theta_p\}, \{\tau_p\}, \{m_b\} | T) = (\prod_p P(\theta_p)) (\prod_p P(\tau_p)) (\prod P(m_b)) (\prod_i P(G_i | T, \{\theta_p\}, \{\tau_p\}, \{m_b\}) P(\mathbf{X}_i | G_i))$$

where the prior  $P(\theta_p)$ ,  $P(\tau_p)$  and  $P(m_b)$  are Gamma distribution;

$P(G_i | T, \{\theta_p\}, \{\tau_p\}, \{m_b\})$  is computed based on coalescent theory and  $P(\mathbf{X}_i | G_i)$  is computed by Felsenstein's substitution model [79]. GPho-CS uses a series of Metropolis-Hastings procedure, to update the layers of 'latent' variables  $(\{G_i\}, \{\theta_p\}, \{\tau_p\}, \{m_b\})$  one by one.

### 3.3.3 HMM Methods

To model the recombination events more effectively and utilize information of whole-genome alignment, hidden Markov models have been favored by researchers. In the following section, we review three HMM-based methods that are able to utilize whole-genome data to infer TMRCA.

#### 3.3.3.1 CoalHMM



CoalHMM is a hidden Markov model that utilizes a pair of whole-genome haploid alignments, one each from the populations to estimate population parameters under an isolation model [88]. 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 describe the ancestry of two adjacent nucleotides back in time. The CTMC helps when computing the transition probability of the coalescent HMM model.
















### CTMC

CoalHMM used 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 is modelled by one-sequence system separately. When two populations merged, 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 5 and Table 6 respectively.

**Table 5. The hidden states of two adjacent nucleotides in one sequence system. Linked edge means the two nucleotides are on the same sequence. This table has been adapted from a similar figure in reference [88].**

Index	1	2
State		

**Table 6. 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.  $\{\Omega_B, \Omega_L, \Omega_R, \Omega_E\}$  represent the state sets of non-coalescence on both nucleotides, coalescence at left nucleotide, coalescence at right nucleotide, coalescence at both nucleotides, respectively. This table has been adapted from a similar figure in reference [88].**

Set	$\Omega_B$							$\Omega_L$			$\Omega_R$			$\Omega_E$	
Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
State															

The CTMCs have transition rate matrices  $\mathbf{Q}_1$  and  $\mathbf{Q}_2$  and transition matrix  $\mathbf{P}_1(\mathbf{t}) = \exp(\mathbf{Q}_1 t)$  and  $\mathbf{P}_2(\mathbf{t}) = \exp(\mathbf{Q}_2 t)$  for the one-sequence and two-sequence system, respectively.

$$Q_1 = \begin{pmatrix} - & R \\ C & - \end{pmatrix}$$

$$Q_2 = \begin{array}{c|c|c|c} & \Omega_B & \Omega_L & \Omega_R & \Omega_E \\ \hline \Omega_B & \begin{pmatrix} - & C & C & 0 & C & C & 0 \\ R & - & 0 & C & 0 & 0 & 0 \\ R & 0 & - & C & 0 & 0 & 0 \\ 0 & R & R & - & 0 & 0 & 0 \\ R & 0 & 0 & 0 & - & 0 & C \\ R & 0 & 0 & 0 & 0 & - & C \\ 0 & 0 & 0 & 0 & R & R & - \end{pmatrix} & \begin{pmatrix} C & 0 & 0 \\ 0 & C & 0 \\ 0 & 0 & C \\ 0 & 0 & 0 \\ 0 & C & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} C & 0 & 0 \\ 0 & C & 0 \\ 0 & 0 & C \\ 0 & 0 & 0 \\ 0 & 0 & C \\ 0 & 0 & C \\ 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ C & 0 \\ 0 & 0 \\ C & 0 \\ 0 & 0 \\ C & 0 \end{pmatrix} \\ \hline \Omega_L & & \begin{pmatrix} - & C & C \\ R & - & 0 \\ R & 0 & - \end{pmatrix} & & \begin{pmatrix} 0 & C \\ C & 0 \\ C & 0 \end{pmatrix} \\ \hline \Omega_R & & & \begin{pmatrix} - & C & C \\ R & - & 0 \\ R & 0 & - \end{pmatrix} & \begin{pmatrix} 0 & C \\ C & 0 \\ C & 0 \end{pmatrix} \\ \hline \Omega_E & & & & \begin{pmatrix} - & R \\ C & - \end{pmatrix} \end{array}$$

where  $C = \frac{N_0^{ref}}{N_0}$  is the coalescence rate,  $R = 2N_0^{ref} r$  is the scaled recombination rate, and  $N_0^{ref}$  is the reference effective population size.

### Coalescent HMM Model

CoalHMM used a discrete state Markov model to depict the coalescent time along the alignments. The hidden states are discretized coalescent time intervals with break points  $\tau_1, \tau_2, \dots, \tau_{k-1}$  and  $\tau_k = \infty$ , where  $\tau_1$  represents the divergence time. State  $i$  is the event that a coalescence occurs in  $[\tau_i, \tau_{i+1}]$ . The distribution of the CTMC states when entering the HMM at state  $i$  is given by

$$\pi_i = \pi_1 \exp(\mathbf{Q}_2(\tau_i - \tau_1)) \text{ for } i = 1, \dots, k. \text{ The transition probability from}$$

$$\text{state } i \text{ to state } j \text{ is then given by } (R \in j | L \in i) = \frac{P(L \in i, R \in j)}{P(L \in i)} = \frac{P(L \in i, R \in j)}{e^{-C\tau_i} - e^{-C\tau_{i+1}}}.$$

When  $i = j$ ,

$$P(L \in i, R \in j) = P(X(\tau_i) \in \Omega_B, X(\tau_{i+1}) \in \Omega_E | P(\tau_1) = \pi_1)$$

$$= \sum_{k \in \Omega_B} \sum_{l \in \Omega_E} (\pi_1 e^{\mathbf{Q}_2(\tau_i - \tau_1)})_k (e^{\mathbf{Q}_2(\tau_{i+1} - \tau_i)})_{kl}$$

When  $i < j$ ,

$$P(L \in i, R \in j) = P(X(\tau_i) \in \Omega_B, X(\tau_{i+1}) \in \Omega_L, X(\tau_j) \in \Omega_L, X(\tau_{j+1}) \in \Omega_E | P(\tau_1) = \pi_1)$$

$$= \sum_{k \in \Omega_B} \sum_{l \in \Omega_L} \sum_{m \in \Omega_L} \sum_{s \in \Omega_E} (\pi_1 e^{\mathbf{Q}_2(\tau_i - \tau_1)})_k (e^{\mathbf{Q}_2(\tau_{i+1} - \tau_i)})_{kl} (e^{\mathbf{Q}_2(\tau_j - \tau_{i+1})})_{lm} (e^{\mathbf{Q}_2(\tau_{j+1} - \tau_j)})_{ms}$$

When  $i > j$ ,

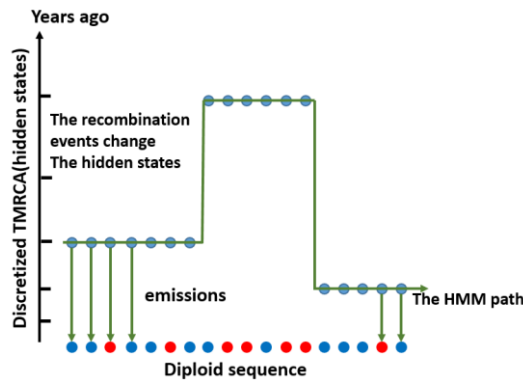
$$P(L \in i, R \in j) = P(L \in j, R \in i)$$

The transition probabilities 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, the mid-point of corresponding time interval is used.

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

### 3.3.3.2 PSMC



**Figure 20. 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 TMRCAs and the transitions are ancestral recombination events.**

**Homozygotes and heterozygotes are colored in red and blue respectively. The figure has been adapted from a similar figure in reference [89].**

PSMC is a case of sequential Markov coalescence model that infers the piecewise constant ancestral effective population size from two chromosomes [89] (Figure 20). When PSMC is applied to a pseudo-diploid sequence which each haploid sequence from one population, the divergence time could be qualitatively inferred as the time when the effective population size increase to infinity.

The hidden states are discretized coalescent time intervals  $[\tau_i, \tau_{i+1}]$ ,  $i = 1, \dots, k$ . Within each coalescent time interval, PSMC has a free parameter representing 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 the SMC model and given by:

$$p(t|s) = (1 - e^{-\rho t})q(t|s) + e^{-\rho s}\delta(t - s)$$

where  $\rho$  is the scaled recombination rate,  $\delta(\cdot)$  is the Dirac delta function and

$q(t|s) = \lambda(t) \int_0^{\min\{s,t\}} \frac{1}{s} \times \exp(-\int_u^t \lambda(v)dv) du$  is the transition probability

conditional on there being a recombination event, where  $\lambda(t) = \frac{N_0^{ref}}{N_0(t)}$  is the

relative population size at state  $t$ .



The emission probability is determined by an exponential distribution of rate  $\theta$  (scaled mutation rate):  $e(1|t) = e^{-\theta t}$ ,  $e(0|t) = 1 - e^{-\theta t}$ , where 1 means heterozygote and 0 means homozygote.

### 3.3.3.3 MSMC

MSMC is a multi-sequence extension of PSMC that also infers the piece-wise constant ancestral effective population size [90]. The hidden state of MSMC is the first coalescence represented by a triplet  $(t, i, j)$ , where  $t$  is the first coalescence time and  $i$  and  $j$  are the labels of the two lineages with regard to the first coalescence. Coalescence time is also discretized into intervals with boundaries  $[\tau_i, \tau_{i+1}]$ ,  $i = 1, \dots, k$ . Suppose there are  $M$  haploid sequences, then MSMC has  $C_2^M k$  hidden states. The transition probability and emission probability are derived under the SMC' framework and parameters are optimized by the Baum-Welch algorithm [82].

When MSMC is applied to two populations, three coalescence rates,  $\lambda_i^1(t)$ ,  $\lambda_i^2(t)$  and  $\lambda_i^{12}(t)$ , are used, where  $\lambda_i^1(t)$  and  $\lambda_i^2(t)$  represent the within population coalescence rates for population 1 and population 2 and  $\lambda_i^{12}(t)$  represents the cross population coalescence rate. MSMC defines the cross-coalescence rate  $(t) = 2\lambda_i^{12}(t)/(\lambda_i^1(t) + \lambda_i^2(t))$  as a measure of relative gene exchange rate between two populations. A population divergence process is shown if the cross-coalescence rate decreases from around one to close to zero.

### 3.3.4 Differential Approximation Methods

### 3.3.4.1 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 [96]. The basic idea is: firstly solve a diffusion equation of relative allele frequency, then calculate the expected AFS and compare it with observed AFS, and iterate the above steps to find the optimal parameters which maximize the likelihood.

Given a number of sequences from  $P$  populations, with  $n_i$  sequences from population  $i$ , AFS is defined as a  $(n_1 + 1) \times (n_2 + 1) \times \dots \times (n_P + 1)$  dimensional matrix with each entry  $S[d_1, d_2, \dots, d_P]$  ( $0 \leq d_i \leq n_i, i = 1, \dots, P$ ) counting the biallelic polymorphic sites that the number of derived allele occurrence is  $d_i$  in population  $i$  [97]. Let  $\phi(\mathbf{x}, t)$  be the process of the density of derived mutations having relative allele frequency  $\mathbf{x}$  ( $x_i \in [0,1], i = 1, \dots, P$ ) at a forward time  $t$ . Under Wright-Fisher model,  $\phi(\mathbf{x}, t)$  follows the diffusion equation:

$$\frac{\partial}{\partial \tau} \phi = \frac{1}{2} \sum_{i=1,2,\dots,P} \frac{\partial^2}{\partial x_i^2} \frac{x_i(1-x_i)}{\lambda_i} \phi - \sum_{i=1,2,\dots,P} \frac{\partial}{\partial x_i} (\gamma_i x_i (1-x_i) + \sum_{j=1,2,\dots,P} M_{i \leftarrow j} (x_j - x_i)) \phi$$

where  $\lambda_i = N_i/N_{ref}$  represents the relative population size of population  $i$ ,  $\gamma_i$  represents the scaled fitness coefficient of variants in population  $i$ ,  $M_{i \leftarrow j}$  represents the scaled migration rate from population  $j$  to population  $i$ .

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  $\phi$ . The diffusion process  $\phi$  can be solved

through a finite different method and the expected AFS can be subsequently derived in the form of:

$$M[d_1, d_2, \dots, d_P] = \int_0^1 \dots \int_0^1 \prod_{i=1, \dots, P} \binom{n_i}{d_i} x_i^{d_i} (1 - x_i)^{n_i - d_i} \phi(x_1, x_2, \dots, x_P) dx_i$$

DADI assumes the entries of AFS to be independent Poisson variables of mean  $\mathbf{M}$ . Hence the likelihood of parameter  $\Theta$  can be derived as below and maximum likelihood parameters can subsequently obtained:

$$L(\Theta|\mathbf{S}) = \prod_{i=1 \dots P} \prod_{d_i=0 \dots n_i} \frac{e^{-M[d_1, d_2, \dots, d_P]} M[d_1, d_2, \dots, d_P]^{S[d_1, d_2, \dots, d_P]}}{S[d_1, d_2, \dots, d_P]!}$$

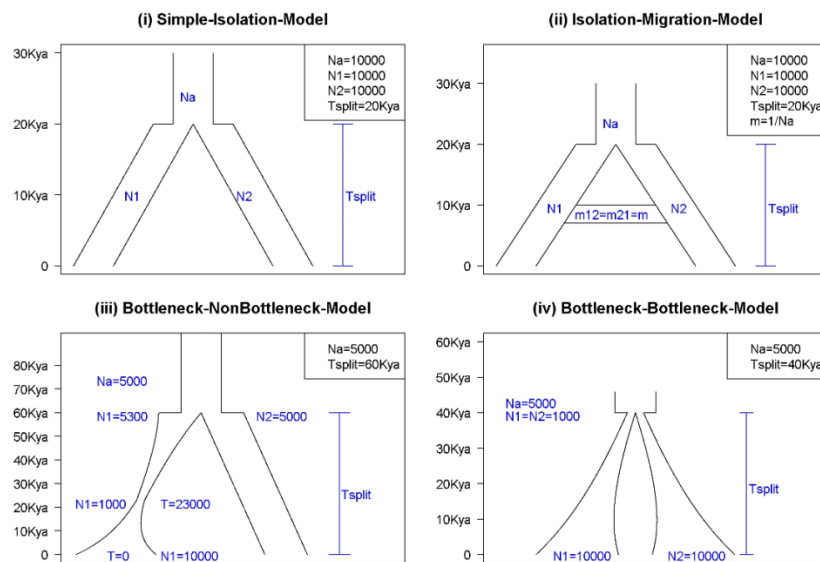
### 3.4 Simulation and Real Data Application

We perform a comparison of the eight methods for estimating the TMRCA (T-LD, T-FST, MIMAR, DADI, GPho-CS, CoalHMM, PSMC, and MSMC) to gauge their relative performance as measured by the robustness and accuracy of the TMRCA estimates. This is achieved through a series of simulations under four different population demography scenarios: (i) a simple-isolation model; (ii) an isolation-migration model; (iii) a bottleneck-nonbottleneck model; and (iv) a bottleneck-bottleneck model. The simple-isolation model is the simplest, which assumes a random mating ancestral population that splits instantaneously into two descendant populations with no subsequent gene flow. The isolation-migration model extends the simple-isolation model by allowing for migration after the population split. The bottleneck-nonbottleneck model simulates the demographic history of African and non-African populations, where studies have suggested the presence of demographic bottlenecks in non-African populations but not in African populations [22, 96,

98]. The bottleneck-bottleneck model simulates the demographic history of two non-African populations. These eight methods are subsequently applied to estimate the TMRCA between Southeast Asian Malays and South Asian Indians, with deep whole-genome sequencing data from these two populations.

### 3.4.1 Methods

#### 3.4.1.1 Simulating demographic models



**Figure 21. Illustration of the four demographic scenarios considered in our simulation study. An ancestral population diverged into two populations (population\_1 and population\_2) at time Tsplit. N1, N2 and Na are the effective population size of population\_1, population\_2 and the ancestral population, respectively. (i) simple-isolation-model: ancestral population split into two populations at 20Kya. (ii) isolation-migration-model: a symmetric migration rate is added after the split. (iii) bottleneck-nonbottleneck-model: ancestral population split into two populations at 60Kya after which population\_2 has constant effective population size and population\_1 experienced a bottleneck. (iv) bottleneck-bottleneck-model: ancestral population split into two populations at 40Kya, after which both population\_1 and population\_2 have population size declined instantly and afterwards increased exponentially.**

We simulated genetic sequences for two populations under four different demographic scenarios (Figure 21) with the ms program [35], where 10 iterations were generated for each scenario. In each iteration, 1001 sequences of length 10Mb are generated, comprising: one sequence from an outgroup

population, and 500 sequences each from the two target populations. Our simulations were specifically designed to evaluate the ability to estimate the TMRCA for two populations that diverged between 20,000 to 60,000 years ago, and we assumed the outgroup population to have diverged from the two target populations 4,100,000 years ago. We assumed a mutation rate per site per year of  $10^{-9}$ , a generation time of 25 years and a recombination rate of  $5 \times 10^{-9}$ . The four demographic models are: (i) Simple-isolation model: that assumed an ancestral population with an effectively population size ( $N_e$ ) of 10,000, which split into two populations 20,000 years ago with the same effective population size of 10,000; (ii) Isolation-migration model: that assumed the same set-up as the simple-isolation model except with the addition of migration (migration rate = 0.01%) between the two populations immediately after the split; (iii) Bottleneck-nonbottleneck model: that assumed an ancestral population with  $N_e = 5,000$ , which split into two populations 60,000 years ago such that one population has an  $N_e = 5,000$  and the other population has  $N_e$  declining exponentially from 5,300 to 1,000 at  $t = 23,000$  years ago, and increasing exponentially to 10,000 at present; (iv) Bottleneck-bottleneck model: that assumed an ancestral population with  $N_e = 5,000$ , which split into two populations 40,000 years ago such that both populations have an  $N_e = 1,000$  immediately after the split, and which increased exponentially to 10,000 at present. Our simulations produced an average of 98,175 SNPs in the simple-isolation model; 98,705 SNPs in the isolation-migration model; 57,677 SNPs in the bottleneck-nonbottleneck model; and 62,920 SNPs in the bottleneck-bottleneck model.

### 3.4.1.2 Estimating TMRCA of Southeast Asian Malays and South Asian

#### Indians with whole genome sequencing data

To estimate the TMRCA of Southeast Asian Malays and South Asian Indians, whole genome sequencing data for 96 Malays from the Singapore Sequencing Malay Project (SSMP) <sup>22</sup> and 36 Indians from the Singapore Sequencing Indian Project (SSIP) <sup>23</sup> were used. These individuals were sequenced on the Illumina HiSeq 2000 at a target depth of 30-fold, where the alignment and variant calling were performed with CASAVA and SAMtools for the Malay data, and with CASAVA and GATK for the Indians. The consensus calls were used as input for T-LD, T-FST, DADI and MIMAR; whereas PSMC, MSMC, GPho-CS and CoalHMM used the variant calls obtained from their individual analysis pipeline. For T-LD, T-FST and DADI, all 96 Malays and 36 Indians were used to estimate the TMRCA. To avoid any effect of uneven sample sizes, we randomly selected 36 Malays to match the 36 Indians for the analysis with MIMAR. For the analysis with PSMC, MSMC, CoalHMM and GPho-CS, one individual each from SSMP (SS6002734) and SSIP (SS6003427) were selected. The analyses were performed independently across 22 autosomal chromosomes, which were subsequently used to derive the mean and 95% confidence interval for the TMRCA estimate.

### 3.4.1.3 Analysis of TMRCA with T-LD and T-FST

In the estimation of the TMRCA between two populations, T-LD and T-FST consider genomic sites that are polymorphic in at least one of the two populations. To minimise the impact of ascertainment bias, the analyses are restricted to SNPs with MAFs  $\geq 5\%$  in the combined set of chromosomes from

both populations as suggested by McEvoy and colleagues [24]. In the four scenarios considered in the simulation, there were on average 29,038 common SNPs in the simple-isolation model; 29,166 common SNPs in the isolation-migration model; 6,772 common SNPs in the bottleneck-nonbottleneck model; and 7,980 common SNPs in the bottleneck-bottleneck model. For the TMRCA of Malays and Indians, there were 295,317 segregating sites shared by the Malays and Indians.

#### 3.4.1.4 Analysis of TMRCA with DADI

DADI estimates TMRCA from the allele frequency spectrum of the variants present in the genomic region. For the simulation study, 500 sequences from each of the two populations were used to derive the allele frequency spectrum, where the outgroup sequence was used to determine the original and derived alleles. We specified three grid sizes (100, 200, 300) to extrapolate to an infinitely fine grid, and we assumed the default setting with an isolation model in all the DADI analyses of the simulation data. In addition, we also applied DADI to the simulation data assuming the specific model setting for the different demographic scenarios, to evaluate how DADI will perform with prior knowledge of the underlying demographic history between the two populations. Specifically, for the bottleneck-nonbottleneck scenario in which two populations split at  $T$ , where the two populations subsequently have effective population sizes of  $Ne_I$  and  $Ne_s$ , where  $Ne_s$  decreases exponentially to  $Ne_b$  at  $T_b$  before increasing exponentially to  $Ne_f$  at present, the parameters ( $Ne_I$ ,  $Ne_s$ ,  $Ne_b$ ,  $Ne_f$ ,  $T_b$ ,  $T$ ) are estimated simultaneously. Similarly, for the bottleneck-bottleneck scenario, two populations split at  $T$ , and the  $i^{\text{th}}$

population has an effective population size of  $Ne_{ib}$  immediately after the split, which increases exponentially to  $Ne_{if}$  at present, for  $i = 1, 2$ , the parameters ( $Ne_{1b}, Ne_{1f}, Ne_{2b}, Ne_{2f}, T$ ) are simultaneously estimated.

For the analyses of the TMRCA between Malays and Indians, all 132 samples (96 Malays, 36 Indians) were used to compute the allele frequency spectrum across the 16,681,861 SNPs, where we ran two separate analyses assuming the isolation model and the bottleneck-bottleneck model. The bottleneck-bottleneck model adopted the same design as in the analysis of the simulation data, except the project sample size was bounded between 70 and 120, and three grid sizes (140, 180, 200) were used for extrapolation.

#### 3.4.1.5 Analysis of TMRCA with GPho-CS

GPho-CS considers neutral loci defined across multiple samples. For the simulations, we consider two haploid sequences from each population. The selected haploid sequences are divided into 10,000 segments each of length 1000 bases, and a constant population size was assumed. We assumed the absence of migration in three scenarios except that of the isolation-migration model where we ran GPho-CS with and without migration bands. For the analyses of the simulated data, a burn in of 100,000 steps and 200,000 samplings were chosen. For estimating the TMRCA of Malays and Indians, 37,563 neutral one kilobase loci were identified, which removed sites under selection, with low sequencing quality and poor alignment [87]. The filtering criteria included removing 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. GPho-CS was applied to five haploid sequences at multiple loci, which included two haplotype sequences from SS6002734, two haplotype sequences from SS6003427, and one chimpanzee reference sequence. The haploid sequences for the two human samples were phased using SHAPEIT [99] against the reference data from Phase 3 of the 1000 Genomes Project [100]. The chimpanzee reference haploid sequence was used to calibrate the mutation rate against the divergence time of 6.5 million years ago for human and chimpanzee, inferring an average mutation rate per site per year of  $6.96 \times 10^{-10}$ , which is consistent with the literature applying GPho-CS to estimate population divergence time [87].

#### 3.4.1.6 Analysis of TMRCA with MIMAR

In the analysis of the simulation data with MIMAR, we considered one hundred haploid sequences from each of the two populations, where the original and derived alleles were determined from the outgroup population. The selected sequences are segmented into regions each of length 1000 bases, where we selected 900 non-adjoining loci (1-1000 bp, 2001-3000 bp, 4001-5000bp, to 1,798,001-1,799,000 bp) for analysis, and further divided them into 30 subsets in order to control the acceptance rate of the MCMC process to be at least 5% as recommended. The MCMC was run with a burn-in of 100,000 runs, and where we recorded 300,000 samplings afterwards. The default demographic model assumed an isolation model that was applied to all four

scenarios in the simulation study, where we further assumed the scaled population mutation rates for the ancestral and two offshoot populations to be sampled from a Uniform[0.0001, 0.002] distribution. The population divergence time in generations was sampled from a Uniform[500, 3000] distribution for three of the four scenarios, except for the bottleneck-nonbottleneck scenario where the population divergence time in generations was assumed to be sampled from a Uniform[1000, 5000] distribution. Separately, we also applied MIMAR under the same demographic model used to simulate the data. Specifically, for the isolation-migration model, we added a prior for the logarithm of scaled migration record  $\log(4Nem)$  as a Uniform [-5, 3] distribution; for the bottleneck-nonbottleneck model, the population size was allowed to decrease exponentially between  $[T, 0.38 T]$  years ago at rate  $4.5 \times 10^{-5}$ , and increasing exponentially between  $[0.38 T, 0]$  years ago at rate  $1 \times 10^{-4}$ ; for the bottleneck-bottleneck model, the population size increased exponentially at a rate of  $5.8 \times 10^{-5}$  immediately after the split.

As MIMAR considers only neutral loci, for the estimation of the TMRCA for Malays and Indians we extracted 37,563 one kilobase loci following the filtering procedure as suggested by the analysis with GPho-CS [87]. As MIMAR is computationally expensive and cannot handle thousands of loci simultaneously, each chromosome is divided into subsets each containing 30 one-kilobase loci. Similarly we assumed a burn-in of 100,000 runs, recorded 300,000 samplings, with a Uniform[0.0001, 0.002] prior for the population mutation rates of the ancestral population and the two populations (Malay, Indian), and a divergence time in generations distributed as Uniform[500,

3000]. A point estimate is derived for each chromosome as the average across the subsets, and the mean divergence time and corresponding 95% confidence interval were obtained from the point estimates of the 22 autosomal chromosomes.

#### 3.4.1.7 Analysis of TMRCA with CoalHMM, PSMC and MSMC

CoalHMM and PSMC consider only two haploid sequences from the two populations. PSMC differs from all the other methods as it does not provide a point estimate for the TMRCA, instead it estimates the effective population size as a step function across time, and the TMRCA is qualitatively determined as the time point when the effective population size increases to infinity. We adopted an effective population size threshold of 100,000 to determine the TMRCA. MSMC is highly similar to PSMC, except that it allows multiple haploid sequences from a population to be considered, where we apply MSMC to two haploid sequences from each population. While MSMC does not provide a point estimate for TMRCA, it provides a “cross-coalescence rate” which measures the relative gene flow between two populations. This is similarly a step function across time, and takes values between 0 and 1. The cross-coalescence rate decreases from 1 to 0, which translates to a decline in gene flow between two populations. As with PSMC, the estimation of TMRCA from MSMC is qualitatively determined, and we adopted a cross-coalescence rate threshold of 0.5 to identify the TMRCA.

For the estimation of the TMRCA between Malays and Indians, the haploid sequences from the same two individuals (SS6002734, SS6003427) were

phased in the manner as described for the analysis with GPho-CS, and were analysed with CoalHMM and PSMC. The same effective population size threshold of 100,000 was used to determine a point estimate for the TMRCA. For the analysis by MSMC, all four phased sequences for the two individuals were used, and a cross-coalescence rate threshold of 0.5 was used to determine the point estimate for the TMRCA.

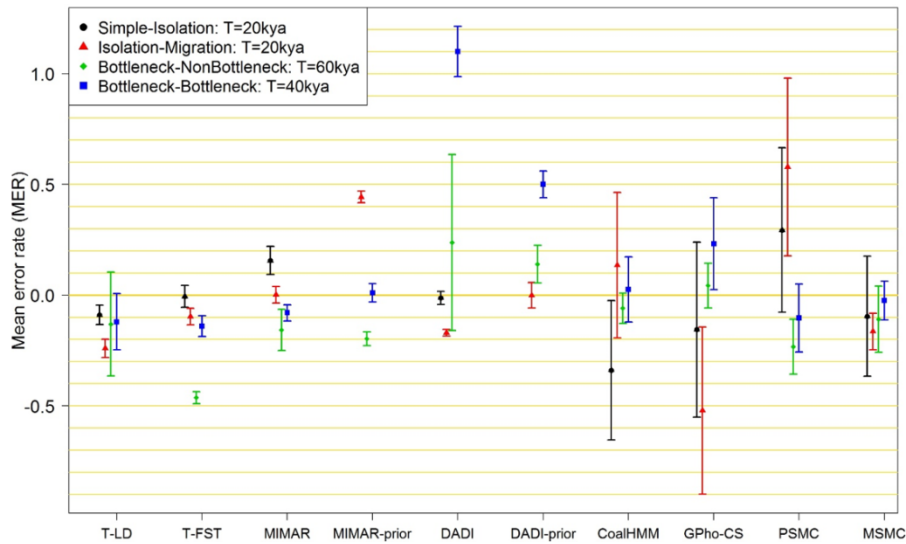
#### 3.4.1.8 Evaluating performance of TMRCA estimation

The estimation of the TMRCA by each of the eight methods is evaluated using the simulation data with two metrics: (i) the mean error rate (expressed in percentage); and (ii) the corresponding 95% confidence interval across the 10 iterations in each of the four demographic scenarios. The error rate for the  $i^{\text{th}}$  iteration is defined as  $\frac{T_i - T_0}{T_0} \times 100\%$ , and  $T_i$ ,  $i = 1, \dots, 10$  represents the TMRCA estimated in the  $i^{\text{th}}$  iteration, and  $T_0$  represents the simulated population divergence time.

All simulation data for the four demographic models, as well as the command line inputs and customized scripts for executing or implementing the eight methods are available for download at <http://www.statgen.nus.edu.sg/~TMRCA/>.

### 3.4.2 Results

#### 3.4.2.1 Comparisons of eight methods with simulations



**Figure 22. Mean error rate and 95% confidence interval are obtained from 10 iterations. Except MIMAR-prior and DADI-prior, the estimations are obtained with simple isolation model. MIMAR-prior and DADI-prior show the results obtained with prior knowledge of the demographic model for scenario (ii), (iii) and (iv).**

We compared the performance of the eight different methods for estimating TMRCA with 10 sets of simulated data from each of four demographic settings, that assumed a: (i) simple-isolation model; (ii) isolation-migration model; (iii) bottleneck-nonbottleneck model; and (iv) bottleneck-bottleneck model. The two simulated populations were designed to diverge 20,000 years ago for the simple-isolation and isolation-migration models; 60,000 years ago for the bottleneck-nonbottleneck model; and 40,000 years ago for the bottleneck-bottleneck model. The performance of the eight methods was then measured using two metrics: (i) the mean error rate (MER); and (ii) the corresponding 95% confidence interval (see Section 3.4.1 Methods for details), where a MER closer to zero with narrower confidence intervals spanning zero is more desirable, across all four scenarios.

We separated the evaluation of the eight methods according to the type of input data considered, such as: (i) genotyping data; (ii) sequencing data across tens of thousands of short loci; and (iii) whole-genome sequencing data.

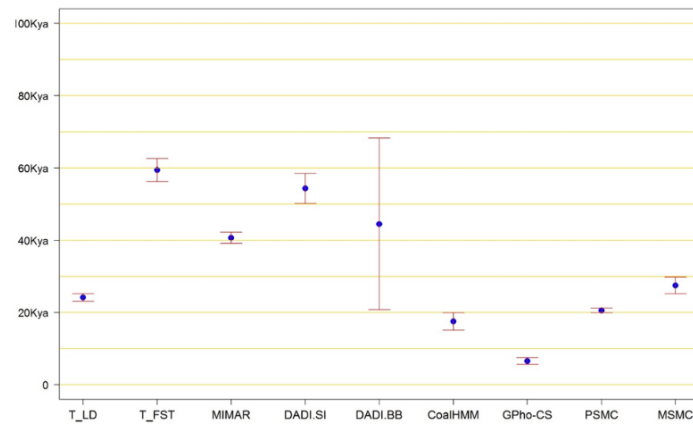
Three methods (T-LD, T-FST, DADI) are applicable when only chip-based genotyping data is available. We observed that T-FST and DADI yielded more accurate TMRCA estimations in the setting assuming a simple-isolation model between two populations (Figure 22), with the former exhibiting the lowest MER of -0.5% (95% CI: -5.5%, 4.4%) and the latter exhibiting a MER of -1.2% (95% CI: -4.2%, 1.8%). T-LD yielded a higher MER (-8.9%, 95% CI: -13.3%, -4.6%). However, in the setting assuming an isolation-migration model, all three methods performed poorly with moderate MERs 9.7%-24%) but with corresponding confidence intervals that were significantly distant to zero. In the setting assuming a bottleneck-nonbottleneck model, while all three models yielded MERs >10%, the confidence intervals for T-LD and DADI encapsulated zero, with that for T-LD narrower than that for DADI. T-FST yielded a significant underestimation of the TMRCA with a MER of -46.3%, and worryingly exhibited a tight 95% confidence interval (-48.9%, -43.6%). For the bottleneck-bottleneck scenario, only the 95% confidence interval from T-LD encapsulated zero MER, whereas DADI yielded a gross overestimation of the TMRCA (MER = 110.0%, 95%CI: 98.6%, 121.4%). In an ideal situation where DADI was implemented knowing what the underlying demographic model was, the error rates and the variability of the TMRCA estimations were reduced, although this did not yield estimates that were close to the true TMRCA except for the isolation-migration model.

When sequence data is available for short regions in the genome, GPho-CS produced TMRCA estimates with moderate error rates for three scenarios (except the isolation-migration model)(MERs<24%), where the corresponding confidence intervals for the simple-isolation and bottleneck-nonbottleneck encapsulated zero MER (Figure 22). Another MCMC-based approach, MIMAR, yielded relatively smaller MER and variability than GPho-CS (MERs<16%), although the estimates tend to be consistently over (simple-isolation) or under (bottleneck-nonbottleneck, bottleneck-bottleneck). Intriguingly, implementing MIMAR with prior knowledge of the underlying demographic model yielded considerably poorer estimates for the isolation-migration and bottleneck-nonbottleneck scenarios, and only improved the estimate for the bottleneck-bottleneck scenario.

For the three HMM-based methods that allow whole-genome sequence data, CoalHMM and MSMC yielded comparable performance where each of the two methods yielded confidence intervals that encapsulated zero for three scenarios and where the corresponding MERs were also small. CoalHMM appeared to be most uncertain in the simple-isolation model, whereas MSMC performed poorer in the isolation-migration scenario. Compared to these two methods, PSMC exhibited greater variability and MERs across all four demographic models.

### 3.4.2.2 Estimating TMRCA of Southeast Asian Malays and South Asian

#### Indians



**Figure 23. Illustrate the point estimation and corresponding 95% confidence interval of TMRCA for Southeast Asian Malays and South Asian Indians by the eight methods. DADI.SI and DADI.BB show the estimates of DADI with isolation model and bottleneck-bottleneck-model respectively.**

The eight methods were applied to whole-genome sequencing data for 96 Southeast Asian Malays and 36 South Asian Indians, where data from the 22 autosomal chromosomes were analyzed independently by each of the eight methods and combined subsequently to derive the mean and 95% confidence intervals of the estimates (Figure 23). DADI was implemented assuming both the simple-isolation model (DADI.SI) and the bottleneck-bottleneck model (DADI.BB). The analyses with the different methods yielded a broad range of TMRCA estimates, with GPho-CS reporting the lowest estimate of 6,594 (95% CI: 5,652, 7,537) years ago (ya), to T-FST reporting the highest estimate of 59,429 ya (95% CI: 56,242, 62,615). Our previous simulation results suggested that T-LD, CoalHMM and MSMC were likely to yield the most robust estimates regardless of the underlying demographic model, and it was reassuring that the TMRCA estimates for Malays and Indians from these three methods were comparable (T-LD = 24,173ya, CoalHMM = 17,546ya, MSMC = 27,508ya). PSMC also yielded a comparable estimate of 20,715ya (95% CI:



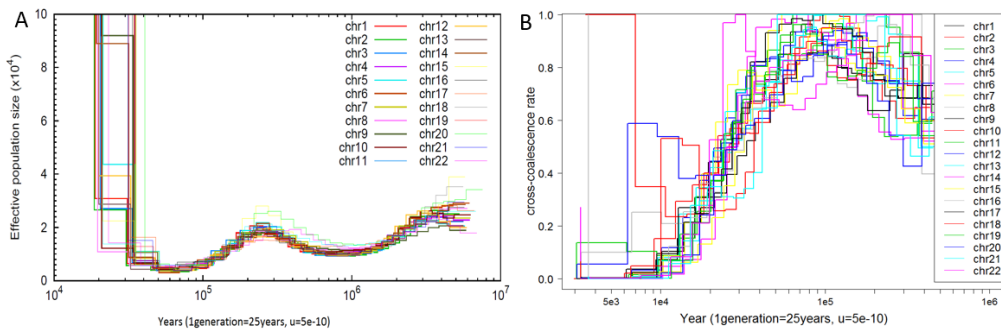
20,011, 21,419), whereas the remaining methods yielded estimates exceeding 30kya.

### **3.5 DISCUSSION**

Estimating the TMRCA between two populations has always been a topic of great interest in population genetics, and there are presently a number of methods that leverage on different genetic features and are built on a variety of statistical frameworks to perform this estimation. We set out to compare the accuracy and robustness of eight of these methods with a series of simulations that assumed different underlying demography between two diverged populations. The results of our simulations suggested that T-LD, CoalHMM and MSMC were more likely to deliver estimates that were robust to a variety of background demography. The consistency in performance and accuracy across different demographic models is important, as often one does not know a priori what the underlying demographic model between two populations will be. The high variability in the TMRCA estimates observed in either the simulations or the analysis of the Malay and Indian data by some of the methods (such as DADI and GPho-CS) is worrying, as this suggests that the derived point estimates by these methods are susceptible to fluctuations even though the independent inputs were essentially from the different chromosomes of the same individuals.

In general, HMM-based methods tend to be more computationally efficient compared to MCMC-based methods. For example, the analysis of the Malay and Indian whole-genome sequencing data using HMM-based methods such

as CoalHMM, PSMC and MSMC can be completed in hours on a standard Linux-based processor, whereas MCMC-based methods such as MIMAR and GPho-CS required several days to a few weeks to complete the same analysis across 22 chromosomes. The computational burden also means that MCMC-based methods could not model recombination effectively, and the analysis was necessarily restricted to short segments. Conversely, the computational dexterity of HMM-based approaches allows both recombination events to be modelled and for full chromosomal data to be analyzed.



**Figure 24. Illustrate the estimation of TMRCA by (A) PSMC and (B) MSMC on whole-genome sequencing data for the 22 autosomal chromosomes from Southeast Asian Malays and South Asian Indians. Both the effective population size (panel A) and the cross-coalescence rate (panel B) are modelled as step functions. The divergence time for the two populations is defined for (A) PSMC as the time when the effective population size increases to infinity, which in practice is implemented as a threshold such as 100,000 in our study; (B) MSMC as the most recent time when the cross-coalescence rate decreases below an arbitrarily selected threshold, which in our study the threshold is selected as 0.5.**

A key challenge in the implementation of PSMC and MSMC is in the selection of the thresholds for the effective population size and cross-coalescence rate respectively to determine divergence time (Figure 4).

Presently there are no recommended or default thresholds for these two approaches, and the TMRCA estimates are sensitive to the choice of the thresholds. For example, the TMRCA estimate for the PSMC analysis of the Malay and Indian data changes from 20,715 ya to 36,824 ya, if the threshold on the effective population size changes from 1,000,000 to 50,000.

GPho-CS produced a considerably lower TMRCA estimate for the Malay and Indian whole-genome sequencing data, and this may be due to two reasons: (i) GPho-CS has previously reported lower accuracy to infer recent events [87]; and (ii) GPho-CS relied on a different mutation rate. Presently, the method calibrates the mutation rate from the number of mutation events from an outgroup species to which the divergence time has to be assumed [87]. By including a chimpanzee sequence in the model and assuming the divergence time from chimpanzee to be 6.5 Mya, this produced an average mutation rate of  $6.96 \times 10^{-10}$ , which is only 70% of the default mutation rate of  $10^{-9}$  for the chimpanzees. While this may be a reasonable calibration given the exclusion of CpG and regions under selection, this is based on the assumption that chimpanzees and modern humans exhibited identical mutation rates per site per year and generation time. A recent study suggested revising the mutation rate to  $5 \times 10^{-10}$  per site per year for studies on modern human evolution [53], which was the value we have used for the genome-wide average mutation rate. As such, a comparable mutation rate for neutral sites should thus be lower than  $5 \times 10^{-10}$ . Assuming we scaled the mutation rate used in GPho-CS to be correspondingly 70% of  $5 \times 10^{-10}$ , this would produce a point estimate for the TMRCA as 13,188 ya (95% CI: 11,304, 15,074). However, this highlights the dependency that TMRCA estimation has on the parameters assumed.

We have evaluated eight statistical methods commonly used in population genetics to estimate TMRCA. The performance of these methods varies

according to the parameter settings assumed, as well as the background demographic model producing the split of the two populations. Our simulations have considered only four relatively simple demographic scenarios, and certainly these are not exhaustive and in no way representative of the complex migration and demographic changes populations undergo in reality. The effective population size is confounding in TMRCA analysis, and an accurate effective population size is crucial for estimating divergence time. Among those methods, DADI, PSMC and MSMC have higher resolution in effective population size. Worryingly, the divergence time estimates of these methods did not always concur. On the basis of our findings, we recommend the use of T-LD, CoalHMM and MSMC for estimating TMRCA with genotyping and whole-genome sequencing data respectively.

## **CHAPTER 4. CONCLUSION**

Microarray genotyping data has dominated genetic research in the last decade.

Mature data analysis protocols and techniques have been established.

Because of the International HapMap Project, large scale genotyping data covering almost all common SNPs in human of most major populations are available for research communities and facilitates relevant research fields such as genome-wide association studies (GWAS) and population genetics. With the advent of next generation sequencing, favorable attention is being drawn towards variants found in the lower frequency spectrum. Burgeoning whole genome sequencing studies provide fine scale genetic data and the majority of the newly discovered variants tend to be rarer in the population. Customized microarrays designed for follow-up studies tend to cover lower frequency variants that are identified through population sequencing efforts. Existing genotyping algorithms are typically designed for common SNPs, and as such the lower frequency variants present significant challenges for the existing genotype calling algorithms.

Our method, iCall, serves as a robust genotyping algorithm for common, low-frequency and rare variants and yields an accuracy at least comparable to, if not better than, existing methods. Accurate genotype calling across the allele frequency spectrum is meaningful for all downstream genetic researches not only for population evolution study, but more importantly for genome-wide association study and personalized molecular diagnostics. Successful molecular diagnostics applications include Human Papillomavirus (HPV) genotyping assay for high-risk and low risk HPV genes [101]; the Cystic

Fibrosis and Hydatidiform Mole genotyping assay for adults of reproductive age in newborn screening [102, 103]; apolipoprotein E genotyping assay for Alzheimer's disease [104]. These clinical applications tend to focus on rare and highly hereditary disorders. They are empirical evidence supporting the common disease rare variant etiology theory, which emphasizes the importance of robust genotyping algorithm for a wide allele frequency spectrum, especially the lower end.

Although there is now a higher proportion of low frequency variants on customized microarrays, microarrays do have some intrinsic limitations. Microarrays are pre-designed according to prior knowledge of the queried genome, and limited in the number of SNPs on each platform. Microarrays thus suffer from ascertainment bias, namely arrays are ineffective in the case of incomplete or outdated genome annotations [105]. The advent of next generation sequencing has mitigated this problem. NGS does not require prior knowledge of the genome, but directly and comprehensively sequences the whole region or whole genome, so it accurately profiles the genome of the individual. Because of NGS, haplotype information of all types of polymorphism on individual genomes has become more affordable and accessible.

Population genetics research infers population history based on heterozygosity, allele frequency, LD and pattern of genetic variation which can be changed by a variety of genetic and demographic forces. Hence fine scale sequencing data of full allele frequency spectrum is one prerequisite for

population genetics research. Microarray-based methods underutilize the genomic information and often have poor statistical power. The sequencing-based methods and whole genome sequencing data for estimating population divergence times while taking into account the historical recombination events have only been available recently. Population parameters, such as population divergence time, effective population size, recombination rate and migration rate are important factors for understanding common historical features in the diversification of human populations. Our work comparing the existing methods for estimating population divergence time provides a reference for future work for elucidating global population structure as well as population evolution history.

#### **4.1 Future work for iCall**

Population stratification and genotype errors are two known confounding factors in association study [106]. Our iCall algorithm is a population-based genotyping method which requires large sample sizes to locate the intensity clouds. Combining datasets from multiple populations for joint calling creates large dataset, but could also introduce calling errors. We have assumed Hardy-Weinberg Equilibrium in iCall, but HWE is not necessarily true for a combined dataset. For example, a SNP that is fixed in allele A in population 1 and fixed in allele B in population 2 presents in genotype AA and BB but not AB in the combined population. Thus iCall could be improved by jointly evaluating population information in its calling process. We could explore a Bayesian hierarchical framework to depict the genotype polymorphism of multiple populations. By using a Dirichlet distribution to model the frequency

of minor alleles in different populations, we might allow variation in allele frequencies across populations.

#### **4.2 Future Work for TMRCA**

Apart from the population divergence time, the recombination rate and the effective population size are two important parameters in population genetics, but their correlation can confound the joint inference of the two parameters. Therefore, accurate TMRCA estimation relies on robust estimations of effective population size and recombination rate. More evaluations on the robustness of the existing methods on other parameters are necessary to boost our confidence in the estimation. To obtain a better understanding about the global population structure, population divergence time among global major populations as well as rare populations is worth exploring through existing robust TMRCA algorithms and will provide interesting insight into human genetic evolution.



## REFERENCE

1. Cohen, S.N., et al., *Construction of biologically functional bacterial plasmids in vitro*. Proc Natl Acad Sci U S A, 1973. **70**(11): p. 3240-4.
2. Maxam, A.M. and W. Gilbert, *A new method for sequencing DNA*. Proc Natl Acad Sci U S A, 1977. **74**(2): p. 560-4.
3. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors*. Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
4. International HapMap, C., et al., *A second generation human haplotype map of over 3.1 million SNPs*. Nature, 2007. **449**(7164): p. 851-61.
5. International HapMap, C., *The International HapMap Project*. Nature, 2003. **426**(6968): p. 789-96.
6. International HapMap, C., et al., *Integrating common and rare genetic variation in diverse human populations*. Nature, 2010. **467**(7311): p. 52-8.
7. Griffith, O.L., et al., *ORegAnno: an open-access community-driven resource for regulatory annotation*. Nucleic Acids Res, 2008. **36**(Database issue): p. D107-13.
8. Information, H.G.P., *How Many Genes Are There?* U.S. Department of Energy Office of Science, 2008.
9. Buchanan, C.C., et al., *A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data*. J Am Med Inform Assoc, 2012. **19**(2): p. 289-94.
10. Rockman, M.V. and L. Kruglyak, *Genetics of global gene expression*. Nat Rev Genet, 2006. **7**(11): p. 862-72.
11. Rapley, R. and S. Harbron, *Molecular analysis and genome discovery*. 2nd ed. 2011, Chichester, West Sussex: John Wiley & Sons. p.
12. Affymetrix. *Affymetrix Genome-Wide Human SNP Array 5.0* [http://media.affymetrix.com/support/technical/datasheets/genomewide\\_snp5\\_datasheet.pdf](http://media.affymetrix.com/support/technical/datasheets/genomewide_snp5_datasheet.pdf). 2007; Available from: [http://media.affymetrix.com/support/technical/datasheets/genomewide\\_snp5\\_datasheet.pdf](http://media.affymetrix.com/support/technical/datasheets/genomewide_snp5_datasheet.pdf).
13. Affymetrix. *GeneChip Human Mapping 10K Array Xba 142 2.0* [http://media.affymetrix.com/support/technical/datasheets/10k2\\_datasheet.pdf](http://media.affymetrix.com/support/technical/datasheets/10k2_datasheet.pdf). 2004; Available from: [http://media.affymetrix.com/support/technical/datasheets/10k2\\_datasheet.pdf](http://media.affymetrix.com/support/technical/datasheets/10k2_datasheet.pdf).
14. Affymetrix, *Genome-Wide Human SNP Array 6.0* [http://media.affymetrix.com/support/technical/datasheets/genomewide\\_snp6\\_datasheet.pdf](http://media.affymetrix.com/support/technical/datasheets/genomewide_snp6_datasheet.pdf). 2009.
15. Affymetrix, ed. *GeneChip Human Mapping 100K Set* [http://media.affymetrix.com/support/technical/datasheets/100k\\_datasheet.pdf](http://media.affymetrix.com/support/technical/datasheets/100k_datasheet.pdf). 2004.
16. Affymetrix. *GeneChip Human Mapping 500K Array Set* [http://media.affymetrix.com/support/technical/datasheets/500k\\_datasheet.pdf](http://media.affymetrix.com/support/technical/datasheets/500k_datasheet.pdf). 2005; Available from: [http://media.affymetrix.com/support/technical/datasheets/500k\\_datasheet.pdf](http://media.affymetrix.com/support/technical/datasheets/500k_datasheet.pdf).

17. illumina. *The Omni Family of Microarrays*  
[http://www.illumina.com/documents/products/datasheets/datasheet\\_gwas\\_roadmap.pdf](http://www.illumina.com/documents/products/datasheets/datasheet_gwas_roadmap.pdf). 2010; Available from:  
[http://www.illumina.com/documents/products/datasheets/datasheet\\_gwas\\_roadmap.pdf](http://www.illumina.com/documents/products/datasheets/datasheet_gwas_roadmap.pdf).
18. illumina. *Infinium HD Assay*  
<http://www.illumina.com/technology/beadarray-technology/infinium-hd-assay.html>. 2015; Available from:  
<http://www.illumina.com/technology/beadarray-technology/infinium-hd-assay.html>.
19. Ritchie, M.E., et al., *BeadArray expression analysis using bioconductor*. PLoS Comput Biol, 2011. **7**(12): p. e1002276.
20. Clayton, D.G., et al., *Population structure, differential bias and genomic control in a large-scale, case-control association study*. Nat Genet, 2005. **37**(11): p. 1243-6.
21. Tavaré, S., et al., *Inferring coalescence times from DNA sequence data*. Genetics, 1997. **145**(2): p. 505-18.
22. Reich, D.E., et al., *Linkage disequilibrium in the human genome*. Nature, 2001. **411**(6834): p. 199-204.
23. Education, S.b.N.; Available from:  
<http://www.nature.com/scitable/definition/haplotype-haplotypes-142>.
24. McEvoy, B.P., et al., *Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs*. Genome Res, 2011. **21**(6): p. 821-9.
25. Tishkoff, S.A., et al., *Global patterns of linkage disequilibrium at the CD4 locus and modern human origins*. Science, 1996. **271**(5254): p. 1380-7.
26. de Roos, A.P., et al., *Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle*. Genetics, 2008. **179**(3): p. 1503-12.
27. Lewontin, R.C., *The Interaction of Selection and Linkage. Ii. Optimum Models*. Genetics, 1964. **50**: p. 757-82.
28. Lewontin, R.C., *The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models*. Genetics, 1964. **49**(1): p. 49-67.
29. Lee, S., et al., *Rare-variant association analysis: study designs and statistical tests*. Am J Hum Genet, 2014. **95**(1): p. 5-23.
30. Holsinger, K.E. and B.S. Weir, *Genetics in geographically structured populations: defining, estimating and interpreting F(ST)*. Nat Rev Genet, 2009. **10**(9): p. 639-50.
31. Rosenberg, N.A., et al., *Genetic structure of human populations*. Science, 2002. **298**(5602): p. 2381-5.
32. Lewontin, R.C., *The apportionment of human diversity*. Evolutionary biology, 1972. **6**: p. 381-398.
33. Wright, S., *The genetical structure of populations*. Ann Eugen, 1951. **15**(4): p. 323-54.
34. Cockerham, C.C., *Variance of gene frequencies*. Evolution, 1969. **23**: p. 72-84.
35. Hudson, R.R., *Generating samples under a Wright-Fisher neutral model of genetic variation*. Bioinformatics, 2002. **18**(2): p. 337-8.

36. B. S. Weir, C.C.C., *Estimating F-Statistics for the Analysis of Population Structure*. *Evolution*, 1984. **38**(6): p. 1358-1370.
37. Cockerham, C.C., *Analyses of gene frequencies*. *Genetics*, 1973. **74**(4): p. 679-700.
38. Cavalli-Sforza, L.L., P. Menozzi, and A. Piazza, *The history and geography of human genes*. 1994, Princeton, N.J.: Princeton University Press. xi, 541, 518 p.
39. Mellars, P., *A new radiocarbon revolution and the dispersal of modern humans in Eurasia*. *Nature*, 2006. **439**(7079): p. 931-5.
40. Cann, R.L., M. Stoneking, and A.C. Wilson, *Mitochondrial DNA and human evolution*. *Nature*, 1987. **325**(6099): p. 31-6.
41. Peng, B., C.I. Amos, and M. Kimmel, *Forward-time simulations of human populations with complex diseases*. *PLoS Genet*, 2007. **3**(3): p. e47.
42. Ewens, W.J., *Mathematical population genetics*. 2nd ed. Interdisciplinary applied mathematics. 2004, New York: Springer. v. <1- >.
43. Vigilant, L., et al., *African populations and the evolution of human mitochondrial DNA*. *Science*, 1991. **253**(5027): p. 1503-7.
44. Prufer, K., et al., *The complete genome sequence of a Neanderthal from the Altai Mountains*. *Nature*, 2014. **505**(7481): p. 43-9.
45. Macaulay, V., et al., *Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes*. *Science*, 2005. **308**(5724): p. 1034-6.
46. Rasmussen, M., et al., *An Aboriginal Australian genome reveals separate human dispersals into Asia*. *Science*, 2011. **334**(6052): p. 94-8.
47. Pinhasi, R., et al., *The genetic history of Europeans*. *Trends Genet*, 2012. **28**(10): p. 496-505.
48. Reich, D., et al., *Reconstructing Native American population history*. *Nature*, 2012. **488**(7411): p. 370-4.
49. Burenhult, G., *Die ersten Menschen*. 2000: Weltbild Verlag.
50. Fu, Y.X. and W.H. Li, *Estimating the age of the common ancestor of men from the ZFY intron*. *Science*, 1996. **272**(5266): p. 1356-7; author reply 1361-2.
51. Griffiths, R.C., *Lines of descent in the diffusion approximation of neutral Wright-Fisher models*. *Theor Popul Biol*, 1980. **17**(1): p. 37-50.
52. Donnelly, P., *Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles*. *Theor Popul Biol*, 1986. **30**(2): p. 271-88.
53. Scally, A. and R. Durbin, *Revising the human mutation rate: implications for understanding human evolution*. *Nat Rev Genet*, 2012. **13**(10): p. 745-53.
54. Kimura, M., *Evolutionary rate at the molecular level*. *Nature*, 1968. **217**(5129): p. 624-6.
55. Frankham, R., *Effective population size/adult population size ratios in wildlife: a review*. *Genet Res*, 2007. **89**(5-6): p. 491-503.
56. Mathieson, I. and G. McVean, *Differential confounding of rare and common variants in spatially structured populations*. *Nat Genet*, 2012. **44**(3): p. 243-6.

57. illumina, *Genotyping Rare Variants: a simulated analysis achieves high call rates and low error rates from loci containing rare variants* [http://www.illumina.com/Documents/products/technotes/technote\\_genotyping\\_rare\\_variants.pdf](http://www.illumina.com/Documents/products/technotes/technote_genotyping_rare_variants.pdf). 2010.
58. Giannoulatou, E., et al., *GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population*. *Bioinformatics*, 2008. **24**(19): p. 2209-14.
59. illumina. *BeadArray Microarray Technology* <http://www.illumina.com/technology/beadarray-technology.html>. 2015; Available from: <http://www.illumina.com/technology/beadarray-technology.html>.
60. Steemers, F.J., et al., *Whole-genome genotyping with the single-base extension assay*. *Nat Methods*, 2006. **3**(1): p. 31-3.
61. Lin, Y., et al., *Smarter clustering methods for SNP genotype calling*. *Bioinformatics*, 2008. **24**(23): p. 2665-71.
62. Zhou, J., et al., *iCall: a genotype-calling algorithm for rare, low-frequency and common variants on the Illumina exome array*. *Bioinformatics*, 2014. **30**(12): p. 1714-20.
63. Di, X., et al., *Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays*. *Bioinformatics*, 2005. **21**(9): p. 1958-63.
64. Nicolae, D.L., et al., *GEL: a novel genotype calling algorithm using empirical likelihood*. *Bioinformatics*, 2006. **22**(16): p. 1942-7.
65. Rabbee, N. and T.P. Speed, *A genotype calling algorithm for affymetrix SNP arrays*. *Bioinformatics*, 2006. **22**(1): p. 7-12.
66. S. Cawley<sup>1</sup>, X.D., E. Hubbell<sup>1</sup>, S. Lincoln<sup>1</sup>, M. Moorhead<sup>1</sup>, W. Short<sup>1</sup>, T.P. Speed<sup>2,3</sup>, C. Sugnet<sup>1</sup>, J. Veitch<sup>1</sup>, T. Webster<sup>1</sup>, A. Williams<sup>1</sup>, G. Yang, *BRLMM: an improved genotype calling method for the genechip human mapping 500k array*. Affymetrix White Paper Publication, 2006.
67. Shah, T.S., et al., *optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants*. *Bioinformatics*, 2012. **28**(12): p. 1598-603.
68. Teo, Y.Y., et al., *A genotype calling algorithm for the Illumina BeadArray platform*. *Bioinformatics*, 2007. **23**(20): p. 2741-6.
69. Li, G., et al., *M(3): an improved SNP calling algorithm for Illumina BeadArray data*. *Bioinformatics*, 2012. **28**(3): p. 358-65.
70. Xiao, Y., et al., *A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays*. *Bioinformatics*, 2007. **23**(12): p. 1459-67.
71. Goldstein, J.I., et al., *zCall: a rare variant caller for array-based genotyping: genetics and population analysis*. *Bioinformatics*, 2012. **28**(19): p. 2543-5.
72. Ionita-Laza, I., et al., *Sequence kernel association tests for the combined effect of rare and common variants*. *Am J Hum Genet*, 2013. **92**(6): p. 841-53.
73. Neale, B.M., et al., *Testing for an unusual distribution of rare variants*. *PLoS Genet*, 2011. **7**(3): p. e1001322.
74. Wu, M.C., et al., *Powerful SNP-set analysis for case-control genome-wide association studies*. *Am J Hum Genet*, 2010. **86**(6): p. 929-42.

75. Wu, M.C., et al., *Rare-variant association testing for sequencing data with the sequence kernel association test*. *Am J Hum Genet*, 2011. **89**(1): p. 82-93.
76. Wong, L.P., et al., *Deep whole-genome sequencing of 100 southeast Asian Malays*. *Am J Hum Genet*, 2013. **92**(1): p. 52-66.
77. Wong, L.P., et al., *Insights into the genetic structure and diversity of 38 South Asian Indians from deep whole-genome sequencing*. *PLoS Genet*, 2014. **10**(5): p. e1004377.
78. Jukes TH, C.C., *Evolution of protein molecules*. *Mammalian Protein Metabolism*, 1969: p. 21-132.
79. Felsenstein, J., *Evolutionary trees from DNA sequences: a maximum likelihood approach*. *J Mol Evol*, 1981. **17**(6): p. 368-76.
80. Xu Shuhua, J.W., *Population Genetics in the Genomics Era, Studies in Population Genetics*. 2012: InTech.
81. McVean, G.A. and N.J. Cardin, *Approximating the coalescent with recombination*. *Philos Trans R Soc Lond B Biol Sci*, 2005. **360**(1459): p. 1387-93.
82. Marjoram, P. and J.D. Wall, *Fast "coalescent" simulation*. *BMC Genet*, 2006. **7**: p. 16.
83. Kingman, J.F., *Origins of the coalescent. 1974-1982*. *Genetics*, 2000. **156**(4): p. 1461-3.
84. Donnelly, P. and S. Tavaré, *Coalescents and genealogical structure under neutrality*. *Annu Rev Genet*, 1995. **29**: p. 401-21.
85. 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.
86. Hobolth, A., et al., *Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model*. *PLoS Genet*, 2007. **3**(2): p. e7.
87. Gronau, I., et al., *Bayesian inference of ancient human demography from individual genome sequences*. *Nat Genet*, 2011. **43**(10): p. 1031-4.
88. 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.
89. Li, H. and R. Durbin, *Inference of human population history from individual whole-genome sequences*. *Nature*, 2011. **475**(7357): p. 493-6.
90. Schiffels, S. and R. Durbin, *Inferring human population size and separation history from multiple genome sequences*. *Nat Genet*, 2014. **46**(8): p. 919-25.
91. 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.
92. Gutenkunst, R.N., et al., *Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data*. *PLoS Genet*, 2009. **5**(10): p. e1000695.
93. Hill, W.G. and A. Robertson, *Linkage disequilibrium in finite populations*. *Theor Appl Genet*, 1968. **38**(6): p. 226-31.
94. Nei, M., *Molecular evolutionary genetics*. 1987, New York: Columbia University Press. x, 512 p.

95. Sved, J.A., *Linkage disequilibrium and homozygosity of chromosome segments in finite populations*. Theor Popul Biol, 1971. **2**(2): p. 125-41.
96. 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.
97. 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.
98. Plagnol, V. and J.D. Wall, *Possible ancestral structure in human populations*. PLoS Genet, 2006. **2**(7): p. e105.
99. Delaneau, O., J. Marchini, and J.F. Zagury, *A linear complexity phasing method for thousands of genomes*. Nat Methods, 2012. **9**(2): p. 179-81.
100. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes*. Nature, 2012. **491**(7422): p. 56-65.
101. Kroupis, C. and N. Vourlidis, *Human papilloma virus (HPV) molecular diagnostics*. Clin Chem Lab Med, 2011. **49**(11): p. 1783-99.
102. Lin, X., et al., *Microtransponder-based multiplex assay for genotyping cystic fibrosis*. Clin Chem, 2007. **53**(7): p. 1372-6.
103. Vang, R., et al., *Diagnostic reproducibility of hydatidiform moles: ancillary techniques (p57 immunohistochemistry and molecular genotyping) improve morphologic diagnosis*. Am J Surg Pathol, 2012. **36**(3): p. 443-53.
104. Mayeux, R., et al., *Utility of the apolipoprotein E genotype in the diagnosis of Alzheimer's disease*. Alzheimer's Disease Centers Consortium on Apolipoprotein E and Alzheimer's Disease. N Engl J Med, 1998. **338**(8): p. 506-11.
105. Hurd, P.J. and C.J. Nelson, *Advantages of next-generation sequencing versus the microarray in epigenetic research*. Brief Funct Genomic Proteomic, 2009. **8**(3): p. 174-83.
106. Miclaus, K., et al., *Batch effects in the BRLMM genotype calling algorithm influence GWAS results for the Affymetrix 500K array*. Pharmacogenomics J, 2010. **10**(4): p. 336-46.