Stony Brook University

Mixture Modeling of Next Generation Sequencing Data and its Applications to Genotyping and Estimating Genotype Frequencies

A Dissertation Presented

by

Jayon Lihm

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Concentration – Statistics)

Stony Brook University

December 2013

Stony Brook University

The Graduate School

Jayon Lihm

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Stephen J. Finch – Dissertation Advisor
Professor, Department of Applied Mathematics and Statistics

Nancy R. Mendell - Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Song Wu
Assistant Professor, Department of Applied Mathematics and Statistics

Seungtai Yoon
Research Assistant Professor, Cold Spring Harbor Laboratory

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Mixture Modeling of Next Generation Sequencing and its Applications to Genotyping and Estimating Genotype Frequencies

by

Jayon Lihm

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Concentration – Statistics)

Stony Brook University

2013

Estimating the probability that an individual has a base pair nucleodite different from the reference nucleotide is important in next generation sequencing (NGS) research. I present a method for modeling the frequency of single nucleotide polymorphism variants in the exome capturing sequence data of an individual. A mixture distribution was used to model the proportion of alternative alleles at a specified base pair position assuming a biallelic single nucleotide polymorphism model. I measured the proportion of alternative alleles for positions in chromosome 1 exome sequencing data fro two trios taken from the Pilot 3 data in the 1000 Genomes Project. The measurements were based on the counts of reference and alternative alleles calculated by the SAMtools genetic software. The mixture model studied here had two point distributions and five continuous distributions. I applied the expectation-maximization algorithm to obtain the maximum likelihood estimates of the mixture model parameters for each

individual. The fitted mixture model well described the properties of the distribution of the alternative allele proportions. The estimates of mixing proportions were used to estimate the genotype frequencies in the data. Each individual had different estimates of model parameters, but the estimates of genotype fractions of the six individuals were similar. The estimated fractions of the members from each trio were similar to each other. I next combined two approaches of clustering and mixture modeling to genotype the exomic base pair positions of an individual using next generation sequencing data. The alternative allele proportion at a position was used to measure the Bayesian posterior probability of single nucleotide polymorphism at a position. I developed software package named "SNVclust" to generate alternative allele proportions and genotypes of an individual. This software was used to make a call set of single nucleotide polymorphism positions and genotypes for each of three members of a trio from the 1000 Genomes Project. The results from this software were compared with the released single nucleotide polymorphisms in the 1000 Genomes Project and results from two other programs. Then I found that minimal average coverage greater than 43 should be to use SNVclust for whole exome sequencing data.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

1KG – 1000 Genomes Project

AD – Adjusted Depth

AAP – Alternative Allele Proportion

AMB – Ambiguous genotypes

BAM – Binary Alignment/Map

BWA – Burrow-Wheeler Alignment

CEU – Utah residents with Northern and Western European ancestry from the CEPH collection

CDF – Cumulative Distribution Function

CLARA – Clustering Large Applications

CV – Coefficient of Variation

EM – Expectation-Maximization

GATK – Genomic Analysis Toolkit

HPCC – High Performance Cluster Computers

HWE – Hardy-Weinberg Equilibrium

MA – Minor Allele

MACOUNT – Count of Minor Allele

MAPQ – Mapping Quality

MLE – Maximum Likelihood Estimate

NGS – Next Generation Sequencing

PAM – Partitioning Around Medoids

PCR – Polymerase Chain Reaction

PDF – Probability Density Function

PM – Possible Multiallelic

PREF – Proabability of having a Reference genotype

PSNP – Probability of being a SNP

Q-Q – Quantile-Quantile

SAM – Sequence Alignment/Map

SNP – Single Nucleotide Polymorphism

SNV – Single Nucleotide Variant

TiTv – Transition versus Transversion

WES – Whole Exome Sequencing

WGS – Whole Genome Sequencing

YRI – Yoruba Population in Ibadan, Nigeria

Acknowledgments

I would like to deeply thank my doctoral advisor, Dr. Stephen J. Finch. Dr. Finch has been an inspiring and generous mentor since I started my research with him four years ago. Without his support and encouragement, I could not have finished this rewarding journey. I would like to express my gratitude to Dr. Seungtai Chris Yoon who gave me great opportunities to work on various research topics. Dr. Yoon helped me to generate the data used throughout this dissertation and consistently gave me a lot of valuable ideas and motivations. Many thanks go to my committee members, Dr. Nancy R. Mendell and Dr. Song Wu for proofreading and making insightful comments.

I specially thank my parents, Joowhan Lihm and Soonkyun Kim, for loving me wherever I am and continuously supporting my study. I wish this dissertation be as meaningful to my parents as it is to me. Also thanks to my brother, Jayso Lihm, who has always been my great buddy. I am grateful to have my parents-in-law, Bogu Kang and Younhee Park, who believed in me and sent me warm-hearted words.

To my dearest friends in Stony Brook who always supported me, stood on my side and cheered me up whenever I needed it, without them it would have been much harder to finish my study: Ruiqi Zhang, Aram Kim, Bora Park, Jing Jin, Eunjung Lim, Unjung Lee, and Soyoun Lee.

Lastly, I cannot thank my loving husband, Kyoungmo Kang, enough. You have always been my greatest inspiration and encouragement. Special thanks to my little Din.

Chapter 1 Introduction

1.1 Research Background

Many types of variants exist in the genome, ranging from single nucleotide variation to structural variants including insertions and deletions, copy number variations, inversions and translocations. The single nucleotide polymorphism (SNP) is one of the most abundant [1] and widely studied genomic variation. A SNP is a single nucleotide change in the DNA sequence of an individual compared to those of other individuals. Recent improvements in next-generation sequencing (NGS) technologies have enabled more reliable identifications of such genomic variations.

NGS is a parallelized sequencing process producing many millions of short sequences of nucleotides called reads. Each read consists of four nucleotides, A, T, C or G. Dealing with so many reads is one of major tasks in NGS data analysis. There are two major approaches to analyzing data from NGS, resequencing and de novo assembly [2]. In de novo assembly, reads are assembled to construct the genome sequence based on overlapping parts of reads. In the resequencing approach, each read is mapped and aligned to the known reference genome. Due to the limited length of reads, the resequencing approach has been used in the majority of NGS research. Considerable bioinformatics software has been developed for resequencing to detect genomic variants. There are software programs developed for mapping to the reference genome such as MAQ (Mapping and Assembly with Quality) [3], BWA (Burrows-Wheeler Aligner) [4], SOAP (Short Oligonucleotide Analysis Package) [5], and Bowtie [6]. Software such as SAMtools (Sequence Alignment/Map Tools) [7], VarScan [8], and GATK (Genomic Analysis

ToolKit) [9] are used to detect SNP or other variants such as copy number variations, short insertions, or deletions.

1.2 Bayesian Formulation for Genotyping

The Bayesian formulation is a useful statistical method for genotyping and discovering SNP positions [10]. Many variant calling algorithms take a Bayesian approach to calculating genotype likelihoods with different settings of prior and posterior probabilities. Let *G* be a genotype and *D* be the data obtained. Then the posterior probability is defined to be

$$p(G|D) = \frac{p(G)p(D|G)}{p(D)} \qquad (1.1)$$

Many existing software programs combine information from sequencing to estimate the prior and posterior probabilities. GATK [9, 11] calculates genotype likelihoods with the base quality scores expressed in the Phred scale [12, 13]. The Phred scaled quality score, q, is logarithmically related to base-calling error probability e as $q = -10\log_{10} e$. The *p(D|G)* in Equation 1.1 is calculated as

$$p(b|A) = \begin{cases} \frac{e}{3}; & b \neq A \\ 1-e; & b = A \end{cases} \qquad (1.2)$$

where *A* is one of the alleles in a genotype, *b* is the base observed at the position, and *e* is the reversed base quality score from the Phred scale. The SOAPsnp [14] uses different information to construct the prior probability. The program uses the ratio of transition versus transversion

(TiTv)[1] to calculate genotype likelihoods. Using the dbSNP database, the prior probability of seeing a specific allele with given reference is calculated in advance by the Ti/Tv ratio.

Some software programs only consider three genotypes from the two most evident alleles out of four nucleotides rather than considering all ten possible genotypes. For example, the MAQ program [3] uses the two most frequent alleles at a position. It assumes that the prior probability *p(G)* is the known proportion of genotypes. The probability of having heterozygotes at a position is set to be 0.001 for new SNP sites and 0.2 for known SNP sites. The SNVMix program [15] also assumes three genotypes. It models the genotype likelihoods using a binomial mixture. Out of total *n* reads covering a position, the probability of having "*a*" non-reference alleles is assumed to follow a binomial distribution as below:

$$p(a|G) \sim \mathrm{Binom}(a|\mu_k, n) \qquad (1.3)$$

where μ_k (*k=1,2,3*) is expected to be 0 for homozygous reference genotype (*k=1*), 0.5 for heterozygous genotype (*k=2*) and 1 for homozygous alternative genotype (*k=3*).

1.3 Alternative Allele Proportion

[1] Four nucleotides A, G, T, and C are categorized as purine or pyrimidine based on their

The raw reads are mapped to the human genome reference sequence. The information from the mapped reads is contained in a Binary Alignment Map (BAM) file format. A BAM file includes information about a sequence read such as start position, end position, mapping quality score and base quality score. SAMtools generates a Pileup file from the BAM file sorted by genomic positions [Table 1.1]. For each position, SAMtools summarizes the number of reads covering the position (depth), its reference allele, a set of nucleotides sequenced there, and a set of base scores represented in ASCII code [Table 1.2] corresponding to each nucleotide. The number of times that each of nucleotides appeared in the pileup file at the position is counted. The sequence column in the pileup file used "." and "," if a sequenced base matches with the reference sequence, and one of "A(or a)", "C(or c)", "T (or t)", or "G (or g)" if a sequenced base is different from the reference sequence. The letter "N (or n)" appears when the sequenced base is too ambiguous to be specified. The signs "+" and "-" are for when insertions and deletions occurred, respectively. Also "*" sign is marked for a base-pair deletion. Each of the letters and the signs were counted. For example, the first row (Chormosome 1 Position 861,207) shows that three reads covered the position. All three reads contained the reference allele at the position (three dots), and each of them has base score letter "D", "D", and "5", respectively. Table 1.2 shows that ASCII code "D" corresponds to a decimal number 68 and ASCII code "5" corresponds to a decimal number 53. Base quality score is a corresponding decimal number minus 33. Thus, the three base score quality scores at this position are 35, 35, and 20 for each read.

Then the proportion of the alternative alleles (AAP) at a position is defined as

$$\frac{the\ count\ of\ the\ alternative\ allele}{the\ sum\ of\ the\ reference\ allele\ and\ the\ alternative\ allele}.$$

The alternative allele is determined as the one having the greatest count among A, G, T, or C. When there are multiple alternative alleles with the same greatest count, one of them is used for calculating AAP at the position. Table 1.3 shows a part of the allele counts in the table. The table consists of sixteen columns; chromosome, position, reference allele at the position, depth at the position, the counts of reference allele, counts of A, C, T, G, a base-pair deletion (DEL), insertion (PLUS) and deletion (MINUS), adjusted depth (AD), the count of minor allele (MACOUNT), and possible minor allele (MA). The AAP can be calculated as $\frac{MACOUNT}{AD}$ from this table.

Table 1.1. An Example of a Pileup File

CHROM	POS	REF	DEPTH	Nucleotides	Base Quality Score
1	861207	C	3	^].^].^].	DD5
1	861208	G	8	...^].^].^].^].^].	BA@DA4B5
1	861209	T	8	.....	FFEDBBBD
1	861210	C	8	.....	KKJJEHFI
1	861211	C	13	.....^].^].^].^].^].	LLGKGIHK'55D=
1	861212	A	14	.....^].	IIHGCDGHFFDFC?
1	861213	C	15	.....^].	KKJJGIHJGDHIFCD
1	861214	G	22	.....^].^].^].^].^].^].^>,	FFFFCEBEDDDE@5A4DDB55.
1	861215	A	26	.....,^].^].^].^].	IIIIADBIHHHHAEGFBECEE.DD55
1	861216	G	30	.....t....^]C^].^].^].	NNNNGJJNMMMNJ7MKLLHKK,JJDJ$/AB
1	865518	T	38	,.....,...c.,...,.c.....^].	GEEF>DGDIJGJJJC:IE#J?<MM3F*MMILLLKKLBD

Table 1.2 ASCII Code to Decimal Number Table

Dec	ASCII	Dec	ASCII	Dec	ASCII	Dec	ASCII	Dec	ASCII	Dec	ASCII
33	!	48	0	64	@	80	P	96	`	112	p
34	"	49	1	65	A	81	Q	97	a	113	q
35	#	50	2	66	B	82	R	98	b	114	r
36	$	51	3	67	C	83	S	99	c	115	s
37	%	52	4	68	D	84	T	100	d	116	t
38	&	53	5	69	E	85	U	101	e	117	u
39	'	54	6	70	F	86	V	102	f	118	v
40	(	55	7	71	G	87	W	103	g	119	w
41	)	56	8	72	H	88	X	104	h	120	x
42	*	57	9	73	I	89	Y	105	i	121	y
43	+	58	:	74	J	90	Z	106	j	122	z
44	,	59	;	75	K	91	[	107	k	123	{
45	-	60	<	76	L	92	\	108	l	124	|
46	.	61	=	77	M	93	]	109	m	125	}
47	/	62	>	78	N	94	^	110	n	126	~
		63	?	79	O	95	_	111	o		

Note: The source for the table is the web site, http://www.asciichart.com.

Table 1.3. An Example of Counting Alternative Alleles

CHROM	POS	REF	DEPTH	REFCOUNT	A	C	T	G	DEL	N	PLUS	MINUS	AD	MACOUNT	MA
1	861216	G	30	28	0	1	1	0	0	0	0	0	29	1	CT
1	861225	G	65	64	1	0	0	0	0	0	0	0	65	1	A
1	861226	G	68	67	0	1	0	0	0	0	0	0	68	1	C
1	861227	G	70	69	0	0	1	0	0	0	0	0	70	1	T
1	861228	G	70	69	0	1	0	0	0	0	0	0	70	1	C
1	861230	A	77	76	0	0	0	1	0	0	0	0	77	1	G
1	861234	G	93	91	0	1	0	0	0	1	0	0	92	1	C
1	861250	C	140	139	1	0	0	0	0	0	0	0	140	1	A
1	861256	G	157	155	0	2	0	0	0	0	0	0	157	2	C
1	865518	T	38	36	0	2	0	0	0	0	0	0	38	2	C

Three genotypes are present in data describing a specified biallelic position: homozygous for the reference allele (genotype 0 for Ref/Ref, called g0 here), heterozygous with reference allele and alternative allele (genotype 1 for Ref/Alt, g1), and homozygous for the alternative allele (genotype 2 for Alt/Alt, g2). Under the biallelic model, the alternative allele is set as the one with the greatest count other than the reference allele. The proportion of the count of the alternative allele at this biallelic position is expected to be 0, 0.5, and 1, respectively for each genotype. After allowing for sequencing errors, the measured AAPs should have a probability distribution with modes near 0, 0.5, and 1.

1.4 Research Objectives

I propose to use the proportion of the alternative allele for describing data at a position. This measure was used in Morin et al. [16] for SNP detection if the following three conditions were met: the denominator is greater than or equal to 6; there are at least 2 non-reference allele reads; and the fraction of the non-reference allele reads is greater than or equal to 33%. This approach, however, cannot assign a probability to each genotype. In chapter 2, I use mixture distributions of the alternative allele proportions from one individual based on exponential and reversed exponential distributions. Initially I start with four components and increase the number of components to seven including two point mass distributions. The Expectation-Maximization (EM) algorithm is used to find the maximum likelihood estimates (MLE) of parameters of each component distribution. Chromosome 1 data (in BAM file) of a CEU trio and a YRI trio from the 1000 Genomes Project Phase 3 is used for modeling. Data information is in Table 1.4. This data

is exome capturing data. Then I apply my fitted model to estimate each genotype fraction in each individual.

Table 1.4 Description of the Two Trios

	ID
CEU-mother	NA12892
CEU-father	NA12891
CEU-child	NA12878
YRI-mother	NA19238
YRI-father	NA19239
YRI-child	NA19240

Second, I propose a SNP calling and genotyping algorithm based on the results of the mixture model. Many of existing algorithms model the prior and the posterior probability based on heuristic information such as base quality score and mapping quality score. Here my objective in chapter 3 is to calculate genotype likelihood using the mixture distribution of the AAP data in a computationally efficient way. Initially I use clustering algorithm CLARA (Clustering Large Applications) [17] to group the AAP data into three components. The clustering method CLARA is based on the clustering algorithm PAM (Partitioning Around Medoids) [17], which is a realization of k-medoids algorithm [18], but is designed to deal with large amount of data. The CLARA algorithm uses subsets of the data set, and each subset is partitioned into a given number of clusters using the PAM algorithm. Each observation is assigned to the cluster containing the nearest medoid. Medoids are representative objects of a data set or a cluster with a data set whose average dissimilarity to all the objects in the cluster is minimal. Medoids are similar in concept to means or centroids, but medoids are always members of the data set (http://en.wikipedia.org/wiki/Medoid). The software for CLARA is implemented in the R

package “cluster” [19]. The CLARA algorithm [19] is applied in a partitioned way suitable for this data set. After the clustering, the measured AAP data excluding 0 and 1 is transformed into the logit scale and modeled by a mixture of normal distributions. A major application of my algorithm is to model the AAP values from the whole exome sequencing. I applied the algorithm to the whole exome capturing of the YRI trio generated in [20] and compared its results with those from GATK and SAMtools.

Chapter 2 Mixture Modeling

2.1 Statistical Models

2.1.1 Model 1: Four Components Mixture

The measured proportions from a set of reads at a biallelic SNP should have a probability distribution with modes near 0, 0.5, and 1 after allowing for sequencing errors. Figure 2.1 is the histogram of AAP values over the entire chromosome 1 for six individuals. The left panel is the histogram of measured AAP values and the right panel is the histogram of the AAPs greater than 0.1, with red line drawn at 0.5. It shows that the AAP values near 0.5 are not symmetric. I observed that the frequency of the measured AAP values near 0 decreased approximately at an exponential rate. Thus my initial model is to use a mixture of four distributions: an exponential distribution for g0 AAP values (Equation 2.2 for component *0*), a "reversed exponential" started at 0.5 for g1 AAP values to the left of 0.5 (Equation 2.3 for component *1l*), an exponential started at 0.5 for g1 AAP values from the right side of 0.5 (Equation 2.4 for component *1r*), and a reversed exponential started at 1 for g2 AAP values (Equation 2.5 component *2*). Note that two separate distributions are used for g1 AAP values.

Let x_i represent a single measured value of AAP, $i = 1, \dots, n$. The probability density function (pdf) of x_i is given by

$$f_{model1}(x_i|\boldsymbol{\theta}_{model1}) = \sum_{j \in J_{model1}} \pi_j f_j(x_i). \qquad (2.1)$$

Here $J_{model1} = \{0, 1l, 1r, 2\}$ and the vector of parameters $\boldsymbol{\theta}_{model1} = (\mu_0, \mu_{1l}, \mu_{1r}, \mu_2, \pi_0, \pi_{1l}, \pi_{1r}, \pi_2)$. The mixing proportions π_j are $0 < \pi_j < 1$ and $\sum_j \pi_j = 1$ for $j \in J_{model1}$. The pdf of each component is given hbbelow:

$$f_0(x_i) = C_0 \frac{1}{\mu_0} e^{-\frac{x_i}{\mu_0}}, \quad 0 \le x_i \le 1 \tag{2.2}$$

$$f_{1l}(x_i) = C_{1l} \frac{1}{\mu_{1l}} e^{-\frac{(0.5-x_i)}{\mu_{1l}}}, \quad 0 \le x_i \le 0.5 \tag{2.3}$$

$$f_{1r}(x_i) = C_{1r} \frac{1}{\mu_{1r}} e^{-\frac{(x_i-0.5)}{\mu_{1r}}}, \quad 0.5 < x_i \le 1 \tag{2.4}$$

$$f_2(x_i) = C_2 \frac{1}{\mu_2} e^{-\frac{(1-x_i)}{\mu_2}}, \quad 0 \le x_i \le 1 \tag{2.5}$$

where μ_j is a positive real number in each distribution and C_j is a scaling constant such that

$$C_0 = \frac{1}{1-e^{-\frac{1}{\mu_0}}}, C_{1l} = \frac{1}{1-e^{-\frac{0.5}{\mu_{1l}}}},$$

$$C_{1r} = \frac{1}{e^{-\frac{0.5}{\mu_{1r}}} - e^{-\frac{1}{\mu_{1r}}}}, \text{and } C_2 = \frac{1}{1-e^{-\frac{1}{\mu_2}}}.$$

Figure 2.1. Histograms of AAP Values for Six Individuals

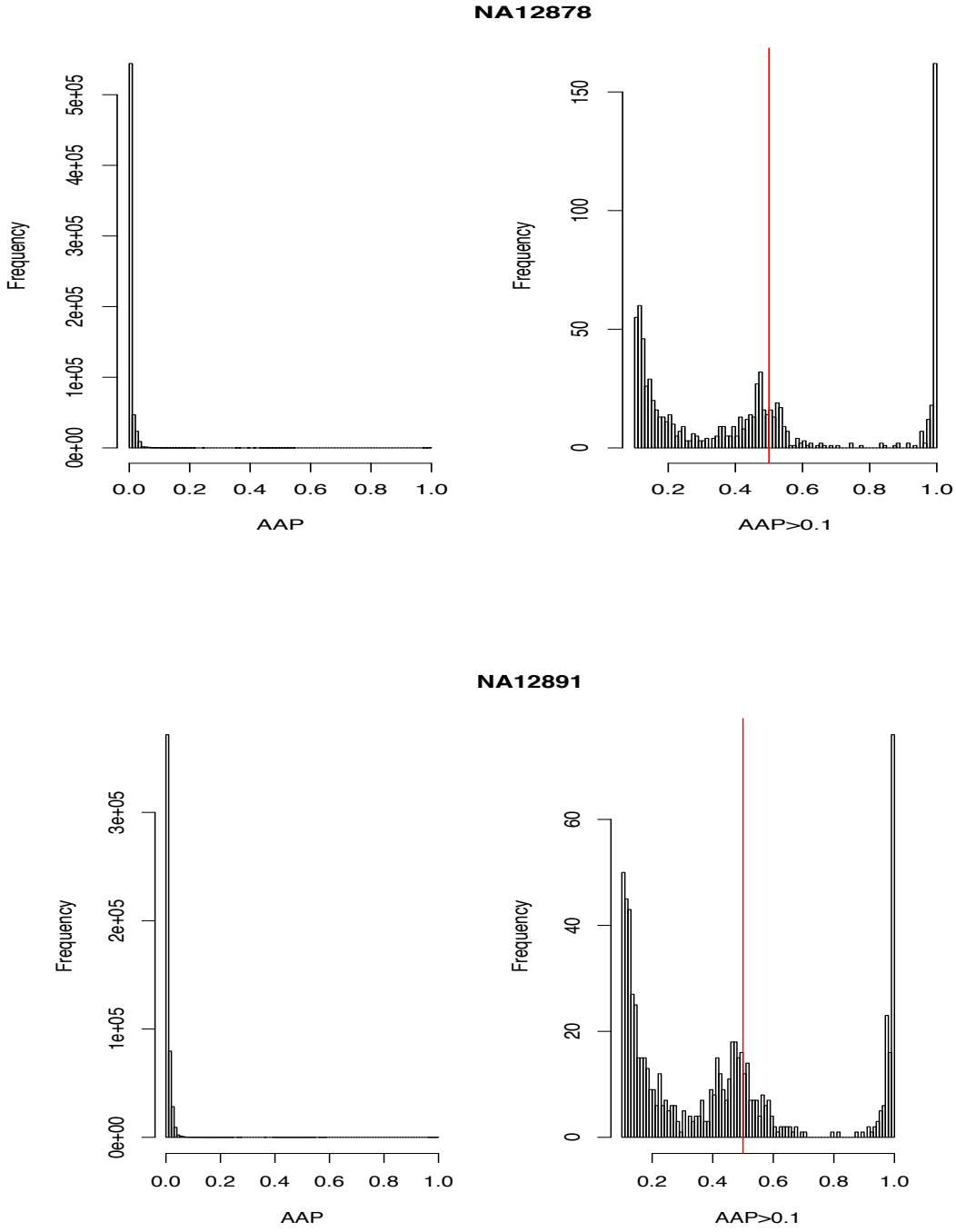


Figure 2.1 (continued)

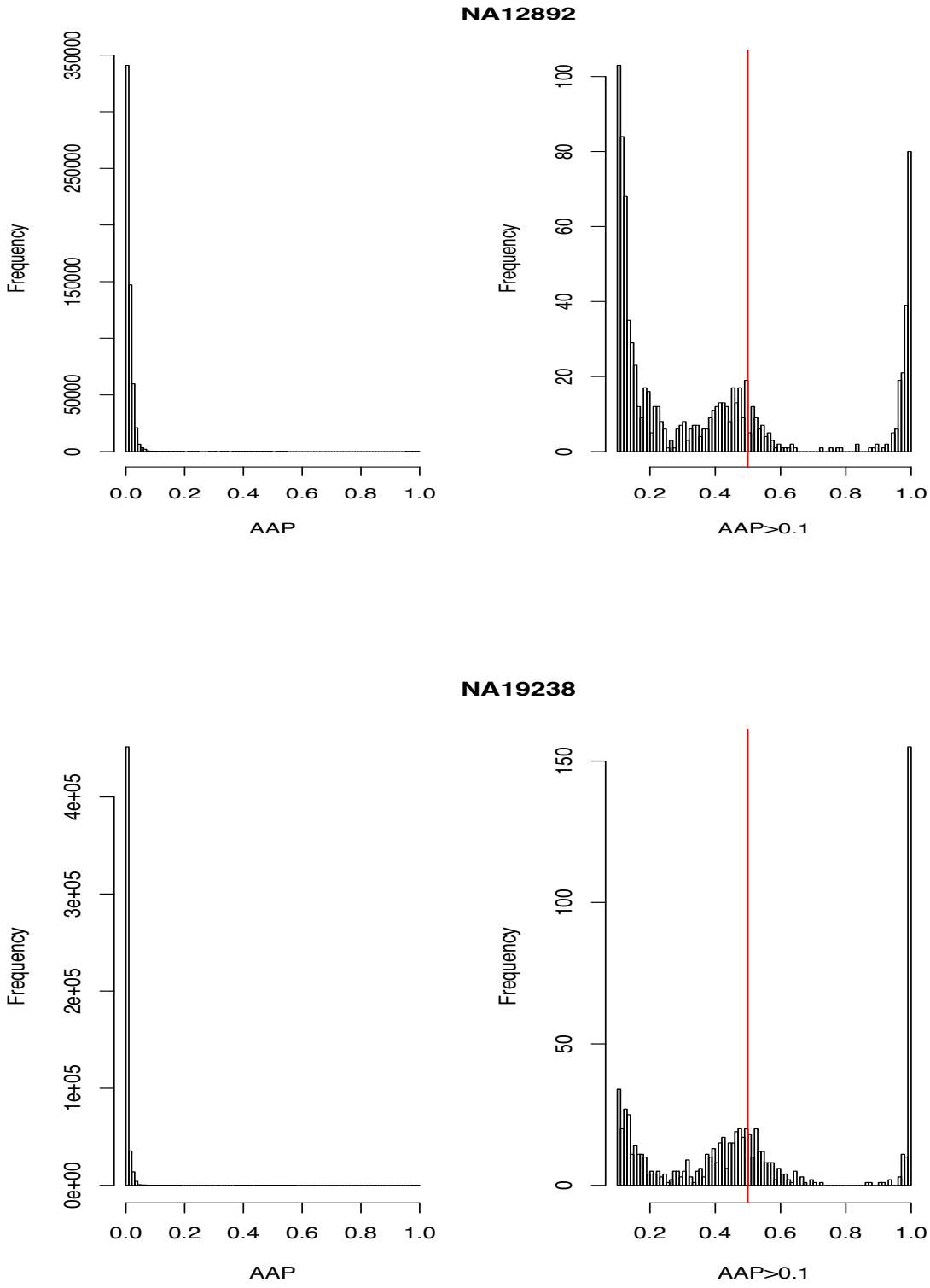
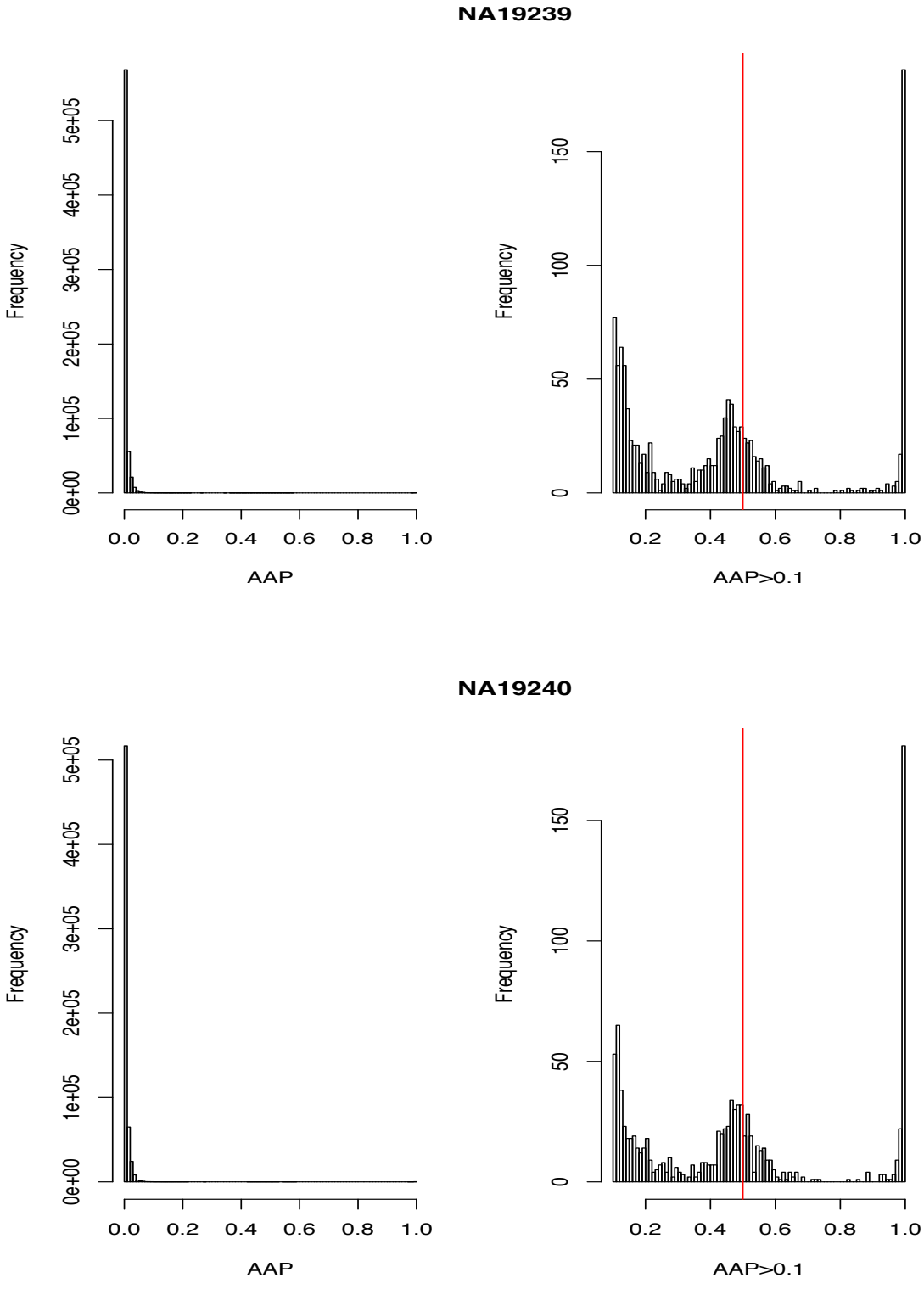


Figure 2.1 (continued)



2.1.2 Model 2: Five Components Mixture

The right panel in Figure 2.1 shows that there were more AAP values between 0.1 and 0.25 than would be expected from a single exponential distribution. For this reason, I added an extra component of an exponential distribution started at 0. The initial four components model is extended to five components model. Now $J_{model2} = \{01, 02, 1l, 1r, 2\}$ and the pdf of x_i is given by

$$f_{model2}(x_i|\boldsymbol{\theta}_{model2}) = \sum_{j \in J_{model2}} \pi_j f_j(x_i) \tag{2.6}$$

where the vector $\boldsymbol{\theta}_{model2} = (\mu_{01}, \mu_{02}, \mu_{1l}, \mu_{1r}, \mu_2, \pi_{01}, \pi_{02}, \pi_{1l}, \pi_{1r}, \pi_2)$ and π_j are $0 < \pi_j < 1$ and $\sum_j \pi_j = 1$ for $j \in J_{model2}$. Components 01 and 02 are for AAP values from the g0 genotype, components 1l and 1r are for AAP values from g1, and component 2 is for AAP values from g2 genotype. The pdf of each component is

$$f_{01}(x_i) = C_{01}\frac{1}{\mu_{01}}e^{-\frac{x_i}{\mu_{01}}}, \quad 0 \le x_i \le 1 \tag{2.7}$$

$$f_{02}(x_i) = C_{02}\frac{1}{\mu_{02}}e^{-\frac{x_i}{\mu_{02}}}, \quad 0 \le x_i \le 1 \tag{2.8}$$

$$f_{1l}(x_i) = C_{1l}\frac{1}{\mu_{1l}}e^{-\frac{(0.5-x_i)}{\mu_{1l}}}, \quad 0 \le x_i \le 0.5 \tag{2.9}$$

$$f_{1r}(x_i) = C_{1r}\frac{1}{\mu_{1r}}e^{-\frac{(x_i-0.5)}{\mu_{1r}}}, \quad 0.5 < x_i \le 1 \tag{2.10}$$

$$f_{2}(x_i) = C_{2}\frac{1}{\mu_{2}}e^{-\frac{(1-x_i)}{\mu_{2}}}, \quad 0 \le x_i \le 1 \tag{2.11}$$

where μ_j is a positive real number in each distribution, μ_{02} is set to be greater than μ_{01}, and C_j is a scaling constant such that

$$C_{01} = \frac{1}{1-e^{-\frac{1}{\mu_{01}}}}, C_{02} = \frac{1}{1-e^{-\frac{1}{\mu_{02}}}},$$

$$C_{1l} = \frac{1}{1-e^{-\frac{0.5}{\mu_{1l}}}}, C_{1r} = \frac{1}{e^{-\frac{0.5}{\mu_{1r}}} - e^{-\frac{1}{\mu_{1r}}}}, \text{and } C_2 = \frac{1}{1-e^{-\frac{1}{\mu_2}}}.$$

The parameter μ_{02} is set to be greater than μ_{01}.

2.1.3 Model 3: Seven Components Mixture

Table 2.1 shows the proportion of positions having the AAP values in each interval and the cumulative proportion. Each row describes an individual. The upper row for an individual is the percentage of the AAP values of chromosome 1 in the interval given, and the lower row is the cumulative percentage of the AAP values in the interval given. On average, 582,036 positions are included in the analysis and more than 99.8% of positions have AAP values less than or equal to 0.25. Table 2.1 shows that the proportion of AAP values at 0 and 1 is very high across six individuals ranging from 32.30% to 71.35%. Point distributions at zero and one (i.e., $I_{\{0\}}$ and $I_{\{1\}}$) were added in a mixed continuous-discrete manner [21] as the frequencies of the values far exceeded the number expected from exponential random variables. The pdf of x_i is

$$f_{model3}(x_i|\tilde{\boldsymbol{\theta}}_{model3}) = \alpha I_{\{0\}}(x_i) + \beta I_{\{1\}}(x_i) + (1-\alpha-\beta)\sum_{j\in J_{model3}} \pi_j f_j(x_i). \qquad (2.12)$$

Here $J_{model3} = \{01, 02, 1l, 1r, 2\}$, $I_{\{0\}}(x_i)$ is the indicator function having 1 if $x_i = 0$ and 0 otherwise, and $I_{\{1\}}(x_i)$ is the indicator function having 1 if $x_i = 1$ and 0 otherwise. The parameters α and β are probabilities of AAP values being zero and one, respectively, such that

$0 < \alpha, \beta < 1$ and $0 < \alpha + \beta < 1$. The vector of parameters $\widetilde{\boldsymbol{\theta}}_{model3}$ is given by $\widetilde{\boldsymbol{\theta}}_{model3} = (\alpha, \beta, \mu_{01}, \mu_{02}, \mu_{1l}, \mu_{1r}, \mu_2, \pi_{01}, \pi_{02}, \pi_{1l}, \pi_{1r}, \pi_2)$. Each of f_j is given in the same formula as in Equations 2.7, 2.8, 2.9, 2.10, and 2.11.

Table 2.1. Distribution of the AAP values from six individuals

Sample	# of positions	[0]	(0, 0.005]	(0.005, 0.25]	(0.25, 0.5]	(0.5, 0.75]	(0.75, 1)	[1]
NA12878 (CEU-child)	629,184	57.1434%	18.5143%	24.2563%	0.0361%	0.0164%	0.0130%	0.0205%
		57.1434%	75.6577%	99.9140%	99.9501%	99.9665%	99.9795%	100.0000%
NA12891 (CEU-father)	494,201	52.7267%	5.8871%	41.3008%	0.0389%	0.0186%	0.0150%	0.0130%
		52.7267%	58.6138%	99.9146%	99.9535%	99.9721%	99.9870%	100.0000%
NA12892 (CEU-mother)	583,253	32.2913%	6.3331%	61.2966%	0.0374%	0.0105%	0.0209%	0.0103%
		32.2913%	38.6244%	99.9210%	99.9583%	99.9688%	99.9897%	100.0000%
NA19238 (YRI-mother)	507,363	71.3276%	8.0873%	20.4773%	0.0461%	0.0252%	0.0104%	0.0260%
		71.3276%	79.4149%	99.8922%	99.9383%	99.9635%	99.9740%	100.0000%
NA19239 (YRI-father)	658,510	49.4758%	25.1105%	25.2954%	0.0580%	0.0254%	0.0156%	0.0193%
		49.4758%	74.5863%	99.8817%	99.9397%	99.9651%	99.9807%	100.0000%
NA19240 (YRI-child)	619,703	48.7109%	20.2024%	30.9753%	0.0489%	0.0255%	0.0171%	0.0198%
		48.7109%	68.9133%	99.8887%	99.9375%	99.9630%	99.9801%	100.0000%

2.2 Methods

2.2.1 Maximum Likelihood Estimation

The EM algorithm [22] is used to obtain MLEs of the parameters. A common practice described in [23] is to use a complete data setting by introducing the latent variables z_{ij}:

$$z_{ij} = \begin{cases} 1, & \text{if } x_i \text{ is from the component } j \\ 0, & \text{otherwise} \end{cases}$$

where $j \in J$.

Let $\underline{\boldsymbol{\theta}} = (\mu_j, \pi_j), j \in J$. For Models 1 and 2, the likelihood function is

$$L(\underline{\boldsymbol{\theta}}|\underline{\mathbf{x}}) = \prod_{i=1}^{n} \prod_{j \in J} \{\pi_j f_j(x_i|\underline{\boldsymbol{\theta}})\}^{z_{ij}} \quad (2.13)$$

and its log-likelihood function is

$$l(\underline{\boldsymbol{\theta}}|\underline{\mathbf{x}}) = \sum_{i=1}^{n} \sum_{j \in J} z_{ij} \{\log \pi_j + \log f_j(x_i|\underline{\boldsymbol{\theta}})\}. \quad (2.14)$$

where J=J_{model1} for Model 1 and J=J_{model2} for model 2. For Model 1, the first derivative of $l(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{x}})$ with respect to μ_j is

$$\frac{\partial l}{\partial \mu_0} = -\frac{1}{\mu_0} \sum_{i=1}^{n} z_{i0} + \frac{1}{\mu_0^2} \sum_{i=1}^{n} z_{i0} x_i \quad (2.15)$$

$$\frac{\partial l}{\partial \mu_{1l}} = -\frac{1}{\mu_{1l}} \sum_{i=1}^{s} z_{i1l} + \frac{1}{\mu_{1l}^2} \sum_{i=1}^{s} z_{i1l}(x_i - 0.5) \quad (2.16)$$

$$\frac{\partial l}{\partial \mu_{1r}} = -\frac{1}{\mu_{1r}} \sum_{i=s+1}^{n} z_{i1r} + \frac{1}{\mu_{1r}^2} \sum_{i=1}^{n} z_{i1r}(0.5 - x_i) \quad (2.17)$$

$$\frac{\partial l}{\partial \mu_2} = -\frac{1}{\mu_2} \sum_{i=1}^{n} z_{i2} + \frac{1}{\mu_2^2} \sum_{i=1}^{n} z_{i2}(1 - x_i) \quad (2.18)$$

where $s = \arg\max_{x_i<0.5} x_i$. By solving the system of equations above [Equations 2.15 to 2.18], the MLEs for Model 1 are

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} z_{i0} x_i}{\sum_{i=1}^{n} z_{i0}} \tag{2.19}$$

$$\hat{\mu}_{1l} = \frac{\sum_{i=1}^{s} z_{i1l}(0.5 - x_i)}{\sum_{i=1}^{s} z_{i1l}} \tag{2.20}$$

$$\hat{\mu}_{1r} = \frac{\sum_{i=s+1}^{n} z_{i1r}(x_i - 0.5)}{\sum_{i=s+1}^{n} z_{i1r}} \tag{2.21}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{n} z_{i2}(1 - x_i)}{\sum_{i=1}^{n} z_{i2}} \tag{2.22}$$

Similarly for Model 2, the MLEs for μ_j are

$$\hat{\mu}_{01} = \frac{\sum_{i=1}^{n} z_{i01} x_i}{\sum_{i=1}^{n} z_{i01}} \tag{2.23}$$

$$\hat{\mu}_{02} = \frac{\sum_{i=1}^{n} z_{i02} x_i}{\sum_{i=1}^{n} z_{i02}} \tag{2.24}$$

$$\hat{\mu}_{1l} = \frac{\sum_{i=1}^{s} z_{i1l}(0.5 - x_i)}{\sum_{i=1}^{s} z_{i1l}} \tag{2.25}$$

$$\hat{\mu}_{1r} = \frac{\sum_{i=s+1}^{n} z_{i1r}(x_i - 0.5)}{\sum_{i=s+1}^{n} z_{i1r}} \tag{2.26}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{n} z_{i2}(1 - x_i)}{\sum_{i=1}^{n} z_{i2}}. \tag{2.27}$$

The signs of the second derivatives show that the estimated μ_j given in in Equations 2.19 to 2.27 maximize the log-likelihood given in [Equation 2.14] (See Appendix).

Model 3 has the likelihood function given by

$$
\begin{aligned}
L(\alpha,\beta,\underline{\boldsymbol{\theta}}|\underline{\mathbf{x}}) &= L_1(\alpha,\beta|\underline{\mathbf{x}})L_2(\underline{\boldsymbol{\theta}}|\underline{\mathbf{x}}) \\
\text{where } L_1(\alpha,\beta|\underline{\mathbf{x}}) &= \prod_{i=1}^{n}\alpha^{I_{\{0\}}(x_i)}\beta^{I_{\{1\}}(x_i)}(1-\alpha-\beta)^{1-I_{\{0,1\}}(x_i)} \\
L_2(\underline{\boldsymbol{\theta}}|\underline{\mathbf{x}}) &= \prod_{i=1}^{n}\prod_{j\in J}\{\pi_j f_j(x_i|\underline{\boldsymbol{\theta}})\}^{z_{ij}}. \qquad (2.28)
\end{aligned}
$$

Let $\alpha = \phi\gamma$ and $\beta = \phi(1-\gamma)$, $0 < \phi, \gamma < 1$ as in [21]. Then the likelihood function can be rewritten as

$$
\begin{aligned}
L(\alpha,\beta,\underline{\boldsymbol{\theta}}|\underline{\mathbf{x}}) &= L(\phi,\gamma,\underline{\boldsymbol{\theta}}|\underline{\mathbf{x}}) = L_{1a}(\phi|\underline{\mathbf{x}})L_{1b}(\gamma|\underline{\mathbf{x}})L_2(\underline{\boldsymbol{\theta}}|\underline{\mathbf{x}}) \\
\text{where } L_{1a}(\phi|\underline{\mathbf{x}}) &= \prod_{i=1}^{n}\phi^{I_{\{0,1\}}(x_i)}(1-\phi)^{1-I_{\{0,1\}}(x_i)} \\
L_{1b}(\gamma|\underline{\mathbf{x}}) &= \prod_{i=1}^{n}\gamma^{I_{\{0\}}(x_i)}(1-\gamma)^{I_{\{1\}}(x_i)} \\
L_2(\underline{\boldsymbol{\theta}}|\underline{\mathbf{x}}) &= \prod_{i=1}^{n}\prod_{j\in J}\{\pi_j f_j(x_i|\underline{\boldsymbol{\theta}})\}^{z_{ij}}. \qquad (2.29)
\end{aligned}
$$

The log-likelihood then is

$$
\begin{aligned}
l(\phi,\gamma,\underline{\boldsymbol{\theta}}|\underline{\mathbf{x}}) &= l_{1a}(\phi\ |\underline{\mathbf{x}}) + l_{1b}(\gamma\ |\underline{\mathbf{x}}) + l_2(\underline{\boldsymbol{\theta}}|\underline{\mathbf{x}}) \\
\text{where } l_{1a}(\phi|\underline{\mathbf{x}}) &= N_0\log\phi + N_1\log\phi + (n-N_0-N_1)\log(1-\phi) \\
l_{1b}(\gamma|\underline{\mathbf{x}}) &= N_0\log\gamma + N_1\log(1-\gamma) \\
l_2(\underline{\boldsymbol{\theta}}|\underline{\mathbf{x}}) &= \sum_{i=1}^{n}\sum_{j\in J} z_{ij}\{\log\pi_j + \log f_j(x_i|\underline{\boldsymbol{\theta}})\} \qquad (2.30)
\end{aligned}
$$

where $N_0 = \sum_{i=1}^{n} I_{\{0\}}(x_i)$, $N_1 = \sum_{i=1}^{n} I_{\{1\}}(x_i)$. The first derivatives of l_{1a} and l_{1b} with respect to ϕ and γ, respectively are

$$\frac{\partial l_{1a}}{\partial \phi} = \frac{N_0 + N_1}{\phi} - \frac{n - N_0 - N_1}{1 - \phi} \tag{2.31}$$

$$\frac{\partial l_{1b}}{\partial \gamma} = \frac{N_0}{\gamma} - \frac{N_1}{1 - \gamma}. \tag{2.32}$$

By solving the normal equations generated by Equations 2.31 and 2.32, the MLEs are

$$\hat{\phi} = \frac{N_0 + N_1}{n} \tag{2.33}$$

$$\hat{\gamma} = \frac{N_0}{N_0 + N_1}. \tag{2.34}$$

Equations 2.33 and 2.34 can be rewritten as

$$\hat{\alpha} = \hat{\phi}\hat{\gamma} = \frac{N_0}{n} \tag{2.35}$$

$$\hat{\beta} = \hat{\phi}(1 - \hat{\gamma}) = \frac{N_1}{n}. \tag{2.36}$$

The estimate of parameters $\underline{\boldsymbol{\theta}}$ has the same formula as in Equations 2.23 to 2.27.

2.2.2 Expectation-Maximization Algorithm for Model 1

The EM algorithm is applied for obtaining MLEs of the parameters $\underline{\boldsymbol{\theta}}$ by maximizing Equation 2.14 for Models 1 and 2 and maximizing $l_2(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{x}})$ in Equation 2.30 for Model 3. The detailed algorithm for Model 1 is below:

A. Initialize z_{ij}'s;

$$z_{i0}^{(0)} = \text{runif}(n, 0, 1), z_{i1l}^{(0)} = \text{runif}\left(s, 0, 1 - z_{i0}^{(0)}\right),$$

$$z_{i1r}^{(0)} = \text{runif}\left(n-s, 0, 1-z_{i0}^{(0)}\right), z_{i1}^{(0)} = \left[\, z_{i1l}^{(0)}, z_{i1r}^{(0)} \right]$$

$$\text{and } z_{i2}^{(0)} = \left(1 - z_{i0}^{(0)} - z_{i1}^{(0)}\right)$$

B. Calculate initial mixing proportion π_j's for $j \in J_{model1}$;

$$\pi_j^{(0)} = \frac{\sum_{i=1}^{n} z_{ij}^{(0)}}{n}$$

C. M-step at *k*-th iteration:

$$\hat{\mu}_0^{(k)} = \frac{\sum_{i=1}^{n} z_{i0} x_i}{\sum_{i=1}^{n} z_{i0}}, \hat{\mu}_{1l}^{(k)} = \frac{\sum_{i=1}^{s} z_{i1l}(0.5 - x_i)}{\sum_{i=1}^{s} z_{i1l}},$$

$$\hat{\mu}_{1r}^{(k)} = \frac{\sum_{i=s+1}^{n} z_{i1r}(x_i - 0.5)}{\sum_{i=s+1}^{n} z_{i1r}}, \text{ and } \hat{\mu}_2^{(k)} = \frac{\sum_{i=1}^{n} z_{i2}(1 - x_i)}{\sum_{i=1}^{n} z_{i2}}.$$

D. E-step:

$$\hat{\pi}_j^{(k)} = \frac{\sum_{i=1}^{n} z_{ij}}{n-s} \text{ and } z_{ij}^{(k)} = \frac{\hat{\pi}_j f_j(x_i)}{\sum_{j \in J} \hat{\pi}_j f_j(x_i)}$$

E. Calculate the log-likelihood $l(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{x}})$ in Equation 2.14 . Go to step C and repeat until

$\left| l(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{x}})^{(k+1)} - l(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{x}})^{(k)} \right| < \tau$, for some $\tau > 0$.

The "runif" [24] is an R-package program to generate random uniform numbers.

2.2.3 Partitioned Expectation-Maximization Algorithm for Models 2 and 3

The EM algorithm searches for a global maximum for the MLEs over the entire data. For long-tailed data, however, it is possible that the tail might not be captured well with the EM algorithm [25]. Riska et al [26] propose a method for using EM algorithm to estimate the parameters in the mixture of two or more exponential distributions. They partition the data by the coefficient of variation (CV) so that each partition has CV greater than a certain threshold (e.g., 1.5). Then they

fit the mixture of exponential distributions separately for each partition. I modify their approach for the AAP data to estimate parameters in Model 2 and Model 3.

The AAP values for most positions are distributed around 0 [Table 2.1]: more than 99% of the AAP values are less than 0.25 for each of the six individuals. My goal is to partition the data at the point where the mixture distribution starts to appear. I use the R software package "fitdistr" [27] to fit the AAP values to a single exponential distribution. The fitted parameter is named $\hat{\mu}_0$. Figure 2.2 shows that the measured AAP values depart from the *x=y* line to the right of $5\hat{\mu}_0$. The x-axis is the random number generated from the estimated $\hat{\mu}_0$ and the y-axis is the measured AAP value. The left panels use all AAP values and the right panels use the AAP values in (0, 1). The blue line is drawn at $5\hat{\mu}_0$. It indicates that there are more AAP values than expected from a single exponential to the right of $5\hat{\mu}_0$. Thus I partitioned the AAP values into two sets at $5\hat{\mu}_0$.

The left partition contains the AAP values less than $5\hat{\mu}_0$ and contains more than 95% of the data for Model 1 and more than 98% of the data for Model 2 for all six individuals [Table 2.2]. I used the mean $\hat{\mu}_0$ for the data to the left of $5\hat{\mu}_0$, i.e., $\hat{\mu}_{01} = \hat{\mu}_0$. The right mixture component for g0 is used for the AAP values greater than or equal to $5\hat{\mu}_0$. Here I assume $\hat{\mu}_0$ is small enough so that $5\hat{\mu}_0 < 0.5$. This algorithm is applied to Model 2 and Model 3 that use two continuous components to describe *g0* genotype.

Table 2.2 The Estimated Values of μ_0 and the Fraction of Each Partition

	Model 2		Model 3	
	μ0	% of AAP < 5μ0 in [0, 1]	μ0	% of AAP < 5μ0 in (0, 1)
NA12878 (CEU-child)	0.0048	95.63%	0.0108	98.55%
NA12891 (CEU-father)	0.0069	98.50%	0.0143	99.44%
NA12892 (CEU-mother)	0.0109	98.97%	0.0160	99.56%
NA19238 (YRI-mother)	0.0038	95.51%	0.0108	99.08%
NA19239 (YRI-father)	0.0051	96.75%	0.0097	98.32%
NA19240 (YRI-child)	0.0057	97.32%	0.0125	98.85%

Figure 2.2 Q-Q Plots of the Measured AAP Values versus an Exponential Distribution

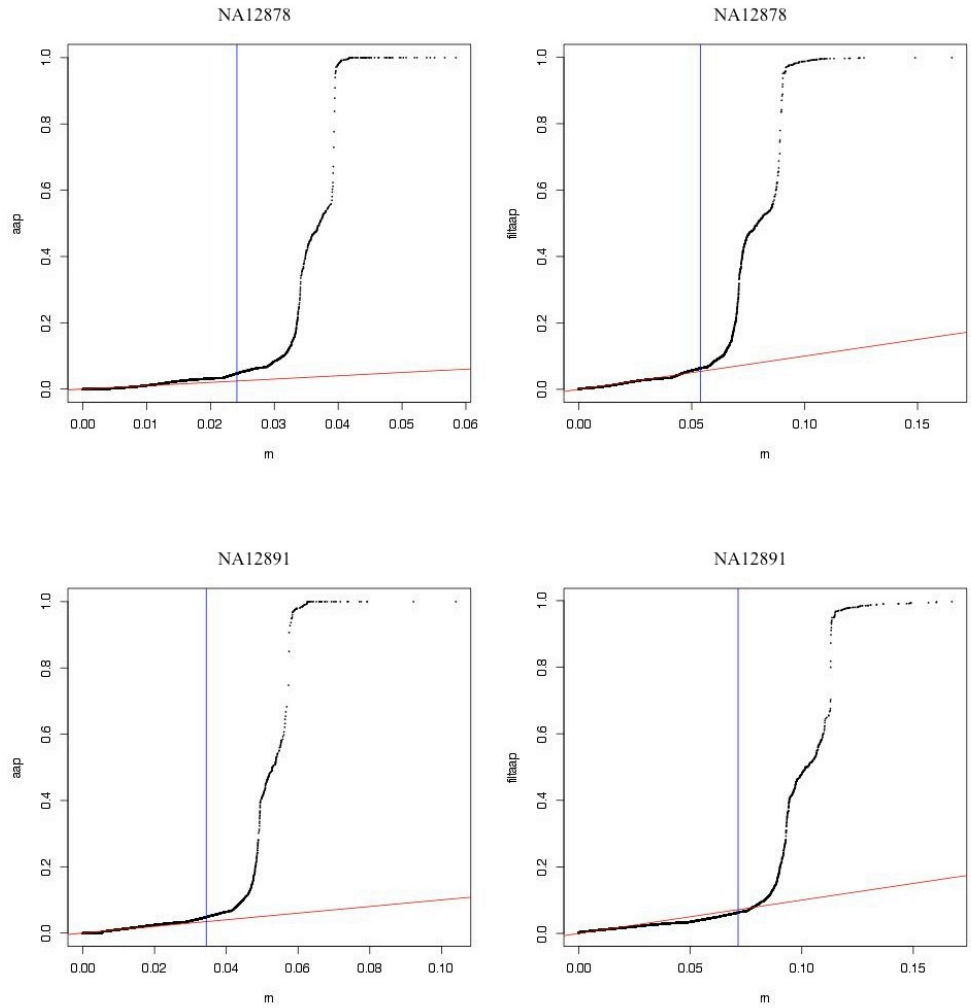


Figure 2.2 (Continued)

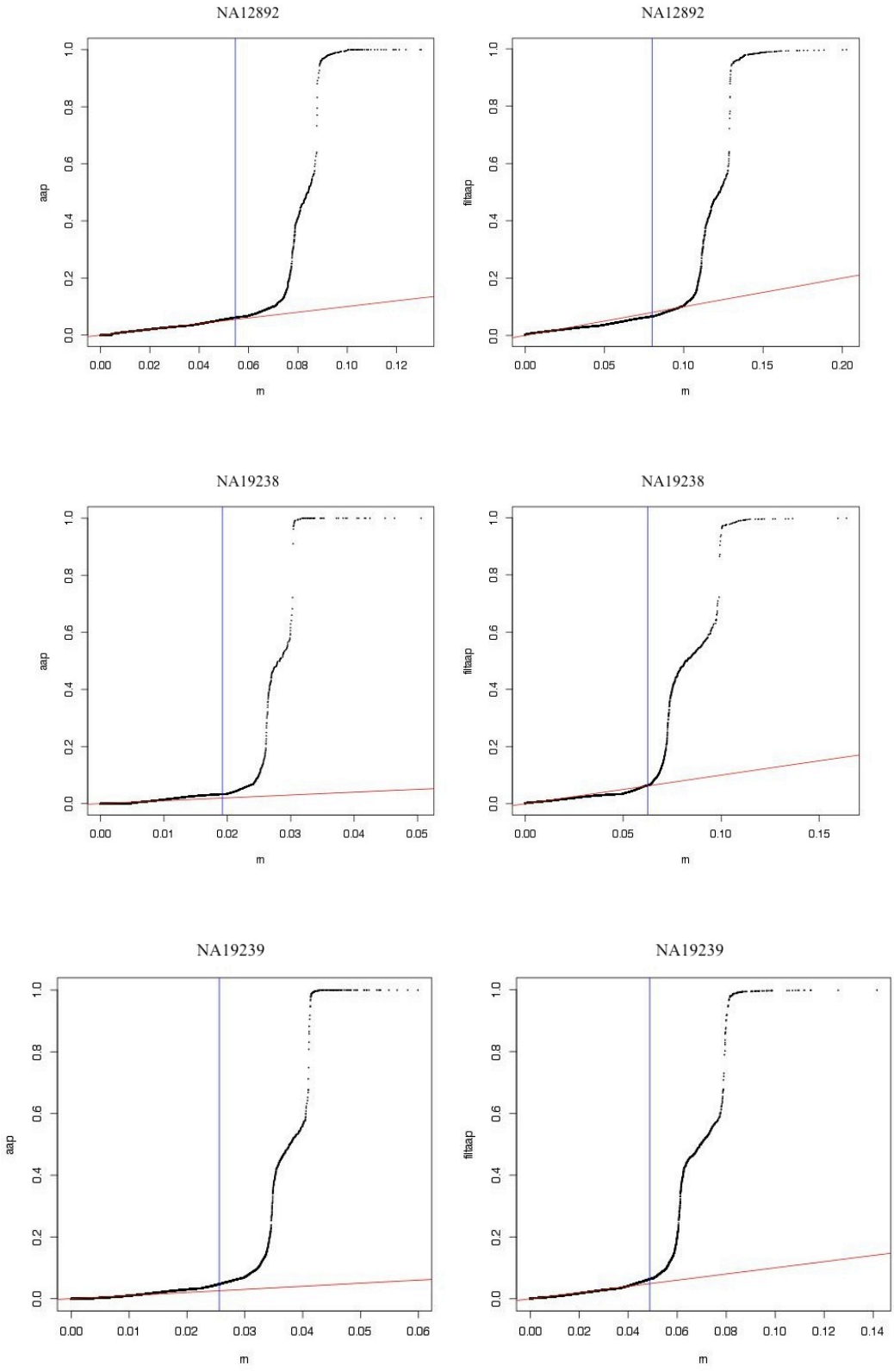
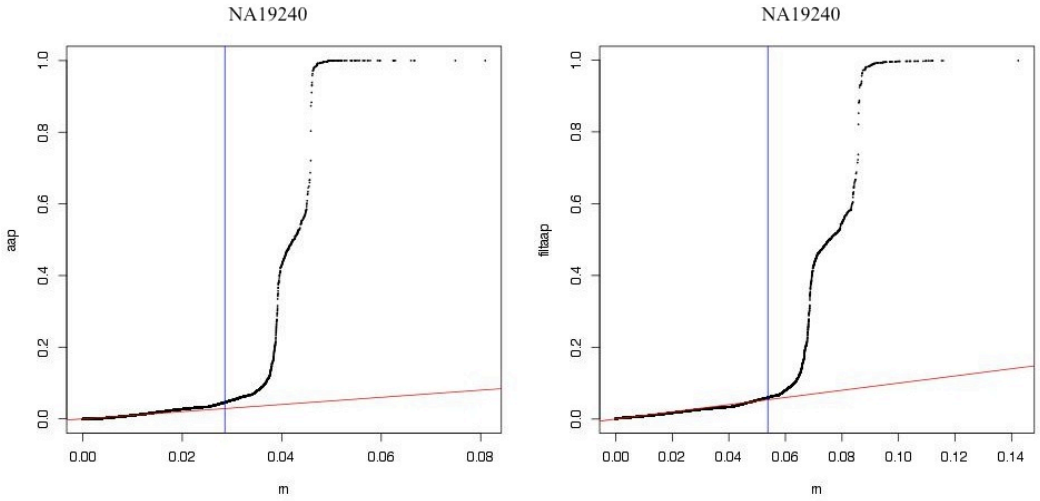


Figure 2.2 (Continued)



The detailed partitioned EM algorithm for Model 2 is as below:

A. Calculate $\hat{\mu}_0$ using the entire data and let $k = \arg\max_{0 \le x_i < 5\hat{\mu}_0} x_i$.

B. Initialize z_{ij}, $i = k+1, \ldots, n$;

$$z_{02}^{(0)} = \text{runif}(n-k, 0, 1)$$

$$z_{1l}^{(0)} = \text{runif}\left(s-k, 0, 1-z_{02}^{(0)}\right), z_{i1r}^{(0)} = \text{runif}\left(n-s+k, 0, 1-z_{02}^{(0)}\right),$$

$$z_{i1}^{(0)} = \left[z_{i1l}^{(0)}, z_{i1r}^{(0)}\right], \text{and } z_{i2}^{(0)} = \left(1 - z_{i02}^{(0)} - z_{i1}^{(0)}\right)$$

where $s = \arg\max_{0 \le x_i \le 0.5} x_i$

C. Calculate initial mixing proportion π'_j's;

$$\pi'^{(0)}_j = \frac{\sum_{i=k+1}^{n} z_{ij}^{(0)}}{n-k}$$

D. M-step at *k*-th iteration:

$$\hat{\mu}_{02}^{(k)} = \frac{\sum_{i=k+1}^{n} z_{i02} x_i}{\sum_{i=k+1}^{n} z_{i02}}$$

$$\hat{\mu}_{1l}^{(k)} = \frac{\sum_{i=k+1}^{s} z_{i1l}(0.5-x_i)}{\sum_{i=k+1}^{s} z_{i1l}}, \hat{\mu}_{1r}^{(k)} = \frac{\sum_{i=s+1}^{n} z_{i1r}(x_i-0.5)}{\sum_{i=s+1}^{n} z_{i1r}},$$

$$\text{and } \hat{\mu}_{2}^{(k)} = \frac{\sum_{i=k+1}^{n} z_{i2}(1-x_i)}{\sum_{i=k+1}^{n} z_{i2}}$$

E. E-step:

$$\hat{\pi}'^{(k)}_{j} = \frac{\sum_{i=k+1}^{n} z_{ij}}{n-k} \text{ and } z_{ij}^{(k)} = \frac{\hat{\pi}_j f_j(x_i)}{\sum_{j\in J} \hat{\pi}_j f_j(y_i)}$$

F. Calculate the log-likelihood in Equation 2.14.

G. Go to step C and repeat until

$$\left| l\left(\underline{\boldsymbol{\theta}}\middle|\underline{\boldsymbol{y}}\right)^{(k+1)} - l\left(\underline{\boldsymbol{\theta}}\middle|\underline{\boldsymbol{y}}\right)^{(k)} \right| < \tau, \text{ for some } \tau > 0.$$

H. Estimate the mean and mixing proportion of component 01 as

$$\hat{\mu}_{01} = \frac{\sum_{i=1}^{k} x_i}{k} \text{ and } \hat{\pi}_{01} = \frac{k}{n}.$$

I. Rescale the mixing proportions $\hat{\pi}_j'$ to the entire data size as

$$\hat{\pi}_j = (1-\hat{\pi}_{01})\hat{\pi}_j'.$$

The partitioned EM algorithm for Model 3 follows similar steps except that $\hat{\mu}_0$ is calculated using the AAP values in (0, 1) in step A and $n' = n - N_0 - N_1$ is used instead of n through the algorithm.

2.3 Results

2.3.1 Parameter Estimates for Model 1

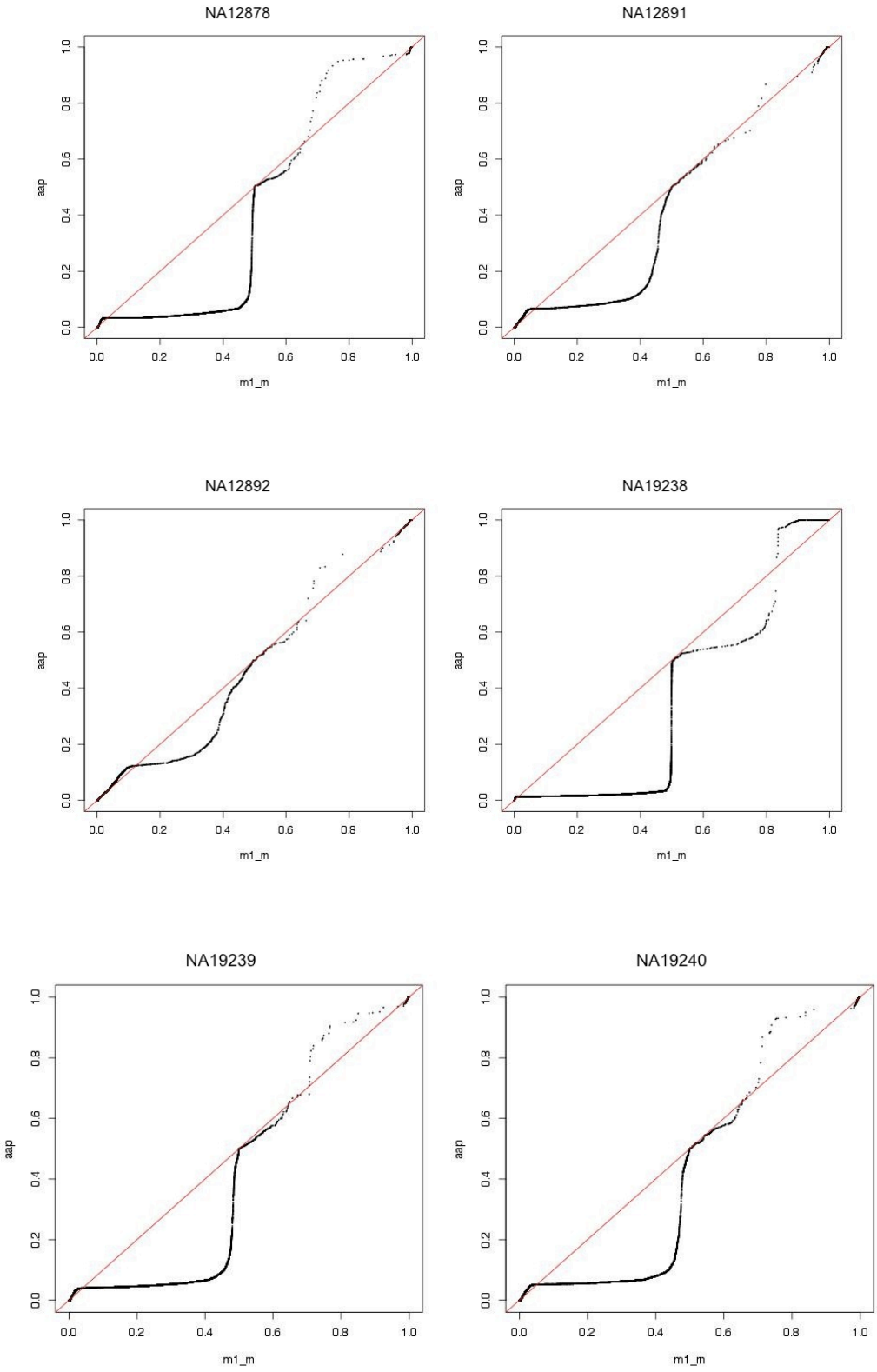
Each individual had somewhat different distributions of the measured AAP values. Table 2.3 shows the estimates of the parameters in Model 1 [Equation 2.1]. The estimated μ_0 varies from 0.0010 to 0.0104. The mixing proportion of component 0 ($\hat{\pi}_0$) has the largest values among four components and ranges from 87.84% to 99.86%. The individual NA12892 has the largest $\hat{\mu}_0$ and $\hat{\pi}_0$, and NA19238 has the smallest $\hat{\mu}_0$ and $\hat{\pi}_0$. The order of $\hat{\mu}_0$ matches with the order of $\hat{\pi}_0$ among six individuals. The mean parameter of component 1l ($\hat{\mu}_{1l}$) ranges from 0.2652 to 0.4798. These are much larger than those of component 1r ($\hat{\mu}_{1r}$), ranging from 0.0335 to 0.1120. The mixing proportion of component 1l ($\hat{\pi}_{1l}$) is much larger than that of component 1r ($\hat{\pi}_{1r}$). The mean parameter $\hat{\mu}_2$ has slightly larger values than $\hat{\mu}_1$, with values between 0.0031 to 0.1261.

Table 2.3 Estimated Parameter Values and the Log-Likelihood for Model 1

sample	loglike	μ0	μ1l	μ1r	μ2	π0	π1l	π1r	π2
NA12892 (CEU-mother)	2,073,690	0.0104	0.2652	0.0692	0.0174	99.8579%	0.1000%	0.0119%	0.0303%
NA12891 (CEU-father)	2,004,106	0.0061	0.3754	0.0684	0.0146	99.6429%	0.3102%	0.0196%	0.0274%
NA12878 (CEU-child)	2,811,123	0.0035	0.4484	0.1120	0.0036	98.1380%	1.8120%	0.0197%	0.0302%
NA19238 (YRI-mother)	2,442,051	0.0010	0.4798	0.0335	0.1261	87.8363%	12.1002%	0.0151%	0.0484%
NA19239 (YRI-father)	2,923,421	0.0038	0.4319	0.0914	0.0031	98.6594%	1.2802%	0.0288%	0.0316%
NA19240 (YRI-child)	2,663,106	0.0046	0.4122	0.0884	0.0042	99.2564%	0.6809%	0.0282%	0.0345%

Figure 2.3 shows the Q-Q plot of a random number generated from the fitted Model 1 versus the measured AAP value. The x-axis is the random number generated from Model 1, and the y-axis is the measured AAP value. The red line is drawn at the $x=y$ line. For the six individuals, the AAP values are under-represented between 0.02 and 0.5. That is, a single exponential distribution estimated fewer AAP values between 0.1 and 0.25 that actually occurred. The AAP values between 0.7 and 0.98 are slightly over-represented.

Figure 2.3 Q-Q plot of Model 1 vs. the AAP Values



2.3.2 Parameter Estimates for Model 2

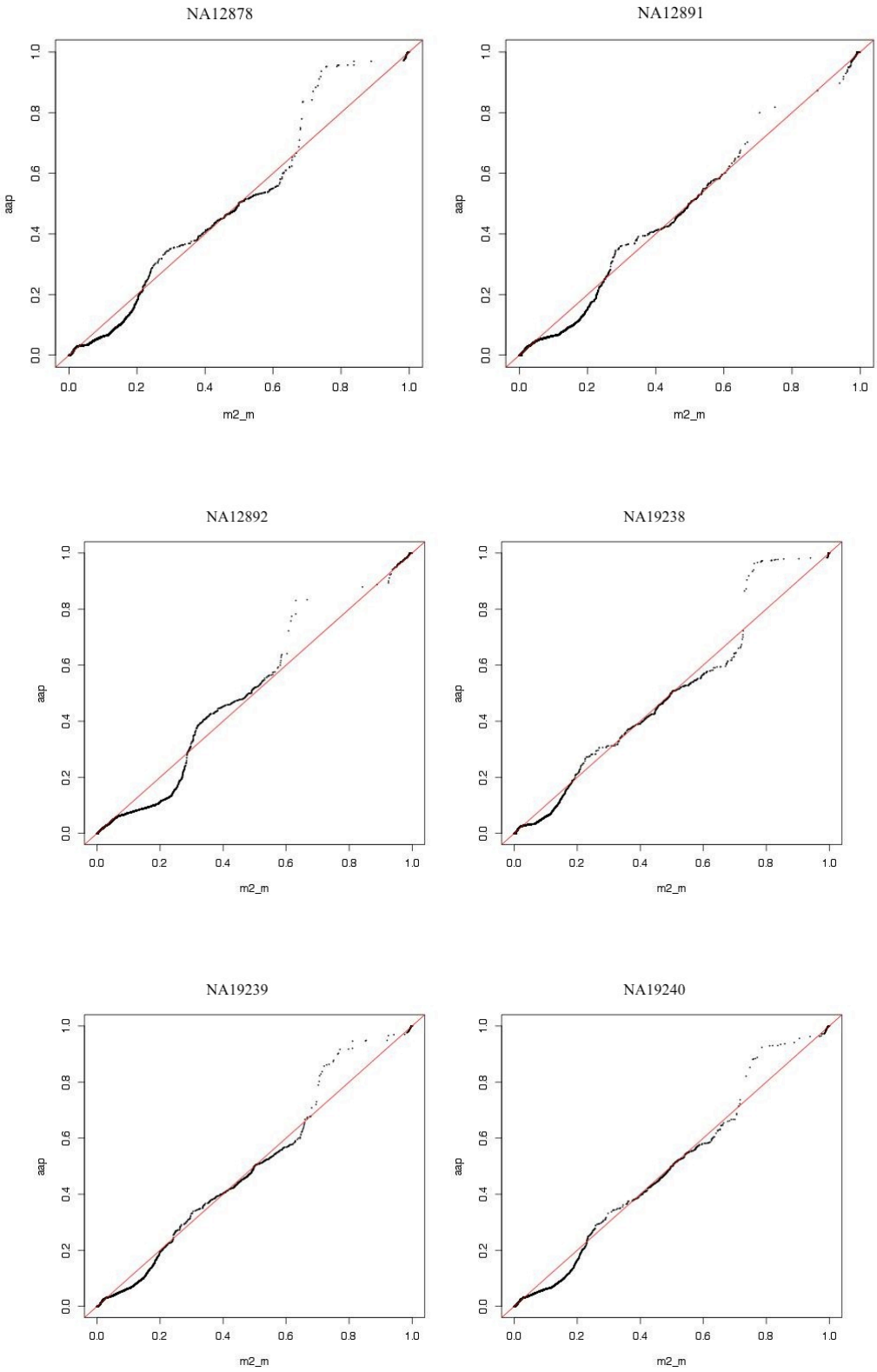
Table 2.4 shows the estimated parameters for Model 2 [Equation 2.6]. The estimated mean parameter of component 01 ranges from 0.0038 to 0.0109. The estimated means of component 02, which is introduced to describe the heavy tail and thus defined to be larger than $\hat{\mu}_{01}$, are from 0.0288 to 0.0787. The ratio of $\hat{\mu}_{02}$ to $\hat{\mu}_{01}$ is 7.34 for NA12878, 7.91 for NA12891, 7.20 for NA12892, 7.49 for NA19238, 7.62 for NA19239, and 7.20 for NA19240. Although each individual has somewhat different $\hat{\mu}_{01}$ and $\hat{\mu}_{02}$ values, the ratios are similar. The estimated mean parameter of component 1l is slightly smaller than that of component 1r. This might be because component 1l uses the truncated AAP values from $5\hat{\mu}_0$ to 0.5 while component 1r uses the AAP values from 0.5 to 1.0. The mixing proportion of component 1l is slightly larger than that of component 1r by 0.01% to 0.02%. The estimated value of μ_2 is smaller than the mean parameters of component 1l and 1r. The mixing proportion of component 2 is approximately the same for all six individuals.

Table 2.4 Estimated Parameter Values and the Log-Likelihood for Model 2

sample	loglike	μ01	μ02	μ1l	μ1r	μ2	π01	π02	π1l	π1r	π2
NA12892 (CEU-mother)	2,062,115.42	0.0109	0.0787	0.0522	0.0709	0.0170	98.9715%	0.9691%	0.0189%	0.0103%	0.0302%
NA12891 (CEU-father)	1,995,567.96	0.0069	0.0545	0.0626	0.0704	0.0142	98.5043%	1.4189%	0.0297%	0.0199%	0.0272%
NA12878 (CEU-child)	2,799,813.67	0.0048	0.0354	0.0724	0.1037	0.0039	95.6259%	4.2907%	0.0333%	0.0195%	0.0306%
NA19238 (YRI-mother)	2,407,630.68	0.0038	0.0288	0.0924	0.1123	0.0016	95.5068%	4.3841%	0.0465%	0.0301%	0.0325%
NA19239 (YRI-father)	2,901,749.07	0.0051	0.0390	0.0713	0.0917	0.0031	96.7508%	3.1348%	0.0534%	0.0293%	0.0316%
NA19240 (YRI-child)	2,642,878.45	0.0057	0.0411	0.0573	0.0862	0.0043	97.3153%	2.5786%	0.0431%	0.0283%	0.0346%

The log-likelihood of Model 2 achieves around 99% of the log-likelihood values in Model 1. Figure 2.4 is the Q-Q plot of random numbers generated from Model 2 vs. actual AAP values. The x-axis is the random number generated from Model 2 and the y-axis is the measured AAP value. The red line is drawn at the $x=y$ line. The Q-Q plots shows that the points of random numbers generated from Model 2 versus the measured AAP values are closer to the $x=y$ line. Most of the points around the AAP value 0.5 are on the $x=y$ line, but the points between 0.01 and 0.3 are under-represented while the points between 0.3 and 0.5 are over-represented. The points around 0.75 are still over-represented in Model 1.

Figure 2.4 Q-Q Plot of Model 2 vs. the AAP Values



2.3.3 Parameter Estimates for Model 3

The estimated parameters of Model 3 [Equation 2.12] are in Table 2.5. The mixing proportions $\hat{\pi}_j$ are scaled as

$$scaled\ \hat{\pi}_j = \left(1 - \hat{\alpha} - \hat{\beta}\right)\hat{\pi}_j\ .$$

Table 2.5 Estimated Parameter Values for Model 3

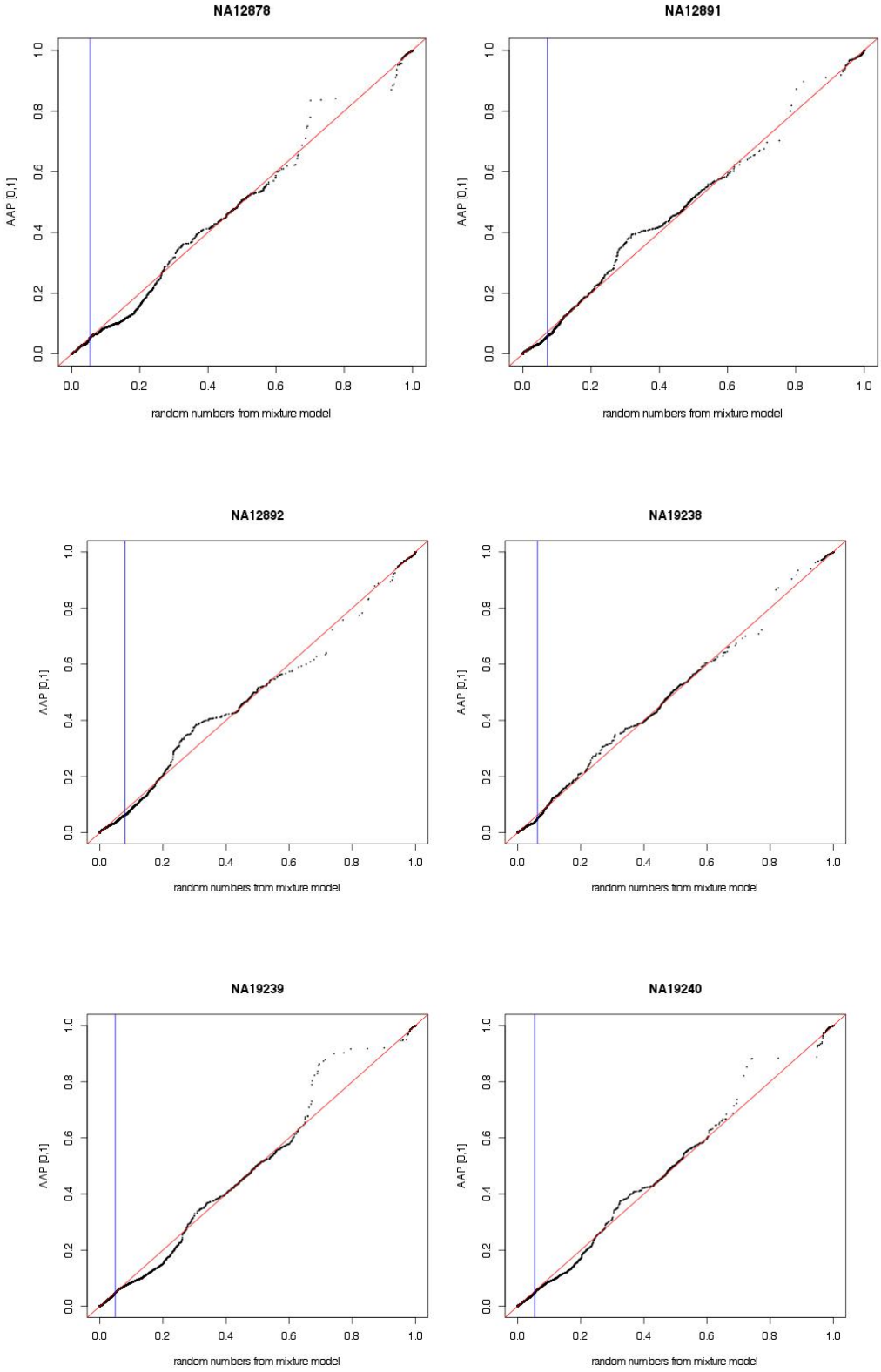
sample	μ01	μ02	μ1l	μ1r	μ2	α	β	π01	π02	π1l	π1r	π2
NA12892 (CEU mother)	0.01600	0.14697	0.05100	0.03584	0.02607	32.2913%	0.0103%	99.7323%	0.2066%	0.0223%	0.0095%	0.0292%
NA12891 (CEU father)	0.01429	0.15627	0.04935	0.05276	0.02776	52.7267%	0.0130%	99.6560%	0.2417%	0.0419%	0.0303%	0.0301%
NA12878 (CEU child)	0.01079	0.08519	0.04984	0.05266	0.02365	57.1434%	0.0205%	99.0976%	0.7805%	0.0558%	0.0368%	0.0293%
NA19238 (YRI mother)	0.01252	0.13859	0.06705	0.05706	0.01861	71.3276%	0.0260%	99.4567%	0.3162%	0.1101%	0.0823%	0.0347%
NA19239 (YRI father)	0.00975	0.08148	0.05666	0.07553	0.01172	49.4758%	0.0193%	99.0048%	0.8291%	0.0859%	0.0534%	0.0268%
NA19240 (YRI child)	0.01076	0.08722	0.04724	0.05874	0.01867	48.7109%	0.0198%	99.3237%	0.5235%	0.0711%	0.0494%	0.0324%

The proportions of zero values ($\hat{\alpha}$) vary from 32% to 71%, and the proportions of one values range from 0.010% to 0.026%. The scaled proportion of component *01* ($\hat{\pi}_{01}$) varied from 28.5% to 67.5%. The scaled proportion of component *02* ($\hat{\pi}_{02}$) is much smaller than the component *01* with values between 0.09% and 0.42%. Among mean parameters, $\hat{\mu}_{01}$has the smallest value of all component mean parameters, and $\hat{\mu}_{02}$ has ratio $\frac{\hat{\mu}_{02}}{\hat{\mu}_{01}}$ ranging between 7.9 to 11.1. The proportion of component *1l* is greater than that of component *1r* for each of the six participants ($\hat{\pi}_{1l} > \hat{\pi}_{1r}$). Both $\hat{\mu}_{1l}$ and $\hat{\mu}_{1r}$ has estimated values between 0.035 and 0.075. The differences between the left and right sets of parameters may reflect the asymmetry of the distribution describing g1. The mean parameter ($\hat{\mu}_2$) for the component has values between 0.011 and 0.028. The estimates of $\hat{\mu}_0$ and $\hat{\mu}_2$ are smaller than the estimates of the other means documenting that

the measured AAP values from g0 and g2 positions are less variable than the AAP values from g1 positions.

The log-likelihood for Model 3 is not given here, because the probability mass function is used at 0 and 1. Since this is always smaller than 1, the log-likelihood is always negative. The Q-Q plots of random numbers generated from Model 3 versus the measured AAP values are in figure 2.5. The red line is drawn at the *x=y* line and the blue line is drawn at $x = 5\hat{\mu}_0$. Compared to Model 1 [Figure 2.3] and Model 2 [Figure 2.4], the points are closer to the *x=y* line and the size of departure from the *x=y* line is smaller. For individuals NA12878, NA19239, and NA19240, the AAP values near 0.2 are underestimated and the AAP values between 0.2 and 0.5 are overestimated. For individuals NA12891, NA12892, and NA19238, the AAP values between 0.35 and 0.5 are over estimated. The AAP values around 0.75 are slightly overestimated for all individuals except NA12892.

Figure 2.5 Q-Q Plot of Model 3 vs. the AAP Values



2.4 Conclusion and Discussion

Out of three models, Model 3 describes the distribution of AAP values best. The Q-Q plot of Model 3 in Figure 2.5 shows that the points of random numbers from the fitted mixture model versus the measured AAP values are on or near the $x=y$ line for most of the values, confirming the adequacy of the fit, except for the points between 0.7 and 0.8. The departure of these points from the line might be explained by other genomic variations such as copy number variations. The fitted distribution adequately described the frequencies of the data for most of the values [Figure 2.6]. In Figure 2.6, the left panel is the histogram of the entire AAP values and the right panel is the AAP values greater than 0.1. The blue line is the fitted frequency from Model 3. The fitted curve showed the pattern of steep slope of g0. The distribution of values greater than 0.1 was well characterized by the four components, *02*, *1l*, *1r*, and *2*. However, there were some regions that were under or over estimated. The observed frequencies around 0.5 somewhat exceeded the values from the fitted distribution, and the observed frequencies near 0 and 1 were slightly overestimated. This model predicted more measured AAP values between 0.1 and 0.2 than occurred for this individual.

Figure 2.6 The Fitted Distribution of Model 3

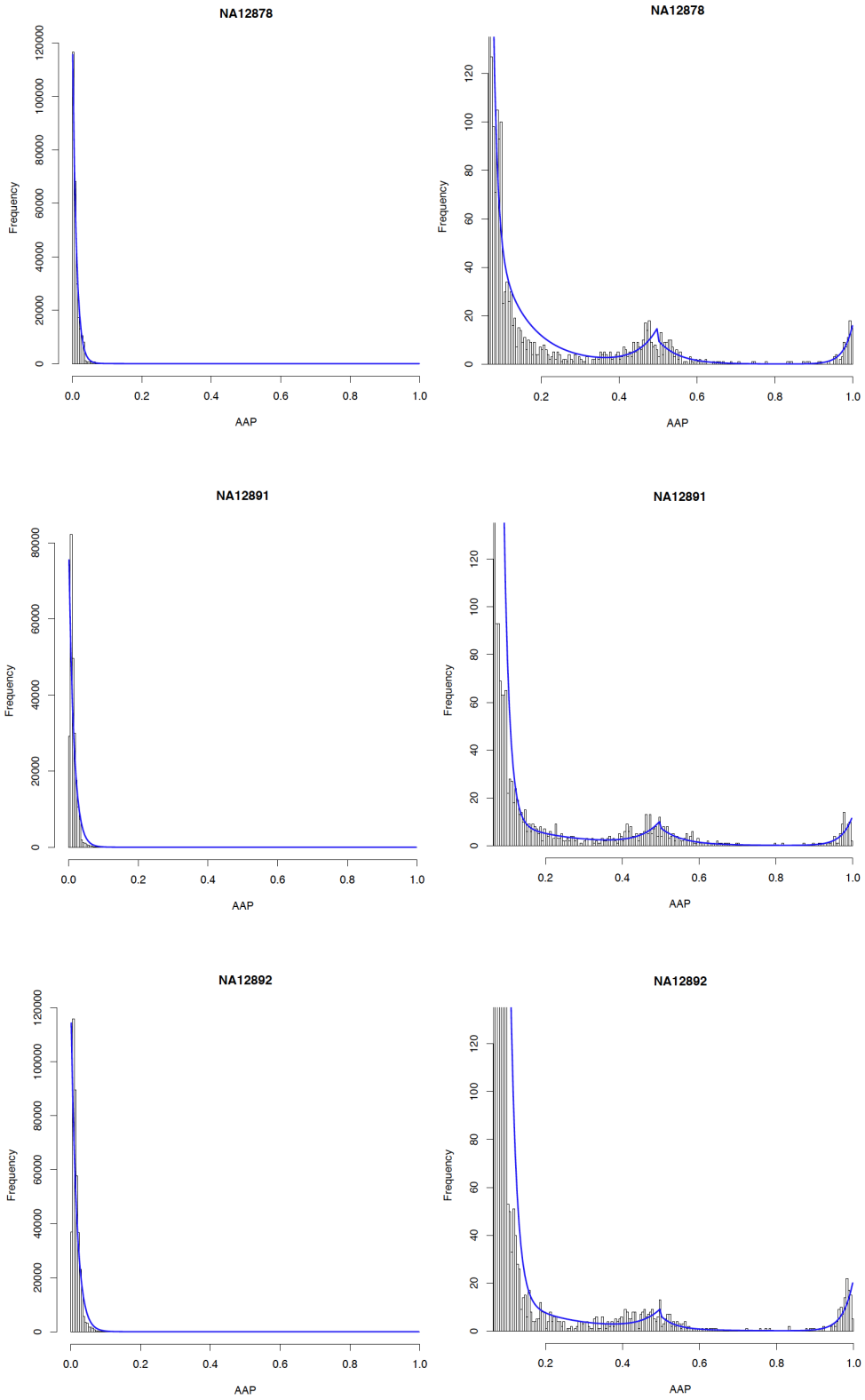
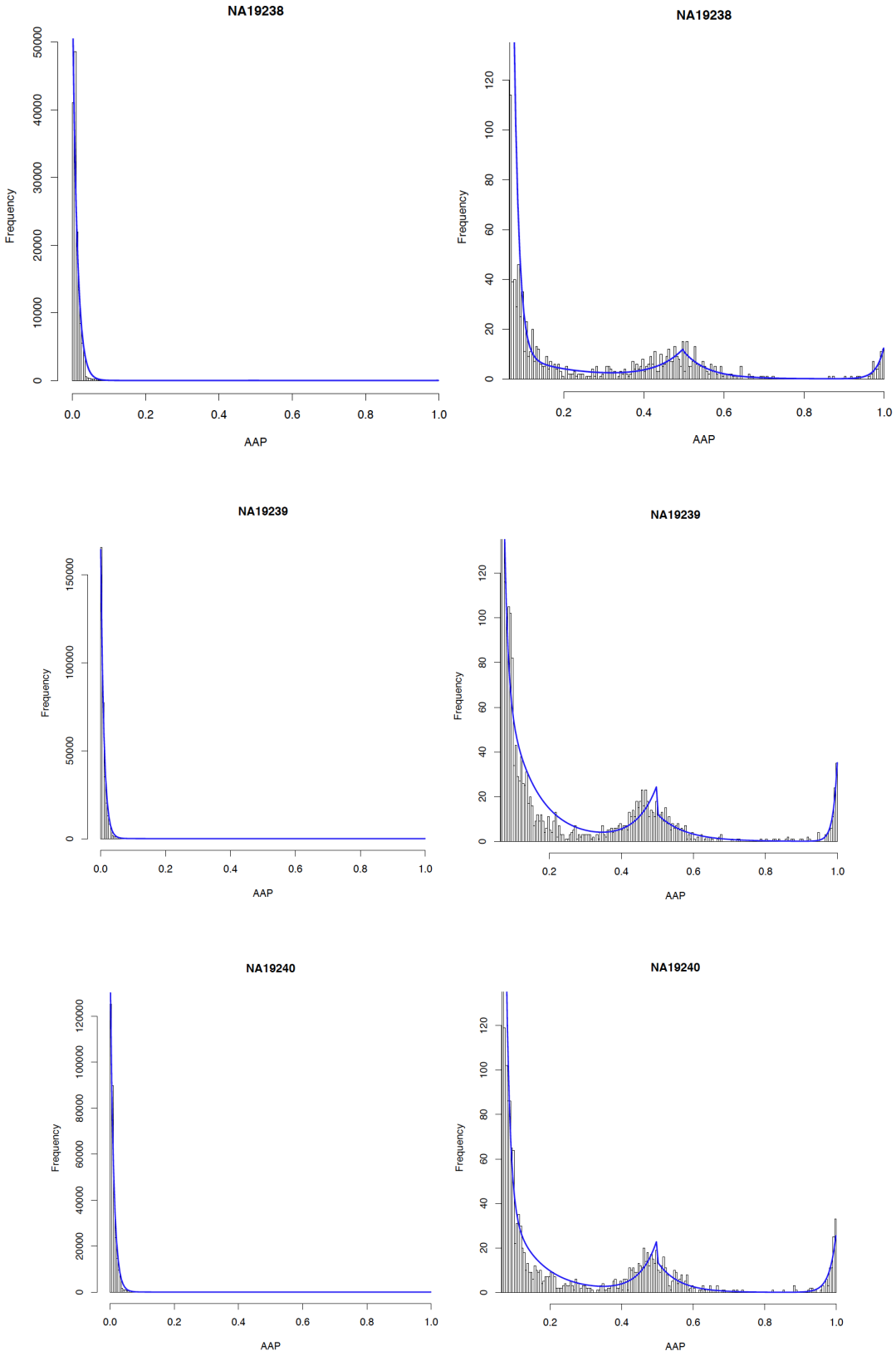


Figure 2.6 (Continued)



2.5 Applications

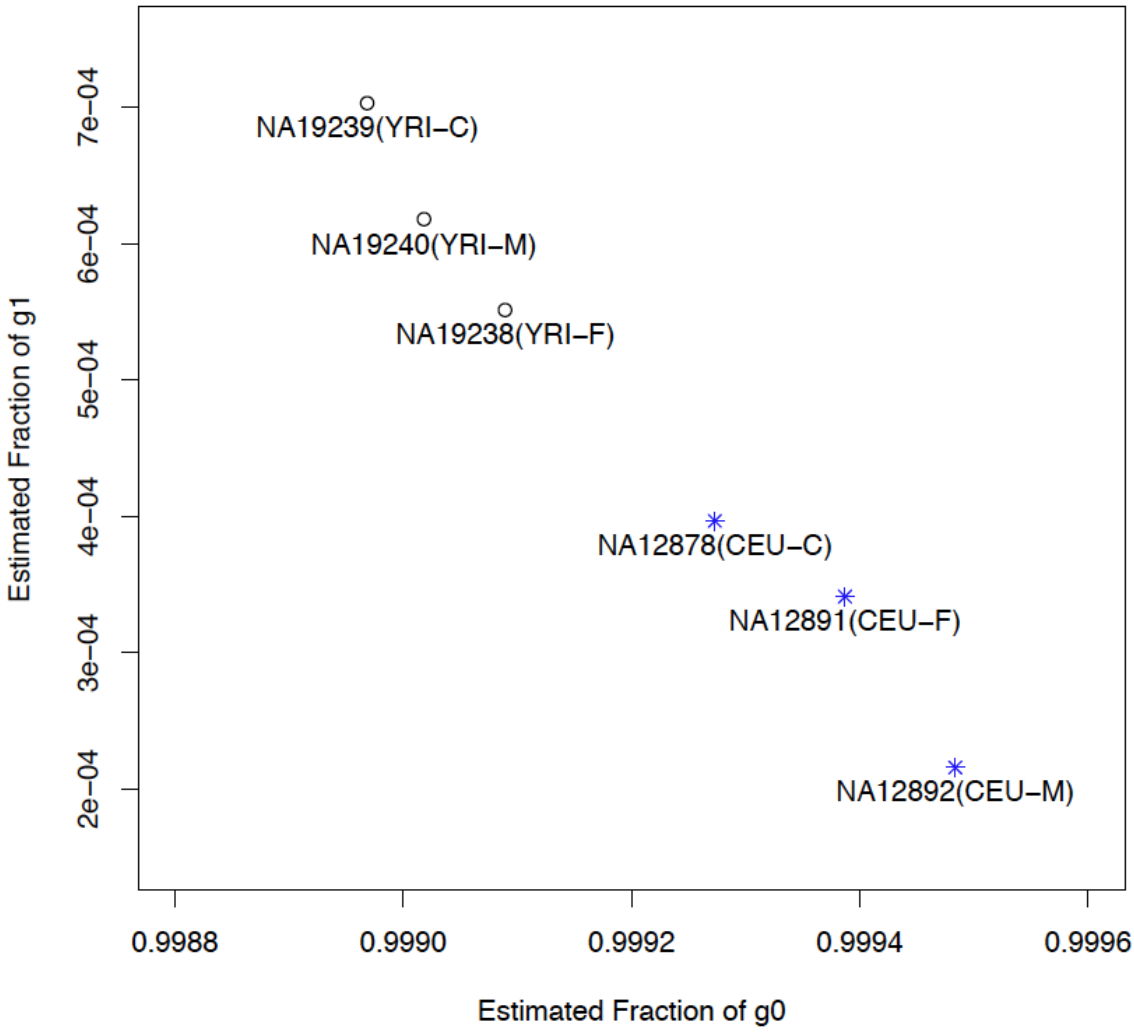
I estimate the fraction of each genotype using the estimated parameters in Model 3 and report them Table 2.6. The total estimated fraction of g0 (named p0) is $(\hat{\alpha} + scaled\ \hat{\pi}_{01} + scaled\ \hat{\pi}_{02})$. The estimate is more than 99.90% for all six individuals. While $\hat{p}_0$ is essentially constant, the component parameters have substantial variation. The estimated fraction for g1 (named p1) is $(scaled\ \hat{\pi}_{1l} + scaled\ \hat{\pi}_{1r})$. The estimated value ranged from 0.022% to 0.070%. The estimated fraction for g2 (named p2) is $(\hat{\beta} + scaled\ \hat{\pi}_2)$. The values ranged from 0.027% to 0.036%, which are smaller than the fractions for g1 for five of the six individuals.

Table 2.6 Estimated Fraction of Each Genotype from Model 3

sample	p0	p1	p2
NA12892 (CEU-mother)	99.9484%	0.0216%	0.0301%
NA12891(CEU-father)	99.9387%	0.0341%	0.0272%
NA12878 (CEU-child)	99.9273%	0.0397%	0.0330%
NA19238 (YRI-mother)	99.9089%	0.0551%	0.0360%
NA19239 (YRI-father)	99.8968%	0.0703%	0.0328%
NA19240 (YRI-child)	99.9018%	0.0618%	0.0364%

Figure 2.7 shows the plot of p0 versus p1 for the two trios studied here. We found each trio to be clustered. The individuals in the YRI trio had a smaller fraction of g0 genotypes and greater fraction of g1 genotypes than those in the CEU trio, suggesting that there is variability of similarity to the reference genome between members of the trios. In both trios, the child had smaller p0 and higher p1 than the parents.

Figure 2.7 Plot of p0 versus p1 for Two Trios



I use the estimated p0, p1, and p2 of parents to calculate the expected fractions of each genotype in a child of each trio. I name the child fractions q0, q1, and q2, and estimate them using the Binomal distribution. That is, $(x+y)^2 = x^2 + 2xy + y^2$. The resulting fractions are:

$$q0 = p_{m0}p_{f0} + \frac{1}{2}\left(p_{m0}p_{f1} + p_{f0}p_{m1}\right) + \frac{1}{4}p_{m1}p_{f1}$$

$$q1 = \frac{1}{2}\left(p_{m0}p_{f1} + p_{f0}p_{m1}\right) + \left(p_{m0}p_{f2} + p_{f0}p_{m2}\right) + \frac{1}{2}p_{m1}p_{f1} + \frac{1}{2}\left(p_{m1}p_{f2} + p_{f1}p_{m2}\right)$$

$$q2 = \frac{1}{4}p_{m1}p_{f1} + \frac{1}{2}\left(p_{m1}p_{f2} + p_{f2}p_{m1}\right) + p_{m2}p_{f2}.$$

The fraction p_{mk} is the fraction of genotype g_k ($k = 0,1, and\ 2$) in the mother and p_{fk} is the fraction in the father of each trio. As shown in Table 2.8, the fitted reference homozygote fraction ($\hat{p}_0$) of a child is approximately equal to the binomial estimate (q_0). The fitted g1 fraction ($\hat{p}_1$) of a child is about half of the binomial estimate (q_1) for both children. The fitted fraction of alternative homozygotes ($\hat{p}_2$) of a child is greater than the binomial estimate (q_2) for both children. The differences between the observed and the expected fractions of g0, g1, and g2 may be due to Mendelian inconsistency in the trios. The effects of Mendelian inconsistency should be analyzed in the future studies.

Table 2.7. Observed and Estimated Genotype Fraction of Children

	sample	p0	p1	p2	q0	q1	q2
CEU child	NA12878	99.92729%	0.03967%	0.03304%	99.91491%	0.08507%	0.00002%
YRI child	NA19240	99.90180%	0.06176%	0.03644%	99.86854%	0.13141%	0.00004%

Chapter 3 Single Nucleotide Polymorphism Calling and Genotyping via Mixture Modeling and Clustering

3.1 Data Description

I study the paired-end sequencing reads of the whole exome sequencing (also known as targeted exome sequencing) of three samples from a YRI trio sequenced in Parla et al. [20]. The selected YRI trio is one of the trios studied in the 1000 Genomes Project and is the YRI trio studied in Chapter 2. I use reads captured by NimbleGen SeqCap EZ Exome Library SR [13]. The data consists of three captures for NA19238 (the mother) and NA19240 (the child) and one capture for NA19239 (the father). The sequencing targeted the whole exome of 33,881,597 positions in chromosomes 1 to 22.

The sequence reads are processed through the GATK (The Genomic Analysis Toolkit) pipeline [9]. The raw reads from each capture, which are in FASTQ file format, are mapped to the human reference genome (HG19) with BWA (Burrow-Wheeler Alignment) [4] and are formatted in a SAM (Sequence Alignment/Map) file format. Thus NA19238 and NA19240 have three SAM files each, and NA19239 has one SAM file. The SAM files of an individual are merged into a single SAM file by PICARD [12]. Then each SAM file is sorted by SAMtools [28] and re-formatted into BAM (Binary Alignment/Map) file format. The PCR (Polymerase Chain Reaction) duplicates are removed by PICARD. Base pair quality recalibration and local realignment are applied to the resulting BAM files using GATK as in [9].

Some of the reads in the BAM files are filtered out before making a pileup file. First, the mapped reads overlapping with the target region are selected. Then reads having mapping quality

score less than 30 are excluded. Lastly, mapped reads that include insertions or deletions are not included in the modeling. The number of filtered reads is summarized in Table 3.1. About 5% of the reads are filtered out. Final BAM files are made with the reads from the last column in Table 3.1.

Table 3.1 Number of Reads Before and After Filtering

Individuals	Average Coverage	# of Mapped Reads	# of Mapped Reads in Target Regions	After Excluding Reads with MAPQ<30	After Excluding Reads with MAPQ<30 & Indel Reads
NA19238 (YRI-mother)	251	209,310,307	143,168,295 (100%)	136,674,117 (95.46%)	136,178,282 (95.11%)
NA19239 (YRI-father)	87	66,198,431	48,113,846 (100%)	45,945,070 (95.49%)	45,776,295 (95.14%)
NA19240 (YRI-child)	221	189,850,073	126,594,931 (100%)	120,450,271 (95.14%)	119,993,854 (94.79%)

Note: “MAPQ” stands for Mapping Quality of a read. “Indel” reads are mapped reads containing insertions and/or deletions.

The filtered BAM files are put into the data processing pipeline described in section 1.3 to produce pileup files. Because some of the reads include out-of-target positions, only the positions in the target region are extracted from the pileup files. The count table of each allele is made from the in-target pileup files described in Section 1.3. The allele count table is separated depending on whether the alternative allele count is equal to zero or greater than zero. The AAPs are only measured when the count of the alternative allele is greater than zero.

The average coverage of a target region (i.e., exomic region) is defined as

$$\frac{\sum_{i=1}^{n} d_i}{n}$$

where n is the number of positions in the target region and d_i is the number of reads covered at the i-th position. The calculated average coverage of the target regions are 251x, 87x, and 221x for NA19238, NA19239, and NA19240, respectively. Due to different numbers of captures, the coverages of NA19238 and NA19240 are about three times the coverage of NA19239. Positions with depth less than 9 are excluded from the analysis, and genotypes are not assigned for these positions. The information about these low depth positions is exported separately. The set of positions covered 9 or more times are called "d9" positions here. Table 3.2 shows the number of positions in the pileup file and the number of d9 positions in target regions. Regardless of the average coverage, 97.01%, 94.58%, and 96.96% of the target positions are covered 9 or more times and those included in the analysis.

Table 3.2 Number of Positions in Pileup Files

Individuals	# of pileup positions	# of d9 positions in-target
NA19238	52,730,500	32,869,503
NA19239	51,847,965	32,045,518
NA19240	52,724,163	32,850,931

Note that the number of target positions is 33,881,597.

3.2 Methods

In this chapter, I use cluster analysis on the AAP values for initial grouping and fit the mixture of three normal distributions to the AAP values expressed in the logit scale, $\log\frac{x}{1-x}, 0 < x < 1$. The measured AAP values of the positions with denominator greater than 30 (named "d30"

positions) are used to calculate the estimates of the parameters in the model. The AAP values equal to 0 and 1 are set to be g0 genotype and g2 genotype, respectively, and not included in the analysis. Cluster analysis is then performed for the initial grouping of d30 positions after excluding positions with AAP values 0 or 1. Parameters of each component are estimated in each cluster. Subsequently, the Bayesian posterior probability is calculated for d9 positions using the parameters estimated from d30 positions. The number of positions for analysis is summarized in Table 3.3. The left column is the number of positions covered 30 or more times having AAP values in the open interval (0, 1) and the right column is the number of positions covered 9 or more times having AAP values in (0, 1). The numbers of d9 and d30 positions vary depending on the average coverage. The percentage of d30 positions is more than 99% for NA19238 and NA19240. For NA19239, the number of d30 positions is 94.48% of the number of d9 positions.

Table 3.3 Number of Positions Included in the Analysis

Samples	Number of d30 positions	Number of d9 positions
NA19238 (YRI-mother)	10,407,572	10,485,726
NA19239 (YRI-father)	2,872,582	3,040,116
NA19240 (YRI-child)	7,352,000	7,426,086

Note: The average coverage of each sample is 251x, 87x, and 221x, respectively.

3.2.1 CLARA Clustering

Under the diploid assumption, the measured AAP values from d30 positions are initially clustered into three components using CLARA (Clustering Large Applications) [17], that

represent the g0, g1, and g2 genotypes. From the clustering results, initial boundaries of measured AAP values for each genotype are determined. The boundary separating genotype g0 and g1 for sample k is called the lower boundary (L_k), and the boundary separating genotype g1 and g2 is called the upper boundary (U_k).

When the entire set of measured AAP values was put into CLARA, the classification was skewed to positions from g0 due to the large proportion of g0 genotypes [Table 3.4]. For NA19238 and NA19240, the lower boundary was less than 0.01 and the upper boundary was 0.01 and 0.08, respectively.

Table 3.4 Results of using CLARA to the Entire Set of AAP values

Samples	L	U
NA19238	0.0049	0.0143
NA19239	0.2853	0.7759
NA19240	0.0072	0.0811

3.2.2 Partitioned CLARA Clustering

The result in section 3.2.1 was different from the expectation that the AAP values from g1 positions were distributed around 0.5 and the AAP values from g2 positions were around 1. Scott and Symons [29] showed that based on their likelihood ratio criteria, when the separation between two clusters was not large or when there was only one underlying cluster, there is a tendency to cluster into two evenly split groups. Thus when two clusters had equal proportions, the clustering performed best. Garcia-Escudero et al. [30] also indicated that the k-means

algorithm, which is close to the k-medoids algorithm, was optimal for clustering groups of roughly equal size. Wu and Yang [31] reported that a cluster with a greater number of observations may be split incorrectly when there was a great difference in the number of observations in each cluster. To determine clusters with less bias against the g0 genotype, I use subsets of AAP values, S_1 and S_2, for clustering:

$$S_1 = \{x \,|c_{1k} < x < 0.5\}, S_2 = \{x|c_{2k} < x < 1\}$$

where x represents an AAP value and c_{1k} is a positive real number between 0 and 0.5 and c_{2k} is a positive real number between 0.5 and 1. The AAP values in S_1 are clustered into two components to get L_k separating g0 and g1 genotypes. Those in S_2 are clustered into two components to get U_k separating g1 and g2 genotypes. The constant c_{1k} may vary for each individual and is determined by the following procedure:

A. Initially let $S_1^0 = \{x\big|c_{1k}^0 < x < 0.5\}$, where $c_{1k}^0 = 0.15$

B. Cluster S_1^0 into two groups using CLARA and calculate an initial value for L_k (l_k^0).

C. Define $T_1(c_{1k}^0, l_k^0) = \{x|c_{1k}^0 < x \leq l_k^0\}$ and $T_2(l_k^0) = \{x|l_k^0 < x < 0.5\}$. Calculate the number of elements: $n_1 = n(T_1)$ and $n_2 = n(T_2)$.

D. If $n_1 \neq n_2$, change c_{1k}^0 to c_{1k}^1 such that $\left|n\left(T_1(c_{1k}^1, l_k^0)\right) - n\left(T_2(l_k^0)\right)\right|$ is minimized. Otherwise, let $c_{1k}^1 = c_{1k}^0$.

Once c_{1k}^1 is calculated, $S_1 = \{x|c_{1k}^1 < x < 0.5\}$ is clustered into two groups for the final value of the lower boundary, l_k for the k-th sample. The procedure for determining the upper boundary, U_k, is similar to that for L_k. That is,

A. Initially let $S_2^0 = \{x \big| c_{2k}^0 < x < 1\}$, where $c_{2k}^0 = 0.5$

B. Cluster S_2^0 into two groups using CLARA and calculate initial U_k, u_k^0.

C. Define $T_1(c_{2k}^0, u_k^0) = \{x | c_{2k}^0 < x \leq u_k^0\}$ and $T_2(u_k^0) = \{x | u_k^0 < x < 1\}$. Calculate the number of elements: $n_1 = n(T_1)$ and $n_2 = n(T_2)$.

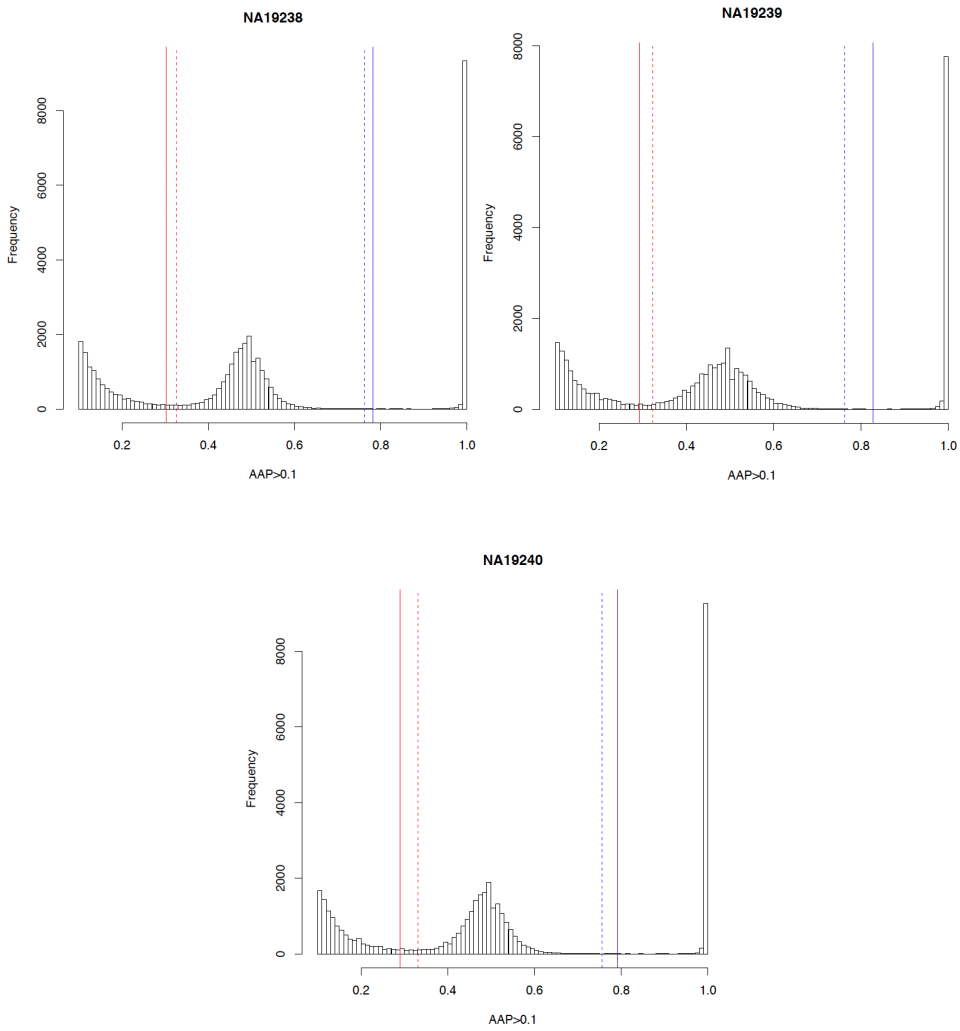
D. If $n_1 \neq n_2$, change c_{2k}^0 to c_{2k}^1 such that $\left| n\left(T_1(c_{2k}^1, u_k^0)\right) - n\left(T_2(u_k^0)\right) \right|$ is minimized. Otherwise, let $c_{2k}^1 = c_{2k}^0$.

Similarly, $S_2 = \{x \big| c_{2k}^0 < x < 1\}$ is clustered into two groups for the final value of U_k. The positions having measured AAP values between 0 and L_k are initially assigned to g0. Those having AAP values between L_k and U_k are assigned to g1. Those having AAP values between U_k and 1 are assigned to g2. Figure 3.1 shows the histogram of the measured AAP values and the boundaries. The red lines are the lower boundary, and the blue lines are the upper boundary. The solid lines are the final values of the boundaries, and the dashed lines are the initial values. The final value of L_k is lower than the initial lower boundary (l_k^0), and the final value of U_k is greater than the initial upper boundary (u_k^0) for each of the three subjects. The final values of L_k and U_k are summarized in Table 3.5. The lower boundary for the three samples ranges from 0.29 to 0.30, and the upper boundary from 0.78 to 0.82. The boundary values vary less than the boundary values obtained in section 3.2.1 across samples.

Table 3.5 Final Values of Boundaries

Samples	L	U
NA19238	0.3018	0.7813
NA19239	0.2921	0.8270
NA19240	0.2906	0.7913

Figure 3.1 Histograms of Measured AAP values with Boundaries



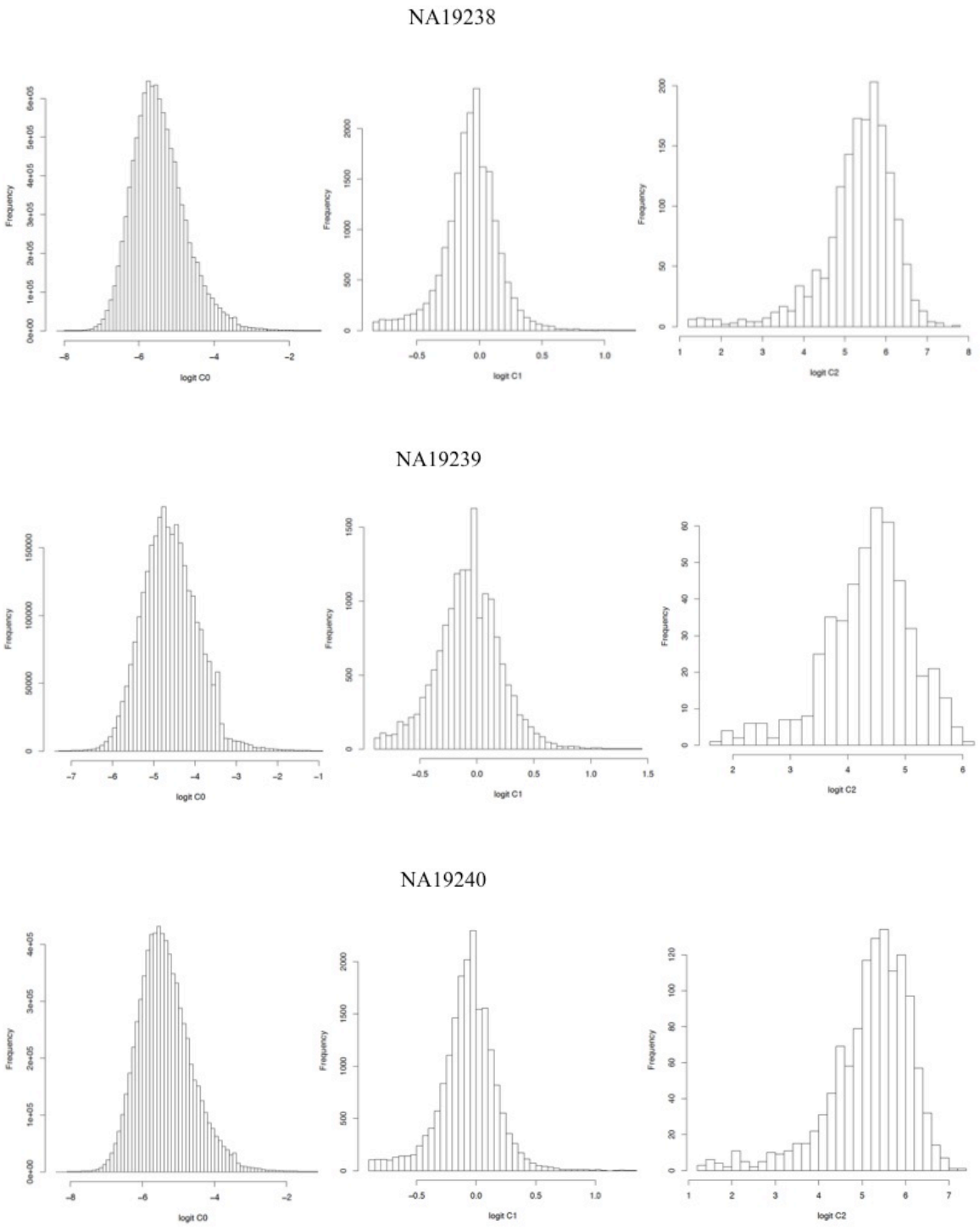
3.2.3 Genotype Likelihood Calculation with Normal Mixture Modeling

The measured AAP values of d30 positions are clustered into three components as

$$C_0 = \{x|0 < x \leq L_k\}, C_1 = \{x|L_k < x \leq U_k\}, C_2 = \{x|U_k < x < 1\}.$$

Component C_0 represents the *g0* genotype, C_1 for the *g1* genotype, and C_2 for the *g2* genotype. The AAP values from each component are expressed in logit scale. Let $y_{ik} = \text{logit}(x_{ik}), i = 1, \ldots, n$ for the *k*-th sample. The histogram of each component in logit scale is in Figure 3.2. The left panel is the histogram of C0, the center panel is the histogram of C1, and the right panel is the histogram of C2.

Figure 3.2 Histograms of Each Component of the Logit of AAP Values



I assume that each component of y_i follows approximately a normal distribution. In fact, each component is unimodal. I calculate the maximum difference between empirical cdf (cumulative distribution function) and normal cdf of each component in Table 3.6. The magnitude of difference varies for each component and sample. The C_0 cluster shows a relatively smaller difference between empirical and fitted cdf (i.e., within 0.05), while the C_2 cluster shows a larger difference (i.e., up to 0.10).

Table 3.6 Difference Between Empirical and Normal cdfs

	NA19238	NA19239	NA19240
C0	0.048	0.024	0.044
C1	0.065	0.029	0.066
C2	0.106	0.059	0.098

The pdf of y_i is given as

$$f(y_{ik}|\underline{\boldsymbol{\theta}}_k) = \sum_{j=0}^{2} p_{jk} g_j(y_{ik}|\mu_{jk}, \sigma_{jk}) \qquad (3.1)$$

where $\underline{\boldsymbol{\theta}}_k = (\mu_{0k}, \mu_{1k}, \mu_{2k}, \sigma_{0k}, \sigma_{1k}, \sigma_{2k}, p_{0k}, p_{1k})$ and $g_j(y_{ik}|\mu_{jk}, \sigma_{jk})$ is the pdf of normal distribution of the *j*-th component with mean μ_{jk} and standard deviation σ_{jk}. The parameter $p_j, j = 0,1,2$ is a mixing proportion such that $\sum_{j=0}^{2} p_j = 1$ and $0 < p_0, p_1 < 1$. The parameters are estimated in each cluster, and each mixing proportion is calculated as the proportion of each component. The estimated parameters are summarized in Table 3.7.

Table 3.7. The Parameter Estimates of Each Component

Sample	μ0	σ0	p0	μ1	σ1	p1	μ2	σ2	p2
NA19238	-5.4566	0.7168	99.8042%	-0.0720	0.2256	0.1804%	5.3149	0.9138	0.0154%
NA19239	-4.5974	0.6713	99.4170%	-0.0757	0.2872	0.5656%	4.3607	0.7706	0.0173%
NA19240	-5.3668	0.7471	99.7256%	-0.0718	0.2443	0.2578%	5.1801	0.9609	0.0166%

The estimates of the parameters vary somewhat across the individuals, but show similar patterns. The estimated means of C_0 and C_2, $\hat{\mu}_0$ and $\hat{\mu}_2$, are nearly symmetric around 0, ranging from -5.37 to -4.60 for $\hat{\mu}_0$ and from 4.36 to 5.31 for $\hat{\mu}_2$. The estimated mean of C_1 ($\hat{\mu}_1$) is slightly smaller than 0; zero in logit scale is equivalent to 0.5 in the measured AAP values. The sizes of the estimated standard deviations are different among three components, but the ordering of the magnitudes of $\hat{\sigma}_j$ is the same: $\hat{\sigma}_2 > \hat{\sigma}_0 > \hat{\sigma}_1$.

The genotype likelihood for the *k*-th sample is calculated using the estimated Bayesian posterior probability:

$$
\begin{aligned}
l_{g0}(y_{ik}|\hat{\underline{\boldsymbol{\theta}}}_k) &= \frac{\hat{p}_{0k} g_0(y_{ik}|\hat{\mu}_{jk}, \hat{\sigma}_{jk})}{\sum_{j=0}^{2} \hat{p}_{jk} g_j(y_{ik}|\hat{\mu}_{jk}, \hat{\sigma}_{jk})} \\
l_{g1}(y_{ik}|\hat{\underline{\boldsymbol{\theta}}}_k) &= \frac{\hat{p}_{1k} g_1(y_{ik}|\hat{\mu}_{jk}, \hat{\sigma}_{jk})}{\sum_{j=0}^{2} \hat{p}_{jk} g_j(y_{ik}|\hat{\mu}_{jk}, \hat{\sigma}_{jk})} \\
l_{g2}(y_{ik}|\hat{\underline{\boldsymbol{\theta}}}_k) &= \frac{\hat{p}_{2k} g_2(y_{ik}|\hat{\mu}_{jk}, \hat{\sigma}_{jk})}{\sum_{j=0}^{2} \hat{p}_{jk} g_j(y_{ik}|\hat{\mu}_{jk}, \hat{\sigma}_{jk})}
\end{aligned}
\tag{3.2}
$$

where $\hat{\underline{\boldsymbol{\theta}}}_k = (\hat{\mu}_{0k}, \hat{\mu}_{1k}, \hat{\mu}_{2k}, \hat{\sigma}_{0k}, \hat{\sigma}_{1k}, \hat{\sigma}_{2k}, \hat{p}_{0k}, \hat{p}_{1k})$.

3.2.4 Single Nucleotide Polymorphism Calling and Genotype Assignment

I assign a genotype to each position based on the clustering and genotype likelihoods in the previous sections. The probability of being a SNP (PSNP) at the *i*-th position is the sum of the Bayesian posterior probabilities for g1 and g2, i.e., $l_{g1}(y_i) + l_{g2}(y_i)$. The Bayesian posterior probability of having the g0 genotype is called PREF here. For convenience, I omit the position notation, *i*. A SNP is called at the *i*-th position of the *k*-th sample if the AAP value at the position is greater than L_k and PREF is less than 0.5. The g0 genotype is assigned for positions with AAP values less than or equal to L_k and PREF$\geq$ 0.5. There are positions that have AAP values greater than L_k but with PREF greater than or equal to 0.5. Also some positions have AAP values that are greater than L_k, but PREF is less than 0.5. For these positions, genotypes are not assigned and separately grouped into "ambiguous g0 or g1 (AMB01)".

For positions of sample *k* that have AAP values that are greater than L_k and PSNP is greater than PREF, either genotype g1 or g2 are assigned by the following rules:

A. If $L_k < x_{ik} \leq U_k$ and $l_{g1}(y_{ik}) \geq l_{g2}(y_i k)$, then the genotype at the *i*-th position is g1.

B. If $x_{ik} > U_k$ and $l_{g2}(y_{ik}) > l_{g1}(y_{ik})$, then the genotype at the *i*-th position is assigned g2.

There are some positions that have the measured AAP values greater than the upper boundary (U_k) but the Bayesian posterior probability of being g2 is smaller than that of being g1. Some positions have Bayesian posterior probability of being g2 is greater than that of being g1, but the measured AAP value is smaller than U_k. These positions are grouped separately and marked "ambiguous g1 or g2 (AMB12). A genotype is not called at such a position, but clearly the position is not genotype g0.

Based on the diploid assumption, I assume that there are two alleles, the reference allele (r_i) and the allele with the greatest alternative count (a_i). This assumption is based on the position that alternative alleles other than a_i are often the result of sequencing errors and that the counts of alternative alleles other than a_i are negligible. For some positions, however, the counts of other alternative alleles than a_i are not negligible. Among SNP positions, I grouped these positions as possible multiple allele positions ("PM") when the allele with the third greatest count is greater than a third of the depth. For example, suppose that there is a position having depth 50 and the reference allele is A. If the reference allele appeared 10 times, allele C appeared 20 times, and allele T appeared 20 times, this position would be marked PM.

3.3 Results

Using the clustering analysis and mixture modeling, I categorized all the d9 positions into six groups; Genotypes g0, g1, and g2, AMB01 (i.e., ambiguous g0 or g1), AMB12 (i.e., ambiguous g1 or g2), and PM (i.e., possible multiallelelic). The steps from pileup generation to assigning genotypes are packaged and named "SNVclust". The results of SNVclust are summarized in Table 3.8. The number of g0 positions is the sum of the number of positions having alternative allele count equal to zero and the number of positions assigned to be g0 from section 3.2.4. Similarly, the number of g2 positions is the sum of the number of positions having alternative allele proportion equal to one and the number of positions assigned to be g2 from section 3.2.4. The sum of the numbers assigned to the six categories (g0, g1, g2, AMB01, AMB12, and PM) is the number of d9 positions. The SNP positions include g1, g2 AMB12, and PM positions. I called 29,690 SNP positions for NA19238, 29,173 SNP positions for NA19239, and 30,034 SNP

positions for NA19240. That is 0.09% of d9 positions for the three individuals, cumulatively. The proportion of SNPs here is close to the known SNP frequency that a SNP occurs in every 1,000 base pairs of DNA sequence; i.e., 0.1% [14, 32]. The number of g1 positions is about two times the number of g2 positions. Similar results were found in chapter 2 (See Table 2.6 for members of YRI trio), where the genotype fraction of g1 was about twice the fraction of genotype g2. Although the total number of d9 positions of NA19239 is somewhat smaller than for the other two individuals, NA19239 has more ambiguous positions than the others for both AMB01 and AMB12, possibly because higher coverage may reduce the number of ambiguous positions.

Table 3.8. The Summary of SNP Calling and Genotyping

Samples	NA19238	NA19239	NA19240
Average Coverage	251x	87x	221x
Number of Positions Not in D9	1,012,138	1,836,123	1,030,710
Number of D9 Positions	32,869,459	32,045,474	32,850,887
Genotype g0	32,838,473	32,014,732	32,819,689
Genotype g1	19,595	19,348	19,902
Genotype g2	10,065	9,789	10,099
Ambiguous g0 or g1	1,296	1,569	1,164
Ambiguous g1 or g2	14	25	23
Possible Multiple Alleles	16	11	10
Number of SNP calls	29,690	29,173	30,034
Proportion of SNP calls from D9 Positions	0.09%	0.09%	0.09%

Note that the number of target positions is 33,881,597

Table 3.9 shows the ranges of the measured AAP values in g0, g1, g2, AMB01, and AMB12 categories. Overall the measured AAP values from g0 positions range from 0 to 0.23. The AAP values from g1 positions are roughly from 0.30 to 0.76 (0.79 for NA19239). The AAP values from g2 positions are from 0.78 (0.83 for NA19239) to 1. Sample NA19239 has wider range of the AAP values from g1 positions and shorter range of the AAP values from g2 positions than the other two samples. Positions having AAP values roughly from a quarter to one third are assigned to the category of ambiguous genotype g0 or g1. Measured AAP values roughly from three quarters to four fifths for NA19238 and NA19240 and from four fifths to five sixths for NA19239 are grouped into the category of ambiguous genotype g1 or g2.

Table 3.9 Ranges of AAP values in Each Category

	NA19238 (Mother)		NA19239 (Father)		NA19240 (Child)	
	min	max	min	max	min	max
Genotype g0	0.0000	0.2313	0.0000	0.2237	0.0000	0.2308
Ambiguous g0 or g1	0.2315	0.3017	0.2238	0.2920	0.2310	0.2906
Genotype g1	0.3019	0.7561	0.2921	0.7917	0.2907	0.7619
Ambiguous g1 or g2	0.7595	0.7798	0.7939	0.8235	0.7647	0.7895
Genotype g2	0.7828	1.0000	0.8333	1.0000	0.7931	1.0000

In order to compare the results from SNVclust to results from GATK and SAMtools, I generated SNP calls and genotypes for the same three individuals using these procedures. The SNP positions from each of three methods were extracted and then compared with the released SNP calls from 1000 Genomes Project (1KG) pilot 2 [33]. The 1000 Genomes Project calls used in this section were from deep whole genome sequencing (WGS) of the YRI trio with 21.8x, 26.4x, and 34.7x coverage, respectively, for NA19238, NA19239, and NA19240.

The comparisons are summarized in Table 3.10. The number of SNP calls from SNVclust was greater than the numbers from GATK and SAMtools. The number of SNP calls from SNVclust was between 1,671 and 2,492 more than the calls from GATK. SNVclust made 8,577 to 10,578 more SNP calls than SAMtools. There were 24,831 SNP positions, 26,462 SNP positions, and 26,883 SNP positions reported in the 1,000 genomes project pilot 2 study. SNVclust found 97.60% of the 1KG calls on average. That is, the sum of the three individuals' numbers of overlapping positions with 1KG calls (the third row of Table 3.10) divided by the sum of total numbers of 1KG calls (the first row of Table 3.10). GATK found 95.91% of the 1KG calls on average, and SAMtools 59.15% on average. The fractions of the overlaps with 1KG calls for each individual are written in the parenthesis. The SNVclust found a greater fraction of 1KG SNP positions than other two methods.

There were SNP positions found by SNVclust that were not included in 1KG call sets. Among these positions, on average 63.4% of them were found by either GATK or SAMtools or both. Similarly, I calculated the average percentage of SNP positions from GATK or SAMtools that were not in 1KG calls but found in the other two methods: 96.59% of SNP positions from GATK not in 1KG calls were found by either SNVclust or SAMtools or both. 53.25% of SNP positions from SAMtools not in 1KG calls were found by either SNVclust or GATK or both. These differences may be due to different sequencing strategies (i.e., WGS or WES), different sequencing times, or the different sequencing technologies. The fraction of positions that only SNVclust called a SNP was 5.18% out of total SNP positions called from SNVclust on average. Similarly, the fraction of positions that only GATK called a SNP is 0.30% on average. The fraction that only SAMtools called a SNP was 10.43% on average. These positions might be possible false discoveries.

Table 3.10 Comparison of SNP Calls from SNVclust with Other Methods

	NA19238	NA19239	NA19240
Total 1KG calls	24,831	26,462	26,883
SNVclust SNP calls	29,690	29,173	30,034
Overlap with 1KG calls	24,448 (98.46%)	25,429 (96.10%)	26,420 (98.28%)
Not in 1KG calls	5,242	3,744	3,614
Overlap with One of Other Methods	3,781	2,165	2,045
SNVclust only	1,461	1,579	1,569
GATK SNP calls	27,272	27,502	27,542
Overlap with 1KG calls	23,839 (96.00%)	25,385 (95.93%)	25,752 (95.79%)
Not in 1KG calls	3,433	2,117	1,790
Overlap with One of Other Methods	3,359	2,013	1,718
GATK only	74	104	72
SAMtools SNP calls	19,498	20,596	19,456
Overlap with 1KG calls	14,379 (57.91%)	16,301 (61.60%)	15,574 (57.93%)
Not in 1KG calls	5,119	4,295	3,882
Overlap with One of Other Methods	3,096	2,104	1,882
SAMtools only	2,023	2,191	2,000

Note: 1KG stands for 1000 Genomes Project.

3.4 Analysis of Coverage and Variant Detection

Even though the overall depth for NA19239 was substantially lower than for the other two, the number of calls for NA19239 was only 2.3% less than the average number of calls for the other two, possibly because the coverage of NA19239 may be high enough for genotyping with the SNVclust algorithm. In order to find the minimal coverage for SNVclust, I randomly selected a range of number of reads in NA19240 (YRI-child) to generate different average coverages (See Table 3.10). Then the SNVclust described in section 3.2.2-3.2.4 was applied for SNP calling.

The total number of reads in NA19240 was 195,317,068, and the average coverage for this individual was 221x. I randomly selected the reads so as to obtain 5x to 220x of average coverage in increments of 5x. That is,

$$\frac{Total\ Number\ of\ Reads}{221} \times Target\ Coverage.$$

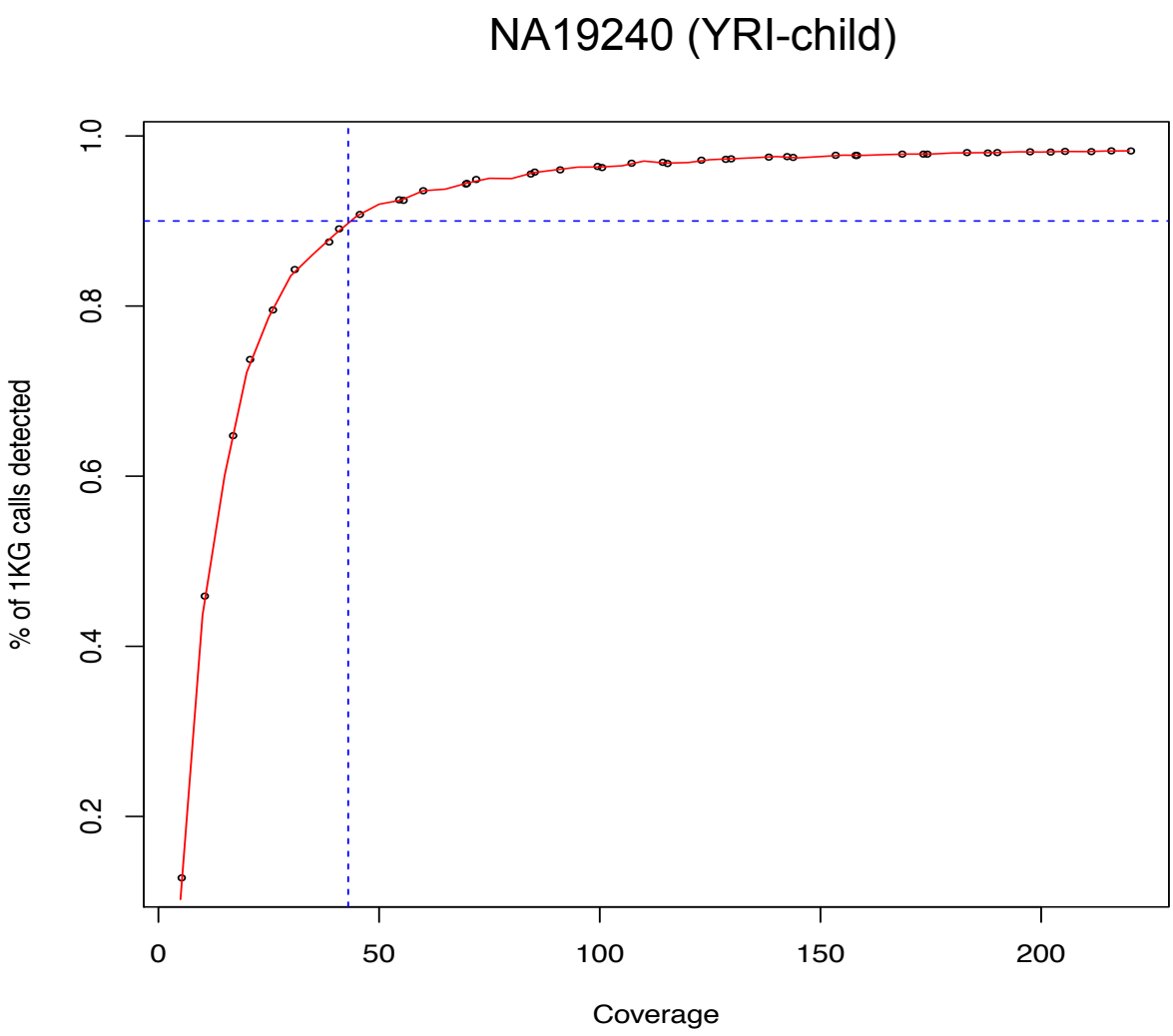
Since some of the reads were not mapped to chromosome 1 to 22, the target coverage was not exactly achieved, but the obtained average coverage was close to the target coverage. With the selected reads, I generated BAM files and Pileup files for the alternative allele counts, and the algorithm was applied. The output SNP calls were then compared with 1KG calls and the overlapping rate was calculated.

Table 3.11 shows the number of calls from BAM files of each coverage generated from SNVclust and its number of overlapping SNP positions with 1KG calls. As the average coverage increased, the number of d9 positions increased, and thus the number of calls of SNVclust increased. Also the number of calls that overlaps with 1KG calls increased as the average coverage grew. Figure 3.3 is the plot of the percentage of the number of calls overlapped with 1KG calls versus the average coverage. The blue horizontal line is drawn at 90% and the blue vertical line is drawn at its corresponding coverage. At 43x, SNVclust found 90% of 1KG SNP positions, where the total number of 1KG calls of NA19240 was 26,883. It suggests that approximately 43x coverage is needed to achieve 90% sensitivity.

Table 3.11 Analysis of Coverage

Target Coverage	Actual Average Coverage	Number of SNP calls	Number of Overlaps with 1KG calls	Number of D9 Positions
5	5.28	3,676	3,435	6,025,363
10	10.51	13,722	12,340	15,720,940
15	16.92	19,089	17,409	22,442,925
20	20.78	22,089	19,818	24,882,537
25	25.95	24,021	21,382	26,882,008
30	30.91	25,571	22,660	28,276,498
35	38.70	26,153	23,531	29,544,940
40	40.95	26,991	23,945	29,846,426
45	45.64	27,513	24,401	30,392,458
50	54.56	27,776	24,863	30,989,819
55	55.55	28,093	24,847	31,007,266
60	60.01	28,390	25,151	31,300,533
65	69.88	28,371	25,382	31,615,166
70	69.70	28,647	25,362	31,603,964
75	72.00	28,711	25,507	31,734,637
80	85.27	28,859	25,740	31,973,661
85	84.38	28,956	25,679	31,962,318
90	91.04	28,961	25,811	32,120,955
95	100.52	29,004	25,884	32,201,911
100	99.51	29,350	25,918	32,202,562
105	107.21	29,312	26,020	32,325,866
110	115.35	29,236	26,011	32,366,860
115	114.28	29,510	26,049	32,366,405
120	123.02	29,365	26,116	32,463,868
125	129.77	29,540	26,159	32,486,127
130	128.55	29,712	26,143	32,493,289
135	138.30	29,591	26,213	32,565,341
140	143.81	29,542	26,204	32,568,126
145	142.45	29,814	26,229	32,579,127
150	153.45	29,643	26,271	32,639,207
155	158.27	29,624	26,267	32,641,643
160	157.97	29,756	26,267	32,653,687
165	168.53	29,723	26,308	32,696,897
170	173.34	29,695	26,308	32,701,809
175	174.21	29,735	26,308	32,719,405
180	183.20	29,828	26,357	32,749,876
185	187.92	29,755	26,344	32,759,557
190	190.07	29,902	26,357	32,770,516
195	197.50	29,933	26,381	32,790,542
200	202.16	29,874	26,373	32,801,042
205	205.42	29,966	26,394	32,813,875
210	211.38	29,913	26,387	32,828,757
215	215.92	30,031	26,412	32,838,637
220	220.34	29,923	26,408	32,848,586

Figure 3.3 Number of Overlaps with 1KG calls vs. Average Coverage



3.5 Conclusion and Discussion

In this chapter, I combined two approaches of clustering and mixture modeling to genotype an individual with NGS data. The AAP value at a position was used for measuring the signal of SNP at a position. After determining the lower and upper boundaries via CLARA clustering, I calculated the Bayesian posterior probability and used it as the genotype likelihood. Based on clustering results and genotype likelihoods, I made a call set of SNP positions and genotyped each d9 position. Then I compared my call set with three external sources, and estimated that the average coverage should be at least 43x for WES data using random read selection.

All parameters needed for the analysis were derived directly from the data of each individual rather than from external information, which should minimize sample-specific biases. For example, among the three individuals analyzed in this chapter, one individual had significantly lower coverage than others. Using internally estimated parameters, SNVclust could successfully genotype the individual with quality roughly comparable to the other members of the trio measured with higher coverage. Importantly, it suggests that batch effects can be successfully removed by calculating the ratio of AAP as a means of normalization indicating the possibility of application of SNVclust to NGS data collected at different times with different technologies.

SNVclust provides information about positions with ambiguous genotypes as well as all of three genotypes. Goldstein et al. [34] described the need of information on reference positions as implemented in gVCF (genome Variant Call Format) file instead of using VCF file (https://sites.google.com/site/gvcftools/). SNVclust includes AAP value and the Bayesian posterior probabilities of having each of g0, g1, and g2 genotypes at each position in the outputs.

A limitation expected in SNVclust is that it is not suitable for data of mixed cell populations such as one from a tumor because it is unlikely that the mixture distribution of AAP values would be as clearly shown as one from homogenous cell populations. The basic assumption of the SNVclust is that the logit of AAP values in each cluster follows a normal distribution. This assumption needs to be assessed before applying the algorithm.

Finally, this method is scalable for large-scale genomic studies as it does not require large memory and computes quickly, especially through parallel processing in high performance cluster computers (HPCC) or cloud computing environment.

Chapter 4 Summary and Future Studies

4.1 Summary

I used the alternative allele proportion measured at each position for modeling. First, I modeled the measured AAP values on chromosome 1 of six individuals from 1000 Genomes Project with seven components, five continuous distributions with two point distributions. This final model well described the patterns of AAP distribution. The estimated parameters in the final model were used for estimating genotype fractions. Second, I developed software package, SNVclust, for calling SNP positions and genotyping. Cluster analysis and mixture modeling were applied together to assign Bayesian posterior probability at each position. The calls from SNVclust found more SNP positions from 1000 Genomes Project positions than GATK and SAMtools.

4.2 Mendelian Inconsistency

Positions that do not follow Mendelian inheritance are of interest in studies of diseases such as autism. Specially de novo mutations, those that the parents do not have but that are in an affected child, are of great interest [35, 36]. As discussed in Chapter 2, genotype fractions of a child were different than the fractions expected from the genotype fractions of the parents under a binomial model. The fraction of g2 genotypes was more than expected from the parents. By applying the algorithm in Chapter 2 to larger number of unaffected trios, I would like to study how much Mendelian inconsistency can occur in normal families.

4.3 Extension to Chromosomes X and Y

The distributions and the algorithms described in chapter 2 and 3 are based on the diploid assumption. For females, it can be easily extended to chromosome X. However, male DNA has one copy of chromosome X and one copy of chromosome Y. A different number of mixture distributions may need to be applied for modeling AAP values from these chromosomes.

4.4 Ambiguous Genotypes

There were considerable numbers of positions having the AAP values around 0.25 and 0.75 in the histograms shown in chapter 2 and 3. The meanings of these positions need to be studied. If those positions are coming from sequencing or alignment errors, modeling such errors separately might increase the accuracy of genotyping in NGS data. Such positions might also come from the effect of other types of genomic variation such as copy number variations.

4.5 Whole Genome Sequencing Application

In Chapter 2 and 3, I assumed a continuous distribution of AAP values, based on the assumption that the denominators of AAP values are large enough to make the assumption of a continuous distribution of AAP values satisfactory. In WGS data, however, the coverage might not be as high as WES and also may not vary as widely as in WES. Thus different distributions may need to be studied when the coverage is not high.

References

1. Collins, F.S., M.S. Guyer, and A. Charkravarti, *Variations on a theme: cataloging human DNA sequence variation.* Science, 1997. **278**(5343): p. 1580-1.

2. Raphael, B.J., *Chapter 6: Structural variation and medical genomics.* PLoS Comput Biol, 2012. **8**(12): p. e1002821.

3. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Research, 2008. **18**(11): p. 1851-8.

4. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

5. Li, R., et al., *SOAP: short oligonucleotide alignment program.* Bioinformatics, 2008. **24**(5): p. 713-4.

6. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

7. Li, H., *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.* Bioinformatics, 2011. **27**(21): p. 2987-93.

8. Koboldt, D.C., et al., *VarScan: variant detection in massively parallel sequencing of individual and pooled samples.* Bioinformatics, 2009. **25**(17): p. 2283-5.

9. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Research, 2010. **20**(9): p. 1297-303.

10. Shoemaker, J.S., I.S. Painter, and B.S. Weir, *Bayesian statistics in genetics: a guide for the uninitiated.* Trends in genetics : TIG, 1999. **15**(9): p. 354-8.

11. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nature Genetics, 2011. **43**(5): p. 491-+.

12. Ewing, B., et al., *Base-calling of automated sequencer traces using phred. I. Accuracy assessment.* Genome Research, 1998. **8**(3): p. 175-85.

13. Ewing, B. and P. Green, *Base-calling of automated sequencer traces using phred. II. Error probabilities.* Genome Research, 1998. **8**(3): p. 186-94.

14. Li, R., et al., *SNP detection for massively parallel whole-genome resequencing.* Genome Research, 2009. **19**(6): p. 1124-32.

15. Goya, R., et al., *SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors.* Bioinformatics, 2010. **26**(6): p. 730-6.

16. Morin, R., et al., *Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing.* BioTechniques, 2008. **45**(1): p. 81-94.

17. Kaufman, L.a.R., P.J., *Finding Groups in Data: An Introduction to Cluster Analysis*. 1990: Wiley.

18. Kaufman, L., Rousseeuw, P., *Clustering by Means of Medoids*. Vol. 87. 1987: Fac., Univ., 1987.

19. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K, *cluster: Cluster Analysis Basics and Extensions*, 2013.

20. Parla, J.S., et al., *A comparative analysis of exome capture.* Genome Biol, 2011. **12**(9): p. R97.

21. Ospina, R. and S.L.P. Ferrari, *Inflated Beta Distributions.* Statistical Papers, 2010. **51**(1): p. 111-126.

22. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm (with discussion).* Journal of the Royal Statistical Society. Series B (Methodological), 1977. **39**(1): p. 1-38.

23. McLachlan, G. and D. Peel, *Finite mixture models*. Vol. 299. 2000: Wiley-Interscience.

24. Team, R.C., *R: A language and environment for statistical computing*, 2012, R Foundation for Statistical Computing, Vienna, Austria.

25. Horvath, A., Telek, M., *Approximating heavy tailed behavior with phase type distribution.* Advances in Algorithmic Methods for Stochastic Models, Notable Publications, 2000: p. pp. 191–214.

26. Riska, A., V. Diev, and S. E., *An EM-based technique for approximating long-tailed data sets with PH distributions.* Performance Evaluation, 2004. **55**(1): p. 147-164.

27. Venables, W.N. and B.D. Ripley, *Modern Applied Statistics with S*, 2002, Springer, New York.

28. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

29. Scott, A.J. and M.J. Symons, *Clustering Methods Based on Likelihood Ratio Criteria.* Biometrics, 1971. **27**(2): p. 387-&.

30. Garcia-Escudero, L.A., et al., *A general trimming approach to robust cluster analysis.* Annals of Statistics, 2008. **36**(3): p. 1324-1345.

31. Wu, K.L., Yang, M.S., *Alternative c-means clustering algorithms.* Pattern Recognition, 2002. **35**: p. 2267 – 2278.

32. Sachidanandam, R., et al., *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.* Nature, 2001. **409**(6822): p. 928-33.

33. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

34. Goldstein, D.B., et al., *Sequencing studies in human genetics: design and interpretation.* Nat Rev Genet, 2013. **14**(7): p. 460-70.

35. Awadalla, P., et al., *Direct measure of the de novo mutation rate in autism and schizophrenia cohorts.* Am J Hum Genet, 2010. **87**(3): p. 316-24.

36. Neale, B.M., et al., *Patterns and rates of exonic de novo mutations in autism spectrum disorders.* Nature, 2012. **485**(7397): p. 242-5.

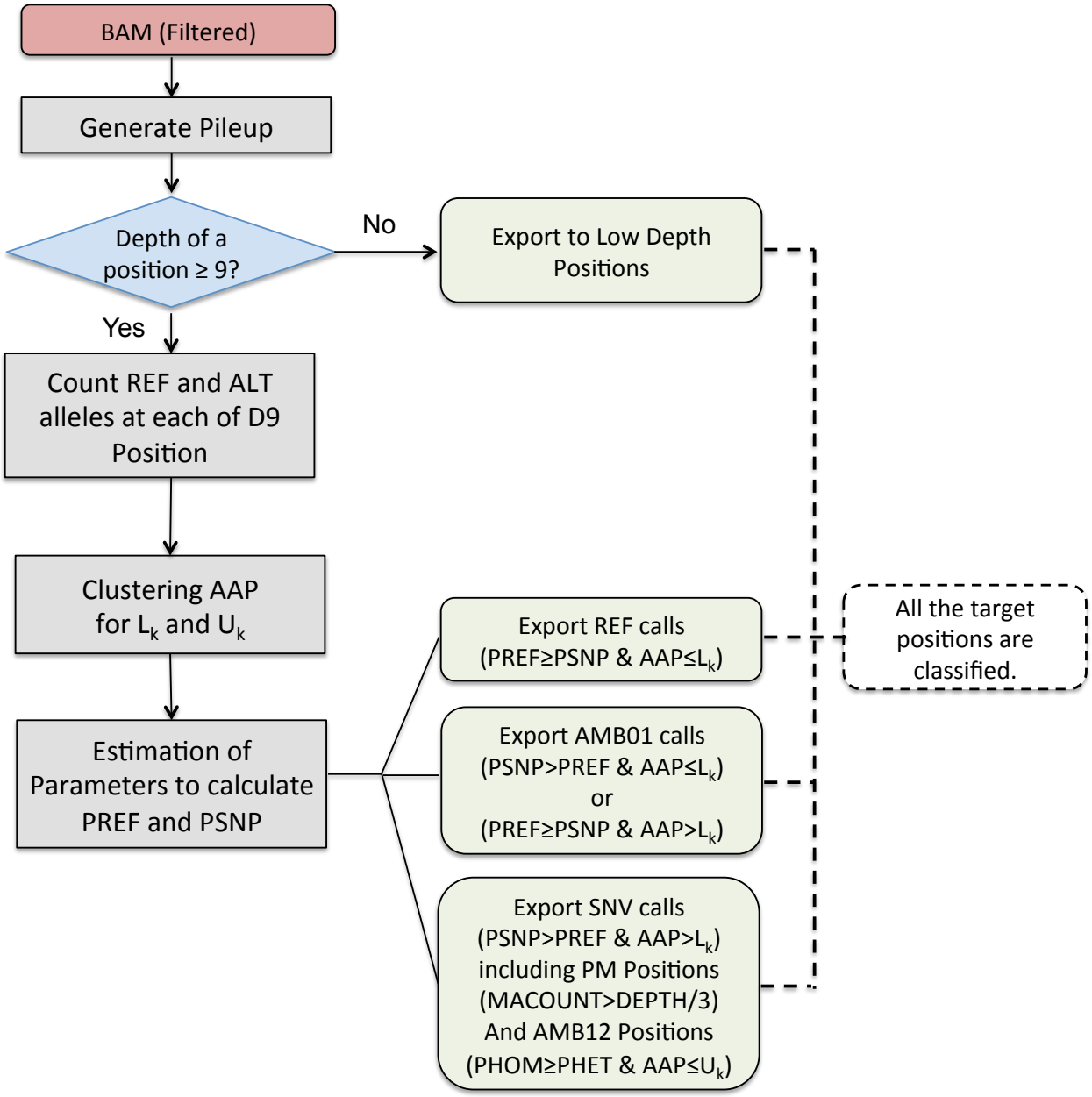
Appendices

A. Checking the Second Derivative Conditions for Maximum for Model 1

$$
\begin{aligned}
\left.\frac{\partial^2 l}{\partial \mu_0^2}\right|_{\mu_0=\hat{\mu_0}} &= \frac{1}{\hat{\mu_0}^2 \sum_{i=1}^{n} z_{i0}} - \frac{2}{\hat{\mu_0}^3} \sum_{i=1}^{n} z_{i0} x_i \\
&= -\frac{(\sum_{i=1}^{n} z_{i0})^3}{(\sum_{i=1}^{n} z_{i0} x_i)^2} < 0 \\
\left.\frac{\partial^2 l}{\partial \mu_{1l}^2}\right|_{\mu_{1l}=\hat{\mu_{1l}}} &= \frac{1}{\hat{\mu_{1l}}^2 \sum_{i=1}^{s} z_{i1l}} - \frac{2}{\hat{\mu_{1l}}^3} \sum_{i=1}^{s} z_{i1l}(0.5 - x_i) \\
&= -\frac{(\sum_{i=1}^{s} z_{i1l})^3}{(\sum_{i=1}^{s} z_{i1l}(0.5 - x_i))^2} < 0 \\
\left.\frac{\partial^2 l}{\partial \mu_{1r}^2}\right|_{\mu_{1r}=\hat{\mu_{1r}}} &= \frac{1}{\hat{\mu_{1r}}^2 \sum_{i=s+1}^{n} z_{i1r}} - \frac{2}{\hat{\mu_{1r}}^3} \sum_{i=s+1}^{n} z_{i1l}(x_i - 0.5) \\
&= -\frac{(\sum_{i=s+1}^{n} z_{i1r})^3}{(\sum_{i=s+1}^{n} z_{i1r}(x_i - 0.5))^2} < 0 \\
\left.\frac{\partial^2 l}{\partial \mu_2^2}\right|_{\mu_2=\hat{\mu_2}} &= \frac{1}{\hat{\mu_2}^2 \sum_{i=1}^{n} z_{i2}} - \frac{2}{\hat{\mu_2}^3} \sum_{i=1}^{n} z_{i1l}(1 - x_i) \\
&= -\frac{(\sum_{i=1}^{n} z_{i2})^3}{(\sum_{i=1}^{n} z_{i2}(1 - x_i))^2} < 0
\end{aligned}
$$

B. Workflow Chart of SNVclust Pipeline

Figure B.1 Workflow Chart of SNVclust



List of Abbreviations in Figure B.1:

REF: a reference allele

ALT: the alternative allele

L_k: Lower boundary of the *k*-th sample

U_k: Upper boundary of the *k*-th sample

PREF: Bayesian posterior probability of having g0 genotype at a position

PHET: Bayesian posterior probability of having g1 genotype at a position

PHOM: Bayesian posterior probability of having g2 genotype at a position

PSNP: Bayesian posterior probability of having g1 or g2 genotype at a position. That is the sum of PHET+PHOM

SNV: Single Nucleotide Variants