Stony Brook University

Statistical Models for SNP Detection

A Dissertation Presented

by

Shengnan Cai

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

December 2010

Stony Brook University

The Graduate School

Shengnan Cai

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Dr. Honshik Ahn – Dissertation Advisor
Professor, Department of Applied Mathematics and Statistics

Dr. Nancy Mendell – Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Dr. Stephen Finch – Member
Professor, Department of Applied Mathematics and Statistics

Dr. Sangjin Hong – Outside Member
Associate Professor, Department of Electrical and Computer Engineering

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Statistical Models for SNP Detection

by

Shengnan Cai

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

2010

Variations in DNA sequences of humans have a strong association with many diseases. Single Nucleotide Polymorphism (SNP) is the most common type of DNA variations. Our research is to detect SNPs from the data generated by Polymerase Chain Reaction (PCR) and next generation sequencing methods. In the first part of the study, we had a relatively small data set with fewer known SNPs as the training data. We developed a classification model based on the cross validation method. From the first part of the research, we gained knowledge of the properties of the data. In the next phase, we obtained a much larger data set with a much larger group of known SNPs. We developed eight measures for every genetic position with these data. Using these eight measures as the predictor variables, we applied several classification methods such as Random Forest (RF), Support Vector Machines (SVM), Single Decision Tree (ST) and Logistic

Regression (LR); then used cross validation to evaluate these classification methods. By comparing the predictive accuracy, sensitivity and specificity, we found the best performing model for the data. To compare the performances of these models while the number of observations for each genetic position (cover depth) is small, we randomly drew out subsets from the whole data and applied these classification models. Variable selection is also used to our study. The result shows, SVM using the selected variables has a significant higher average accuracy than the other methods in general, but RF using the selected variables performs the best when the cover depth is as small as 20.

Table of Contents

List of Figures

List of Tables

Chapter 1

Background

1.1 SNPs and Traditional Approaches to Detect SNPs

The genetic sequences of different people are remarkably similar. When the chromosomes of two humans are compared, their DNA sequences can be identical for thousands of bases. But at about one in every 1,000 bases, on average, the sequences will differ (Cooper et al., 1985). Differences in individual bases are by far the most common type of genetic variation (Lander, 1996).

A single nucleotide polymorphism (SNP, pronounced "snip") is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual) (Wang et al., 1998; Cargill et al., 1999; Halushka et al., 1999). For example, two sequenced DNA fragments from

different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide. In this case we say that there are two alleles: C and T. Almost all common SNPs have only two alleles.

Variations in the DNA sequences of humans can affect how humans develop diseases and respond to pathogens, chemicals, drugs, vaccines, and other agents. SNPs are the most common type of genetic variation (Risch et al., 1996; Chakravarti, 1999). Moreover, SNPs have a very low mutation rate, which is the number of mutations per generation expressed as a decimal value or a percentage. The general average mutation rate was estimated to be only between 2.5×10^{-8} and 4.4×10^{-8} per base pair per generation (Pitman 2001). Since SNPs are mostly inherited from parents, two related individuals have relatively small difference in their SNPs. However, the whole human population is huge; hence there are a large number of SNPs in the genome. Nowadays, millions of SNPs are documented. Hence, SNPs are thought to be the ideal markers for the dissection of complex traits in association studies and linkage disequilibrium mapping (Collins et al., 1997).

For most of the current research, researchers focus on finding the association between genotype and disease phenotype. As described above, SNPs are ideal to be the genetic markers for disease diagnosis. Hence it is necessary to identify very large numbers of SNPs. Conventionally, direct sequencing methods, especially gel-based fluorescent sequencing methods, are widely used to identify the large-scale SNP. In many recent strategies, overlapping sequences from multiple individuals are computationally aligned to identify high-quality mismatches (Taillon-Miller et al., 1998). Information about alignment depth, sequence context and read quality is summarized to identify if the site is a SNP or a sequencing error. Base-calling quality analysis programs have been widely developed to

obtain such information (Ewing et al., 1998). For example, PolyPhred which is integrated with the use of three other programs: Phred (Brent Ewing and Phil Green), Phrap (Phil Green), and Consed (David Gordon and Phil Green) is a typically traditional recently developed base-calling program. It compares fluorescence-based sequences across traces obtained from different individuals to identify heterozygous sites for single nucleotide substitutions (see URL: http://chum.gs.washington.edu/).

Using the primary data generated by the Human Genome Project, 75% of the 1,500,000 SNPs currently in the database dbSNP were identified from overlapping regions of genomic clones (Carlson et al., 2001). These SNPs are clustered in 10-50 kb regions. A much smaller fraction of dbSNP (4% of all dbSNP entries) comes from similar data-mining efforts using single-pass sequence reads from expressed sequence tags (ESTs) as the raw data (Clifford et al., 2000). Although the large-scale SNP identification efforts generate most of the SNPs in dbSNP, directed SNP identification within candidate genes also provides a small but important fraction of SNPs (Cargill et al., 1999). Polymerase chain reaction (PCR) then is introduced to amplify the genomic DNA for targeted SNP discovery (Nakajima et al., 1998). Once amplified, many techniques are available to scan for SNPs in the PCR products obtained from different individuals (Kwok et al., 1994).

1.2 PCR and Next-generation Sequencing Method

Our data are from a targeted genomic DNA region, which is 55kb human PAK4 gene region. This region is related to neuronal disease, in which the

research interest lies. Our data are based on PCR and next generation sequencing technology. Since the next-generation sequencing method is pretty novel, our data format is distinct from others. The following are details of PCR and next-generation sequencing method.

PCR is a technique widely used in molecular biology. It derives its name from one of its key components, a DNA polymerase used to amplify a piece of DNA. This DNA polymerase enzymatically assembles a new DNA strand from DNA building blocks, the nucleotides, by using single-stranded DNA as a template and DNA oligonucleotides (also called DNA primers), which are required for initiation of DNA synthesis. The vast majority of PCR methods use thermal cycling, i.e., alternately heating and cooling the PCR sample to a defined series of temperature steps. These thermal cycling steps are necessary to physically separate the strands (at high temperatures) in a DNA double helix (DNA melting) used as template during DNA synthesis (at lower temperatures) by the DNA polymerase to selectively amplify the target DNA. The selectivity of PCR results from the use of primers that are complementary to the DNA region targeted for amplification under specific thermal cycling conditions. As PCR progresses, the DNA thus generated is itself used as a template for replication. Because both strands of DNA could be functioned as templates, this sets in motion a chain reaction in which the DNA template is exponentially amplified.

The term DNA sequencing encompasses biochemical methods for determining the order of the nucleotide bases. Since 1970s, sequencing methods have evolved from relatively labor-intensive gel-based procedures to modern automated protocols based on dye labeling and detection in capillary electrophoresis that permit rapid large-scale sequencing of genomes and transcriptomes. There are several sequencing methods in biological research

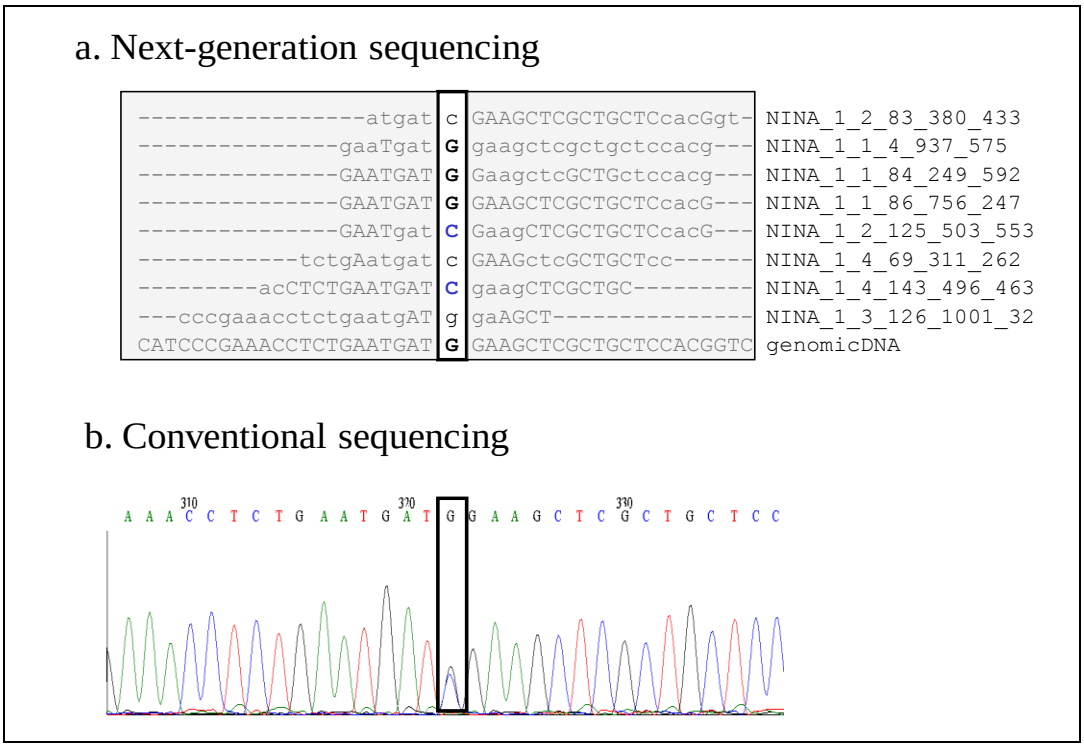
history, such as Maxam-Gilbert sequencing (Jou et al., 1972), Chain-termination methods, Large-scale sequencing strategies, high-throughput sequencing and next-generation sequencing.

'Next-generation' is used in reference to the various implementations of cyclic-array sequencing that have recently been realized in many biological investigations (Margulies et al., 2005). In our research, we get the sequencing data using reversible terminator chemistry. Single molecules of DNA are attached to a flat surface, amplified *in situ* and used as templates for synthetic sequencing with fluorescent reversible terminator deoxyribonucleotides. Images of the surface are analyzed to generate a high-quality sequence (Bentley et al., 2008). The major difference between conventional sequencing and next-generation sequencing technologies is that the former one uses the fluorescence-signals from a group of DNA molecules to detect SNPs (Kwok, 2001), while the latter one is a single molecule sequencing method and the SNPs are detected by the analysis of a set of genotypes from many molecules sequenced separately. Figure 1 shows the comparison of next-generation sequencing and conventional sequencing results. In Figure 1b, the lines with four different colors stand for four nucleotides. The height of the peaks is correlated to the signals of specific base call, which are generated from the whole group of experimented DNA molecules. Using this technology, we can only determine the base by the strength of the signals, but have no idea about the ratio of different nucleotide reads. However, we can get this kind of information by using the next-generation sequencing method. As shown in Figure 1a, since we analyze every DNA molecule separately, we get every sequence as a read to summarize the genotype and detect the SNP in further. Actually, the marked position in Figure 1 is a heterozygous SNP candidate, which can be detected by both approaches.

Moreover, in order to identify those disease associated SNP markers, we need a large enough sample size from both patients and normal controls with available genotypes. The conventional SNP detection method is too expensive in both cost and time, which limit efficient SNP detection. However, next-generation sequencing technology is more efficient. That is they could detect genotypes for each individual fast accurately with less cost and time.

Therefore, data analysis for SNP detection based on next-generation sequencing technology is feasible and valuable. Since the new SNP model is based on a newly developed technology, it requires building a new algorithm to analyze this new type of data.

Figure 1: SNPs detection example



Chapter 2

Data Format

2.1 Data Source

The data source is from the International HapMap Project, which is a multi-country effort to identify and catalog genetic similarities and differences in human beings. The project studies 270 DNA samples: 90 samples from a US Utah population with Northern and Western European ancestry (samples collected in 1980 by the Centre d'Etude du Polymorphisme Humain (CEPH)), and new samples collected from 90 Yoruba people in Ibadan, Nigeria (30 trios), 45 unrelated Japanese in Tokyo and 45 unrelated Han Chinese in Beijing. All donors gave specific consent for their inclusion in the project. Our data are generated from one person (CEU-NA12762) of CEPH. The CEPH samples are available from the non-profit Coriell Institute of Medical Research (International HapMap Consortium, 2003).

2.2 Raw Data

From the PCR and next-generation sequencing method, the data contain the following information. In Table 1, every row is a read of a typical position. The first column is the position; the second column is the reference base from the previous experiments; the third column is the location on the sequence which ranges from 0 to 31; the fourth column is the strand which stands for the direction of the sequence, either forward or backward; the fifth through eighth columns are the quality scores of the four nucleotides; the ninth column is the predicted base by comparing the scores; the last column is the highest score among the four scores.

Quality score (QS) is the measure used to evaluate the base call. In our data, it ranges from -40 to 40, with a higher value corresponding to a higher accuracy. The following formula can explain the relationship between QS and the probability that a base call is incorrect.

$$QS = -10\log_{10}[\frac{P(Error)}{1-P(Error)}] \ ,$$ where P(Error) is the probability of a wrong base call.

For example, if the quality score for A in a particular observation is 40, then the probability that this base is not A is approximately 0.0001. Every observation has four quality scores for the four possible nucleotides: A, T, C, or G. In all cases, there is only one positive score among the four scores. The second highest score is at most the negative absolute value of the highest score (HS). We use HS to predict the bases. In the later analysis, we will only focus on the highest score since the base is predicted by this score.

Table 1: Raw Data

position	reference base	location on the sequence	Strand	score for A	score for C	score for G	score for T	predicted base	HS
44308116	C	29	1	10	-12	-16	-25	A	10
44308116	C	31	1	-13	-20	-15	10	T	10
44308116	C	31	1	-21	-24	-15	14	T	14
44308117	C	28	1	-20	20	-28	-40	C	20
44308117	C	30	1	-13	10	-13	-28	C	10
44308117	C	30	1	-17	13	-17	-21	C	13

According to the QS formula, when the HS is small, the probability of erroneous base prediction will be high. Hence, we analyzed the distribution of the frequency of the predicted base given known reference base. We randomly selected 1,000 positions with 2,907,483 observations as a testing data set. Figures 2-5 show the results given the reference base is A. The horizontal axis is the highest score of a typical nucleotide and the vertical axis is the frequency of the reads. The figures with reference bases C, G and T are almost the same as those of base A.

Figure 2 shows that if the reference base is A, 57.1% of the reads with predicted base A have HS of 40. The lower scores have substantially lower scores than 40. The reads with predicted base A have HS greater than 10 is 84.6%. Moreover, 2.89% of the reads contain HS less than 0. This suggests that not all the observations are valid even when they have the same reference base.

Figure 2: Distribution of score A given A is the reference base (HS)

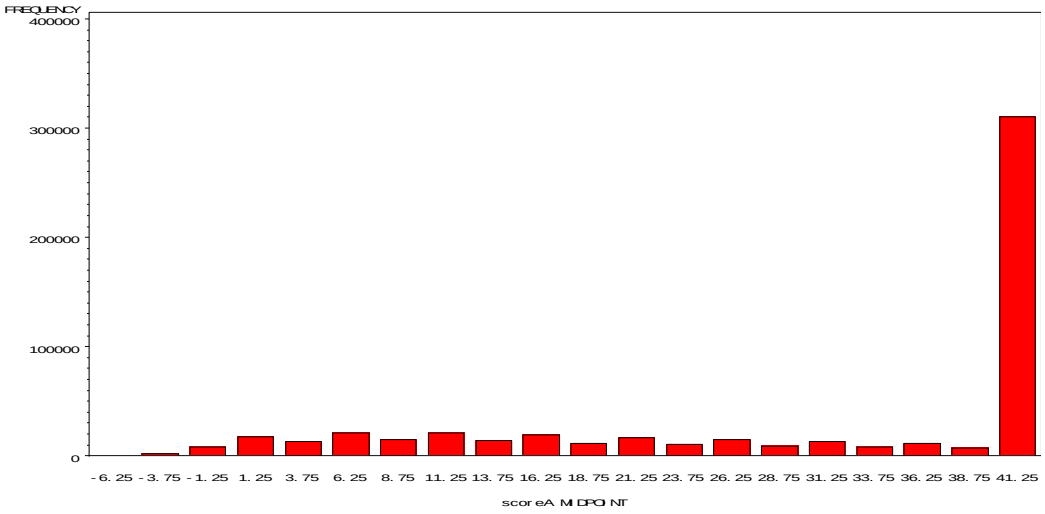


Figure 3 through Figure 5 show distributions of predicted bases different from A when the reference base is A. We can see that the distribution is skewed to the right, which implies that the reads with low HSs have higher chance to be predicted erroneously. In these figures, the read counts tend to decreas as HS increases for HS less than 40, but the frequency slightly picks up at HS=40. A large portion of these (HS=40) are suspected to be heterozygous SNPs.

Figure 3: Distribution of score C given A is the reference base (HS)

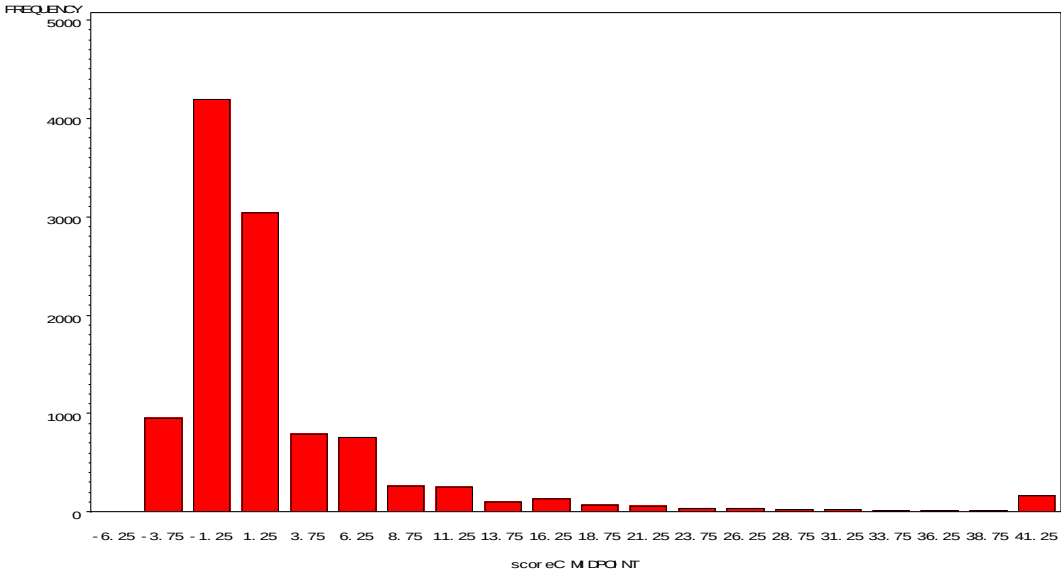


Figure 4: Distribution of score G given A is the reference base (HS)

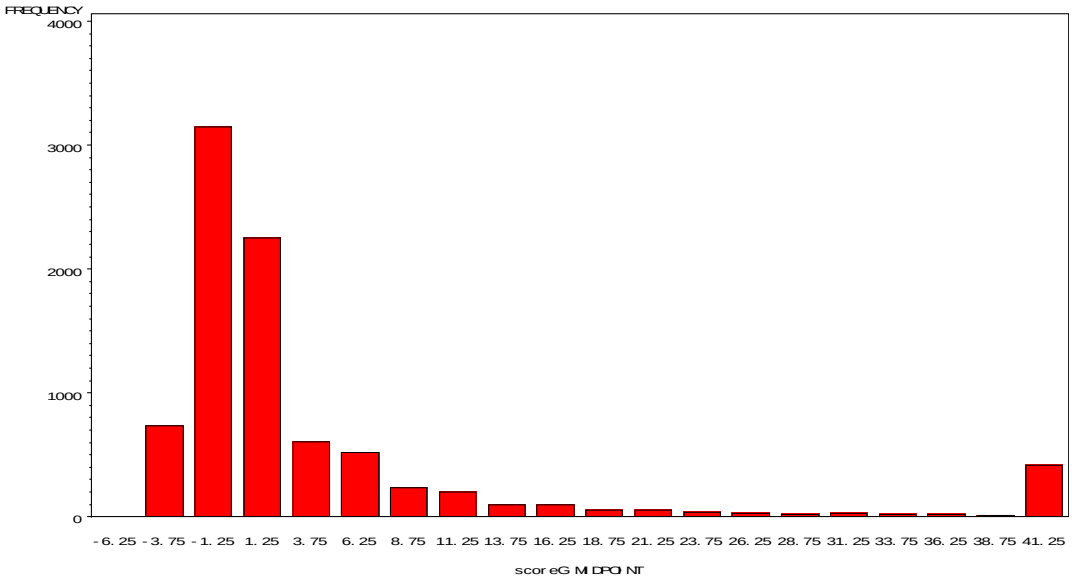
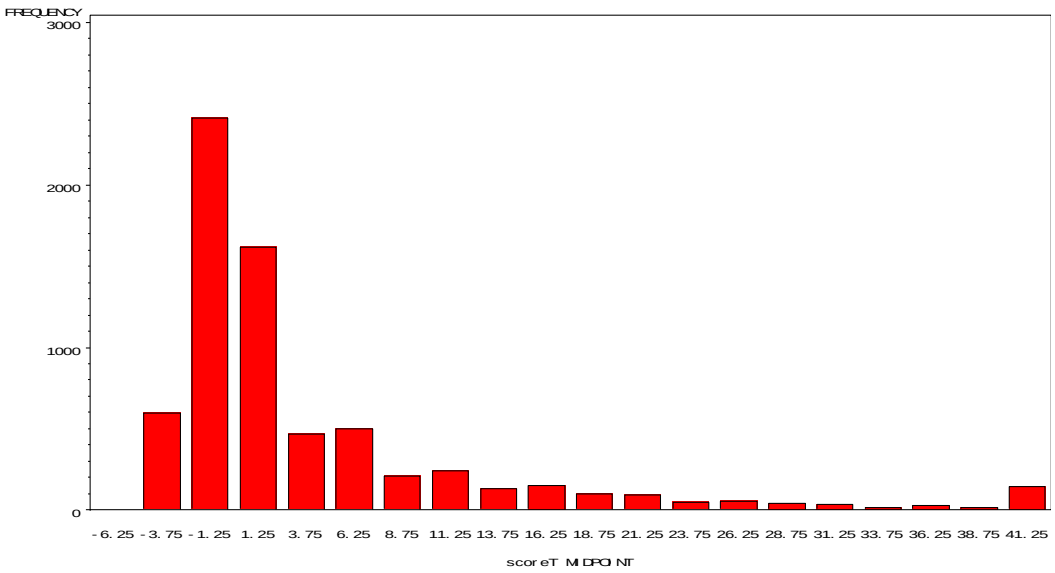


Figure 5: Distribution of score T given A is the reference base (HS)



To improve the accuracy of the SNP detection, we must keep relatively high base call accuracy. We excluded reads with a low HS from the analysis by implementing a threshold on HS, namely cut-off of the quality score.

We analyzed the sample with 7,319,675 reads of 1999 positions to find the proportion of the reads and the accuracy of the prediction by comparing with the reference base given a constraint on the value of the highest score. Table 2 shows the result of quality score cut-off analysis. The first column is the threshold of the HS; the second one is the number of reads with the HS above the threshold; the third column is the number of reads with the same predicted base as the reference base; the fourth column is the proportion of the selected reads out of the total reads; the last column is the accuracy of predicted base. In Table 2, we can see that if we use 10 as the threshold we retain a prediction accuracy of 0.995. Although we lose approximately 16% of the data, it is important to maintain high prediction accuracy in the SNP detection. In fact, the cut-off score of 10 is

commonly used in related research. Hence, we use 10 as the cut-off score to control the accuracy of the base prediction.

Table 2: Quality score cut-off analysis

HS	#Examined reads	#reads with same predicted base as reference base	Exam/Total	Correct/Exam
40	4059686	4049960	0.55462654	0.997604248
>=35	4308131	4297476	0.58856862	0.99752677
>=30	4585700	4573911	0.62648956	0.997429182
>=25	4900523	4887180	0.66950008	0.997277229
>=20	5259955	5244264	0.71860499	0.997016895
>=15	5676192	5656557	0.7754705	0.996540815
>=10	**6148450**	**6120469**	**0.83998948**	**0.995449097**
>=5	6642928	6592923	0.90754412	0.992472446
>=0	7015485	6910867	0.95844214	0.98508756
>=-5	7079719	6943928	0.96721767	0.980819719

2.3 Summarized Data

In the raw data, we have thousands of reads for one position. Hence it is necessary to summarize these and reduce the size of data. For every position, we summarize the information by the strand and the sequence location. Table 3 shows a sample of summarized data for a position. The contents of the columns are as follows:

Column 1: position ID
Column 2: DNA strand

Column 3: location on the sequence (from 0 to 31)
Columns 4-7: total of the highest scores of A, C, G and T, respectively
Columns 8-11: number of reads of A, C, G and T, respectively
Columns 12: total number of reads for a specific strand and sequence location

Researchers commonly compare the numbers of reads of the four nucleotides to find SNPs. However, we have the quality scores containing more information than only the number of reads. The total of the quality scores of each of the four nucleotides turned out to be useful to identify the correct base.

A remaining problem is that it is difficult to identify duplicated observations. Since the observations with different strands or sequence locations are definitely from different sequences, we organized the data by the strand and the sequence location to obtain a measure which can alleviate the problem of duplication in the detection of SNPs.

Table 3: Summarized data for a position

position	strand	location on the sequence	sum of score A	sum of score C	sum of score G	sum of score T	# reads of A	# reads of C	# reads of G	# reads of T	# total reads
44309757	0	0	0	0	0	1304	0	0	0	34	34
44309757	0	1	0	0	0	975	0	0	0	25	25
44309757	0	2	0	0	0	784	0	0	0	20	20
44309757	0	3	0	0	0	1227	0	0	0	32	32
44309757	0	4	0	0	0	718	0	0	0	19	19
44309757	0	5	0	0	0	1058	0	0	0	27	27
44309757	0	6	0	0	0	1590	0	0	0	40	40
44309757	0	7	0	0	0	1648	0	0	0	42	42
44309757	0	8	0	0	0	1584	0	0	0	40	40
44309757	0	9	0	0	0	1680	0	0	0	42	42
44309757	0	10	0	20	0	1120	0	1	0	28	29
44309757	0	11	0	0	0	1211	0	0	0	31	31

44309757	0	12	0	0	0	1760	0	0	0	44	44
44309757	0	13	0	0	0	1670	0	0	0	42	42
44309757	0	14	0	0	0	1262	0	0	0	32	32
44309757	0	15	0	0	0	1431	0	0	0	36	36
44309757	0	16	0	0	0	960	0	0	0	24	24
44309757	0	17	0	0	0	1360	0	0	0	35	35
44309757	0	18	0	0	0	1298	0	0	0	33	33
44309757	0	19	0	0	0	954	0	0	0	24	24
44309757	0	20	0	0	0	791	0	0	0	20	20
44309757	0	21	0	0	0	1030	0	0	0	26	26
44309757	0	22	0	0	0	1510	0	0	0	38	38
44309757	0	23	0	0	0	581	0	0	0	15	15
44309757	0	24	0	0	0	988	0	0	0	26	26
44309757	0	25	0	0	0	1567	0	0	0	40	40
44309757	0	26	0	0	0	936	0	0	0	24	24
44309757	0	27	0	0	0	966	0	0	0	29	29
44309757	0	28	0	0	0	757	0	0	0	20	20
44309757	0	29	0	0	0	1506	0	0	0	41	41
44309757	0	30	0	14	0	933	0	1	0	27	28
44309757	0	31	0	0	0	776	0	0	0	26	26
44309757	0	32	0	0	0	936	0	0	0	24	24
44309757	0	33	0	0	0	966	0	0	0	29	29
44309757	0	34	0	0	0	1298	0	0	0	33	33
44309757	0	35	0	0	0	954	0	0	0	24	24
44309757	1	0	0	13	0	2000	0	1	0	50	51
44309757	1	1	0	40	0	1476	0	1	0	37	38
44309757	1	2	0	0	0	1491	0	0	0	38	38
44309757	1	3	0	0	0	920	0	0	0	23	23
44309757	1	4	0	40	0	996	0	1	0	25	26
44309757	1	5	0	0	0	1377	0	0	0	35	35
44309757	1	6	0	0	0	1010	0	0	0	26	26
44309757	1	7	0	0	0	339	0	0	0	9	9
44309757	1	8	0	0	0	753	0	0	0	20	20
44309757	1	9	0	0	0	938	0	0	0	26	26
44309757	1	10	0	0	0	480	0	0	0	13	13
44309757	1	11	0	24	0	710	0	1	0	22	23
44309757	1	12	0	0	0	1118	0	0	0	33	33

44309757	1	13	0	0	0	976	0	0	0	30	30
44309757	1	14	0	0	0	619	0	0	0	18	18
44309757	1	15	0	0	0	307	0	0	0	12	12
44309757	1	16	0	0	0	977	0	0	0	36	36
44309757	1	17	0	0	0	582	0	0	0	23	23
44309757	1	18	0	0	0	587	0	0	0	20	20
44309757	1	19	0	0	11	287	0	0	1	15	16
44309757	1	20	0	0	0	738	0	0	0	32	32
44309757	1	21	0	0	40	673	0	0	1	28	29
44309757	1	22	0	0	40	628	0	0	1	22	23
44309757	1	23	0	0	0	627	0	0	0	26	26
44309757	1	24	0	16	0	610	0	1	0	29	30
44309757	1	25	0	0	0	476	0	0	0	21	21
44309757	1	26	0	12	0	224	0	1	0	12	13
44309757	1	27	0	0	0	497	0	0	0	26	26
44309757	1	28	0	0	0	168	0	0	0	9	9
44309757	1	29	0	12	0	287	0	1	0	16	17
44309757	1	30	0	26	0	92	0	2	0	5	7
44309757	1	31	0	0	0	134	0	0	0	6	6
44309757	1	32	0	0	0	753	0	0	0	20	20
44309757	1	33	0	0	0	938	0	0	0	26	26
44309757	1	34	0	0	0	307	0	0	0	12	12
44309757	1	35	0	0	0	977	0	0	0	36	36
Total			0	217	91	68161	0	11	3	1929	1943

Chapter 3

First Part of the Study on SNP Classification

There are two general types of DNA positions: one is homozygous position and the other one is heterozygous position. A homozygous position has most of reads with the same predicted base, which means that most of the highest scores appear on the same nucleotide. More than 95% of the positions are homozygous and not SNPs. However, if the predicted base with the highest reads is different from the reference base, this position is a potential homozygous SNP. Similarly, a heterozygous position has most of its reads with one of two different predicted bases. A heterozygous position is very likely a heterozygous SNP.

To detect the SNPs from our data, we simply need to classify all the positions into two groups: homozygous positions and heterozygous positions. Those predicted heterozygous positions will be treated as heterozygous SNPs. After comparing the observed bases with the corresponding reference bases, we can also detect the homozygous SNPs. Therefore, we focus on the two nucleotides

with the highest and second highest values of features, such as total sum of scores, sum of absolute counts and independent counts.

At the beginning of this research, we have the data including 51,473 positions after using the highest score cut-off of 10. Among them, 397 positions with too much artificial information due to experimental reasons are deleted. Thus the size of the data reduced to 51,076. In this data set, we had 21 known SNPs from other sources. Among these 21 known SNPs, 2 were homozygous SNPs, 17 were heterozygous SNPs and the other 2 were ambiguous ones which could not be detected by our method. We decided the threshold of the measures by examining the properties of these known SNPs.

3.1 Introducing Three Measures

To find potential SNPs, first we separated the positions into several groups. We used three measures to realize the goal.

Measure 1: For the summarized data (total 72 rows), if the number of reads of a base for a specific position is not 0, take the log with base 2 to the number of reads and then add 1 to this value. If there is no read for a base, then we set it to 0. Finally, we add the 72 values for each base on this position. We defined this measure as log-transformed value. This measure is designed according to the property of PCR and sequencing method. In the PCR process, DNA sequences are amplified exponentially with base 2. The more reads with the specific strand and location on the sequence, the more chance that the reads are duplicated.

Measure 2: For each position, we added all the scores, and divided it into #total reads from all 4 nucleotides. This measure turned out to be reliable to

identify the correct base of the position, but we cannot identify duplications by using this measure.

Measure 3: As we see in the introduction to the data above, one position may have many reads of the four alleles ACGT with different strands and locations on the sequence. For the same observed nucleotide, if two reads have different strands or sequence locations, they are independent of each other. For the summarized data, we have 72 different combinations of strand and sequence location. If the number of reads of an allele of a combination for a specified position is not 0, set the value as 1; otherwise, set it as 0. Finally, we add the 72 values of each allele for a position and define it as the "independent count". For example, in Table 3, the independent counts for A, C, G and T are 0, 10, 3 and 72 respectively. We find the ratio of the counts for the highest and second highest frequent bases. This measure is used for the second step of screening. To differentiate the independent counts and the observed reads, we also defined the reads as absolute counts.

For measure 1 and measure 2, we calculate the proportions of four nucleotides. We use these proportions to classify the positions into different groups. Table 4 shows an example of the three measures of a specific position.

Table 4: Three Measures of a Position

	A	C	G	T	Proportion of A	Proportion of C	Proportion of G	Proportion of T
Measure 1	3	249.2	270.1	0	0.006	0.477	0.517	0
Measure 2	0.069	15.38	18.46	0	0.002	0.454	0.544	0
Measure 3	3	61	64	0				

3.2 SNP Detection Steps

After the data preparation, we tried to detect the SNPs. The three measures introduced above are used to classify the positions into three groups: clearly homozygous group, clearly heterozygous group and group with the remaining positions.

A clearly homozygous position has most of reads with the same predicted base, which means that most of the highest scores appear on the same nucleotide. More than 95% of the positions are clearly homozygous and not SNPs. However, if the predicted base is different from the reference base, this position is a potential homozygous SNP. Similarly, a clearly heterozygous position has most of reads with one of two different predicted bases. A clearly heterozygous position is very likely a heterozygous SNP. The remaining group, which contains hard-to-classify positions, is set aside to investigate further.

If the count in measure 3 is small, then most of the reads are from only a few sequences. Thus the base call may not be reliable. Thus we excluded the positions with highest count<5 based on measure 3 before classifying the clearly homozygous group and the clearly heterozygous group. After this exclusion, we determine the homozygous positions first. If the highest proportion in measure (1)

is greater than 0.9 or the highest proportion in measure (2) is greater than 0.95, we move the position into the homozygous group. We compare the predicted base of these positions with the reference base. If they coincide, then the position is a normal base; otherwise, it is a homozygous SNP. For the remaining positions, we check if they are heterozygous using the constraint described as following: $0.3 \leq$ the highest and the second highest in measure (1) ≤ 0.7 or $0.2 \leq$ the highest and the second highest in measure (2) ≤ 0.8. After classifying the positions to the clearly heterozygous group, the remaining positions are classified to the hard-to-classify group.

We classified 50,901 positions as clearly homozygous positions. Among them, 8 positions were identified as homozygous SNP candidates because the predicted bases were different from the reference bases. The two known homozygous SNPs were included in these 8 positions. We also identified 63 heterozygous SNP candidates and 503 positions were classified to the hard-to-classify group. All the 17 known heterozygous SNPs were included in the set of 63 heterozygous SNP candidates.

3.3 Cross-Validation

3.3.1 Use of Cross-Validation

In Section 2.2, we predicted homozygous and heterozygous SNPs by the threshold on HS based on the observation from the known SNPs. However, the thresholds we used are not adequately validated. Hence cross-validation is used to evaluate the method for finding the optimal thresholds to predict SNPs. Using the

information of known SNPs, we propose a method for finding an optimal combination of thresholds.

We are using K-fold cross-validation. In K-fold cross-validation, the original sample is randomly partitioned into K subsamples. Of the K subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $K - 1$ subsamples are used as training data. The cross-validation process is then repeated K times (the folds), with each of the K subsamples used exactly once as the validation data. The K results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once (Blum et al., 1999). Ten-fold cross-validation is widely used even if computation power allows using more folds (Kohavi 1995). Hence, we use ten-fold cross-validation in our research.

3.3.2 Preparing the Data Set for Cross-Validation

In Section 2.2, we made the prediction of SNPs. Among all the bases analyzed, most of them were classified as non-SNP homozygous positions. Hence, we drew 100 positions out of these 50,509 positions randomly. We treated these 100 positions as known non-SNPs. Together with the 19 known heterozygous SNPs; we have a data set consisting of 119 positions for cross-validation. We randomly partitioned the 119 positions into 10 subsets. Except for the 10^{th} subset, every subset has 12 positions.

3.3.3 Threshold Search

Step 1

Of the 10 groups of positions, take 1st through 9th groups as the training set and take the 10th group as the test set.

Step 2

Apply various combinations of the thresholds for each measure (rule) which are used to determine the clearly homozygous positions and clearly heterozygous SNP candidates to the training set. We calculate the accuracy of every combination as shown in Table 5.

An example of the combinations is given below:

Constraint for clearly homozygous base:

Highest proportion in measure (1) ≥ 0.95 or highest proportion in measure (2) ≥ 0.95

Constraint for clearly heterozygous base:

$0.2 \leq$ both highest and second highest in measure (1) ≤ 0.8 or $0.2 \leq$ both highest and second highest in measure (2) ≤ 0.

Table 5: Accuracy of the proposed approach

	#Real SNP	#Real Non-SNPs	
Predicted #SNPs:	**A**	**B**	
Predicted #Non-SNPs:	**C**	**D**	#in middle data: **E**

Accuracy: (A+D)/(A+B+C+D+E)

Sensitivity: A/(A+C)

Specificity: D/(B+D)

Step 3

For each combination of thresholds we search for optimal thresholds for classifying homozygous positions and heterozygous SNPs using the training set in a sequential fashion. We determine clearly homozygous positions first. Among these positions, SNP candidates are identified by comparing with the reference base. For the remaining positions, determine clearly heterozygous positions. The remaining positions are considered "hard to classify" or positions in the "middle area". For the thresholds for a clearly heterozygous position, we search optimal thresholds for each of measure 1 and measure 2 for the proportion of the most frequent base. All possible pairs of thresholds for measure 1 and measure 2 are searched starting with 0.95, and decrement of 0.05. We choose the pair of thresholds resulting in the highest classification accuracy of the prediction. Optimal combinations of thresholds for searching heterozygous SNPs are obtained as follows: First, we find the combinations of thresholds with the highest accuracy. Second, among these combinations, we select all non-nested most-balanced combinations. After this, find the least-balanced pair in each measure from the selected combinations. Third, among the combinations with the highest accuracy, select all non-nested least-balanced combinations, and find the most-balanced pair in each measure from the selected combinations. Finally, take the average of the values obtained from the previous steps. Further details are given in Section 3.3.4.

Step 4

Apply the threshold combinations obtained in Step 3 to the test set, and find the classification accuracy.

Step 5

Use the 9^{th} subset as the new test set and repeat the step 2 through step 4 similarly until all the subsets are used as test sets.

Step 6

Evaluate the cross-validation accuracy by averaging the accuracies obtained from steps 1 through 6.

Step 7

Implement step 2 and step 3 to the whole data (119 bases), and obtain an optimal combination of thresholds, and find the accuracy.

3.3.4 Result of the Cross-Validation

Table 6 shows the result from training sets. The first column is the number of combination; the second column is the threshold for clearly homozygous position; the third column is the threshold for clearly heterozygous position; the fourth through thirteenth columns are the accuracies of every combination with the training set corresponding to the specific test set, in which 1 means the 100% classification accuracy and blank means accuracy less than 100%. For example, the threshold pair of (0.95, 0.95) for clearly homozygous position gave 100% accuracy for all ten training sets.

Table 6: Accuracy of training sets using different combination of thresholds

Combinations	Constraint for clearly homozygous base	Constraint for clearly heterozygous base	Test set									
	Homo	Heter	Group10	G9	G8	G7	G6	G5	G4	G3	G2	G1
combination1	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.1<=both<=0.9 or measure(2)0.1<=both<=0.9	1	1	1	1	1	1	1	1	1	1
combination2	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.1<=both<=0.9 or measure(2)0.15<=both<=0.85	1	1	1	1	1	1	1	1	1	1
combination3	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.1<=both<=0.9 or measure(2)0.2<=both<=0.8	1	1	1	1	1	1	1	1	1	1
combination4	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.1<=both<=0.9 or measure(2)0.25<=both<=0.75	1	1	1	1	1	1	1	1	1	1
combination5	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.1<=both<=0.9 or measure(2)0.3<=both<=0.7	1	1	1	1	1	1	1	1	1	1
combination6	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.1<=both<=0.9 or measure(2)0.35<=both<=0.65	1	1	1	1	1	1	1	1	1	1
combination7	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.1<=both<=0.9 or measure(2)0.4<=both<=0.6	1	1	1	1	1	1	1	1	1	1
combination8	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.1<=both<=0.9 or measure(2)0.45<=both<=0.55	1	1	1	1	1	1	1	1	1	1
combination9	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.15<=both<=0.85 or measure(2)0.1<=both<=0.9	1	1	1	1	1	1	1	1	1	1
combination10	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.15<=both<=0.85 or measure(2)0.15<=both<=0.85	1	1	1	1	1	1	1	1	1	1
combination11	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.15<=both<=0.85 or measure(2)0.2<=both<=0.8	1	1	1	1	1	1	1	1	1	1
combination12	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.15<=both<=0.85 or measure(2)0.25<=both<=0.75	1	1	1	1	1	1	1	1	1	1
combination13	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.15<=both<=0.85 or measure(2)0.3<=both<=0.7	1	1	1	1	1	1	1	1	1	1
combination14	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.15<=both<=0.85 or measure(2)0.35<=both<=0.65	1	1	1	1	1	1	1	1	1	1
combination15	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.15<=both<=0.85 or measure(2)0.4<=both<=0.6	1	1	1	1	1	1	1	1	1	1
combination16	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.15<=both<=0.85 or measure(2)0.45<=both<=0.55	1	1	1	1	1	1	1	1	1	1
combination17	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.2<=both<=0.8 or measure(2)0.1<=both<=0.9	1	1	1	1	1	1	1	1	1	1
combination18	measure(1) >=0.95	measure(1)0.2<=both<=0.8	1	1	1	1	1	1	1	1	1	1

	or measure(2) >=0.95	or measure(2)0.15<=both<=0.85										
combination19	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.2<=both<=0.8 or measure(2)0.2<=both<=0.8	1	1	1	1	1	1	1	1	1	1
combination20	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.2<=both<=0.8 or measure(2)0.25<=both<=0.75	1	1	1	1	1	1	1	1	1	1
combination21	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.2<=both<=0.8 or measure(2)0.3<=both<=0.7	1	1	1	1	1	1	1	1	1	1
combination22	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.2<=both<=0.8 or measure(2)0.35<=both<=0.65	1	1	1	1	1	1	1	1	1	1
combination23	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.2<=both<=0.8 or measure(2)0.4<=both<=0.6	1	1	1	1	1	1	1	1	1	1
combination24	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.2<=both<=0.8 or measure(2)0.45<=both<=0.55	1	1	1	1	1	1	1	1	1	1
combination25	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.25<=both<=0.75 or measure(2)0.1<=both<=0.9	1	1	1	1	1	1	1	1	1	1
combination26	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.25<=both<=0.75 or measure(2)0.15<=both<=0.85	1	1	1	1	1	1	1	1	1	1
combination27	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.25<=both<=0.75 or measure(2)0.2<=both<=0.8	1	1	1	1	1	1	1	1	1	1
combination28	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.25<=both<=0.75 or measure(2)0.25<=both<=0.75	1	1	1	1	1	1	1	1	1	1
combination29	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.25<=both<=0.75 or measure(2)0.3<=both<=0.7	1	1	1	1	1	1	1	1	1	1
combination30	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.25<=both<=0.75 or measure(2)0.35<=both<=0.65	1	1	1	1	1	1	1	1	1	1
combination31	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.25<=both<=0.75 or measure(2)0.4<=both<=0.6	1	1	1	1	1	1	1	1	1	1
combination32	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.25<=both<=0.75 or measure(2)0.45<=both<=0.55	1	1	1	1	1	1	1	1	1	1
combination33	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.3<=both<=0.7 or measure(2)0.1<=both<=0.9	1	1	1	1	1	1	1	1	1	1
combination34	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.3<=both<=0.7 or measure(2)0.15<=both<=0.85	1	1	1	1	1	1	1	1	1	1
combination35	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.3<=both<=0.7 or measure(2)0.2<=both<=0.8	1	1	1	1	1	1	1	1	1	1
combination36	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.3<=both<=0.7 or measure(2)0.25<=both<=0.75	1	1	1	1	1	1	1	1	1	1
combination37	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.3<=both<=0.7 or measure(2)0.3<=both<=0.7	1	1	1	1	1	1	1	1	1	1
combination38	measure(1) >=0.95 or	measure(1)0.3<=both<=0.7 or	1	1	1	1	1	1	1	1	1	1

	measure(2) >=0.95	measure(2)0.35<=both<=0.65										
combination39	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.3<=both<=0.7 or measure(2)0.4<=both<=0.6	1	1	1	1	1	1	1	1	1	1
combination40	measure(1) >=0.95 or measure(2) >=0.95	**measure(1)0.3<=both<=0.7 or measure(2)0.45<=both<=0.55**	1	1	1	1	1	1	1	1	1	1
combination41	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.35<=both<=0.65 or measure(2)0.1<=both<=0.9	1	1	1	1	1	1	1	1	1	1
combination42	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.35<=both<=0.65 or measure(2)0.15<=both<=0.85	1	1	1	1	1	1	1	1	1	1
combination43	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.35<=both<=0.65 or measure(2)0.2<=both<=0.8	1	1	1	1	1	1	1	1	1	1
combination44	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.35<=both<=0.65 or measure(2)0.25<=both<=0.75	1	1	1	1	1	1	1	1	1	1
combination45	measure(1) >=0.95 or measure(2) >=0.95	**measure(1)0.35<=both<=0.65 or measure(2)0.3<=both<=0.7**	1	1	1	1	1	1	1	1	1	1
combination46	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.35<=both<=0.65 or measure(2)0.35<=both<=0.65						1				
combination47	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.35<=both<=0.65 or measure(2)0.4<=both<=0.6						1				
combination48	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.35<=both<=0.65 or measure(2)0.45<=both<=0.55						1				
combination49	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.4<=both<=0.6 or measure(2)0.1<=both<=0.9	1	1	1	1	1	1	1	1	1	1
combination50	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.4<=both<=0.6 or measure(2)0.15<=both<=0.85	1	1	1	1	1	1	1	1	1	1
combination51	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.4<=both<=0.6 or measure(2)0.2<=both<=0.8	1	1	1	1	1	1	1	1	1	1
combination52	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.4<=both<=0.6 or measure(2)0.25<=both<=0.75	1	1	1	1	1	1	1	1	1	1
combination53	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.4<=both<=0.6 or measure(2)0.3<=both<=0.7			1							
combination54	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.4<=both<=0.6 or measure(2)0.35<=both<=0.65										
combination55	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.4<=both<=0.6 or measure(2)0.4<=both<=0.6										
combination56	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.4<=both<=0.6 or measure(2)0.45<=both<=0.55										
combination57	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.45<=both<=0.55 or measure(2)0.1<=both<=0.9	1	1	1	1	1	1	1	1	1	1
combination58	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.45<=both<=0.55 or measure(2)0.15<=both<=0.85	1	1	1	1	1	1	1	1	1	1

combination59	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.45<=both<=0.55 or measure(2)0.2<=both<=0.8	1	1	1	1	1	1	1	1	1	1
combination60	measure(1) >=0.95 or measure(2) >=0.95	**measure(1)0.45<=both<=0.55 or measure(2)0.25<=both<=0.75**	1	1	1	1	1	1	1	1	1	1
combination61	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.45<=both<=0.55 or measure(2)0.3<=both<=0.7			1							
combination62	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.45<=both<=0.55 or measure(2)0.35<=both<=0.65										
combination63	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.45<=both<=0.55 or measure(2)0.4<=both<=0.6										
combination64	measure(1) >=0.95 or measure(2) >=0.95	measure(1)0.45<=both<=0.55 or measure(2)0.45<=both<=0.55										

For threshold for clearly heterozygous positions, we found {(0.1, 0.9), (0.1, 0.9)} as the non-nested least-balanced combination for every training set. Consider the training set corresponding to the 10^{th} test set for example. We find {(0.3, 0.7), (0.45, 0.55)}, {(0.35, 0.65), (0.3, 0.7)} and {(0.45, 0.55), (0.25, 0.75)} as the non-nested most-balanced combinations. Among these, the least-balanced pair for measure 1 is (0.3, 0.7), and the least-balanced pair for measure 2 is (0.25, 0.75). Combining these two pairs, we obtain {(0.3, 0.7), (0.25, 0.75)}. By taking the average thresholds for each measure between the least-balanced and most-balanced thresholds, we obtain $\{(\frac{0.3+0.1}{2}, \frac{0.7+0.9}{2}), (\frac{0.25+0.1}{2}, \frac{0.75+0.9}{2})\}$, or {(0.2, 0.8), (0.175, 0.825)}. We use this combination of thresholds to the corresponding test set to obtain the accuracy for the training set. Table 7 shows the accuracies for the test set using the result of training sets above.

Table 7: Cross-validation accuracy

Group 10 as the test set.
measure(1) >=0.95 or measure(2) >=0.95 && measure(1)0.2<=both<=0.8 or measure(2)0.175<=both<=0.825

	Real SNP	Real Non-SNP	
Test SNP:	3	0	
Test Non-SNP:	0	8	Middle data: 0

Sensitivity: 1.000000 Specificity: 1.000000 Overall Accuracy: 1.000000

Group 9 as the test set.
measure(1) >=0.95 or measure(2) >=0.95 && measure(1)0.2<=both<=0.8 or measure(2)0.175<=both<=0.825

	Real SNP	Real Non-SNP	
Test SNP:	2	0	
Test Non-SNP:	0	10	Middle data: 0

Sensitivity: 1.000000 Specificity: 1.000000 Overall Accuracy: 1.000000

Group 8 as the test set.
measure(1) >=0.95 or measure(2) >=0.95 && measure(1)0.2<=both<=0.8 or measure(2)0.2<=both<=0.8

	Real SNP	Real Non-SNP	
Test SNP:	4	0	
Test Non-SNP:	0	8	Middle data: 0

Sensitivity: 1.000000 Specificity: 1.000000 Overall Accuracy: 1.000000

Group 7 as the test set.
measure(1) >=0.95 or measure(2) >=0.95 && measure(1)0.2<=both<=0.8 or measure(2)0.175<=both<=0.825

	Real SNP	Real Non-SNP	
Test SNP:	0	0	
Test Non-SNP:	0	12	Middle data: 0

Sensitivity: -1.#IND00 Specificity: 1.000000 Overall Accuracy: 1.000000

Group 6 as the test set.

measure(1) >=0.95 or measure(2) >=0.95 && measure(1)0.2<=both<=0.8 or measure(2)0.175<=both<=0.825

	Real SNP	Real Non-SNP	
Test SNP:	1	0	
Test Non-SNP:	0	11	Middle data: 0

Sensitivity: 1.000000 Specificity: 1.000000 Overall Accuracy: 1.000000

Group 5 as the test set.

measure(1) >=0.95 or measure(2) >=0.95 && measure(1)0.225<=both<=0.775 or measure(2)0.175<=both<=0.825

	Real SNP	Real Non-SNP	
Test SNP:	2	0	
Test Non-SNP:	0	10	Middle data: 0

Sensitivity: 1.000000 Specificity: 1.000000 Overall Accuracy: 1.000000

Group 4 as the test set.

measure(1) >=0.95 or measure(2) >=0.95 && measure(1)0.2<=both<=0.8 or measure(2)0.175<=both<=0.825

	Real SNP	Real Non-SNP	
Test SNP:	2	0	
Test Non-SNP:	0	10	Middle data: 0

Sensitivity: 1.000000 Specificity: 1.000000 Overall Accuracy: 1.000000

Group 3 as the test set.

measure(1) >=0.95 or measure(2) >=0.95 && measure(1)0.2<=both<=0.8 or measure(2)0.175<=both<=0.825

	Real SNP	Real Non-SNP	
Test SNP:	3	0	
Test Non-SNP:	0	9	Middle data: 0

```
Sensitivity: 1.000000        Specificity: 1.000000        Overall Accuracy: 1.000000
--------------------------------------------------------------------------------

Group 2 as the test set.
measure(1) >=0.95  or  measure(2) >=0.95  && measure(1)0.2<=both<=0.8  or
measure(2)0.175<=both<=0.825
              Real SNP                    Real Non-SNP
Test SNP:          1                     0
Test Non-SNP:    0                              11                          Middle data:  0
Sensitivity: 1.000000        Specificity: 1.000000        Overall Accuracy: 1.000000
--------------------------------------------------------------------------------

Group 1 as the test set.
measure(1) >=0.95  or  measure(2) >=0.95  && measure(1)0.2<=both<=0.8  or
measure(2)0.175<=both<=0.825
              Real SNP                    Real Non-SNP
Test SNP:          1                     0
Test Non-SNP:    0                              11                          Middle data:  0
Sensitivity: 1.000000        Specificity: 1.000000        Overall Accuracy: 1.000000
```

We see that all the accuracies are 1. Hence the average accuracy for all the test sets, which is the cross validation accuracy, is 100%.

3.3.5 Summary and Application to the Entire Data

We found that the cross-validation accuracy using a 119 subject data set was 100%. This is based on limited information. We believe the result will be more reliable if we can use a bigger data set with known SNP information. We applied the optimal combination of thresholds obtained from the data set of size 119 to the entire data set of size 51,076 to make a prediction as explained in Section 2.2.

We excluded the positions with count less than 5 based on measure 3 before applying our optimal thresholds. We first checked if the position is homozygous, using the following constraint obtained as optimal thresholds using our approach: the highest proportion in measure (1) $\geq$0.95 or the highest proportion of measure (2) $\geq$0.95. We obtained 50,509 homozygous positions by these thresholds. For these positions, we compared the predicted base with the reference base to determine if it is a homozygous SNP or not. If it is not clearly homozygous, then we checked if it is heterozygous, using the constraint: 0.2$\leq$ the highest and second highest proportions in measure (1) $\leq$0.8 or 0.175$\leq$ the highest and second highest proportions in measure (2) $\leq$0.825. The positions which do not belong to either clearly homozygous bases or clearly heterozygous bases were assigned to the middle group.

This process resulted in 8 homozygous SNP candidates, which is the same as the earlier prediction. Next we identified 68 heterozygous SNP candidates including 5 new ones from the previous prediction in Section 2.2. All the known homozygous and heterozygous SNPs are included in our prediction. We classified 499 positions to the hard-to-classify group. Among them, 454 positions have the count in measure 3 less than 5.

Chapter 4

Advanced Classification Modeling

In the previous study, we analyzed a relatively small data set with around fifty-one thousand positions. Among them, there were only 21 known SNPs. In this chapter, we will introduce our further work on data set including 1,278,923 positions and 308 known SNPs. Since the size of the training data is substantially increased, we introduced some popular classification methods, such as random forest (Breiman, 2001), Support Vector Machines (SVM) (Vapnik, 1995; Cortes, Vapnik, 1995), single decision tree and logistic regression.

4.1 Developing Measures

To make the prediction more accurate, we developed eight measures according to the property of the data. This time we further focus on the independent counts.

Measure 1: Ratio of sum of absolute counts of the top two alleles (the second highest sum of absolute counts divided by the highest one). If the ratio is close to 0, it means that this position has the most of reads with the same allele, which indicates possible homozygosity. If the ratio is close to 1, it means that this position has similar amount of absolute counts for two alleles and is possibly a heterozygous SNP. Taking Table 3 as an example, the ratio of measure 1 for position 44309757 is 11/1929, or 0.0057.

Measure 2: Ratio of sum of quality scores of the top two alleles (the second highest sum of scores divided by the highest one). In Table 3, measure 2 is 217/68161, or 0.0032.

Measure 3: Ratio of sum of independent counts of the top two alleles (the second highest sum of independent counts divided by the highest one). Since independent counts avoid duplication, this ratio is a very important measure although it loses some information. In Table 3, measure 3 is 10/72, or 0.1389.

Measure 4: Ratio of log-transformed values of the top two alleles. The log-transformed value is introduced in Section 3.1. In Table 3, measure 4 is 11/404.9, or 0.0272.

Measure 5: If we assume a position is a heterozygous SNP, then the independent counts for the two alleles are expected to be equally distributed to the two strands. We defined a chi-square value expressed below as a measure.

$$Chi-square=\frac{(A1^{-}-N/4)^2}{N/4}+\frac{(A1^{+}-N/4)^2}{N/4}+\frac{(A2^{-}-N/4)^2}{N/4}+\frac{(A2^{+}-N/4)^2}{N/4}$$

where $A1^{-}$ denotes the independent counts of the most frequent allele with the negative strand;

$A1^+$ denotes the independent counts of the most frequent allele with the positive strand;

$A2^-$ denotes the independent counts of the second most frequent allele with the negative strand;

$A2^+$ denotes the independent counts of the second most frequent allele with the positive strand;

N denotes the total observed independent counts.

If the position is heterozygous, this chi-square value should be small; otherwise it should be large. Hence, we can use it as a classifier.

Measure 6: Similar to measure 5, if a position is heterozygous, we expect to observe the same independent counts for the two alleles. Therefore we define another chi-square value which is expressed below:

$$Chi-square = \frac{(A1 - N/2)^2}{N/2} + \frac{(A2 - N/2)^2}{N/2}$$

Where $A1$ denotes the independent counts of the most frequent allele;

$A2$ denotes the independent counts of the second most frequent allele;

N denotes the total observed independent counts.

If the assumption holds, this chi-square value should also be small; otherwise it should be large.

Measure 7: If we assume that a position is homozygous, then only one allele is expected to be observed. However, it is almost impossible to observe only one nucleotide in the experiment even if the position is truly homozygous. Hence, we assume that we have 99% probability of having only one allele if the position is homozygous. Then the probability of observing one of the other three possible

nucleotides is approximately 0.003. Using this assumption, we define the following chi-square value:

$$Chi-square=\frac{(A1-N\times 0.99)^2}{N\times 0.99}+\frac{(A2-N\times 0.003)^2}{N\times 0.003}+\frac{(A3-N\times 0.003)^2}{N\times 0.003}+\frac{(A4-N\times 0.003)^2}{N\times 0.003}$$

and

$$Chi-square=\frac{(A2-N\times 0.99)^2}{N\times 0.99}+\frac{(A1-N\times 0.003)^2}{N\times 0.003}+\frac{(A3-N\times 0.003)^2}{N\times 0.003}+\frac{(A4-N\times 0.003)^2}{N\times 0.003}$$

Where $A1$ denotes the independent counts of the most frequent allele;

$A2$ denotes the independent counts of the second most frequent allele;

$A3$ denotes the independent counts of the third most frequent allele;

$A4$ denotes the independent counts of the least frequent allele;

N denotes the total observed independent counts.

Since $A1$ and $A2$ are sometimes close to each other and we do not know which allele is the true homozygous nucleotide, we calculate the two chi-square values and choose the larger one. If the position is homozygous, this chi-square value should be small; otherwise it should be large.

Measure 8: Proportion of independent counts of all non-reference alleles (the sum of the independent counts for three alleles other than the reference base divided by the total independent counts for the four alleles). A homozygous position can be either homozygous SNP or normal genome position. Hence, there are three possible ranges of this proportion. If a position is normal homozygous, the proportion will be close to 0; if it is a homozygous SNP, the proportion will be close to 1; if it is a heterozygous SNP, the proportion will be close to 0.5. In our classification models, we just want two classes: homozygous group and heterozygous group. Hence, we made a transformation to the proportion. We take

the absolute value of the difference between the proportion and 0.5. In Table 3, since the reference base is T, measure 8 is $\left|0.5-(10+3)/(10+3+72)\right|$ = 0.32.

4.2 Classification Methods

4.2.1 Single Decision Tree

A decision tree is a tree-structured plan of a set of attributions to test in order to predict the output. Classification and Regression Trees (CART) is a widely used tree algorithm (Breiman et al., 1984). R package "rpart" running CART is used in this study. The tree is built by the following process: first the single variable is found which best splits the data into two groups. The data are separated, and then this process is applied separately to each sub-group, and so on recursively until the subgroups either reach a minimum size or until no improvement can be made (Breiman et al., 1984; De'ath et al., 2000).

4.2.2 Random Forest

Random Forest (Breiman, 2001) uses the result by combining multiple decision trees using the bagging algorithm. Bagging algorithm is a widely used ensemble based algorithm (Breiman et al., 1996). Different training data sets are randomly drawn with replacement from the entire training data set. Each training data set is used to generate a classifier. The result is then given as a combination of individual classifiers by taking a simple majority vote of their decisions.

In random forest, from the root of a tree, the given object follows the relevant branches and arrives at a leaf. Branches are features and leaves are classes. If the number of cases in the original data set is N, a bootstrap sample of size N is generated as a training set to grow each tree. If there are M input variables, a fixed number m which should be much less than M is specified. At each node of a tree, m variables are randomly selected out of the M variables and the best split on these values is used to split the node. All trees are grown to their largest extent possible without pruning. Each tree gives a classification for a new object from an input vector, and we say the tree votes for that class. The forest chooses the class having the most votes over all the trees in the forest. Each tree is constructed using a different training set obtained from the original data set. When a training set for a tree is selected with replacement from the original data set, approximately one-third of the cases are left and not used in the construction of the tree as explained earlier. This out-of-bag data can be used to get the estimates of the classification error or variable importance.

4.2.3 Support Vector Machine

Support vector machines (SVM) are a set of related supervised learning methods that analyze data and recognize patterns (Vapnik, 1995; Cortes et al., 1995). The standard SVM is a non-probabilistic binary linear classifier, i.e. it predicts, for each given input, which of two possible classes the input is a member of. Since SVM is a classifier, given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. More formally, SVM is a method to find a separating hyperplane in data space, which maximizes the margin between the two separated data sets. If the data are

nonseparable in the original feature space, they are transformed to a higher dimensional space, where the data become linearly separable.

4.2.4 Logistic Regression

Logistic regression is a model that fits the log odds of the response to linear combinations of the explanatory variables (Alpaydin et al., 2004). It is used mainly for binary responses, although there are extensions for multinomial responses as well. Regression coefficients are determined by maximizing the likelihood function. Usually the coefficients are estimated by numerical methods such as the Newton-Raphson algorithm. Logistic regression is known as a robust model for classification and the model is presented clearly and succinctly, but on the flip side, it might not be able to produce complex models, leading to under-fitting.

4.3 Preparing the Data

We have 285 known SNPs including 137 heterozygous and 148 homozygous SNPs. These known SNPs are used as training set to create models. Before applying the four classification models to the whole data, we used 10-fold cross validation method to evaluate and compare the four methods. This time we took the 148 known homozygous SNPs in the training set instead of randomly selecting homozygous positions from the whole data. We conducted 10-fold cross validation using the 285 known SNPs

4.4 Results

We calculated the generalized accuracies of the four methods. In order to compare the performance of them, we repeated 10-fold cross validation for 30 times. After gathering all the generalized accuracies, we conducted paired t-tests to see if there was significant difference between every pair of the classification methods. Table 8 is the matrix of the p-values of the paired t-tests and Table 9 shows the mean accuracies with standard deviation in parentheses of the four methods.

Table 8: P-values of paired t-tests

P-values	SVM	ST	LR
RF	<0.0001 (SVM > RF)	<0.0001 (ST >RF)	0.0005 (RF>LR)
SVM		<0.0001 (SVM > ST)	<0.0001 (SVM > LR)
ST			<0.0001 (ST > LR)

Table 9: Mean accuracies, sensitivities and specificities (sd) of the four methods

Methods	RF	SVM	ST	LR
Mean accuracy (sd)	0.991 (0.0018)	0.9955 (0.0021)	0.993 (<0.0001)	0.988 (0.0034)
Mean sensitivity (sd)	0.9924 (0.0014)	0.9927 (0.0000)	0.9927 (0.0000)	0.9877 (0.0052)
Mean specificity (sd)	0.9897 (0.0034)	0.9981 (0.0040)	0.9932 (0.0000)	0.9884 (0.0047)

From Table 8, we see that the mean cross-validation accuracies of the four methods are significantly different from each other at the significance level of 0.05. From Table 9, we see that SVM has the highest values of mean accuracy, mean sensitivity and mean specificity. Hence, we conclude that SVM performs the best among the four methods.

We applied SVM to the whole data to predict the potential SNPs. We classified 213 unidentified positions as heterozygous SNP candidates. For the remaining positions classified as homozygous, the observed bases were compared with the reference bases, and we identified 100 homozygous SNP candidates.

Chapter 5

Classification Based on a Subset of the Data

5.1 Preparing the Data

In this study, the analyzed DNA positions have the numbers of absolute reads ranging from 200 to 5,000, which provide clear and sufficient information to identify the genome type. However, in the real biological experiment, due to the limitation of time or funding, researchers often cannot obtain the data with sufficient reads. Hence, it is also important to find some efficient and accurate models to deal with data with fewer observations.

In the previous section, we evaluated the four classification methods based on the whole data. In this section, we will conduct the similar process using the randomly selected subset of data and compare the performances of the four models.

For every position, we randomly selected 10, 20, 30 and 40 absolute reads from the raw data separately. We defined the number of randomly picked reads as cover depth. Under the different cover depths, we calculated the eight measures and generate the classification matrix. After that, we conducted 10-fold CV to the 285 known SNPs to obtain accuracies. We repeated the processes for 30 times. In order to statistically compare the accuracies obtained from the four classification methods, we kept the same group members for the known SNPs with each cover depth.

5.2 Results and Conclusion

We calculated the mean accuracies and conducted paired t-test to compare performance of the four classification methods. Table 10 shows the average accuracies of the four classification methods for different cover depths and Table 11 through Table 15 are the p-value matrices.

Table 10: Mean accuracies (sd) of the four classification methods

	Cover depth 10	Cover depth 20	Cover depth 30	Cover depth 40	Whole data
Random Forest	0.9776 (0.0083)	0.9885 (0.0045)	0.9908 (0.0034)	0.9915 (0.0029)	0.991 (0.0018)
SVM	0.9739 (0.0086)	0.9891 (0.0053)	0.9927 (0.0040)	0.994 (0.0031)	0.9955 (0.0021)
Single Decision Tree	0.9675 (0.0101)	0.9835 (0.0097)	0.988 (0.0058)	0.9908 (0.0073)	0.993 (<0.001)
Logistic Regression	0.9736 (0.0088)	0.9883 (0.0085)	0.992 (0.0053)	0.9912 (0.0051)	0.988 (0.0034)

Table 11: P-values of paired t-test for cover depth 10

Cover Depth 10	SVM	ST	LR
RF	0.0336 (RF > SVM)	<0.0001 (RF > ST)	0.0061 (RF > LR)
SVM		0.0129 (SVM > ST)	0.87
ST			0.0188 (LR > ST)

Table 12: P-values of paired t-test for cover depth 20

Cover Depth 20	SVM	ST	LR
RF	0.5018	0.0012 (RF > ST)	0.8539
SVM		0.0025 (SVM > ST)	0.5702
ST			0.0138 (LR > ST)

Table 13: P-values of paired t-test for cover depth 30

Cover Depth 30	SVM	ST	LR
RF	0.0237 (SVM > RF)	0.0031 (RF > ST)	0.2887
SVM		<0.0001 (SVM > ST)	0.4299
ST			0.0021 (LR > ST)

Table 14: P-values of paired t-test for cover depth 40

Cover Depth 40	SVM	ST	LR
RF	0.0023 (SVM > RF)	0.5798	0.7389
SVM		0.0197 (SVM > ST)	0.0072 (SVM > LR)
ST			0.7566

Table 15: P-values of paired t-test for the whole data

Whole data	SVM	ST	LR
RF	<0.0001 (SVM >RF)	<0.0001 (ST > RF)	0.0005 (RF > LR)
SVM		<0.0001 (SVM > ST)	<0.0001 (SVM > LR)
ST			<0.0001 (ST > LR)

From Table 10 through Table 15, we observed the following:

1. When the cover depth is 10, the mean accuracy of RF is significantly greater than that of SVM, ST and LR.
2. When the cover depth is 20, the mean accuracies of RF, SVM and LR are significantly greater than that of ST. The accuracies of RF, SVM and LR are not significantly different.
3. When the cover depth is 30, the mean accuracy of SVM is significantly greater than that of RF and ST, while the mean accuracy of LR is significantly greater than that of ST. Although SVM has a slightly higher mean accuracy than LR, the difference is not significant.

4. When the cover depth is 40, the mean accuracy of SVM is significantly greater than that of the other three methods.
5. When we use the whole data, the mean accuracy of SVM is significantly greater than that of the other three methods.

From Figures 6 and 7, we can see that when the cover depth grows, the mean accuracies of the four classification methods have an increasing trend in general. Among the four, SVM and single decision tree have a strictly increasing trend. When the cover depth is 10, RF has the highest accuracy. When the cover depth is greater than or equal to 20, SVM has the highest accuracy.

Table 16 and Table 17 show the mean sensitivity and specificity, respectively, of the four methods. From the two tables, we see that SVM always has the highest mean sensitivities, which implies that SVM performs the best to detect the known heterozygous SNPs. SVM also has the highest specificity when the cover depth is greater than or equal to 40. However, RF has the highest specificity when the cover depth is 10 and LR has the highest specificity when cover depth is 20 or 30.

Figure 6: Average accuracies for different cover depths

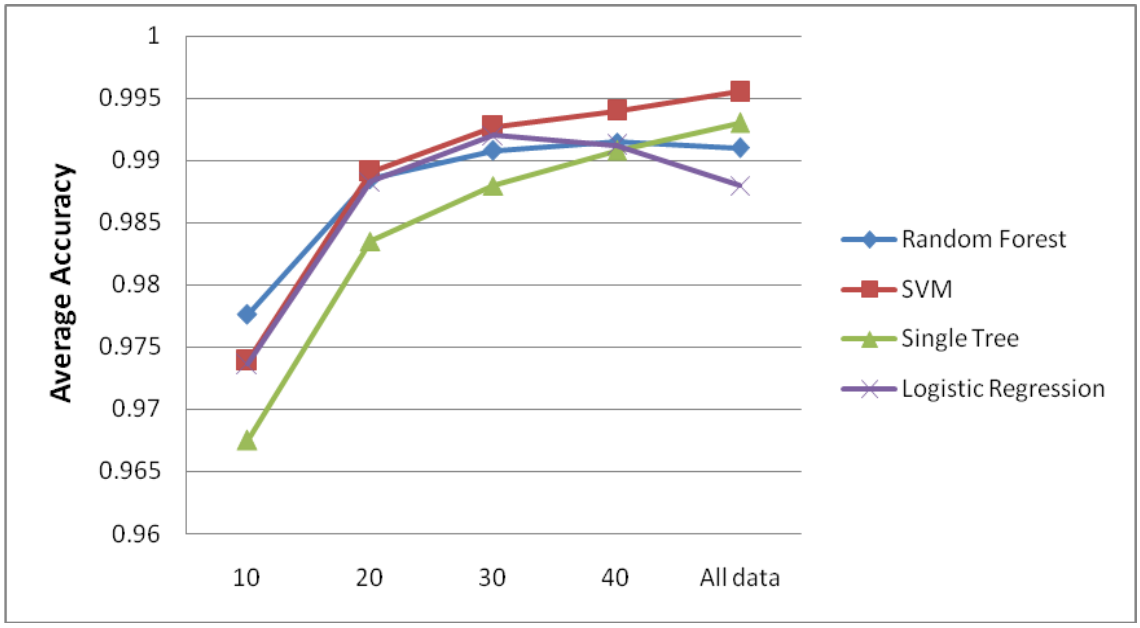


Figure 7: Average accuracies for different cover depths

Table 16: Mean sensitivities (sd) of the four classification methods

	Cover depth 10	Cover depth 20	Cover depth 30	Cover depth 40	Whole data
Random Forest	0.9738 (0.0116)	0.9904 (0.0062)	0.9934 (0.0049)	0.9937 (0.0047)	0.9924 (0.0014)
SVM	0.9763 (0.0106)	0.9924 (0.0060)	0.9957 (0.0050)	0.9965 (0.0046)	0.9927 (0.0000)
Single Decision Tree	0.9726 (0.0127)	0.9887 (0.0084)	0.9922 (0.0058)	0.9942 (0.0053)	0.9927 (0.0000)
Logistic Regression	0.9718 (0.0099)	0.9889 (0.0111)	0.9937 (0.0064)	0.9944 (0.0057)	0.9877 (0.0052)

Table 17: Mean specificities (sd) of the four classification methods

	Cover depth 10	Cover depth 20	Cover depth 30	Cover depth 40	Whole data
Random Forest	0.9811 (0.0082)	0.9867 (0.0056)	0.9884 (0.0044)	0.9895 (0.0039)	0.9897 (0.0034)
SVM	0.9716 (0.0111)	0.9860 (0.0088)	0.9900 (0.0059)	0.9916 (0.0056)	0.9981 (0.0040)
Single Decision Tree	0.9627 (0.0143)	0.9788 (0.0126)	0.9842 (0.0087)	0.9877 (0.0110)	0.9932 (0.0000)
Logistic Regression	0.9753 (0.0113)	0.9877 (0.0085)	0.9904 (0.0066)	0.9881 (0.0074)	0.9884 (0.0047)

Therefore, we conclude that SVM performs the best overall in this study, although RF performs the best when the cover depth is as small as 10.

Chapter 6

Variable Selection

6.1 Method

We ran the four classification models using the 8 measures we developed. Although the number of variables is not large, these variables are highly correlated. To enhance the generalization performance of the classification models, we conducted a variable selection. In this section, we use BW ratio and variable important ranking technique in random forest.

BW ratio is a widely used approach to determine the importance of the variables (Dudoit et al., 2002). By computing the between-group sum of squares (BSS) divided by the within-group sum of squares (WSS), we obtain the BW ratio. BW ratio is calculated for every variable, where the groups are the different classes of the response variable. The higher the BW ratio, the more important the variable is. For a particular variable j, the BW ratio is given as follows:

$$BW\ \ ratio(j) = \frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{\cdot j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2}$$

$$where\ \bar{x}_{\cdot j} = \frac{\sum_i x_{ij}}{N}, \quad \bar{x}_{kj} = \frac{\sum_{I(y_i=k)} x_{ij}}{N_k}.$$

Random Forest also calculates the variable importance ranking (Van der Laan et al., 2006). For each tree, the prediction accuracy on the out-of-bag portion of the data is recorded. Then the same is done after permuting one predictor variable at a time. The difference between the two accuracies is then averaged over all trees, and normalized by the standard error. These mean decreases in accuracy are the importance scores for variables. An important variable has a high importance score.

6.2 Variable Selection based on BW Ratio

Table 18 shows the BW ratios for the eight measures introduced in this study. From the table, we can see that the measures of chi-square values have significantly low scores. Hence, we selected measures 1, 2, 3, 4 and 8 and included in the classification models.

Table 18: BW ratios

Variable	Variable Label	BW ratio
Measure 8	non-reference allele ratio	17.97
Measure 3	ratio of indep. counts	16.80
Measure 4	ratio of log2 values	16.52
Measure 2	ratio of quality scores	12.81
Measure 1	ratio of absolute counts	10.20
Measure 5	chi-square of (A1+, A1-, A2+, A2-)	3.48
Measure 6	chi-square of (A1A2)	3.47
Measure 7	chi-square of (A1, A2, A3, A4)	3.14

Using the selected variables instead of all the eight measures, we conducted the same procedures to the same subset data. After calculating the mean accuracies, we conducted paired t-test to compare the performance of the four classification methods.

Table 19: Mean accuracies (SD) of the four classification methods

	Cover depth 10	Cover depth 20	Cover depth 30	Cover depth 40	Whole data
Random Forest	0.9802 (0.0096)	0.9890 (0.0038)	0.9909 (0.0034)	0.9924 (0.0028)	0.9929 (0.0007)
SVM	0.9723 (0.0079)	0.9875 (0.0065)	0.9926 (0.0048)	0.9938 (0.0033)	0.9961 (0.0020)
Single Decision Tree	0.9677 (0.0104)	0.9837 (0.0094)	0.9880 (0.0058)	0.9909 (0.0072)	0.9930 (<0.0001)
Logistic Regression	0.9742 (0.0102)	0.9880 (0.0065)	0.9915 (0.0046)	0.9909 (0.0052)	0.9906 (0.0017)

Table 20: P-values of paired t-test for cover depth 10

Cover Depth 10	SVM	ST	LR
RF	0.0001 (RF>SVM)	<0.0001 (RF>ST)	0.0006 (RF>LR)
SVM		0.0481 (ST>SVM)	0.2173
ST			0.0194 (ST>LR)

Table 21: P-values of paired t-test for cover depth 20

Cover Depth 20	SVM	ST	LR
RF	0.2115	0.0011 (RF>ST)	0.3408
SVM		0.0174 (SVM>ST)	0.6775
ST			0.0319 (LR>ST)

Table 22: P-values of paired t-test for cover depth 30

Cover Depth 30	SVM	ST	LR
RF	0.0551	0.0034 (RF>ST)	0.5382
SVM		0.0001 (SVM>ST)	0.2562
ST			0.0015 (LR>ST)

Table 23: P-values of paired t-test for cover depth 40

Cover Depth 40	SVM	ST	LR
RF	0.0257 (SVM>RF)	0.1609	0.1166
SVM		0.0205 (SVM>ST)	0.0192 (SVM>LR)
ST			1.0000

Table 24: P-values of paired t-test for the whole data

Whole data	SVM	ST	LR
RF	<0.0001 (SVM>RF)	0.3259	<0.0001 (RF>LR)
SVM		<0.0001 (SVM>ST)	<0.0001 (SVM>LR)
ST			<0.0001 (ST>LR)

Table 19 shows the average accuracies and standard deviations of the four classification methods using the selected variables by BW ratio for different cover depths and Table 20 through Table 24 show the p-value matrices. From these tables, we observed the following:

1. When the cover depth is 10, the mean accuracy of RF is significantly greater than that of SVM, ST and LR.
2. When the cover depth is 20 or 30, the mean accuracies of RF, SVM and LR are significantly greater than that of ST. The accuracies of RF, SVM and LR are not significantly different.

3. When the cover depth is 40, the mean accuracy of SVM is significantly greater than that of the other three methods.
4. When we use the whole data, the mean accuracy of SVM is significantly greater than that of the other three methods.

From Figures 8 and 9, we can see that the trend of the average accuracy is similar with that of Figures 6 and 7. When the cover depth grows, the mean accuracies increase in general. Among the four models, RF, SVM and ST have a strictly increasing trend. ST has the lowest accuracy for all the cover depths. Except for cover depths 10 and 20, SVM has the highest accuracy. When the cover depth is 10 and 20, the accuracy of RF is relatively greater than that of the other three.

Figure 8: Average accuracies for different cover depths (selected variables by BW ratio)

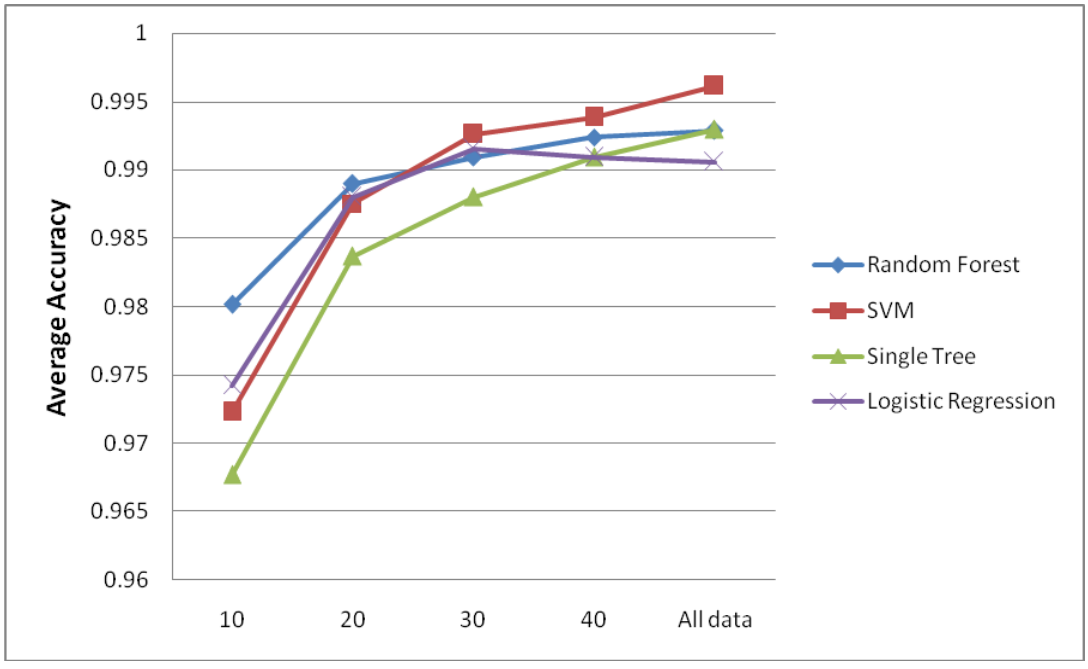


Figure 9: Average accuracies for different cover depths (selected variables by BW ratio)

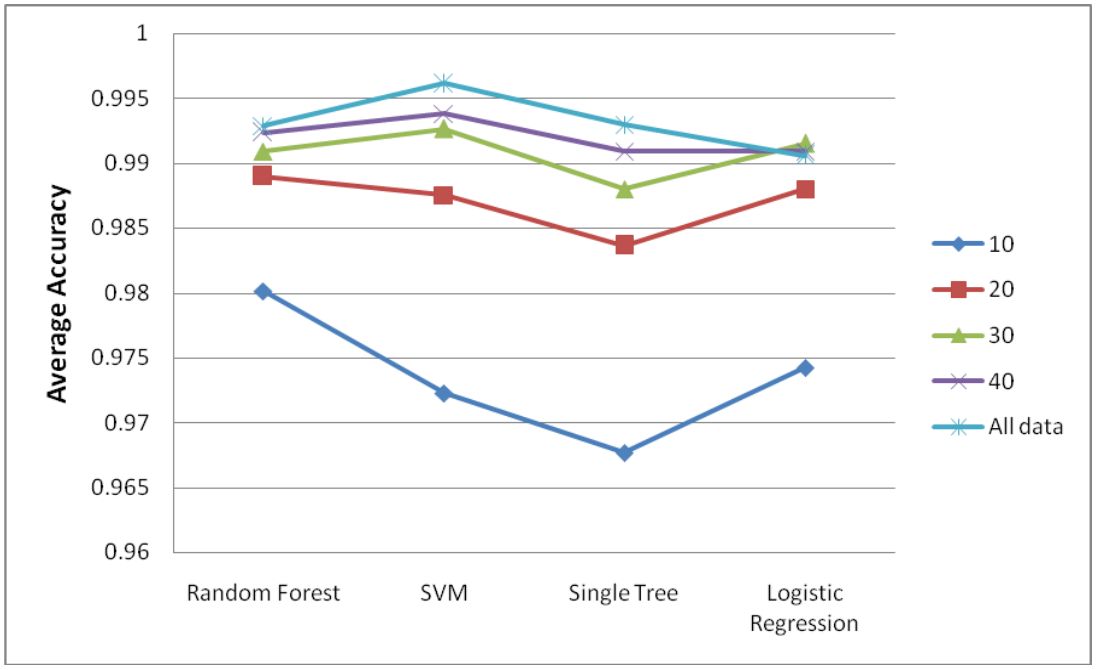


Table 25: Mean sensitivities (sd) of the four classification methods

	Cover depth 10	Cover depth 20	Cover depth 30	Cover depth 40	Whole data
Random Forest	0.9753 (0.0100)	0.9899 (0.0053)	0.9937 (0.0042)	0.9947 (0.0051)	0.9924 (0.0014)
SVM	0.9784 (0.0116)	0.9894 (0.0063)	0.9957 (0.0041)	0.9950 (0.0052)	0.9927 (0.0000)
Single Decision Tree	0.9726 (0.0131)	0.9891 (0.0077)	0.9922 (0.0058)	0.9942 (0.0053)	0.9927 (0.0000)
Logistic Regression	0.9733 (0.0119)	0.9859 (0.0093)	0.9912 (0.0063)	0.9899 (0.0090)	0.9867 (0.0028)

Table 26: Mean specificities (sd) of the four classification methods

	Cover depth 10	Cover depth 20	Cover depth 30	Cover depth 40	Whole data
Random Forest	0.9846 (0.0115)	0.9881 (0.0056)	0.9884 (0.0047)	0.9902 (0.0034)	0.9897 (0.0000)
SVM	0.9667 (0.0076)	0.9858 (0.0112)	0.9897 (0.0078)	0.9928 (0.0060)	0.9981 (0.0038)
Single Decision Tree	0.9632 (0.0139)	0.9786 (0.0128)	0.9842 (0.0087)	0.9879 (0.0109)	0.9932 (0.0000)
Logistic Regression	0.9751 (0.0125)	0.9900 (0.0067)	0.9918 (0.0064)	0.9918 (0.0052)	0.9884 (0.0024)

Table 25 and Table 26 show the mean sensitivity and specificity, respectively, of the four methods. From the two tables, we see that SVM always has the highest sensitivity. SVM also has the highest specificity when the cover depth is greater than or equal to 40. However, RF has the highest specificity when the cover depth is 10 and LR has the highest specificity when cover depth is 20 or 30.

In this section, we selected five variables based on the BW ratio. These selected variables were used to generate the classification models. According to the results above, we conclude that when the cover depth of the data is high, SVM

performs the best, although RF performs the best when the cover depth is as small as 20.

6.3 Variable Selection based on RF Variable Importance Ranking

Table 27 shows the RF mean decrease in accuracy for measuring variable importance ranking for the eight measures. Different from BW ratios, RF mean decrease in accuracy has an apparent trend. Except for measure 2, all the other measures have the score lower than 1. Since there is no formal inference (p-value) available for the random forest variable important measures, the variable selection is flexible. We wanted to check the performance of the chi-square values, but we also noticed that the measure 5 has the lowest value which is substantially smaller than the second lowest one; hence, we excluded measure 5 and used the others in the classification models.

Using the selected variables, we conducted the same procedures to the same subset data. We again conducted paired t-test to compare the performance of the four classification methods with the 7 selected variables.

Table 27: RF Variable importance ranking

Variable	Variable Label	RF Mean Decrease in Accuracy
Measure 2	ratio of quality scores	1.365
Measure 4	ratio of log2 values	0.950
Measure 1	ratio of absolute counts	0.948
Measure 8	non-reference allele ratio	0.852
Measure 7	chi-square of (A1, A2, A3, A4)	0.733
Measure 3	ratio of indep. counts	0.667
Measure 6	chi-square of (A1A2)	0.624
Measure 5	chi-square of (A1+, A1-, A2+, A2-)	0.203

Table 28: Mean accuracies (SD) of the four classification methods

	Cover depth 10	Cover depth 20	Cover depth 30	Cover depth 40	Whole data
Random Forest	0.9783 (0.0084)	0.9890 (0.0038)	0.9909 (0.0036)	0.9917 (0.0029)	0.9921 (0.0024)
SVM	0.9733 (0.0085)	0.9887 (0.0050)	0.9926 (0.0038)	0.9944 (0.0026)	0.9960 (0.0018)
Single Decision Tree	0.9675 (0.0102)	0.9835 (0.0097)	0.9880 (0.0058)	0.9908 (0.0073)	0.9930 (<0.0001)
Logistic Regression	0.9742 (0.0102)	0.9880 (0.0065)	0.9915 (0.0046)	0.9909 (0.0052)	0.9906 (0.0017)

Table 29: P-values of paired t-test for cover depth 10

Cover Depth 10	SVM	ST	LR
RF	0.0026 (RF>SVM)	<0.0001 (RF>ST)	0.0161 (RF>LR)
SVM		0.0327 (ST>SVM)	0.4749
ST			0.0155 (ST>LR)

Table 30: P-values of paired t-test for cover depth 20

Cover Depth 20	SVM	ST	LR
RF	0.7579	0.0009 (RF>ST)	0.3680
SVM		0.0036 (SVM>ST)	0.4765
ST			0.0315 (LR>ST)

Table 31: P-values of paired t-test for cover depth 30

Cover Depth 30	SVM	ST	LR
RF	0.0240 (SVM>RF)	0.0017 (RF>ST)	0.5621
SVM		0.0002 (SVM>ST)	0.2934
ST			0.0015 (LR>ST)

Table 32: P-values of paired t-test for cover depth 40

Cover Depth 40	SVM	ST	LR
RF	0.0002 (SVM>RF)	0.4823	0.4483
SVM		0.0142 (SVM>ST)	0.0008 (SVM>LR)
ST			0.9327

Table 33: P-values of paired t-test for cover depth data

Whole data	SVM	ST	LR
RF	<0.0001 (SVM>RF)	0.0698	0.0028 (RF>LR)
SVM		<0.0001 (SVM>ST)	<0.0001 (SVM>LR)
ST			<0.0001 (ST>LR)

Table 28 shows the average accuracies and standard deviations of the four classification methods using the selected variables by RF variable importance ranking for different cover depths and Table 29 through Table 33 show the p-value matrices. From these tables, we observed the following:

1. When the cover depth is 10, the mean accuracy of RF is significantly greater than that of SVM, ST and LR.
2. When the cover depth is 20 or 30, the mean accuracies of RF, SVM and LR are significantly greater than that of ST. The accuracies of RF, SVM and LR are not significantly different.
3. When the cover depth is 40, the mean accuracy of SVM is significantly greater than that of RF and LR.
4. When we use the whole data, the mean accuracy of SVM is significantly greater than that of the other three methods.

Figures 10 and 11 are similar to the previous average accuracy figures. When the cover depth grows, the mean accuracies increase in general. Among the four, RF, SVM and ST have a strictly increasing trend. Single decision tree has the lowest accuracy for all the cover depths. Except for cover depth 10, SVM has the highest accuracy. When the cover depth is 10, the accuracy of RF is relatively greater than those of the other three.

Figure 10: Average accuracies for different cover depths (selected variables by RF variable selection)

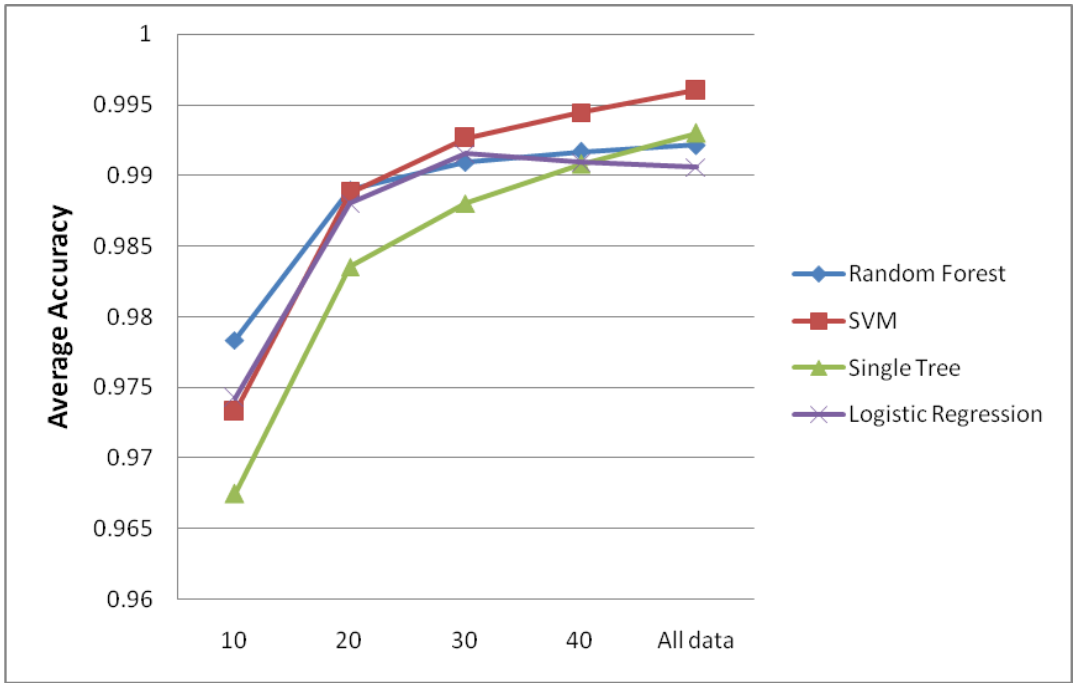


Figure 11: Average accuracies for different cover depths (selected variables by RF variable selection)

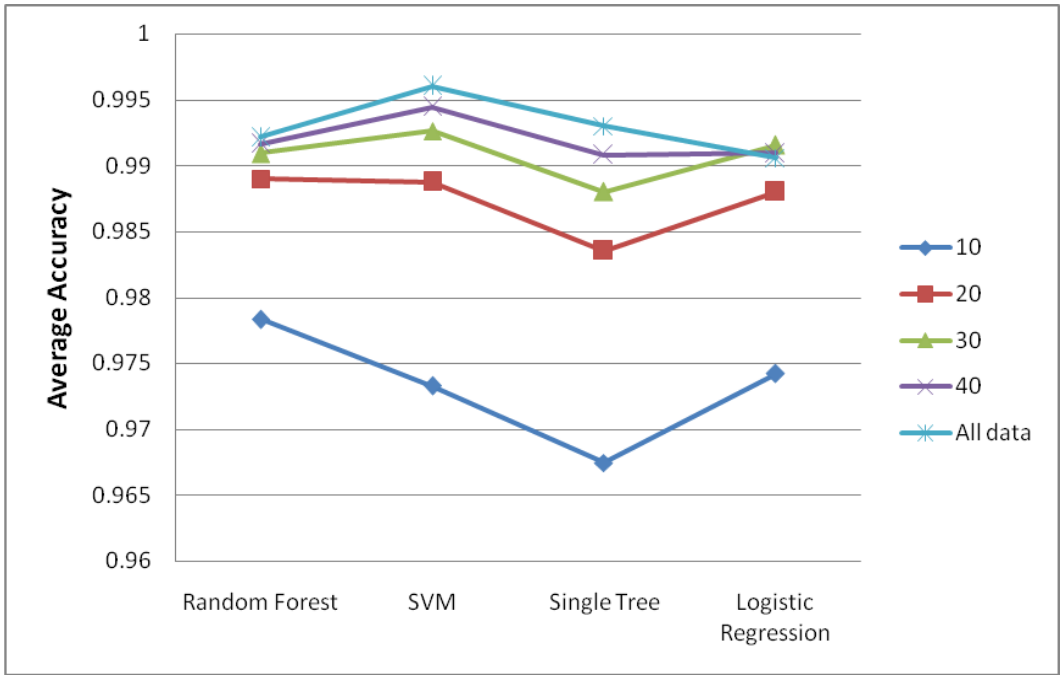


Table 34: Mean sensitivities (sd) of the four classification methods

	Cover depth 10	Cover depth 20	Cover depth 30	Cover depth 40	Whole data
Random Forest	0.9746 (0.0104)	0.9904 (0.0059)	0.9937 (0.0047)	0.9940 (0.0048)	0.9924 (0.0014)
SVM	0.9766 (0.0116)	0.9924 (0.0046)	0.9957 (0.0041)	0.9962 (0.0042)	0.9927 (0.0000)
Single Decision Tree	0.9726 (0.0127)	0.9887 (0.0084)	0.9922 (0.0058)	0.9942 (0.0053)	0.9927 (0.0000)
Logistic Regression	0.9733 (0.0119)	0.9859 (0.0093)	0.9912 (0.0063)	0.9899 (0.0090)	0.9867 (0.0028)

Table 35: Mean specificities (sd) of the four classification methods

	Cover depth 10	Cover depth 20	Cover depth 30	Cover depth 40	Whole data
Random Forest	0.9818 (0.0090)	0.9877 (0.0057)	0.9884 (0.0047)	0.9895 (0.0043)	0.9918 (0.0042)
SVM	0.9702 (0.0111)	0.9853 (0.0085)	0.9897 (0.0059)	0.9927 (0.0048)	0.9991 (0.0035)
Single Decision Tree	0.9627 (0.0143)	0.9788 (0.0126)	0.9842 (0.0087)	0.9877 (0.0110)	0.9932 (0.0000)
Logistic Regression	0.9751 (0.0125)	0.9900 (0.0067)	0.9918 (0.0064)	0.9918 (0.0052)	0.9942 (0.0024)

Table 34 and Table 35 show the mean sensitivity and specificity, respectively, of the four methods. From the two tables, we see that SVM always has the highest sensitivity. SVM also has the highest specificity when the cover depth is greater than or equal to 40. However, RF has the highest specificity when the cover depth is 10 and LR has the highest specificity when cover depth is 20 or 30.

In this section, we selected 7 variables based on the RF variable important ranking. According to the results above, we conclude that when the cover depth of the data is high, SVM performs the best, although RF performs the best when the cover depth is as small as 10.

6.4 Variable Selection based on BW Ratio and RF Variable Importance Ranking

After conducting variable selection based on BW ratio and random forest variable importance ranking, we also attempted to combine the two methods. From Table 36, we can see the measures of the chi-square values have relatively low scores for both BW ratio and RF mean decrease in accuracy. Moreover, RF

mean decrease in accuracy for measure 3 is 0.667, which is not significantly large compared to the other four out of the five measures with high BW ratio. Hence, we select measures 8, 4, 2 and 1 for the classification models.

Table 36: BW ratios and RF variable importance ranking

Variable	Variable Label	BW ratio	RF Mean Decrease in Accuracy
Measure 8	non-reference allele ratio	17.96628	0.852
Measure 3	ratio of indep. Counts	16.80413	0.667
Measure 4	ratio of log2 values	16.51956	0.950
Measure 2	ratio of quality scores	12.81418	1.365
Measure 1	ratio of absolute counts	10.19642	0.948
Measure 5	chi-square of (A1+, A1-, A2+, A2-)	3.476573	0.203
Measure 6	chi-square of (A1A2)	3.468348	0.624
Measure 7	chi-square of (A1, A2, A3, A4)	3.140116	0.733

Table 37: Mean accuracies (SD) of the four classification methods

	Cover depth 10	Cover depth 20	Cover depth 30	Cover depth 40	Whole data
Random Forest	0.9804 (0.0085)	0.9899 (0.0046)	0.9914 (0.0036)	0.9932 (0.0034)	0.9930 (<0.0001)
SVM	0.9728 (0.0075)	0.9881 (0.0070)	0.9934 (0.0048)	0.9948 (0.0033)	0.9965 (0.0016)
Single Decision Tree	0.9683 (0.0108)	0.9839 (0.0095)	0.9881 (0.0057)	0.9908 (0.0073)	0.9930 (<0.0001)
Logistic Regression	0.9737 (0.0110)	0.9904 (0.0061)	0.9920 (0.0050)	0.9938 (0.0045)	0.9930 (0.0016)

Table 38: P-values of paired t-test for cover depth 10

Cover Depth 10	SVM	ST	LR
RF	<0.0001 (RF>SVM)	<0.0001 (RF>ST)	0.0005 (RF>LR)
SVM		0.0516	0.5440
ST			0.0710

Table 39: P-values of paired t-test for cover depth 20

Cover Depth 20	SVM	ST	LR
RF	0.1381	0.0004 (RF>ST)	0.6018
SVM		0.0055 (SVM>ST)	0.0468 (LR>SVM)
ST			0.0005 (LR>ST)

Table 40: P-values of paired t-test for cover depth 30

Cover Depth 30	SVM	ST	LR
RF	0.0087 (SVM>RF)	0.0008 (RF>ST)	0.5315
SVM		<0.0001 (SVM>ST)	0.1548
ST			0.0026 (LR>ST)

Table 41: P-values of paired t-test for cover depth 40

Cover Depth 40	SVM	ST	LR
RF	0.0620	0.0479 (RF>ST)	0.3935
SVM		0.0030 (SVM>RF)	0.3615
ST			0.0213 (LR>ST)

Table 42: P-values of paired t-test for the whole data

Whole data	SVM	ST	LR
RF	<0.0001 (SVM>RF)	1	1
SVM		<0.0001 (SVM>ST)	<0.0001 (SVM>LR)
ST			1

Table 37 shows the average accuracies and standard deviations of the four classification methods using the selected variables for different cover depths and Table 38 through Table 42 show the p-value matrices. From these tables, we observed the following:

1. When the cover depth is 10, the mean accuracy of RF is significantly greater than that of SVM, ST and LR.
2. When the cover depth is 20, the mean accuracies of RF, SVM and LR are significantly greater than that of ST. The accuracies of RF, SVM and LR are not significantly different.

3. When the cover depth is 30, the mean accuracy of SVM is significantly greater than that of RF and ST, while the mean accuracy of LR is significantly greater than that of ST. Although SVM has a slightly higher mean accuracy than LR, the difference is not significant.
4. When the cover depth is 40, the mean accuracy of SVM is significantly greater than that of ST. There are no significant differences among RF, SVM and LR.
5. When we use the whole data, the mean accuracy of SVM is significantly greater than that of the other three methods.

Figures 12 and 13 are similar to the previous average accuracy figures. When the cover depth grows, the mean accuracies increase in general. Among the four models, SVM and ST have a strictly increasing trend. ST has the lowest accuracy for all the cover depths. Except for cover depth 10 and 20, SVM has the highest accuracy. When the cover depth is 10, the accuracy of RF is relatively greater than the other three. We also noticed that when we use the whole data, RF, ST and LR have exactly the same mean accuracy.

Figure 12: Average accuracies for different cover depths (selected variables)

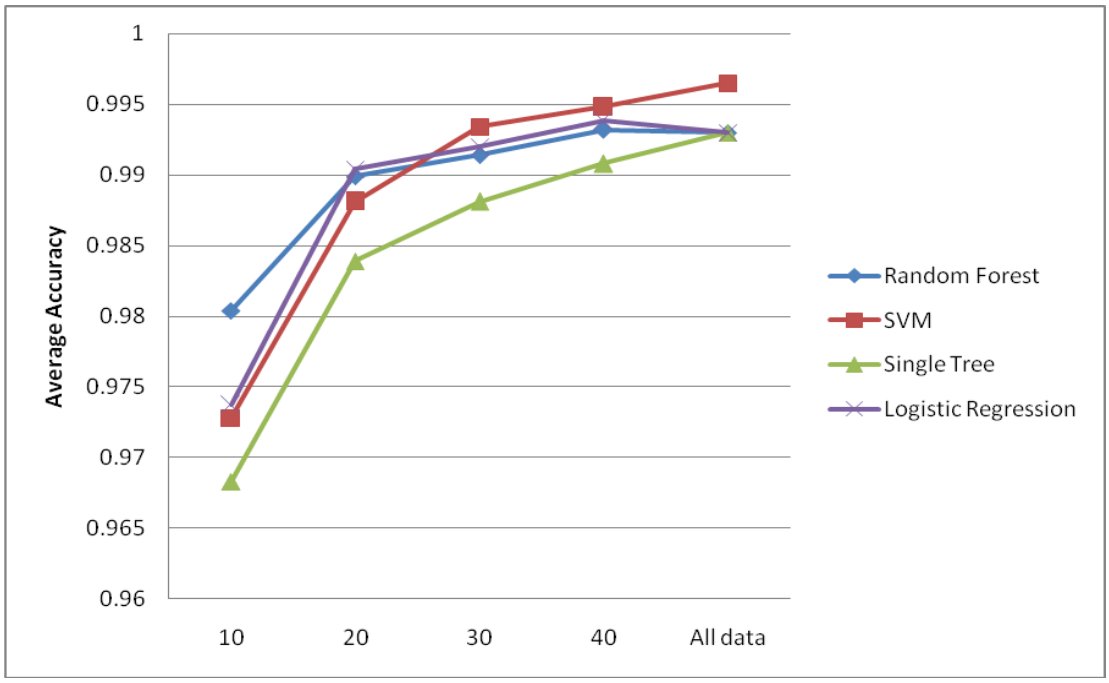


Figure 13: Average accuracies for different cover depths (selected variables)

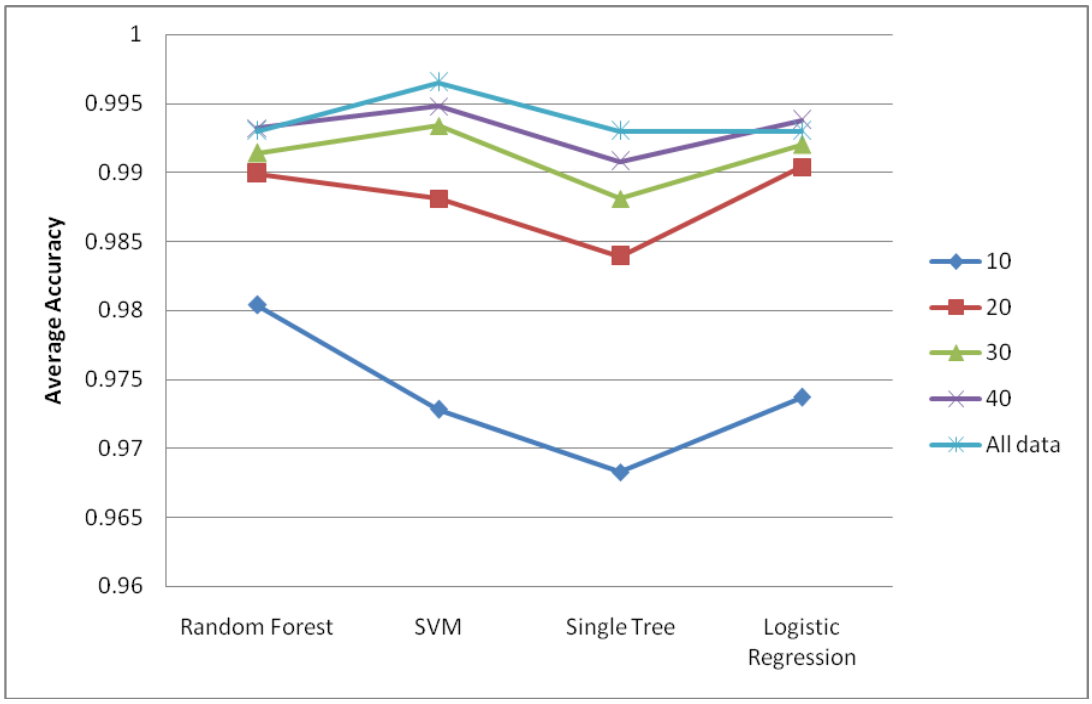


Table 43: Mean sensitivities (sd) of the four classification methods

	Cover depth 10	Cover depth 20	Cover depth 30	Cover depth 40	Whole data
Random Forest	0.9771 (0.0095)	0.9904 (0.0062)	0.9939 (0.0034)	0.9955 (0.0045)	0.9927 (0.0000)
SVM	0.9786 (0.0122)	0.9912 (0.0066)	0.9955 (0.0045)	0.9955 (0.0049)	0.9932 (0.0019)
Single Decision Tree	0.9728 (0.0138)	0.9889 (0.0082)	0.9924 (0.0057)	0.9940 (0.0059)	0.9927 (0.0000)
Logistic Regression	0.9728 (0.0126)	0.9892 (0.0072)	0.9914 (0.0062)	0.9945 (0.0046)	0.9917 (0.0026)

Table 44: Mean specificities (sd) of the four classification methods

	Cover depth 10	Cover depth 20	Cover depth 30	Cover depth 40	Whole data
Random Forest	0.9835 (0.0115)	0.9895 (0.0056)	0.9890 (0.0047)	0.9911 (0.0034)	0.9932 (0.0000)
SVM	0.9674 (0.0076)	0.9853 (0.0112)	0.9916 (0.0078)	0.9942 (0.0060)	0.9995 (0.0038)
Single Decision Tree	0.9641 (0.0139)	0.9793 (0.0128)	0.9842 (0.0087)	0.9879 (0.0109)	0.9932 (0.0000)
Logistic Regression	0.9746 (0.0125)	0.9916 (0.0067)	0.9925 (0.0064)	0.9932 (0.0052)	0.9942 (0.0024)

Table 43 and Table 44 show the mean sensitivity and specificity, respectively, of the four methods. The result is almost the same as that of the previous sensitivity and specificity tables. We see that SVM always has the highest sensitivity. SVM also has the highest specificity when the cover depth is greater than or equal to 40. However, RF has the highest specificity when the cover depth is 10 and LR has the highest specificity when cover depth is 20 or 30.

According to the results above, we conclude that when the cover depth of the data is high, SVM performs the best, although RF performs the best when the cover depth is 10.

6.5 Comparison of the Models using All Variables and Selected Variables

In the previous sections, we made the conclusion about the best model using either all variables or selected variables while cover depth is fixed. When the cover depth is greater than 20, SVM has the best performance among the four classification methods in general. When the cover depth is 20, SVM, RF and

logistic regression are equally good. When the cover depth is 10, RF always performs the best.

For a fixed cover depth data, we need to decide whether to conduct variable selection before finding the optimal classification model. Therefore, we compared the mean accuracies of the best models using all variables or selected variables for each cover depth level, by conducting paired t-test.

Table 45: Paired t-test for all variables and selected variables

Cover Depth	All Variables		BW ratio		RF variable selection		Both methods	
	Best Model	Mean Accuracy (SD)	Best Model	Mean Accuracy (SD)	Best Model	Mean Accuracy (SD)	Best Model	Mean Accuracy (SD)
10	RF	0.9776 (0.0083)	RF	0.9802 (0.0096)	RF	0.9783 (0.0084)	RF	0.9804 (0.0085)
20	SVM	0.9891 (0.0053)	RF	0.9890 (0.0038)	RF	0.9890 (0.0038)	RF	0.9899 (0.0046)
30	SVM	0.9927 (0.0040)	SVM	0.9926 (0.0048)	SVM	0.9926 (0.0038)	SVM	0.9934 (0.0048)
40	SVM	0.9940 (0.0031)	SVM	0.9938 (0.0033)	SVM	0.9944 (0.0026)	SVM	0.9948 (0.0033)
All Data	SVM	0.9955 (0.0021)	SVM	0.9961 (0.0020)	SVM	0.9960 (0.0018)	SVM	0.9965 (0.0016)

Table 45 shows the best model and the corresponding mean accuracy and standard deviation for a particular combination of cover depth and variable selection.

Table 46: P-values of paired t-test for cover depth 10

Cover depth 10	BW ratio	RF variable selection	Both Method
All variables	0.0035	0.2460	0.0056
BW ratio		0.0701	0.7869
RF variable selection			0.0509

Table 47: P-values of paired t-test for cover depth 20

Cover depth 20	BW ratio	RF variable selection	Both Method
All variables	0.4238	0.1609	0.0257
BW ratio		1.0000	0.0433
RF variable selection			0.1033

Table 48: P-values of paired t-test for cover depth 30

Cover depth 30	BW ratio	RF variable selection	Both Method
All variables	0.8513	0.7689	0.3517
BW ratio		1	0.0698
RF variable selection			0.2144

Table 49: P-values of paired t-test for cover depth 40

Cover depth 40	BW ratio	RF variable selection	Both Method
All variables	0.8316	0.2930	0.2434
BW ratio		0.2829	0.0029
RF variable selection			0.5864

Table 50: P-values of paired t-test for whole data

Whole data	BW ratio	RF variable selection	Both Method
All variables	0.0961	0.0433	0.0088
BW ratio		0.6626	0.0831
RF variable selection			0.1033

Tables 46 through Table 50 show the p-value matrices of paired t-test for different cover depths. By checking the p-values and the average accuracies, we observed the following:

1. When the cover depth is 10, the mean accuracy of RF using the selected variables by BW ratios or both variable selection methods is significantly greater than using all variables or RF variable selection.
2. When the cover depth is 20, the mean accuracy of RF using both variable selection methods is significantly greater than using all variables or each of the other two variable selection methods.
3. When the cover depth is 30, there is no significant difference among the four different sets of variables.
4. When the cover depth is 40, the mean accuracy of SVM using both variable selection methods is significantly greater than using BW ratio. But there is no significant evidence to conclude that using both variable selection methods is the best.
5. When we use the whole data, the mean accuracy of SVM using both variable selection methods is significantly greater than using the other three methods.

From the result above, we found that the models perform better when we apply variable selection rather than using all variables. Among the three variable selection methods, the combination of BW ratio and RF variable importance ranking always improve the performance of each classification model. We conclude that for the small cover depth data such as 20 or less, the best model is RF using variables selected by the combination of BW ratio and RF variable importance ranking, but for the large cover depth data, the best model is SVM using variables selected by the combination of BW ratio and RF variable importance ranking.

Chapter 7

Conclusion and Discussion

In the first part of the study, we had very small training data including only 19 known heterozygous SNPs and 2 homozygous SNPs. We developed three measures based on the data format and then decided the threshold of the measures by examining the properties of these known SNPs. Using 10-fold cross validation, we developed an algorithm to decide the optimal thresholds of the three measures. The thresholds are then used to classify the genomic positions. The positions with the highest proportion in measure (1) ≥ 0.95 or the highest proportion of measure (2) ≥ 0.95 are classified as homozygous position; the positions with $0.2 \leq$ the highest and second highest proportions in measure (1) ≤ 0.8 or $0.175 \leq$ the highest and second highest proportions in measure (2) ≤ 0.825 are classified as heterozygous positions. The positions which do not belong to either clearly homozygous bases or clearly heterozygous bases are assigned to the middle group. Using this approach we predicted 8 homozygous SNP candidates and 68 heterozygous SNP candidates in the previous data set. All the known 13 SNPs are detected among these SNP candidates.

In the second part of the study, we had a substantially larger data set with 308 known SNPs. We applied four widely used classification methods: random forest, SVM, single decision tree and logistic regression. Before generating the models, we developed five more measures based on the properties of the SNPs. Since biologists often cannot obtain the data with sufficient reads in practice, we evaluated the four classification methods based on different cover depths. As expected, when the cover depth grows, the mean accuracies of the four classification methods have an increasing trend in general. According to the result of the paired t-tests, we found that SVM has the best performance when the cover depth is greater than 10 and random forest performs the best when the cover depth is 10.

Since the eight measures developed based on the original information of positions are highly correlated, we conducted variable selection to enhance the generalization performance of the classification models. BW ratio and random forest variable importance ranking were used in the study. By applying both the BW ratios and RF mean decrease in accuracy, we selected four measures for the classification models. Using the selected variables, we evaluated the performances of the four classification methods based on the subset data with different cover depths. The result is similar to that using all the variables. SVM performs the best when the cover depth is relatively large and RF performs the best when the cover depth is small. Moreover, after the variable selection, the average accuracies of the best models for different cover depths are statistically higher than those using all the variables in general. Therefore, we conclude that when the cover depth of the data is large, such as greater than 20, SVM using the variables selected by the combination of the two variable selection methods has the best performance, and random forest performs the best when the cover depth is small.

We applied the SVM using the selected measures 1, 2, 4 and 8 to the whole data to predict the potential SNPs. There are 213 unidentified positions classified as heterozygous SNP candidates and 100 are classified as the homozygous SNP candidates.

Chapter 8

Future Study

Using the same data format as described in earlier chapters, biologists often predict SNPs only depending on the counts of nucleotides but not using quality scores. Our result is close to their prediction and it provides more information. The predicted SNP candidates are being tested in the lab. After obtaining the result, we will evaluate the accuracy of our prediction and then modify our method. Moreover, for the data with limited number of known SNPs, we still have positions in the hard-to-classify group. In future study, we will develop a probability model to identify these positions.

To make the model developed in this study accessible, we will work on combining the procedures including data format transformation, measure calculation, model determination and prediction to an executable web-based program. After biologists use our program to classify their data, we will gather the feedback and make improvement of our model.

We will also try to add some more measures based on the properties of the SNPs. We also plan to study other classification models and compare them with the four classification methods used in this study.

References

Alpaydin, E. (2004), "*Introduction to machine learning*", MIT Press, Cambridge, MA.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., *et al.* (2008), "Accurate whole human genome sequencing using reversible terminator chemistry", *Nature*, 456: 53-59.

Blum, A., Kalai, A., Langford, J. (1999), "Beating the hold-out: bounds for K-fold and progressive cross-validation", *Association for Computing Machinery*, 203-208.

Breiman, L., Friedman, J., Olshen, R., Stone. J. (1983), "*Classification and Regression Trees"*, Wadsworth, Belmont, CA.

Breiman., L. (2001), "Random forests", *Machine Learning*, 45(1):5-32.

Breiman., L. (1996), "Bagging predictors", *Machine Learning*, 24:123-140.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C.R., Lim, E.P., Kalayanaraman, N., Nemesh, J. *et al.* (1999), "Characterization of single-nucleotide polymorphisms in coding regions of human genes", *Nat Genet*, 22:231-238.

Carlson, C.S., Newman, T.L., Nickerson, D.A. (2001), "SNPing in the human genome", *Chemical Biology*, 5:78–85.

Chakravarti, A. (1999), "Population genetics–making sense out of sequence", *Nature Genet*, 21:56–60.

Clifford, R., Edmonson, M., Hu, Y., Nguyen, C., Scherpbier, T., Buetow, K.H. (2000), "Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project", *Genome Res*, 10:1259-1265.

Collins, F.S., Guyer, M.S., Charkravarti, A. (1997), "Variations on a theme: cataloging human DNA sequence variation", *Science*, 278:1580–1581

Cooil, B., Winer, R.S., Rados, D.L., (1987), "Cross-Validation for Prediction", *Journal of Marketing Research*, Vol. 24, No. 3, pp. 271-279.

Cooper, D.N., Smith, B.A., Cooke, H.J., Niemann, S., Schmidtke, J. (1985), "An estimate of unique DNA sequence heterozygosity in the human genome", *Hum. Genet.*, 69:201–205.

Cortes, C. and Vapnik, V.N. (1995), "Support vector networks", *Machine Learning*, 20:273-297.

De'ath, G., Fabricius, K.E. (2000), "Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis", *Ecology*, 81(11): 3178-3192.

Dudoit, S., Fridlyand, J., Speed., T.P. (2002), "Comparison of discrimination methods for the classi_cation of tumors using gene expression data", *Journal of the American Statistical Association*, 97:77-87.

Ewing, B., Hillier, L., Wendl, M.C., Green, P. (1998), "Base-calling of automated sequencer traces using Phred. I. Accuracy assessment", *Genome Res*, 8:175-185.

Ewing, B., Green, P. (1998), "Base-calling of automated sequencer traces using Phred. II. Error probabilities", *Genome Res*, 8:186-194.

Galas, D.J., McCormack, S.J. (2002), "*Genomic Technologies: Present and Future*". Caister Academic Press, Wymondham, UK.

Halushka, M., Fan, J.B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., Chakravarti, A. (1999), " Patterns of single nucleotide polymorphisms in candidate genes regulating blood pressurehomeostasis", *Nature Genet*, 22:239–247.

International HapMap Consortium (2003), "The International HapMap Project", *Nature*,426(6968):789-96.

Jou, W.M., Haegeman, G., Ysebaert, M., Fiers, W. (1972), "Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein", *Nature*. 237(5350):82-88.

Kohavi, R. (1995), "A study of cross-validation and bootstrap for accuracy estimation and model selection. In C. S. Mellish (Ed.)", *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1137–1143.

Kwok, P.Y. (2001), "Methods for genotyping single nucleotide polymorphisms", *Annu. Rev. Genomics Hum. Genet.*, 2:235–258.

Kwok, P.Y., Carlson, C., Yager, T.D., Ankener, W., Nickerson, D.A. (1994), "Comparative analysis of human DNA variations by fluorescencebased sequencing of PCR products", *Genomics*, 23:138-144.

Lander, E.S. (1996), "The new genomics: global views of biology", *Science*, 274, 536-539.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z.T., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., *et al*. (2005), "Genome sequencing in microfabricated high-density picolitre reactors", *Nature*, 437, 376–380.

Marnellos, G. (2003), "High-throughput SNP analysis for genetic association studies", *Curr Opin Drug Discov Devel.*, 6(3):317-321.

Nakajima, H., Furutama, D., Kimura, F., Shinoda, K., Ohsawa, N., Nakagawa, T., Shimizu, A. (1998), "Herpes simplex Virus Myelitis: Clinical Manifestations and Diagnosis by the Polymerase Chain Reaction Method", *Eur Neurol* ,39:163-167.

Pitman, S.D. (2001), "DNA mutation rates and Evolution", *American Journal of Human genetics.*

Risch, N., Merikangas, K. (1996), "The future of genetic studies of complex human diseases", *Science*, 273:1516–1517.

Shendure, J., Ji, H. (2008), "Next-generation DNA sequencing", *Nature Biotechnology*, 26:1135 – 1145.

Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L., Kwok, P.Y. (1998), "Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms", *Genome Res*, 8:748-754.

Van der Laan, M.J. (2006) "Statistical Inference for Variable Importance," *The International Journal of Biostatistics*, Vol. 2.

Vapnik, V.N. (1995), "*The Nature of Statistical Learning Theory",* Springer Verlag, NY.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., et al. (1998), "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome." *Science*, 280:1077–1082.