Robust Genotype Classification Using Dynamic Variable Selection

by

Mohua Podder

MSTAT, Indian Statistical Institute, 2003
B.Sc., Calcutta University, 2001

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August, 2008

Abstract

Single nucleotide polymorphisms (SNPs) are DNA sequence variations, occurring when a single nucleotide -A, T, C or G - is altered. Arguably, SNPs account for more than 90% of human genetic variation. Dr. Tebbutt's laboratory has developed a highly redundant SNP genotyping assay consisting of multiple probes with signals from multiple channels for a single SNP, based on arrayed primer extension (APEX). The strength of this platform is its unique redundancy having multiple probes for a single SNP. Using this microarray platform, we have developed fully-automated genotype calling algorithms based on linear models for individual probe signals and using dynamic variable selection at the prediction level. The algorithms combine separate analyses based on the multiple probe sets to give a final confidence score for each candidate genotypes.

Our proposed classification model achieved an accuracy level of $> 99.4\%$ with 100% call rate for the SNP genotype data which is comparable with existing genotyping technologies. We discussed the appropriateness of the proposed model related to other existing high-throughput genotype calling algorithms.

In this thesis we have explored three new ideas for classification with high dimensional data: (1) ensembles of various sets of predictors with built-in dynamic property; (2) robust classification at the prediction level; and (3) a proper confidence measure for dealing with failed predictor(s).

We found that a mixture model for classification provides robustness against outlying values of the explanatory variables. Furthermore, the algorithm chooses among different sets of explanatory variables in a dynamic way, prediction by prediction. We analyzed several data sets, including real and simulated samples to illustrate these features. Our model-based genotype calling algorithm captures the redundancy in the system considering all the underlying probe features of a particular SNP, automatically down-weighting any 'bad data' corresponding to image artifacts on the microarray slide or

failure of a specific chemistry.

Though motivated by this genotyping application, the proposed methodology would apply to other classification problems where the explanatory variables fall naturally into groups or outliers in the explanatory variables require variable selection at the prediction stage for robustness.

Table of Contents

List of Tables

List of Figures

Glossary of Some Genetic Terms

(All the definitions are in either the website
http://www.genome.gov/glossary.cfm or in the Glossary of
Gene-Environment Meeting at iCAPTURE.)

Allele One of the variant forms of a gene at a particular locus, or location, on a chromosome. Different alleles produce variation in inherited characteristics such as hair color or blood type. In an individual, one form of the allele (the dominant one) may be expressed more than another form (the recessive one).

Nucleotide One of the structural components, or building blocks, of DNA and RNA. A nucleotide consists of a base (one of four chemicals: adenine, thymine, guanine, and cytosine) plus a molecule of sugar and one of phosphoric acid.

Adenine One of the four bases in DNA that make up the letters ATGC, adenine is the "A". The others are guanine, cytosine, and thymine. Adenine (A) always pairs with thymine (T) and cytosine (C) always pairs with guanine (G).

APEX (arrayed primer extension) A type of genotyping method that allows multiple genotypes to be generated in one experiment. The method utilizes several oligonucleotide probes that are specific for a particular SNP and match the sequence just before the polymorphic site. The end of each oligonucleotide is modified to allow its covalent attachment to a glass slide. Each SNP is amplified by PCR and the DNA added to the bound oligonucleotides on the glass slide. An extension reaction is carried out using a thermostable DNA polymerase. Four unique, fluorescently-labelled nucleotides are used to extend each probe by only one base, dependent on the individual's template DNA. Following removal of all unincorporated dye as well as the template DNA, the SNPs are detected by the wavelength of fluorescence from the dyes at each site on the array.

Asper Biotech Asper Biotech is a genetic testing company with an established set of robust and efficient DNA tests. Asper is also a reliable partner

for the scientific and commercial communities in their custom genotyping projects (http://www.asperbio.com/).

Base pair Two bases which form a "rung of the DNA ladder." A DNA nucleotide is made of a molecule of sugar, a molecule of phosphoric acid, and a molecule called a base. The bases are the "letters" that spell out the genetic code. In DNA, the code letters are A, T, G, and C, which stand for the chemicals adenine, thymine, guanine, and cytosine, respectively. In base pairing, adenine always pairs with thymine, and guanine always pairs with cytosine.

Candidate gene A gene, located in a chromosome region suspected of being involved in a disease, whose protein product suggests that it could be the disease gene in question.

Chromosome One of the threadlike "packages" of genes and other DNA in the nucleus of a cell. Different kinds of organisms have different numbers of chromosomes. Humans have 23 pairs of chromosomes, 46 in all: 44 autosomes and two sex chromosomes. Each parent contributes one chromosome to each pair, so children get half of their chromosomes from their mothers and half from their fathers.

Coriell The Coriell Cell Repositories provide research reagents to the scientific community by establishing, maintaining and distributing cell cultures and DNA derived from the cell cultures.

Deoxyribonucleic acid (DNA) The molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between base pairs of nucleotides. The four nucleotides in DNA contain the bases: adenine (A), guanine (G), cytosine (C), and thymine (T). In nature, base pairs form only between A and T and between G and C; thus the base sequence of each single strand can be deduced from that of its partner.

DNA sequencing Determining the exact order of the base pairs in a segment of DNA.

Gene The term coined by Johannsen (1909) for the fundamental physical and functional unit of heredity. The word gene was derived from De Vries' term pangen, itself a derivative of the word pangenesis which Darwin (1868) had coined. A gene is an ordered sequence of nucleotides located in a particular position (locus) on a chromosome that encodes a specific functional product (the gene product, *i.e.*, a protein or RNA molecule). It includes regions involved in regulation of expression and regions that code for a specific functional product.

Gene expression The process by which a gene's coded information is converted into the structures present and operating in the cell. Expressed genes include those that are transcribed into mRNA and then translated into pro-

tein and those that are transcribed into RNA but not translated into protein (*e.g.*, transfer and ribosomal RNAs).

Genetic marker A segment of DNA with an identifiable physical location on a chromosome and whose inheritance can be followed. A marker can be a gene, or it can be some section of DNA with no known function. Because DNA segments that lie near each other on a chromosome tend to be inherited together, markers are often used as indirect ways of tracking the inheritance pattern of a gene that has not yet been identified, but whose approximate location is known.

Genome The entire complement of genetic material in an organism.

Genotype A somewhat poorly defined term. Most often it refers to the set of alleles at a single point on a chromosome. Confusingly, it is also used to mean an organism's entire genetic makeup (thus overlapping with the term "genome").

Haplotype The alleles of a set of closely linked genetic markers present on one chromosome which tend to be inherited together.

Haplotype tag SNP A SNP that can be used to identify, *i.e.*, tag, a haplotype in a given region of the genome. The information from a haplotype tag SNP (htSNP) can be used to infer the presence of all the alleles that form the haplotype.

Homozygous An individual who has inherited two identical copies of an allele at a particular locus.

Heterozygous An individual who has inherited two different alleles at a particular locus.

Human Genome Project An international research project to map each human gene and to completely sequence human DNA.

Linkage Disequilibrium Linkage disequilibrium (often termed "allelic association") is a situation when alleles at two separate loci occur together on chromosomes more frequently than expected by chance alone. Linkage disequilibrium tends to only occur between loci that are very close to each other on a chromosome.

Locus The position of a gene or a genetic marker on a chromosome. Plural: loci.

Microarray Spotter A high-precision robot with metal pins that dip into a DNA solution, suck up a specific volume and deposit the DNA onto a glass slide in a pre-arranged pattern.

Microarray technology A new way of studying how large numbers of genes interact with each other and how a cell's regulatory networks control vast batteries of genes simultaneously. The method uses a robot to precisely apply tiny droplets containing functional DNA to glass slides. Researchers

then attach fluorescent labels to DNA from the cell they are studying. The labeled probes are allowed to bind to complementary DNA strands on the slides. The slides are put into a scanning microscope that can measure the brightness of each fluorescent dot; brightness reveals how much of a specific DNA fragment is present, an indicator of how active it is.

Normalised DNA A group of DNA samples that has been adjusted so that each sample is at the same concentration, thus making subsequent genotyping easier and more accurate.

Nucleotides The building blocks of nucleic acids such as DNA.

Oligo Oligonucleotide, short sequence of single-stranded DNA or RNA. Oligos are often used as probes for detecting complementary DNA or RNA because they bind readily to their complements.

PHASE A program that estimates haplotypes from a set of genotypes found in a given person. Haplotypes are usually difficult to determine experimentally and therefore we rely on statistical means to estimate what the haplotypes are in a person. PHASE can be accessed on the
web (http://www.stat.washington.edu/stephens/phase.html).

Phenotype The term coined by Johannsen (1909) for the appearance (Gk. phainein, to appear) of an organism with respect to a particular character or group of characters (physical, biochemical, and physiologic), as a result of the interaction of its genotype and its environment. We most often use the different diseases that we study as the phenotype of interest.

Polymerase chain reaction (PCR) A fast, inexpensive technique for making an unlimited number of copies of any piece of DNA. Sometimes called "molecular photocopying," PCR has had an immense impact on biology and medicine, especially genetic research.

Polymorphism Difference in DNA sequence among individuals. Usually the term only applies to genetic variations occurring at a frequency of more than 1%.

Primer short oligonucleotide sequence used in a polymerase chain reaction.

Probe A piece of labeled DNA or RNA or an antibody used to detect the function of a gene.

Ribonucleic acid (RNA) A chemical similar to a single strand of DNA. In RNA, the letter U, which stands for uracil, is substituted for T in the genetic code. RNA delivers DNA's genetic message to the cytoplasm of a cell where proteins are made.

Sepsis A serious bacterial infection of the blood. Sepsis is more common in the elderly and in neonates. Symptoms include high fever, chills, decreased urine output, and a decreased level of consciousness.

SIRS The full form is the systemic inflammatory response syndrome. SIRS

can be the inflammatory response to sepsis or the response to non-infectious stimuli such as cardiopulmonary bypass.

SNP (single nucleotide polymorphism) A DNA sequence variation that involves a change in a single nucleotide (SNP is often pronounced as "snip").

TaqMan TaqMan is a type of assay patented by Roche Molecular Systems. We use this assay for genotyping the samples in our cohorts. In general, TaqMan assays utilize an oligonucleotide probe that is specific for the target gene. This probe is labelled with a fluorescent tag and a quenching molecule. During the extension step of a PCR the Taq enzyme will disrupt probe bound to the target separating the fluorescent tag from its quencher molecule thus permitting fluorescence. For genotyping assays, we use two probes rather than one. Each probe is specific for a given allele of a polymorphism. The two probes can be distinguished because they are labelled with different tags that fluoresce at different wavelengths.

Acknowledgements

First and foremost, I want to acknowledge the role of my supervisory committee. My thesis was co-supervised by Professors William J. Welch and Ruben H. Zamar. The other member of my committee, Dr. Scott J. Tebbutt, was my *de facto* co-supervisor on biomedical aspects. It is very difficult to describe in words how much I have learned from them in the last four years. Indeed, without their invaluable guidance, constant encouragement, precise criticism and sustained interest throughout the entire period of my Ph.D. work, I would not have been able to complete this thesis.

My indebtedness to the faculty members of the Statistics department of University of British Columbia is immense. It has been a joy learning from all of them. I would also like to thank Ms. Elaine Salameh, Ms. Christine Graham, Ms. Rhoda Morgan, Ms. Peggy Ng and Ms. Viena Tran for their sustained help during my early days in Canada and with administrative matters.

It has been a learning experience at the iCAPTURE Centre and I would like to acknowledge Dr. Keith R. Walley, James Russell and Jian Ruan for providing me the data sets, which became an important part of my thesis.

I am grateful to my colleagues in the Department of Statistics for making my time in the department a memorable one. I especially thank Guohua Yan for his illuminating discussions in computer programming.

My heartfelt thanks to my family - my father, mother and my husband - for their patience, support and understanding. They allowed me to stay preoccupied with my research neglecting my family duties, and without their love, affection and encouragement, I would have found it very difficult to pull along.

Financial support from the iCAPTURE Centre, Allergen NCE, National Sanitarium Association and NSERC, MITACS are gratefully acknowledged.

To my daughter Renisa

Statement of Co-Authorship

This thesis is completed under the supervision of Prof. William J. Welch, Prof. Ruben H. Zamar and Dr. Scott J. Tebbutt. They directed the research from start to finish through frequent research meetings; they contributed numerous critical ideas both globally and in research details. Dr. Tebbutt designed the microarray SNP genotyping chip and provided the data. They also guided me enormously in revising the manuscripts. Individual chapter has the following co-authorship statements.

Chapter 2: Mohua Podder designed and developed the algorithms, performed the statistical analysis of the data sets and drafted the manuscript; all the authors contributed to the writing of the final version. William J. Welch and Ruben H. Zamar supervised in developing the algorithms. Scott J. Tebbutt designed the APEX microarray chip and provided the data sets. All authors read and approved the final manuscript.

Chapter 3: Mohua Podder performed the linear discriminant analyses (LDA) using dynamic variable selection. Jian Ruan performed the wetlab experiments described in this study, and assisted in the design of the initial multiplex PCR. Ben W. Tripp performed the image analysis and MACGT auto-calling and analysis steps, and assisted Scott J. Tebbutt in the manual genotype calling. Zane E. Chu helped design the 50-plex PCR primers, and undertook initial experimental evaluation of these primers. Mohua Podder, Jian Ruan, Ben W. Tripp and Scott J. Tebbutt discussed the results and contributed to the preparation of this manuscript. Scott J. Tebbutt designed and supervised the experiments and analyses, and wrote the paper.

Chapter 4: Mohua Podder designed and developed the algorithms, performed the statistical analysis of the data sets and drafted the manuscript; all the authors contributed to the writing of the final version. William J. Welch and Ruben H. Zamar supervised in developing the algorithms.

Chapter 1

Introduction

In 2004, I joined the iCAPTURE Centre, St. Paul's Hospital as a summer student under the supervision of Scott Tebbutt. There I started work in a project of genotyping single nucleotide polymorphisms (SNPs) using Arrayed Primer Extension technology. Gradually this project became an essential part of my PhD thesis, and collaboration between my supervisors William J. Welch and Ruben H. Zamar in the Statistics Department and Scott Tebbutt in the iCAPTURE Centre has formed. At present I am a member of Dr. Tebbutt's lab at the iCAPTURE center and the main focus of my research has been to develop automated SNP-genotype calling algorithms for the APEX-based microarray genotyping platform. In this chapter, I will first give a brief description of SNPs and their relevance in Biomedical research. I will also describe various genotyping technologies as well as the proposed genotyping algorithm and its significance with respect to the current genotyping platform.

1.1 Single Nucleotide Polymorphisms and Their Relevance in Biomedical Research

Single Nucleotide Polymorphisms or SNPs (pronounced as "snips") are DNA sequence variations, occurring when a single nucleotide:— adenine (A), thymine (T), cytosine (C) or guanine (G) — in the genome is altered. An example of a SNP is a possible change in the nucleotide sequence aagcCta to aagcTta. Here the fifth letter "C" is replaced with "T". The two possible bases for a biallelic SNP are called SNP alleles. For a variation to be considered a SNP, it must occurs in at least 1% of the population. SNPs, which make up about 90% of all human genetic variations, occur every 100 to 300 bases along the 3-billion-base long human genome. Two of every three SNPs involve the replacement of cytosine (C) with thymine (T). The International HapMap Consortium has already reported discovery of approximately 10 million SNPs
(http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp).

More than 99% of the human DNA sequence is the same across the population. However, variations in the DNA sequence are responsible for responsiveness of humans towards disease; environmental interaction through bacteria, viruses, toxins, and chemicals; drugs and other therapies (Janssens et al., 2004). This makes SNPs of great importance in Biomedical research (Fan et al., 2006).

Many common diseases in humans like cancer, diabetes, vascular diseases, and various kinds of allergy are not caused only by a genetic variation within a single gene, but are influenced by complex interactions among multiple genes and various environmental factors (Yang et al., 2003). The main purpose of any genetic study is to investigate the potential of an individual to develop a disease based on different genetic variations (Candidate gene association studies: (Risch and Merikangas, 1996). For this purpose, it is of utmost importance to develop proper methodology to investigate the genetic variations among different samples (patients).

1.2 Genotyping SNPs and Various Microarray Platforms

The exact determination of the base sequence at a specific SNP site is called genotyping. A number of medium to high-throughput genotyping methods have been developed. Among these various techniques the most popular and oldest is TaqMan (Livak et al., 1995), which was designed to be optimal for genotyping a large number of individuals for a single SNP. But in the context of personalized medicine (use of genomic information to improve the diagnosis of disease, as well as the prevention and treatment of disease) one needs a system capable of genotying multiple SNPs simultaneously for a single individual.

Such a system can be achieved through a device known as a "genotyping microarray". Through this mechanism, one can display hundreds, or even thousands of specific oligonucleotide probes, precisely located on a small glass slide. These array-based technologies offer both an economic and patient specific application allowing the simultaneous genotyping of multiple SNPs. There are a number of microarray genotyping protocols with leading technologies including Affymetrix GeneChips (*Kennedy et al.*, 2003) and Illumina's BeadArray system (Oliphant et al., 2002), (Fan et al., 2006); these two technologies are designed to analyze thousands if not hundreds of thousands of SNPs simultaneously. Compared to these two systems, Arrayed primer extension (APEX), commonly known as mini-sequencing, is

an evolving technology, potentially suitable for rapid genetic diagnostics in clinical settings for its fast on-chip chemistry reaction. In this platform two- or four-coloured channel data for signal intensities are analyzed (Pastinen et al., 1997), (Tebbutt et al., 2004). We will give very brief descriptions of these three main microarray genotyping platforms in the following subsections. Notice that each array refers to a single individual (sample) and will be used interchangeably in the entire thesis.

1.2.1 Affymetrix GeneChips

For the widely used Affymetrix GeneChip system, a system based on the discriminatory power of nucleic acid hybridization to generate the genotyping signals, sophisticated autocalling algorithms have been developed. We will give a brief review of the algorithms published over the last decade by Affymetrix research group in Section 1.3. The Affymetrix GeneChips platform uses multiple sets of short oligonucleotide probe quartets for each known SNP, which have been combined through various statistical methodologies (both clustering and classification approaches have been applied) to generate reliable and accurate genotype calls. Each quartet consists of a perfect match (PM) cell and a mismatch (MM) cell with 25-mer probe, corresponding to both alleles (generically known as allele X and allele Y) for each SNP, which consequently generates the basic unit of probe quartet with four different probes: PMX, MMX, PMY and MMY for allele-specific hybridization. There are multiple quartets corresponding to different strands (both sense and antisense strands) and shifts (seven in total) surrounding the polymorphic site. Seven probe quartets per strand give 56 probe cells per SNP (Liu et al., 2003).

1.2.2 Illumina BeadArray

The Illumina BeadArray genotyping platform is based on the hybridization of a dual-purpose oligonucleotide probe (carrying a tag sequence as well as an SNP-specific sequence) to a complementary probe on an array (www.illumina.com). Each array in the Illumina HumanMap 550K SNP microarray consists of lateral strips of 55000 different beadtypes. Each beadtype assays a single SNP and is represented by 20 beads on average. On average 20 allele measurements per SNPs are obtained through single base extension biochemistry reaction on the locus-specific 50-mer probes associated with the nucleotide sequence directly adjacent to the SNP. Thus, similar to Affymetrix multi probe feature set, Illumina has multiple beads (20 on

average for each SNP in each array) corresponding to two possible SNP alleles, generating dual coloured signal intensities for each allele, visualized through labelled ddNTP corresponding to the complement of the assayed SNP (Steemers et al., 2007). Using the BeadArray genotyping assay, the automatic genotype calls are obtained through a proprietary genotype calling software: GenCall. Unfortunately, to our knowledge, the exact details of this genotype calling algorithm are not available in the public domain. However, (Teo et al., 2007) published an algorithm for the Illumina BeadArray platform. Brief details on the algorithm are given in the Section 1.3.

1.2.3 Mini-sequencing reaction on Microarray

Arrayed primer extension (APEX), commonly known as mini-sequencing is based on the allelic discriminating primer extension reaction in multiplex, separating the SNP alleles by pre-arraying the probes on a solid support (glass slide). In this method, the identity of the base incorporated is provided by the dye-labeled fluorescent terminators and the identity of the SNP assay is provided by the specifically designed probe (Pastinen et al., 1997), (Tebbutt et al., 2004). Our robust mixture model based genotype classification algorithm has been motivated by a complex data set generated in Dr. Tebbutt's laboratory using this APEX technology. Details of this APEX platform are described in Chapter 2. In order to discuss our genotyping model in the context of other genotype calling algorithms, we now provide a very brief description of our set of explanatory variables.

We have multiple probe sets corresponding to two different chemistries (classical APEX probes and allele-specific APEX probes) for both sense and antisense strands. Each probe has multiple replicates, which are randomly allocated on the microarray chip, each generating four coloured allele specific signals. In summary we have a set of four separate probe chemistries which produces four pair of variables that measure the intensity of the two possible SNP alleles (Podder et al., 2006).

1.3 Available Genotype Calling Algorithms

For any genotyping platform, the main idea has been to design an appropriate mapping between the signal intensities of the several predictor sets (for Affymetrix they are probe quartets; for Illumina BeadArray they are unique beadtypes; and for APEX mini-sequencing assays they are four different probe chemistries) and the three possible genotype classes (generically defined as XX, YY and XY) for all the biallelic SNPs analyzed in the array.

I will briefly describe several Affymetrix algorithms and also an algorithm based on Illumina BeadArray.

The genotype mapping can be designed through two basic approaches: a standard single model using all probes together to give the final genotype class; and a combination of multiple models each exploring a single probe.

For the standard models, information from multiple probes are combined to produce a single set of explanatory variables and a single prediction model. Whereas, in the multiple model set up, in a first stage each probe set generates a separate set of explanatory variables and probe-based genotyping model. Next, the first stage genotyping models are assembled in various effective manners to give a final genotype call.

1.3.1 Genotyping using a Single Model

A single model-based approach has been successfully implemented in the clustering-based MPAM algorithm (Liu et al., 2003) for Affymetrix first-generation (10K) SNP GeneChip microarrays. Here the probe-level aggregated predictor set, a point on a unit square, is formed based on the relative allele signal (RAS). Based on this predictor set, a clustering algorithm using modified partitioning around medoids has been used to cluster samples (arrays) into three possible genotype classes for each SNP. A large set of samples has been interrogated to label the appropriate genotype clusters. More recently another single model-based approach for the Affymetrix GeneChips has been proposed by Rabbee and Speed (2006). They designed a classification-based RLMM (robust linear model with Mahalanobis distance) algorithm, which used the prior knowledge from a large number of publicly available SNP calls from the HapMap project to build the genotype classification model. In this algorithm, for each sample (i) a two-dimensional predictor set is formed using robust linear model giving estimated probe effects for each SNP (j) corresponding to two possible SNP alleles. Decision regions for each genotype class (XX, XY and YY) are formed assuming bivariate Gaussian or Mahalanobis regions based on the two-dimensional predictor set using the training data. A new test data is assigned to a genotype class after computing the Mahalanobis distance of the point (two-dimensional predictor set), estimated from the test data, w.r.t. the center of the genotype class and using Mahalanobis distance as a minimum distance classifier.

Xiao et al. (2007) proposed a multi-array multi-SNP genotyping algorithm for the Affymetrix SNP microarrays using a normal M-cluster mixture model-based approach to cluster the SNP data. Here again a pooled estimate

of multiple probe quartet intensities for two possible SNP alleles has been used as a single predictor set. Another single classifier based approach has been developed by Hua et al. (2007) based on an expectation-maximization (EM) clustering algorithm (defined as SNiPer-HD) with parameters estimated using a large training set assuming again an M-cluster based Gaussian mixture model. Here a single predictor set consists of multiple relative-allele-signal (RAS) values corresponding to multiple probe quartets. Both of these mixture models have the following general form

$$L(\underline{\tau}, \underline{\pi}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i; \tau_g) \tag{1.1}$$

where, f_g is the predefined density function of the gth genotype class; τ_g denotes the corresponding parameter set (*e.g.*, location parameter, variance-covariance matrix, etc.); π_g is the proportion of the gth class and $\mathbf{x}_i$ is the predictor set corresponding to the ith array.

For the Illumina BeadArray platform, the information from multiple beads is pooled through proprietary normalization technique to provide a pair of variables (x_{ij}, y_{ij}: for the jth SNP in the ith array) corresponding to a single SNP for each array. The genotype calling algorithm using the data from Illumina BeadArray platform, published by Teo et al. (2007), is mainly based on a three component bivariate mixture model (see Equation 1.1). This approach is similar to the standard M-cluster mixture model. They assumed a t-distribution for their data (x, y), with three different sets of parameters (location, variance-covariance matrix and degrees of freedom) specific to three genotype classes. These parameters are estimated by applying an EM algorithm using a large training set.

1.3.2 Genotyping Combining Multiple Models

Assembling of individual probe level models has been first proposed by Di et al. (2005) in their dynamic model (DM) genotyping algorithm for 100K and 500K Affymetrix arrays. DM is a single sample based algorithm, which has been implemented through pixel intensities corresponding to a single SNP. Likelihood functions corresponding to four possible genotype states (XX, YY, XY and Non-call) have been estimated for each probe separately. Accordingly, individual probe level score functions are defined for n probe quartets and later combined through a non-parametric Wilcoxon signed rank test. In this way the majority of good quality probes are chosen to give the genotype class with smallest p-value, obtained from the above mentioned

test. This p-value is also used as a confidence measure. An improvement of the DM algorithm has been proposed by Nicolae et al. (2006) in their empirical likelihood based genotype calling model. For each SNP, two sufficient statistics corresponding to two possible alleles are formed for each probe quartet giving the basic predictor set. Three separate distributions for each predictor set are estimated using empirical likelihood functions. Genotypes are called based on the Bayesian posterior probability, which is calculated using a weighted likelihood of the previously defined individual empirical likelihood functions. The weights are predetermined using a reliability score based on the training set.

1.4 Proposed Genotype Calling Algorithm

I have mentioned in the previous section that there are two possible ways for modeling the multi probe level signal intensities to predict the genotype class: single model approach and multiple model approach. Based on exploratory data analysis of the APEX microarray genotyping data (obtained from Dr. Tebbutt's laboratory), we found that combining information from all available probes at the initial stage and then fitting a single prediction model gives a higher error rate compared with the proposed two-stage genotype classification model. This phenomenon is justifiable since functionality of different probe chemistries is highly sequence specific, *i.e.*, varying from sample to sample and from SNP to SNP. Even if we design a different model for each SNP, still probe chemistry variation with respect to the arrays might give erroneous results.

We believe that the best way for designing a genotyping mapping would be a two-stage modeling where each model in the first stage captures one individual probe chemistry. These models are then assembled in a second stage in such a way that the actual strength of each probe chemistry is reflected in the final genotype calling through an objective confidence measure. Initially, two independent data sets of different sizes with the same set of SNPs have been provided by Dr. Tebbutt to build an appropriate genotyping model for the APEX platform. We will call them the Coriell and the SIRS data sets. Details on these data sets are given in Chapter 2. We can then treat one of them as the training set and the other as the test set.

We began our journey by proposing a supervised learning algorithm which treats each probe set as an individual classifier using simple linear discriminant scores. The four prediction models are dynamically combined. We

take a weighted average of the four posterior probabilities provided by the four individual LDA (linear discriminant analysis) classifiers. The weights are sample/SNP-specific and defined so that the best working probe set gets maximum weight. Whereas, the only other genotyping algorithm which involves weighted likelihood of the available Affymetrix probe data, Nicolae et al. (2006) defines SNP-specific weights based on the training set. These weights remain constant for all the test samples. If some probe does not work well for a given sample this would not be reflected in the genotype call for that test sample. Details of our weighting scheme are described in Chapter 2 and Chapter 4.

We found that it is convenient but not sufficient to use a genotyping model with dynamic weights. Each model built on individual probe chemistry should also downweight outliers which occur in both the training and the test set. To deal with the outliers in the training set we can replace the estimates of the underlying model parameters by their robust counterparts. To deal with the outliers in the testing set we propose a class-specific mixture model for each base classifier, in which one part models the signal from good working probes (good signal distribution) and another part models possible outliers, *i.e.* generated by poorly working probes. We notice that our mixture model is totally different from the mixture model defined in Equation 1.1 (proposed by Xiao et al. (2007), Hua et al. (2007) and Teo et al. (2007)). Our mixture model is defined as follows

$$P_j(\mathbf{x}_j|g) = (1-\alpha) f_g(\mathbf{x}_j;\tau_g) + \alpha h(\tau) \tag{1.2}$$

where f_g is the predefined density function of the gth genotype class; τ_g denotes the corresponding parameter set (*e.g.*, location parameter, variance-covariance matrix, etc.); $\mathbf{x}_j$ is the jth predictor set and α is a user defined constant which allows α fraction mixing of the outlier distribution $h(\tau)$ with the actual distribution f_g. Our proposed mixture model is also different from the classical contamination model, since for this APEX based genotyping platform each predictor set is contaminated independently from the other sets. Therefore, some of the four sets could be outliers, while the others are perfectly good data. Thus, class-specific mixture models for the individual predictor sets makes sense.

After obtaining a posterior probability for each base classifier using the above model (defined in Equation 1.2), a weighted average of the four basic posterior probabilities gives a confidence measure for all possible genotype classes. Then a new test sample is assigned to a genotype class with the maximum confidence measure. The weights are defined dynamically and

depend on the actual strength of each individual probe chemistry. This modeling approach is described in Chapter 4.

1.5 Linear Discriminant Analysis

In this section, I will give a brief description of the linear discriminant analysis (LDA: Fisher (1936) and Hastie et al. (2001)). LDA, introduced by Fisher (1936), is one of the first and simplest statistical classification methods. I will describe LDA with two classes, where Fisher's discriminant function (DF) is a single linear combination of d explanatory variables $\mathbf{X}$ and the coefficients are estimated optimally to give the maximum separation between the classes for the training set. There are several ways of deriving the same DF and here I will describe both methods: maximizing the variance between classes relative to within classes and the likelihood ratio.

1.5.1 Maximizing the Variance-ratio

Let $x_1, \ldots, x_d$ be the d-dimensional observed predictor set and DF is a linear combination of the variables $X_1, \ldots, X_d$ defined as

$$D = \sum_{i=1}^{d} w_i X_i \tag{1.3}$$

We assume that the weights $\mathbf{w} = (w_1, \ldots, w_d)^T$ are estimated in such a way that the two classes get maximum separation w.r.t their location parameters and the population covariance matrix for $X_1, \ldots, X_d$ is the same for both classes. Then the conditional distribution of $X_1, \ldots, X_d$ given class k is

$$X_1, \ldots, X_d | k \sim (\mu^{(k)}, \Sigma) \qquad (k = 0,\ 1) \tag{1.4}$$

where $\mu^{(k)}$ is the d-dimensional population mean for class k, and Σ is the common $d \times d$ population covariance matrix. Now the distribution of D given class k can be defined through the mean(D) and Var(D):

$$\mathrm{E}(D|\text{class } k) = \mathrm{E}(\mathbf{w}^T\mathbf{X}|\text{class } k) = \mathbf{w}^T\mu^{(k)} \tag{1.5}$$

and

$$\mathrm{Var}(D) = \mathrm{Var}(\mathbf{w}^T\mathbf{X}) = \mathbf{w}^T\Sigma\mathbf{w} \tag{1.6}$$

Now a good DF should have weights which would maximize the following t-like ratio for comparing the two means of DF corresponding to two classes.

$$t = \frac{\mathbf{w}^T(\mu^{(0)} - \mu^{(1)})}{\sqrt{(\frac{1}{n_0} + \frac{1}{n_1})\mathbf{w}^T\Sigma\mathbf{w}}} \tag{1.7}$$

An estimate of the above quantity would be

$$\hat{t} = \frac{\mathbf{w}^T(\bar{\mathbf{x}}^{(0)} - \bar{\mathbf{x}}^{(1)})}{\sqrt{(\frac{1}{n_0} + \frac{1}{n_1})\mathbf{w}^T\mathbf{S}\mathbf{w}}} \tag{1.8}$$

Where S is the pooled estimate of Σ based on the individual estimates $\mathbf{S}^{(0)}$ and $\mathbf{S}^{(1)}$

$$\mathbf{S} = \frac{(n_0 - 1)\mathbf{S}^{(0)} + (n_1 - 1)\mathbf{S}^{(1)}}{n_0 + n_1 - 2}$$

Maximizing $\hat{t}$ in Equation 1.8 is equivalent to maximize the $\hat{t}^2$, which is a ratio of between-class to within-class variation (apart from the constant part in Equation 1.8) and derived as follows:

$$\hat{t}^2 = \frac{\mathbf{w}^T(\bar{\mathbf{x}}^{(0)} - \bar{\mathbf{x}}^{(1)})(\bar{\mathbf{x}}^{(0)} - \bar{\mathbf{x}}^{(1)})^T\mathbf{w}}{\mathbf{w}^T\mathbf{S}\mathbf{w}} \tag{1.9}$$

The optimizing $\mathbf{w}$ is then

$$\mathbf{w} \propto \mathbf{S}^{-1}(\bar{\mathbf{x}}^{(0)} - \bar{\mathbf{x}}^{(1)}) \tag{1.10}$$

The solution is arbitrary up to a constant of proportionality. Finally Fisher's DF is given by

$$\mathbf{w}^T x = (\bar{\mathbf{x}}^{(0)} - \bar{\mathbf{x}}^{(1)})^T\mathbf{S}^{-1}\mathbf{x}, \tag{1.11}$$

or any multiple thereof. Note that no assumption regarding the exact distribution of $X_1, \ldots, X_d$ has been made beyond means and covariances. Now for more than two classes (suppose there are c classes), the idea of Fisher's LDA is to find the directions that maximize between-class variability relative to within-class variability, *i.e.*, to maximize the following quantity

$$\frac{\mathbf{w}^T\mathbf{B}\mathbf{w}}{\mathbf{w}^T\mathbf{S}\mathbf{w}}$$

with respect to w. Where

$$\mathbf{B} = \frac{1}{c}\sum_{k=0}^{c-1}(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})^T$$

is the between-class variance with $\bar{\mathbf{x}} = \frac{1}{n}\sum_{k=0}^{c-1} n_k \bar{\mathbf{x}}^{(k)}$ as the overall mean vector, and the within-class variance is given by

$$\mathbf{S} = \frac{\sum_{k=0}^{c-1}(n_k - 1)\mathbf{S}^{(k)}}{n - c}$$

The optimizing w is the eigenvector of $\mathbf{S}^{-1}\mathbf{B}$ with the largest eigenvalues. The same DF can be derived through the Likelihood Ratio method (with exact distributional assumptions) and Regression model.

1.5.2 Likelihood Ratio Optimization

Suppose we want to test the hypotheses

$$H_0: \text{ Object comes from Class 0}$$

versus

$$H_1: \text{ Object comes from Class 1}$$

with the assumption that $X_1, \ldots, X_d$ have a multivariate normal (MN) distribution. Then the assumption in (1.4) become

$$X_1, \ldots, X_d | k \sim \text{MN}(\mu^{(k)}, \Sigma) \qquad (k = 0,\ 1) \tag{1.12}$$

The underlying density function would be

$$L(\mathbf{x}; \mu, \Sigma) = \frac{1}{(1\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp[-\frac{1}{2}(\mathbf{x} - \mu)^T\Sigma^{-1}(\mathbf{x} - \mu)] \tag{1.13}$$

Using the likelihood ratio of the two distributions as the scoring function for discrimination:

$$\begin{aligned}\frac{L(\mathbf{x}|\text{ Class 0})}{L(\mathbf{x}|\text{ Class 1})} &= \frac{\exp[-\frac{1}{2}(\mathbf{x} - \mu^{(0)})^T\Sigma^{-1}(\mathbf{x} - \mu^{(0)})]}{\exp[-\frac{1}{2}(\mathbf{x} - \mu^{(1)})^T\Sigma^{-1}(\mathbf{x} - \mu^{(1)})]} \\ &= \exp[(\mu^{(0)} - \mu^{(1)})^T\Sigma^{-1}x]\exp(\frac{1}{2}\mu^{(0)^T}\Sigma^{-1}\mu^{(0)} - \frac{1}{2}\mu^{(1)^T}\Sigma^{-1}\mu^{(1)})\end{aligned} \tag{1.14}$$

Only the first factor of the above scoring function involves $\mathbf{x}$ and after taking the natural logarithm it leads to the following classification criterion

$$(\mu^{(0)} - \mu^{(1)})^T\Sigma^{-1}\mathbf{x}$$

After replacing $\mu^{(k)}$ with $\bar{\mathbf{x}}^{(k)}$ and Σ by $\mathbf{S}$ we obtain the same Fisher's DF again

$$(\bar{\mathbf{x}}^{(0)} - \bar{\mathbf{x}}^{(1)})^T \mathbf{S}^{-1} \mathbf{x}$$

Details on LDA can be found in Hastie et al. (2001).

Organization of the Thesis

This is a manuscript based PhD thesis as allowed by UBC guidelines. The main objective of this thesis is to build a classification algorithm for SNP genotype data obtained through APEX based microarray genotyping platform. Chapter 2 to Chapter 4 are three independent manuscripts which are self contained and each has its own reference list. Work under Chapter 2 and Chapter 3 has been published in peer-reviewed journals and Chapter 4 will be submitted to a refereed journal and is being formatted accordingly. Chapter 5 concludes the thesis with a summary and discussion of possible future research directions.

Bibliography

Di, X., H. Matsuzaki, T. Webster, E. Hubbell, G. Liu, S. Dong, D. Bartell, J. Huang, R. Chiles, G. Yang, et al. (2005). Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics 21*(9), 1958–1963.

Fan, J., K. Gunderson, M. Bibikova, J. Yeakley, J. Chen, E. Wickham Garcia, L. Lebruska, M. Laurent, R. Shen, and D. Barker (2006). Illumina universal bead arrays. *Methods Enzymol 410*, 57–73.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*(2), 179–188.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hua, J., D. Craig, M. Brun, J. Webster, V. Zismann, W. Tembe, K. Joshipura, M. Huentelman, E. Dougherty, and D. Stephan (2007). SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. *Bioinformatics 23*(1), 57.

Janssens, A., M. Pardo, E. Steyerberg, and C. van Duijn (2004). Revisiting the Clinical Validity of Multiplex Genetic Testing in Complex Diseases. *The American Journal of Human Genetics 74*(3), 585–588.

Liu, W., X. Di, G. Yang, H. Matsuzaki, J. Huang, R. Mei, T. Ryder, T. Webster, S. Dong, G. Liu, et al. (2003). Algorithms for large-scale genotyping microarrays. *Bioinformatics 19*(18), 2397–2403.

Livak, K., S. Flood, J. Marmaro, W. Giusti, and K. Deetz (1995). Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *Genome Research 4*(6), 357.

Nicolae, D., X. Wu, K. Miyake, and N. Cox (2006). GEL: a novel genotype calling algorithm using empirical likelihood. *Bioinformatics 22*(16), 1942.

Oliphant, A., D. Barker, J. Stuelpnagel, and M. Chee (2002). BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques 32*, S56–S61.

Pastinen, T., A. Kurg, A. Metspalu, L. Peltonen, and A. Syvanen (1997). Minisequencing: A Specific Tool for DNA Analysis and Diagnostics on Oligonucleotide Arrays. *Genome Research 7*(6), 606.

Podder, M., W. Welch, R. Zamar, and S. Tebbutt (2006). Dynamic variable selection in SNP genotype autocalling from APEX microarray data. *BMC Bioinformatics 7*(1), 521.

Rabbee, N. and T. Speed (2006). A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics 22*(1), 7–12.

Risch, N. and K. Merikangas (1996). The Future of Genetic Studies of Complex Human Diseases. *Science 273*(5281), 1516.

Steemers, F., K. Gunderson, I. Illumina, and C. San Diego (2007). Whole genome genotyping technologies on the BeadArray platform. *Biotechnol J 2*(1), 41–49.

Tebbutt, S., J. He, K. Burkett, J. Ruan, I. Opushnyev, B. Tripp, J. Zeznik, C. Abara, C. Nelson, and K. Walley (2004). Microarray genotyping resource to determine population stratification in genetic association studies of complex disease. *Biotechniques 37*(6), 977–85.

Teo, Y., M. Inouye, K. Small, R. Gwilliam, P. Deloukas, D. Kwiatkowski, and T. Clark (2007). A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics 23*(20), 2741.

Xiao, Y., M. Segal, Y. Yang, and R. Yeh (2007). A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics 23*(12), 1459.

Yang, Q., M. Khoury, L. Botto, J. Friedman, and W. Flanders (2003). Improving the Prediction of Complex Diseases by Testing for Multiple Disease-Susceptibility Genes. *The American Journal of Human Genetics 72*(3), 636–649.

Chapter 2

Dynamic Variable Selection

2.1 Background

2.1.1 Genotyping SNPs

Determination of the alleles at a specific single nucleotide polymorphism (SNP) site is called genotyping. An optimal genotyping technology should be capable of genotyping any number of SNPs for a large number of individuals satisfying the following criteria: 1. easy and quick development of an assay from the sequence information; 2. over-all low cost; 3. the data analysis must be simple, transparent, fully-automated and robustly give accurate genotype-calls for all kinds of samples; and 4. the study design must be flexible and scalable in all respects (e.g., number of SNPs investigated). Automated genotype calling is an essential part of such a system. A number of medium to high-throughput genotyping methods have been developed. Among these various techniques, TaqMan (Livak et al., 1995) was designed optimally to give genotypes of large numbers of individuals for one SNP at a time. But from a clinically relevant, personalized medicine point of view, we require a system which can genotype multiple SNPs simultaneously for any single patient sample.

Such a system can be achieved through a device known as a genotyping microarray. Through this mechanism, one can display thousands of specific oligonucleotide probes, precisely located on a small glass slide. These array-based technologies offer both economic and patient specific applications allowing the genotyping of multiple SNPs simultaneously. There are a number of microarray genotyping protocols, including Affymetrix GeneChips$^{(R)}$ (Kennedy et al., 2003) and Illumina's BeadArray^{TM} system (Oliphant et al., 2002). For the widely used Affymetrix GeneChip system, a system based on

A version of this chapter has been published as: Mohua Podder, William J. Welch, Ruben H. Zamar and Scott J. Tebbutt: Dynamic variable selection in SNP genotype autocalling from APEX microarray data. *BMC Bioinformatics* 2006, **7**:521.

the discriminatory power of nucleic acid hybridization to generate the genotyping signals, sophisticated autocalling algorithms have been developed (Di et al., 2005). Over the last five to six years Affymetrix has developed and tested a series of algorithms using their platform. The Affymetrix GeneChip is suitable for very large scale genotyping, e.g., 10,000 or more SNPs at a time, but is expensive for medium to small scale genotyping (e.g., 100 to 200 SNPs). The Illumina BeadArray genotyping platform provides a powerful combination of high-throughput and accuracy with low cost per SNP analysis. Based on the GoldenGate^{TM} genotyping assay, Illumina designed a genotype calling algorithm using a Bayesian model, taking the ratio of two single colored intensity signals corresponding to two possible SNP alleles, to give the genotype for a single SNP (Shen et al., 2005). The automatic calling of genotypes is performed by proprietary software, GenCall, which is based on a custom-designed clustering algorithm (Shen et al., 2005). To our knowledge, exact details of the algorithm are not available in the public domain.

Compared to these systems, our laboratory has developed a robust and redundant chemistry platform using the technology of single base extension which produces multiple signals from multiple probes [APEX and allele-specific APEX (ASO) probes for both DNA strands] corresponding to a single SNP (Tebbutt et al., 2004). To our knowledge, APEX is the only chemistry in which the on-chip assay can be performed in 20 minutes, making APEX potentially suitable for rapid genetic diagnostics in clinical settings: the Affymetrix assay takes several hours for hybridization on the chip, and Illumina's assays also takes longer compared to APEX.

Commercial software called Genorama (www.asperbio.com) can detect all the four colors of fluorescence emitted from the dyes used in an APEX experiment, and then automatically call the base(s) corresponding to a specific probe spot. The problem with this system is that the underlying scoring algorithm treats all probes equally and thus requires considerable inspection of the original array data to produce the final genotype call (Kamiński et al., 2005), (Kurg et al., 2000). Using the Genorama base-calling data for both APEX and AS-APEX probes, Gemignani et al. (2002) developed a simple matrix-score based algorithm and made the calls corresponding to the most likely genotype, but with considerable manual inspection.

2.1.2 Current Genotype Calling System: SNP Chart

SNP Chart is a Java based visualization tool, developed by our research group (Tebbutt et al., 2005). In this integrated platform, spot intensity

data from different and/or replicate probes (randomly scattered across the microarray slide) that interrogate the same SNP are imported, together with a multi-channel TIFF image of the original array experiment. This system is capable of calling any SNP genotype with the help of individual manual data inspection. The main problem with this genotype-calling system is that it is time-consuming and exposed to user subjectivity bias.

Examples of SNP Charts are shown in Figure 2.1. Here, template DNA from three Coriell samples with three possible genotypes (CC, CT and TT) and one negative control (NN) are shown in four different charts. Each chart shows four-channel fluorescent intensity data (A, C, G, and T) on the vertical axes, from 12 rs1106577-specific array spots (duplicate spots for six different probes). On the horizontal axes, 12 probe-names corresponding to 12 spots are given sequentially. The first and second spots from the left ("LEFT C/T") refer to the left-hand APEX probe that will give either a single C (green) signal (for homozygous CC genotypes) or a T (blue) signal (for homozygous TT genotypes) or a mixture of C and T (heterozygous CT). The third and fourth spots from the left ("RIGHT G/A") refer to the right-hand APEX probe that interrogates the DNA strand nucleotide complementary to that of the left-hand APEX probe, thus giving a single G (red) signal (for CC), a single A (yellow) signal (for TT), or a mixed G and A signal (for CT). From the left, spots 5 to 12, inclusive, represent allele-specific APEX probes in which a base-specific fluorescence signifies the presence of the allele. Among them, spots 5 to 8 refer to the "_1" probes corresponding to the first allele (C in the case of rs1106577) and spots 9 to 12 refer to the "_2" probes corresponding to the second allele (T). The details of the APEX chemistry are explained by Tebbutt et al. (2004, Section "APEX Microarrays"). The redundancy and consistency of the data across different probes give high confidence in the assigned genotypes.

2.1.3 Data Composition

We built our genotyping model based on the training set of 32 Coriell DNA samples (http://coriell.umdnj.edu/) and 3 negative PCR controls (Tebbutt et al., 2004), (Tebbutt et al., 2005). Each sample comes from a single microarray experiment, conducted on a small glass slide, and contains information on all the SNPs under study. Our laboratory has developed a robust microarray platform for each sample patient, generating multiple signals for approximately one hundred SNPs using two kinds of probes, namely, classical APEX probes and allele specific APEX (ASO) probes (Tebbutt et al., 2004). There are six probes in total for each biallelic SNP and each probe

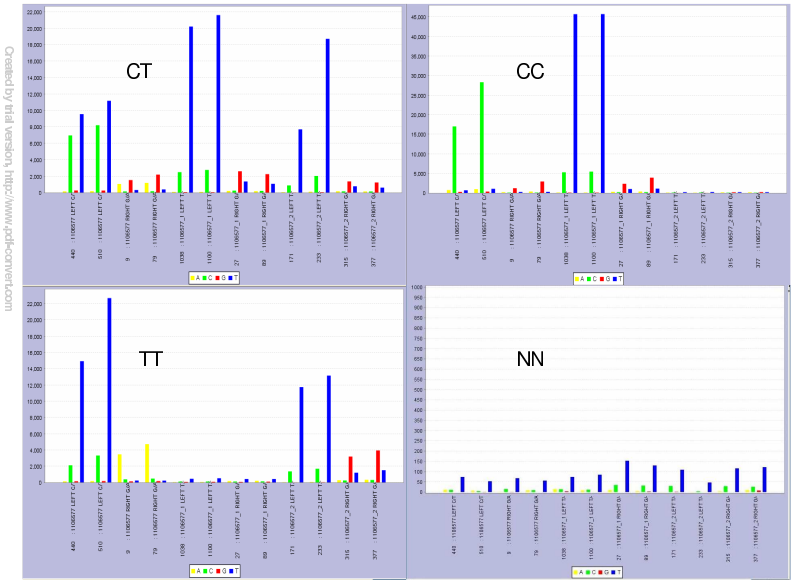

Figure 2.1: An Example of SNP Chart Application for the SNP rs1106577.

has two replicates which make twelve different spots for a single SNP on the microarray slide. All these spots are randomly scattered across the microarray slide but with known coordinates. Multiple sets of probes of these types along with their replicates make this genotyping platform unique and redundant. Each spot in the microarray slide produces signals from four different channels, corresponding to A, C, G and T. In our current genotyping method, we only considered the expected foreground signals and will consider all the background, non-expected signals for further development of genotyping model (see below).

An example of a data source for a single Coriell sample and a single SNP is given in Table 2.1. For the SNP rs1106577, the two possible alleles are C or T. Each row of this table represents a single spot. The first column is the spot ID; the second column is the probe name; the third column is the expected allele ID for the appropriate spot; and the last four columns are the signal intensity values for the four channels corresponding to each spot.

In the second column of Table 2.1, "APEX_LEFT" refers to the left-hand APEX probe on the sense strand, and "APEX_RIGHT" refers to the right-hand APEX probe on the anti-sense strand that interrogates the DNA strand nucleotide complementary to that of the left-hand APEX probe. For all

the APEX probes, the fluorescent signals come from the base position of the SNP allele. In contrast, for all the ASO probes, fluorescent signals come from the base adjacent (3') to the actual SNP site (Tebbutt et al., 2004). For the SNP rs1106577 considered in Table 2.1, the base 3' adjacent to the SNP allele is always T on the sense strand. The left probes, ASO_1LEFT and ASO_2LEFT, are designed to signal at this adjacent base, T, if the SNP site has the first allele (C here) and/or the second allele (T here), respectively. Similarly, SNP rs1106577 has G in the adjacent position 3' to the SNP side on the anti-sense strand. The right probes, ASO_1RIGHT and ASO_2RIGHT, signal at this adjacent base, G, again for C and/or T at the SNP site, respectively. It is merely the presence or absence of the signal that indicates the SNP allele. According to the probe structure, the signals corresponding to the expected alleles are highlighted. The data represented in Table 2.1 come from the DNA sample Coriell NA17102 and here the true genotype is CC (see the top-right CC-chart in Figure 2.1). According to the APEX chemistry, for the genotype CC the dominating signals from spots 1 and 2 should be C among the two expected channels C and T. Similarly the dominating signals from spots 3 and 4 should be G (complementary to C in the left-strand) among the two expected channels G and A. Rows 5–12 represent the ASO probes in which a base-specific fluorescence signifies the presence of the allele. Since the genotype is CC, all the expected signals corresponding to allele 1 (C) should dominate over the other channels, i.e., expected foreground (expected channel corresponding to all allele 2 probes) and background signals (Tebbutt et al., 2004). Note that for spots 11 and 12, the expected signal (G), corresponding to the presence of the T allele (which is absent in this particular case), is comparable to the background signals. In Table 2.1, all the signals which are not highlighted in bold are considered as background signals, often due to the spectral overlap between dyes, and/or a general background.

In fact, this is a very good source of data, as all the signals corresponding to allele 2 (T in this case) are comparable to the level of background signals. Now suppose the true genotype is TT, then we should expect dominating signals only from the expected channels corresponding to all allele 2 probes. For a heterozygous CT genotype, we should expect dominating signals from all the expected channels corresponding to both allele 1 probes and allele 2 probes. These features of our redundant and robust platform can also be represented through our data visualization tool: SNP Chart (Tebbutt et al., 2005). In Figure 2.1, four SNP Charts corresponding to three different genotypes (CT, CC and TT) and a negative control (NN) are shown for the same SNP (rs1106577). In our study we use the 32 Coriell samples plus

Table 2.1: Data structure for SNP rs1106577 and DNA sample Coriell NA17102 (CC) (CC-chart in Figure 2.1)

Spot ID	*Probe ID*	*Expected allele*	A	C	G	T
Spot 1	APEX_LEFT	C and/or T	732	**17003**	258	**667**
Spot 2	APEX_LEFT	C and/or T	965	**28290**	348	**1046**
Spot 3	APEX_RIGHT	G and/or A	**190**	85	**1198**	233
Spot 4	APEX_RIGHT	G and/or A	**353**	104	**2923**	269
Spot 5	ASO_1LEFT	T	109	5284	80	**45700**
Spot 6	ASO_1LEFT	T	107	5456	83	**45713**
Spot 7	ASO_2LEFT	T	90	88	20	**182**
Spot 8	ASO_2LEFT	T	76	106	22	**222**
Spot 9	ASO_1RIGHT	G	288	182	**2346**	992
Spot 10	ASO_1RIGHT	G	369	209	**3908**	1098
Spot 11	ASO_2RIGHT	G	138	68	**166**	187
Spot 12	ASO_2RIGHT	G	151	68	**212**	193

three negative PCR controls for model building. These 35 samples will be called the Coriell training set. To test the performance of the calling algorithm we also have a completely independent set of 270 SIRS (systematic inflammatory response syndrome) DNA samples from the ICU of St. Paul's hospital, plus one test negative control sample. This set of 271 samples will be called the SIRS test data. Note that the SIRS data are not used in model building and come from a separate study, so they provide a very rigorous test. For the training data, there are 123 SNPs on the microarray slide, but only 96 were usable: (1) 15 SNPs had PCR chemistry failure and (2) 12 SNPs had one of the three possible genotypes missing among the training set.

2.1.4 Formation of Classifiers

Ideally, the genotype call could be solely based on just one of four sets of probes: (1) APEX_LEFT, (2) APEX_RIGHT, (3) ASO_1LEFT and ASO_2LEFT, and (4) ASO_1RIGHT and ASO_2RIGHT (see Table 2.1). Accordingly, we have developed four sets of classifiers, named APEX.L, APEX.R, ASO.L and ASO.R, based on the respective probe sets. Each classifier is based on two explanatory variables, generically denoted by X

and Y, measuring the signal intensities for the two candidate alleles in the SNP position. In Table 2.1, for example, X and Y corresponds to the C and T alleles, respectively.

Between them the four classifiers have four pairs of such explanatory variables: (APEX.XL, APEX.YL); (APEX.XR, APEX.YR); (ASO.XL, ASO.YL) and (ASO.XR, ASO.YR). They are derived from the signal intensities in rows 1–2, 3–4, 5–8, and 9–12, respectively, in data structures exemplified in Table 2.1. All these variables take the sum of the relevant signals. From the example data in Table 2.1, the values of the variables for the classifier APEX.L are APEX.XL $= 17,003 + 28,290 = 45,293$ and APEX.YL $= 667 + 1,046 = 1,713$, and so on, as summarized in Table 2.2. Our main objective is to automatically select from these four sets of variables those pairs which give "good" signals for genotype calling. Moreover, the variables and hence the classifier(s) used will be chosen dynamically, i.e., for a specific SNP and sample. In this paper we use Fisher's (Fisher, 1936) linear discriminant analysis (LDA) to build the classifiers, but the method of dynamic variable selection would apply to any linear or nonlinear classifier. Figure 2.2 and Figure 2.3 illustrate how dynamic variable selection exploits the redundancy in the chemistry. The figures are based on the 32 Coriell samples plus three negative PCR controls, where the true genotypes are known. We plot the X and Y signals for each of the four probe sets. Ideally, any pair of variables would form well separated clusters for the three possible genotypes, XX, XY and YY (plotted with different colors and symbols) [red, green, blue and black colored symbols respectively denote the classes YY, XY, XX and NN]. There is a fourth cluster corresponding to the negative controls (NN). Any reasonable classifier based on these variables should make correct calls under ideal conditions. Figure 2 shows an ideal SNP, where all four probe sets produce good separation of the three genotypes and the negative controls. In Figure 2.2, all the classifiers give

Table 2.2: Values of the explanatory variables for SNP rs1106577 and DNA sample Coriell NA17102

Classifier	*Variables used by classifier*		*Values*	
APEX.L	APEX.XL	APEX.YL	45,293	1,713
APEX.R	APEX.XR	APEX.YR	4,121	543
ASO.L	ASO.XL	ASO.YL	91,413	404
ASO.R	ASO.XR	ASO.YR	6,254	378

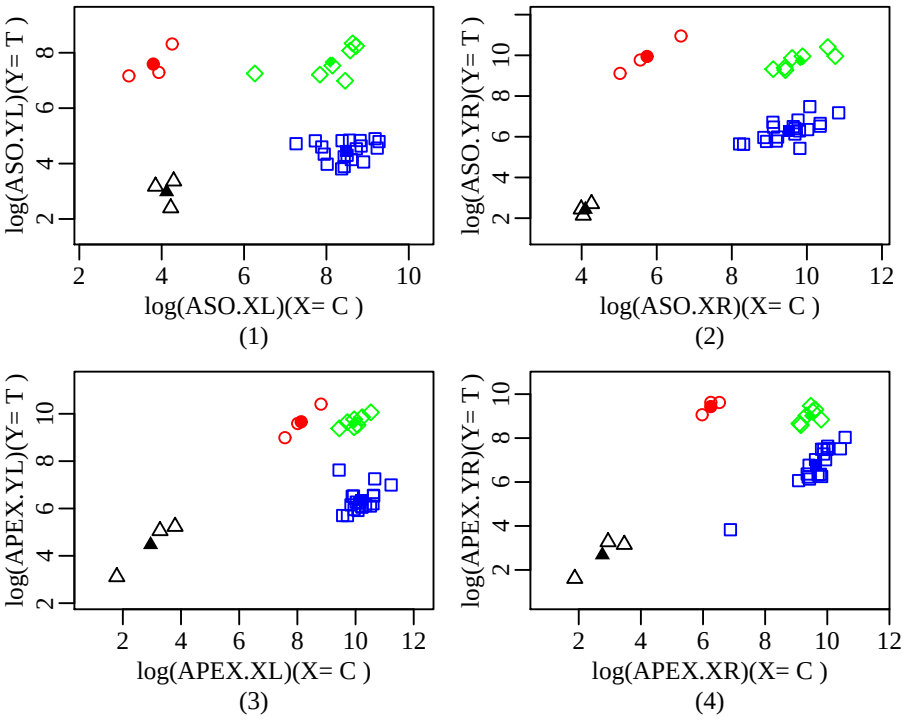

Figure 2.2: Example of a well-behaved SNP: rs1932819

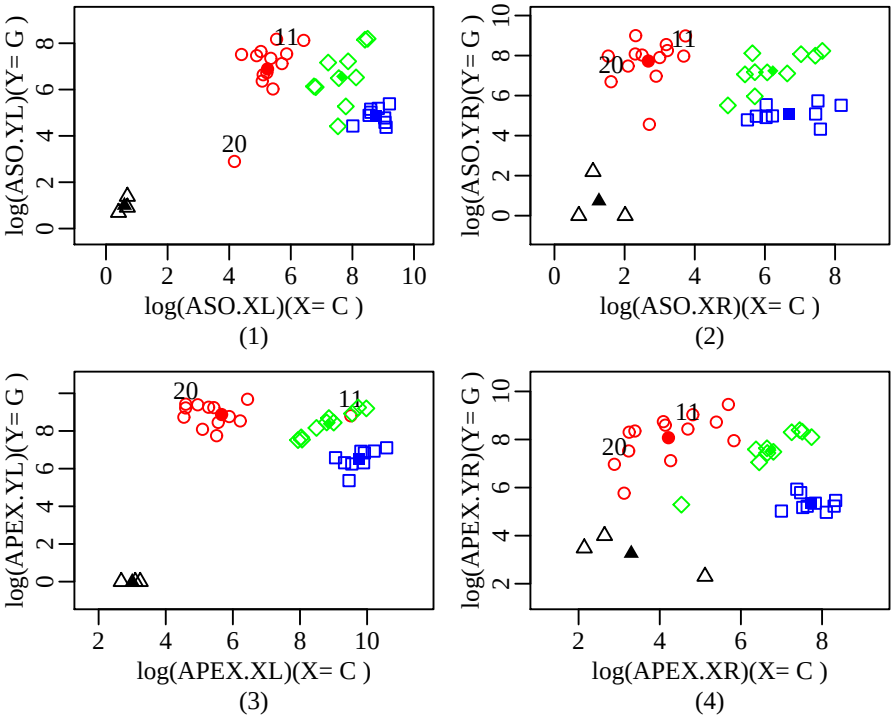

Figure 2.3: Example of a critical SNP: rs1003399.

three well separated clusters for the SNP rs1932819; whereas in Figure 2.3, sample 11 is correctly classified by both ASO probes and APEX.R probe but wrongly classified by APEX.L probe for the SNP: rs1003399, whereas for sample 20, APEX.L probe works the best.

Conversely, problems with the samples or the chemistry may lead to overlap in the four clusters, making calling difficult. In Figure 2.3 for SNP rs1003399, for example, sample 11 is a GG genotype which falls in the CG cluster for the left APEX probe set. Fortunately, the other three probe sets correctly place sample 11 in the GG cluster. So three out of the four probe sets work, and classifiers based on them would make the correct call for sample 11. For sample 20 (NA07341), however, the left APEX probe set works the best, placing the GG sample clearly in the GG data cluster. Thus, different probe sets may be effective for different samples, even for the same SNP. Our algorithm attempts to identify effective probe sets automatically, sample by sample, and it is in this sense that it chooses variables dynamically.

2.2 Results and Discussion

2.2.1 Dynamic-variable LDA Based Genotyping Model

For each SNP we build four separate LDA classification models; the models are based on the pairs of explanatory variables in Table 2.2 corresponding to the four probe sets. For this stage the training data are the 32 Coriell samples and the three negative PCR controls described under Data Composition. As test data to evaluate the calling performance we use the 271 SIRS test samples also described under Data Composition. Within each SNP, sample by sample the four classifiers are combined using the weighting algorithm described later in the Methods Section, to give one call for the particular test sample. The calls are checked for concordance with the validated genotypes in the SIRS data, leading to the results in the first row of Table 2.3. In 0.4% of samples, the called genotype is NN (non-call), hence the call rate of less than 100% in the table. As detailed under Methods, by changing the threshold for calling, a modest reduction in the call rate to 94.9% yields a 99.6% concordance rate. We also reverse the roles of the training and test data sets, leading to the second row of Table 2.3.

The results are stronger in terms of the number of SNPs called, call rate and concordance rate, because in this second analysis a much larger set of data is used for training the models.

Row 3 of Table 2.3 reports the results from applying the method of cross

Table 2.3: Results from Dynamic-variable LDA

Training set	*Test set*	*No. of SNPs*	*High call.rate*		*Lower call.rate*	
			Call.rate	*Accuracy*	*Call.rate*	*Accuracy*
Coriell	SIRS	96	99.6%	98.9%	94.9%	99.6%
SIRS	Coriell	102	99.9%	99.3%	95.6%	99.8%
Coriell	CV	96	100.0%	98.7%	94.2%	99.2%
SIRS	CV	102	99.9%	99.3%	96.0%	99.8%

validation (CV) (Hand et al., 2001) to the Coriell data set. Here, each sample is removed in turn from the data, and its genotype is predicted based on retraining the four classifiers using *only the remaining data*. The results are similar to those in row 1. For the SIRS data, row 4 reports analogous cross validation performance estimates, and there is very close agreement with row 2.

2.2.2 Simple LDA Based Genotyping Model

For comparison, for each SNP we use the training data to build a single LDA classification model using all eight variables available in Table 2.2. For each SNP, simple LDA applied to the training data assigns weights to the eight variables and these weights are constant for every test sample. Thus, this more standard modeling approach does not allocate weights dynamically. The same comment applies to MACGT from our research group (Walley et al., 2006), which also requires greater levels of manual inspection of the APEX data. In fact, a simple LDA is expected to work well if all the underlying variables are good (without contamination) so that simple LDA can assign optimal weights treating all variables simultaneously. However, in this genotype classification problem, we have seen that occasionally some probe fails thus introducing outliers in the system.

The results from simple linear discriminant analysis are given in Table 2.4. In row 1 the concordance rate for the SIRS test set is 97.3%, which might be considered good for other applications but for clinical purposes a much smaller concordance error is desirable. Modifying the calling threshold makes negligible difference to the concordance rate. Reversing the training and test data shows an even worse outcome: (1) again changing the threshold value does not control the call rate and (2) the concordance rate deteriorates dramatically. Therefore the performance is not competitive

against dynamic-variable LDA.

As shown in rows 3 and 4 of Table 2.4, the performance of simple linear discriminant analysis is better when measured by cross validation, particularly when predicting the SIRS data. It seems that the method is not robust to using samples from different sources for training and testing.

Table 2.4: Results from Simple LDA

Training set	*Test set*	*No. of SNPs*	*High call.rate*		*Lower call.rate*	
			Call.rate	*Accuracy*	*Call.rate*	*Accuracy*
Coriell	SIRS	96	99.4%	97.3%	98.1%	97.3%
SIRS	Coriell	102	99.5%	93.0%	99.5%	93.0%
Coriell	CV	96	99.8%	98.4%	99.7%	98.5%
SIRS	CV	102	99.4%	99.5%	98.9%	99.6%

We also analyze the SNP specific performance of the two models: dynamic-variable LDA and dimple LDA. For this we plot the misclassification rates corresponding to 96 SNPs using Coriell as a training set and predicting the genotypes of 270 SIRS samples (see Figure 2.4) for 100% call rate. Figure 2.4 shows that there are many points with higher misclassification rates with simple LDA model as compared to the dynamic-variable LDA model.

2.2.3 Discussion

We also tried classifiers based on different sets of variables. For example, we built an ASO classifier using the variables ASO.XL, ASO.YL, ASO.XR and ASO.YR and an APEX classifier using the variables APEX.XL, APEX.YL, APEX.XR and APEX.YR. The calls from the two classifiers were then combined using the dynamic variable methodology. Little improvement in concordance rate was found relative to eight-variable simple LDA. Similar results were obtained when combining left and right classifiers, based on the left variables (ASO.XL, ASO.YL, APEX.XL and APEX.YL) and the right variables (ASO.XR, ASO.YR, APEX.XR and APEX.YR), respectively.

2.3 Conclusions

We have developed a robust automated genotype calling method based on an ASO and APEX microarray platform. Multiple, qualitatively different probes provide redundancies in the event that a probe does not provide

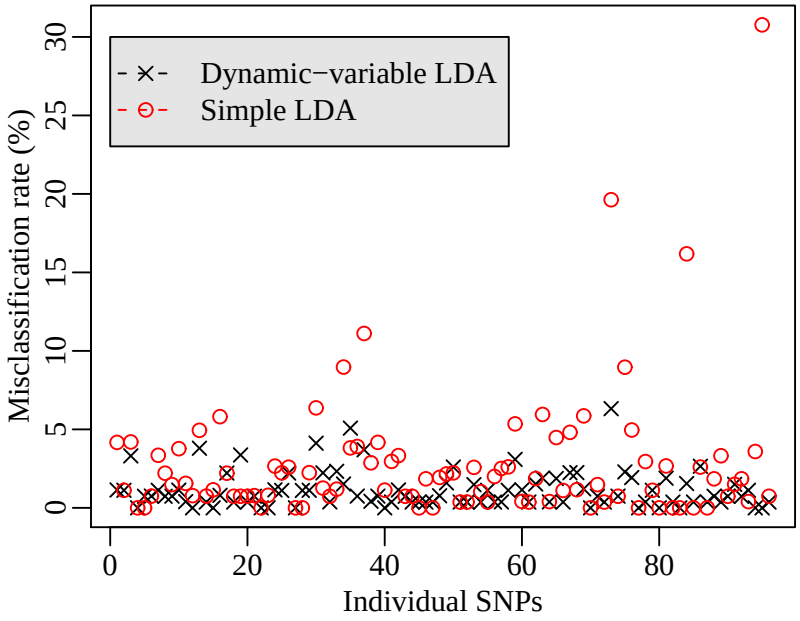

Figure 2.4: SNP specific performance of two models

a reliable signal. The dynamic-variable calling algorithm respects these redundancies, building up an overall call from classifiers based on subsets of variables, with more weight given to seemingly more reliable classifiers. The weights change from one test sample to another; it is in this sense that the method is dynamic. Standard methods of variable selection (also called feature extraction) as described by, for example, Hand, Mannila, and Smyth (Hand et al., 2001) or Hastie, Tibshirani, and Friedman (Hastie et al, 2001), would select or filter the variables and use the same set of reduced variables for every call. Such a strategy would be appropriate if the *same* probe sets are reliable from sample to sample.

For a call rate of approximately 95%, we were able to achieve a concordance rate of 99.6% in a large, independent test set of validated genotypes. The probe data for those samples/SNPs that are not automatically called would be manually inspected within SNP Chart; unlike 100% manual inspection, this does not impose an unreasonable time burden. The method of combining classifiers is not specific to linear discriminant analysis; other statistical classifiers could be used. Similarly, the method could be applied to other microarray platforms with complex redundancies.

2.4 Methods

2.4.1 LDA

Linear discriminant analysis (LDA), due to Fisher (Fisher, 1936), is one of the oldest methods of discrimination between classes or classification. It is described in virtually every text book that includes classification (e.g., Hastie, Tibshirani, and Friedman) (Hastie et al, 2001). LDA is applied to each SNP separately. It is assumed that the variables (probe signals) used to classify have a multivariate normal distribution, with a within-class covariance matrix that is common to all classes (the genotypes and a negative control class) but within-class mean vectors that vary from one class to another. These quantities are estimated from the Coriell training data. For any test sample, the values of the same variables lead to posterior probabilities for the various classes. The genotype called is the class with the highest posterior probability. The method also requires the prior probabilities of belonging to the various classes. We assume priors based on observed frequencies in the training data. This basic LDA methodology is common to all the strategies we use.

2.4.2 Simple LDA

In *Simple LDA* we train a single LDA genotyping model using the logarithms of all eight variables described in Table 2.2. Among the validated genotypes of the 32 Coriell samples, there are some cases where the exact genotype is unknown, denoted by NN (non-call). The three negative controls added to the Coriell data are also treated as NN as well. Thus, for each SNP there may be up to four classes present in the training data, corresponding to the three candidate genotypes and NN. Thus, LDA may call NN. The call rate is the proportion of calls that are not NN.

2.4.3 Dynamic-variable LDA

For each SNP we apply LDA to each pair of variables in Table 2.5. For example, the classifier ASO.L is based on the left ASO variables, log(ASO.XL) and log(ASO.YL). Here we consider the log transformation of the original values as the microarray signal intensities are usually transformed to the log scale (e.g. Di et al. (2005), Rabbee and Speed (2006)). Moreover we found from the initial data analysis that log-transformation works better as compared with the raw values. For generic alleles X and Y, the classes

Table 2.5: Applying LDA using four sets of classifiers

Classifier	*Variables*
ASO.L	log(ASO.XL), log(ASO.YL)
ASO.R	log(ASO.XR), log(ASO.YR)
APEX.L	log(APEX.XL), log(APEX.YL)
APEX.R	log(APEX.XR), log(APEX.YR)

are XX, XY, YY, and NN (if all are present). Table 2.6 sets out the notation for the Bayesian posterior probabilities for the possible classes from each of the four possible classifiers. For example, $P_{\mathrm{XX}}^{(\mathrm{ASO.L})}$ is the posterior probability for the XX genotype from the classifier, ASO.L. The posterior probabilities for the four classifiers are combined using an entropy weighting scheme. Entropy is a measure of uncertainty or dispersion of a random variable, which can give more weight to seemingly more confident classifiers. Other measures could be used, such as the variance of the posterior probabilities, but entropy performs well for this application. Another issue is the uncertainty in the *estimates* of the individual posterior probabilities, which

Table 2.6: Posterior probabilities from four LDA classifiers

Classifier/Class	XX	XY	YY	NN
ASO.L	$P_{\mathrm{XX}}^{(\mathrm{ASO.L})}$	$P_{\mathrm{XY}}^{(\mathrm{ASO.L})}$	$P_{\mathrm{YY}}^{(\mathrm{ASO.L})}$	$P_{\mathrm{NN}}^{(\mathrm{ASO.L})}$
ASO.R	$P_{\mathrm{XX}}^{(\mathrm{ASO.R})}$	$P_{\mathrm{XY}}^{(\mathrm{ASO.R})}$	$P_{\mathrm{YY}}^{(\mathrm{ASO.R})}$	$P_{\mathrm{NN}}^{(\mathrm{ASO.R})}$
APEX.L	$P_{\mathrm{XX}}^{(\mathrm{APEX.L})}$	$P_{\mathrm{XY}}^{(\mathrm{APEX.L})}$	$P_{\mathrm{YY}}^{(\mathrm{APEX.L})}$	$P_{\mathrm{NN}}^{(\mathrm{APEX.L})}$
APEX.R	$P_{\mathrm{XX}}^{(\mathrm{APEX.R})}$	$P_{\mathrm{XY}}^{(\mathrm{APEX.R})}$	$P_{\mathrm{YY}}^{(\mathrm{APEX.R})}$	$P_{\mathrm{NN}}^{(\mathrm{APEX.R})}$

could be assessed, for example, via bootstrap variances. We do not pursue this here.

Denote the four posterior probabilities from any classifier (*) in any row of Table 2.6 by P_c^*, where c indexes one of the classes (genotypes) in the set

$$C = \{\mathrm{XX}, \mathrm{XY}, \mathrm{YY}, \mathrm{NN}\}.$$

Posterior probabilities P_c^* are calculated using the function *lda* in R from library "MASS". The prior probabilities for the four classes are based on the training-data frequencies. The underlying calculations associated with the function *lda* follow similar steps as described in Section 1.5.2. Then the entropy for this probability distribution over classes is defined to be

$$-\sum_{c \in C} P_c^* \log(P_c^*).$$

Entropy or uncertainty is maximized when all the P_c^* are equal and minimized (taking the value 0) when one of the P_c^* is 1 and the others are zero.

Entropy is computed for each of the four classifiers in Table 2.6. We will be giving more weight to a classifier with *less* entropy (uncertainty). Thus, we define for the ASO.L classifier in row 1 of the table, for example,

$$E_{\mathrm{ASO.L}} = -\log(\frac{1}{4}) - [-\sum_{c \in C} P_c^{(\mathrm{ASO.L})} \log(P_c^{(\mathrm{ASO.L})})],$$

which is a quantity which is large if ASO.L's entropy is small compared to the maximum possible entropy. Analogous quantities are computed for $E_{\mathrm{ASO.R}}$, $E_{\mathrm{APEX.L}}$, and $E_{\mathrm{APEX.R}}$ in Table 2.6. The weights for the four classifiers are obtained by normalizing them so that they sum to 1, i.e.,

$$W_{\mathrm{ASO.L}} = \frac{E_{\mathrm{ASO.L}}}{E_{\mathrm{ASO.L}} + E_{\mathrm{ASO.R}} + E_{\mathrm{APEX.L}} + E_{\mathrm{APEX.R}}},$$

with analogous computations for $W_{\text{ASO.R}}$, $W_{\text{APEX.L}}$, and $W_{\text{APEX.R}}$. Note that the probabilities in Table 2.6 and hence the weights will vary from one test sample to another.

The weights for the four classifiers are applied to the posterior probabilities for each class (column) in Table 2.6 to obtain the final class posterior probabilities. For example, the final probability for XX is

$$P_{\text{XX}} = \sum_j W_j P_{\text{XX}}^{(j)},$$

where $j \in \{\text{ASO.L, ASO.R, ASO.L, ASO.R}\}$ with similar calculations for XY, YY, and NN. A sample is assigned to the class with maximum weighted probability.

To increase the concordance with the validated test samples (at the expense of reducing the call rate), a call is made if and only if the maximum probability across the classes exceeds a threshold. For instance, the results in the last two columns of the first row of Table 2.3 are obtained by requiring the maximum probability to be at least 0.6 for a call.

Example corresponding to SNP rs1003399 and sample Coriell NA17111

Figure 2.3 relates to SNP rs1003399 and the point labeled 11 is Coriell sample NA17111. To check how dynamic-variable LDA works for a sample with complex redundancy, we predict the genotype of that sample based on the remaining 31 Coriell samples plus three negative PCR controls. Underlying calculations for both dynamic-variable LDA and simple LDA are shown here.

The posterior probabilities from dynamic-variable LDA corresponding to Table 2.6 but specific to this example are given in Table 2.7. The fi-

Table 2.7: Posterior probabilities from Table 2.6 for SNP rs1003399 and target sample Coriell NA17111

Classifier/Class	CC	CG	GG	NN
ASO.L	< 0.001	0.001	0.999	< 0.001
ASO.R	< 0.001	0.003	0.997	< 0.001
APEX.L	< 0.001	1.000	< 0.001	< 0.001
APEX.R	< 0.001	0.005	0.995	< 0.001

nal posterior probabilities from Dynamic-variable LDA and Simple LDA are given in Table 2.8. So from Table 2.8, it is clear that the sample

Table 2.8: Resultant Posterior probabilities from Two Methods

Classes/Methods	Dynamic-variable LDA	Simple LDA
CC	< 0.001	< 0.001
CG	0.253	1.000
GG	0.746	< 0.001
NN	< 0.001	< 0.001

Coriell NA17111 (with validated genotype GG) is correctly classified only by dynamic-variable LDA with confidence measure .75, but simple LDA fails to do so. Moreover simple LDA wrongly classifies the sample as CG with high confidence score (posterior probability 1.000).

This example also illustrates potential problems with a simple 0/1 weighting scheme, where weight 1 is given to the most confident classifier. In Table 2.7, APEX.L assigns 1.00 posterior probability to GG, and this is the most confident classifier, but it makes the wrong call. The right call is made here by also taking account of ASO.L and ASO.R.

Bibliography

Di, X., H. Matsuzaki, T. Webster, E. Hubbell, G. Liu, S. Dong, D. Bartell, J. Huang, R. Chiles, G. Yang, et al. (2005). Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics 21*(9), 1958–1963.

Gemignani, F., C. Perra, S. Landi, F. Canzian, A. Kurg, N. Tonisson, R. Galanello, A. Cao, A. Metspalu, and G. Romeo (2002). Reliable Detection of {beta}-Thalassemia and G6PD Mutations by a DNA Microarray. *Clinical Chemistry 48*(11), 2051.

Hand, D., H. Mannila, P. Smyth, et al. (2001). Principles of Data Mining [M]. *Massachusetts Institute of Technology*, 362–363, 359–360.

Kamiński, S., A. Ahman, A. Ruœć, A. Wójcik, and T. Malewski (2005). MilkProtChip–a microarray of SNPs in candidate genes associated with milk protein biosynthesis–development and validation. *J Appl Genet 46*(1), 45–58.

Kennedy, G., H. Matsuzaki, S. Dong, W. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, et al. (2003). Large-scale genotyping of complex DNA. *Nature Biotechnology 21*(10), 1233–1237.

Kurg, A., N. Tonisson, I. Georgiou, J. Shumaker, J. Tollett, and A. Metspalu (2000). Arrayed Primer Extension: Solid-Phase Four-Color DNA Resequencing and Mutation Detection Technology. *Genetic Testing 4*(1), 1–7.

Livak, K., S. Flood, J. Marmaro, W. Giusti, and K. Deetz (1995). Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *Genome Research 4*(6), 357.

Oliphant, A., D. Barker, J. Stuelpnagel, and M. Chee (2002). BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques 32*, S56–S61.

Rabbee, N. and T. Speed (2006). A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics 22*(1), 7–12.

Shen, R., J. Fan, D. Campbell, W. Chang, J. Chen, D. Doucet, J. Yeakley, M. Bibikova, E. Wickham Garcia, C. McBride, et al. (2005). High-throughput SNP genotyping on universal bead arrays. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis 573*(1-2), 70–82.

Tebbutt, S., J. He, K. Burkett, J. Ruan, I. Opushnyev, B. Tripp, J. Zeznik, C. Abara, C. Nelson, and K. Walley (2004). Microarray genotyping resource to determine population stratification in genetic association studies of complex disease. *Biotechniques 37*(6), 977–85.

Tebbutt, S., I. Opushnyev, B. Tripp, A. Kassamali, W. Alexander, and M. Andersen (2005). SNP Chart: an integrated platform for visualization and interpretation of microarray genotyping data.

Walley, D., B. Tripp, Y. Song, K. Walley, and S. Tebbutt (2006). MACGT: multi-dimensional automated clustering genotyping tool for analysis of microarray-based mini-sequencing data.

Chapter 3

Validation of Genotype Calling Algorithm

3.1 Background

If 'personalized medicine', using genomic knowledge, is to become a reality, then the ability to determine the most appropriate clinical intervention for a patient will require the genotyping of several tens to hundreds of single nucleotide polymorphisms (SNPs) across many genes and their regulatory sequences for that individual patient (Yang et al., 2003), (Janssens et al., 2004), rapidly and at the point-of-care. Of many genotyping methods, those based on microarrays offer the greatest potential for economic, patient-specific application (Hirschhorn et al., 2000), (Kennedy et al., 2003), (Oliphant et al., 2002), (Pastinen et al., 2000), (Steemers et al., 2007), due to their ability to simultaneously interrogate multiple SNPs. Arrayed primer extension (APEX: Kurg et al. (2000), Shumaker et al. (1996)) is a minisequencing microarray assay based on a two-dimensional array of oligonucleotide probes that are immobilized, via their 5' ends, on a glass surface. The probes (25-mers) are designed so that they are complementary to the gene up to, but not including, the base where the SNP exists. The Sanger-based sequencing chemistry of APEX allows genotyping of hundreds of SNPs, with the array chemistry taking only fifteen to twenty minutes to complete. APEX achieves this clinically relevant speed because it uses the catalytic ability of a DNA polymerase to carry out a single nucleotide base extension (SBE) at the 3' end of the arrayed probes, specific to the SNP sites of interest in amplified patient DNA that is temporarily hybridized to these probes. The dideoxynucleotide (ddNTP) 'terminator' bases are la-

A version of this chapter has been published as Mohua Podder, Jian Ruan, Ben W Tripp, Zane E Chu and Scott J Tebbutt: **Robust SNP genotyping by multiplex PCR and arrayed primer extension**; *BMC Medical Genomics*, 2008: 1-5.

belled with tags containing distinct fluorescent chromophores, specific for each of the four bases of DNA (A,C,G,T). Hence, the fluorescent 'colour' at each of the probe sites (array spots) will give SNP-specific genotypic information. As a discovery research tool, APEX has been used to detect -thalassemia (Gemignani et al. (2002), p53 of Tonisson et al. (2002)), and BRCA1 mutations (Tõnisson et al., 2000). Importantly, APEX has also been shown to be efficient at simultaneously genotyping SNP markers that are widely dispersed across the human genome (Dawson et al. (2002), Tebbutt et al. (2004)); such capability is essential for future 'individualized' genomic diagnostic analysis across multiple genes and pathways that are relevant to disease. In a recent quality assessment survey of SNP genotyping laboratories (Lahermo et al., 2006), in which up to 18 SNPs were genotyped across 47 DNA samples, APEX performed well against other methods, and the authors concluded that a "conservative approach for calling the genotypes should be used to achieve a high accuracy at the cost of a lower genotyping success rate." Whilst such a conservative approach may be applicable for research studies, it may not be appropriate for clinical diagnostics, in which life-saving medical decisions might require extremely accurate genotyping across all SNPs of interest.

Given the potential utility of APEX for rapid clinical diagnostics, we have developed robust assay design, chemistry and analysis methodologies, and have sought to determine just how effective APEX is in comparison to leading 'gold-standard' genotyping platforms, including Perlegen and Illumina. Our objective was to achieve 100% assay completion rate, call rate and genotyping accuracy rate, for multiple SNPs across multiple samples. Previous studies from our laboratory have reported APEX genotyping accuracies ranging from 98% to 99.8% (Tebbutt et al. (2004), Tebbutt et al. (2006), Walley et al. (2006), Podder et al. (2006)), though the call rates in these studies have always been significantly lower than 100%, and usually do not include a proportion of the originally selected SNPs that fail the assay. Similarly, other laboratories that use APEX and equivalent technology have reported genotyping accuracies ranging from 98% to $> 99\%$, with call rates varying from 84.4% to 96.8% (Gemignani et al. (2002), Tonisson et al. (2002), Dawson et al. (2002), Lahermo et al. (2006), Cremers et al. (2007), Zernant et al. (2005) and Jaakson et al. (2003)).

3.2 Results and Discussion

We selected 50 SNPs from the HapMap database that had been previously genotyped and analyzed as part of the third quality control exercise on Illumina and Perlegen platforms, arguably the most accurate and best validated high-throughput methodologies for SNP genotyping to date. The randomly selected SNPs were located across multiple chromosomes and are listed in Additional Table 1 online, along with details of the APEX probe sequences and PCR primer sequences. The genotyping arrays that are currently being developed and tested in our laboratory incorporate multiple redundant measures consisting of sense and antisense DNA-strand APEX probes plus allele-specific oligonucleotide (ASO) APEX probes for a total of six different probes per SNP (Tebbutt et al., 2004), with each replicated five times on the array grid, which allows for more robust statistical averaging. Optimal PCR primer pairs were designed for each of the 50 SNP loci (Additional Table 1 online) and seven multiplex PCR groups were set up that, together, would amplify all 50 loci (Additional Table 2 online). We obtained a set of 287 DNA samples from McGill University and Gnome Qubec Innovation Centre (one of the HapMap Project's genotyping centers). This set comprised 270 DNA samples from the Coriell Institute for Medical Research (http://coriell.org/) plus hidden duplicates and negative controls, all of which our laboratory was blinded to. PCR (Fig. 3.1a and Fig. 3.1b) and APEX assays were performed on each of the samples, plus a 10% repeat set which was randomly selected by us to allow internal quality control and an initial assessment of genotyping concordance. Fig. 3.1 describes multiplexing PCR and subsequent amplicon fragmentation results, prior to APEX reaction on HapMap Chip. (a) Standard multiplex PCR from a single Coriell DNA sample using optimally-designed primers within seven unique multiplex groups (lanes 1-7; lane M shows 100 bp DNA ladder markers), showing wide range of amplicon sizes across the 50 SNP loci. (b) Purification, concentration and fragmentation of standard PCR amplicons. Lane 1 represents an aliquot of concentrated mixture of all seven multiplex products shown in Fig. 3.1a. Lane 2 shows the fragmentation result, generating single-stranded nucleic acid of 30-100 base length. (c) Multiplex PCR amplification of all 50 SNP loci in a single reaction tube using new PCR primer set (Additional Table 4 online), showing 50-plex PCR products (individual SNP loci amplicons are unresolvable by agarose gel electrophoresis) from two Coriell DNA samples (lanes 1 & 2), plus a negative PCR control (lane 3). (d) Fragmentation of 50-plex PCR amplicons from aliquots of lane 1 & lane 2 samples shown in Fig. 3.1c.

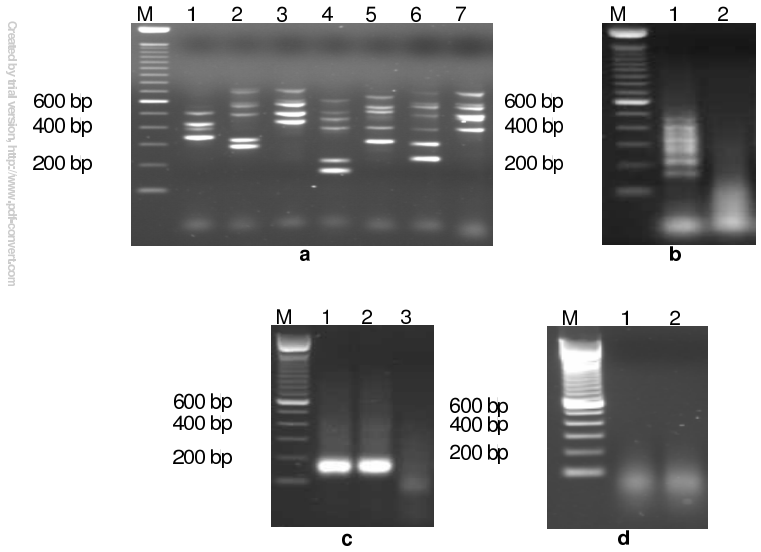

Figure 3.1: Multiplexing PCR and subsequent amplicon fragmentation results, prior to APEX reaction on HapMap Chip.

Microarray image data were imported into SNP Chart (Tebbutt et al., 2005) and analyzed using previously described image analysis algorithms (Abbaspour et al. (2006) and Abbaspour, Abugharbieh, Podder, and Tebbutt (Abbaspour et al.)). Genotypes were called using two previously published methods: 1. MACGT software (Walley et al., 2006), which is a multidimensional clustering tool; 2. simple linear discriminant analysis (LDA) using dynamic variable selection (Podder et al., 2006), which is a classification algorithm. Results are shown in Table 3.1 and Additional Table 3 online. Briefly, a training set was established using SNP Chart, followed by auto-calling in MACGT. Nine SNPs did not pass quality control due to assay failure or inconsistent PCR amplification. For all remaining SNPs that were auto-called by MACGT, any genotypes that had a 'fit' score of less than 0.001 (approximately 9%) were checked by manual scoring in SNP Chart and either validated, or changed to a different genotype or to a non-call (NN). The final results using MACGT showed highly accurate genotyping (99.94% concordance with HapMap) with good call rates (90% auto-called plus 9% manual scoring). Importantly, of the 1,013 genotypes called manually, the accuracy was 99.87%, even in cases where the array spot signal intensities were up to an order of magnitude lower than for higher quality genotype

data, and only slightly higher than background signals (Additional Fig. 1 and Fig. 2 online). Using the same training set, we then analyzed the data set with simple linear discriminant analysis (LDA) using dynamic variable selection (Podder et al., 2006). Results (Table 3.1 and Additional Table 3 online) also showed accurate genotyping (99.91% HapMap concordance), and with higher automated call-rates (94.91% - using a confidence score threshold of 0.75). We also calculated the homozygous and heterozygous performance for the set of 270 HapMap samples with the previously selected 41 SNPs out of 50 SNPs (See Table 3.1). For a threshold of 0.75, we were able to call 6883 cases out of 7214 homozygous cases (95.41% call rate) with 6880 correct calls (99.96% HapMap concordance). Whereas, with the same threshold, out of 3873 heterozygous cases, we were able to call 3640 cases (93.98% call rate) with 3634 correct calls (99.84% HapMap concordance). Therefore, in common with other genotyping platforms, our methodology has a slight bias that favours the calling of homozygous genotypes. These

Table 3.1: Results summary for 287 HapMap samples and 41 SNPs

Method	Call rate	HapMap Accuracy
MACGT (0.001) + manual	98.90% (9% manual)	99.94%
LDA (0.75) Total	94.91%	99.91%
LDA (0.75) Homozygous	95.41%	99.96%
LDA (0.75) Heterozygous	93.98%	99.84%

Table 3.2: Results summary for 49 HapMap samples and 50 SNPs

Method	Call rate	HapMap Accuracy
Manual calling only	100.0%	99.92%
MACGT (no cut-off)	100.0%	99.84%
LDA (0) Total	100.0%	99.89%
LDA (0) Homozygous	100.0%	100.0%
LDA (0) Heterozygous	100.0%	99.7%
MACGT (0.001)	94.04%	99.94%
LDA (0.75) Total	99.18%	99.90%
LDA (0.75) Homozygous	98.91%	100.0%
LDA (0.75) Heterozygous	99.7%	99.7%

results, although promising and at least as accurate as any previously reported for APEX-based methodologies, did not deliver on our objective of 100% call rate and 100% accuracy, and several of the 50 SNPs failed quality control. However, two important lessons were learnt from the study: 1. our on-chip assay chemistry is extremely robust and specific, allowing accurate genotype calls (at least by manual inspection of the array spot data within SNP Chart) even at very low sensitivities (i.e., when the sequence-specific spot intensities are only slightly higher than background signals); 2. non-calls (NNs) generally resulted from sporadic PCR failure for certain amplicons, especially those of a length greater than 650-700 base pairs (bp). Taken together, our results suggested that even if specific SNPs give high NN rates across multiple samples, the genotypes for the remaining samples for these SNPs (for which APEX assay data can be obtained) are still very accurate, despite low signal to noise. We believe that this is due to the redundancy in the genotyping probe design: two classical APEX probes (one probe per DNA strand), plus four allele-specific (ASO) APEX probes (two probes per strand), each replicated five times, for each SNP site. When this redundant data is displayed in a SNP Chart, it is relatively straightforward to interpret the genotype manually (Additional Fig. 1 and Fig. 2 online). From these conclusions we reasoned that the PCR design itself needed to be addressed, so that sporadic failures (despite good primer design algorithms) could be consistently minimized or even eliminated.

For SNP genotyping, only the immediate sequence around the SNP site is of interest. Therefore, keeping the PCR amplicon size to a minimum ensures short extension times and minimal use of reagents. However, sequence-context issues, especially in multiplex PCR, necessitate the design of unique primers that have balanced annealing temperatures. This requirement can result in individual amplicon sizes in a multiplex mix ranging from 100 to >700 bp (Tebbutt et al., 2004). Large amplicons are optimal neither for fast PCR nor for the subsequent APEX assay, which requires amplicons to be fragmented to 50-100 base lengths (Fig. 3.1b). In addition, the degree of multiplexing is usually limited to between four and ten amplicons per individual multiplex PCR: e.g., for our original HapMap chip, the 50 SNP loci are amplified in a total of seven separate multiplex reactions (Fig. 3.1a and Additional Table 2 online). We initially tested multiplex PCR using all original PCR amplicon primer pairs in a single reaction. As expected, several experimental attempts all failed to amplify even a modest proportion of the 50 amplicons (typically, less than 20 amplicons would be successful; data not shown). Thus, our new objectives were to increase the degree of multiplexing and shorten the amplicon lengths to less than 200 bp, so that all 50 SNP loci

could be simultaneously and robustly amplified in a single reaction vessel. New PCR primers were designed for the 50 HapMap SNP loci, with amplicon sizes restricted to between 100 and 200 bp (Additional Table 4 online). Because of this limitation, we were not able to optimally design the primers based on a balanced melting temperature (Tm). To try to compensate for this potential problem, each new PCR primer had a common linker sequence designed at its 5' end ($^{5'}$TACGACTCACTTAGGGAG$^{3'}$ for each of the left hand PCR primers / $^{5'}$CGATGTAGGTGACACTAG$^{3'}$ for each of the right hand PCR primers). These linkers have two properties: a balanced and reasonably high GC content to increase the melting temperature of the primer and a unique sequence not found in the human DNA template (Wang et al., 1998). After the first few cycles of PCR, the linker sequence becomes incorporated into the amplicon sequence and is amplified along with the template sequence. This approach helps reduce primer-dimer formation during the PCR (Brownie et al., 1997). Because the primers have balanced GC content, primer annealing in later cycles of PCR should become much more sensitive and robust (Wang et al., 2005). We randomly selected 50 of the HapMap Coriell DNA samples from our initial study, for 50-plex PCR using the pool of linker-modified primers. Specific PCR cycling conditions were adopted from a previously published study by Wang et al. (2005). We also attempted 50-plex PCR using the redesigned PCR primers, but without the common 5' linker sequences. We managed to amplify only a modest number of the 50 SNPs, and this multiplex PCR was not robust and we could never amplify all 50 SNPs (data not shown).

PCR (Fig. 3.1c and Fig. 3.1d) and APEX assays (Fig. 3.2) were performed on each of the samples, including negative controls. Microarray image data were imported into SNP Chart and analyzed as described previously. Genotype calling was performed using three independent methods: 1. manual calling in SNP Chart; 2. auto-calling with MACGT; and 3. auto-calling by LDA using dynamic variable selection. Genotypes were compared to HapMap data for concordance. One SNP (rs7693776) was monomorphic (TT) across all samples genotyped. Results are presented in Table 3.2 and Additional Tables 7-10 online. Manual genotype calling, although time-consuming and vulnerable to user-subjectivity issues (Tebbutt et al. (2004), Tebbutt et al. (2005)), is nevertheless an accurate and validated way to interpret APEX data, especially at low spot intensity levels (see above). In addition, manual calling does not require the use of a training set. Of the 49 Coriell DNA samples (one sample out of the random set was a blinded negative control sample) assayed across 50 SNPs, manual calls were made for all possible 2,450 genotypes (100% assay completion and 100% call rate).

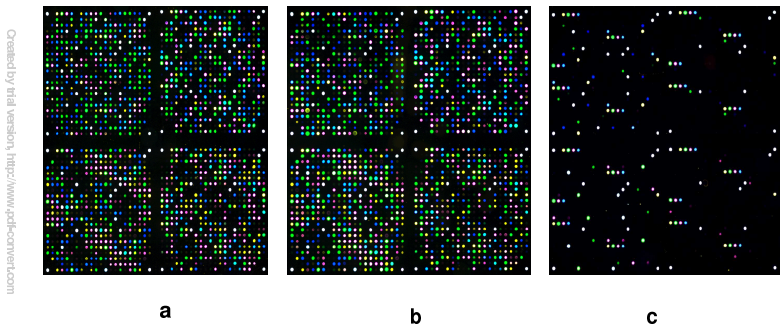

Figure 3.2: HapMap Chip four colour microarray images showing successful de-multiplexing of 50-plex PCR from two Coriell DNA samples (a, b), plus a negative control sample (c), prior to image analysis and automated genotyping. The spots on the negative control image represent positive control probes.

Of these, 2,448 were concordant with HapMap data (99.92%). The two discrepant genotypes were for two different samples each at different SNP loci. Interestingly, the SNP Charts for these two genotypes showed high quality data, and the same samples/genotypes had previously been concordant with HapMap in the initial data set (Additional Table 3 online, and discussed further below).

Auto-calling was independently undertaken. Initially, MACGT cluster plots and quality control using SNP Chart were used to allow manual selection of a limited training set of samples from the data set (Walley et al., 2006). Using this training set, MACGT auto-calling of the test set with a 0.001 fit threshold resulted in a call rate of 94.04% and a concordance rate of 99.94%. When the fit threshold was relaxed to achieve a 100% call rate, three genotypes were discordant with HapMap data. Two of these genotypes (both with high fit values - good confidence scores) were the same as the two that had been identified as part of the manual calling data. The third discrepancy had a relatively poor fit confidence score. LDA with dynamic variable selection, using a slightly reduced sized training set, yielded identical genotyping results to manual calling, at a 100% call rate across all 50 SNPs (16 NNs at a 0.65 confidence score threshold). Again, the two discrepant genotypes, both of which were incorrectly called as homozygous, had high confidence scores, consistent with high quality APEX assay data. Separate analysis of homozygous and heterozygous cases showed that for a 0.0 threshold, homozygous cases (1289 in total) achieved a call rate of

100% with 100% HapMap concordance, whereas heterozygous cases (652 in total) achieved a call rate of 100% with 99.7% HapMap concordance (two heterozygous errors with high confidence scores). Surprisingly, with a 0.65 threshold, among 16 non-calls 14 were homozygous with 11 cases (all TT genotypes) from a single SNP rs1891403, which gives a homozygous call rate of 98.9% and a heterozygous call rate of 99.7%. Interestingly, the LDA-called genotype that had the lowest score (but nevertheless was still called correctly) was the same genotype as the third MACGT-called discordant genotype (see above and Additional Table 7 online). Subsequent inspection of the SNP Chart for this genotype (heterozygous CT) showed that the ASO-APEX probe intensity signals for the C allele were somewhat lower than the T allele signals. Again, this same sample/genotype had previously been concordant with HapMap in the initial data set, using the original PCR primer pairs. (See below for further discussion of this genotype and the other two discrepant genotypes.)

In summary, we have shown that a combination of multiplex PCR, redundant and robust APEX design and assay, and statistically-robust auto-calling (simple LDA using dynamic variable selection) can achieve 100% completion and call rate with >99.9% accuracy, for multiple SNPs and multiple samples. We believe that this is a significant improvement over other published APEX methodologies. The strength of our methodology is not based on the quality of a single measurement but on the redundancy obtained from measuring the allele intensities by using multiple chemistries. To take advantage of this inherent robustness of the assay we use robust statistical methods that automatically select the most reliable measurements for each SNP to make the genotype call, sample by sample (Podder et al., 2006). Redundancy in genotyping arrays is associated with higher costs per SNP, concomitant with lower numbers of SNPs able to be interrogated in a given area of the microarray. For research studies, a trade-off may need to be taken into consideration, given the ever-increasing need to genotype as many SNPs as possible, at minimal cost per SNP, and a recent article by Smemo and Borevitz (2007) cogently argues for a reduction in the approximately 40-fold probe redundancy currently featured on Affymetrix GeneChips, which only use hybridization for allelic signal generation. For clinical diagnostics however, we believe that genotyping accuracy, call rate and completion rate are paramount.

To further determine the effect of probe redundancy in our APEX methodology, we used LDA to reanalyze both data sets (original and 50-plex) but using non-redundant and partially-redundant probe-specific data (Additional Tables 8-10 online). Fig. 3.3 and Additional Fig. 3 online show simple

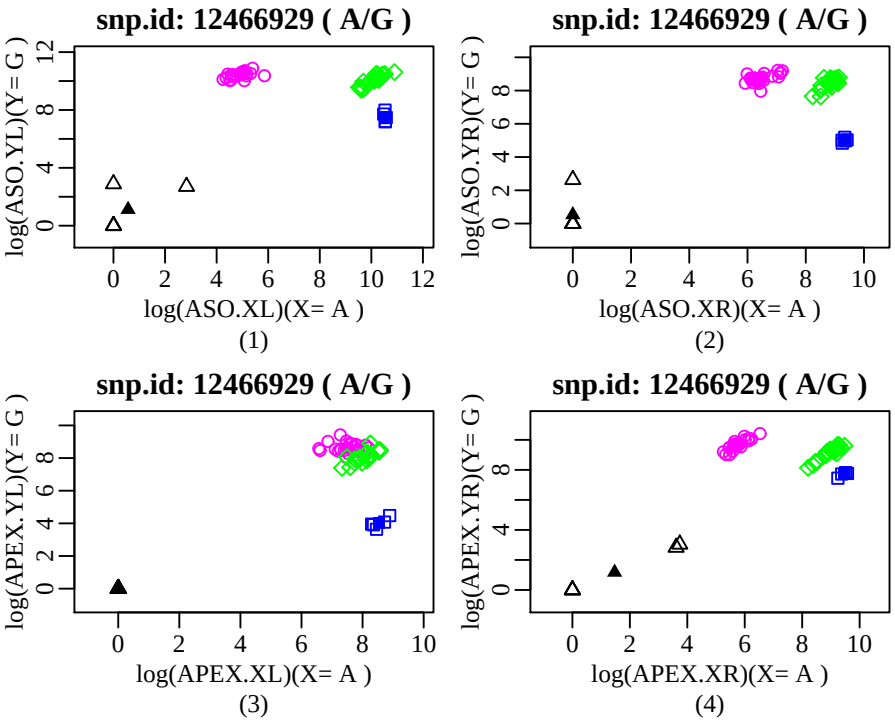

Figure 3.3: Simple scatter plots for SNP rs12466929 (A/G) from 50-plex data set (this SNP is representative of the entire set of 50 HapMap SNPs).

four-panel scatter plots of the probe data for the 50-plex experiment. For each plot the x-axis represents signal values for X allele (A for this SNP) and the y-axis represents signal values for Y allele (G for this SNP). All values are in log scale. Magenta, green, blue and black coloured symbols denote the classes YY (GG), YX (AG), XX (AA) and NN (negative control samples), respectively. Plot (1) combines the two ASO-APEX Left probes (one for each allele); plot (2) combines the two ASO-APEX Right probes (one for each allele); plot (3) is for the APEX Left probe; plot (4) is for the APEX Right probe. All the classifiers except APEX Left (plot 3) give well separated genotype clusters for this SNP. Dynamic variable selection is able to automatically weight these LDA classifiers in such a way that the homozygous AA cluster in plot (3) (blue) is able to contribute to the final call for such genotypes, even though AG (green) and GG (magenta) genotype clusters overlap somewhat for this Left APEX probe. Additional Fig. 3 online shows four-panel scatter plots for all 50 SNPs from the 50-plex data set.

In particular, Fig. 3.3 represents the four separate scatter plots for the SNP rs12466929 corresponding to the four different probe chemistries: ASO.LEFT, ASO.RIGHT, APEX.LEFT and APEX.RIGHT. For each scatter plot, the three possible genotype clusters (previously known from the HapMap data set) are presented with three different colours: blue for allele 1 homozygous; magenta for allele 2 homozygous; and green for allele 1 and allele 2 heterozygous. For the SNP rs12466929, allele 1 is A and allele 2 is G, and the scatter plots are representative of the entire set of 50 HapMap SNPs. The four scatter plots indicate that three out of the four probe chemistries work perfectly well and produce well separable (informative) clusters corresponding to the three genotype classes (AA, AG and GG), whereas one probe chemistry, namely APEX.LEFT, fails to work properly and gives overlapping clusters for AG and GG genotype classes (plot (3) in Fig. 3.3). Nevertheless, this probe chemistry gives a well separable cluster for the AA genotype class. This phenomenon conveys the point of considering each probe chemistry separately during the building of the genotype classification model, and in the next stage of the genotype calling algorithm, combining the four genotype models with proper weights adjusted dynamically with the quality of each of the four classifiers (four probe chemistries) specific to each SNP and sample. If all four probes failed to produce informative clusters, then our LDA-based genotype calling algorithm would flag that SNP as a failed SNP, which clearly is not the case for the SNP rs12466929. This is how the redundancy amongst our APEX based genotyping platform is captured through the proposed LDA-based genotype

calling algorithm with dynamic variable selection. Viewing the four-panel scatter plots, we would also like to emphasize the point that for most of the SNPs the homozygous clusters show some significant signal intensities corresponding to the other allele, due to spectral overlap within the APEX fluorescent ddNTP chemistry, thus inducing background to the homozygous clusters. Particularly for this reason, we do not often see a homozygous cluster close to either of X- or Y-axes. Here, the aim is to compare the allele 1 and allele 2 signal intensities for the three possible genotype classes, and then assign a test sample to the appropriate class based on the prior knowledge of the available training set. We would also like to mention that the initial signal intensities corresponding to each allele for all four probe chemistries are converted into the log-scale in order to reduce the variability between several microarray slides.

Performance analyses for the different data sets are described below, addressing the redundant probe chemistry (Table 3.3, 3.4, 3.5). The extreme left hand column of each table indicates the combination of four classifiers (APEX.L; APEX.R; ASO.L and ASO.R) used to build the LDA model. For example, in the first row, all four classifiers were used to give the final genotype call, and in the fourth row, only the left classifiers were used. In the last four rows, only one classifier was used at a time to give independent genotype calls using the simple LDA model (with no dynamic variable selection). For the complete set of 287 HapMap samples and the set of 41 SNPs, the training data had in total 807 genotype cases (among which 519 genotypes were from HapMap Coriell samples and 288 genotypes were from other Coriell samples) and the test data had in total 11,248 genotype cases (among which 163 had no validated genotypes from HapMap for comparison). For the set of 270 HapMap DNA samples, applying a 0.65 threshold improved the concordance rate (0.31% miss-classification rate) with a reduced call rate of 97.30%. We further checked the performance of the same data set applying a stringent threshold of 0.75, which gave 99.91% concordance (0.06% miss-classification rate) for a reduced call rate of 94.91%. Applying different level of thresholds, we can control the call rates and, given the validated genotype set, we can also check the performance level by calculating the miss-classification rates. The underlying supposition is that, with reduced call rate, accuracy should increase successively until it reaches its maximum limit. For the improved 50-plex PCR chemistry, we were able to achieve a high concordance rate (99.89% using all four classifiers) with 100% call rate (see Table 3.3). If we apply a 0.65 threshold to the set of 50-plex PCR HapMap samples, then the automated call rate reduced to 99.18%, leaving only 16 non calls (below threshold value) to be verified manually using SNP

Table 3.3: 270 HapMap samples on the subset of 41 SNPs

	LDA (0)		LDA (0.65)		LDA (0.75)	
Classifiers	Call.rate	Error	Call.rate	Error	Call.rate	Error
All	100	0.8	97.30	0.31	94.86	0.06
APEX	100	1.58	94.52	0.44	92.37	0.42
ASO	100	1.83	95.35	0.69	93.08	0.60
LEFT	100	1.57	94.65	0.46	92.68	0.40
RIGHT	100	2.49	94.48	0.82	92.57	0.69
APEX.L	100	5.16	97.42	4.02	95.85	3.41
APEX.R	100	4.84	98.66	4.4	97.59	4.05
ASO.L	100	4.30	97.53	3.65	96.41	3.37
ASO.R	100	5.05	97.58	4.03	95.57	3.34

Table 3.4: 50-plex HapMap samples on 50 SNPs using smaller training set including three negative control samples

	LDA (0)		LDA (0.65)	
Classifiers	Call.rate	Error	Call.rate	Error
All	100	0.12*	99.26	0.12*
APEX	100	0.18	97.28	0.12*
ASO	99.50	1.42	96.33	0.72
LEFT	99.51	0.96	95.92	0.30
RIGHT	99.67	0.24	97.93	0.12*
APEX.L	99.37	2.52	98.27	2.04
APEX.R	99.72	1.98	96.88	1.38
ASO.L	99.32	2.40	98.66	2.10
ASO.R	99.28	1.38	97.50	0.96

Table 3.5: 50-plex HapMap samples on 50 SNPs using minimal training set including three negative control samples

	LDA (0)		LDA (0.65)	
Classifiers	Call.rate	Error	Call.rate	Error
All	99.61	0.10*	96.96	0.10
APEX	99.02	0.70	93.87	0.10
ASO	98.57	4.80	87.1	1.00
LEFT	98.67	3.20	91.54	0.85
RIGHT	98.39	1.05	91.28	0.25
APEX.L	98.93	3.60	98.18	3.35
APEX.R	98.77	1.95	96.88	1.35
ASO.L	98.92	5.85	95.84	5.35
ASO.R	98.34	5.20	95.30	4.90

Chart (all of which were correct).[1]

Therefore, we have determined that reliance on any single probe type alone [*i.e.*: APEX Left probe; APEX Right probe; 2 x ASO-APEX Left probes (one for each allele); 2 x ASO-APEX Right probes (one for each allele)] resulted neither in as high an accuracy of genotyping nor in as high a call rate, compared to the dynamic use of multiple probes.

We were interested in further study of the two discrepant genotype cases, since both had previously been concordant with HapMap in the 7reaction-multiplex PCR data set, and both showed high quality, unambiguous SNP Charts in the 50plex PCR data set. A third genotype case (concordant with HapMap by manual calling and simple LDA, but with a low quality score of 0.4876) was also discrepant when called by MACGT. We re-amplified these three individual SNP loci from their respective Coriell DNA samples, using the original PCR primers (Additional Table 1 online), and sequenced each amplicon from both ends. The two discrepant genotypes were: 1. DNA sample 192 (NA18502) at SNP rs3776720 50plex genotype GG / HapMap & 7reaction-multiplex genotype GA; 2. DNA sample 101 (NA18621) at SNP rs12472674 50plex genotype CC / HapMap & 7reaction-multiplex genotype CT. The third genotype case (concordant with HapMap by manual calling

[1]*The only two discrepancies occurred due to the presence of hidden SNPs within the PCR primer sites. Otherwise, manual inspection of the data corresponding to those two cases was completely agreeable with the predicted genotypes by the automated genotype calling algorithm.

and simple LDA, but with a low quality score of 0.4876) was also discrepant when called by MACGT (DNA sample 228 (NA19210) at SNP rs4739199 50plex genotype (MACGT) TT / HapMap, 7reactionmultiplex, and 50plex (manual call & LDA) genotype CT).

As expected, we identified additional polymorphic sites that coincided with the positions delimited by the PCR primer sequences used for the 50-plex reaction. One of the sites was identified as an existing SNP (rs6871885). To our knowledge, the other two sites represent genetic variants not previously reported. For each of these cases, it appears that the sequence variation within the PCR primer site has caused allelic drop-out, resulting in homozygous genotype calls for the two discrepant cases, and a poor quality heterozygous genotype call for the third case (partial allelic drop-out). Specifically, for discrepant genotype case 1 (Coriell NA18502 at SNP rs3776720), we found a neighbouring SNP (T/A) which is located at the 3' end of the anti-sense PCR primer site ($^{5'}$CGA TGT AGG TGA CAC TAG TAT TGC AGG CAG ACG TGA$^{3'}$ - Additional Table 4 online) - this polymorphic site (30 bp downstream of rs3776720) is reported in dbSNP as rs6871885, with the A base (sense strand) being described as a rare allele (0.083) in sub-Saharan African populations only (Coriell NA18502 is indeed a sub-Saharan African, Yoruba, and is heterozygote for this SNP).

For discrepant genotype case 2 (Coriell NA18621 at SNP rs12472674), we found a sequence variant (G/A) 52 bp downstream of SNP rs12472674, located within the anti-sense PCR primer site ($^{5'}$CGA TGT AGG TGA CAC TAG CTC AAT ATG TTA CCA CAA$^{3'}$ - Additional Table 4 online) - this variant (heterozygous in Coriell NA18621 - Asian, Han Chinese) has not been previously reported in dbSNP and may represent a novel polymorphism. For the low quality genotype case 3 (Coriell NA19210 at SNP rs4739199), we found a sequence variant (G/A) 45 bp downstream of SNP rs4739199, located within the anti-sense PCR primer site ($^{5'}$CGA TGT AGG TGA CAC TAG TCC ACT TCA TTA GGT GAA$^{3'}$ - Additional Table 4 online) - this variant (heterozygous in Coriell NA19210 - sub-Saharan African, Yoruba) has also not been previously reported in dbSNP and may represent a novel polymorphism.

Whilst more stringent due-diligence at the 50-plex PCR primer design stage would have alerted us to one of these SNPs (rs6871885), the evidence that we have identified two hitherto unreported SNPs provides a cautionary tale (Quinlan and Marth, 2007). Elimination of such 'sporadic' genotyping errors due to novel or unaccounted-for SNPs, as well as due to structural variation in the genome (e.g., copy number variants - CNVs) (Feuk et al., 2006), will need to be addressed in future clinical diagnostic genotyping

technologies, and possibly even in research discovery studies where any sporadic errors due to hidden SNPs will not cause significant departure from Hardy-Weinberg equilibrium (Lahermo et al., 2006). In preliminary studies we have been able to correct all three discrepancies previously described, using a redundant 50-plex PCR assay that includes two primer pairs for each SNP loci (data not shown).

Finally, due to the low amount (5 ng) of genomic DNA required for the 50-plex PCR (compared to 25 ng for each of the 7-reaction-multiplex PCRs), we have attempted APEX genotyping using our improved methodology on DNA derived from plasma samples. A pilot project was performed on five plasma samples (stored for up to ten years). Comparing the plasma-derived genotyping data with data obtained from high quality genomic DNA for the same five individuals, the call rate was >99% (100% for high quality DNA) and the concordance was >99%, which opens up the possibility of robust and accurate genotyping of clinical plasma samples without any need for prior whole genome amplification.

3.3 Conclusion

We report significant improvements to arrayed primer extension (APEX) genotyping methodology that may show utility in future point-of-care genetic diagnostic applications. Our methods have been validated against industry-leading technologies in a blinded experiment based on Coriell DNA samples and SNP genotype data from the International HapMap Project. Modifications to PCR amplification design have allowed robust 50-plex genotyping from as little as 5 ng of DNA, with 100% call rate and >99.9% accuracy.

3.4 Materials and Methods

3.4.1 DNA Samples and Validated Genotypes

A set of 287 DNA samples were obtained from McGill University and Gnome Qubec Innovation Centre (one of the HapMap Project's genotyping centers). This set comprised 270 DNA samples from the Coriell Institute for Medical Research (http://coriell.org/) plus hidden duplicates and negative controls, all of which our laboratory was blinded to. We were given access to the validated HapMap genotyping data for these samples only after we had finished the main genotyping experiment (287 samples / 50 SNPs), and

after we had sent a file of our genotyping results to McGill University.

3.4.2 HapMap APEX Chip - Probe Design and Printing

Six oligonucleotide probes (25 mers) for each SNP were designed using Biodata algorithms (Biodata Ltd., Tartu, Estonia - www.biodata.ee) (Additional Table 1 online): two classical APEX probes (one probe per DNA strand), plus four allele-specific (ASO) APEX probes (two probes per strand) which include the actual SNP site at the 3' end of the probe. Allele-specific single base extension of these ASO-APEX probes during the reaction is contingent on the presence of the actual complementary base at the SNP site in the sample template DNA (Pastinen et al. (2000) and Gemignani et al. (2002)). Probes were synthesized at a 25 nmol scale and aliquotted into 96-well plates by Integrated DNA Technologies (Coralville, IA, USA). We diluted each probe at 200 pmol/μL as stock concentration in pure water (resistivity of 18.2 MΩ-cm and total organic content of less than five parts per billion) using a Biomek FX robot (Beckman Coulter, Fullerton, CA, USA).

Arrays were generously printed for us at the Microarray Facility of The Prostate Centre at Vancouver General Hospital (University of British Columbia, Vancouver, BC, Canada). Briefly, the APEX and ASO-APEX probe oligonucleotides (50 pmol/μL in 150 mM sodium phosphate printing buffer, pH 8.5) were printed to specific grid positions on CodeLink^{TM} Activated Microarray Slides (Amersham Biosciences/GE Healthcare, Piscataway, NJ, USA) following the manufacturer's recommended protocols. The 5' end of each oligonucleotide probe was amino-modified during synthesis, allowing its covalent attachment to the slide's pre-applied surface chemistry. Each grid consisted of five spot replicates of each of the six probes per SNP, as well as multiple buffer-only spots and positive control normalization spots. The latter comprised an oligonucleotide probe based on a plantspecific gene sequence that will extend by a single N base due to the presence of an exogenous complementary template oligonucleotide in the APEX reaction mixture (Npg1) (Tebbutt et al., 2004). Each Npg1 positive control probe was spotted 40 times onto the grid, at regular physical intervals. Each one of the six probes for each SNP was printed at a reasonably wide distance apart from any other probe for the same SNP within the grid (as were their replicate spots). This enabled a useful degree of robustness in the system, especially helpful in cases of high local background and hybridization problems (Tebbutt et al., 2004). Each spot was approximately 110 μm in diameter. Three replicated grids were printed on each slide, enabling three samples to be genotyped per slide. Follow-

ing the printing of the arrays, the slides were incubated overnight at room temperature at 75% relative humidity (saturated NaCl chamber) to drive the covalent coupling reaction between the probes' 5' amino group and the CodeLink^{TM} slide chemistry to completion. Blocking of the arrays was in 50 mM ethanolamine, 0.1 M Tris, pH 9.0, 0.1% SDS, at 50°C for 20 min, according to the manufacturer's protocol.

3.4.3 PCR Amplification and Fragmentation

For the first experiment, PCR primers were designed to amplify the regions across the 50 SNPs, based on a melting temperature (T_m) of 62°C ± 3°C (at 20 mM monovalent salt concentration in PCR buffer Additional Table 1 online). All primers were computationally tested against the human genome and found to amplify single product (Biodata Ltd., Tartu, Estonia www.biodata.ee). Multiplex PCR amplifications were performed on the Coriell genomic DNA samples (plus several negative PCR control samples that contained no genomic DNA). The multiplex PCR group had a unique combination of the primer pairs among 7 reactions (Additional Table 2 online). Each PCR was performed in a total volume of 15 μL, containing 1.5 μL 10 × PCR buffer [Tris-Cl, $(NH_4)_2SO_4$, 15 mM $MgCl_2$, pH 8.7], 1.5 mM $MgCl_2$, 200 μM dNTPs without dTTP, 160 μM dTTP, 40 μM dUTP, 0.75 U HotStar *Taq* DNA polymerase (5 U/μL; Qiagen, Valencia, CA, USA), 1 μL 10 μM primer mixtures (each primer), and 25 ng genomic DNA. Incorporation of the dUTP allowed for the amplified DNA to be enzymatically sheared by uracil N-glycosylase (UNG, InterScience, Troy, NY, USA) to produce a DNA size of approximately 50100 bases, optimal for hybridization to the oligonucleotides on the microarray (see below). Genomic DNA and PCR master mixture were transferred into ABI 384well reaction plates (Applied Biosystems, Foster City, CA, USA) using a Biomek FX robot (Beckman Coulter, USA). PCR reactions were performed in a GeneAmp PCR System 9700 ThermoCycler (Applied Biosystems, USA). PCRs were initiated by a 15 min polymerase activation step at 95°C and completed by a final 10 min extension step at 72°C. The PCR cycles were as follows: 35 cycles of 30 s denaturation at 95°C, 30 s annealing at 58°C, and 50 s extension at 72°C.

For the second experiment, in order to increase the efficiency of PCR, we designed 50x 5' linker PCR primer pairs (Additional Table 4 online) based on a Tm of 65°C ± 7°C and performed 50-plex PCR in one single reaction per sample. Each new PCR primer had a common linker sequence designed at its 5' end ($^{5'}$TACGACTCACTTAGGGAG$^{3'}$ for each of the left hand PCR primers / $^{5'}$CGATGTAGGTGACACTAG$^{3'}$ for each of

the right hand PCR primers). The 3' ends of the primers were chosen to have non-complementary bases with respect to each other (i.e., all primers ended with on or two A bases), in order to reduce the probability of primer interactions and primer-dimer formation. All primers were computationally tested against the human genome and found to amplify single product. The new amplicon sequences were located within the amplicon sequences from the original primer pairs. The multiplex PCR was carried out in a 25 μL reaction containing 20 nM (final) of each primer plus 20 nM of left and right linker-only primers (left linker: $^{5'}$TACGACTCACTTAGGGAG$^{3'}$ / right linker: $^{5'}$CGATGTAGGTGACACTAG$^{3'}$), 200 μM dNTPs without dTTP, 160 μM dTTP, 40 μM dUTP, 6 units of HotStar *Taq* DNA polymerase (5 U/μL; Qiagen, USA), 1.5 mM MgCl2 in 1x PCR reaction buffer [100 mM Tris-HCl, 50 mM KCl, 100 μg/mL Gelatin, pH 8.3] with 5 ng of genomic DNA. PCR was performed using a MJR PTC 200 ThermoCycler (MJ Research, Waltham, MA, USA). PCR was initiated by a 15 min polymerase activation step at 95°C and completed by a final 3 min extension step at 72°C. The reaction procedure consisted of 40 cycles of denaturation at 95°C for 40 s, primer annealing at 55°C for 2 min and one ramping-up step from 55°C to 70°C for 2.5 min (0.1°C/s) (Wang et al., 2005).

Aliquots of PCR products were visualized with Gel Red fluorescent nucleic acid dye (Biotium, Hayward, CA, USA) staining under ultraviolet (UV) illumination on a 2% agarose gel, following electrophoresis in 0.5x Tris-borate EDTA (TBE) buffer. The 7 subgroup multiplex PCR products were pooled for each individual Coriell sample and precipitated by adding 2.5 volumes of ice-cold 100% ethanol and 0.25 volumes of 10 M ammonium acetate solution. After precipitation at -20°C overnight, the mixture was centrifuged at 20,800 g at 4°C for 20 min. The supernatant was carefully removed, and the DNA pellet was washed with 400 μL of ice-cold 70% ethanol. The DNA pellet was then dissolved in 15 μL pure water. 10 μL of this DNA (or 10 μL of unpurified 50plex PCR products; amplified to a concentration of approximately 300 400 ng/μL.) were then fragmented by 1 UuracilNglycosylase (UNG; Inter Science Inc., Troy, NY, USA) and unincorporated dNTPs were simultaneously inactivated by digestion with 1 U shrimp alkaline phosphatase (SAP; Amersham Biosciences / GE Healthcare, USA) for 15 min at 37°C, in a 20 μL reaction mixture containing 2 μL 10 digestion buffer [0.5 M Tris-HCl, 0.2 M $(HN_4)_2SO_4$, pH9.0], followed by enzyme inactivation for 10 min at 95°C.

3.4.4 Microarray-based Minisequencing: Arrayed Primer Extension (APEX)

The APEX reaction was performed in a total volume of 40 μL by the addition of 17 μL fragmented DNA template, 1 mL of 2 pmol/μL Npg1-positive control template oligonucleotide, 1.25 μM of each fluorescently labeled dideoxynucleotide triphosphate (Texas Red-ddATP, Cy3-ddCTP, Cy5-ddGTP, R110-ddUTP; Perkin Elmer Life Sciences, Boston, MA, USA), 5 U Thermo Sequenase^{TM} DNA polymerase (Amersham Biosciences / GE Healthcare, USA) diluted in its dilution buffer, 2 Thermo Sequenase reaction buffer [10, 260 mM Tris-HCl, 65 mM $MgCl_2$, pH 9.5]. The reaction mixture was applied to the grid of APEX and ASO-APEX probes previously printed on the CodeLink slide that had been washed two times in 95°C pure water and placed on a Thermo Hybaid HyPro20 incubation plate (Thermo Electron, Waltham, MA, USA) set at 58°C. The reaction mixture was covered with a small piece of Parafilm^{TM}, and the APEX reaction allowed to proceed at 58°C with agitation (setting 1) for 20 min. Following the incubation period, slides were washed with 95°C water to remove the template DNA, enzyme, and excess ddNTPs. Further washing in 0.3% Alconox (Alconox Inc., White Plains, NY, USA) and 95°C pure water ensured low background on the array images.

3.4.5 DNA Sequencing

As described in the main paper, we directly sequenced three SNP loci in three independent samples: 1. sample 192 (NA18502) at SNP rs3776720; 2. sample 101 (NA18621) at SNP rs12472674; 3. sample 228 (NA19210) at SNP rs4739199. We performed three single-plex PCR reactions using primer pairs from the first experimental design and methods (Additional Table 1 online) to obtain the DNA fragments including the SNP sites on these three Coriell DNA samples. PCR primers pairs used were: 1. rs3776720 sense $^{5'}$GGC CAA GGA AAA GAA ATG AAT CTG CT$^{3'}$, anti-sense $^{5'}$AAC TTT AGT GCA GGA TTT GCC ATC CA$^{3'}$ - PCR amplicon size of 389 bp; 2. rs12472674 sense $^{5'}$TAA AAT CCA ATC AGG CCA ACT GTT CA$^{3'}$, anti-sense $^{5'}$TCA ATG CCA TTA TAT GTG CCA GCC A$^{3'}$ - PCR amplicon size of 388 bp; 3. rs4739199 sense $^{5'}$TCC AGC CAG CAA AAG ATC CTC AAA$^{3'}$, anti-sense $^{5'}$TCA AGC ACA TGT TAC CAG TTT CCC AA$^{3'}$ - PCR amplicon size of 587 bp. PCR products were purified using a QIAquick PCR Purification Kit (Qiagen, Valencia, CA, USA) according to the manufacture's instructions. DNA sequenc-

ing reactions were performed by the Nucleic Acid Protein Service Unit (http://www.michaelsmith.ubc.ca/services/NAPS/) at the University of British Columbia (Vancouver, BC, Canada). For each amplicon, sense and anti-sense PCR primers were used as sequencing primers.

3.4.6 Microarray Imaging and Spot Intensity Calculation

Slide microarrays were imaged using an *arrayWoRx*^{e} Auto Biochip Reader (Applied Precision, LLC, Issaquah, WA, USA), fitted with the following filter sets: 1. A488 - Ex. 480/15x - Em. 530/40 (R110 dye); 2. Cy3 (narrowband) - Ex. 546/11 - Em. HQ570/10m (Cy3); 3. Texas Red - Ex. 602/13 - Em. 631/23 (Texas Red); 4. Cy5 - Ex. 635/20 - Em. 685/40 (Cy5) (Chroma Technology,Rockingham, VT, USA). Exposure times for each dye were set up to give approximately 60-70% pixel saturation for selected Npg1 positive control probe spots. Resolution of the imager was set to 10 μm. Four 16-bit TIFF files for each array were obtained (one from each channel) and these were imported into SNP Chart, a data management and visualization tool for array-based genotyping by primer extension from multiple probes (http://www.snpchart.ca) (Tebbutt et al., 2005). This software generates visual patterns of spot intensity values, from multiple channels across a multiple probe set specific for a given SNP, allowing easy calling of the genotype. All the images were gridded in SNP Chart by manually selecting four pre-defined spots that, combined with knowledge of the layout of the grid, allows SNP Chart to locate every spot (Tebbutt et al., 2005). Spot segmentation and background subtraction were based on hybrid segmentation algorithms previously published by our laboratory (Abbaspour et al. (2006) and Abbaspour, Abugharbieh, Podder, and Tebbutt (Abbaspour et al.)). Spot intensity values were normalized by setting the 40 Npg1 positive control spots, widely distributed across each array grid, to an average value of 20,000 units per channel, with the exported normalized intensity value calculated from the scale factor x median signal) (Tebbutt et al., 2006).

3.4.7 Genotyping - Manual Calling

Manual genotype calling within SNP Chart was carried out as previously described (Tebbutt et al. (2004), Tebbutt et al. (2005) and Tebbutt et al. (2006)).

3.4.8 Genotyping - Automated Calling using MACGT

The training set for MACGT (multi-dimensional automated clustering genotyping tool) (Walley et al., 2006) was selected by manually inspecting SNP Charts for each of the SNPs across some of the 287 samples. For the 50 SNPs, up to ten high-quality charts were chosen as 'prototypes' (Tebbutt et al., 2005) for each genotype. All prototype data were exported from SNP Chart into a format readable by MACGT. MACGT was run on just the training data, and the clusters for each SNP were manually inspected to ensure there where no errors in the training set. Genotyping was performed by MACGT using the parameters NORMALIZE_GROUP_OF_4=1, GROUP_OF_4_MEAN_CUTOFF=10,
PATCH_GROUPS_OF_4=1, DROP_NNS=1. A 'fit' statistical cut-off of 0.001 was used to identify poor quality genotypes as non-calls (NNs) (Walley et al., 2006). Any SNP or sample with a high rate of NNs was subject to further inspection. We identified nine SNPs that the PCR assay performed poorly on and which MACGT could not confidently score, although manual inspection of SNP Charts did show that the assays were somewhat successful, albeit non-reproducibly. The final training set for the 41 SNPs was made up of 519 genotypes (Additional Table 3 online). All NNs were inspected within SNP Chart and manually called if possible. The final genotypes from MACGT and from those manually called were combined, and compared to the validated genotypes from HapMap using a Microsoft Excel macro (Additional Table 3 online).

3.4.9 Genotyping - Automated Calling using Simple LDA with Dynamic Variable Selection

Detailed descriptions of the algorithms used in simple linear discriminant analysis (LDA) with dynamic variable selection have previously been published by our laboratory (Podder et al., 2006). A brief descriptive example follows, using the data structure for SNP rs12466929 and DNA sample 101 (Coriell NA18621 - genotype AA - Additional Table 5 online).

Ideally, for variable construction, each genotype call could be based on just one of the four sets of probes: (1) APEX_LEFT; (2) APEX_RIGHT; (3) ASO_1LEFT and ASO_2LEFT; and (4) ASO_1RIGHT and ASO_2RIGHT (Additional Table 5 online). Considering the underlying chemistry, we have developed four sets of classifiers, named: APEX.L, APEX.R, ASO.L and ASO.R. Each of these classifiers consists of a pair of explanatory variables, generically denoted by X and Y, corresponding to two candidate alleles in

the SNP position (Additional Table 6 online). In Additional Table 5 online, for example, X and Y correspond to the A and G alleles, respectively. Since there are five realizations (replicates) for each of the two entries in each classifier, we summarized the information for each allele, by taking a robust average: median of the relevant signals from five spots, for each of the classifiers. From the example data in Additional Table 5 online, the values of the variables for the classifier APEX.L are
APEX.XL = median (1394, 1148, 597, 1106, 1504) = 1148, and
APEX.YL = median (29, 27, 43, 27, 32) = 29, and so on, as summarized in Additional Table 6 online. In our subsequent analyses, we have considered different combinations of the above mentioned classifiers.

Our automated genotype calling algorithm is based on the simple linear discriminant analysis (LDA), using dynamic variable selection as a special criteria for various classifiers related to multiple probes. LDA is a supervised learning technique which requires a valid training set in order to build the classification (genotyping) model for each SNP. For the complete set of 287 HapMap samples, our dynamic variable LDA-based genotype calling algorithm used the same training set as used by MACGT above (i.e., 519 genotypes across the 41 SNPs - Additional Table 3 online) and predicted the genotypes for the remainder of the samples.

For LDA analysis of the 50-plex PCR chemistry, performed on a subset of 50 HapMap samples which were chosen randomly out of the original 287 samples, we selected prototypes to build a new training set using MACGT clusters, verifying the chosen cases with SNP Chart. We considered two different training sets, one with a small number of prototypes (at most 3 to 4 prototypes in each class) and the other with a minimal number of prototypes (at most 2 prototypes in each class) for each SNP. The two different training sets yielded different performances for the respective test data sets (see Table 3.4, 3.5).

For automated genotype calling, we started our analysis by fitting the simple LDA-based genotype model using each classifier separately, and then comparing the predicted genotypes with the validated genotypes. Subsequently, we applied our dynamic-variable LDA-based genotyping model on different combinations of the four classifiers.

For automated genotype calling, we followed the same steps as described in Podder et al. (2006).

Bibliography

Abbaspour, M., R. Abugharbieh, M. Podder, and S. Tebbutt. Fully-Automated Analysis of Multi-Resolution Four-Channel Microarray Genotyping Data. *Proc. of SPIE Vol 6144*, 61443M–1.

Abbaspour, M., R. Abugharbieh, M. Podder, B. Tripp, and S. Tebbutt (2006). Hybrid Spot Segmentation in Four-Channel Microarray Genotyping Image Data. *Signal Processing and Information Technology, 2006 IEEE International Symposium on*, 11–16.

Brownie, J., S. Shawcross, J. Theaker, D. Whitcombe, R. Ferrie, C. Newton, and S. Little (1997). The elimination of primer-dimer accumulation in PCR. *Nucleic Acids Research 25*(16), 3235–3241.

Cremers, F., W. Kimberling, M. Külm, A. de Brouwer, E. van Wijk, H. te Brinke, C. Cremers, L. Hoefsloot, S. Banfi, F. Simonelli, et al. (2007). Development of a genotyping microarray for Usher syndrome. *British Medical Journal 44*(2), 153–160.

Dawson, E., G. Abecasis, S. Bumpstead, Y. Chen, S. Hunt, D. Beare, J. Pabial, T. Dibling, E. Tinsley, S. Kirby, et al. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature 418*(6897), 544–548.

Feuk, L., A. Carson, S. Scherer, et al. (2006). Structural variation in the human genome. *Nat Rev Genet 7*(2), 85–97.

Gemignani, F., C. Perra, S. Landi, F. Canzian, A. Kurg, N. Tonisson, R. Galanello, A. Cao, A. Metspalu, and G. Romeo (2002). Reliable Detection of {beta}-Thalassemia and G6PD Mutations by a DNA Microarray. *Clinical Chemistry 48*(11), 2051–2054.

Hirschhorn, J., P. Sklar, K. Lindblad-Toh, Y. Lim, M. Ruiz-Gutierrez, S. Bolk, B. Langhorst, S. Schaffner, E. Winchester, and E. Lander (2000).

SBE-TAGS: An array-based method for efficient single-nucleotide polymorphism genotyping. *Proceedings of the National Academy of Sciences*, 210394597.

Jaakson, K., J. Zernant, M. Kulm, A. Hutchinson, N. Tonisson, D. Glavac, M. Ravnik-Glavac, M. Hawlina, M. Meltzer, R. Caruso, et al. (2003). Genotyping microarray (gene chip) for the ABCR (ABCA4) gene. *Hum Mutat 22*(5), 395–403.

Janssens, A., M. Pardo, E. Steyerberg, and C. van Duijn (2004). Revisiting the Clinical Validity of Multiplex Genetic Testing in Complex Diseases. *The American Journal of Human Genetics 74*(3), 585–588.

Kennedy, G., H. Matsuzaki, S. Dong, W. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, et al. (2003). Large-scale genotyping of complex DNA. *Nature Biotechnology 21*(10), 1233–1237.

Kurg, A., N. Tonisson, I. Georgiou, J. Shumaker, J. Tollett, and A. Metspalu (2000). Arrayed Primer Extension: Solid-Phase Four-Color DNA Resequencing and Mutation Detection Technology. *Genetic Testing 4*(1), 1–7.

Lahermo, P., U. Liljedahl, G. Alnaes, T. Axelsson, A. Brookes, P. Ellonen, P. Groop, C. Halldén, D. Holmberg, K. Holmberg, et al. (2006). A quality assessment survey of SNP genotyping laboratories. *Hum Mutat 27*(7), 711–714.

Oliphant, A., D. Barker, J. Stuelpnagel, and M. Chee (2002). BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques 32*, S56–S61.

Pastinen, T., M. Raitio, K. Lindroos, P. Tainola, L. Peltonen, and A. Syvanen (2000). A System for Specific, High-throughput Genotyping by Allele-specific Primer Extension on Microarrays. *Genome Research 10*(7), 1031.

Podder, M., W. Welch, R. Zamar, and S. Tebbutt (2006). Dynamic variable selection in SNP genotype autocalling from APEX microarray data. *BMC Bioinformatics 7*(1), 521.

Quinlan, A. and G. Marth (2007). Primer-site SNPs mask mutations. *Nature Methods 4*(3), 192.

Shumaker, J., A. Metspalu, and C. Caskey (1996). Mutation detection by solid phase primer extension. *Hum Mutat 7*(4), 346–54.

Smemo, S. and J. Borevitz (2007). Redundancy in Genotyping Arrays. *PLoS ONE 2*(3), e287.

Steemers, F., K. Gunderson, I. Illumina, and C. San Diego (2007). Whole genome genotyping technologies on the BeadArray platform. *Biotechnol J 2*(1), 41–49.

Tebbutt, S., J. He, K. Burkett, J. Ruan, I. Opushnyev, B. Tripp, J. Zeznik, C. Abara, C. Nelson, and K. Walley (2004). Microarray genotyping resource to determine population stratification in genetic association studies of complex disease. *Biotechniques 37*(6), 977–85.

Tebbutt, S., G. Mercer, R. Do, B. Tripp, A. Wong, and J. Ruan (2006). Deoxynucleotides can replace dideoxynucleotides in minisequencing by arrayed primer extension. *BioTechniques 40*(3), 331–338.

Tebbutt, S., I. Opushnyev, B. Tripp, A. Kassamali, W. Alexander, and M. Andersen (2005). SNP Chart: an integrated platform for visualization and interpretation of microarray genotyping data.

Tõnisson, N., A. Kurg, K. Kaasik, E. Lõhmussaar, and A. Metspalu (2000). Unravelling Genetic Data by Arrayed Primer Extension. *Clinical Chemistry and Laboratory Medicine 38*(2), 165–170.

Tonisson, N., J. Zernant, A. Kurg, H. Pavel, G. Slavin, H. Roomere, A. Meiel, P. Hainaut, and A. Metspalu (2002). Evaluating the arrayed primer extension resequencing assay of TP53 tumor suppressor gene. *Proceedings of the National Academy of Sciences 99*(8), 5503–5508.

Walley, D., B. Tripp, Y. Song, K. Walley, and S. Tebbutt (2006). MACGT: multi-dimensional automated clustering genotyping tool for analysis of microarray-based mini-sequencing data.

Wang, D., J. Fan, C. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, et al. (1998). Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science 280*(5366), 1077–1082.

Wang, H., M. Luo, I. Tereshchenko, D. Frikker, X. Cui, J. Li, G. Hu, Y. Chu, M. Azaro, Y. Lin, et al. (2005). A genotyping system capable of simultaneously analyzing >1000 single nucleotide polymorphisms in a haploid genome. *Genome Research 15*(2), 276–283.

Yang, Q., M. Khoury, L. Botto, J. Friedman, and W. Flanders (2003). Improving the Prediction of Complex Diseases by Testing for Multiple Disease-Susceptibility Genes. *The American Journal of Human Genetics 72*(3), 636–649.

Zernant, J., M. Kulm, S. Dharmaraj, A. den Hollander, I. Perrault, M. Preising, B. Lorenz, J. Kaplan, F. Cremers, I. Maumenee, et al. (2005). Genotyping Microarray (Disease Chip) for Leber Congenital Amaurosis: Detection of Modifier Alleles. *Investigative Ophthalmology & Visual Science 46*(9), 3052–3059.

Chapter 4

Robust Dynamic Variable Selection

4.1 Introduction

4.1.1 Genetics Background

The success of the Human Genome Project and the International HapMap Project is inspiring much research on the effects of genes on various complex diseases, e.g., asthma-allergy, heart disease and sepsis. Genetic variability has been a primary focus for many of these studies. SNPs are the most abundant form (90%) of genetic variability, and are defined as DNA sequence variations that occur when a single base (A, C, G or T) in the genome sequence is altered. Different combinations of SNPs in single or multiple genes are partly responsible for disease susceptibility, the variability in how individuals respond to illness and to medical therapy, and for whether they develop adverse drug responses.

The determination of a given person's base sequence at a specific SNP site is called genotyping. Many medium to high throughput genotyping techniques have been developed and tested on various populations. Affymetrix GeneChips (Kennedy et al., 2003) and Illumina's bead-array system (Oliphant et al. (2002), Fan et al. (2006)) are regarded as the leading technologies in this field and are optimally designed to analyze thousands if not hundreds of thousands of SNPs simultaneously.

One challenge of the Human Genome Project is how to transfer research-based genetic knowledge to the benefit of society at large. In the field of biomedical research this translates into how to apply the knowledge obtained from SNP-related research to medical and clinical settings. Successful translation requires technological advancement and real-time innovation. In clini-

A version of this chapter will be submitted for publication. Authors: Mohua Podder, William J. Welch, Ruben H. Zamar and Scott J. Tebbutt.

cal settings, a genotyping technology should be designed to classify hundreds of SNPs simultaneously for a patient in a rapid, accurate, robust and cost effective manner. The above mentioned technologies are mostly suitable for pure research discovery; they are not optimally designed for rapid genetic diagnosis of an individual patient. For example, a single intensive care unit (ICU) patient with severe sepsis might require rapid genetic diagnosis within one hour in order to receive optimal treatment based on his or her underlying genetic variability. Such diagnostics cannot be provided through existing research-based technologies, since their underlying chemistries require more than 24 hours for completion.

The James Hogg iCAPTURE Centre for Cardiovascular and Pulmonary Research is a translational research unit, with scientists and medical practitioners working in the field of gene-environment interactions and their combined effects on various complex diseases involving heart, lung and blood vessels. Tebbutt's laboratory is engaged in developing a microarray genotyping assay based on arrayed primer extension, abbreviated APEX (Shumaker et al. (1996); Kurg et al. (2000); Pastinen et al. (2000)). APEX is a mini-sequencing assay where the array chemistry takes only fifteen to twenty minutes to complete, allowing rapid genotyping of hundreds of SNPs simultaneously for an individual patient.

4.1.2 Redundant Microarray Genotyping Platform using APEX Probe Chemistry

Tebbut's genotyping array chip design is based on a robust and redundant probe chemistry platform. The technology involves multiple probes: classical APEX probes and allele-specific APEX (ASO) probes for both DNA strands (generically denoted as left and right strand) corresponding to a single SNP (Tebbutt et al., 2004). Each probe has several replicates (two to five) and each probe-replicate in this system generates signals for all four channels (A, C, G, and T). According to the underlying chemistry, some of these signals are considered as foreground signals and some are considered as background signals. For each of the two DNA strands, each individual probe provides signals corresponding to the two possible alleles (generically denoted as X and Y allele) at a specific SNP site.

For a specific SNP, the multiple probes generate four separate pairs of explanatory variables based on the expected foreground signals: (ASO.XL, ASO.YL); (ASO.XR, ASO.YR); (APEX.XL, APEX.YL); and (APEX.XR, APEX.YR). Each set has two signals, corresponding to the X and Y alleles, respectively. The four sets are all combinations of APEX versus ASO

chemistries and left versus right strands. Details may be found in Podder et al. (2006). We will call the four sets of signals ASO.L, ASO.R, APEX.L and APEX.R, respectively. The main objective of this study is to classify the SNP as one of the three genotype classes, XX, YY and XY, based on the these four independent probe sets (pairs of explanatory variables). This can be done manually for each SNP and individual patient using a graphical display called SNP Chart (Tebbutt et al., 2005). However, in any SNP related genetic study, hundreds of SNPs are analyzed simultaneously, which necessitates automated genotype calling. Such an algorithm should be simple, fast, and robust to poor values of the explanatory variables. We will build multiple classifiers in this article, each based on one of the ASO.L, ASO.R, APEX.L or APEX.R variable sets.

To build and test models for this APEX-based microarray genotyping platform, we have two independent data sets: 32 Coriell DNA samples (http://coriell.umdnj.edu/); and 270 SIRS (systematic inflammatory response syndrome) DNA samples from the ICU of St. Paul's hospital. These two sets will be called Coriell and SIRS, respectively. We will use SIRS as training data and Coriell as test data. For each sample (i.e., patient), there are about 100 SNPs on the microarray chip, which are genotyped simultaneously.

Ideally, any one of the probe sets should provide information to genotype a particular SNP and sample. In practice, however, SNP by SNP and sample by sample, a probe set may fail. In other words, some probe set(s) may provide useful information for a particular SNP and sample, whereas the same probe(s) might give misleading information for another SNP or sample. This complexity will be clearer if we examine the data for a few critical SNPs.

SNP rs1360590 and subject 12. Figure 4.1 shows data for SNP rs1360590. The two alleles generically called X and Y are A and G, here, and the known AA, AG, or GG genotypes are shown by different symbols in the four panels corresponding to the four probe sets. Sample 12 is known to have genotype AA for this SNP, but suppose we treat sample 12 as an unknown test case to be classified based on the remaining samples. For the ASO.L probe set, data for the three genotypes overlap considerably and there are many outliers. ASO.L does not provide good information for this SNP in general. Similar comments may be made about the other three probe sets here, though the data are more informative in general. For classifying sample 12 specifically, we see that the ASO.R and APEX.L probe sets are not informative as this sample falls between the AA and AG clusters evident.

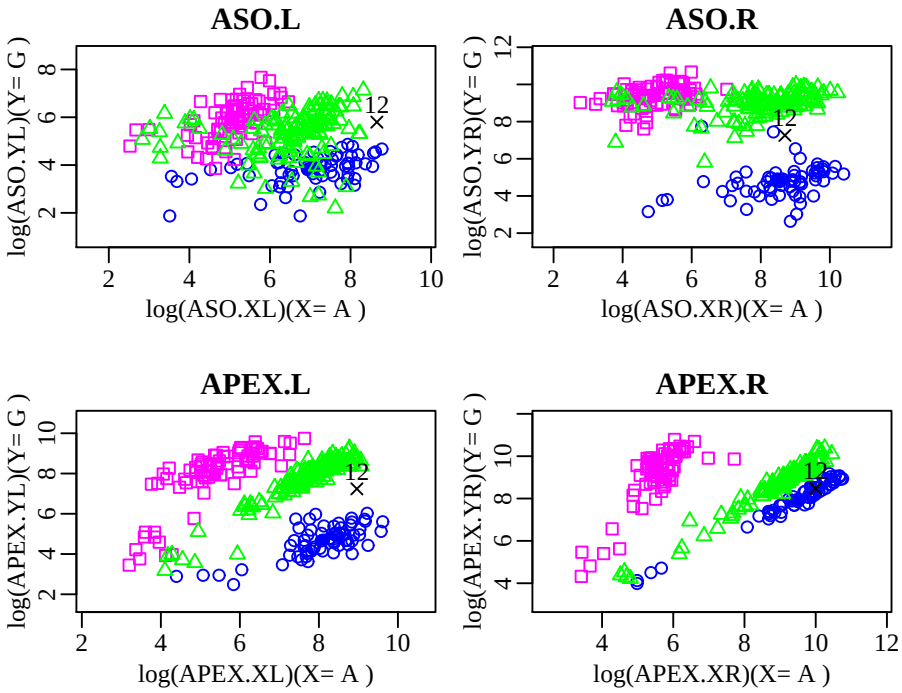

Figure 4.1: Data from the four probe sets are shown in the four panels for SNP rs1360590 (alleles A/G). The AA, AG, and GG genotypes are denoted by circles, triangles, and squares, respectively. Coriell sample 12 is denoted by ×; its genotype is AA.

APEX.R is the best behaving probe chemistry here, placing subject 12 well within the AA data. We next look at a different SNP where the situation is completely different even for the same sample.

SNP rs1981278 and subject 12. Figure 4.2 shows analogous plots for classifying subject 12 but for SNP rs1981278. Here the three possible genotype classes are CC, CT and TT, and sample 12 is TT. The four probe

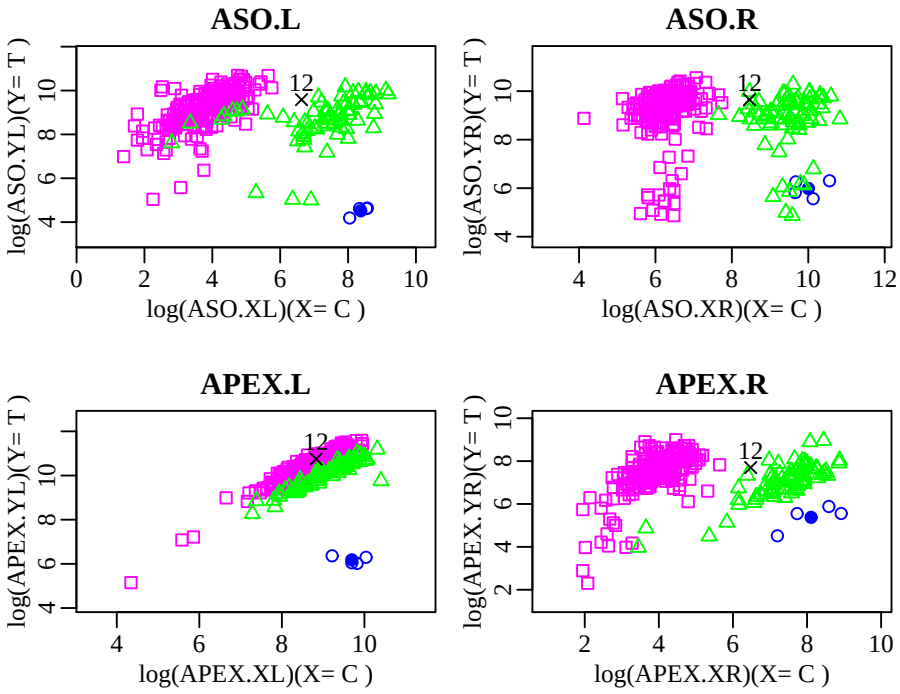

Figure 4.2: Data from the four probe sets are shown in the four panels for SNP rs1981278 (alleles C/T). The CC, CT, and TT genotypes are denoted by circles, triangles and squares, respectively. Coriell sample 12 is denoted by ×; its genotype is TT.

chemistries again give some overlapping clusters and some outliers. Sample 12 falls on the edge of a wrong class (CT) in the ASO.L, ASO.R, and APEX.R signal spaces. The APEX.L chemistry works best here, placing subject 12 in the correct class (TT).

4.1.3 Implications for Statistical Modeling

The APEX-based genotyping platform is deliberately redundant, anticipating the occasional failure of one or more probe-set chemistries. Conventional variable selection would use the training data to select a fixed set of variables optimal in some sense for a particular classification method from the eight available in the four probe sets. The resulting classifier would be applied to all test cases.

From the illustrative examples in Section 4.1.2 it is clear, however, that variable selection needs to be dynamic to exploit the redundancy in the probe chemistry. Our approach is to design *four* SNP-specific classification models, each based on the variables in one probe set. At the prediction stage, these base classifiers are combined with weights according to the measures of confidence for the four chemistries *specific to that test sample.* Details of this modeling approach are described in Section 4.2.

The examples in Section 4.1.2 also illustrate that outliers are fairly frequent. Indeed, the anticipation of outliers—from failed chemistries—is the reason for redundancy in the data. Thus, there is a need for robustness in statistical modeling, particularly for validity of the measures of confidence used to combine classifiers.

For the base classifiers in the dynamic ensemble we start with linear discriminant analysis (LDA). It is straightforward to use robust estimates of the means and covariances (Croux and Dehon, 2001) required for LDA . Paradoxically, as we illustrate in Section 4.6.2, training the base classifiers in a robust way leads to *less* robustness in assessing the confidence of correct classification. For this reason, the base LDA classifiers in Section 4.2 are not trained in a robust way.

For robustness to outliers, it is necessary to apply robust methods *at both the training and prediction stages.* In other words, it is essential also to model the possible presence of outliers in the signals associated with a new sample when classifying that sample. A mixture model allowing for "good" and "failed" signal distributions is developed in Section 4.3 for this purpose. In combination with robust training of the models via robust estimates of location and scale, the entire training-prediction modeling process is made resistant to outliers. The mixture model leads to a better estimate of confidence for each base classifier, and hence better weights in the dynamic ensemble.

4.1.4 Outline of the article

Section 4.2 introduces the methodology for building a dynamic ensemble of classifiers. The underlying base classifiers are made robust in Section 4.3 through a mixture model, which allows a "good signal" distribution to be contaminated by outliers from a failed chemistry. In Section 4.4 we revisit the examples of Section 4.1.2 to illustrate the dynamic ensemble of classifiers, in both non-robust and robust forms. Section 4.5 gives results for the two methods in terms of overall classification accuracy when genotyping the Coriell samples using the SIRS data for model training. Section 4.6 provides some insight via simulation into the performance of the ensemble methods, in particular how robustness at only the training stage may be harmful. Finally, Section 4.7 makes some concluding remarks.

4.2 Dynamic Ensemble of Models

In Section 4.1.2 we discussed four independent pairs of explanatory variables. A pair of variables for a particular probe, generically denoted X and Y, summarize the signal intensities for the two candidate alleles at a particular SNP position. For each pair of variables and each SNP we apply Fisher's linear discriminant analysis (LDA) (Fisher (1936); Hastie et al. (2001), Section 4.3). This is implemented in R via the function lda in the library MASS with the prior class probabilities estimated by the training class frequencies. The parameters: μ_c and Σ are estimated here using the class-specific sample mean vectors and the common sample covariance matrix. In this way, for each SNP, four base classifiers are available, leading to four sets of posterior probabilities for any test subject. These probabilities are denoted by $P_c^{(i)}$, where $i = 1, 2, 3, 4$ indexes the four base classifiers (1 =ASO.L, 2 =ASO.R, 3 =APEX.L and 4 =APEX.R) and $c \in \mathcal{C} = \{\text{XX,XY,YY}\}$ indexes the three possible genotype classes. They are set out in Table 4.1, where, for example, $P_{\text{XX}}^{(1)}$ is the posterior probability for the XX genotype using the base classifier ASO.L.

A single probability for each genotype class, P_c, can now be obtained as a weighted average of the posterior probabilities $P_c^{(1)}, \ldots, P_c^{(4)}$, with the weights chosen dynamically for each SNP and test subject. These weights are estimated using individual test sample data. Ideally, the weight assigned to a base classifier should reflect its degree of "confidence". A confident base classifier assigns a large probability to one of the three possible classes and low probabilities to the other two. Entropy is a measure that captures this

Table 4.1: Posterior probabilities for the three genotypes from four LDA classifiers

Classifiers/Classes	XX	XY	YY
LDA(ASO.L)	$P_{\text{XX}}^{(1)}$	$P_{\text{XY}}^{(1)}$	$P_{\text{YY}}^{(1)}$
LDA(ASO.R)	$P_{\text{XX}}^{(2)}$	$P_{\text{XY}}^{(2)}$	$P_{\text{YY}}^{(2)}$
LDA(APEX.L)	$P_{\text{XX}}^{(3)}$	$P_{\text{XY}}^{(3)}$	$P_{\text{YY}}^{(3)}$
LDA(APEX.R)	$P_{\text{XX}}^{(4)}$	$P_{\text{XY}}^{(4)}$	$P_{\text{YY}}^{(4)}$

property, but it has to be changed such that larger entropy leads to smaller weight (confidence). Hence, we define

$$E_i = -\log\left(\frac{1}{3}\right) - \left[-\sum_{c\in\mathcal{C}} P_c^{(i)} \log(P_c^{(i)})\right], \tag{4.1}$$

where $-\sum_{c\in\mathcal{C}} P_c^{(i)} \log(P_c^{(i)})$ is the entropy of the probability distribution $P_c^{(i)}$ over $c \in \mathcal{C}$ and $-\log(1/3)$ is the corresponding maximum entropy. Note that E_i in (4.1) is minimized—equal to zero—when $P_{XX}^{(i)} = P_{XY}^{(i)} = P_{YY}^{(i)} = 1/3$, and E_i is maximized—equal to $-\log(1/3)$—when one of the three probabilities is 1 and the other two are 0.

The weights for the four classifiers are obtained by normalizing the E_i so that their sum is 1, i.e.,

$$W_i = \frac{E_i}{\sum_{i=1}^{4} E_i}.$$

Note that the weights will vary from one test sample to another as they depend on the sample/SNP specific probabilities in Table 4.1. Finally, for each $c \in \mathcal{C}$,

$$P_c = \sum_{i=1}^{4} W_i P_c^{(i)}.$$

Two applications of this modeling approach have been reported in Podder et al. (2006); Podder et al. (2008). We would like to mention at this point that other approaches (e.g., logistic regression, quadratic discriminant analysis, classification trees or support vector machines) could also be used

for the base classifier models. We choose LDA because it works well for the given application, has a simple interpretation, has been widely used in similar classification problems (Guo et al., 2007), and it is easy to robustify.

4.3 Dynamic Ensembles of Robust Mixture Models

4.3.1 Robust mixture model

We now present a robustification of the LDA-based classifiers described in Section 4.2, recognizing that one or more of the redundant chemistries may fail and produce noninformative outliers. Instead of modeling the signals X and Y by a bivariate normal distribution (as in LDA) we use the mixture density

$$f(x, y|C = c) = (1 - \lambda)g_c(x, y) + \lambda h(x, y), \tag{4.2}$$

where g_c is an informative class-specific density and h is a noninformative (failed-chemistry) density. Specifically, we take g_c as $N(\mu_c, \Sigma)$ and h as uniform over the data range. The informative and noninformative distributions are mixed via a user-adjustable weight parameter, $\lambda \in (0, 1)$, which represents the proportion of times the chemistry is expected to fail. Now, μ_c, Σ are global parameters estimated using the training data and λ is a global parameter specified by the user. We will see that the inclusion of the noninformative background distribution has the crucial effect of producing noninformative (high entropy) posterior probabilities whenever the test case signals (x, y) seem to come from a failed chemistry. In this way, we formally acknowledge the deliberate redundancy in the chemistry and ultimately get valid robust genotype probability estimates.

Let (x_i, y_i) be the intensity signals corresponding to one of the four chemistries ($i = 1, \ldots, 4$). By Bayes rule, the posterior probability of class c given by the mixture-model classifier for chemistry i is

$$\begin{aligned} P_c^{(i)} = P(C = c|x_i, y_i) &= \frac{p_c f(x_i, y_i|C = c)}{\sum_{c'} p_{c'} f(x_i, y_i|C = c')} \\ &= \frac{p_c[(1 - \lambda)g_c(x_i, y_i) + \lambda h(x_i, y_i)]}{\sum_{c'} p_{c'}[(1 - \lambda)g_{c'}(x_i, y_i) + \lambda h(x_i, y_i)]}, \end{aligned} \tag{4.3}$$

where p_c is the prior probability of class c estimated from the training data. For each class c, four posterior probabilities, $P_c^{(i)}$ for $i = 1, \ldots, 4$, are estimated from the four base classifiers. These sets of probabilities will be

combined dynamically (Section 4.3.2) to give one final probability for each class.

We note from (4.3) that when (x_i, y_i) is an outlier with respect to all classes, $g_c(x_i, y_i)$ is very small for all c and so λ almost cancels. On the other hand, when (x_i, y_i) is not an outlier, λ should be small enough so that it does not unduly affect the posterior probability calculation. The parameter λ can be estimated using the training data or simply set equal to a small value. In this article we take $\lambda = 0.1$ for the SNP genotyping data application and also for the simulation study. We have observed (see Figure 4.7 in Section 4.6.2) that for small values of λ (between 0.005 and 0.15), the overall performance of the classification model shows little change. Which indicates that our genotyping model is robust with respect to the small values of lambda.

The remaining model parameters (μ_c, Σ) are robustly estimated using the training data. In the first step, the location parameters μ_c in $f(x_i, y_i|C = c)$ in (4.2) are estimated by applying the robust fast minimum covariance determinant (MCD) estimator proposed by Rousseeuw and Van Driessen (1999) and implemented in the function covMcd in the R library Robustbase. In the next step, a common estimate of the dispersion matrix Σ of the three g_c densities ($c \in \{\text{XX,XY,YY}\}$) is obtained by a second application of the robust MCD estimator after centering the data for each class with respect to $\hat{\mu}_c$ from the first step.

4.3.2 Dynamic Ensemble based on Robust Mixture Models

For the test set, posterior probabilities corresponding to all three genotype classes are calculated for all four base classifiers (see Table 4.2). As before, the four classifiers are combined using an entropy-based weighting scheme. If any one of the four redundant chemistries generates signals that appear to come from the background distribution (h) for a particular test sample, the corresponding robust classifier will assign roughly equal $P_c^{(i)}$ to the three classes and therefore receive small weight in the final probability calculation for that sample.

We slightly modify our weighting scheme to have better performance across all SNPs (compared to (4.1)). First, we introduce a new threshold parameter, Q, to disqualify a base classifier if the maximum posterior probability for the three predicted classes is less than Q. This attempts to disqualify failed classifiers. Second, we apply the entropy-based weighting scheme to the binary distribution formed by the combination of the maximum posterior probability and its complement. This has the effect of lowering the weight of classifiers with maximum probability close to 1/2. In

Table 4.2: Posterior probabilities for the three genotypes from four mixture model (MM) classifiers

Classifiers/Classes	XX	XY	YY
MM(ASO.L)	$P_{XX}^{(1)}$	$P_{XY}^{(1)}$	$P_{YY}^{(1)}$
MM(ASO.R)	$P_{XX}^{(2)}$	$P_{XY}^{(2)}$	$P_{YY}^{(2)}$
MM(APEX.L)	$P_{XX}^{(3)}$	$P_{XY}^{(3)}$	$P_{YY}^{(3)}$
MM(APEX.R)	$P_{XX}^{(4)}$	$P_{XY}^{(4)}$	$P_{YY}^{(4)}$

summary, the modified weights are defined as follows: For the ith classifier, let

$$P^{(i)} = \text{Max}(P_{XX}^{(i)}, P_{XY}^{(i)}, P_{YY}^{(i)})$$

and

$$\bar{P}^{(i)} = 1 - P^{(i)}.$$

Define

$$E^{(i)} = -[\log(\frac{1}{2}) - \{P^{(i)} \log(P^{(i)}) + \bar{P}^{(i)} \log(\bar{P}^{(i)})\}]$$

and

$$E_i = \begin{cases} E^{(i)} & \text{if } P^{(i)} > Q, \\ 0 & \text{otherwise.} \end{cases}$$

Here Q can be adjusted to maintain the quality of the classification.

As before, the weight for the ith classifiers is defined as

$$W_i = \frac{E_i}{\sum_i E_i},$$

and the overall class probabilities are given by the weighted average over the four classifiers:

$$P_c = \sum_{i=1}^{4} W_i P_c^{(i)}.$$

Finally, an object is assigned to the class c with the maximum P_c. A further threshold can be applied on $\max(P_c)$ if we wish to accept a lower call rate in exchange for a higher accuracy level.

4.4 Illustrative Examples Revisited

We gain further insight into the working of the robust ensemble of robustified mixture models (and its potential advantages) by revisiting the two examples introduced in Section 1.2 and examining the computations leading to the genotyping of *test sample number 12*. In this case, we will show that SNP 1360590 is incorrectly genotyped by the dynamic ensemble of LDA classifiers but correctly classified by the robust ensemble. On the other hand SNP rs1981278 is incorrectly genotyped by both ensembles. In the latter case, however, the robust approach has a lower maximum probability and could be easily thresholded down as a "noninformative" case.

SNP rs1360590 and subject 12 For the SNP 1360590, the mixture model ensemble places subject 12 in the correct class (AA) after combining the four classifiers, whereas the non-robust LDA ensemble assigns subject 12 to the wrong class (AG). The results will be clearer if we analyze the raw posterior probability matrices showing the behavior of the individual base classifiers as well as the weights assigned to these classifiers.

The posterior probability matrix and the respective weights from the ensemble of LDA classifiers are given in Table 4.3. From Figure 4.1 it is seen that the probe data for ASO.L, ASO.R, and APEX.L are not good for predicting subject 12. Particularly for ASO.R and APEX.L, subject 12 is an outlier, but the ASO.R classifier gives high probability (0.972) to the wrong class, AG, thus assigning high weight to the misleading classifier. This causes the final, wrong genotype call (AG) shown in the last row of Table 4.3.

Table 4.4 gives the posterior probability matrix for the robust ensemble of mixture models. We can see that the ASO.R and APEX.L classifiers give roughly equal probability to the three classes and therefore they get small weights. The good classifier, APEX.R, gets large weight because of the high probability (0.998) for the correct class (AA). Thus after taking the weighted average, the ensemble of mixture models assigns the correct class with a high confidence score 0.919 (last row of Table 4.4).

Moreover, for the good working probe (APEX.R), the robust mixture model assigns higher posterior probability (0.998) to the right class (Table 4.4). Whereas, the non-robust LDA model assigns relatively small posterior probability (0.951) to the right class (Table 4.3). This phenomenon is actually addressing the point that the robust mixture model works more efficiently for an individual probe chemistry with good signal for the test

sample in the presence of several outliers from the training set (see Figure 4.1). This also indirectly helps to improve the quality of the weights for the associated classifiers and thus assigning relatively more weight to a good classifier.

Table 4.3: Posterior probabilities for the three genotypes of SNP rs1360590 from four LDA classifiers

Classifier/Class	AA	AG	GG	Weight
LDA(ASO.L)	0.128	0.864	0.008	0.216
LDA(ASO.R)	0.028	0.972	< 0.001	0.412
LDA(APEX.L)	0.579	0.421	< 0.001	0.009
LDA(APEX.R)	0.951	0.049	< 0.001	0.363
LDA(Ensemble)	0.389	**0.609**	0.002	—

SNP rs1981278 and subject 12 Here we illustrate a situation where both models make the wrong call. However, the non-robust ensemble predicts the wrong class with relatively high confidence, whereas the robust ensemble makes a "don't know" call, which is reasonable in light of the poor data in Figure 4.2.

From Figure 4.2, we see that only the data from APEX.L are reliable for subject 12, indicating the correct class (TT). For the other three probes, subject 12 falls on the edge of the CT class, and the corresponding LDA classifiers produce high posterior probabilities for CT (see Table 4.5). Thus, the ensemble assigns the wrong class (CT) with relatively high confidence score (.71). On the other hand, the mixture model assigns high but smaller

Table 4.4: Posterior probabilities for the three genotypes for SNP rs1360590 from four robust mixture model classifiers

Classifier/Class	AA	AG	GG	Weight
MM(ASO.L)	0.625	0.357	0.018	0.040
MM(ASO.R)	0.304	0.392	0.304	0.030
MM(APEX.L)	0.331	0.341	0.328	0.065
MM(APEX.R)	0.998	0.001	0.001	0.865
MM(Ensemble)	**0.919**	0.049	0.032	—

probabilities to the wrong class (CT) for these misleading probes (see Table 4.6). Overall, the robust ensemble assigns less weight (.55) to the wrong CT call (see the last row of Table 4.6).

In summary, if failed chemistries produce outliers, the robust ensemble of classifiers can discard them. However, even the robust ensemble will be adversely affected in the (hopefully less likely) case that several failed chemistries produce signals that fall consistently close to the same wrong cluster (as compared with the estimated scatter in the training data).

Table 4.5: Posterior probabilities for the three genotypes of SNP rs1981278 from four LDA classifiers

Classifier/Class	CC	CT	TT	Weight
LDA(ASO.L)	< 0.001	0.938	0.062	0.186
LDA(ASO.R)	< 0.001	0.996	0.004	0.268
LDA(APEX.L)	< 0.001	0.001	0.999	0.276
LDA(APEX.R)	< 0.001	0.997	0.003	0.270
LDA(Ensemble)	< 0.001	**0.711**	0.289	—

4.5 Accuracy and Call Rate Results

For each of the 100 SNPs, we applied the non-robust and robust algorithms taking the SIRS data as the training set and the Coriell data as the test set. The true genotypes for both Coriell and SIRS are known, so we can validate the models against the actual genotypes. We measure the performance of

Table 4.6: Posterior probabilities for the three genotypes for SNP rs1981278 from four robust mixture model classifiers

Classifier/Class	CC	CT	TT	Weight
MM(ASO.L)	0.089	0.819	0.092	0.128
MM(ASO.R)	0.013	0.965	0.022	0.316
MM(APEX.L)	0.001	0.002	0.997	0.391
MM(APEX.R)	0.071	0.858	0.071	0.165
MM(Ensemble)	0.027	**0.552**	0.420	—

the two models in terms of concordance rate of the predicted and the true genotypes. Different call rates are obtained by applying a threshold to the final probability (confidence measure) of the selected class.

The robust mixture model ensemble achieves an overall concordance rate of 99.44% for 100% call rate (for all 100 SNPs over the test set), whereas the non-robust LDA model had an overall concordance rate of 99.28% for 100% call rate. Moreover, we can see from Figure 4.3, that using the robust ensemble we can achieve 99.56% concordance rate in exchange for a very small reduction in the call rate.

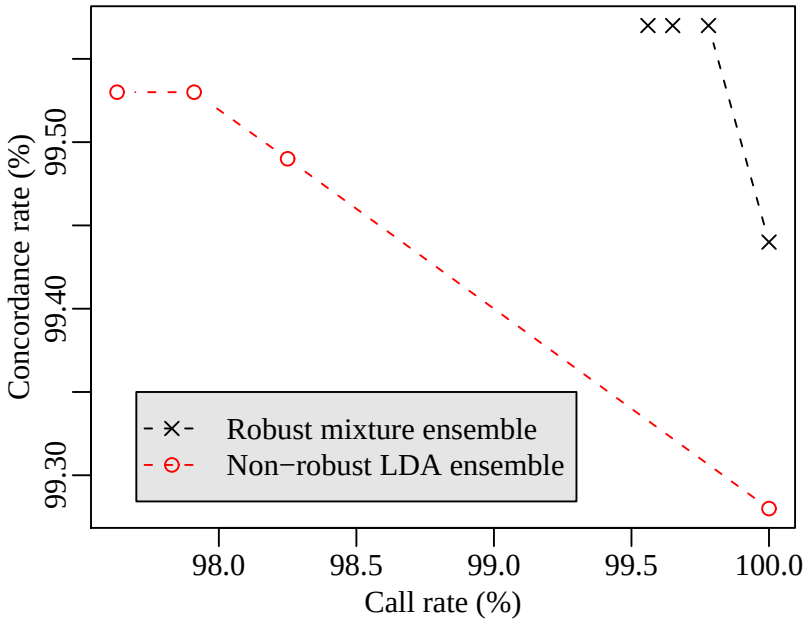

Figure 4.3: Concordance rate versus call rate for 100 SNPs. The SIRS data are used for training and the Coriell data for testing.

4.6 Simulation and Numerical Studies

4.6.1 Controlling the Amount of Contamination

We conducted a simulation study to further compare the performance of the robust and non-robust ensembles. To mimic the motivating SNP classifica-

tion problem, we consider three classes and four pairs of variables. For each pair, the class-specific bivariate distributions are normal with means (0,25), (25,25) and (25,0), respectively, and common covariance matrix

$$\Sigma = \begin{pmatrix} 45 & 25 \\ 25 & 45 \end{pmatrix}.$$

The training data for each pair of variables and each of the three classes have 100 observations in total, with 90 coming from the class-specific bivariate normal distribution, and 10 outliers drawn from a uniform distribution on the square $(-40, 100) \times (-40, 100)$. This is repeated independently for four pairs of explanatory variables. Thus, there are four separate sets of training samples, each with 3×100 labeled observations. The realization simulated from these contaminated distributions and used for the study is shown in Figure 4.4, where classes 1, 2, and 3 are denoted by $\square$, $\triangle$, and $\circ$, respectively. Note the resemblance with the real data sets in Figure 4.1 and 4.2.

For the simulation of the test data, we consider the following design. Of the four pairs of variables, $0, 1, \ldots, 4$ may be contaminated. Consequently, we draw test samples of size 200 under each of these five situations. We fix the probability of contamination at $\lambda = 0.1$. We combine the performances of these five types of test samples taking a weighted average of the individual call rates and concordance rates. Here, the performances for $0, 1, \ldots, 4$ contaminated pairs are weighted using the binomial probabilities $(0.9)^4, 4(0.9)^3(0.1), \ldots, (0.1)^4$, which assumes the pairs are contaminated or not independently (an assumption not made elsewhere in this article).

Figure 4.5 shows the results of our simulation study for the two classification ensembles. The robust model clearly dominates the non-robust one in terms of the concordance/call-rate trade-off, especially when there are two or three contaminated pairs of variables.

4.6.2 Training Versus Prediction Robustness

To gain further insight into the behavior of robust and non-robust classifiers, we consider a simple one-dimensional, two-class problem, with no redundancy in the single explanatory variable (x).

Training sets of size 20 are generated for each class: 18 observations from either $N(0, 16)$ (class 1) or $N(60, 16)$ (class 2), and two observations from the noninformative $U(-60, 120)$ distribution (see Figure 4.6). At the prediction stage, we calculate the posterior probability of class 1 for $x \in (-60, 120)$.

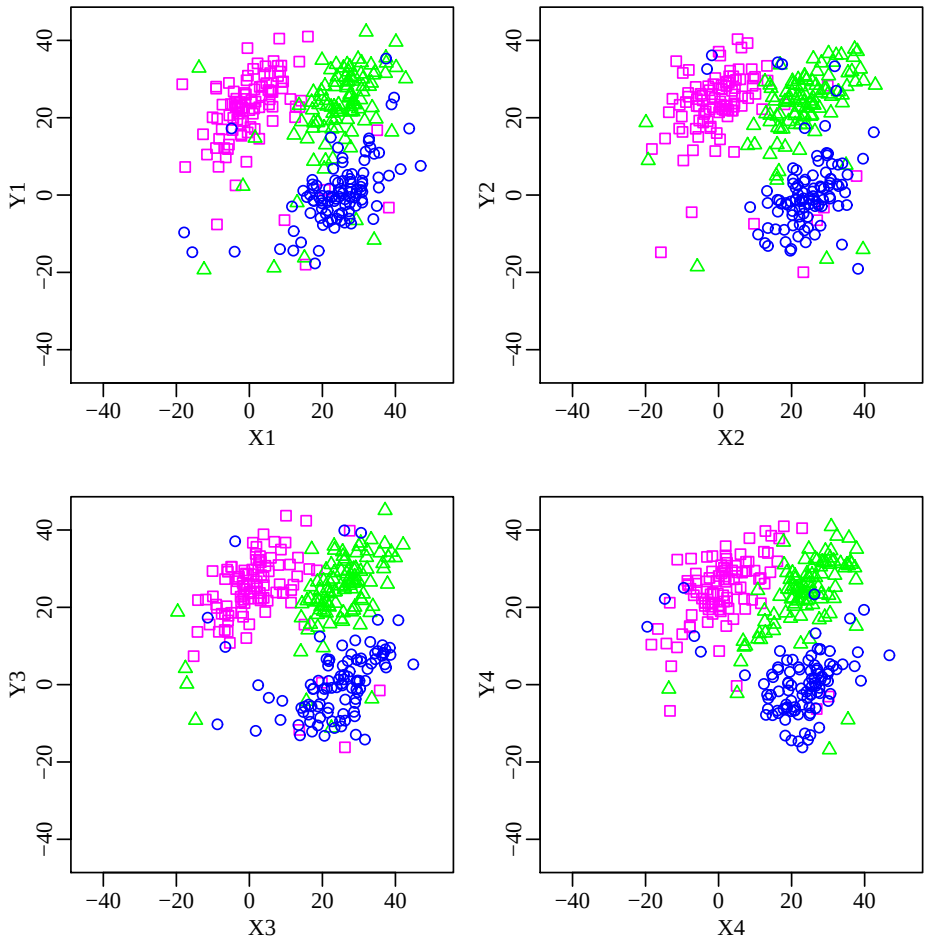

Figure 4.4: Simulated training data for four pairs of variables.

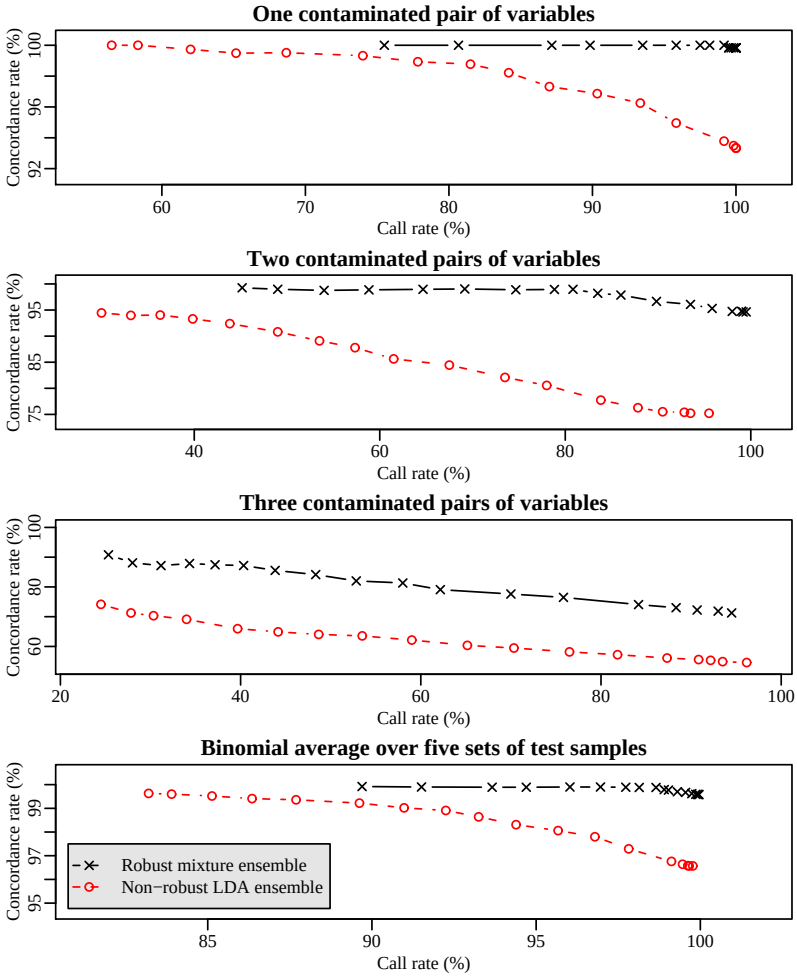

Figure 4.5: Concordance versus call rate trade-off with 1, 2, or 3 contaminated pairs of variables for prediction, and the overall average performance.

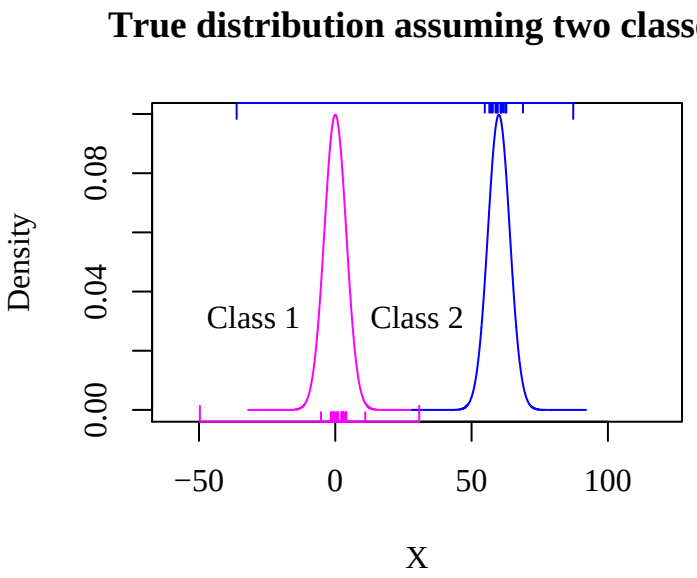

Figure 4.6: Good-signal distributions for two classes are denoted by the normal-density curves. Contaminated realized values of x for the two classes are shown in the rug plots at the bottom (Class 1) or top (Class 2).

Three classifiers are considered: LDA, robustified LDA, and a robust mixture model (Section 4.3.1). The robustified LDA is obtained by replacing sample means and standard deviations in training the model by sample medians and MADs (median absolute deviations). That is, robustified LDA is robust only at the training stage and not at the testing stage.

The posterior probability of class 1 is shown in Figure 4.7. In the top panel, regular LDA switches from predicting class 1 to predicting class 2 as x increases from the class 1 mean of 0 to the class 2 mean of 60. Note that when $x = 20$, say, which is five standard deviations away from the class 1 mean of 0, the estimated probability of class 1 is still 0.8. Paradoxically, the curve for the partially robustified LDA exhibits an even less desirable behavior. It remains very confident about calling class 1 up to $x = 30$ and then switches abruptly to a very confident class 2 call for larger values of x. The explanation is that regular LDA "benefits" from the gross inflation of the sample variances (214 and 280 in the realization leading to these results) from outliers, which leads to less confident posterior probabilities. The lower panel of Figure 4.7 shows that the robust mixture model behaves well, allotting roughly 50% probability to noninformative test cases (values of x lying more than three standard deviations from both population means). Moreover, the posterior probability function does not vary much for values of λ in the interval $(0.005, 0.15)$.

4.7 Conclusions

Our approach to using multiple redundant sets of explanatory variables is to train separate classifiers and then dynamically combine their calls test sample by test sample. Entropy based measures of confidence are used as weights in the ensemble. The alternative strategy of using all the variables in one classifier was shown to be inferior by Podder et al (2006).

For data contaminated by outliers at the training and test stages, it is important to have robustness at both levels. In fact, as illustrated by Figure 4.5, robustifying only at the training level may lead to an even less robust call when the test sample is also contaminated. Robustness in training is obtained by replacing sample means and covariances by their robust counterparts (MCD estimates in this paper). Robustness at the testing stage follows from a mixture model that allows for a fixed (relatively small) fraction of outliers.

The fully robustified approach is shown to outperform the non-robust method in an extensive simulation study and in a real data example.

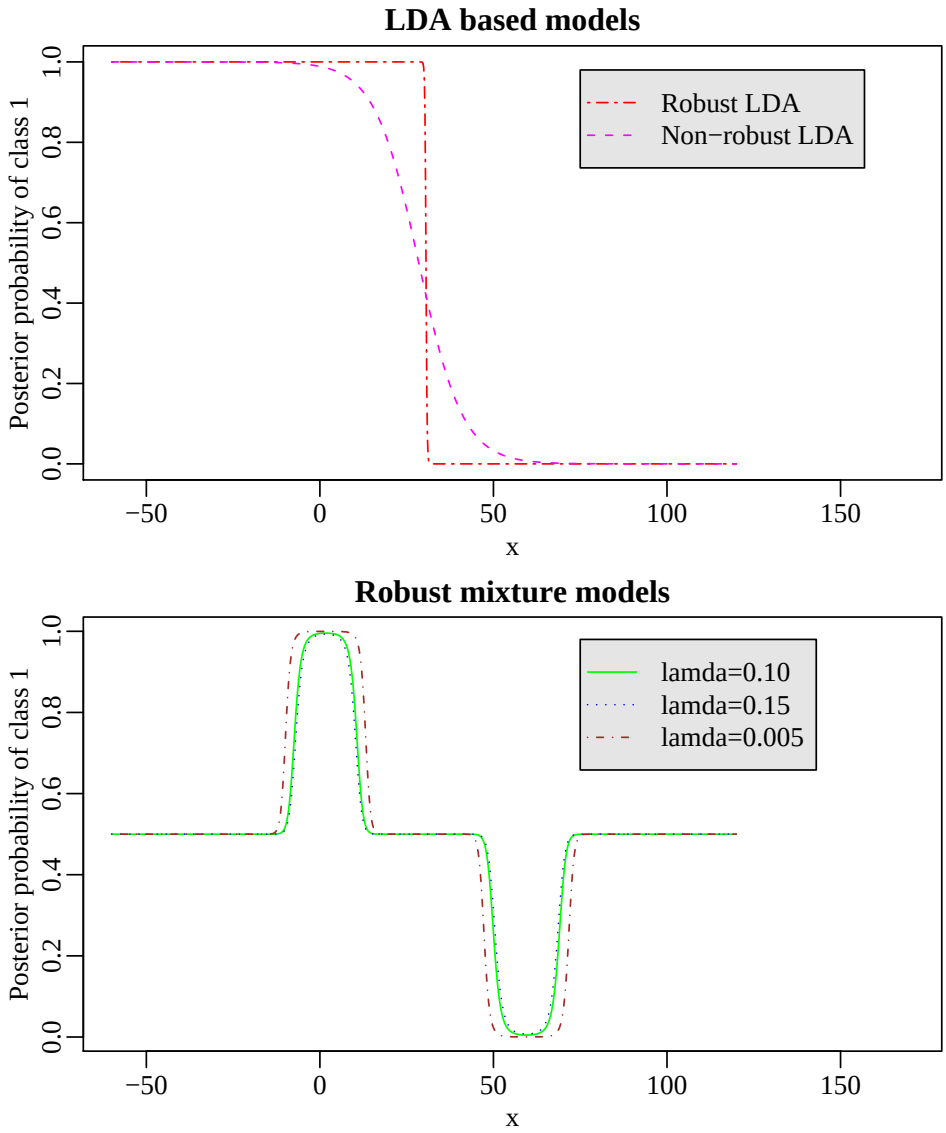

Figure 4.7: Posterior probability of class 1 as a function of the test value x

Bibliography

Croux, C. and C. Dehon (2001). Robust linear discriminant analysis using S-estimators. *The Canadian Journal of Statistics 29*, 473–492.

Fan, J., K. Gunderson, M. Bibikova, J. Yeakley, J. Chen, E. Wickham Garcia, L. Lebruska, M. Laurent, R. Shen, and D. Barker (2006). Illumina universal bead arrays. *Methods Enzymol 410*, 57–73.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*(2), 179–188.

Guo, Y., T. Hastie, and R. Tibshirani (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics 8*(1), 86.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Kennedy, G., H. Matsuzaki, S. Dong, W. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, et al. (2003). Large-scale genotyping of complex DNA. *Nature Biotechnology 21*(10), 1233–1237.

Kurg, A., N. Tonisson, I. Georgiou, J. Shumaker, J. Tollett, and A. Metspalu (2000). Arrayed Primer Extension: Solid-Phase Four-Color DNA Resequencing and Mutation Detection Technology. *Genetic Testing 4*(1), 1–7.

Oliphant, A., D. Barker, J. Stuelpnagel, and M. Chee (2002). BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques 32*, S56–S61.

Pastinen, T., M. Raitio, K. Lindroos, P. Tainola, L. Peltonen, and A. Syvanen (2000). A System for Specific, High-throughput Genotyping by Allele-specific Primer Extension on Microarrays. *Genome Research 10*(7), 1031.

Podder, M., J. Ruan, B. Tripp, Z. Chu, and S. Tebbutt (2008). Robust snp genotyping by multiplex pcr and arrayed primer extension. *BMC Medical Genomics 1*, 1–5.

Podder, M., W. Welch, R. Zamar, and S. Tebbutt (2006). Dynamic variable selection in SNP genotype autocalling from APEX microarray data. *BMC Bioinformatics 7*(1), 521.

Shumaker, J., A. Metspalu, and C. Caskey (1996). Mutation detection by solid phase primer extension. *Hum Mutat 7*(4), 346–54.

Tebbutt, S., J. He, K. Burkett, J. Ruan, I. Opushnyev, B. Tripp, J. Zeznik, C. Abara, C. Nelson, and K. Walley (2004). Microarray genotyping resource to determine population stratification in genetic association studies of complex disease. *Biotechniques 37*(6), 977–85.

Tebbutt, S., I. Opushnyev, B. Tripp, A. Kassamali, W. Alexander, and M. Andersen (2005). SNP Chart: an integrated platform for visualization and interpretation of microarray genotyping data.

Chapter 5

Further Extensions

In this thesis, we have proposed a robust, mixture–model based classification approach for SNP genotyping. The proposed method introduces three main new ideas for addressing the problem of automatic SNP genotyping: (1) dynamic ensemble of several base classifiers; (2) construction of robust classifiers at the testing level (mixture model); and (3) robust training of base classifiers. This concluding chapter discusses some venues for future research.

- *Unknown grouping of the variables.* The present APEX based SNP genotype classification problem can be viewed as a particular case of a general high dimensional classification problem where the set of explanatory variable contains several naturally grouped subsets (e.g. for the SNP genotyping problem, each base classifier corresponds to a different probe chemistry). The base classifiers have been constructed using prior knowledge of the APEX chemistry. A more general application can be considered in the absence of such prior knowledge. Given a set of explanatory variables, it would be interesting to consider different ways of grouping these variables to form separate base-classifiers. The base classifiers would then be robustly trained and dynamically combined to classify each test sample.

- *Using the background data.* The APEX microarray platform produces four-channel microarray data which includes several background channels, depending on the allelic probe type. This background information has not been used in our genotype classification model. A natural extension is to investigate the utility of the background data.

- *Pixel level data.* For SNP genotype classification, it would be desirable to design a model based on just a single sample and classify the SNP according to three possible genotypes for a biallelic SNP. More precisely, in the APEX based genotyping problem, the explanatory variables have been constructed based on the summary measurements

of hundreds of pixel intensity values. These pixel values can be incorporated directly, e.g. in likelihood based genotype models like Di et al. (2005) and Nicolae et al. (2006). The incorporation of the pixel-level data would be straightforward but computationally expensive.

- *Unequal mixture parameters and covariances.* We have proposed a robust mixture model, which combines an informative distribution (when the underlying model is true) and a non-informative distribution (for the presence of outliers). The mixture parameter λ has been assumed known and equal for all the classes. Naturally, λ could be assumed to be unknown and different across classes. In such case, λ_c should be estimated using the training data. We have also assumed that the informative components are bivariate normal distributions with equal covariances. A straight forward extension would be to assume class specific covariances instead of the same covariances across classes. This would lead to a robustified quadratic discriminant model. This extension would still allow a natural combination of the base classifiers based on the corresponding posterior probability matrix.

- *Unsupervised learning approach.* We have considered a supervised learning setup to take advantage of the available genotyped data. An obvious next stage of modeling would be to consider a unsupervised learning approach. In the absence of training data, clustering based methods could be used to label the data first. These labeled data would then be used to train the base classifiers. A worthwhile extension would be to consider a hierarchical Bayes model for the base classifiers.

- *Nonlinear classification.* We have used a robustified linear discriminant function for each base classifier and it has been straightforward to combine the effects of the classifiers based on the matrix of posterior probabilities. Instead, other classification models can be tried, e.g. random forests, support vector machines, neural networks, etc. Combining and robustifying the base classifier models, however, would not be as straightforward as in the present modeling approach.

- *Assessing the variability of the confidence measure.* From the clinical point of view, it would be useful to have a measure of the variability of the confidence score associated with each genotype call. This can be achieved in two ways: either using bootstrap or using individual pixel-level probe data instead of their average (Di et al., 2005).

Bibliography

Di, X., H. Matsuzaki, T. Webster, E. Hubbell, G. Liu, S. Dong, D. Bartell, J. Huang, R. Chiles, G. Yang, et al. (2005). Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics 21*(9), 1958–1963.

Nicolae, D., X. Wu, K. Miyake, and N. Cox (2006). GEL: a novel genotype calling algorithm using empirical likelihood. *Bioinformatics 22*(16), 1942.