UNIVERSITY OF CINCINNATI

Date: 1-Feb-2010

I, Prachi Kothiyal, hereby submit this original work as part of the requirements for the degree of:
Doctor of Philosophy
in Biomedical Engineering

It is entitled:
Detection and Classification of Sequence Variants for Diagnostic Evaluation of Genetic Disorders

Student Signature: Prachi Kothiyal

This work and its defense approved by:

Committee Chair: Bruce Aronow, PhD
Marepalli Rao, PhD
John Greinwald, Jr., MD

5/24/2010
Detection and Classification of Sequence Variants for Diagnostic Evaluation of Genetic Disorders

A dissertation submitted to the Graduate School
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the department of Biomedical Engineering
of the College of Engineering, University of Cincinnati

by

Prachi Kothiyal

B.Eng., Nanyang Technological University, Singapore, 2002

Dissertation Committee Chair and Adviser: Dr. Bruce J. Aronow
Co-adviser: Dr. John H. Greinwald
Co-adviser: Dr. Marepalli B. Rao
Abstract

Identifying and cataloguing individual and population-level DNA sequence variations is a critical step towards understanding the genetic basis of disease and clinically significant human variation. Recent advances in molecular microarray technology have made it feasible to rapidly screen DNA samples for possible genetic mutations. This dissertation focuses on evaluating the efficacy of resequencing arrays as a tool for variant detection and proposes mechanistic bases and computational algorithms that can be employed for an improvement in performance. We present results from hearing loss arrays developed in two different research facilities and highlight some of the approaches we adopted to enhance the applicability of the arrays in a clinical setting.

We leveraged sequence and intensity pattern features responsible for diminished coverage and accuracy and developed a novel algorithm, sPROFILER, which resolved >80% of no-calls from Affymetrix™ GSEQ and allowed 99.6% (range: 99.2-99.8%) of sequence to be called, while maintaining overall accuracy at >99.8% based upon dideoxy sequencing comparison. We implemented a bioinformatics pipeline that incorporated sPROFILER to support clinical genetic testing of hearing loss patients at the Cincinnati Children’s Hospital Medical Center.

The utility of any molecular diagnostic tool in determining the genetic basis of a disease is fully realized only when an effective variant detection method is complemented by a rigorous framework for evaluating the potential clinical significance of these
variants. We evaluated the contribution of various properties related to amino acid substitution in determining whether a residue change is damaging or not. We developed a machine learning-based framework to assess the functional impact of missense variants using childhood Sensorineural Hearing Loss and Hypertrophic/Dilated Cardiomyopathy as specific instances of application of the methodology. We compared our method with some of the representative tools for missense variant classification and present results that demonstrate the improvement in classification accuracy. Additionally, we developed customized classifiers trained on proteins sharing Gene Ontology terms with the protein being tested and observed a smaller training set could be used to provide better or same prediction accuracy as compared to utilizing a large generic training set for all proteins.
Acknowledgements

I would like to thank my advisor, Dr. Bruce J. Aronow, for his guidance, support and patience during the course of my dissertation research. It was working with Dr. Aronow and seeing his enthusiasm for his work that motivated me to explore bioinformatics and eventually adopt it as my primary field of research. He has been a constant source of encouragement and inspiration. His knowledge and deep understanding of the direction in which bioinformatics research is headed made working with him a privilege. Dr. Aronow motivated me through times of success and failure and was a mentor to me in the true sense of the word. I consider myself fortunate to have his mentorship and friendship during my research and thereafter.

I would like to thank the members of my dissertation committee, Dr. John H. Greinwald and Dr. Marepalli B. Rao. Dr. Greinwald provided me with a framework to conduct research while working towards an objective that has clinical relevance. He provided invaluable suggestions and feedback and helped me fully understand the motivation behind the projects I worked on. I would like to thank Dr. Rao for his guidance and critique on the development of the computational methods. Additionally, the courses he offered were extremely helpful in preparation for my dissertation. I would like to express my gratitude to Dr. Heidi L. Rehm at Harvard Partners HealthCare Center for Personalized Genetic Medicine for providing me with an opportunity to work with her, albeit remotely. Dr. Rehm’s guidance has been an integral part of my research and I am fortunate to have collaborated with her at various stages of my graduate career. I
would also like to thank other faculty at Division of Biomedical Informatics and University of Cincinnati.

I would like to acknowledge and thank Jonathan Ebert, Stephanie Cox and Ammar Husami for providing me with the experimental data I needed for my dissertation and for always being willing to help. I would also like to thank Sivakumar Gowrisankar for sharing his knowledge and providing me with useful feedback along the way.

I am grateful for the support my friends have provided me with. I would especially like to thank Amit, Vivek, Mythri, Janet, Megha and Mukta for helping me at critical junctures in their own unique ways. My family’s support and belief in me have helped me through all my endeavors. I would like to acknowledge the encouragement I received from my in-laws in working towards the completion of my dissertation. I would like to thank my parents for helping me get to this point, for teaching me that getting ahead at the cost of compromising fundamental values is taking a step back, and for their unconditional encouragement to pursue my interests, even when my interests took me beyond the boundaries of geography. And finally, the most patient, supportive, logical and encouraging person I have ever known, who I am fortunate to be married to. Thank you for believing in me even when I was not so sure.
Publications arising from this thesis

Parts of the dissertation are based on the following manuscripts:

**Chapters 3 and 4:**


**Chapters 6 and 7:**

Table of Contents

ABSTRACT ............................................................................................................................... III
ACKNOWLEDGEMENTS ........................................................................................................... VI
PUBLICATIONS ARISING FROM THIS THESIS ...................................................................... VIII
LIST OF FIGURES ................................................................................................................ XI
LIST OF TABLES .................................................................................................................. XII

CHAPTER 1. OVERVIEW ............................................................................................................ 1
  1.1 MOTIVATION .................................................................................................................... 1
  1.2 CONTRIBUTIONS OF THIS THESIS ............................................................................. 2

CHAPTER 2. GENETIC TESTING FOR SENSORINEURAL HEARING LOSS (SNHL) USING RESSEQUENCING MICROARRAYS .......................................................................................... 5
  2.1 SIGNIFICANCE OF GENETIC TESTING FOR SENSORINEURAL HEARING LOSS ....... 5
  2.2 APPLICATION OF RESSEQUENCING MICROARRAYS FOR GENETIC TESTING ......... 6
  2.3 THE HEARING LOSS GENE CHIP .................................................................................. 7

CHAPTER 3. BASE-CALLING FOR THE HEARING LOSS GENE CHIP ...................................... 10
  3.1 INTRODUCTION ................................................................................................................ 10
  3.1.1 Allele-specific hybridization and base-calling ......................................................... 10
  3.1.2 Performance measures for base-calling assessment ................................................. 12
  3.1.3 Base-calling with Affymetrix GeneChip Sequence Analysis Software (GSEQ) ......... 13
  3.2 MATERIALS AND METHODS ....................................................................................... 15
    3.2.1 Experimental design ............................................................................................... 15
    3.2.2 Generation of base calls with GSEQ ................................................................. 16
    3.2.3 Analysis of hybridization intensities .................................................................... 16
  3.3 RESULTS ....................................................................................................................... 17
    3.3.1 Performance of the Hearing Loss Gene Chip ...................................................... 17
    3.3.2 Impact of GC content on probe performance ................................................... 19
  3.4 DISCUSSION ................................................................................................................ 24

CHAPTER 4. BASE-CALLING OPTIMIZATION WITH SPROFILER AND IMPLEMENTATION OF A BIOINFORMATICS PIPELINE FOR CLINICAL APPLICATION OF RESSEQUENCING MICROARRAYS .......................................................... 29
  4.1 INTRODUCTION ............................................................................................................. 29
  4.1.1 Leveraging sequence-specific hybridization properties for improved base-calling ... 29
  4.1.2 Reduction of false positive variant calls ............................................................... 30
  4.2 MATERIALS AND METHODS .................................................................................... 30
    4.2.1 Development of sPROFILER .............................................................................. 30
    4.2.2 sPROFILER: Methodology and algorithm design ............................................. 32
  4.3 RESULTS .................................................................................................................... 34
    4.3.1 Determination of optimal threshold ratios ......................................................... 34
    4.3.2 Array performance with sPROFILER ............................................................... 36
    4.3.3 Bioinformatics pipeline for clinical application of the Hearing Loss Gene Chip ... 39
  4.4 DISCUSSION ............................................................................................................... 41

CHAPTER 5. SURVEY OF EXISTING METHODS FOR FUNCTIONAL INTERPRETATION OF MISSENSE GENETIC VARIANTS .............................................................. 43
  5.1 SIGNIFICANCE OF FUNCTIONAL INTERPRETATION OF VARIANTS .................... 43
  5.2 EXISTING TOOLS FOR ANNOTATION OF MISSENSE VARIANTS ......................... 44
  5.3 DISCUSSION .............................................................................................................. 47
List of Figures

Figure 1: Summary of steps involved in variant detection using resequencing array. .................... 9
Figure 2: CustomSeq Array Design ............................................................................................... 11
Figure 3: Examples of hybridization patterns ................................................................................ 12
Figure 4: Overview of Affymetrix GSEQ base-calling algorithm .................................................... 14
Figure 5: Experimental design for resequencing array study ........................................................ 15
Figure 6: Relationship between probe GC content and array call rate and accuracy ................. 20
Figure 7: Relationship between probe GC content and peak feature intensity and ratio of peak to
next highest intensity. .................................................................................................................... 20
Figure 8: Differential dependence of peak feature intensity on probe G- and C-content .......... 22
Figure 9: Relationship between number of bases in a continuous stretch of C's/G's and peak
feature intensity and ratio of peak to next highest intensity. ......................................................... 22
Figure 10: Differential impact of high probe G-content and C-content on probe performance .... 23
Figure 11: Probe consensus sequences for no-calls, false positive and wild-type calls......... 24
Figure 12: Sample output file with results for a single chip (ChipNum_final.txt) ................. 34
Figure 13: Output file containing calls made by different programs for positions known to have
variants (ChipNum_OutKnownSnp.txt) ......................................................................................... 34
Figure 14: Improvement in array call rates with protocol optimization and application of
sPROFILER to GSEQ calls (data shown for Cincinnati arrays) .................................................. 37
Figure 15: False positive calls with and without protocol optimization/sPROFILER .......... 38
Figure 16: False negative calls with and without protocol optimization/sPROFILER represented as
a portion of total true variants ........................................................................................................ 38
Figure 17: Extraction of positive and negative instances for training set. .................................... 53
Figure 18: Classification agreement among SIFT, PolyPhen and PANTHER for variants in training
set ................................................................................................................................................. 61
Figure 19: Receiver Operating Characteristic (ROC) curves and Area Under Curve (AUC) for
SVM, SIFT, PolyPhen and PANTHER. ......................................................................................... 62
Figure 20: Classification agreement among SIFT, PolyPhen and PANTHER for variants in
validation dataset ........................................................................................................................... 70
Figure 21: ROC curves and AUC for SVM, SIFT, PolyPhen and PANTHER for variants in
validation dataset ........................................................................................................................... 70
Figure 22: Performance comparison of SVM trained with genes belonging to different GO
categories ........................................................................................................................................ 71
Figure 23: ROC curve and AUC for SVM trained with genes sharing “Molecular Function” GO
terms, SIFT, PolyPhen and PANTHER. ......................................................................................... 75
List of Tables

Table 1: Genes selected for the Hearing Loss Gene Chip .............................................................. 8
Table 2: Performance measures for determining efficacy of resequencing arrays ....................... 13
Table 3: Algorithm settings used for Affymetrix GSEQ ................................................................. 16
Table 4: Number of exons in each array with no-calls or variant calls ........................................ 18
Table 5: Comparison of schemes for calculation of site-specific threshold for peak to next highest intensity ratio .............................................................................................................. 36
Table 6: Overall array performance with and without application of sPROFILER to GDAS/GSEQ base calls .................................................................................................................................. 39
Table 7: Fields included in the output report for array based analysis ........................................... 40
Table 8: Predictive features evaluated for development of classifier ............................................. 54
Table 9: Accuracy, AUC and F-score with one feature held out; F-score quantifies importance of each feature based on training data ...................................................................................... 60
Table 10: Performance comparison of SVM-based classifier against SIFT, PolyPhen and PANTHER ................................................................................................................................ 62
Table 11: Classification performance with SIFT, PANTHER, PolyPhen, and SVM .......................... 72
Table 12: Prediction for individual genes in the test set. Entries in bold indicate additional variants that were detected by training the classifier on genes sharing GO terms .................................................. 74
Table 13: Misclassified pathogenic variants .................................................................................... 76
Table 14: Variants with low MAF and benign predictions selected to test the dependence on MAF of the classifier ......................................................................................................................... 77
1.1 Motivation

As more and more genes, pathways, polymorphisms and allelic interactions are understood with respect to their roles in disease, the need for cost-effective high-throughput DNA resequencing for clinically affected individuals has risen. The identification of over 45 genes causative for Sensorineural Hearing Loss makes it imperative to develop a high-throughput resequencing assay that would allow for a more comprehensive and therefore higher yield diagnostic evaluation of the etiology of hearing loss in patients. Several studies have been published describing the application of resequencing arrays for screening DNA samples for genetic mutations (Liu et al. 2007, Denning et al. 2007, Fokstuen et al. 2008, Waldmuller et al. 2008, Lebet et al. 2008). These studies also underscore that inadequate call coverage associated with the base-calling software (GSEQ) provided by Affymetrix can be a significant limitation. Affymetrix, GSEQ is known to produce very few false negatives, thus providing a highly sensitive test. However, follow-up dideoxy sequencing for resolution of no-calls leads to an additional variable cost, a factor which needs to be carefully considered for clinical application of the technology. To this effect, we propose a novel algorithm for resolution of no-calls from GSEQ.

One of the critical steps towards developing an effective genetic test focuses on understanding the biological significance of genetic variants that are obtained upon screening a subject. An interpretation method becomes especially necessary when novel
variants with unknown significance are discovered in a patient. We present some of the existing methods for missense variant classification and discuss their strengths and limitations. We propose the development of a machine learning-based framework to assess the functional impact of missense variants using childhood Sensorineural Hearing Loss and Hypertrophic/Dilated Cardiomyopathy as specific instances of application of the methodology.

1.2 Contributions of this thesis

Chapter 2 discusses the genetic factors associated with sensorineural hearing loss and presents the motivation behind developing a genetic test for the disorder. We present some of the applications of resequencing arrays for variant detection and discuss the features that make arrays effective for genetic testing. We also present an overview of the Hearing Loss Gene Chip and describe the design of the array.

Chapter 3 discusses base-calling for resequencing arrays and performance of the Hearing Loss Gene Chip in terms of accuracy and coverage in interrogating the DNA sequence of interest. All results are compared to dideoxy sequencing data, which is used as the gold standard for validation of the arrays. Sub-section 3.1.1 introduces the concept of allele-specific hybridization and how it is applied to resequencing arrays. Sub-section 3.1.2 describes some of the performance measures used to evaluate the efficacy of the arrays and 3.1.3 discusses the current base-calling algorithm available for Affymetrix CustomSeq resequencing arrays. Sections 3.2 and 3.3 describe in detail the experimental design and results for performance evaluation of the Hearing Loss Gene Chip. Section
Chapter 4 introduces a novel algorithm, sPROFILER, for the reduction of unresolved calls from Affymetrix GSEQ. 4.1.1 presents some of the properties associated with sequence-specific hybridization that can be leveraged to improve base-calling and 4.1.2 presents past work on reduction of false positive calls. Section 4.2 describes in detail the methodology and design of the algorithm. Section 4.3 presents results to compare performance of the algorithm against Affymetrix GSEQ and dideoxy sequencing results. The algorithm is evaluated in terms of associated coverage and accuracy in base-calling. 4.3.2 describes the implementation of a bioinformatics pipeline to support clinical application of Hearing Loss Gene Chip at Cincinnati Children’s Hospital Medical Center.

Chapter 5 presents a survey of existing tools for analysis of missense variants. Section 5.1 discusses the motivation behind functional interpretation of missense variants. Section 5.2 reviews literature on the existing computational methods and describes some of the representative tools for missense variant classification while discussing their individual strengths and weaknesses.

Chapter 6 describes in detail the development of a machine learning based method for classification of missense variants. Section 6.1 presents the rationale behind using a combination of various predictors to make an inference about the pathogenicity of a
variant and discusses some of the features that have been used previously. Section 6.2 describes the process used for selection of training dataset, extraction of predictive features and the implementation of a Support Vector Machine based classifier. Section 6.3 presents cross-validation results on the training set and discusses the performance of the classifier.

Chapter 7 presents a case study for validation of the classifier introduced in Chapter 6. Hypertrophic Cardiomyopathy and Hearing Loss variants are used as a test dataset and are analyzed with the classifier. Section 7.2 presents the methods used for performing the analysis and 7.3 describes in detail the performance of the classifier and provides a comparison with three other existing tools for variant classification. Chapter 7 also describes the development and evaluation of an approach using gene-specific customized training sets based on Gene Ontologies.

Chapter 8 gives an overview of the goals achieved in this dissertation and presents specific aspects of the work that require further research. The chapter provides concluding remarks on the performance achieved with the two computational methods discussed in the dissertation and also explains some of the limitations that are associated with these methods.
Chapter 2. Genetic testing for sensorineural hearing loss (SNHL) using resequencing microarrays

2.1 Significance of genetic testing for Sensorineural Hearing Loss

SNHL is characterized by abnormalities in the cochlea, auditory nerve or central processing centers of the brain. It affects nearly 1 to 3 for 1000 newborn babies (Gurtler et al. 2003) and at least 50% of pre-lingual stage hearing loss can be attributed to genetic causes (Schrijver 2004). In the absence of rehabilitation, hearing loss can adversely affect speech and language development (Matsunaga 2009). Early diagnosis of SNHL can lead to an early treatment and improvement in communication outcomes (Verhaert et al. 2008). Determining the etiology of hearing loss is important for its management and can aid in determining prognosis, interventions and recurrence risks to future children (http://www.hpcgg.org/LMM/comment/OtoChip_Info_Sheet.jsp). It can also help in identifying a syndrome that may be present without showing any manifestations of other problems (e.g. adolescent-onset retinitis pigmentosa in Usher syndrome). Identification of genetic causes can facilitate understanding the underlying mechanism of hearing loss and can lead to better management of the disorder. Early identification of a genetic mutation may preempt costly and invasive procedures that may be otherwise required (http://www.cincinnatichildrens.org/svc/alpha/m/molecular-genetics/hearing-loss/tier-one.htm).

Mutations in the GJB2 gene account for nearly 20% of childhood non-syndromic SNHL cases (Putcha et al. 2007). In addition to GJB2, there are other genes that have
been reported to play a significant role in childhood hearing loss including *SLC26A4* (Everett et al. 1997), Usher type 1 genes (e.g. *MYO7A, CDH23*), (Weil et al. 1997, Bork et al. 2001) and *OTOF* (Yasunaga et al. 1999). With the identification of over 45 genes causative for SNHL, a unified genetic test can allow for the detection of mutations in multiple genes, thus providing a more comprehensive diagnostic evaluation of the genetic etiology of hearing loss in patients.

### 2.2 Application of resequencing microarrays for genetic testing

Serial molecular techniques (e.g., direct sequencing, single-strand conformation analysis, denaturing gradient gel electrophoresis and denaturing high-performance liquid chromatography) have been employed for detection of mutations associated with disorders showing high genetic and allelic heterogeneity but they can be laborious requiring high turnaround times and show little difference in their direct costs per base, which are high (Fokstuen et al. 2008). Conventional serial methods can be especially ineffective for screening large genes without definite hot spots for disease-associated mutations (Liu et al. 2007). Although new advancements in next generation sequencing will soon replace all large scale sequencing platforms, these technologies are still too costly for medium size applications of targeted disease sequencing. High-density oligonucleotide microarrays provide an efficient and economically competitive method for genetic screening of heterogeneous disorders by allowing parallel resequencing of multiple genes in a single experiment. Since the first study reporting detection of known genomic variants using oligonucleotide arrays (Saiki et al. 1989), several others have been published describing the principles of resequencing array technology (Hacia 1999,

Resequencing arrays are most effectively used in disease-specific studies where multiple samples have to be analyzed for variant detection against an established reference sequence (Kothiyal et al. 2009). Resequencing arrays allow for the organization of multiple genes on a single platform. Studies on development of customized resequencing arrays for various disorders (e.g. intrahepatic cholestasis, myasthenic syndromes, hypertrophic cardiomyopathy and severe combined immunodeficiency) underscore the clinical utility of the tool (Liu et al. 2007, Denning et al. 2007, Fokstuen et al. 2008, Waldmuller et al. 2002, Lebet et al. 2008). However, these studies also highlight limitations that compromise the effectiveness of the technology, some of which will be discussed in detail (see section 3.4).

2.3 The Hearing Loss Gene Chip

The Hearing Loss Gene Chip contains 8 genes, namely GJB2, GJB6, CDH23, KCNE1, KCNQ1, MYO7A, OTOF, and SLC26A4 (Table 1). The array contains probes to query nearly 26,000 DNA bases. Due to space constraints only the most conserved region of CDH23 has been tiled on the array, thus covering 59 out of 69 exons or 80.3% of the coding sequence. In addition to the coding sequence, 2 base pairs of each flanking splice site have been tiled for most of the exons. The genes were selected based on prevalence
of published mutations, known inheritance patterns of mutations, and the impact of the genetic variation on patient outcome. Since most genes had limited data on relative contribution to hearing loss, genes were also selected based upon minor characteristics (e.g. SLC26A5) for discovering their roles in the disorder. In order to validate the resequencing assay against an established gold standard, conventional dideoxy sequencing has been performed for all exons for 25 hearing loss arrays to allow assessment of sensitivity and specificity.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Locus</th>
<th>Accession Number</th>
<th>Length (bp)</th>
<th>Exons</th>
</tr>
</thead>
<tbody>
<tr>
<td>GJB2</td>
<td>DFNB1/A3</td>
<td>NM_004004</td>
<td>681</td>
<td>2</td>
</tr>
<tr>
<td>OTOF</td>
<td>DFNB9</td>
<td>NM_194248</td>
<td>7051</td>
<td>42</td>
</tr>
<tr>
<td>GJB6</td>
<td>DFNB1</td>
<td>NM_006783</td>
<td>783</td>
<td>1</td>
</tr>
<tr>
<td>SLC26A4</td>
<td>DFNB4/PDS</td>
<td>NM_000441</td>
<td>2342</td>
<td>19</td>
</tr>
<tr>
<td>KCNQ1</td>
<td>JLN51</td>
<td>NM_000218</td>
<td>1746</td>
<td>15</td>
</tr>
<tr>
<td>KCNE1</td>
<td>JLN52</td>
<td>NM_000219</td>
<td>390</td>
<td>1</td>
</tr>
<tr>
<td>MYO7A</td>
<td>DFNB2/A11/USH1B</td>
<td>NM_000260</td>
<td>6647</td>
<td>41</td>
</tr>
<tr>
<td>CDH23</td>
<td>DFNB12/USH1D</td>
<td>NM_022124</td>
<td>10065</td>
<td>59</td>
</tr>
</tbody>
</table>

**Table 1**: Genes selected for the Hearing Loss Gene Chip.

**Figure 1** gives an overview of the steps involved in the detection of variants using resequencing array. The Hearing Loss Gene Chip was designed with the intent of screening candidate genes in patients with pediatric SNHL and identifying causative variations for diagnostic evaluation of the hearing loss etiology in these patients. This dissertation discusses bioinformatics approaches applied towards the development and clinical application of resequencing arrays at the Cincinnati Children’s Hospital for detection of variants in children diagnosed with SNHL. Subsequent sections describe the design of the array, the challenges faced in terms of determination and analysis of genetic variations in light of some of the known limitations of resequencing arrays, and the
impact of a novel algorithm developed for improved resequencing coverage and accuracy.

Figure 1: Summary of steps involved in variant detection using resequencing array.
Chapter 3. Base-calling for The Hearing Loss Gene Chip

3.1 Introduction

3.1.1 Allele-specific hybridization and base-calling

Oligo-hybridization based resequencing arrays rely on the principle of allele-specific hybridization for determining the sequence of interest (Lipshutz et al. 1999). Every base in the DNA sequence being interrogated is represented by a set of four oligomers for each strand, thus producing a total of 8 oligomers per base (Figure 2). Multiple copies of each oligomer are synthesized on the array and are contained within discrete features. The oligomers are 25 base pairs long for the Affymetrix GeneChip CustomSeq arrays and differ only at the central position with the remaining 24 positions being complementary to the reference DNA sequence. Based on the DNA base at the position of interest, each of the probes binds differently to the target DNA. The patient sample is labeled with fluorescent molecules and the feature producing maximum fluorescence corresponds to the complementary probe that produces strongest binding with the target DNA, thus indicating the base at that position.

For any given position, there exist a total of eight feature intensities corresponding to hybridization between target DNA and four different probes for each strand where the central position is A, C, G or T (http://www.affymetrix.com/support/technical/datasheets/customseq_program_datasheet.pdf). A base-calling algorithm assigns a call based on the most discernible intensity
signature that is produced at a position in the target sequence being interrogated. The assigned call can be homozygous (identical alleles at a site), heterozygous (two different alleles at a site) or left as a no-call (indeterminate intensity signature). A position in the sequence of a haploid organism can be assigned one of five calls (A, C, G, T or no-call) whereas diploid organisms can be assigned one of six calls (AC, AG, AT, CG, CT, TG) in addition to the above five calls. A homozygous call corresponds to single peak intensity whereas a heterozygous call involves two peak intensities on either strand and in one of the peaks corresponding to the reference sequence allele at the site (Figure 3). A variant is detected when the probe producing maximum fluorescence does not correspond to the reference sequence (wild-type allele) at the site. The final call could be a wild-type call, a homozygous variant, a heterozygous variant, or a no-call.

![CustomSeq Array Design](image)

**Figure 2:** CustomSeq Array Design. Eight 25-mers are tiled to probe each individual base pair. A single position differs so that each of the four possible nucleotides can be queried on both the sense and antisense strands. *Reproduced from the Affymetrix Data Sheet: GeneChip CustomSeq Resequencing Array Program.* (Kothiyal et al. 2009)
3.1.2 Performance measures for base-calling assessment

In order to determine the effectiveness of a base-calling algorithm, certain performance measures need to be considered. For the purpose of validation, dideoxy sequencing is used to generate “true data” against which base-calling results are assessed. Table 2 outlines some of these measures.

An effective base-calling algorithm would produce minimal false positives and false negatives while maximizing call rate. Performance of the base-calling algorithm directly affects call rate, consistency and accuracy of the called genotypes. In cases where dideoxy sequencing is used for confirmation of variant calls or for following up on array no-calls, false positives do not jeopardize sensitivity of the test but the cost-effectiveness of the test is compromised if a large amount of follow-up sequencing is required. On the other hand, a true variant that has been called wild-type by a base-calling algorithm will not be detected and will subsequently affect accuracy of the test.
Table 2: Performance measures for determining efficacy of resequencing arrays (Kothiyal et al. 2009)

<table>
<thead>
<tr>
<th>Performance measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average call rate</td>
<td>Total bases called/total bases interrogated by array</td>
</tr>
<tr>
<td>Base call accuracy</td>
<td>Correct base calls/total bases called</td>
</tr>
<tr>
<td>Total false positive rate</td>
<td>False variant calls/total bases called</td>
</tr>
<tr>
<td>Total false negative rate</td>
<td>Variants falsely called WT/total bases called</td>
</tr>
<tr>
<td>Variant false positive rate</td>
<td>False variants called/total variant calls</td>
</tr>
<tr>
<td>Variant false negative rate</td>
<td>Variants falsely called WT/total true variants</td>
</tr>
</tbody>
</table>

3.1.3 Base-calling with Affymetrix GeneChip Sequence Analysis Software (GSEQ)

Resequencing data from Affymetrix CustomSeq arrays is analyzed with the help of two software applications, GCOS and GSEQ. GCOS generates a DAT file, which contains an image of the scanned probe array. The DAT file is further processed to generate a CEL file containing the mean signal intensities for each discrete feature. The base-calling algorithm, GSEQ, analyzes intensity data from CEL file and generates a CHP file, which contains final calls corresponding to each position in the DNA sequence being interrogated (Affymetrix GSEQ User Guide: http://www.affymetrix.com/support/downloads/manuals/gseq_user_guide.pdf). GSEQ version 4.0 uses Resequencing Algorithm Version 1.0 (RA v1.0) for assigning base calls after analyzing intensity data. RA v1.0 is an extension of the Adaptive Background Genotype Calling Scheme (ABACUS) (Cutler et al. 2001). The algorithm has been described in detail in the Affymetrix GSEQ User Guide and Figure 4 provides an overview of the steps involved in generation of base calls with GSEQ.
GSEQ uses a combination of user-modifiable parameters in making a base call. These parameters have been described in detail in the Affymetrix GSEQ User Guide. In addition to assigning base calls, GSEQ also produces quality scores (QS) to indicate the level of confidence for the corresponding call. GSEQ allows the user to set the QS threshold (QS) below which a site will be automatically assigned a no-call. A higher QST increases accuracy at the cost of reduced call rate. The QST can be assigned a higher value when a low false positive rate is preferred over a low false negative rate.

Figure 4: Overview of Affymetrix GSEQ base-calling algorithm
3.2 Materials and Methods

3.2.1 Experimental design

Results from 25 Hearing Loss arrays have been used for this study. Figure 5 gives an overview of the experimental design for generation of resequencing data for patients with SNHL.

The arrays were prepared and hybridized in the Molecular Diagnostics Laboratory at Cincinnati Children’s Hospital Medical Center and once the arrays had been run, results were subsequently obtained in the form of intensity data, which was analyzed with GSEQ to produce final base calls. After initial experience, the experimental protocol was modified to include only short range PCR in order to increase PCR and fragmentation consistency (see 3.3.1). Dideoxy sequencing was performed for all 25 samples for comparison of array results against a “gold standard”.

### Selection of Cohort
- Patients with SNHL and treated at CCHMC
- Exclude patients with:
  - biallelic GJB2 mutation causing DFNB1
  - other symptoms indicating syndromic hearing loss
  - diagnosis of non-genetic etiology of hearing loss

### Selection of Genes
- Genes that cause syndromic SNHL but where hearing loss is the only observable symptom during infancy
- Genes known to cause non-syndromic SNHL
- Genes reported in multiple populations, indicating clinical significance

### Identification of Variants
- 25 samples have been selected for the study
- 8 genes per sample were sequenced using dideoxy sequencing
- Each individual was assigned a hearing loss phenotype (mild, moderate, moderately-severe, severe, profound) based on audiometric data

**Figure 5:** Experimental design for resequencing array study
3.2.2 Generation of base calls with GSEQ

A total of 25 arrays were analyzed with the pilot batch consisting of 12 Hearing Loss arrays. Affymetrix GSEQ 4.0 with optimized settings (Table 3) was used for assignment of base calls. Array hybridization data was used to generate DAT files with GCOS, which were converted to CEL files with average feature intensities and GSEQ subsequently analyzed the intensities to produce base calls. All 12 arrays were analyzed in a single batch in GSEQ to maximize the sample size of the batch. The settings had been characterized for a batch of 26 Harvard arrays and the same settings were then used for Cincinnati arrays. Optimal settings were determined after evaluating different quality score thresholds and comparing array calls with dideoxy sequencing results to characterize coverage and accuracy.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Signal Threshold (probe signal/noise ratio)</td>
<td>1</td>
</tr>
<tr>
<td>Weak Signal Fold Threshold (mean/probe ratio)</td>
<td>20</td>
</tr>
<tr>
<td>Large SNR Threshold (probe signal/noise ratio)</td>
<td>20</td>
</tr>
<tr>
<td>Strand Quality Threshold (quality score)</td>
<td>-4</td>
</tr>
<tr>
<td>Total Quality Threshold (quality score)</td>
<td>30</td>
</tr>
<tr>
<td>Maximum Fraction of Heterozygote Calls (0-1)</td>
<td>0.9</td>
</tr>
<tr>
<td>Model Type (0=Heterozygote, 1=Homozygote) = 0</td>
<td>0</td>
</tr>
<tr>
<td>Perfect Call Quality Threshold (quality score)</td>
<td>2</td>
</tr>
<tr>
<td>Min Fraction of Calls in Neighboring Probes (0-1)</td>
<td>0.5</td>
</tr>
<tr>
<td>Min Fraction of Calls of Samples (0-1)</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 3: Algorithm settings used for Affymetrix GSEQ

3.2.3 Analysis of hybridization intensities

In addition to assessing overall array coverage and accuracy, raw hybridization intensities were also analyzed to discover new or ascertain known correlations that exist
in the hybridization data but are not evident if only the eventual calls are inspected. CEL files were extracted for the 13 optimized arrays and were used for the analysis; however, 12 arrays from the pilot batch were excluded from the analysis as they performed poorly due to sub-optimized experimental protocol. Perl scripts were written to analyze raw intensity data.

3.3 Results

3.3.1 Performance of the Hearing Loss Gene Chip

**Overall call rate and accuracy.** Base call rates for 12 arrays (315,504 bases) ranged from 82.5% to 96.9% with an average call rate of 91.3% using GSEQ 4.0. Comparison of dideoxy sequencing data with array calls gave a call accuracy of 99.23% with nearly 180 false positives per array. Upon comparing results from the pilot batch to Harvard arrays, it was found that the number of no-calls and false positives were found to be higher for Cincinnati arrays. The average base call rate across 26 Harvard arrays (654,862 bases) was 96.9% and the call accuracy was 99.82% with 41 false positives per array on average. Data from Harvard arrays had shown that incomplete fragmentation resulted in low call rates and that the bases most affected were those within the long range PCR fragments. Additionally, reproducibility of the quantity of product from long range PCR was less.

Based on the above observations, the target amplification process for Cincinnati arrays was modified. Instead of using a combination of long and short range PCR (54 fragments ranging from 434 to 13,282 bases in length), as was employed in the pilot batch, all long range PCRs were converted to short range resulting in 180 fragments (1
per exon) ranging from 315 to 980 bases in length. The impact of using shorter PCR products was evaluated by comparing array data across the two protocols. A total of 13 arrays (341,796 bases) were run with the optimized “short range PCR only” protocol and an average call rate of 97.9% (range 96.7% to 98.4%) was obtained with GSEQ. Dideoxy sequencing was performed for 336,171 of the bases and comparison of this data with array calls gave an array call accuracy of 99.83%. The average number of false positives dropped from 180 to 42 per array. The modified protocol with shorter PCR products was then adopted for subsequent arrays.

**Analysis of no-calls and variant calls.** Affymetrix GSEQ produces few false negatives because the algorithm is conservative while making wildtype calls as an incorrect call would amount to a completely missed variant allele. However, unresolved bases need visual inspection or have to be interrogated with dideoxy sequencing.

<table>
<thead>
<tr>
<th></th>
<th>006</th>
<th>159</th>
<th>162</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>010</td>
<td>150</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>030</td>
<td>164</td>
<td>165</td>
</tr>
<tr>
<td></td>
<td>032</td>
<td>162</td>
<td>162</td>
</tr>
<tr>
<td></td>
<td>035</td>
<td>151</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>044</td>
<td>129</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td>050</td>
<td>135</td>
<td>139</td>
</tr>
<tr>
<td></td>
<td>051</td>
<td>144</td>
<td>147</td>
</tr>
<tr>
<td></td>
<td>058</td>
<td>140</td>
<td>145</td>
</tr>
<tr>
<td></td>
<td>059</td>
<td>134</td>
<td>137</td>
</tr>
<tr>
<td></td>
<td>061</td>
<td>134</td>
<td>141</td>
</tr>
<tr>
<td></td>
<td>153</td>
<td>154</td>
<td>157</td>
</tr>
<tr>
<td></td>
<td>241</td>
<td>151</td>
<td>155</td>
</tr>
</tbody>
</table>

**Table 4:** Number of exons in each array with no-calls or variant calls (total exons=180).
False positives or no-calls do not represent a lack of test sensitivity, as they are followed up by dideoxy sequencing but they compromise the cost-effectiveness of the technology if a large amount of sequencing is required. Table 4 lists the number of exons for each array that need to be interrogated with dideoxy sequencing to follow-up on no-calls and/or variant calls.

3.3.2 Impact of GC content on probe performance

**Probe GC-content and variant detection.** For each of the 13 optimized arrays, variants were classified based on whether they could be identified correctly or if they were called wild-type or not called by GSEQ. A total of 240 variants (59 unique) detected in the 13 samples by dideoxy sequencing were used for the analysis. GC-content was calculated for probes that interrogated variant positions on the array. GSEQ correctly identified 184 of these variants while 45 were not called and 2 were called wild-type. Mean and median GC-contents were (55.7%, 56%) for correctly identified variants and (64.8%, 64%) for not called and missed variants. Anderson-Darling test on the two groups indicated a non-normal distribution, thus a non-parametric test was used. Wilcoxon rank sum test rejected the null hypothesis of equal median GC-content for the two groups with a p-value of 1e-07 at 5% significance level, indicating a potential impact of GC-content on variant detection.

**Differential impact of high C- and G-content on probe performance.** Average probe call rate and call accuracy were evaluated as functions of probe GC-content (Figure 6).
As GC-content increases, average probe call rate decreases and the variance increases.

Call accuracy also showed a slight degradation and higher variance for GC>68%.

Since efficient target-specific hybridization is critical for a base to be called, we evaluated the impact of probe GC-content on peak feature intensity. Average peak intensity is low at very low and very high GC-content but probes with high GC-content show higher variance (Figure 7).

**Figure 6:** Relationship between probe GC content and array call rate and accuracy.

**Figure 7:** Relationship between probe GC content and peak feature intensity and ratio of peak to next highest intensity.
To determine if probe C- and G-contents had similar impact on hybridization intensity, we plotted average peak intensity with respect to probe C- and G-contents instead of a combined GC-content. We used complementary feature quartets to determine intensities associated with the C- or G-content of a probe. In agreement with previously reported findings (Cutler et al. 2001, Zhan, Kulp 2005), we observed that an increase in probe G- and C-contents have differential impact on performance. Intensity characteristics vary differently with respect to an increase in C- or G-content and average peak intensity is affected more severely by a high G-content than by an equally high C-content (Figure 8).

We sought to determine if continuous stretches of G’s and C’s show distinct hybridization properties. We selected all runs of G’s and C’s (≥1 bases) within probes on both strands and calculated the average peak intensity and ratio of peak to next highest intensity for all constituent bases within the run. We then plotted these values as functions of the number of continuous bases in C- and G-stretches. Hybridization intensity appears to be able to maintain robustness despite an increase in C’s whereas an equivalent increase in the number of G’s in a stretch causes a drop in intensity (Figure 9).
Figure 8: Differential dependence of peak feature intensity on probe G- and C-content. G-richness of a probe has a more severe impact on hybridization intensity than C-richness.

Figure 9: Relationship between number of bases in a continuous stretch of C’s/G’s and peak feature intensity and ratio of peak to next highest intensity.
In order to assess if our data demonstrated previously reported debilitating effect of stretches of Gs (G-stacks) on probe performance, we compared hybridization intensities among probes with the same G-content grouped based on the presence or absence of G-stretches ($\geq 4$ Gs in a continuous stretch). For the same G-content, probes with G-stretches produce lower peak intensities than probes with C-stretches or without any stretches of Gs or Cs (Figure 10).

We also generated consensus sequences for probes corresponding to positions with no-calls, false positive calls and true wild-type. Figure 11 shows that unresolved and false positive positions were frequently surrounded by G/C bases whereas no such consistent pattern was observed for true wild-type positions.

**Figure 10:** Differential impact of high probe G-content and C-content on probe performance; G-stretches degrade peak intensity. Peak feature intensity for probes with same G-content grouped based on presence of G-stretch, C-stretch and no continuous stretches. Error bars represent one standard deviation.
Figure 11: Probe consensus sequences for no-calls, false positive and wild-type calls. Positions with no-call or false positive calls (top) have G or C as the neighboring bases more often than correctly identified wild-type bases (bottom).

3.4 Discussion

Previous resequencing array-based studies (Liu et al. 2007, Denning et al. 2007, Fokstuen et al. 2008, Waldmüller et al. 2008, Lebet et al. 2008) have reported call rates ranging from 93.5% to 98% with GSEQ, which employs the adaptive background genotyping-calling scheme developed by Cutler and colleagues (Cutler et al. 2001). While our call rates are within the high end of reported ranges, a large percentage (~80%) of the tiled exons required sequencing to follow-up on ambiguous calls, representing a limitation to clinical application of the technology under the current methodologies.

The efficiency of sequence-specific hybridization is dependent on the properties of the probe and target sequences (Hacia 1999). High GC-content, presence of a nearby SNP and cross-hybridizing sequences are known to affect base-calling, thus limiting the capacity of resequencing microarrays. In agreement with our findings, mutations in GC-
rich regions have previously been reported to be missed (Waldmuller et al. 2008). While it is well known that the GC-content of a probe can impact hybridization (http://www.affymetrix.com/support/technical/technotes/customseq_arraybase_technote.pdf), (Denning et al. 2007) C-rich probes perform better than G-rich probes for the identical site when complementary strand quartets are compared (Zhan, Kulp 2005) and fluorescence intensity declines with G-richness of a probe (Cutler et al. 2001). Additionally, bases within the G-stretch of a probe produce lower peak intensities, especially for stretches with \( \geq 4 \) continuous Gs. It has previously been suggested that probes with multiple Gs in a row (G-stacks) tend to have higher cross-hybridization signals possibly caused by formation of G-quartets due to multiplex binding (Wu et al. 2007). Taken together, these observations imply that sites interrogated with G-rich probes may show stronger signal on the complementary strand employing C-rich probes.

Another study reports that nearly 80% of no-calls can be resolved by visual inspection of the intensities as one of the strands provides a clear signature for these positions (Waldmuller et al. 2008). However, there are no existing computational approaches that leverage such sequence-specific characteristics in an attempt to resolve GSEQ no-calls that have a distinct signature on one strand but are still ruled no-call due to improper hybridization on the complementary strand.

GSEQ is known to produce very few false negatives, thus providing a highly sensitive test. However, follow-up dideoxy sequencing for resolution of no-calls leads to an additional variable cost, a factor which needs to be carefully considered for clinical application of the technology. To this effect, we propose a novel algorithm for resolution
of no-calls from GSEQ. It should be noted that the algorithm is not designed to be an alternative to GSEQ. Instead, it provides an optional step for salvaging unresolved bases from GSEQ before initiating confirmatory dideoxy sequencing.

The critical performance characteristics we attempted to understand and optimize are call rate, sensitivity and specificity. Affymetrix GSEQ is an upgraded version of the GDAS base-calling software and offers some additional features as described in the GSEQ technical datasheet (http://www.affymetrix.com/support/technical/datasheets/gseq_datasheet.pdf). However, they both employ a base-calling algorithm built upon the adaptive background genotyping-calling scheme (ABACUS) developed by Cutler and colleagues (Cutler et al. 2001). GDAS and GSEQ produce few false negatives because these algorithms are conservative in making wild-type calls. False positives or no-calls do not represent a lack of test sensitivity when they are followed up by dideoxy sequencing; however, they compromise the cost-effectiveness of the technology if a large amount of sequencing is required. In our hands, the cost reduction is roughly a 25%-50% reduction compared to traditional capillary sequencing when thorough follow-up is employed to resolve all variant calls and rare no-calls. The exact reduction depends on the degree of multiplexing employed in the up-front PCR step and the amount of follow-up sequencing that is needed. The latter factor is unique to each test depending on the sequences included, PCR robustness, the amount of DNA variation in the regions tested and degree of bioinformatics and test optimization that has been achieved.

It has previously been suggested that large PCR amplicons do not hybridize efficiently to immobilized probes possibly due to steric constraints on the approach of the
target DNA (Shchepinov, Case-Green & Southern 1997) and this finding has been taken into consideration during design of nucleic acid amplification strategies (Vora et al. 2004). Optimizing the target amplification process to include only short range PCR improved the overall array performance in the Cincinnati arrays, thus providing further evidence for the relationship between PCR amplicon length and hybridization efficiency. It should be noted that the data generated by the Harvard arrays was based upon a combination of short range and long range PCR. However, the Harvard group has also discontinued use of long range PCR in subsequently developed array-based sequencing tests. This is because the efforts needed to continually optimize the fragmentation of long range PCR fragments and the additional limitations caused by diminished DNA quantity and variable amplification efficiency in long range PCR do not outweigh the benefits. For most nuclear genes with dispersed exons, only one to a handful of exons can be combined into a long range PCR reaction limiting the efficiency gained by this approach. In contrast, amplification of long stretches of contiguous interrogated DNA, such as that present in the mitochondrial genome, enables the highest efficiency savings for long-range PCR approaches.

Average call rates of 96.9% and 97.7% for all of Harvard arrays and the optimized Cincinnati arrays respectively were achieved using GDAS/GSEQ. Previous resequencing array-based studies (Liu et al. 2007, Denning et al. 2007, Fokstuen et al. 2008, Waldmuller et al. 2008, Lebet et al. 2008) have reported call rates ranging from 93.5% to 98% with GDAS/GSEQ. While our call rates are within the high end of reported ranges, a large percentage (~80%) of the tiled exons required sequencing to
follow-up on ambiguous calls, representing a limitation to clinical application of the technology under the current methodologies.

While it is well known that the GC-content of a probe can impact hybridization (http://www.affymetrix.com/support/technical/technotes/customseq_arraybase_technote.pdf), (Denning et al. 2007), C-rich probes perform better than G-rich probes for the identical site when complementary strand quartets are compared (Zhan, Kulp 2005) and fluorescence intensity declines with G-richness of a probe (Cutler et al. 2001). Additionally, bases within the G-stretch of a probe produce lower peak intensities, especially for stretches with \( \geq 4 \) continuous Gs. It has previously been suggested that probes with multiple Gs in a row (G-stacks) tend to have higher cross-hybridization signals possibly caused by formation of G-quartets due to multiplex binding (Wu et al. 2007). We also observed that positions that could not be called or were incorrectly called variant by GSEQ showed a higher likelihood of having G/C as the neighboring bases as compared to correctly identified wild-type positions.

The observations taken together indicate some of the limitations of Affymetrix GSEQ that lead to an overhead when the technology is applied clinically. As discussed earlier, oligonucleotide-based hybridization is sensitive to attributes such as GC content and presence of G-stretches. Chapter 4 discusses how we leveraged some of these properties to develop a method for improved call coverage with resequencing microarrays.
Chapter 4. Base-calling optimization with sPROFILER and implementation of a bioinformatics pipeline for clinical application of resquencing microarrays

4.1 Introduction

4.1.1 Leveraging sequence-specific hybridization properties for improved base-calling

C-rich probes perform better than G-rich probes for the identical site when complementary strand quartets are compared. The observation implies that sites interrogated with G-rich probes may show stronger signal on the complementary strand employing C-rich probes. Additionally, a previous study suggests that majority of no-calls (> 80%) can be resolved by inspecting intensities on one of the strands (Waldmuller et al. 2008). No-calls can potentially be resolved by utilizing a distinct signal on either strand to complement the GDAS/GSEQ base-calling scheme for 20 μm arrays, which requires high quality signals on both strands. By adopting such an approach, a position can be called even if one strand performs poorly but the complementary strand produces a clear signal. Stretches of no-calls indicate improper hybridization and therefore base calls lying within such regions are sometimes unreliable. Such calls can be converted into no-calls so that these positions can be confirmed by dideoxy sequencing instead of being assigned a possibly inaccurate call.

In the subsequent sections, we discuss the development of a computational algorithm to utilize some of the properties associated with sequence-specific
hybridization that can potentially lead to a reduction in unresolved calls obtained from Affymetrix GDAS/GSEQ.

**4.1.2 Reduction of false positive variant calls**

False positive calls represent another limitation to the clinical application of resequencing arrays. As discussed in Chapter 3, we observed ~40 false positive calls per array for our data. If each of these calls needs to be confirmed with dideoxy sequencing and the calls are dispersed across several unique exons, it would lead to an overhead in terms of cost and time associated with analysis of each sample. A scheme for reducing false positive calls has been previously developed for array-based whole-genome resequencing of *Francisella tularensis* (Pandya et al. 2007). Pandya et al developed a set of bioinformatic filters targeting systematic base-calling errors caused by cross-hybridization between the sample and the probes. We discuss some of these filters and their eventual adoption in reduction of false positive calls produced with resequencing arrays.

**4.2 Materials and Methods**

**4.2.1 Development of sPROFILER**

To improve call rates and, in some cases, reduce false positives, a new computer algorithm (sPROFILER) was developed that analyzes feature intensity profiles of no-call and variant positions obtained after GSEQ analysis. sPROFILER attempts to find distinct intensity signatures on either strand so that a position can be called even if one strand performs poorly but the complementary strand produces a clear signal. Base call quality
scores are assigned to new calls to indicate the strength of the evidence available for making the call.

The algorithm uses all wild-type bases within the array to determine threshold for peak to next highest intensity ratio on either strand and uses the base call at the position of interest across all arrays to determine the proportion of wild-type calls that are made on that position by GSEQ. The latter is used for scaling the threshold ratio and thus, in effect, making the algorithm more conservative while attempting to assign a wild-type call to a position that is being called variant in a large number of samples and vice-versa. We also adopted two additional bioinformatics filters developed for the Francisella tularensis whole-genome resequencing platform (Pandya et al. 2007). The filters were designed for reduction in false positives by screening variant calls that 1) are in regions rich in variant calls and no-calls, and 2) have low quality scores for the corresponding base call. With the exception of wild-type calls within no-call stretches, sPROFILER does not attempt to re-examine any base calls conforming to the reference sequence because GSEQ is conservative in making wild-type calls and thus achieves low false negative rates. The output calls were compared against sequencing results and against GSEQ calls for validation. sPROFILER was implemented in MATLAB. The workflow for analysis of no-calls with sPROFILER encompasses the following steps:

- Analyze hybridization intensity data with GSEQ to generate base-calls.
- Extract intensity data, reference sequence and calls made by GSEQ for all arrays to be analyzed with sPROFILER.
Run sPROFILER on array data to resolve GSEQ no-calls.

4.2.2 sPROFILER: Methodology and algorithm design

sPROFILER analyzes every chip individually and takes the reference sequence, the final sequence obtained from GSEQ, and the raw intensity values for all sites as input. We have incorporated validation capability in the program so that it gives us all the data pertinent to its performance when compared against the “Gold Standard” which are dideoxy sequencing results. It should be kept in mind that sequencing information is not used in any way to affect the resequencing output. It is solely used for validation to determine how many true variants are missed or detected by GDAS/GSEQ and the program. The program requires input files in the form of CEL and CHP data for each array to be analyzed. Following are the different stages involved in analysis of a single array:

1. Read intensities and calls from CEL and CHP files for the array.
2. If >75% of neighboring calls are variant or No-calls in a +/-12-bp window, convert the base call to a no-call and do not attempt to resolve it. If GSEQ Quality Score (QS) for a variant call is <55, convert it to No-call (false positive filters adapted from (Pandya et al. 2007)).
3. Check if ONE of the strands shows a clear peak AND corresponds to the refseq. The site is called based on the peak intensity if the following conditions are met:
   - Maximum intensity on sense or anti-sense matches the base on refseq.
- Maximum value exceeds the next highest by a certain factor (see the next section on comparison of threshold ratios).

4. If the above condition is not met, assign a ‘No Call’ to the site and if the site was being called by GSEQ, retain that call.

5. Assign a flag to each call as follows. This would help the user in filtering the calls while analyzing the data.

   - If the call from GSEQ matches refseq, Flag=0;
   - If no-call is converted to refseq, Flag=1;
   - If GSEQ detected a Het on the site, Flag=2;
   - If GSEQ detected a homozygous variant on the site, Flag=3;
   - If the site is still a ‘No Call’, Flag =4;

6. Print results for each array comprising of GSEQ calls, refseq, calls made by the program, quality score, flag value, and the 8 intensities. Figure 12 illustrates a sample of the output file.

When sequencing results are available for an array, the program can be used to validate the output against sequencing data. The user needs to create a file, which comprises of all the locations at which a variant was found with sequencing along with the actual base called at that location. All the steps remain the same as in the case where we do not have sequencing data available. However, additional information is generated which indicates how many true variants are missed or
detected in the final resequencing output. All the known variant locations are listed along with the call obtained from sequencing, GSEQ and sPROFILER (Figure 13).

### 4.3 Results

#### 4.3.1 Determination of optimal threshold ratios

Five different methods for determining site-specific peak to next highest intensity ratios for making wild-type calls with sPROFILER were evaluated, namely:

---

**Figure 12:** Sample output file with results for a single chip (ChipNum\_final.txt)

**Figure 13:** Output file containing calls made by different programs for positions known to have variants (ChipNum\_OutKnownSnp.txt)
- average threshold ratio obtained from using all wild-type calls from GSEQ within the same array (Within-array)
- average threshold ratio obtained from all wild-type calls on the site of interest across all arrays (Across-arrays; exception: if no wild-type call was made at a particular position across arrays, Within-array ratio was used)
- combination of Within-array ratio and a scaling factor, obtained from the proportion of base calls (exclude no-calls) for the site that were called wild-type by GSEQ across all arrays
- combination of Across-arrays and the scaling factor
- combination of Within- and Across-arrays

Table 5 presents results from comparing different schemes for calculation of site-specific threshold ratios for making wild-type calls. We used 13 optimized arrays for the analysis. GSEQ provided highest sensitivity but lowest specificity and call coverage. Across all methods, highest sensitivity was obtained when the threshold was calculated by using average peak to next highest intensity ratio across all wild-type calls within the same array and scaling it with the proportion of base calls at the site that were called wild-type across all arrays (Within-array + Scaling in Table 5). The remaining methods provided higher specificity and fewer no-calls but lower sensitivity. We chose the most conservative method (Within-array+Scaling) for our analysis in order to minimize the increase in number of false negatives while reducing no-calls and false positives.
4.3.2 Array performance with sPROFILER

sPROFILER was only used on no-calls from GDAS/GSEQ. Examined bases were called wild-type based on single-strand evidence or were left as no-calls depending upon the feature intensity profile. Base calls were subsequently compared against GSEQ and dideoxy sequencing calls. Table 6 and Figures 14-16 provide detailed comparison of call rates, number of false positives and false negatives before and after analyzing GSEQ calls with sPROFILER. For Cincinnati arrays run with short and long range PCR fragments, the average call rate increased to 96.7% (99.82% call accuracy). The average number of no-calls dropped from 2350 to 902 per array and the number of false positives dropped from 180 to 44 bases per array. Improvement was obtained at the cost of incorrectly assigning an additional 0.38 true variants per array as wild-type. For the optimized Cincinnati arrays, average call rate increased to 99.6% (range 99.5% to 99.8%) with 99.88% call accuracy and number of no-calls dropped from 563 to 103 (Table 6). The number of false positives dropped from 42 to 30 per array with the application of filters to screen variant calls based on low quality scores and the number of no-calls and variant calls in the vicinity. Improvement was achieved at the cost of an additional 0.4 false negative per array. We also calculated the number of exons per array that would need to

<table>
<thead>
<tr>
<th></th>
<th>No. of no-calls*</th>
<th>FP*</th>
<th>FN*</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSEQ</td>
<td>7317 (563)</td>
<td>549 (42)</td>
<td>6 (0.46)</td>
<td>96.92</td>
<td>97.77</td>
</tr>
<tr>
<td>Within-array</td>
<td>810 (62)</td>
<td>398 (30)</td>
<td>15 (1.15)</td>
<td>93.04</td>
<td>99.66</td>
</tr>
<tr>
<td>Across-arrays</td>
<td>738 (57)</td>
<td>398 (30)</td>
<td>17 (1.30)</td>
<td>92.44</td>
<td>99.68</td>
</tr>
<tr>
<td>Within-array+Scaling</td>
<td>1338 (103)</td>
<td>398 (30)</td>
<td>11 (0.85)</td>
<td>94.61</td>
<td>99.51</td>
</tr>
<tr>
<td>Across-arrays+Scaling</td>
<td>745 (57)</td>
<td>398 (30)</td>
<td>17 (1.30)</td>
<td>92.44</td>
<td>99.68</td>
</tr>
<tr>
<td>Within-array+Across-arrays</td>
<td>456 (35)</td>
<td>398 (30)</td>
<td>17 (1.30)</td>
<td>92.44</td>
<td>99.76</td>
</tr>
</tbody>
</table>

Table 5: Comparison of schemes for calculation of site-specific threshold for peak to next highest intensity ratio
*values represented as Total bases and (per chip average).
be interrogated with dideoxy sequencing in order to resolve no-calls and confirm positive calls. After processing GSEQ calls with sPROFILER, the average number of exons to be sequenced in order to clarify no-calls and confirm variants dropped from 150 (range 132 to 165) to 68 (range 52 to 82) per array.

Figure 14: Improvement in array call rates with protocol optimization and application of sPROFILER to GSEQ calls (data shown for Cincinnati arrays). Data is separated into two categories based upon protocol (short and long range PCR vs. short range only PCR) and then arranged in ascending order of GSEQ call rates. (Kothiyal et al. 2010)
Figure 15: False positive calls with and without protocol optimization/sPROFILER. No-calls and positive calls were processed for the first 12 while only no-calls were processed for the remaining 13 chips. No-calls were converted to wild-type, left as no-call, or were assigned a variant call. Chips that were analyzed only for no-calls may show an increase in false positives due to conversion of a fraction of no-calls to variant calls, some of which are not true variants. (Kothiyal et al. 2010)

Figure 16: False negative calls with and without protocol optimization/sPROFILER represented as a portion of total true variants. (Kothiyal et al. 2010)
Table 6: Overall array performance with and without application of sPROFILER to GDAS/GSEQ base calls. (Kothiyal et al. 2010)

Percentages are obtained by averaging individual percent values over all arrays.
A: Bases called/total bases on array
B: Correct calls/total calls
C: Wild-type bases incorrectly identified as variants/total variant calls * 100%
D: True variants incorrectly called wild-type/total true variants * 100%
E: Wild-type bases incorrectly identified as variants/total wild-type calls * 100%
F: True variants incorrectly called wild-type/total true calls * 100%
G: Average number of bases not called per array
H: Number of exons that need follow-up sequencing to interrogate no-calls or variant calls
I: 14 Harvard arrays with full dideoxy sequencing results were used for determination of false negatives and overall accuracy. However, no-calls and variant calls across all 26 Harvard arrays were confirmed by sequencing.

4.3.3 Bioinformatics pipeline for clinical application of the Hearing Loss Gene Chip

A pipeline was developed for the application of the Hearing Loss Gene Chip in the clinical laboratory at Cincinnati Children’s Hospital Medical Center. The 13 optimized arrays have been used as the baseline reference batch for calculation of threshold ratios in sPROFILER and for generation of calls in GSEQ. Once the input CEL and CHP files have been saved by the user, an MS-Dos batch file can be executed to run
all analysis steps including processing of input data, analysis of no-calls with sPROFILER and generation of a final report, contents of which are described below in detail. The pipeline has been implemented in Perl and sPROFILER runs as a Matlab program. Once execution has completed, the program automatically generates a report file, which appears automatically after the process is completed successfully in notepad or the default text editor as shown. The file lists all variant and no-calls that remain after sPROFILER analysis. It should be noted that some of the variant calls in GSEQ will be converted to No-calls with sPROFILER due to low quality score or if they are in a window with neighboring no-calls. The final report has the information shown below.

<table>
<thead>
<tr>
<th>Names</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>AffyID</td>
<td>Affymetrix base ID</td>
<td>5892</td>
</tr>
<tr>
<td>Frag</td>
<td>Reference fragment name</td>
<td>CDH23_ex6_68-43</td>
</tr>
<tr>
<td>CDS</td>
<td>Aligned Coding Sequence (CDS) and Database significance information</td>
<td>366c.366T&gt;C,p.Val122Val</td>
</tr>
<tr>
<td>Refseq</td>
<td>Reference base</td>
<td>t</td>
</tr>
<tr>
<td>exp.1_sPROFILER</td>
<td>All variant calls and no-calls that remain after sPROFILER application to GSEQ calls</td>
<td>y</td>
</tr>
<tr>
<td>exp.1_GSEQ</td>
<td>The most likely base according to the Affymetrix resequencing algorithm</td>
<td>y</td>
</tr>
<tr>
<td>exp.1_Qual</td>
<td>Quality score</td>
<td>232.348</td>
</tr>
<tr>
<td>exp.1_exonCallRate</td>
<td>Exon call rate (%)</td>
<td>97.9</td>
</tr>
<tr>
<td>GC</td>
<td>Probe GC content (%)</td>
<td>48</td>
</tr>
<tr>
<td>NSeq</td>
<td>nearby sequence of 2 nt</td>
<td>gg</td>
</tr>
<tr>
<td>NCalls</td>
<td>Total number of samples without calls at this position (out of 13)</td>
<td>12</td>
</tr>
<tr>
<td>NCips</td>
<td>Samples without calls at this position</td>
<td>0</td>
</tr>
<tr>
<td>NumWT</td>
<td>Number of wild type samples sequence verified</td>
<td>3_13</td>
</tr>
<tr>
<td>FPcalls</td>
<td>Total number of samples with False Positive Calls</td>
<td>2</td>
</tr>
<tr>
<td>FPChips</td>
<td>Samples with False Positive Calls</td>
<td>030,051</td>
</tr>
<tr>
<td>BaseExCallRate</td>
<td>Exon call rate (%) GSEQ</td>
<td>97.2</td>
</tr>
<tr>
<td>VarOrN</td>
<td>Variants and no-calls in the analyzed chip</td>
<td>Var</td>
</tr>
<tr>
<td>Unique</td>
<td>Unique co-calls</td>
<td>UniqueN</td>
</tr>
</tbody>
</table>

**Table 7:** Fields included in the output report for array based analysis.
4.4 Discussion

By applying additional computer algorithm (sPROFILER) to reduce GSEQ no-calls by utilizing a distinct signal on either strand, 86% and 82% of the no-calls could be resolved for Harvard and Cincinnati arrays respectively while maintaining overall accuracy at $\geq 99.8\%$. The implementation of the sPROFILER led to a substantial reduction in exons needing follow-up (average of 153 down to 52).

Small indels constitute nearly 24% of disease-causing mutations in the Human Gene Mutation Database as of November 2009 and have been shown to cause severe phenotypic variability (Lin et al. 2004, Othman et al. 2005). The inability of resequencing arrays to have high sensitivity for detecting novel indels, especially those involving only a few base pairs, presents a significant limitation (Denning et al. 2007, Liu et al. 2007, Fokstuen et al. 2008). Regions showing aberrant hybridization patterns can be selected for confirmatory dideoxy sequencing to potentially detect variations (including indels) that are missed with resequencing arrays. However, small indels present challenges as they sometimes do not lead to easily discernible variability in hybridization patterns. After interrogating three different deletions on our arrays, we could detect only the largest deletion through a series of no-calls and a variant call in the region. Further algorithmic and technical improvements could entail development of a scheme for detection of indels by virtue of identifying a regional drop in signal intensity.
We developed a bioinformatics pipeline to support clinical application of the Hearing Loss Gene Chip at Cincinnati Children’s Hospital Medical Center. We integrated sPROFILER into the pipeline to reduce the number of unresolved bases and subsequently the number of follow-up assays with dideoxy sequencing.

Irrespective of the technology used for detection of sequence variants, their interpretation in light of their clinical impact is the next obvious step required for a genetic test. Chapter 6, 7 and 8 discuss some of the existing methods for variant classification, their associated strengths and weaknesses, and we finally describe the development and validation of a computational method for classification of missense variants.
Chapter 5. Survey of existing methods for functional interpretation of missense genetic variants

5.1 Significance of functional interpretation of variants

High-throughput sequencing methods and rapid progress in the identification and mapping of Single Nucleotide Variants (SNVs) have generated an opportunity for exploring the functional and possibly phenotypic effects of genetic variation (Mooney 2005). Interpreting the putative impact of a genetic variant on the encoded protein is a critical step in designing disease-centric genetic tests and determining variants that affected individuals should be screened for. An interpretation mechanism becomes especially necessary in cases where novel variants with unknown significance are discovered in a patient. Determining the genetic etiology of a phenotype requires identification of putatively pathogenic variants in the patient’s genotype. Association studies are usually undertaken to identify variants that contribute to a complex disease. A common strategy adopted in design of such studies is limiting markers to coding regions of genes that have shown some evidence of association with the trait of interest (Nakken, Alseth & Rognes 2007). A way of prioritizing markers that have a higher likelihood of being functional is critical for direct association studies.
5.2 Existing tools for annotation of missense variants

Causal variants could fall within exonic or intronic regions of a gene or outside of a gene region. Variants within a gene can be classified as coding or non-coding depending on whether they overlap with the coding segment of the gene or not. Coding SNVs occur in the protein coding region (exons) of the genome. Missense or non-synonymous SNVs are a subset of coding SNVs that cause a change in the coded amino acid of the protein product. A change in amino acid may or may not affect protein structure depending on the properties of the wild-type and mutated amino acid residues. If a substitution results in a mutated residue that has very different properties from the original residue, the likelihood of protein structure, function and interactions being affected increases. Whereas a large number of such variants may be functionally neutral, other may have a debilitating effect on the protein function and hence may be disease associated. High-throughput methods of variant detection lead to the next challenge of predicting which of them are potentially disease-associated. So far, in vitro functional tests for confirmation of pathogenic mutations are restricted to only a few genes. Given the unavailability of and complications associated with such functional tests, in silico methods have been developed to discriminate disease-associated missense variants from neutral variants. Subsequent sections present an overview of some of the available methods and discuss limitations associated with each one of them.

Bioinformatics approaches have been applied to the analysis of disease-associated mutations and there have been several recent studies focusing on the properties and
localization of disease-causing mutations (Mooney 2005). Studies have shown that positions harboring disease-causing mutations are evolutionarily conserved and that they can potentially affect protein structure (Mooney, Klein 2002, Saunders, Baker 2002, Sunyaev et al. 2001, Chasman, Adams 2001, Ng, Henikoff 2002, Krishnan, Westhead 2003, Sunyaev, Ramensky & Bork 2000). Previous evidence suggests that regulatory and coding variants have a higher likelihood of being disease-causing (Chakravarti 1998, Collins, Guyer & Charkravarti 1997). Therefore, several studies have been undertaken towards development of methods for predicting the impact of nsSNPs. A comprehensive discussion of various methods for annotation of human SNVs has been provided in (Karchin et al. 2007). According to the survey, SNV annotation tools can broadly be divided into the following three categories:

- **Methods servers** generate results using their own computational methods (e.g. SIFT, PolyPhen, PANTHER, nsSNPAnalyzer, SNAP, PMUT, etc.).
- **Meta servers** extract information from different servers but do not generate original results with a computational method of their own (e.g. PolyDoms, MutDB, Snap, SNP@Domain, SNPSelect etc.).
- **Hybrids** extract information from other servers and generate results using their own computational methods (e.g. LS-SNP, SNPeffect, SNPs3D, FAST-SNP, etc.).

None of the Meta servers or Hybrids, with the exception of FAST-SNP, allow protein sequences to be submitted. For the purpose of this dissertation, we focus only on tools
that allow analysis of novel variants since a substantial portion of missense variants we attempt to annotate as part of the work presented here have not been identified before and therefore can only be analyzed by the Methods Servers.

SIFT, PolyPhen and PANTHER are some of representative tools for classification of missense variants. The following section discusses these methods and highlights their associated strengths and limitations.

5.2.1 SIFT (Sorting Intolerant From Tolerant)

SIFT uses sequence homology to predict whether an amino acid substitution will have an impact on protein function (Ng, Henikoff 2003). It is based on the assumption that amino-acid positions that are important for the correct biological function of the protein are conserved across protein family and/or evolutionary history. SIFT accepts protein sequence and the position of substitution as the input and predicts whether the residue is tolerated or deleterious. However, there has been evidence that SIFT fails to identify residues that are vital for protein function but are not conserved throughout the family (Ng, Henikoff 2001). As is the case with all methods using multiple sequence alignment, the accuracy of SIFT heavily depends on availability of a large number of homologous sequences.

5.2.2 PolyPhen (Polymorphism Phenotyping)

PolyPhen uses protein structure data (from Protein Data Bank), evolutionary information, data from DSSP (Dictionary of Secondary Structure in Proteins), and 3D structure (when available) to determine if a variant may have an effect on secondary
structure of the protein, interchain contacts, functional sites and binding sites (Ramensky, Bork & Sunyaev 2002b). PolyPhen accepts protein sequence and amino acid substitution as the input. Although PolyPhen uses protein structural information when it is available, majority of the proteins do not have 3D structures available. For example, only 2 out of the 8 hearing loss genes have resolved structures. In such cases, predictions are based on amino acid residue properties and sequence alignment scores (Sunyaev et al. 1999). Similar to the principle behind the SIFT algorithm, PolyPhen then calculates the probability that a substituted amino acid will be tolerated, based on a multiple sequence alignment.

5.2.3 PANTHER (Protein Analysis Through Evolutionary Relationships)

PANTHER scores nsSNPs against Hidden Markov Model for protein families to estimate their likelihood of disrupting conserved amino acid elements, and thus protein function (Mi et al. 2007). It calculates the subPSEC (substitution position-specific evolutionary conservation) score based on an alignment of evolutionarily related proteins.

5.3 Discussion

Although various in silico methods exist for interpretation of missense variants, each is associated with its own strengths and drawbacks. All of the Methods Servers can analyze novel missense variants by allowing submission of protein sequence and corresponding amino acid change. However, none of the Meta servers or Hybrids allow analysis of novel variants. Only FAST-SNP can analyze novel variants by allowing
submission of DNA sequence and the corresponding nucleotide change. The inability of Meta Servers and Hybrids to handle novel variants restricts their application.

SIFT, PolyPhen and PANTHER are some of the representative tools used for prediction of deleterious missense variants. SIFT (Ng, Henikoff 2003) relies on alignment of orthologous and/or paralogous protein sequences to predict the likelihood with which a site of substitution can mutate. However, there has been evidence that SIFT fails to identify residues that are vital for protein function but are not conserved throughout the family (Ng, Henikoff 2001). PolyPhen (Ramensky, Bork & Sunyaev 2002b) utilizes a combination of conservation scores and structural data when available to assess the functional significance of an amino acid substitution. Although PolyPhen includes structural information, number of known protein sequences significantly surpasses the number of protein structures (Nakken, Alseth & Rognes 2007). When structural data is absent, predictions are based on amino acid residue properties and sequence alignment scores (Sunyaev et al. 1999). PANTHER (Mi et al. 2007) scores nsSNPs against Hidden Markov Model for protein families to estimate their likelihood of disrupting conserved amino acid elements, and thus protein function(Mi et al. 2007).

Since each method is associated with its own strengths and limitations, we assessed different prediction results and properties that are potential indicators of the severity of an amino acid substitution to develop an integrated framework for classification of missense variants. Chapter 7 discusses the development of a machine-learning based prediction tool and Chapter 8 presents the results of testing the method on a validation dataset.
6.1 Introduction

High-throughput sequencing methods and rapid progress in the identification and mapping of Single Nucleotide Variants (SNVs) have generated an opportunity for exploring the functional and possibly phenotypic effects of genetic variation (Mooney 2005). Several genes underlying human genetic diseases have been characterized and are available for genetic screening in diagnostic laboratories. As a result of such genetic tests, many variations including SNPs and pathogenic mutations are routinely identified in affected individuals. Missense mutations arising from non-synonymous SNPs (nsSNPs) account for nearly half of all allelic variations associated with inherited disorders in humans (Hamosh et al. 2005, Stenson et al. 2003). Evaluating the putative impact of a genetic variant on the encoded protein is a critical step in designing and interpreting disease-centric genetic tests.

In vitro functional tests for validation of pathogenic mutations are limited to only a few genes and are not easily available in clinical diagnostic settings. Various attempts have been made to develop prediction tools that can ascertain if an nsSNP affects protein function and subsequently contributes to genetic disease. Studies have shown that positions harboring disease-causing mutations are evolutionarily conserved and that they can potentially affect protein structure (Mooney, Klein 2002, Saunders, Baker 2002,
Sunyaev et al. 2001, Chasman, Adams 2001, Ng, Henikoff 2002, Krishnan, Westhead 2003, Sunyaev, Ramensky & Bork 2000). Previous evidence suggests that regulatory and coding variants have a higher likelihood of being disease-causing (Chakravarti 1998, Collins, Guyer & Chakravarti 1997). Therefore, several studies have been undertaken towards development of methods for predicting the impact of nsSNPs. Almost all methods for prediction of deleterious variants utilize different properties related to amino acid properties, protein structure, and evolutionary conservation. These properties have been shown to differ between disease-causing mutations and neutral polymorphisms (Saunders, Baker 2002). This premise has been used in the development of various classification models (Saunders, Baker 2002, Krishnan, Westhead 2003, Bao, Cui 2005, Ferrer-Costa, Orozco & de la Cruz 2005, Ferrer-Costa, Orozco & de la Cruz 2004, Cai et al. 2004). A comprehensive discussion of methods for annotation of human SNVs has been provided in (Karchin et al. 2007).

Several tools using a combinatorial approach comprising of multiple features and prediction results from different sources in a statistical or machine learning framework have been developed in the past (Nakken, Alseth & Rognes 2007, Ng, Henikoff 2003, Bao, Cui 2005, Frederic et al. 2009, Karchin, Kelly & Sali 2005, Ramensky, Bork & Sunyaev 2002a, Won et al. 2008). It has been shown that combining predictions from different in silico algorithms yields better performance than the individual algorithms (Won et al. 2008). We have evaluated SIFT scores, PolyPhen scores, PANTHER scores, minor allele frequency, solvent accessibility at the site of substitution, Grantham score, Yu’s biochemical score, BLOSUM62 score, and changes in mass, volume and
hydrophobicity, as predictors of the debilitating effect of an amino acid substitution using a machine learning framework.

6.2 Materials and Methods

6.2.1 Selection of training data

Swiss-Prot dataset. The Swiss-Prot variant webpage provides a list of single amino acid variants (SAV) associated with human proteins in the knowledgebase (Yip et al. 2004). Each variant is manually categorized as ‘disease’ (associated with disease), ‘polymorphism’ (no known disease association) or ‘unclassified’ (insufficient data to infer classification). Release 56.7 (Jan 2009) of the Swiss-Prot variant webpage consists of 21408 disease variants on 2502 proteins and 29646 polymorphisms on 9728 proteins.

dbSNP and OMIM dataset. Mutations with known disease association were extracted from OMIM and neutral variants were extracted from dbSNP Build 129. In order to obtain a reliable training set, only validated missense variants that did not have any OMIM or Clinical/LSDB annotations were selected for the neutral training set. These additional limits were enforced to ensure that dbSNP variants with any disease association were not incorrectly included as neutral polymorphisms in the training set. Once individual records satisfying the above criteria were obtained from dbSNP (rsID’s) and OMIM (allelic variants in OMIM records), annotations for variants were downloaded as XML files from the two databases. The XML files were subsequently processed to generate a list of all variants that resulted in a single amino acid change in the
corresponding protein. A total of 4508 SAVs were obtained from OMIM and 14327 neutral variants were obtained from dbSNP.

**Generation of training dataset.** It has previously been demonstrated that Swiss-Prot variant set proves to be a reliable training set for human nsSNP predictions (Care et al. 2007). However, the same study also states that variants annotated as neutral polymorphisms can sometimes have an unknown association with disease. While complete exclusion of incorrect annotations in a large training set is difficult to achieve, it is possible to validate variants from one source against another to ensure that classifications are consistent irrespective of the database they are obtained from. Retaining cases that are repeatedly annotated as neutral or disease-causing across different sources and discarding those that have inconsistent annotations can be used to refine the training data at the cost of sacrificing the volume of training instances available. For the purpose of this study, it was stipulated that fewer instances with accurate annotations would produce better classification than a large number of training variants with unreliable class labels.

Only variants that were classified as ‘neutral polymorphism’ in Swiss-Prot and existed in the dbSNP dataset (OMIM variants had already been excluded) were selected for the negative training set. Similarly, the positive training set comprised of variants that existed in OMIM as being related to disease and were classified as ‘disease variant’ in Swiss-Prot. After matching results from the two sources and excluding cases where
protein identifiers could not be mapped, we obtained a training set of 2355 positive instances and 6578 negative instances (Figure 17).

![Figure 17: Extraction of positive and negative instances for training set.](image)

### 6.2.2 Extraction of features

The learning algorithm was trained by associating each variant in the training set with 11 predictive features depicting different attributes of the amino acid change, as listed in Table 8. Each of the features was extracted as outlined below.

**SIFT scores.** SIFT (Ng, Henikoff 2003) scores were obtained by running the program locally on a Linux-based computational cluster. The input provided to the program comprised of the amino acid sequences in FASTA format and the variant positions along with the residue change. A Java program was written to process the input data, generate sets of files with single amino acid sequence and the associated substitutions and then
analyze the variants with SIFT. The SIFT program uses multiple alignment alignments to predict whether a substitution is tolerated or deleterious for every position of interest in the amino acid sequence. SIFT searches for similar sequences using PSI-BLAST, selects sequences that are closely related, generates alignment for the selected sequences, and

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>PANTHER score</td>
<td>Likelihood of variant to cause a functional impact based on alignment of evolutionarily related proteins</td>
<td><a href="http://www.pantherdb.org/">http://www.pantherdb.org/</a> (Thomas et al. 2006)</td>
</tr>
<tr>
<td>Yu’s biochemical score</td>
<td>Derived from the Biochemical matrix based on physiochemical data; a score ≥0.05 represents neutral substitution</td>
<td><a href="http://cmgm.stanford.edu/bioc">http://cmgm.stanford.edu/bioc</a> hem218/Projects%202001/Yu.pdf</td>
</tr>
<tr>
<td>Grantham score</td>
<td>Physiochemical difference between sidechains</td>
<td>(Grantham 1974)</td>
</tr>
<tr>
<td>Change in volume</td>
<td>Volume shift upon substitution</td>
<td>(Zamyatnin 1972)</td>
</tr>
</tbody>
</table>

Table 8: Predictive features evaluated for development of classifier.
calculates likelihood for the amino acid substitutions based on conservation of the site in
the alignment. A likelihood of less than 0.05 indicates a ‘deleterious’ prediction and those
greater than or equal to 0.05 correspond to a ‘tolerated’ prediction. The prediction is
‘unknown’ the query sequence does not share any conserved regions with sequences
found in PSI-BLAST and the program returns an error. Prediction results were available
for 2337 of 2355 positive training instances and for 6546 of 6578 negative instances.

**PANTHER scores.** PANTHER (Thomas et al. 2006) scores were generated by running
the programs locally on a Linux-based computational cluster and providing protein
sequences and corresponding amino acid substitutions as the input. PANTHER calculates
a subPSEC (Substitution Position-specific Evolutionary Conservation) score based on
evolutionarily related proteins. The subPSEC score ranges from 0 (neutral) to -10
(predicted deleterious with high confidence). A score of -3.0 corresponds to a 50%
likelihood of being deleterious. PANTHER is unable to produce a prediction when the
substitution position cannot be aligned with the multiple sequence alignment or related
proteins. PANTHER prediction results were obtained for 1979 of 2355 positive training
instances and for 4219 of 6578 negative instances.

**PolyPhen scores.** Batch query processing scripts were requested and run on the
computational cluster for analyzing variants with PolyPhen (Ramensky, Bork & Sunyaev
2002b). Protein sequences and the corresponding amino acid changes (position and
residue change) were provided as the input. PolyPhen classifies variants as “benign”,
“possibly damaging”, “probably damaging” or “unknown and also includes a PSIC
(Position-Specific Independent Counts) score difference in the output. An “unknown” status is assigned when PolyPhen is unable to make a prediction. PolyPhen predictions could be obtained for 2205 of 2355 positive training instances and for 6308 of 6578 negative instances.

**Minor Allele Frequency.** An XML file consisting of variants within all genes in the training and test sets was extracted from dbSNP and was used to associate each variant with dbSNP rsID based on annotations like protein identifier and amino acid change. The rsID’s were then used to obtain XML files with allele frequencies from dbSNP using NCBI’s E-Utilities (http://eutils.ncbi.nlm.nih.gov) and the XML files were subsequently parsed for calculation of average MAF for each variant. This method was adopted for training and test sets. Additionally, patient and control frequency data acquired in individual studies included in the test sets were used when available.

**Solvent Accessibility.** Solvent accessibility for all wild-type residues was predicted by running the Sable program (Adamczak, Porollo & Meller 2004) on the computational cluster. The query sequences for which predictions need to be generated are provided to the program in FASTA format. The output file consists of predicted secondary structure and relative solvent accessibility for each amino acid residue in the protein sequence. Solvent accessibility for a residue ranges from 0 (fully buried) to 9 (fully exposed).

**BLOSUM62, Grantham and Yu’s biochemical scores.** All three matrices were downloaded and Perl scripts were used to obtain difference between wild-type and
variant residue for each instance in the training and test sets. A Yu’s Biochemical Matrix score of \( \geq 0.05 \) indicates a neutral substitution, a positive BLOSUM62 score is associated with a conservative change and vice-versa, and a high Grantham difference (>100) depicts a significant change in amino acids.

**Changes in mass, volume and hydrophobicity.** These properties were extracted for each amino acid from sources in Table 8 and Perl scripts were written to calculate the shift in mass, volume and hydrophobicity due to amino acid changes in the training and test sets.

### 6.2.3 Support Vector Machine

We trained a Support Vector Machine (SVM) using the LIBSVM library (Chih-Chung, Chih-Jen 2001). SVM maps the training data into a higher dimensional space and then finds a linear plane separating the data with maximal margin. The input data was formatted to be compatible with LIBSVM and all features in the training and test sets were linearly scaled. Selection of an appropriate kernel function is a critical issue for SVM training. The RBF (radial basis function) kernel non-linearly maps instances into a higher dimensional space and then linearly separates the instances whereas a linear kernel separates the instances in the input space without mapping them into a higher dimensional space. Unlike the linear kernel, the RBF kernel can compute a model in cases where class labels and features are not linearly related.
The parameters g (radial basis kernel width) and C (penalty imposed for violating the margin) need to be chosen so that the radial or linear classifier can accurately predict unknown instances. The parameters were optimized on the training data using a grid search (Python script grid.py in LIBSVM) with 10-fold cross validation. The chosen values for C and g were 32.0 and 1.0 respectively. Radial and linear kernel functions were evaluated for classification of training and test data.

6.2.4 Performance measures

We calculated overall classification accuracy, area under curve (AUC), Matthews Correlation Coefficient (Matthews 1975), and coverage associated with each of the tools. Performance for the SVM-based classifiers was assessed with a 10-fold cross-validation study. Training set was divided into 10 subsets while keeping equal proportion of positive and negative instances in each subset. Ten different models were constructed and for each iteration, one subset was used as a test set and remaining nine were combined together to generate a training set. The final accuracy reported was mean of the accuracies of the 10 different models trained as described above. The probability score for a variant being pathogenic (belonging to the positive class) derived from 10-fold cross validation was used to generate the ROC curves. We also calculated the Matthews Correlation Coefficient, which is defined as:

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]
where TP, TN, FP, FN are the number of True Positives, True Negatives, False Positives, False Negatives, respectively. MCC is a balanced measure of accuracy with a value of 1, 0 and -1 representing perfect, random and worst possible classifications, respectively. For any specific tool, we define coverage as the proportion of variants in the training set for which the prediction tool can produce a classification.

6.3 Results

6.3.1 Feature selection

In order to determine the optimal combination of features, we performed 10-fold cross validation with each of the 11 features held out. Accuracy and area under curve (AUC) were calculated for each SVM model with a combination of 10 features and 1 feature left out (Table 9). In addition to performing leave-one-out cross-validation, we also used a feature selection tool (fselect) from the LIBSVM library. The tool generates an F-score as a measure of the importance of each of the features (Chen, Lin 2005). Excluding MAF, SIFT score and PolyPhen score had a negative impact on accuracy but leaving other features out either did not affect accuracy or yielded increased performance. The top 4 features with highest F-scores were MAF, SIFT, PolyPhen and PANTHER scores. Training the classifier with these 4 features resulted in 95.1% accuracy, which decreased when any one of these 4 features were excluded. Including additional features did not yield additional improvement and were therefore rejected.
### 6.3.2 Classification accuracy on training set

Variants in the training dataset were first analyzed with SIFT, PolyPhen and PANTHER to determine the concordance among these methods. Figure 18 represents the overlap in the classification produced by the three methods. 2481 out of 6578 (37.7%) neutral variants and 1242 out of 2355 (52.7%) disease variants were classified correctly by all three methods. Each method had a unique set of variants that were classified correctly only by that method.

Five different classifiers were used on the training set to predict the functional significance of the positive and negative instances. Prediction results were obtained with SIFT, PolyPhen, PANTHER, SVM with linear kernel, and SVM with radial kernel (Figure 19). We evaluated linear as well as radial kernels. The non-linear SVMs did not achieve significant improvement over the linear model. Hence, we used linear SVM because of its simplicity.

<table>
<thead>
<tr>
<th>Feature left out</th>
<th>Accuracy</th>
<th>AUC</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOSUM62</td>
<td>95.10</td>
<td>0.9942</td>
<td>0.0884</td>
</tr>
<tr>
<td>Solvent Accessibility</td>
<td>95.10</td>
<td>0.9942</td>
<td>0.0872</td>
</tr>
<tr>
<td>Grantham Score</td>
<td>95.04</td>
<td>0.9942</td>
<td>0.0603</td>
</tr>
<tr>
<td>Biochemical Score</td>
<td>95.03</td>
<td>0.9944</td>
<td>0.0384</td>
</tr>
<tr>
<td>Hydrophobicity</td>
<td>95.06</td>
<td>0.9943</td>
<td>0.0364</td>
</tr>
<tr>
<td>Volume</td>
<td>95.04</td>
<td>0.9943</td>
<td>0.0006</td>
</tr>
<tr>
<td>Mass</td>
<td>95.06</td>
<td>0.9942</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Table 9: Accuracy, AUC and F-score with one feature held out; F-score quantifies importance of each feature based on training data.
SVM-based classifiers produced the highest overall accuracy (96%), AUC (0.99) and MCC (0.91) compared to the other prediction tools (Table 10 & Figure 19).

PolyPhen, SIFT and PANTHER showed comparable accuracies (81%, 80% and 79% respectively) and AUC (0.85, 0.88 and 0.86 respectively) whereas PolyPhen showed highest MCC (0.63) as compared to SIFT (0.56) and PANTHER (0.55). The SVM-based classifiers had 100% coverage because the SVM made predictions on all variants.

PANTHER had the lowest coverage (70%) when compared to SIFT (99%) and PolyPhen (97%).

Figure 18: Classification agreement among SIFT, PolyPhen and PANTHER for variants in training set.
<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>AUC</th>
<th>MCC</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (radial kernel)</td>
<td>0.96</td>
<td>0.99</td>
<td>0.91</td>
<td>1.00</td>
</tr>
<tr>
<td>SIFT</td>
<td>0.80</td>
<td>0.88</td>
<td>0.56</td>
<td>0.99</td>
</tr>
</tbody>
</table>

**Table 10:** Performance comparison of SVM-based classifier against SIFT, PolyPhen and PANTHER.

![ROC curves and AUC](image)

**Figure 19:** Receiver Operating Characteristic (ROC) curves and Area Under Curve (AUC) for SVM, SIFT, PolyPhen and PANTHER.

### 6.4 Discussion

High-throughput methods of variant detection lead to the next challenge of predicting which of the variants are potentially associated with disease. So far, in vitro
functional tests for confirmation of pathogenic mutations are restricted to only a few genes. Given the unavailability of and complications associated with such functional tests, in silico methods have been developed to discriminate disease-associated missense variants from neutral variants. Bioinformatics approaches have been applied to the analysis of disease-associated mutations and there have been several studies focusing on the properties and localization of disease-causing mutations (Mooney 2005).

It has been shown that using a combinatorial approach comprising of multiple features and prediction results from diverse sources can improve prediction performance when compared to individual prediction tools (Nakken, Alseth & Rognes 2007, Ng, Henikoff 2003, Bao, Cui 2005, Frederic et al. 2009, Karchin, Kelly & Sali 2005, Ramensky, Bork & Sunyaev 2002a, Won et al. 2008). In this study, we assessed different properties hypothesized to indicate the severity of an amino acid substitution, and prediction results from SIFT, PolyPhen and PANTHER to develop a machine learning based approach for classification of missense variants.

Learning-based classification models require careful selection of a “gold standard” dataset of ‘deleterious’ and ‘neutral’ variants for training and/or validation. One of the challenges in the functional annotation of variants is that it is difficult to obtain a set of true neutral polymorphisms and most methods for collecting neutral variants end up retaining some false positives (Mooney 2005). The Online Mendelian Inheritance In Man (OMIM) database catalogs human genes and genetic phenotypes (Hamosh et al. 2005). The Single Nucleotide Polymorphism database (dbSNP) presents a broad collection of
simple genetic polymorphisms for a variety of organisms (Wheeler et al. 2008). The Swiss-Prot variant page contains manually annotated variants labeled as ‘polymorphism’, ‘disease’ or ‘unclassified’ (Yip et al. 2004). Datasets from different sources have different characteristics and when used to train machine-learning methods, they produce classifiers with varying error rates (Care et al. 2007). As is true with any of the learning-based prediction algorithms, performance of a classifier largely depends on the training data used. In order to exclude instances that have conflicting classifications in different databases, we selected ‘disease’ variants common to OMIM and SwissProt and ‘neutral’ variants labeled as such in SwissProt and with no OMIM or Clinical/LSDB associations in dbSNP.

Amino acid properties, protein structure, evolutionary conservation have been used as features for prediction of deleterious variants. The Grantham matrix quantifies the chemical distance between different amino acid residues (Grantham 1974). Kyte and Dolittle hydropathy scale can be used to determine hydropobicity along a protein (Kyte, Doolittle 1982). Yu’s “Biochemical Matrix” (http://cmgm.stanford.edu/biochem218/Projects%202001/Yu.pdf) has been developed based on 48 qualitative physicochemical properties and has previously been used for deleterious variant prediction (Frederic et al. 2009). Solvent accessibility has been reported to be another important feature in structure-based prediction of deleterious variants (Chasman, Adams 2001). Comparative evolutionary analysis can be used to determine if a variant is located in a conserved genomic region. The BLOSUM62 matrix
(Henikoff, Henikoff 1992) is an amino acid substitution matrix that uses multiple alignments of a protein set of related sequences.

Minor Allele Frequency (MAF) quantifies the proportion of a population that has the less common allele in the specific position on the chromosome. According to Hardy Weinberg principle, the frequency of specific alleles and that of specific diseases are directly proportional to each other when matings are random. Therefore, MAF can theoretically be associated with the functional significance of a variant allele. Montgomery et al. have analyzed sequence and population-based properties including minor allele frequency (MAF) of the variant for classification of regulatory SNVs (Montgomery et al. 2007). MAF has been used as a feature for classification of coding SNVs (Gowrisankar 2008).

SIFT, PANTHER, PolyPhen predictions and MAF contributed most to the performance of the predictor (Table 9). A possible reason for the remaining attributes not having a large impact could be that the prediction results from SIFT, PolyPhen and PANTHER already encapsulate these properties. The final classifier was trained using these attributes and the remaining features were excluded.

The SVM-based classifier produced highest cross-validation accuracy (96%), MCC (0.91) and area under ROC curve (0.99) for the training set (Table 10). PolyPhen showed better sensitivity (95%) than SIFT (82%) and PANTHER (75%). However, PANTHER gave better specificity (82%) compared to SIFT (80%) and PolyPhen (75%).
PANTHER provided the lowest coverage (69%), producing predictions for 6198 out of 8933 variants. Similar results have been reported earlier and low coverage of PANTHER can be attributed to stricter selection of sources for sequences and structures (Won et al. 2008).
Chapter 7. Case study: Prediction of SNHL and HCM/DCM variants using gene-specific customized training sets

7.1 Introduction

7.1.1 Hypertrophic Cardiomyopathy

Hypertrophic cardiomyopathy (HCM) is defined by the presence of unexplained asymmetric left ventricular hypertrophy with predominant involvement of the interventricular septum (Maron et al. 2003). With a prevalence of about 1 in 500 (Maron et al. 1995), it represents the most common inherited cardiovascular disorder, and it is also considered the primary cause of sudden cardiac death (SCD) in young athletes (Maron et al. 2000). In the past decade, 4450 pathogenic mutations in at least 16 different genes have been identified for HCM (Blair et al. 2001, Bos et al. 2006, Charron, Komajda 2006, Geier et al. 2003, Minamisawa et al. 2003, Van Driest et al. 2005). *MYH7, MYBPC3, TNNT2, TNNI3, TPM1, ACTC, MYL2*, and *MYL3* are proven pathogenic HCM disease genes. Mutations in PRKAG2 are now considered to cause a glycogen storage disorder characterized by ventricular hypertrophy, ventricular pre-excitation, and progressive conduction system disease (Murphy et al. 2005). The majority of reported mutations (86%) are single nucleotide substitutions causing missense or nonsense codons. Although a few mutation hotspots have been described in *MYH7* and *MYBPC3* (Charron, Komajda 2006), most affected families segregate a unique mutation. HCM is characterized by a high degree of phenotypic variability among and within
affected families, indicating the important role of environmental and modifying genetic factors.

Given the genetic etiology of HCM and the availability of previously studied and annotated variants, we selected variants associated with Hypertrophic Cardiomyopathy and Sensorineural Hearing Loss (see 2.1) as our validation dataset. The details of the dataset are provided in 7.2.

### 7.1.2 Gene Ontologies

The Gene Ontology project provides controlled vocabularies (ontology) of terms that describe gene products (http://www.geneontology.org). The ontology encompasses the biological processes, molecular functions and cellular components associated with gene products. The ontology is structured as a directed acyclic graph (DAG) where each term can be related to one or more terms within the same domain or in other domains. An earlier study has proposed evaluating the efficacy of a decision rule learned with a training set of variants from one protein when applied to other proteins (Karchin et al. 2007). We aim to assess the impact of deriving decision rules from customized training sets of proteins sharing Gene Ontology terms with the protein of interest.

### 7.2 Materials and Methods

#### 7.2.1 Test dataset

The HCM/DCM dataset consisted of 57 missense variants from 9 genes (GLA, LAMP2, MYBPC3, MYH7, MYL2, PRKAG2, TNNI3, TNNT2 and TPM1) that had been
classified as benign (7) or pathogenic (50) (http://cardiogenomics.med.harvard.edu). The hearing loss dataset consisted of manually annotated missense variants from 3 genes (CDH23, MYO7A and OTOF) implicated in sensorineural hearing loss. We analyzed 8 variants without any evidence of disease association and 84 variants that were known to be associated with hearing loss.

7.2.2 Gene Ontology Associations

Gene Ontology (GO) associations for human genes were downloaded from the GO FTP site (ftp://ftp.geneontology.org/pub/go/gene-associations/). The tab delimited file contains Swiss-Prot accession IDs for proteins, the associated GO terms and citation and evidence data. As of Nov 2009, the database contains 18587 gene products and 165764 annotations. The file was then processed to obtain a mapping between unique genes and all corresponding GO terms and another for unique GO terms and all genes associated with each term. It was verified that all genes in the training and test sets were represented in the gene-GO associations.

7.2.3 Feature extraction

As described in 6.2.2, SIFT, PolyPhen, and PANTHER predictions were extracted for all variants in the test set. All HCM/DCM variants had frequency data available either from literature or from mutation-discovery projects, as cataloged in the Harvard Medical School CardioGenomics database (http://cardiogenomics.med.harvard.edu). MAF was extracted from dbSNP for SNHL variants with available frequency data.
7.3 Results

7.3.1 Classification performance on validation set

Variants in the test dataset were first analyzed with SIFT, PolyPhen and PANTHER to determine the concordance among these methods. Figure 20 represents the overlap in the classification produced by the three methods. 1 out of 15 (6%) neutral variants and 53 out of 134 (40%) disease variants were classified correctly by all three methods.

Figure 20: Classification agreement among SIFT, PolyPhen and PANTHER for variants in validation dataset.
The variants were further analyzed with the SVM-based classifier. Receiver Operating Characteristics (ROC) analysis was performed on all 149 test variants (15 neutral and 134 deleterious) to compare prediction results from SIFT, PANTHER, PolyPhen and the SVM-based classifier.

The area under the ROC curve (AUC) gives an estimate of the likelihood that the classifier ranks a randomly selected positive example higher than a randomly selected negative example. An ideal test would produce an AUC of 1.0 and a random classifier would lead to an expected AUC of 0.5. The AUC values are 0.76 for SIFT, 0.83 for PANTHER, 0.70 for PolyPhen and 0.92 for Support Vector Machine (Figure 21).

Figure 21: ROC curves and AUC for SVM, SIFT, PolyPhen and PANTHER for variants in validation dataset. Decision thresholds for each classifier were varied over a large range and sensitivity and specificity were calculated at each threshold.
Based on sensitivity, specificity, MCC and coverage, the SVM-based classifier produced the best classification results for training and test sets (Table 11). It should be noted that PANTHER and PolyPhen predictions could not be obtained for all variants and an “unknown” status was assigned to variants that could not be classified. Therefore, PolyPhen and PANTHER provided lower coverage than SIFT or SVM-based classifier. The choice of kernel did not affect performance on test set (data not shown).

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>MCC</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.88 (44/50)</td>
<td>1.00 (7/7)</td>
<td>0.69</td>
<td>100 (57/57)</td>
</tr>
<tr>
<td>SIFT</td>
<td>0.72 (36/50)</td>
<td>0.57 (4/7)</td>
<td>0.21</td>
<td>100 (57/57)</td>
</tr>
<tr>
<td>PolyPhen</td>
<td>0.65 (28/43)</td>
<td>0.80 (4/5)</td>
<td>0.28</td>
<td>84 (48/57)</td>
</tr>
<tr>
<td>PANTHER</td>
<td>0.51 (20/39)</td>
<td>0.67 (2/3)</td>
<td>0.09</td>
<td>74 (42/57)</td>
</tr>
</tbody>
</table>

Table 11: Classification performance with SIFT, PANTHER, PolyPhen, and SVM.

7.3.2 Training on genes sharing GO terms

The idea for evaluating a customized classifier where genes in the training set share attributes with the gene of interest has been proposed by Karchin et al (Karchin et al. 2007). We aimed to test the hypothesis that using a method where the training set is comprised of genes that share GO terms with the gene being tested would result in comparable or better prediction results than using a generic classifier for all genes. Since GO terms can be assigned to three major categories namely Molecular Function (MF),
Cellular Component (CC) and Biological process (BP), we wanted to test if training the classifier on proteins sharing terms from different categories produced varying results. We performed cross-validation on the training set with and without customized training sets and observed that training on genes sharing GO terms improved the AUC (Figure 22). Classifier trained with genes sharing MF (AUC 0.996), any GO category (AUC 0.993), or BP (AUC 0.991) terms performed better than classifier trained with CC terms (AUC 0.90) (Figure 22). For subsequent analyses, we chose MF as the GO category for developing customized training sets.

Figure 22: Performance comparison of SVM trained with genes belonging to different GO categories. Variants in training set were analyzed by selecting genes sharing GO terms with gene of interest and using these for training the classifier.
A Perl script was written to generate separate training sets for each gene in the test set for training the SVM classifier. All hearing loss and HCM/DCM variants were tested using this approach with an SVM (linear kernel, C=32). Table 12 lists variant predictions for individual genes in the test set. Predictions for all genes in the hearing loss dataset remained the same and performance measures did not change from those listed in Table 11. However, 2 additional deleterious variants were detected in the HCM/DCM dataset (Table 12) using customized training data. The last column in Table 12 lists the number of variants that were included in training set for each gene. It is interesting to note that using a smaller training set of variants from genes sharing GO terms resulted in similar, or better, performance when compared to a classifier trained on a much larger generic set (8933 variants). ROC curves were generated for SVM trained with MF terms and compared with SIFT, PolyPhen and PANTHER (Figure 23). The SVM produced the largest AUC (0.996) for the validation set.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Total True Positives</th>
<th>True Positives detected</th>
<th>Total True Negatives</th>
<th>True Negatives detected</th>
<th>No. of variants in training set</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAMP2</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>3507</td>
</tr>
<tr>
<td>MYH7</td>
<td>27</td>
<td>23 (+1)</td>
<td>-</td>
<td>-</td>
<td>4164</td>
</tr>
<tr>
<td>PRKAG2</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>3102</td>
</tr>
<tr>
<td>TNNT2</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>515</td>
</tr>
<tr>
<td>CDH23</td>
<td>34</td>
<td>32</td>
<td>7</td>
<td>7</td>
<td>5662</td>
</tr>
<tr>
<td>OTOF</td>
<td>7</td>
<td>7</td>
<td>-</td>
<td>-</td>
<td>4590</td>
</tr>
</tbody>
</table>

Table 12: Prediction for individual genes in the test set. Entries in bold indicate additional variants that were detected by training the classifier on genes sharing GO terms.
7.3.3 Misclassified variants

The classifier produced incorrect predictions for 6 pathogenic HCM mutations and 4 pathogenic SNHL mutations (Table 13). *MYBPC3* E258K had a score of 0.47, thus making it a borderline benign prediction. It was found in 0.9% of controls and had borderline deleterious SIFT prediction and was called “possibly damaging” by PolyPhen. R502W had deleterious SIFT prediction but had a MAF of 1.7% resulting in a benign prediction (score 0.07). *MYH7* R663H and D906G had benign predictions from all three tools and while the former showed a low likelihood of being pathogenic (score 0.17), the latter had a borderline score of 0.47 due to a lower MAF (0.2%). L908V (MAF 0.7) was assigned a 0.47 probability of being pathogenic even though all three tools classified it as damaging. *MYL2* R58Q and all SNHL variants were classified as benign by all tools that could provide a prediction for these variants.
Table 13: Misclassified pathogenic variants.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Gene</th>
<th>AA</th>
<th>SVM</th>
<th>MAF (%)</th>
<th>SIFT</th>
<th>PANTHER</th>
<th>PolyPhen</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCM</td>
<td>MYBPC3</td>
<td>E258K</td>
<td>0.47</td>
<td>0.9</td>
<td>0.05</td>
<td>-</td>
<td>PD 1.58</td>
</tr>
<tr>
<td>HCM</td>
<td>MYBPC3</td>
<td>R502W</td>
<td>0.07</td>
<td>1.7</td>
<td>0.00</td>
<td>-</td>
<td>Unknown</td>
</tr>
<tr>
<td>HCM</td>
<td>MYH7</td>
<td>R663H</td>
<td>0.17</td>
<td>1.1</td>
<td>0.09</td>
<td>-1.90</td>
<td>B 0.11</td>
</tr>
<tr>
<td>HCM</td>
<td>MYH7</td>
<td>D906G</td>
<td>0.47</td>
<td>0.2</td>
<td>0.41</td>
<td>-2.39</td>
<td>B 0.55</td>
</tr>
<tr>
<td>HCM</td>
<td>MYH7</td>
<td>L908V</td>
<td>0.47</td>
<td>0.7</td>
<td>0.00</td>
<td>-3.36</td>
<td>PD 1.72</td>
</tr>
<tr>
<td>HCM</td>
<td>MYL2</td>
<td>R58Q</td>
<td>0.12</td>
<td>0.2</td>
<td>1.00</td>
<td>-2.13</td>
<td>B 0.14</td>
</tr>
<tr>
<td>SNHL</td>
<td>CDH23</td>
<td>E541K</td>
<td>0.13</td>
<td>-</td>
<td>1.00</td>
<td>-2.40</td>
<td>B 1.26</td>
</tr>
<tr>
<td>SNHL</td>
<td>CDH23</td>
<td>R1746Q</td>
<td>0.24</td>
<td>-</td>
<td>0.81</td>
<td>-</td>
<td>B 1.35</td>
</tr>
<tr>
<td>SNHL</td>
<td>MYO7A</td>
<td>A1628S</td>
<td>0.26</td>
<td>-</td>
<td>0.74</td>
<td>-</td>
<td>B 0.15</td>
</tr>
<tr>
<td>SNHL</td>
<td>MYO7A</td>
<td>G955S</td>
<td>0.30</td>
<td>-</td>
<td>0.69</td>
<td>-</td>
<td>B 0.03</td>
</tr>
</tbody>
</table>

In order to assess if the classifier was heavily dependent on MAF and produced misclassifications for any benign variant with a low MAF but was classified as benign correctly by the three tools, we selected a test dataset of benign HCM/DCM variants and SNPs from dbSNP with a low MAF.

We observed that all 11 variants were classified correctly as benign with low probabilities of being pathogenic, which showed that a low MAF did not necessarily skew the classifier to produce pathogenic predictions irrespective of the remaining prediction results.
Table 14: Variants with low MAF and benign predictions selected to test the dependence on MAF of the classifier.

<table>
<thead>
<tr>
<th>Class</th>
<th>Gene</th>
<th>AA/rsID</th>
<th>MAF</th>
<th>SIFT</th>
<th>PANTHER</th>
<th>PolyPhen</th>
<th>SVM pathogenic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ben</td>
<td>LAMP2</td>
<td>V391I</td>
<td>0.007</td>
<td>1.00</td>
<td>-0.95</td>
<td>-</td>
<td>Ben 0.02</td>
</tr>
<tr>
<td>Ben</td>
<td>MYBPC3</td>
<td>A833V</td>
<td>0.004</td>
<td>0.64</td>
<td>-2.37</td>
<td>-</td>
<td>Ben 0.23</td>
</tr>
<tr>
<td>Ben</td>
<td>LAMP2</td>
<td>A100V</td>
<td>0.001</td>
<td>0.82</td>
<td>-1.35</td>
<td>0.68</td>
<td>Ben 0.24</td>
</tr>
<tr>
<td>Ben</td>
<td>PRKAG2</td>
<td>G100S</td>
<td>0.009</td>
<td>0.06</td>
<td>-</td>
<td>0.39</td>
<td>Ben 0.29</td>
</tr>
<tr>
<td>Ben</td>
<td>MYBPC3</td>
<td>V189I</td>
<td>0.006</td>
<td>0.42</td>
<td>-</td>
<td>0.12</td>
<td>Ben 0.29</td>
</tr>
<tr>
<td>Ben</td>
<td>MEGF11</td>
<td>rs333550</td>
<td>0.005</td>
<td>0.84</td>
<td>-3.10</td>
<td>0.90</td>
<td>Ben 0.05</td>
</tr>
<tr>
<td>Ben</td>
<td>SULT2B1</td>
<td>rs2302947</td>
<td>0.005</td>
<td>0.51</td>
<td>-0.81</td>
<td>0.33</td>
<td>Ben 0.16</td>
</tr>
<tr>
<td>Ben</td>
<td>GSTA4</td>
<td>rs4147617</td>
<td>0.007</td>
<td>0.74</td>
<td>-1.17</td>
<td>0.39</td>
<td>Ben 0.05</td>
</tr>
<tr>
<td>Ben</td>
<td>ZSCAN20</td>
<td>rs4403594</td>
<td>0.002</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>Ben 0.05</td>
</tr>
<tr>
<td>Ben</td>
<td>CUL7</td>
<td>rs9381231</td>
<td>0.002</td>
<td>1.00</td>
<td>-</td>
<td>0.19</td>
<td>Ben 0.07</td>
</tr>
<tr>
<td>Ben</td>
<td>TRPV3</td>
<td>rs322965</td>
<td>0.002</td>
<td>0.77</td>
<td>-</td>
<td>0.78</td>
<td>Ben 0.13</td>
</tr>
</tbody>
</table>

7.4 Discussion

For the validation sets, SVM produced the best results in terms of overall accuracy (93.3%), AUC (0.92), sensitivity (93%), specificity (100%) and MCC (0.75) for the HCM/DCM and hearing loss variants. PANTHER provided lowest coverage (63%) compared to SIFT (100%) and PolyPhen (94%). SIFT and PANTHER provided slightly better overall accuracy and sensitivity than PolyPhen (Table 11).

A study by Karchin et al proposes assessing the impact of using classifiers trained with proteins that have properties similar to the protein being tested instead of a generic training set for analyzing all variants (Karchin et al. 2007). We generated customized training sets for every protein in the validation set by including variants from only the
proteins that share GO terms with the protein being analyzed. This reduced the training data size for individual proteins and led to the identification of an additional pathogenic variant in \textit{MYH7} and \textit{TPM1} (Table 12). We observed that classifier trained with proteins sharing “Molecular Function” and “Biological Process” Gene Ontology terms produced larger AUC’s than the classifier trained with “Cellular Component” terms. The cellular component ontology describes locations of gene products at the levels of sub-cellular structures and macromolecular complexes whereas biological process and molecular function are indicative of the series of events a gene is involved in and the abilities it possesses (http://www.geneontology.org). Disruptions in biological processes, which are collections of molecular functions, are often associated with phenotypes resulting from mutations. Therefore, biological process and molecular function have a stronger association with the functions a gene is involved in than the cellular component it is related to. This could explain the observation that training classifiers with genes sharing similar functions and processes produces better results than cellular component. However, these findings need to be validated on a larger scale with more variants from a large number of proteins.

The classifier showed a high degree of sensitivity to MAF associated with each variant. Studies recorded in dbSNP and SwissProt often use frequency data to label variants (few have functional data) and if a variant is found in a high number of controls, it is likely that it will be labeled neutral. Such cases are highly represented in the training set and thus affect classification on test data significantly. Three pathogenic HCM variants (\textit{MYH7} L908V, \textit{MYBPC3} E258K and \textit{MYBPC3} R502W) were assigned low
probabilities of being pathogenic (0.47, 0.47 and 0.07 respectively) even though prediction tools classified them as damaging (Table 13). MAF associated with these variants are 0.7%, 0.9% and 1.7% and therefore the classifier determined a low probability of these variants being deleterious. To avoid skewing the classifier with a biased MAF, future work needs to focus on obtaining population-specific frequencies for all training instances based on the ethnicity of the subjects for which variants need to be analyzed.

In summary, we have used a machine learning framework to evaluate the contribution of various properties related to amino acid substitution in determining whether a residue change is damaging or not. We have applied a combinatorial approach to analyze HCM/DCM and hearing loss variants and obtained more accurate prediction results than some of the representative tools for missense variant classification. We also assessed the impact of developing customized classifiers trained on proteins sharing GO terms with the protein being tested and observed that a smaller number of related proteins can be used for training the classifier to provide better or same prediction accuracy as compared to employing a large generic training set for all proteins.
Chapter 8. Conclusions and Future Work

In summary, we have developed a novel computational algorithm, sPROFILER, for improved detection of sequence variants with resequencing arrays. The algorithm utilizes properties such as differential dependence on G- and C-content of probe-target hybridization. We validated the algorithm on a set of arrays from two different laboratories and demonstrated that it improved call coverage while maintaining high overall accuracy. We compared our results against dideoxy sequencing to determine the accuracy of base calls. By applying sPROFILER to reduce Affymetrix GSEQ no-calls, >80% of the no-calls could be resolved while maintaining overall accuracy at ≥99.8%. This improvement led to a lower overhead associated with each sample by reducing number of exons needing follow-up dideoxy sequencing assays. We implemented a bioinformatics pipeline that also incorporated sPROFILER to support clinical genetic testing of hearing loss patients at the Cincinnati Children’s Hospital Medical Center.

The inability of resequencing arrays to detect small insertions and deletions presents a limitation to their clinical application. The regional drop in signal intensity associated with small indels can sometimes be very small to allow easy detection. Algorithmic and technical improvements need to be developed for detection of indels.

We developed a computational framework to assess the deleterious impact of missense variants. We used a combination of SIFT, PolyPhen, PANTHER and Minor
Allele Frequency in a machine learning framework to determine the likelihood of a variant being pathogenic. We validated our approach on Hypertrophic Cardiomyopathy and Hearing Loss missense variants and obtained higher accuracies when compared to some of the existing representative tools for variant classification. We also evaluated the contribution of various properties related to amino acid substitution in determining whether a residue change is damaging or not. Additionally, we developed customized classifiers trained on proteins sharing GO terms with the protein being tested and observed a smaller training set could be used to provide better or same prediction accuracy as compared to utilizing a large generic training set for all proteins.

Preliminary analysis showed that training the classifier on proteins sharing “Molecular Function” GO terms produced better classification results than training on “Biological Process” and “Cellular Component”. However, further work with a larger dataset is needed to confirm this hypothesis. Our results also indicate that the classification is sensitive to MAF of the variant so future work needs to focus on obtaining population-specific MAF to avoid introducing bias in the classifier.
References


associated with hypertrophic cardiomyopathy", *Biochemical and biophysical research communications*, vol. 304, no. 1, pp. 1-4.


Mooney, S. 2005, "Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis", *Briefings in bioinformatics*, vol. 6, no. 1, pp. 44-56.


Ng, P.C. & Henikoff, S. 2002, "Accounting for human polymorphisms predicted to affect protein function", *Genome research*, vol. 12, no. 3, pp. 436-446.

Ng, P.C. & Henikoff, S. 2001, "Predicting deleterious amino acid substitutions", *Genome research*, vol. 11, no. 5, pp. 863-874.


Sunyaev, S.R., Eisenhaber, F., Rodchenkov, I.V., Eisenhaber, B., Tumanyan, V.G. & Kuznetsov, E.N. 1999, "PSIC: profile extraction from sequence alignments with
position-specific counts of independent observations", *Protein engineering*, vol. 12, no. 5, pp. 387-394.


1B syndrome are allelic defects of the myosin-VIIA gene", *Nature genetics*, vol. 16, no. 2, pp. 191-193.


