January 2011

Clinical Interpretation of Novel Copy Number Variations

Clifton Carey
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science

Department of Computer Science and Engineering

Thesis Examination Committee:
S. Joshua Swamidass, Chair
Michael Brent
Jeremy Buhler

CLINICAL INTERPRETATION OF NOVEL COPY NUMBER VARIATIONS

by

Clifton M. Carey, Jr.

A thesis presented to the School of Engineering
of Washington University in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

May 2011
Saint Louis, Missouri

ABSTRACT OF THE THESIS

Clinical Interpretation of Novel Copy Number Variations

by

Clifton M. Carey, Jr.

Master of Science in Computer Science

Washington University in St. Louis, 2011

Research Advisor: Dr. S. Joshua Swamidass

Copy Number Variations (CNVs) are a significant source of human genetic diversity and are believed to be responsible for a wide variety of phenotypic variation. Recent advances in microarray-based genomic hybridization techniques have facilitated CNV analysis as a viable diagnostic technique in the clinic, and several public databases of well-characterized CNVs are being compiled, but a standard for interpreting uncharacterized CNVs has yet to emerge.

This thesis examines the clinical interpretation of uncharacterized CNVs as a multiple instance binary classification problem. We analyze the current state of clinical techniques, then present and test a novel, statistical approach to the problem.

Acknowledgments

I would like to thank Dr. S. Joshua Swamidass, at the Washington University School of Medicine's Pathology department, for his wisdom and patience guiding me through the entire research process.

Thanks also to Dr. Colin Pritchard, at the University of Washington's Department of Pathology, for providing the CNV scoring rubric used extensively in this thesis.

A special thanks to Brad Calhoun, Michael Browning, and to Drs. Jeremy Buhler and Michael Brent in the Computer Science department, for reviewing this thesis and providing essential feedback.

Clifton M. Carey, Jr.

Washington University in Saint Louis
May 2011

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Copy Number Variations (CNVs) are a form of structural variation in a genome, where large segments of DNA have different numbers of copies between individual organisms of the same species. Segments of the genome with increased copy number are called duplications; decreases in copy number are referred to as deletions. An organism's CNVs are either inherited or the result of a *de novo* mutation.

CNVs are common in humans[1] and account for a significant amount of genetic variability[2]. Some hypothesize, therefore, that CNVs might explain phenotypic variation among humans, specifically as they relate to an individual's susceptibility to disease[3, 4]. Consequently, patient CNV data is becoming for frequently used in the clinic as a tool for diagnosis. To this end, several databases of characterized CNVs are maintained, including the Database for Genomic Variants[5], the UCSC Genome Browser[6], CNVVdb[7], CHOPPY[8], DECIPHER[9], and the NCBI dbVar database[10]. However, the majority of CNVs encountered in the clinic are still uncharacterized, that is, with no proven association to disease.

1.2 Case Study

This case, taken from a 2009 study by Tsuchiya et al. [11], represents a common clinical scenario. A patient (case 2 in Table 1.1) with a family history of learning disabilities arrives

at the clinic showing symptoms of developmental delay. After running several other tests, the physician orders a screen of the patient's CNVs. A deletion of 810 kilobases is identified on chromosome 8, then confirmed with a secondary test. This CNV is entirely contained in CSMD1 gene, thought to be a tumor suppressor, but involves multiple exons. Searching the laboratory's internal database of known CNVs yields no results, because no CNV of that size and location has been reported. The laboratory must now attempt to interpret this uncharacterized CNV, determining whether the mutation is a cause for concern.

1.3 Interpretation Variability

While the International System for Chromosome Nomenclature's (ISCN) 2009 report provides a system for reporting CNVs [12], no standard for interpeting uncharacterized CNVs exists. Several factors are commonly thought to be important (e.g., a deletion is considered to be more concerning than a duplication), but it is unclear to what extent each factor is predictive. Furthermore, many factors known to be predictive are often unavailable in initial screenings: CNV inheritance has been shown to strongly correlate with phenotypic normalcy, but running full CNV screens on a patient's parents is expensive, time-consuming, and in many cases infeasible. This has led laboratories to consider various types of CNV metrics, which may be combined in several ways to arrive at an interpretation.

Most concerningly, one study showed no concordance between interpretations of uncharacterized CNVs reported by different labs.[11] This study presented 11 different laboratories with 13 patient cases, each with a clinical indication and a CNV. The labs were asked to classify the CNVs using their standard practices. The results from each laboratory were then aggregated to assess the variability of interpretation. The CNV from our case study was interpreted as normal by three labs, abnormal by another three, and of "uncertain clinical significance" by the remaining five. See Table 1.1 for the full listing of results, which demonstrate a general lack of agreement between laboratories for uncharacterized CNV interpretation. These results raise questions about the reproducibility and validity of the tested techniques, and by extension, the current state of CNV characterization methods.

Table 1.1: Variablility in CNV interpretation. The varied responses from independent CNV classification efforts [11] demonstrate a lack of concordance between results.

Case	CNV	Normal	Likely Benign	Uncertain	Abnormal
1	0.19/0.87[a] Mb dup., 3p26.3	6	0	4	1
2	0.81/0.89 Mb del., 8p23.2	3	0	5	3
3	0.17/0.39 Mb del., 3p26.2	5	0	6	0
4	0.46/1.0 Mb del., 6p11.2	6	1	3	1
5	0.19/2.9 Mb dup., 6q26	3	0	7	1
6	0.26/0.92 Mb dup., 6q27	5	0	5	1
7	0.63/0.79 Mb dup., 3p12.3	8	1	2	0
8	0.04/0.33 Mb del., 22q11.21	6	0	5	0
9	0.07/0.09 Mb del., 16p13.3	4	1	6	0
10	0.28/0.36 Mb dup., Xp11.23[b]	6	1	4	0
11	0.42/0.58 Mb del., 7q11.23	10	0	1	0
12	0.54/0.55 Mb dup., Xp22.33[b]	8	2	1	0
13	1.0/1.1 Mb dup., 5q21	4	2	5	0

Cases 1 through 6 are BAC array cases; 7 through 13 are oligonucleotide array cases.

[a]Minimum/maximum size,[b]Female.

1.4 Goals

The goals of this research are two-fold: we seek to perform an evaluation of the CNV characterization attempts made by others, then to apply more rigorous statistical and computational methods in an attempt to improve on the state of the art techniques.

Chapter 2

Data

2.1 Copy Number Variation Identification

The CNVs used in our study were detected by hybridizing genomic DNA to a chromosomal microarray (CMA)[13]. Each probe on the array maps to a specified location in the genome. A set of probes that map to a contiguous segment of the genome are simultaneously perturbed in the presence of a CNV. CNVs are identified by finding contiguous probes with elevated or reduced intensities (see Figure 2.1). The sign of the magnitude of the copy number change differentiates between deletions and duplications. CNV boundaries cannot be measured exactly, but range within the sampling interval of the CMA.

The accuracy of detecting each CNV depends on the probe density, CNV size, and the magnitude of the copy number change. Arrays with high probe density have more potential to demonstrate evidence of CNVs, and achieve more accurate bounds on start and end points. Similarly, larger CNVs will cross more probes, which also yields higher confidence. CNVs with a large magnitude of copy number change are more accurately detected.

Several algorithms are commonly used for identifying CNVs in CMA data, including PennCNV[14], QuantiSNP[15], HMMSeg[16], and cnvPartition (Illumina). Each software package formulates the task as an optimization problem, using a Hidden Markov Model to find optimal start and end locations for likely CNVs. A recent study by Tsuang et al. [17] shows that each of these algorithms produce many false positives and false negatives, and stresses the importance of the establishment of a standard for CNV validation.

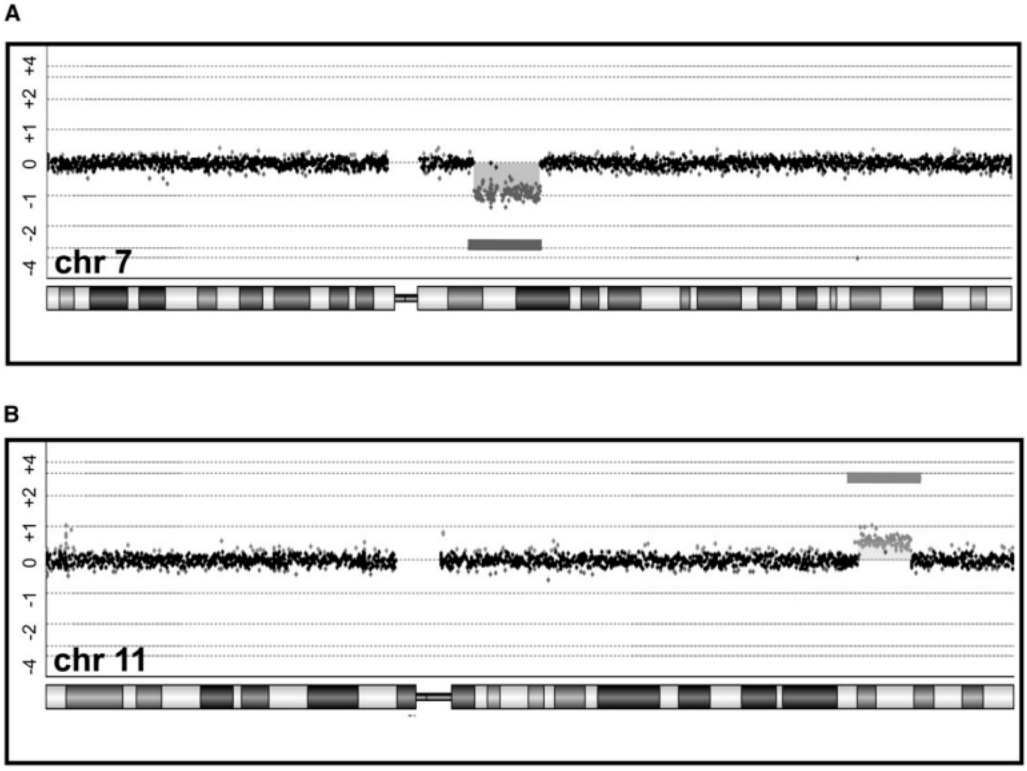


Figure 2.1: A visualization of two sections of CMA results, adapted from Miller et al. [13]. Each point corresponds to a probe on the CMA, and the thick gray bars indicate an identified CNV. The vertical axis indicates the deviation of probe readings from a normal genome, and the horizontal axis indicates position in the chromosome. Section A shows an identified deletion on chromosome 7; section B shows an identified duplication on chromosome 11.

2.2 CNV Data

This study's CNV data is taken from the set released by Itsara et al. [18] as well as the International Standards for Cytogenomic Arrays (ISCA) consortium's data in the restricted-access database of Genotypes and Phenotypes (dbGaP)[19]. The Itsara set contains only benign CNVs, while the ISCA set contains CNV data from affected patients. These two sets, combined, account for 15,389 CNVs from 3167 patients. (See Table 2.1 for details). This combination requires several assumptions about the makeup of each consitituent data set. These assumptions are addressed in the discussion section.

Table 2.1: CNV statistics. Summary statistics for the two sources of CNV data and the combined set.

	Patients	Cases	Mean CNVs/Patient	Mean Size	Std Dev Size
Itsara	2154	0	6.43	740.12 kb	1242.85 kb
ISCA	1013	1013	1.53	92.01 kb	140.22 kb
Combined	3167	1013	4.86	157.17 kb	459.29 kb

2.3 Supplemental Data

In addition to the raw CNV data, we integrate supplemental data from outside sources. These sources are: the National Center for Biotechnology Information (NCBI) Entrez listing of human genes [20], the Biomart webservice [21], the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) protein interaction database [22], and the Gene Set Enrichment Analysis (GSEA) database [23].

Chapter 3

Methods

In this section, we present a CNV scoring system currently in use in clinical settings, then describe and justify our general approach.

3.1 Expert CNV Scoring

One current system for CNV interpretation, currently in use at the University of Washington[24], is an expert-constructed, worksheet-based rubric that places a CNV into one of four classes. The worksheet tabulates points for each of several CNV features, as shown in Table 3.1. Points are added for features that indicate a higher likelihood of pathogenicity. Conversely, points are subtracted for features that suggest the CNV is more likely benign.

We implement a facsimile of this scoring function in software to test its efficacy in uncharacterized CNV interpretation, and to serve as a baseline against which other approaches can be compared. Our version mimics the scoring of the original for the first four features only. We disregard features that are relevant only to characterized CNVs, because our study focuses on the problem of interpreting uncharacterized CNVs. We also disregard subjectively scored features, which are not meaningully implementable in software.

Table 3.1: Worksheet point allocation. Each CNV feature has the potential to contribute points from the listed ranges, but the lack of secondary data (features 6,7,9) for uncharacterized CNVs limits the features under consideration.

	Range	Uncharacterized
Deletion/Duplication	0 to 1	0 to 1
Size	0 to 3	0 to 3
Number of Genes	-1 to 4	-1 to 4
Genes Interruption	0 to 1	0 to 1
Gene Roles*	0 to 2	0 to 2
Known Del/Dup Syndrome	0 to 1	0
Signature/Decipher Cases	-1 to 2	0
Segmental Duplications*	-1 to 0	-1 to 0
Benign CNV Overlap	-3 to 1	0

*These features, while available for uncharacterized CNVs, are scored too subjectively to be used in a software implementation.

Table 3.2: Worksheet score ranges. A CNV's score is the sum of the assigned points for each feature, which provides a basis for labelling.

Score	CNV Label
≤ 0	Do Not Report (Normal)
1-4	No Known Pathogenic Imbalances
5-8	Unclear Clinical Significance
> 8	Likely Pathogenic

3.2 Reference Range Approach

We hypothesize that significant improvements can be made to the expert scoring system discussed in the previous section by using statistical methods to construct a reference range (or reference interval)[25] for benign CNVs. Our basic approach is to,

1. Model the probability density of each CNV descriptor amongst the set of benign CNVs.

2. Score each CNV by the sum of the negative log likelihood according these densities.

$$score\,(CNV) = -\sum_{feature} \log P\,(feature\,(CNV)\,|benign)\,. \tag{3.1}$$

3. Assign the score of the highest scoring CNV to each patient.

$$score\,(patient) = \max_{CNV \in patient} [score\,(CNV)]\,, \tag{3.2}$$

4. Declare a patient's CNV abnormal if their score is above the 95th percentile (that is, outside the reference interval) of scores for normal patients.

The reference range is a well-established strategy in medical practice[25, 26, 27, 28, 29]. A reference range is simply a confidence interval (generally 95%) for the distribution of test results for a control group, which provides a common-sense framework for interpreting the results of clinical tests. The interval is used to provide context for the results of the same test when performed on a patient. A measurement within the reference range is likely to be normal (that is, unlikely to be diagnostically relevant), and a measurement outside the range may merit further investigation.

Our approach measures the feature densities of benign CNVs, whereas a more complete Bayesian treatment is to score CNVs by their likelihood of being pathogenic given their features,

$$score\,(CNV) = P\,(pathogenic|F) = \frac{P(F|pathogenic)P(pathogenic)}{P(F)}. \tag{3.3}$$

However, direct estimation of the distribution of pathogenic CNVs is difficult, as we do not have a clear strategy for labeling CNV-level data as pathogenic. Instead, each patient's set of CNVs is associated with the patient's known clinical phenotype. As such, while all the CNVs of a normal patient can be safely labeled benign, labels for the CNVs of pathogenic patients are unknown.

3.3 Likelihood Scoring

In an attempt to improve the accuracy of CNV interpretation, we consider several features in this probabilistic framework. For each feature, we define the feature density function $P\,(feature\,(CNV)\,|benign)$ used in Equation 3.1. We begin by examining the four CNV features used in the expert scoring rubric, then move on to novel CNV features.

3.3.1 Duplication vs Deletion

The simplest CNV feature notes the sign of the magnitude of the copy number change. We model this binary feature with a Bernoulli distribution, counting the number of each type in the set of benign CNVs, B. This gives rise to the probabilility density function

$$P(dup|benign) = \frac{|duplications \cap B|}{|B|} = 1 - P(del|benign). \tag{3.4}$$

3.3.2 Size

In a similar fashion, we consider the probability density for the size of a benign CNV. Visualizing the histogram of benign CNV sizes suggests an exponential distribution (Figure 3.1), which, unlike a Gaussian distribution, ensures zero probability of any negative value.

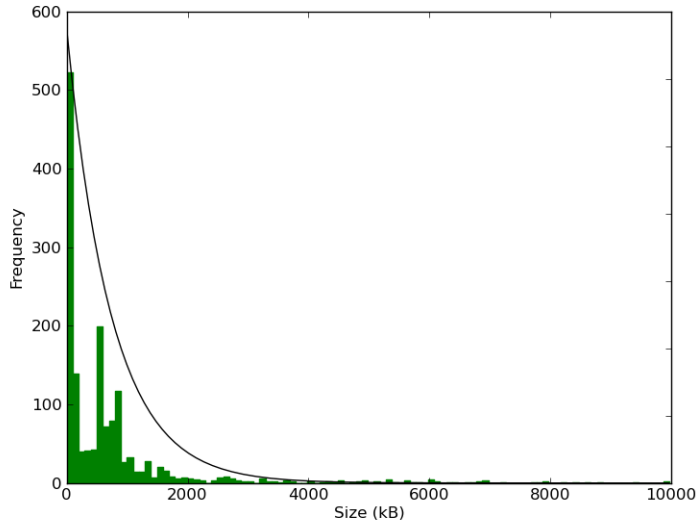


Figure 3.1: Frequency of benign CNV sizes. $\lambda \approx \frac{1}{740}$.

This produces the density function

$$P(size = x|benign) = \lambda e^{-\lambda x}, \tag{3.5}$$

where the rate parameter λ is determined using fitting of first moment,

$$\lambda = E\left(sizes\right)^{-1} = \frac{|B|}{\sum_{CNV \in B}\left[size\left(CNV\right)\right]}. \tag{3.6}$$

Better estimates might be obtained using a maximum likelihood estimator for λ, but we found this fit good enough for our purposes.

3.3.3 Split and Spanned Genes

The last two features from our implementation of the expert worksheet consider the number of genes a CNV overlaps, broken into counts for spanned (complete overlap) and split (partial overlap) genes. As with the previous feature, we visualize the distribution of spanned genes among the benign CNV set (Figure 3.2), and observe that the exponential distribution fits resonably well. The split genes feature can clearly only take on one of three values, so we use an empirical distribution, counting the frequency of each number of gene splits over the set of benign CNVs.

As such, the density function from Equation 3.5 may be reused for the spanned genes feature, with a different rate parameter λ based on the expected value of the number of spanned genes in the benign set. The density function for the split genes feature resembles Equation 3.4, taking the form

$$P\left(splits = x|benign\right) = \frac{|\{CNV \in B|splits\left(CNV\right) = x\}|}{|B|}. \tag{3.7}$$

3.3.4 Gene Density Scoring

We now turn to new CNV features, to test whether additional types of information are useful for discriminating benign from pathogenic CNVs. The first such feature considers the specific genes disrupted by a CNV, computing a density for each benign CNV's disrupted genes. The least likely of a CNV's genes is selected to represent the CNV, resulting in the function

$$P\left(genes|benign\right) = 1 + \min_{g \in genes}\left[count_B\left(g\right)\right] \tag{3.8}$$

where $count_B(g)$ is the number of benign CNVs that overlap gene g. One is added to the minimum count to ensure that the function's range is strictly greater than zero, because the overall CNV scoring function (Eq. 3.1) takes the logarithm of this likelihood.

3.3.5 Network Density Scoring

The gene density scoring in the previous section is perhaps too restrictive, counting only direct matches between overlapped genes. To blur this relationship somewhat, we now consider the gene network of a CNV, a superset of the CNV's directly disturbed genes. Gene networks are the union of the neighborhoods of a CNV's disturbed genes. A gene's neighborhood is constructed using the connection information from the STRING database[22], finding likely, immediate neighbors in the connectivity graph. (That is, with edge p-value of at least 0.95 and a graph distance of at most 1.)

We modify the density function described in Equation 3.8 only slightly, producing

$$P(network|benign) = 1 + \min_{g \in network} [count_B(g)] \tag{3.9}$$

where $count_B(g)$ now represents a weighted number of benign CNV networks that include gene g. As before, any benign CNV that overlaps g contributes 1 to the count, but CNVs with g as a neighbor gene contribute $\frac{1}{2}$. This simple weighting scheme encodes the relative importance of the genes in a CNVs network.

3.4 Gene Set Scoring

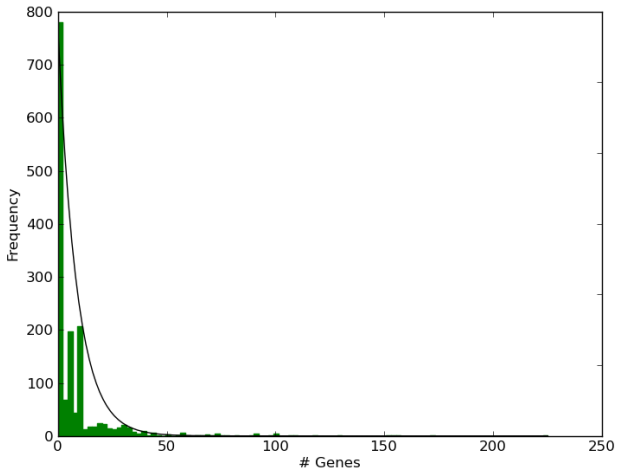
Similar to the networks created from STRING data in the previous section, Gene Set Enrichment Analysis (GSEA)[23] can be used to produce a set of statistically significant gene sets that may be representative of phenotypic effects of the CNV. This gene-set feature provides another way of observing the effect of disturbed genes in a less-restricted manner.

We modify the density function in Equation 3.8 again, creating

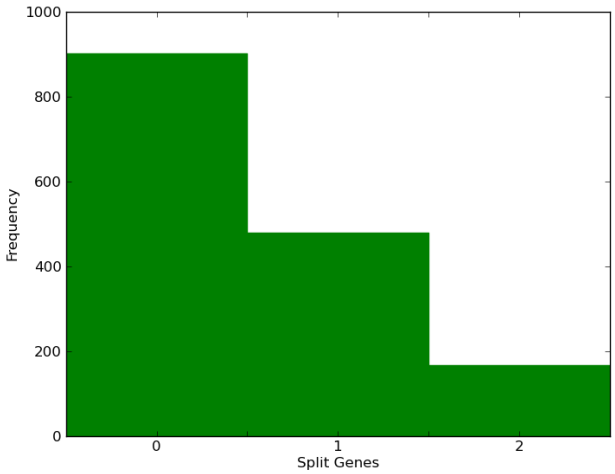
$$P(geneset|benign) = 1 + \min_{gs \in geneset} [count_B(gs)] \tag{3.10}$$

where $count_B(gs)$ is the number of benign CNVs with a disturbed gene in gene-set gs.

This method of scoring operates at a higher level of abstraction than the previous two, focusing not on individual genes but instead on the greater biological processes at work. We hypothesize that this will be beneficial in cases where an incomplete model of benign CNVs hinders the performance of the other gene-based scoring functions.



(a) Spanned gene frequency. $\lambda \approx \frac{1}{8.6}$.



(b) Split gene frequency.

Figure 3.2: Frequency of benign CNV gene interactions.

Chapter 4

Results

The following experimental results are generated using leave-one-out cross validation to assess patient classification accuracy. The classification results are visualized on Receiver Operating Characteristic (ROC) plots for comparison, with a dotted line indicating a 95% specificity level and a dashed no-indication line.

4.1 How Effective is Expert Scoring?

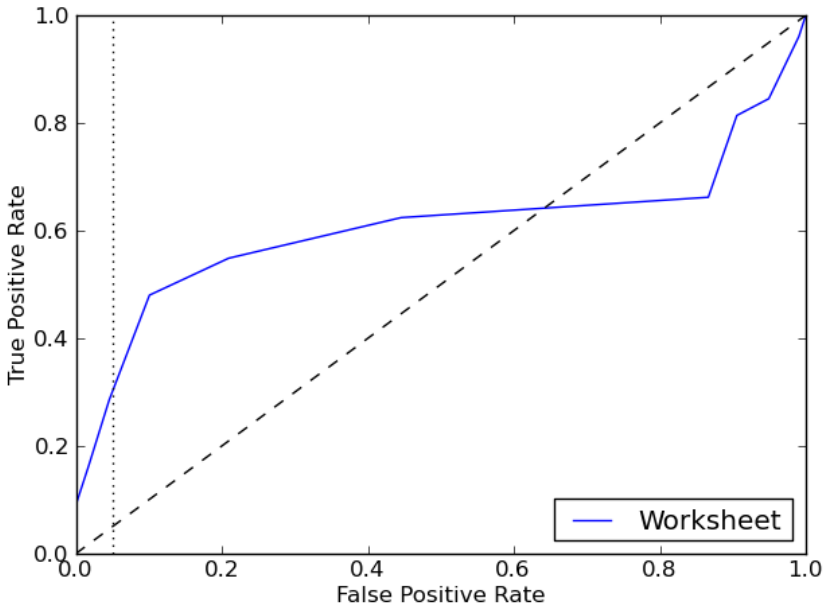
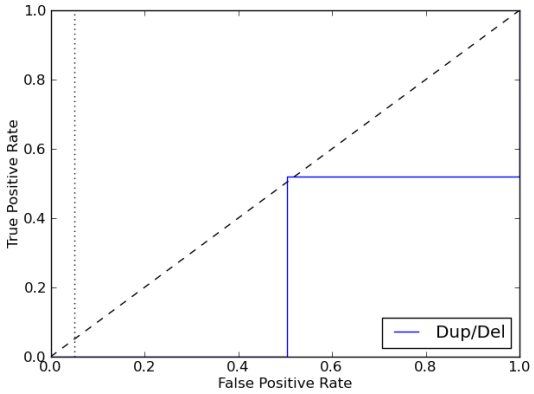


Figure 4.1: Expert worksheet scoring classification results.

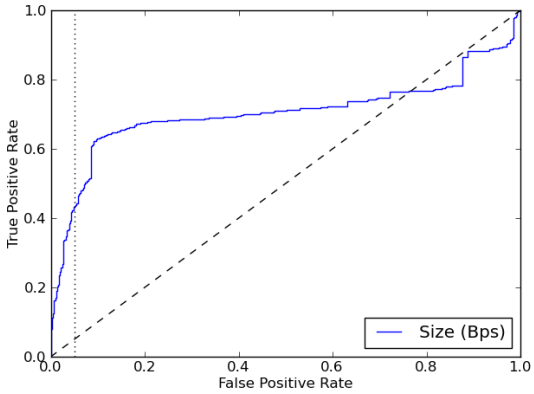
This experiment shows that our version of the expert worksheet classifier does not adequately separate pathogenic from benign CNVs. As shown in Figure 4.1, only 30.0% sensitivity is achieved at a specificity of 95%. This result aligns with the survey data presented in Figure 1.1, in which a general lack of concordance between classification attempts is observed. This confirms our suspicion that, for the problem of assessing uncharacterized CNVs, current clinical practices are not adequate.

4.2 Which Features are Most Effective?

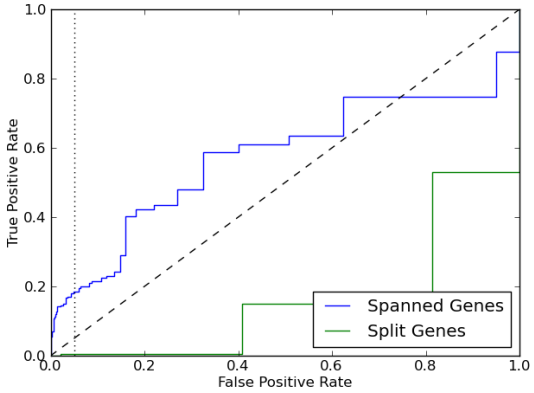
This experiment attempts to improve on the expert worksheet scoring system by applying our more statistically rigorous approach (described in Section 3.2) to an expanded set CNV features. First, we score each feature independently (Figure 4.2, Table 4.1).



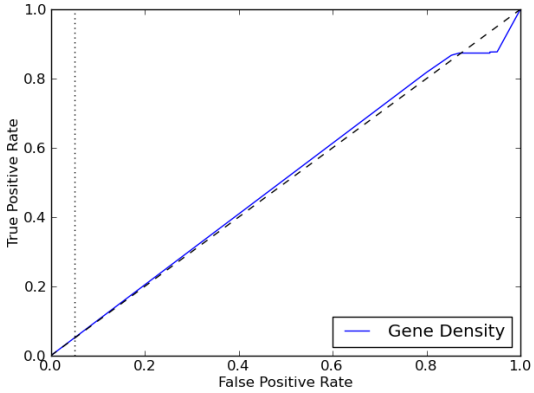
(a) Duplication vs Deletion



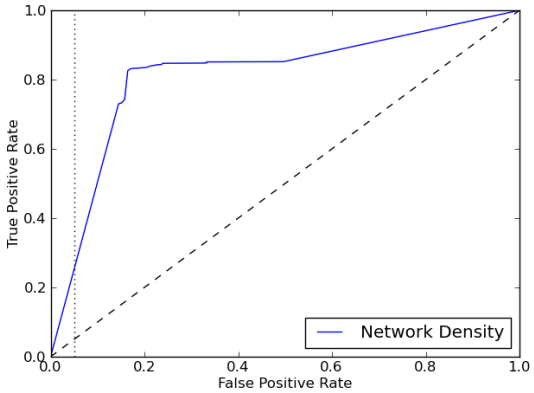
(b) Size



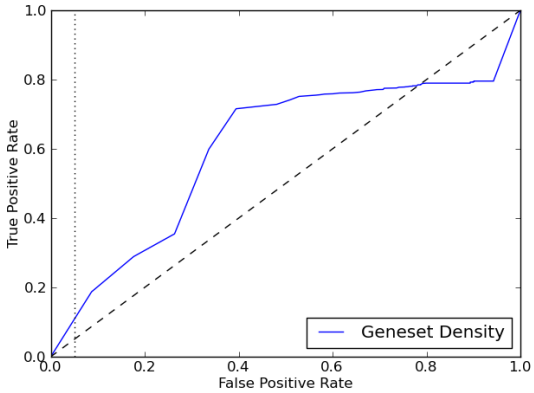
(c) Gene Overlap Counts



(d) Gene Density



(e) Network Density



(f) Gene Set Density

Figure 4.2: Independent feature scoring classification results.

Table 4.1: Independent feature scoring classification results. Sensitivity values are measured at the 95% specificity level.

Feature	Sensitivity	Area Under ROC
Dup/Del	0.0%	0.259
Size	43.3%	0.704
Genes Spanned	18.5%	0.576
Genes Split	0.4%	0.162
Gene Density	5.1%	0.504
Network Density	25.3%	0.815
Gene Set Density	10.9%	0.601

These results provide a clear look at which features best discriminate between pathogenic and benign CNVs. We observe improvement over the expert scoring function using only CNV size, a result that underscores the utility of our statistical approach. This result follows the intuition that, generally, large CNVs have more of an effect on biological processes than small CNVs. Other features, like the number of spanned genes, show some discriminative ability but fail to stand on their own.

As expected, the network density and gene set density features outperform the basic gene density feature. This confirms our hypothesis that restricting comparison to shared gene overlaps does not capture the full complexity of the biological interaction.

4.3 Worksheet Feature Correlation

To further compare our approach against the expert scoring system, we examine the correlation between the expert CNV scores and a function in our framework composed of the same four features, which we refer to as the "probabilistic worksheet" scoring function (Figure 4.3).

We measure a positive correlation between the two CNV scoring functions for both sets of CNVs. This demonstrates that our statistical approach does indeed emulate the behavior of the expert scoring worksheet. We further examine this relationship by plotting the classification results of the two functions on a ROC curve in Figure 4.4. While this result indicates that the probabilistic model failed to improve upon our baseline with this feature

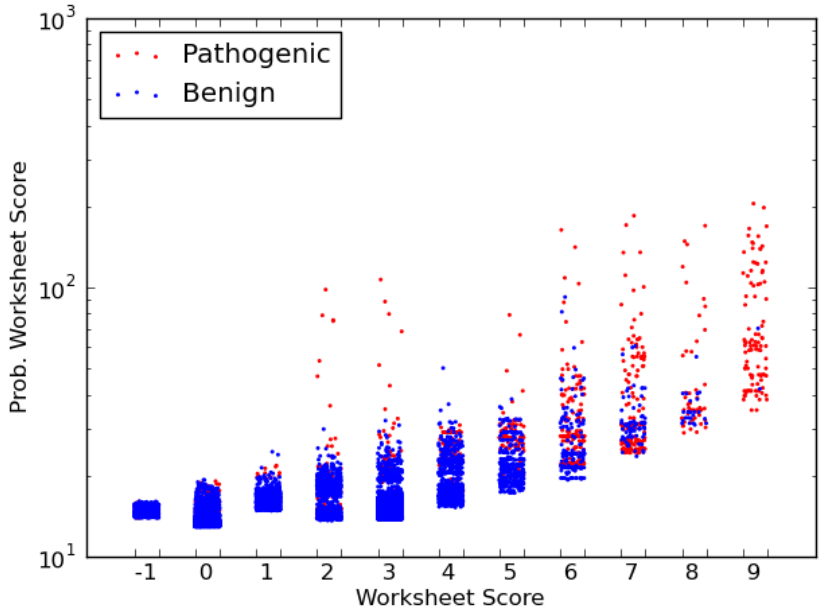


Figure 4.3: CNV scoring correlation. A ± 0.2 random jitter is applied to the integer worksheet scores for added clarity. Pearson correlation coefficients (R) for pathogenic and benign sets are 0.593 and 0.552, respectively.

combination, it also provides evidence that our framework produces reasonable CNV scoring functions.

4.4 Feature Combination

We now select the most promising features from the independent feature experiments for use in a combined-feature "fusion" scoring function. The resulting function follows the form of Equation 3.1, summing the negative log-likelihoods of each feature.

Figure 4.5 reveals that an unweighted sum of size, spanned genes, network, and gene-set density scores actually performs worse than any of its individual constituent features. This indicates that the sum of feature log-likelihoods will require non-uniform weight coefficients, to model the relative influence of each log-likelihood. Many methods exist for learning the coefficients of this weighted sum, but that problem is beyond the scope of this thesis and is discussed more fully below.

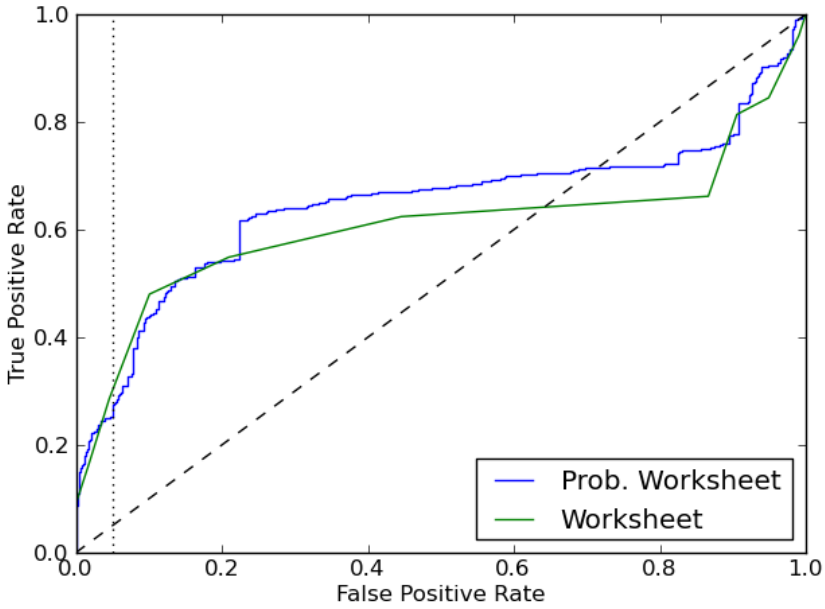


Figure 4.4: Probabilistic and expert worksheet scoring classification results. The probabilistic worksheet achieves slightly lower sensitivity at 95% specificity, but generally follows the expert scoring function closely in the critical 80%-100% specificity range.

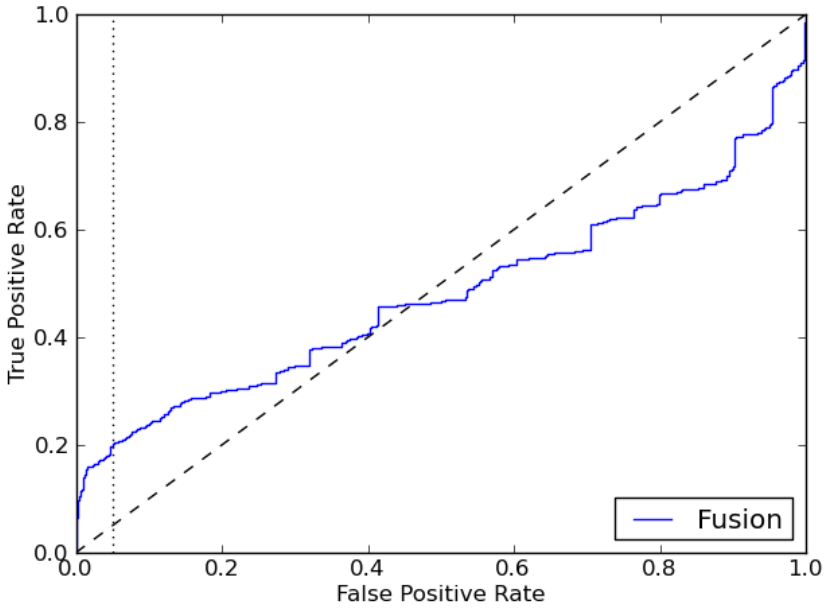


Figure 4.5: Fusion scoring classification results. The combined-feature scoring function, made up of size, spanned genes, network density, and gene-set density features, achieves only 19.8% sensitivity at 95% specificity. Total area under the ROC curve is 0.480.

Chapter 5

Discussion

The results of our experiments provide insight into the current state of uncharacterized CNV classification. It is clear that current clinical techniques do not yield trustworthy interpretations, but the results of our experiments with CNV features in a statistical framework show promise. Attempts to combine features reveal the need for non-uniform feature likelihood weights, an additional learning problem discussed in more detail below.

The methods of this study have some limitations, stemming from mixed data sources, unknown labels at the CNV level, and independence assumptions made on CNV features.

5.1 Heterogeneous Data

The first limitation relates to the method of combining datasets, as it is difficult to tell whether observed distinctions between CNVs of affected and normal patients are true differences or simply artifacts of different CNV identification and reporting methods. CNVs from both the Itsara and ISCA sets were generated using CMA, but not necessarily with the same sampling parameters. As discussed in Section 2.1, the density of probe coverage on each chromosome introduces variability in CNV identification, affecting the accuracy of CNV boundaries as well as the resolution at which CNVs can be reliably measured. More samples from the array afford higher confidence, so different screens from the same patient tend to agree the larger CNVs, but may not be able to reproduce smaller mutations.

Future work would benefit from a standardized database of characterized CNVs as proposed by Tsuchiya et al. [11], with guarantees about the methods of data collection and CNV identification. Such a database would have immediate utility in the clinic, while also providing a reliable data set for large-scale CNV studies. This task is nontrivial, however. Analysis of the Database of Genomic Variants revealed that 88% of previously reported CNVs were inaccurately sized[30], a symptom of the imprecise and unstandardized CNV detection methods used in various laboratories. CMA whole-genome hybridization techniques have improved CNV detection significantly, but data inaccuracies will continue to plague CNV studies until a gold standard is established.

5.2 Multiple Instance Problem

The second limitation arises from the lack of ground-truth labels for individual CNVs. In an ideal scenario, the problem of CNV characterization could be simplified to a binary classification: a CNV either contributes to phenotypic abnormalities or does not. Many well-known machine learning techniques are suited to binary classification, and are easy to apply to a variety of problems. However, in the case of CNV characterization, instance labels are available only at the patient level. This introduces the multiple instance problem: a pathogenic patient may have many CNVs, and it is unknown which or how many of them are contributory to the observed phenotypic abnormality.

Multiple instance learning (MIL) was first introduced as such by Dietterich and Lathrop [31] in 1997, and remains an active research area today[32, 33]. Early applications of MIL focused on drug activity prediction tasks, then expanded to scene classification in images[34], and now MIL techniques pervade machine learning.

Several algorithms of varying complexities have been proposed for addressing the multiple instance problem[35, 36], but the scoring functions in this study deal with the multiple instance problem much more simply: scoring each of a patient's CNVs individually, then choosing the most extreme score for patient classification. This practice follows from a simple interpretation of the underlying biological processes: one key breakdown in an important pathway will produce detrimental effects, regardless of other mutations. This allows one model to be constructed for both patient phenotype prediction and CNV characterization.

However, the approach does not account for cases in which pathogenicity arises from a combination of otherwise benign CNVs. Few such cases have been identified, and the relevance of this "combinatorial pathogenicity" is unclear. As such, further study will be required to establish the full importance of these potential CNV interactions.

5.3 Reference Ranges and Feature Independence

The multiple instance problem complicates a full Bayesian treatment of the data, for which several methods of statistical MIL have been developed [37, 38, 39]. Our approach, however, draws from the precedent set by medical practice of using a reference range to model the set of known benign CNVs, as described in Section 3.2. This relies on the assumption that the features used to construct the reference range are conditionally independent, given benign CNVs. Clearly, this is not true for the size of a CNV and number of genes it spans. This is a significant limitation of our method, but, as is the case in many machine learning applications, the benefit of a simple model– a result of the naive independence assumption– outweighs the costs of mathematical correctness.

5.4 Likelihood Weight Learning

As previously discussed, our model suffers from the assumption that all feature log-likelihoods contribute equally to a CNV pathogenicity score. To learn these weights, a second layer of learning should be added to the model. Under our framework, using k different CNV features, it is simple to consider CNVs as k-dimensional vectors in feature log-likelihood space. Our current approach, therefore, scores a CNV by the negative sum of its vector. A more effective approach might instead use the MIL algorithms referred to in Section 5.2, like the support vector machine method described in Andrews et al. [36]. A less rigorous treatment might sidestep the multiple instance problem by choosing the least likely benign CNV with the naive method, associating the patient label with that CNV, then learning weights using standard techniques. The discriminative power of individual feature scoring in this study suggests that future work on feature combination may lead to significant improvements in CNV interpretation.

Chapter 6

Conclusion

Copy Number Variation analysis has recently become a viable tool for patient diagnosis in clinical settings, thanks to the development of relatively inexpensive and accurate array-based detection methods. The wealth of CNV data being generated is underutilized, however, especially in the case of uncharacterized CNVs. This thesis performs a detailed examination of one assessment technique currently in clinical use, then presents a novel statistical method for CNV interpretation. Experimental results confirm previous findings that current clinical CNV interpretation methods are inadequate and that there is significant room for improvement. The results indicate the effectiveness of a principled, feature-based, statistical framework for uncharacterized CNV interpretation, which future studies can expand upon to construct more reliable classifiers.

Bibliography

[1] Redon, R.; Ishikawa, S.; Fitch, K.; Feuk, L.; Perry, G.; Andrews, T.; Fiegler, H.; Shapero, M.; Carson, A.; Chen, W. et al.Redon, R. and Ishikawa, S. and Fitch, K.R. and Feuk, L. and Perry, G.H. and Andrews, T.D. and Fiegler, H. and Shapero, M.H. and Carson, A.R. and Chen, W. and others *Nature* **2006**, *444*, 444–454.

[2] Stranger, B. E. et al. *Science* **2007**, *315*, 848–853.

[3] Sebat, J. et al. *Science* **2007**, *316*, 445–449.

[4] St Clair, D. *Schizophrenia Bulletin* **2009**, *35*, 9–12.

[5] Iafrate, A.; Feuk, L.; Rivera, M.; Listewnik, M.; Donahoe, P.; Qi, Y.; Scherer, S.; Lee, C. *Nature genetics* **2004**, *36*, 949–951.

[6] Karolchik, D.; Baertsch, R.; Diekhans, M.; Furey, T.; Hinrichs, A.; Lu, Y.; Roskin, K.; Schwartz, M.; Sugnet, C.; Thomas, D. et al.Karolchik, D. and Baertsch, R. and Diekhans, M. and Furey, T.S. and Hinrichs, A. and Lu, YT and Roskin, K.M. and Schwartz, M. and Sugnet, C.W. and Thomas, DJ and others *Nucleic acids research* **2003**, *31*, 51.

[7] Chen, F.; Chen, Y.; Chuang, T. *Bioinformatics* **2009**, *25*, 1419.

[8] Shaikh, T.; Gai, X.; Perin, J.; Glessner, J.; Xie, H.; Murphy, K.; O'Hara, R.; Casalunovo, T.; Conlin, L.; D'arcy, M. et al.Shaikh, T.H. and Gai, X. and Perin, J.C. and Glessner, J.T. and Xie, H. and Murphy, K. and O'Hara, R. and Casalunovo, T. and Conlin, L.K. and D'arcy, M. and others *Genome research* **2009**, *19*, 1682.

[9] Firth, H.; Richards, S.; Bevan, A.; Clayton, S.; Corpas, M.; Rajan, D.; Vooren, S.; Moreau, Y.; Pettett, R.; Carter, N. *The American Journal of Human Genetics* **2009**, *84*, 524–533.

[10] Church, D.; Lappalainen, I.; Sneddon, T.; Hinton, J.; Maguire, M.; Lopez, J.; Garner, J.; Paschall, J.; DiCuccio, M.; Yaschenko, E. et al.Church, D.M. and Lappalainen, I. and

Sneddon, T.P. and Hinton, J. and Maguire, M. and Lopez, J. and Garner, J. and Paschall, J. and DiCuccio, M. and Yaschenko, E. and others *Nature genetics* **2010**, *42*, 813–814.

[11] Tsuchiya, K.; Shaffer, L.; Aradhya, S.; Gastier-Foster, J.; Patel, A.; Rudd, M.; Biggerstaff, J.; Sanger, W.; Schwartz, S.; Tepperberg, J. et al.Tsuchiya, K.D. and Shaffer, L.G. and Aradhya, S. and Gastier-Foster, J.M. and Patel, A. and Rudd, M.K. and Biggerstaff, J.S. and Sanger, W.G. and Schwartz, S. and Tepperberg, J.H. and others *Genetics in Medicine* **2009**, *11*, 866.

[12] on Human Cytogenetic Nomenclature, I. S. C.; Shaffer, L.; Slovak, M.; Campbell, L. *ISCN 2009: an international system for human cytogenetic nomenclature (2009)*; Karger, 2009.

[13] Miller, D.; Adam, M.; Aradhya, S.; Biesecker, L.; Brothman, A.; Carter, N.; Church, D.; Crolla, J.; Eichler, E.; Epstein, C. et al.Miller, D.T. and Adam, M.P. and Aradhya, S. and Biesecker, L.G. and Brothman, A.R. and Carter, N.P. and Church, D.M. and Crolla, J.A. and Eichler, E.E. and Epstein, C.J. and others *The American Journal of Human Genetics* **2010**, *86*, 749–764.

[14] Wang, K.; Li, M.; Hadley, D.; Liu, R.; Glessner, J.; Grant, S.; Hakonarson, H.; Bucan, M. *Genome Research* **2007**, *17*, 1665.

[15] Colella, S.; Yau, C.; Taylor, J.; Mirza, G.; Butler, H.; Clouston, P.; Bassett, A.; Seller, A.; Holmes, C.; Ragoussis, J. *Nucleic acids research* **2007**, *35*, 2013.

[16] Day, N.; Hemmaplardh, A.; Thurman, R.; Stamatoyannopoulos, J.; Noble, W. *Bioinformatics* **2007**, *23*, 1424.

[17] Tsuang, D.; Millard, S.; Ely, B.; Chi, P.; Wang, K.; Raskind, W.; Kim, S.; Brkanac, Z.; Yu, C. *PloS one* **2010**, *5*, e14456.

[18] Itsara, A.; Cooper, G.; Baker, C.; Girirajan, S.; Li, J.; Absher, D.; Krauss, R.; Myers, R.; Ridker, P.; Chasman, D. *The American Journal of Human Genetics* **2009**, *84*, 148–161.

[19] Mailman, M.; Feolo, M.; Jin, Y.; Kimura, M.; Tryka, K.; Bagoutdinov, R.; Hao, L.; Kiang, A.; Paschall, J.; Phan, L. *Nature genetics* **2007**, *39*, 1181–1186.

[20] Maglott, D.; Ostell, J.; Pruitt, K.; Tatusova, T. *Nucleic Acids Research* **2005**, *33*, D54.

[21] Haider, S.; Ballester, B.; Smedley, D.; Zhang, J.; Rice, P.; Kasprzyk, A. *Nucleic acids research* **2009**, *37*, W23.

[22] Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P. *Nucleic Acids Research* **2011**, *39*, D561.

[23] Subramanian, A.; Tamayo, P.; Mootha, V.; Mukherjee, S.; Ebert, B.; Gillette, M.; Paulovich, A.; Pomeroy, S.; Golub, T.; Lander, E. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 15545.

[24] Pritchard, C. unpublished.

[25] Gräsbeck, R.; Saris, N. *Scand J Clin Lab Invest* **1969**, *26*, 62–3.

[26] Ceriotti, F.; Hinzmann, R.; Panteghini, M. *Ann Clin Biochem* **2009**, *46*, 8–17.

[27] Barth, J. *Annals of Clinical Biochemistry* **2009**, *46*, 1.

[28] Shine, B. *Ann Clin Biochem* **2008**, *45*, 467–475.

[29] Tate, J. R.; Ferguson, W.; Bais, R.; Kostner, K.; Marwick, T.; Carter, A. *Ann Clin Biochem* **2008**, *45*, 275–288.

[30] Perry, G.; Ben-Dor, A.; Tsalenko, A.; Sampas, N.; Rodriguez-Revenga, L.; Tran, C.; Scheffer, A.; Steinfeld, I.; Tsang, P.; Yamada, N. et al.Perry, G.H. and Ben-Dor, A. and Tsalenko, A. and Sampas, N. and Rodriguez-Revenga, L. and Tran, C.W. and Scheffer, A. and Steinfeld, I. and Tsang, P. and Yamada, N.A. and others *The American Journal of Human Genetics* **2008**, *82*, 685–695.

[31] Dietterich, T. G.; Lathrop, R. H. *Artificial Intelligence* **1997**, *89*, 31–71.

[32] Amar, R. A.; Dooly, D. R.; Goldman, S. A.; Zhang, Q. *Multiple-Instance Learning of Real-Valued Data*, 2001.

[33] Babenko, B.; Yang, M.; Belongie, S. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2010**.

[34] Maron, O.; Ratan, A. L. Multiple-Instance Learning for Natural Scene Classification. *In The Fifteenth International Conference on Machine Learning*, 1998; pp 341–349.

[35] Maron, O.; Lozano-Pérez, T. A framework for multiple-instance learning. *Advances in neural information processing systems*, 1998; pp 570–576.

[36] Andrews, S.; Tsochantaridis, I.; Hofmann, T. *Advances in neural information processing systems* **2003**, 577–584.

[37] Wang, J.; Zucker, J. Solving the multiple-instance problem: A lazy learning approach. *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, 2000; pp 1119–1126.

[38] Zhang, Q.; Goldman, S. *Advances in neural information processing systems* **2002**, *2*, 1073–1080.

[39] Raykar, V.; Krishnapuram, B.; Bi, J.; Dundar, M.; Rao, R. Bayesian multiple instance learning: automatic feature selection and inductive transfer. *Proceedings of the 25th international conference on Machine learning*, 2008; pp 808–815.

Vita

Clifton M. Carey, Jr.

Date of Birth	July 30, 1989
Place of Birth	Gaithersburg, Maryland
Degrees	B.S. Computer Science, May 2011
Professional Societies	Association for Computing Machinery

May 2011

CNV Interpretation, Carey, M.S. 2011