Text Based Methods for Variant Prioritization

by

Michael Gottlieb

B.A. Asian Studies & Psychology, The University of British Columbia, 2009

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

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Bioinformatics)

The University of British Columbia
(Vancouver)

January 2017

Abstract

Despite improvements in sequencing technologies, DNA sequence variant interpretation for rare genetic diseases remains challenging. In a typical workflow for the Treatable Intellectual Disability Endeavor in B.C. (TIDE BC), a geneticist examines variant calls to establish a set of candidate variants that explain a patient's phenotype. Even with a sophisticated computation pipeline for variant prioritization, they may need to consider hundreds of variants. This typically involves literature searches on individual variants to determine how well they explain the reported phenotype, which is a time consuming process. In this work, text analysis based variant prioritization methods are developed and assessed for the capacity to distinguish causal variants within exome analysis results for a reference set of individuals with metabolic disorders.

Preface

Contributions

The division of TIDE cases into training and test sets was proposed by Drs. Wyeth Wasserman and Maja Tarailo-Graovac. The cases in each set were selected by Drs. Maja Tarailo-Graovac and Allison Matthews.

The idea to do a simulated test of Synverita's capabilities for variant prioritization was conceived of by Dr. Steven Jones.

The original idea to use Synverita for variant prioritization was my own. The methods described in this work were developed, implemented and tested by me with advice from Emily Hindalong.

Publications

There are no publications based on this work at this time.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank Drs. Wyeth Wasserman and Steven Jones for supporting me in pursuing this project. They provided me the resources of the CMMT and GSC, enabling the work to move rapidly. I thank Dr. Anna Lehman for serving on my committee, providing insightful feedback and helping me ground this work with a clinical perspective.

My thanks to Dr. Allison Matthews for providing me with the whole genome and exome data I needed and feedback on my progress, and to Jake Lever for granting me access and support in using Synverita and updating Synverita for the needs of variant prioritization.

Special thanks go to Dr. Maja Tarailo-Graovac, whose guidance and support ensured that I stayed on track in all aspects of this work. I owe my success to her patience and incredible ability to make sense of medical genomics.

I would like to thank my wife, Emily Hindalong for encouraging me and refining my work through many discussions.

I acknowledge the Canadian Institute of Health Research for funding my studies.

Chapter 1

Introduction

1.1 Background

Rare Mendelian diseases are caused in most cases by DNA sequence variations within one or two alleles of a gene in a patient. Despite individual disease rarity, it is estimated that 8% of people have rare disease worldwide [22] [6]. The altered genes for about 50% of rare diseases have been discovered, but few have known treatments [40] [11].

Within this thesis the research is focuses upon a subset of rare diseases characterized by both deficits in intellectual development and metabolite processing. Intellectual developmental disorders (IDD) are a group of disorders with heterogeneous etiologies, including Mendelian diseases, that occur in about 2-3% of children worldwide [13] [38]. They are marked by lifelong disturbances in multiple cognitive domains that first manifest early in life [34]. IDD are amongst the costliest disorders due to the degree of impairment and lifelong duration [27]. Inborn errors of metabolism (IEM) are a category of Mendelian diseases in which a single protein's functioning is disrupted in a metabolic pathway. IEMs can be difficult to diagnose, as they may present in patients as diverse phenotypes, for instance as immunodeficiency [10] or IDD [33]. Individuals with the same IEM may present different disease phenotypes due to other genetic differences and interactions with other metabolic pathways [4]. Conversely, IEMs caused by distinct genetic mechanisms may present with the same phenotype. For example, at present over 90 IEMs are known to cause IDD [44].

IEMs are amongst (if not the top) most likely to be treatable Mendelian diseases [3]. IDDs due to IEM are the largest group of treatable IDDs with genetic causes [43] [44]. The prototypic example of IDD due to IEM is phenylketonuria, which occurs when mutations in the PAH gene disrupt the metabolism of phenylalanine. This in turn allows dietary phenylalanine to accumulate, eventually reaching toxic levels and causing IDD. As phenylketonuria is well known, it is often screened for, and individuals with phenylketonuria are able to develop normally by adhering to a low phenylalanine diet [2].

For an individual with an IEM that can lead to IDD, timely discovery of the causal variant(s) is needed for effective treatment and management [43][41]. However, for many IEMs, causal genes are either unknown or the observed phenotypes could arise from defects in any of multiple genes. Due to the inexact relationship between disease phenotype and genotype of an individual, single gene tests can be inconclusive even when caused by a known disease gene variant [38]. Early diagnosis of treatable conditions can be important to avoid tissue damage from toxic compounds.

Causal variant discovery of rare Mendelian diseases has been revolutionized by next generation sequencing (NGS). Analysis of whole genome sequencing (WGS) and targeted sequencing of protein coding regions, known as whole exome sequencing (WES), have revealed the causal variants for many rare diseases [11] [7]. WGS and WES provide a solution to the problem of targeted testing as they check all genes for mutations at a cost that is approaching single gene tests [38] [29].

The broad application of exome and WGS to metabolic disorders has been shown to have a high diagnostic success rate [38] [39]. In the Treatable Intellectual Disability Endeavor in B.C. (TIDE BC), the DNA sequences from a patient are analyzed using a semi-automated bioinformatics pipeline [41] to establish a list of variations that may explain the patient's symptoms. Based on availability, genomic data from close relatives may be included, most commonly as a trio (mother-father-child). Inclusion of close relatives allows for variations to be deprioritized if present in healthy individuals (for dominant models), or for the inheritance patterns of recessive candidates to be confirmed. The results are then reviewed by an interdisciplinary team, leveraging their respective expertise, to establish the top causal candidate(s). The experts consider factors such as variant inheritance pattern, the relationship between the gene function and patient phenotype (and known phenotypes of individuals with disruptions to the gene) and pathogenicity estimates [41].

This workflow, outlined in Figure 1.1, is executed in two cycles. After receiving the output of automated analysis, the bioinformatics team semi-manually reviews each variant and relevant literature to form a hypothesis that explains the patient's phenotype. The number of variants per patient can range from dozens in a typical WES analysis of a trio to a few hundred in the event of a WES analysis restricted to the patient. As manual analysis is time consuming, the number of variants to consider and the order in which they are considered can result in diagnosis and treatment delays for individuals with treatable IDD.

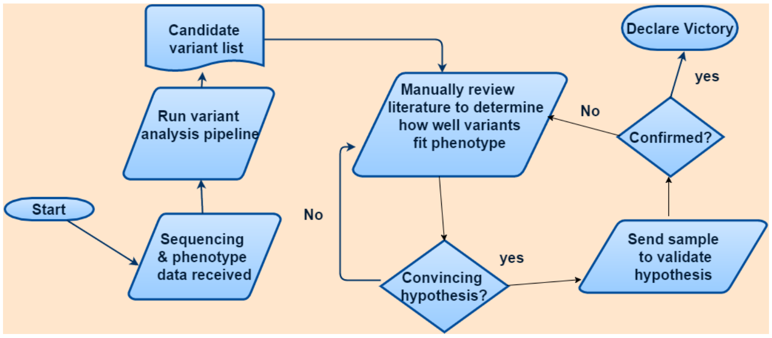
Candidate variant list
Declare Victory
yes
Manually review literature to determine how well variants fit phenotype
No
Run variant analysis pipeline
Confirmed?
Start
No
Sequencing & phenotype data received
Send sample to validate hypothesis
yes
Convincing hypothesis?

Figure 1.1: Overview workflow for establishing best candidate variants used by TIDE BC.

1.2 Automated Variant Prioritization

Automated variant prioritization is a set of techniques that aim to establish a ranking amongst a set of variants. The goal of variant prioritization in this case is to rank variants that are more likely to explain a patient's phenotype higher than those that are not. Such approaches are expected to accelerate diagnosis. There have been numerous prioritization procedures introduced in the literature, which can be roughly categorized into four broad (and occasionally overlapping) categories: features of genes; properties of the specific sequence altered; human phenotype ontology (HPO) leveraging; or text mining.

1.2.1 Properties of Specific Sequence Alterations

Approaches that prioritize variants based on properties of the specific sequence alteration do so without taking phenotype into consideration. This includes features such as the type of variation and its predicted or known effect on its **transcript**, as in the SNP Effect Predictor [26]. This category also includes prioritization based on how frequently specific variants appear in the general population from sources such as ExAC [25]. Given an observed disease phenotype and a list of variants, these methods will predict how likely a variant is related to disease or deleterious. However, these methods will not provide any measure of how well a variant fits the observed disease phenotype.

Combined Annotation-Dependent Depletion (CADD): CADD [21] is a generalized approach for scoring single-nucleotide variant and insertion-deletion events. It combines multiple measures and annotations in order to produce a single score that estimates the likelihood of a specific variant being deleterious. This alleviates the knowledge-biases that individual metrics of pathogenicity suffer from. CADD scores have been pre-calculated for all possible single-nucleotide variants in the hg19 reference, which enables quick ranking of variants with a low computational threshold for use.

1.2.2 Features of Genes

Approaches that priortize variants using features of genes also do not take phenotype into consideration. These approaches include features such as selective pressure or degree of functional diversity as in RVIS [30]. They share the same limitations as the preceding category as they do not estimate how likely a variant gene is involved in a particular disease phenotype.

Residual Variation Intolerance Score (RVIS): RVIS [30] is a gene level approach for scoring variants. Unlike CADD, which can produce multiple scores for different variants in one gene, RVIS produces one score per gene. RVIS aims to measure the degree to which a gene can tolerate functional variation. This is estimated by regressing on the ratio of commonly observed missense and stop-gained mutations to total number of observed variants in the gene. Genes that have fewer commonly observed missense and stop-gained mutations than expected are considered less tolerant of variation and have lower RVIS scores. That is, non-synonymous variants in a gene with a low RVIS are more likely to be deleterious.

FLAGS: Unlike RVIS and CADD, [35] focuses on identifying features of genes in which non-synonymous mutations are less likely to be pathogenic in rare diseases. FLAGS is a ranking system that deprioritizes genes that have more associated publications, longer coding sequences, more paralogs and less selective pressure as defined by the dN/dS ratio [45]. Not to be confused with filtering, FLAGS identifies genes that should be interpreted with caution when examining their variants for pathogenicity.

Ongoing work in variant prioritization continues and extends well beyond the above mentioned projects. For example, recent work has investigated methods for prioritization based on biological networks [28], but these have been met with limited success, likely due to gaps in our current understanding of biological networks [31].

1.2.3 HPO Leveraging

The Human Phenotype Ontology (HPO) [24] is a hierarchically structured set of human phenotypic abnormalities. Its structure supports a number of computational approaches to phenotype comparisons for rare disease diagnostics [23][18][20][12][19]. Computational approaches to phenotype comparisons have also been used for variant prioritization in conjunction with other methods. An in-depth review is presented in Smedley et al. [37].

Exomiser: Exomiser[36] is a variant prioritization system that combines several methods of phenotype comparison with a variety of gene/variant features and measures. Given a patient phenotype and variants, it ranks the variants based on how well they match established disease phenotypes for that gene and related genes. Unlike other variant HPO leveraging prioritization systems, Exomiser also considers established disease phenotypes of model organisms. Exomiser is available as a standalone tool with modest requirements and requires VCF files and phenotypes described using the HPO [9].

1.2.4 Text Mining

Text mining approaches use automatic analysis of relevant literature in order to rank links between disease phenotypes and genes. The two examples below function like a search engine over primary literature and curated sources. Both use the Unified Medical Language System (UMLS)[8] for biomedical term identification. The UMLS provides a consolidated identification system that spans biomedical concepts across many biomedical vocabularies.

FindZebra: FindZebra [14] is a web-based tool focused on the rare disease domain. FindZebra uses a small corpus that only includes online sources relevant to the rare disease domain, such as Online Mendelian Inheritance in Man and the Genetic and Rare Diseases Information Center. Using a Google-like user interface, FindZebra accepts free text biomedical queries. It then ranks the documents in its corpus based on relevance to the query and clusters them based on UMLS concepts. It also provides the option to rank UMLS concepts related to the query, which can help researchers prioritize genes for variant analysis.

FindZebra's small corpus gives it good specificity as it mines from highly curated sites. This means that it does not include low quality associations that might be available in a general web search. However, this negatively affects its sensitivity because curated sources can be slow to update and less complete. Nonetheless, it provides a simple application programming interface (API) that allows it to be used in the context of automated analysis pipelines.

Beegle: Beegle[15] is a web-based tool built on top of the Endeavour gene prioritization tool [1][42]. Like FindZebra, Beegle uses a Google-like search interface and accepts free text user queries in the biomedical domain. Beegle then finds a list of associated MEDLINE abstracts via PubMed, including 10,000 PubMed Ids. It uses these articles to construct a UMLS profile from approximately 67,000 UMLS concepts. Beegle calculates similarity between the query and genes by counting abstract level co-occurrences. It generates a strength of association by comparing the number of co-occurrences to the total number of abstracts in which each term occurs. It also compares the number of concepts associated with each term as a second metric of similarity. The result of this stage is a list of genes known to be associated with a given query term. The user selects some number of these known associated genes and selects a number of candidate genes that they are interested in. This allows Beegle to produce a list of genes it predicts to be associated with the query.

Beegle's larger corpus and association predictions offer greater sensitivity than FindZebra at the cost of specificity. However, Beegle's computations are performed on-the-fly, which can result in long delays. Furthermore, it does not provide an API, which makes it difficult to use for automated analysis pipelines.

1.3 Toward Improved Text-Based Variant Prioritization

Individual cases that are referred for TIDE BC's WES or WGS bioinformatics analysis have inconclusive findings from single gene and panel tests. For a given case, automated variant analysis may result in dozens of variant genes to consider with respect to a clinician provided free text deep phenotype. As there are many variant genes to consider and rareness of the disease phenotype, successful variant prioritization in TIDE BC's bioinformatics analysis workflow should have minimal running time or user interaction and be sufficiently sensitive to link a variant gene to the observed disease phenotype.

RVIS and CADD are currently incorporated into TIDE BC's variant analysis pipeline. These scores require no user input to generate and are produced quickly. However, they do not provide any indication as to how well a variant or variant gene fits with the observed disease phenotype. Exomiser, Beegle, and FindZebra all score how well variant genes fit a disease phenotype. Exomiser, especially in its leveraging of model organisms, promises more sensitive performance. Its requirement of HPO terms provides a challenge, as referring clinicians provide patient phenotype using free text. Beegle and FindZebra can both score phenotype gene associations and accept free text phenotypes. Beegle's large corpus, vocabulary and predictive capacity appear to be a good fit for this application, but its lack of a back end and long on the fly computation prevent it from being incorporated into the current pipeline. FindZebra's simple API allow it to be integrated into the current pipeline, but its small corpus size may result in reduced sensitivity.

A candidate solution to this problem is a current biomedical text mining project at Canada's Michael Smith Genome Sciences Centre, Synverita. Synverita is under development to predict future biomedical discoveries. It does this by building a matrix of sentence level co-occurrences between UMLS concepts across the entirety of PubMed's Open Access Subset. This includes co-occurrences of almost 300,000 concepts across 1.2 million full text articles and over 24 million abstracts. Synverita produces predictions using singular value decomposition [16] on its mined matrix of sentence level concept co-occurrences.

Synverita may provide a highly sensitive measure of how well a variant gene fits a disease phenotype, given its use of predictions and large corpus. Its large vocabulary provides a means to map clinician free text phenotype to UMLS concepts. As Synverita's text mining results are precomputed and it allows for backend access, it can produce scores between phenotype UMLS concepts and variant genes efficiently.

The research presented here aims to augment variant prioritization by Synverita. Its raw data and predictions are used to prioritize variants for pediatric patients of IDD due to IEM. This targets the workflow outlined in (Figure: 1.1). (Figure: 1.2)shows the updated workflow that includes text mining based variant prioritization. The performance at variant prioritization is compared with FindZebra, RVIS and CADD as well as with an ensemble method of all methods.

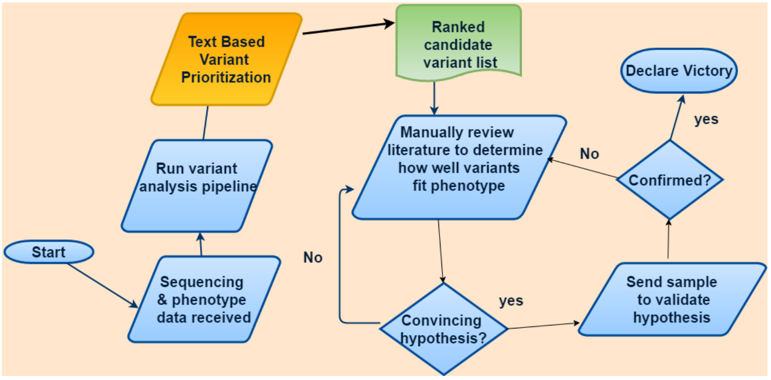
Text Based Variant Prioritization
Ranked candidate variant list
Declare Victory
yes
No
Confirmed?
Manually review literature to determine how well variants fit phenotype
Run variant analysis pipeline
Start
Sequencing & phenotype data received
No
yes
Convincing hypothesis?
Send sample to validate hypothesis

Figure 1.2: Proposed modification to existing workflow. Text based variant prioritization produces a ranked candidate variant list aims to reduce number of passes through the two cycles.

Chapter 2

Methods

Given a set of phenotype terms describing a patient, we seek to return a ranking of genes containing rare variations where genes ranked highly are more likely to be causal for the observed phenotype (or a subset of the phenotype terms). The returned ranking is based on a strength of association score between each gene and the entire set of phenotype terms. Both phenotype terms and gene symbols are represented using UMLS concept unique identifiers (CUIDs), which are described below.

In broad strokes, strength of association scores are generated between a set of gene symbols $G = \{g_1, g_2, .., g_m\}$ and a set of disease phenotype terms $D = \{P_1, P_2, ...P_r\}$ in the following steps:

1. Map each $g \in G$ to a CUID
2. Map each $P \in D$ to a set of all relevant CUIDs, $P = \{p_1, p_2, ..., p_n\}$
3. Calculate an association score, $s_{g,P}$, between each $g \in G$ and each $P \in D$ as follows:
    (a) Calculate an association score, s_{g,p_i}, between g and each $p_i \in P$
    (b) Aggregate the scores from (a), weighting each s_{g,p_i} according to the weighting scheme
4. Calculate an association score, $s_{g,D}$, between each gene $g \in G$ and each D by summing over the scores from (3)

5. Rank all genes in G according to $s_{g,D}$

It is possible to mix-and-match strategies for phenotype-to-CUIDs mapping, association scoring, and term weighting. These are described in detail below.

2.1 Term Mapping

UMLS CUIDs provide standard labels for synonymous noun phrases that are found in the biomedical literature as well as a relational structure based on semantic categories. An example of this is the CUID C0085997 which is defined as

> C0085997 T048 child development dis specific—child development disorders, specific—developmental delay dis—developmental delay disorder—developmental delay disorders.

In this example, the CUID refers to five different pipe-delimited synonyms that belong to category T048, which contains concepts related to mental or behavioural dysfunction.

Gene symbols are each mapped to a single UMLS CUID in the genes category in two passes through Synverita's wordlist. In the first pass, each gene symbol is matched to the CUID that contains the gene symbol followed by the string ' gene' in its synonyms list. If no matches are found in the first pass for a gene symbol, this is attempted again without the ' gene' string. This is done to resolve cases in which a gene symbol maps to multiple CUIDs representing homologs or alleles.

Phenotype terms do not always correspond one-to-one to entries in Synverita's word list. In the event that no exact match for a phenotype term is found in Synverita's wordlist, a semi-automated approach is used to identify all potentially relevant terms. An example of this is the oft-reported ”global developmental delay,” which is not within Synverita's wordlist. In this case, ”developmental delay” is used instead, and all CUIDs with ”developmental delay” in their synonym list are used. Management of multiple CUIDs per phenotype term is reflected in the scoring methods.

2.2 Gene-Phenotype Association Scoring

Synverita provides two broad approaches to generating scores of association between two CUIDs. The first is via its raw data matrix. Synverita's raw data matrix captures the number of sentences in which pairs of CUIDs both appear in the biomedical literature. This matrix is large and sparse. That is, it contains many 0 values for pairs of CUIDs. Synverita's raw data provides a quantitative measure of how strongly two CUIDs associate based on the current state of the literature. However, this scoring makes no assumptions with respect to the directionality of association. That is, the following two hypothetical sentences would both be considered an association between $Gene1$ and $SymptomA$:

> "Mutations in Gene 1 have been found to cause Symptom A."
>
> "Mutations in Gene 1 do not cause Symptom A."

The second approach uses Synverita's predictions to score associations between pairs of CUIDs. Performing singular value decomposition on Synverita's raw data matrix infers many of the missing values based on the values of similar terms. The idea is that if there is a strong association between $Gene1$ and $SymptomA$ and between $Gene1$ and $Gene2$, Synverita would predict an association between $Gene2$ and $SymptomA$.

2.3 Raw Data Association Strength

The strength of association between two CUIDs, t_1 and t_2, in Synverita's raw data is calculated by using the Jaccard similarity coefficient, js. The value of js is calculated using the number of sentences in which t_1 and t_2 appear together ($S_{t_1 \cap t_2}$) and the total number of sentences that each appears in (S_{t_1} and S_{t_2}). This is similar to how association strength is calculated in Beegle [15] but is based on sentence-level co-occurrence rather than abstract-level co-occurrence. The value of js can range from 0 when t_1 and t_2 never occur in the same sentence to 1 when they always occur together. js is defined as

$$js(t_1,t_2) = \frac{S_{t_1 \cap t_2}}{S_{t_1} + S_{t_2} - S_{t_1 \cap t_2}}.$$

2.4 Predicted Association Strength

Predicted association strength between two terms, ps, is calculated using the results of singular value decomposition of the binarized form of Synverita's data matrix. That is, the non-zero values in Synverita's data matrix are replaced with 1 to create a binary matrix that contains values of only 0 and 1. Singular value decomposition is performed using the graphlab implementation with best parameters found in Lever et al. (2016).

2.5 Weighting Methods

Four different weighting methods are used for both raw data and association calculations between gene CUID g and phenotype CUID p. All four weighting methods are based on the total number of unique terms that g and p co-occur with, A_g and A_p respectively. The four weighting methods, w_g, w_p, w_{g+p} and w_{g*p}, are defined as follows:

$$w_g = log_{10}(A_g + 1)$$

$$w_p = log_{10}(A_p + 1)$$

$$w_{g+p} = log_{10}(A_g + A_p + 1)$$

$$w_{g*p} = log_{10}(A_g * A_p + 1)$$

Weighting is applied to both predicted and raw data association scores by division. Given the generalized association score $s \in \{js, ps\}$ and weight $w \in \{w_g, w_p, w_{g+p}, w_{g*p}\}$, the generalized weighted score, ws is defined as

$$ws = \frac{s}{w}.$$

2.6 Score Aggregation

Four different methods are used for scoring association strength between a gene term g and a phenotype term P with CUIDs $\{p_1, p_2, ..., p_n\}$. Average scoring between g and P averages $s(g, p_i)$ across all p_i in P and is defined as

$$avg_s(g,P) = \frac{\sum_{i=1}^{n} s(g,p_i)}{n}.$$

Forgiving scoring averages only those values of $s(g, p_i)$ that have a non-zero value and is defined as

$$for_s(g,P) = \frac{\sum_{i=1}^{n} s(g,p_i)}{max(1, \sum_{i=1}^{n} \begin{cases} 1 & \text{if } s(g,p_i) > 0 \\ 0 & \text{otherwise} \end{cases})}.$$

Best scoring selects the highest single value of $s(g, p_i)$ and is defined as

$$best_s(g,P) = \max_{1 \leq i \leq n} s(g,p_i).$$

The fourth scoring method, representative scoring, requires the definition of a representative CUID of phenotype term P, p_r. The representative CUID of a phenotype term is the CUID that is most closely associated to the other phenotype terms in its case, based on Synverita's raw data. For case C with phenotype terms $\{P_1, .., P_m\}$, the representative CUID of P_j, $p_r j$, is defined as

$$p_{rj} = \underset{p_{ij} \in P_j}{argmax} \sum_{k=1}^{m} \begin{cases} 0 & \text{if } k = j \\ avg_j(p_{ij}, P_k) & \text{otherwise} \end{cases}.$$

With a phenotype term's representative CUID defined, the representative scoring for gene g and phenotype term P with representative CUID p_r is defined as

$$rep(g,P) = s(g,p_r).$$

Synverita's prediction values do not produce values of zero, which means that forgiving scoring degenerates to average scoring. This results in a total of seven methods for scoring the strength of association between g and P: four scoring methods for raw data $\{avg_j, for_j, best_j, rep_j\}$ and three scoring methods for predictions $\{avg_p, best_p, rep_p\}$. Given case C, the total strength of association between gene g and C is calculated for scoring method $meth_s \in \{avg_j, for_j, best_j, rep_j, avg_p, best_p, rep_p\}$

as

$$\sum_{j=1}^{m} meth_s(g, P_j).$$

2.7 Score Ranking

For each set of gene symbols and phenotype terms, scores were ranked for each gene. For a given gene g_x with score s_x, in a case with genes $G = \{g_1, ..., g_m\}$ and vector of corresponding scores $S = \{s_1, ..., s_m\}$, the ranked score rs_x of g_x is the proportion of scores that s_x is greater than in S. This is calculated as

$$rs_x = \frac{\sum_{i=1}^{m} \begin{cases} 1 & \text{if } s_x > s_i \\ 0 & \text{otherwise} \end{cases}}{m-1}$$

2.8 Comparison to Other Methods

Three established methods for variant prioritization were compared to a subset of the methods introduced above. Two of these, RVIS and CADD, are currently in use in the standard TIDE BC workflow. FindZebra is not currently used in the standard TIDE BC worfklow, but it can be used to score variants with respect to a rare disease phenotype. The following section describes how these scores were obtained.

2.8.1 RVIS Scoring

The pipeline described in Tarailo-Graovac et. al [41] is used to generate RVIS version 2 scores. If a gene does not have an RVIS value, 1 is used instead, which is the lowest possible RVIS value. This value is then subtracted from 1 to produce a non-inverted scoring.

2.8.2 CADD Scoring

The pipeline described in Tarailo-Graovac et. al [41] is used to generate CADD scores. If a gene has multiple CADD scores, the maximum value across all of its variants in a given case is used. If a gene has no CADD scores, 0 is used instead, which is the lowest possible CADD score.

2.8.3 FindZebra Scoring

FindZebra scores are calculated by initializing every gene in a case to a score of 0. FindZebra's API is then queried with each of a case's phenotype terms. Each time a gene appears in a query's associated gene list, the association score is added to the gene's score.

2.8.4 Score Comparison

The above three methods are used to create a ranked score as described above. These ranked scores are then compared with best performing raw data and prediction methods.

2.8.5 Ensembling

Methods are ensembled by training random forests on multiple methods using the Caret R package [17]. Trained random forests are tuned using 20 repeats of 10-fold crossvalidation with post-sampling downsampling to control for class imbalance. Performance is then optimized on receiver operating characteristic.

Chapter 3

Results

This chapter presents the details of the tests performed using the preceding methods and their results.

3.1 Simulation

As Synverita is a general purpose biomedical text mining tool (as opposed to a rare disease focused resource), the first test was to compare the contents of its raw data matrix to FindZebra. FindZebra is a rare disease specific text mining project, so it provides a benchmark for Synverita's domain applicability. Because FindZebra is able to produce relevant disease CUIDs given a gene symbol, term selection and scoring methods for multiple CUIDs per phenotype term were not used. Similarly, prediction scoring was not used as FindZebra does not produce predictive scores.

A set of 922 known mendelian disease gene symbols from Tarailo-Graovac et al. (submitted) were queried using FindZebra's API (on September 12, 2016). The CUIDs of diseases for each of these genes were collected in order to create a disease profile for each. Querying FindZebra with the 922 known Mendelian gene symbols resulted in 922 disease profiles. The 922 disease profiles contained between zero and ten CUIDs with a median of nine CUIDs (Figure: 3.1). Five disease profiles contained no CUIDs and were excluded from further analysis.

Synverita's raw association matrix was used to create a profile of candidate genes for each disease profile by selecting CUIDs from the UMLS genes semantic category, G that had a non-zero avg_j association score with that disease profile. Given a gene symbol g_x and its corresponding disease profile d_x, its corresponding candidate gene profile gp_x is defined as

$$gp_x = \{g \in G | avg_j(g, d_x) > 0 \vee g = g_x\}.$$

The 917 disease profiles that contained at least one CUID produced 909 candidate gene profiles with at least one CUID based on Synverita's raw data matrix. The candidate gene profiles ranged from a minimum of one CUID to a maximum of 13480 CUIDs with a median of 1981 CUIDs (Figure: 3.2). The distribution of gene profile size is presented in Figure 4.

For each gene in a candidate gene profile, avg_j association score was calculated with and without w_p weighting between it and its corresponding disease profile. This produced two vectors of association scores between each disease profile and each gene in its candidate gene profile. These vectors were then ranked as outlined above. Weighted ranked scores performed similarly to ranked scores without weighting (Figure : 3.3). Weighting produced a median ranked score of .9972 whereas the median ranked scores without weighting was .9970. No meaningful correlation was found between either set of ranked scores and size of gene profile or disease profile. Both methods ranked the target gene in the top 10% of its gene profile over 90% of the time. These results suggest that Synverita's raw data matrix is consistent with FindZebra with respect to associating rare disease genes to their established set of disorders.

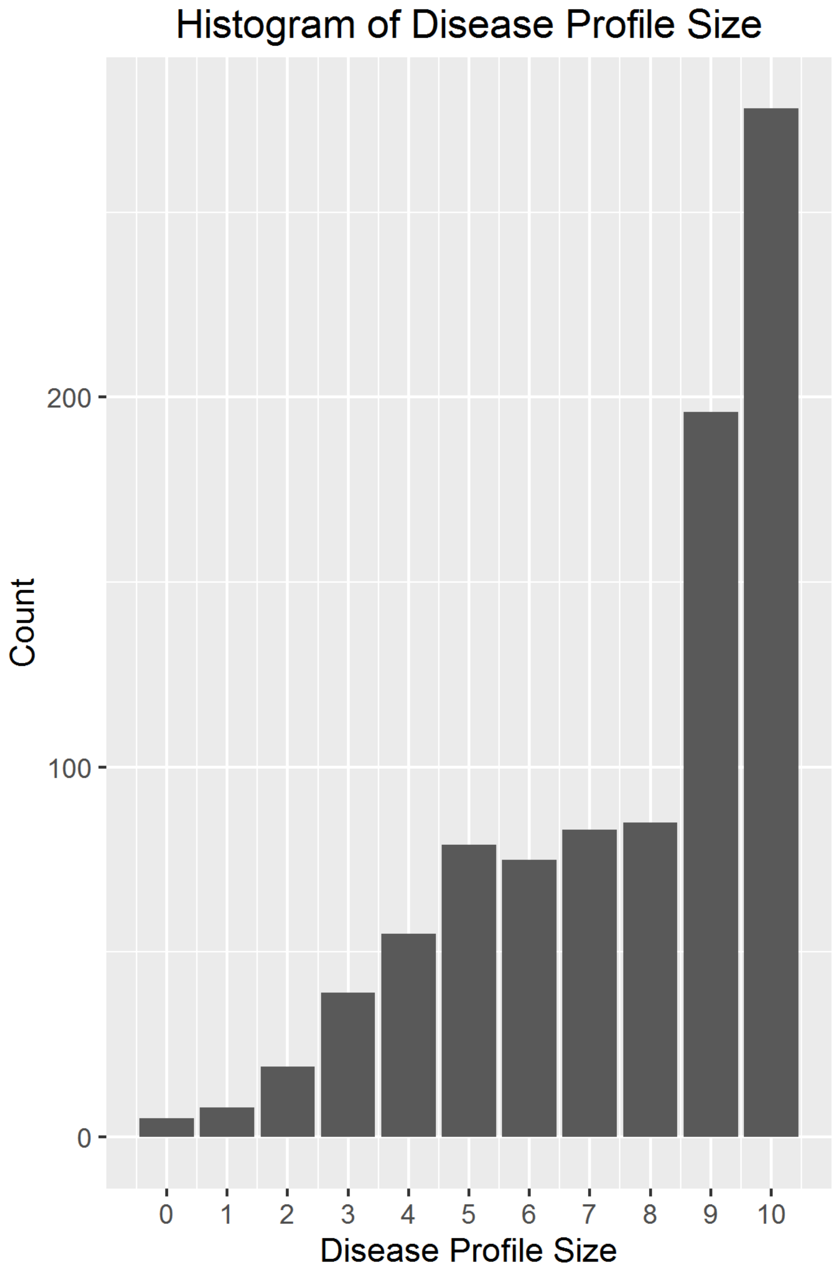

Figure 3.1: Histogram of disease profile size as represented by number of CUIDs. Median of 9 CUIDs per disease profile.

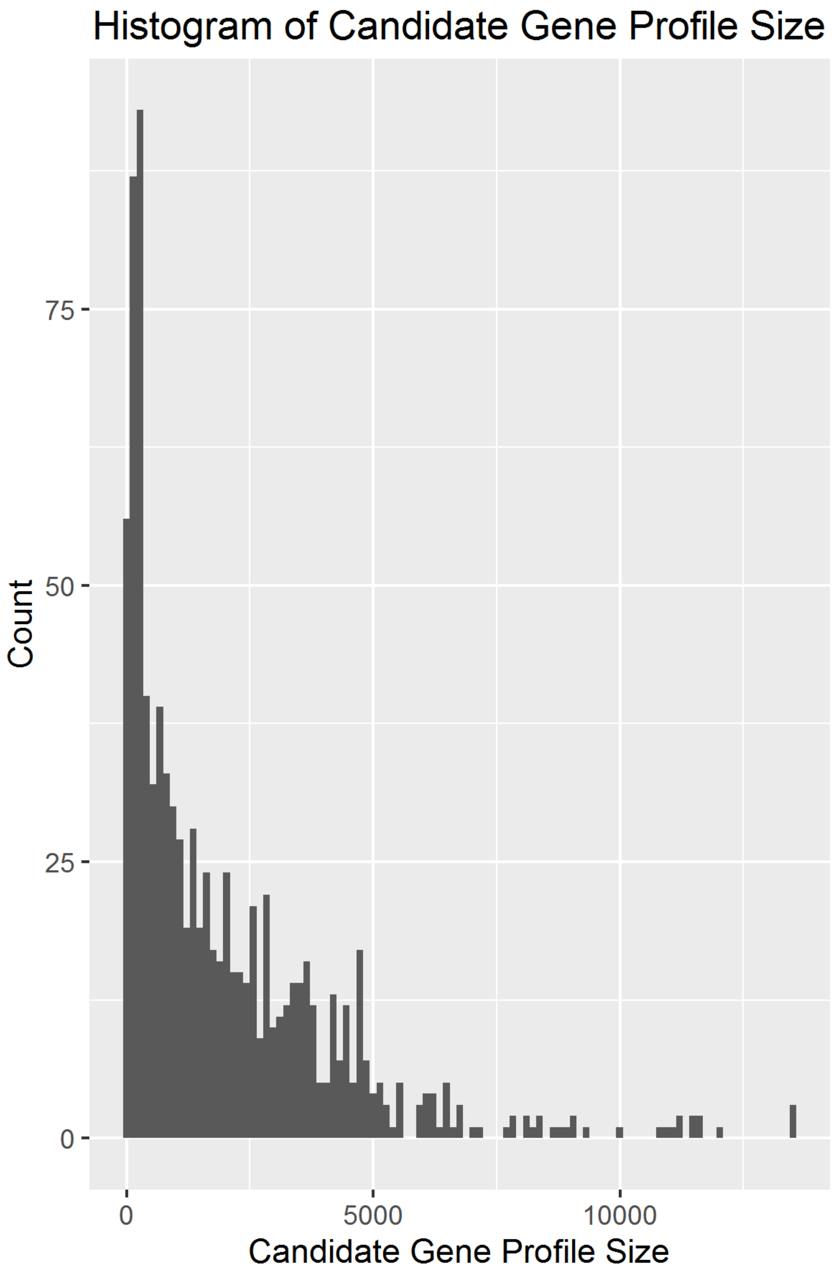

Figure 3.2: Histogram of candidate gene profile size as represented by number of CUIDs with non-zero avg_j scores. Median of 1981 CUIDs per disease profile.

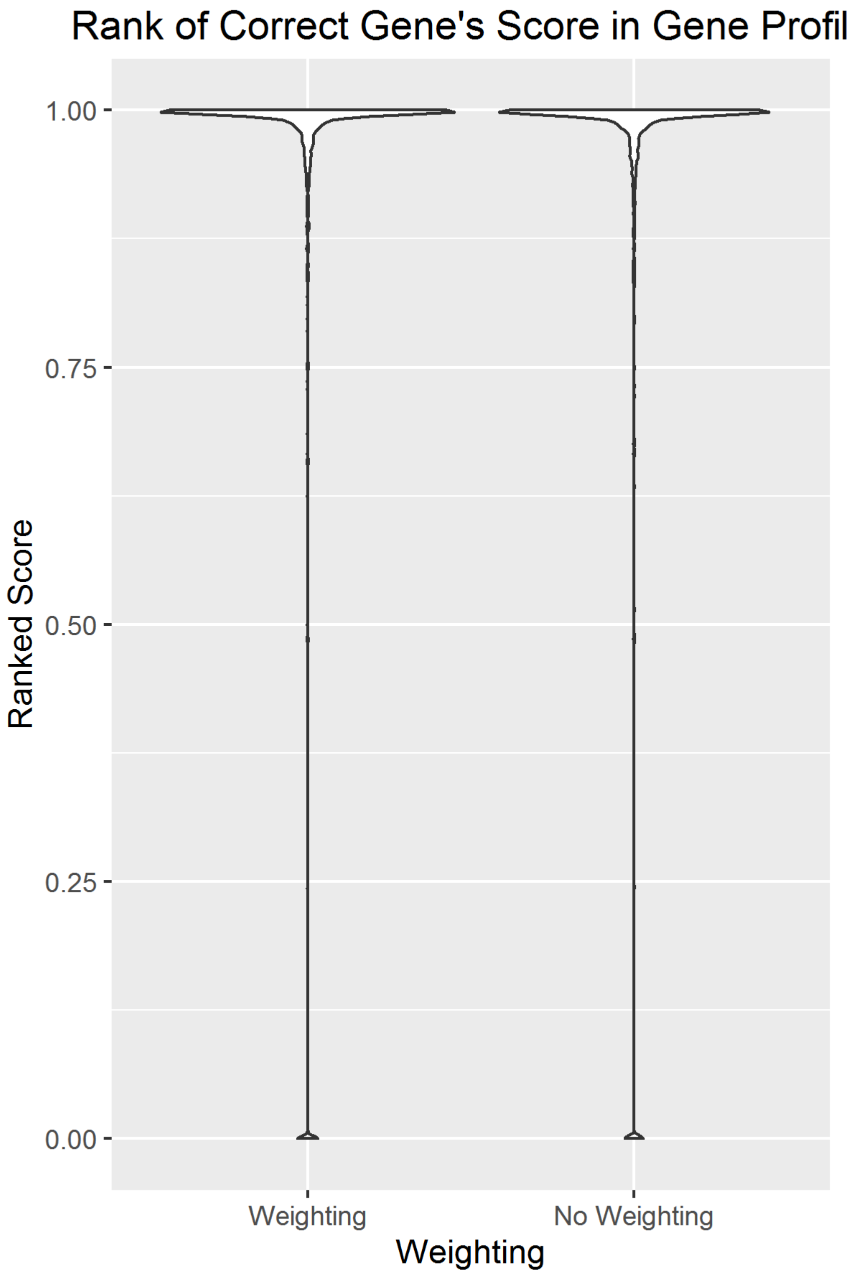

Figure 3.3: Distribution of disease phenotype association rank scores of candidate genes in their candidate gene profile with and without w_p weighting. Performance is almost identical between the two approaches.

3.2 Application of text ranking to clinical case examples

Two groups of TIDE BC cases were run through the current analysis pipeline. These two groups had established best candidate variants and diagnoses. The first set, the training set, was used to develop Synverita based variant prioritization methods. The second set, the test set, was subsequently analyzed using the methods developed on the training set in order to establish the generalizability of these methods. For both groups, pipeline results and clinician-reported phenotype were used to produce raw data and predicted association scores, as well as CADD, RVIS and FindZebra scores as outlined above. All association, weighting and ensembling methods were used on the training set. The best performing methods were then used on the test and discovery sets.

3.2.1 Training Set

Data

The training set was composed of candidate variant lists and clinical reports from 41 individuals from 38 TIDE families with established best candidate variants recently analysed in Tarailo-Graovac et al. [41]. Three pairs of two siblings shared common best candidate variants. Five of the individuals had digenic etiology while the remaining 36 had monogenic etiology. The automatically generated candidate variant lists of four individuals, one with digenic etiology, did not contain the established candidate variants (indicating that manual steps such as reducing expected allele frequency had been performed to determine the candidates) and were excluded from further analysis. A further two individuals with digenic etiology were missing one of the two established best candidate variants (again due to customized minor-allele frequency settings), which resulted in these variants being excluded.

The remaining 37 individuals had between 30 and 352 variant genes in the candidate variant lists with a median of 49 variant genes. The total number of variants ranged between 38 and 440 with a median of 62. Clinician reports for these patients contained a median of 12 phenotype terms and ranged between 5 and 35 terms. Clinician reported phenotype terms mapped to a median of 343 CUIDs per case with a minimum of 42 and maximum of 2303.

Unweighted Ranking

Ranked scores for each score aggregation method outlined in 2.6 were first computed without weighting in order to establish a performance baseline. Success of a method was evaluated by its median and minimum ranking of a best candidate variant across all cases.

Raw Data All four raw data aggregation methods shared the same minimum and maximum performance. Best scoring produced the highest median results, ranking the established best candidate variant gene ahead of over 92% of the other variant genes in half of the training cases (Figure:3.4). All methods produced a number of 0 rankings for several best candidate variants. This occurs when there are no co-occurrences between a gene and any of the phenotype CUIDs in Synverita's data matrix. Representative aggregation, rep_j, is most prone to rank a best candidate variant 0, as it only uses one CUID per phenotype term (Figure : 3.5). However, rep_j also has the largest number of cases in which the best candidate is ranked ahead of all other variants. Comparison of best candidate variant ranking to rankings of all other variants produced significant Kolmogorov-Smirnoff (as implemented in R [32]) test p-values for all four methods (Table : 3.1).

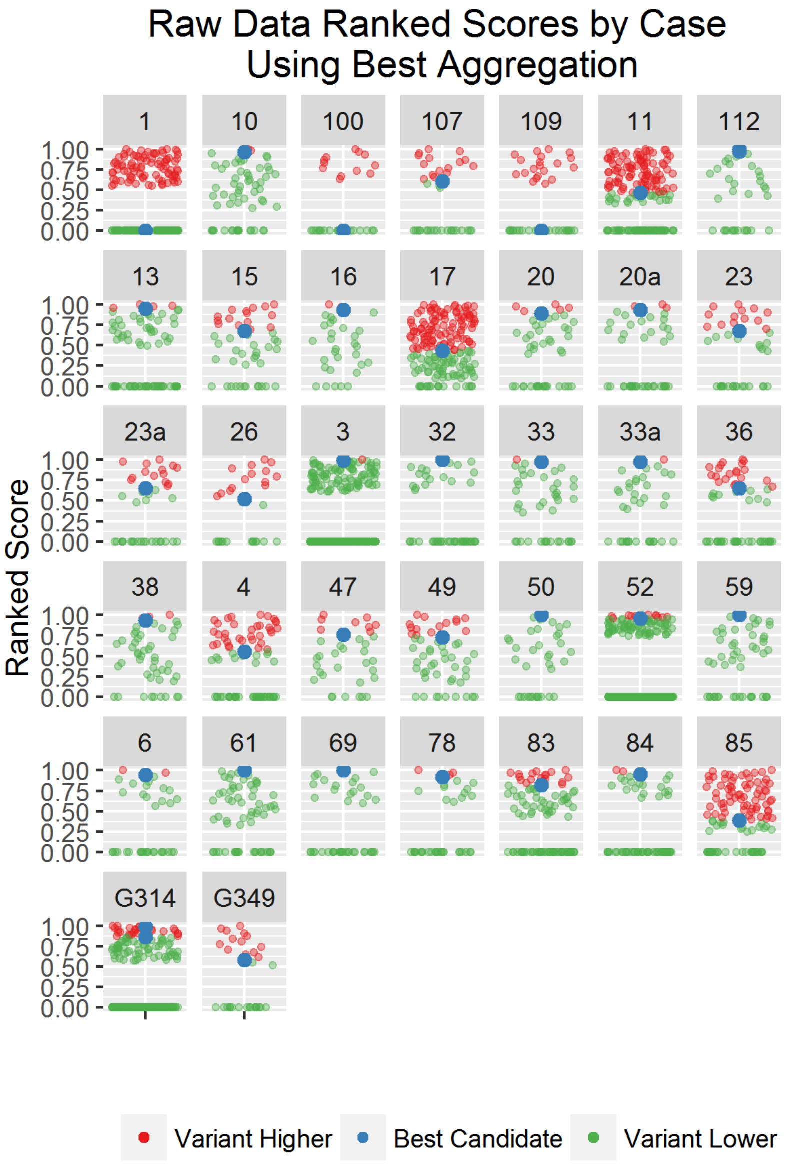

Figure 3.4: Ranked scores of all variant genes in each case in training set using $best_j$ aggregation. Variant higher and variant lower are variants ranked higher or lower than their case's lowest ranked best candidate variant.

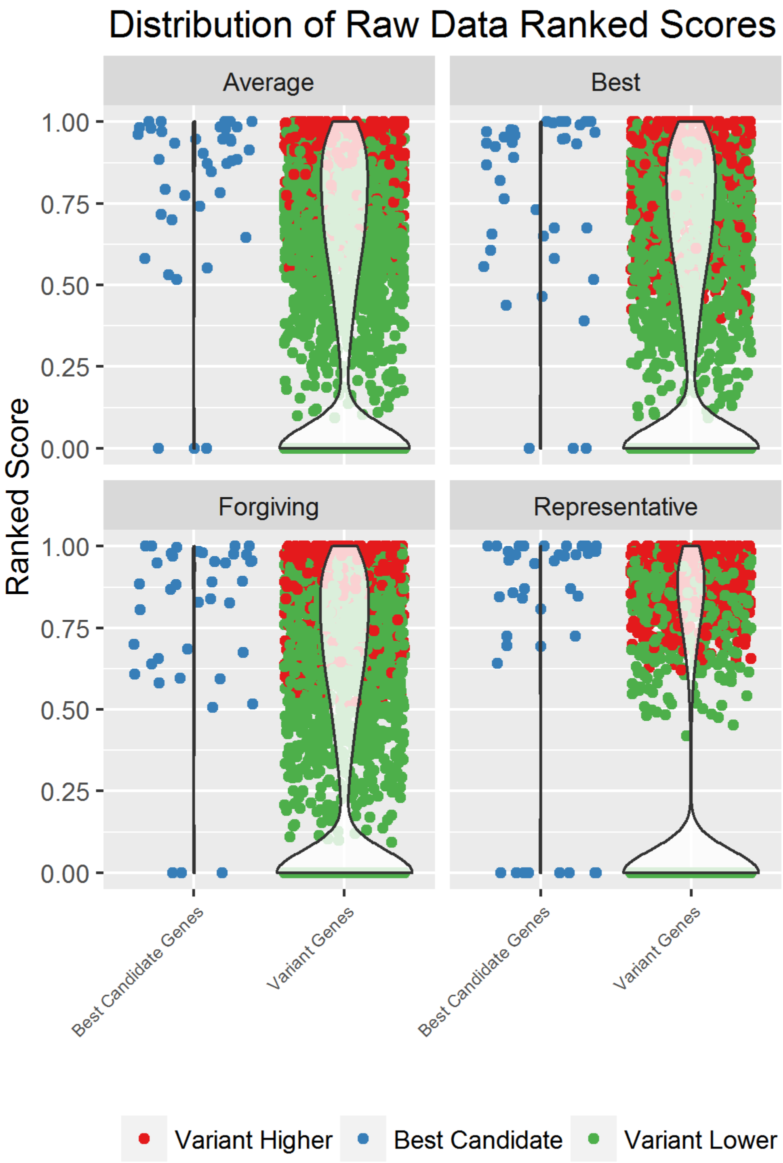

Figure 3.5: Distribution of ranked scores for raw data aggregation methods across all 37 individuals in the training set.

Prediction The prediction based aggregation methods had worse median performance and higher minimum performance than their respective raw data counterparts (Table : 3.1). Representative aggregation, rep_p, had the best median performance (Figure : 3.7). Notably, representative scoring always places a case's best candidate variant ahead of at least 36.7% of the other variants in the case. This indicates that inclusion of the approach in a pipeline could reduce the burden on expert reviewers. Comparison of best candidate variant ranking to rankings of all other variants produced significant Kolmogorov-Smirnoff test p-values for all three methods (Table 3.1).

	Method	Median	Mean	Min	KS p-value
Raw Data	avg_j	0.89	0.80	0.00	$6.24e^{-09}$
	$best_j$	0.92	0.76	0.00	$7.85e^{-07}$
	for_j	0.88	0.77	0.00	$1.98e^{-08}$
	rep_j	0.87	0.72	0.00	$3.87e^{-10}$
Prediction	avg_p	0.80	0.74	0.03	$1.24e^{-05}$
	$best_p$	0.78	0.74	0.29	$3.37e^{-06}$
	rep_p	0.83	0.77	0.37	$3.25e^{-06}$

Table 3.1: Unweighted ranked scoring results.

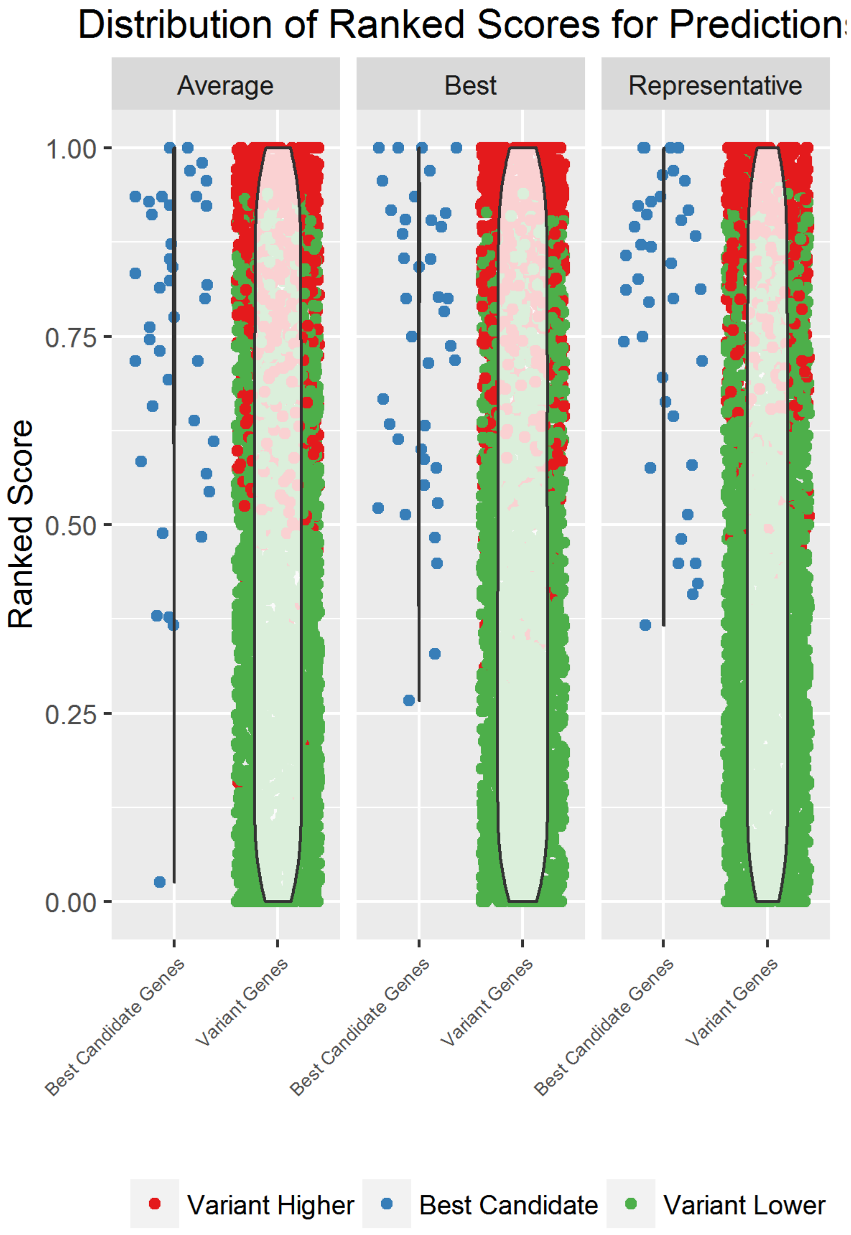

Figure 3.6: Distribution of ranked scores for prediction aggregation methods across all 37 individuals in the training set.

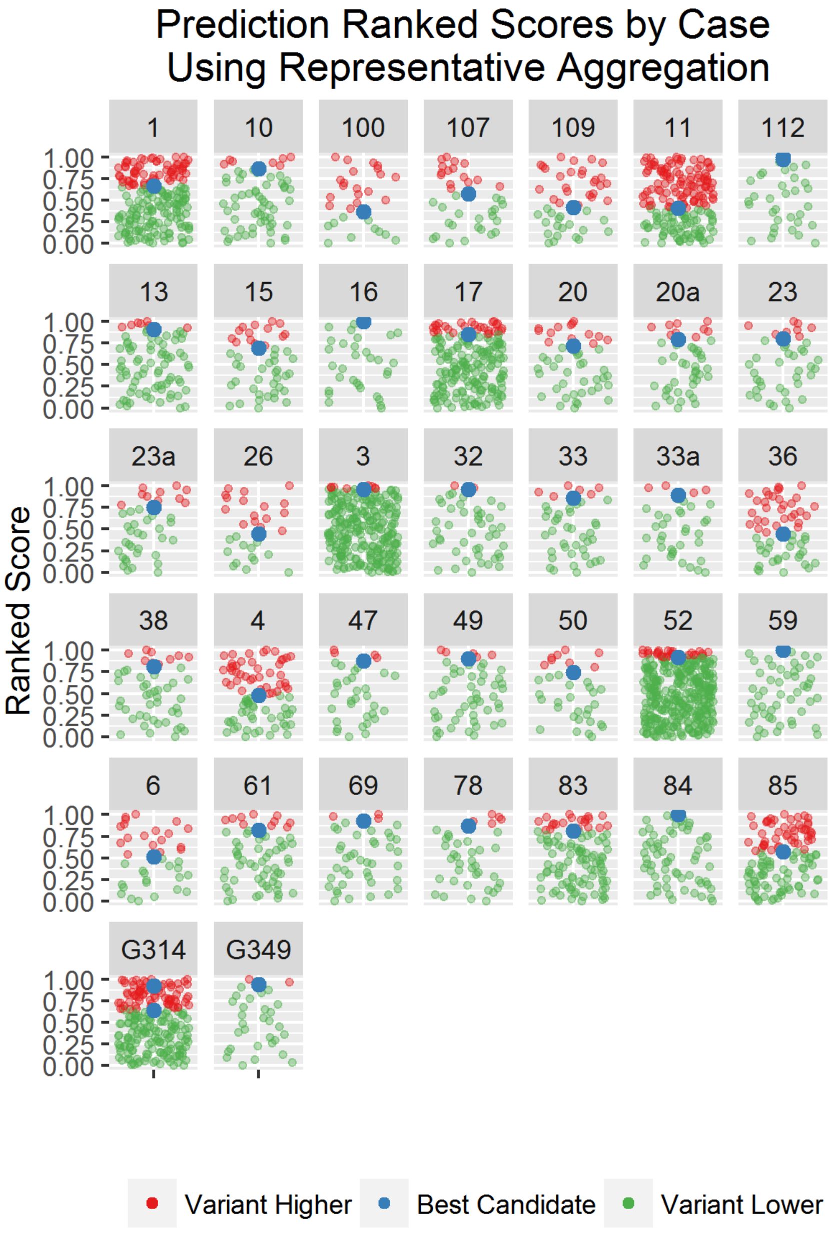

Figure 3.7: Ranked scores of all variant genes in each case in training set using rep_p aggregation.

Unweighted Raw Data vs. Prediction Ranking The rankings of each raw data aggregation method was compared to its corresponding prediction method. Raw data methods using best and average aggregation ranked the best candidate variant higher than the prediction cases more often than not (Figure : 3.8). Representative selection was divided more evenly, with the raw data ranking the best candidate higher 49% of the time, predictions ranking the best candidate variant higher 41% of the time, and producing equal ranking 10% of the time.

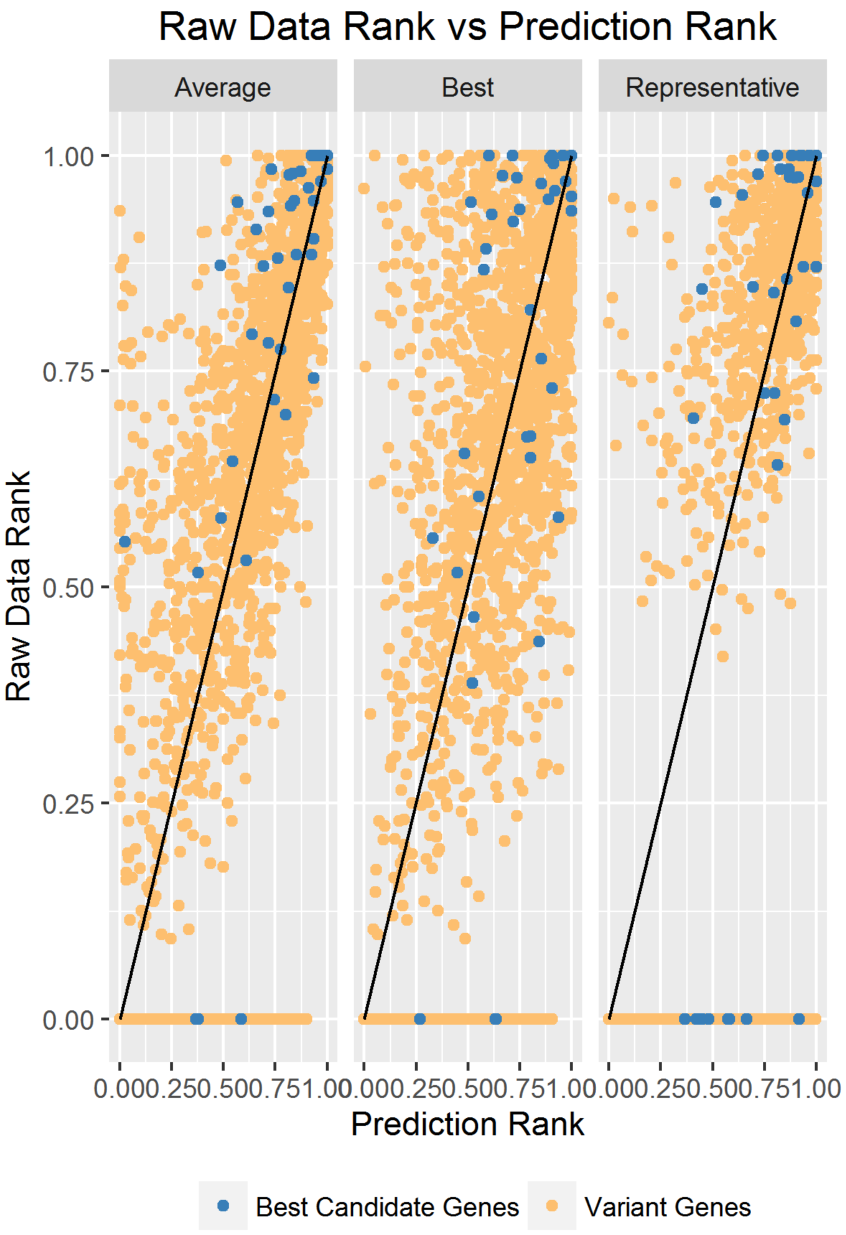

Figure 3.8: Ranked scores of each of the prediction aggregation methods compared to their respective raw data aggregation methods. Points above the black line rank higher using raw data. Those below the black line rank higher using prediction.

Weighted Ranking

In order to quantify the effects of each weighting method on each aggregation method, all combinations of each aggregation method and each weighting method were used to rank candidate variants in the training set. Additionally, Spearman's ρ was calculated between each variant gene g and the number of unique terms it co-occurrs with in Synverita's raw data matrix(A_g in the previous chapter). That is, the correlation was calculated for all genes, not just best candidate variants. This was done in order to determine if weighting is able to moderate the effects of A_g on ranked score.

Of all of the weighting methods, w_{g*p} is notable as it produces the highest median performance of 0.95 with rep_j and the lowest Spearman's ρ value of 0.566 (see Table A.1 for full results). Similarly, it produces the highest median performance of all prediction methods with rep_p and the lowest correlation. Based on the training data, w_{g*p} produces the best median ranking of best candidate variants and the ranked scores it produces are least dependent on how well studied a gene is. Accordingly, w_{g*p} weighted rep_j and rep_P are used in all following association calculations.

Comparison To Other Methods

The rankings of the best performing prediction and raw data aggregation methods, rep_p and rep_j were compared to rankings produced by CADD, RVIS and FindZebra. CADD and rep_j had the best performance for 17 variant genes, RVIS for nine genes, rep_p for seven genes and FindZebra for seven genes. Compared to these three existing methods, Synverita based methods perform well. rep_p had the highest minimum performance and rep_j tied CADD in the number of cases that it performed the best.

Rankings for each of these methods show low correlation to each other (see Table 3.2) The highest correlation is between the two Synverita based methods, rep_j and rep_p with a value of .55. In all, the degree of correlation is limited, which supports the idea that they are measuring underlying properties.

Method	RVIS	FindZebra	rep_p	rep_j
CADD	0.010	0.172	−0.191	−0.190
RVIS	1.000	−0.013	0.250	0.122
FindZebra	−0.013	1.000	0.118	0.156
rep_p	0.250	0.118	1.000	0.546

Table 3.2: Spearman correlation of ranked scores of best candidate variants across 5 metrics.

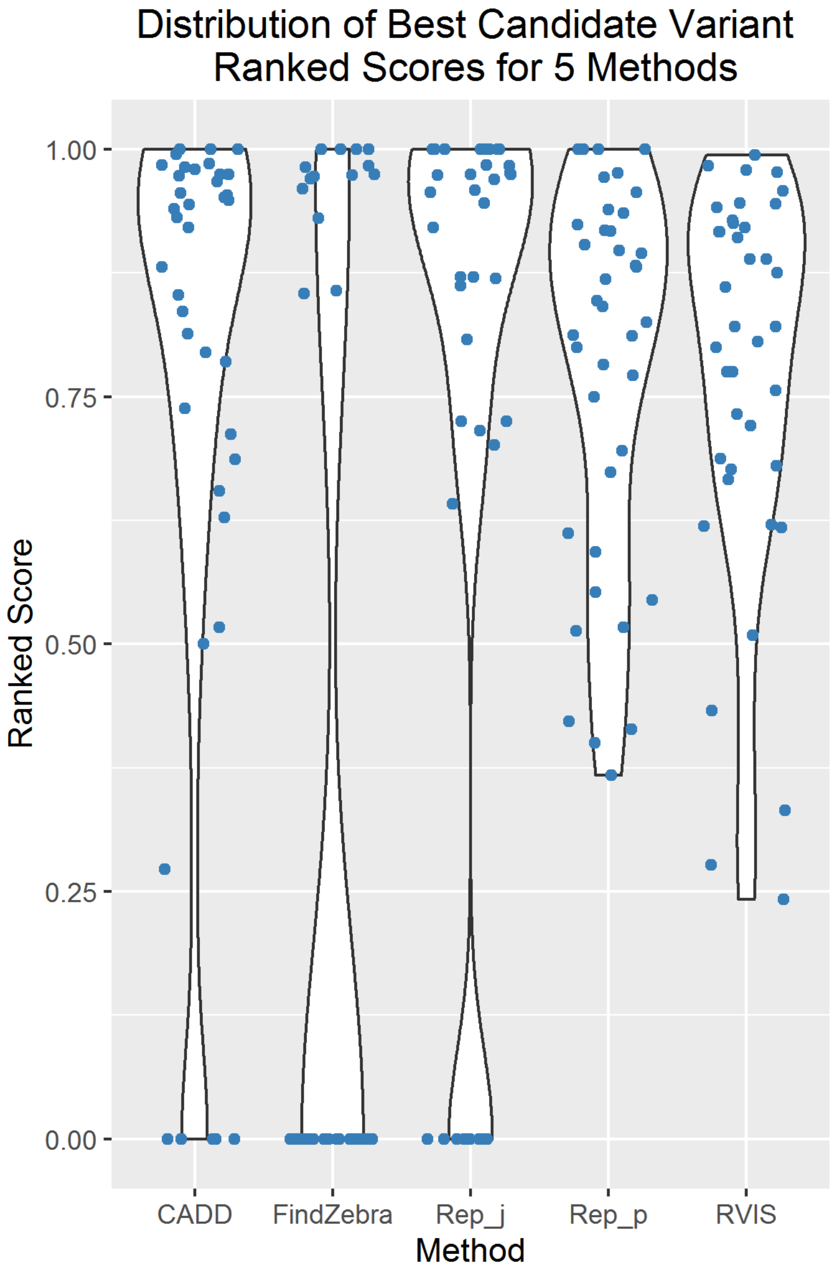

Figure 3.9: Distribution ranked scores of best candidate variants in training set using FindZebra, CADD, RVIS, rep_j and rep_p.

Ensembling

In order to test synergism of these methods, thirty-one random forest classifiers were trained using all possible combinations of the above five metrics. Performance for each classifier was estimated using 10-fold cross-validation as described in section 2.8.5. Unsurprisingly, estimated sensitivity increased with the number of metrics trained on (see Figure: 3.10). The most sensitive classifier was trained using CADD, rep_j and rep_p, followed by the classifier trained on all metrics. All of the most sensitive classifiers used CADD and rep_j (see Figure: 3.11). The most specific classifiers incorporated either rep_j or FindZebra (see Figure: 3.12). Inclusion of FindZebra tended to result in worse sensitivity and was estimated as the least important metric in the random forest cross validation.

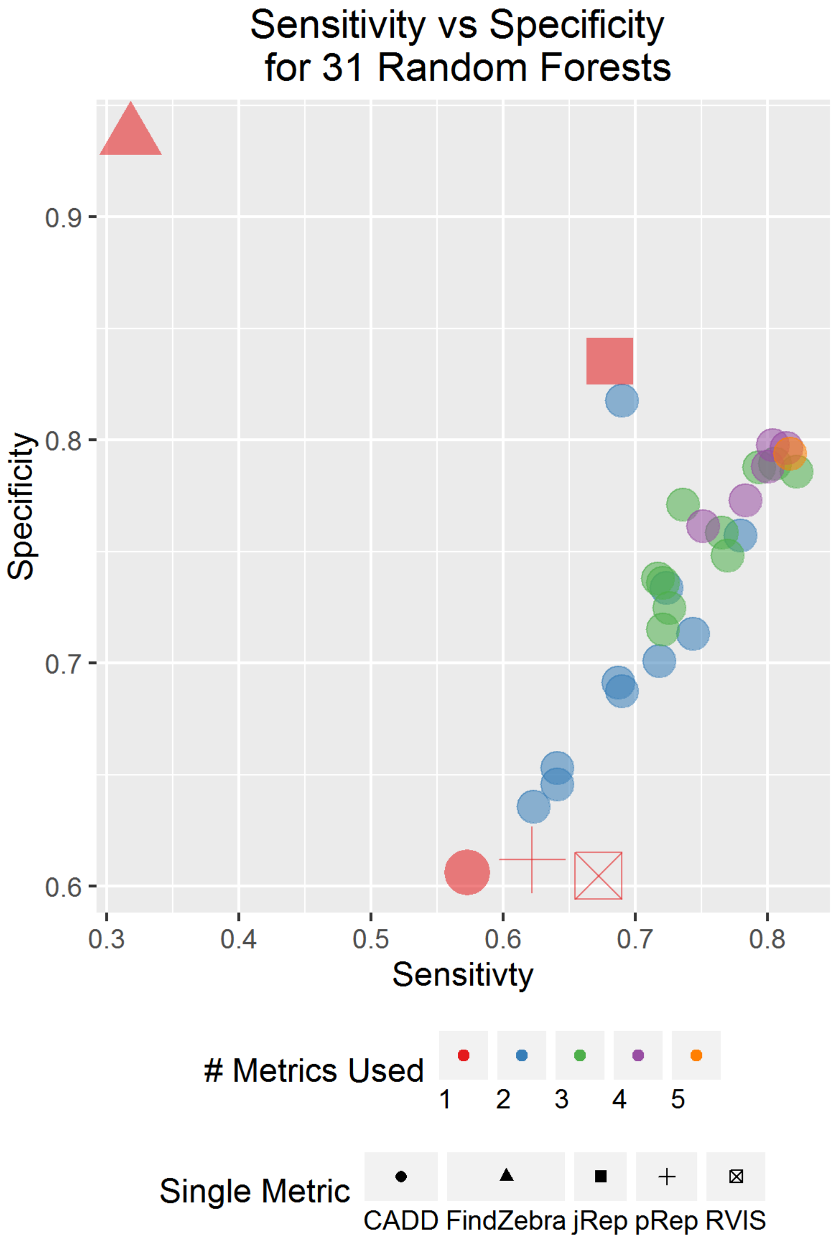

Figure 3.10: Estimated sensitvity and specificity of 31 classifiers trained on all combinations of 5 metrics.

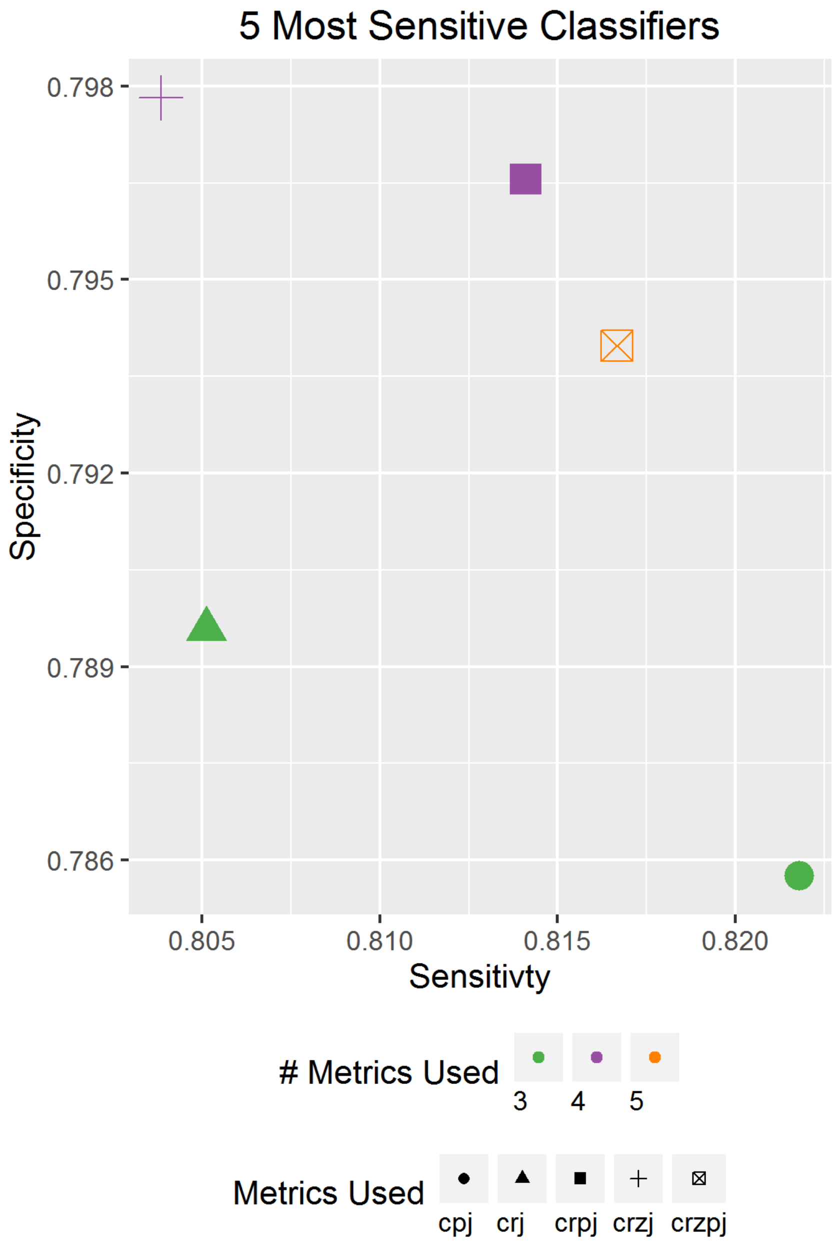

Figure 3.11: Top 5 classifiers based on estimated sensitivity. Shapes denote the metrics used. The symbols used to denote which metrics a classifier used are c for CADD, z for FindZebra, j for rep_j and p for rep_p.

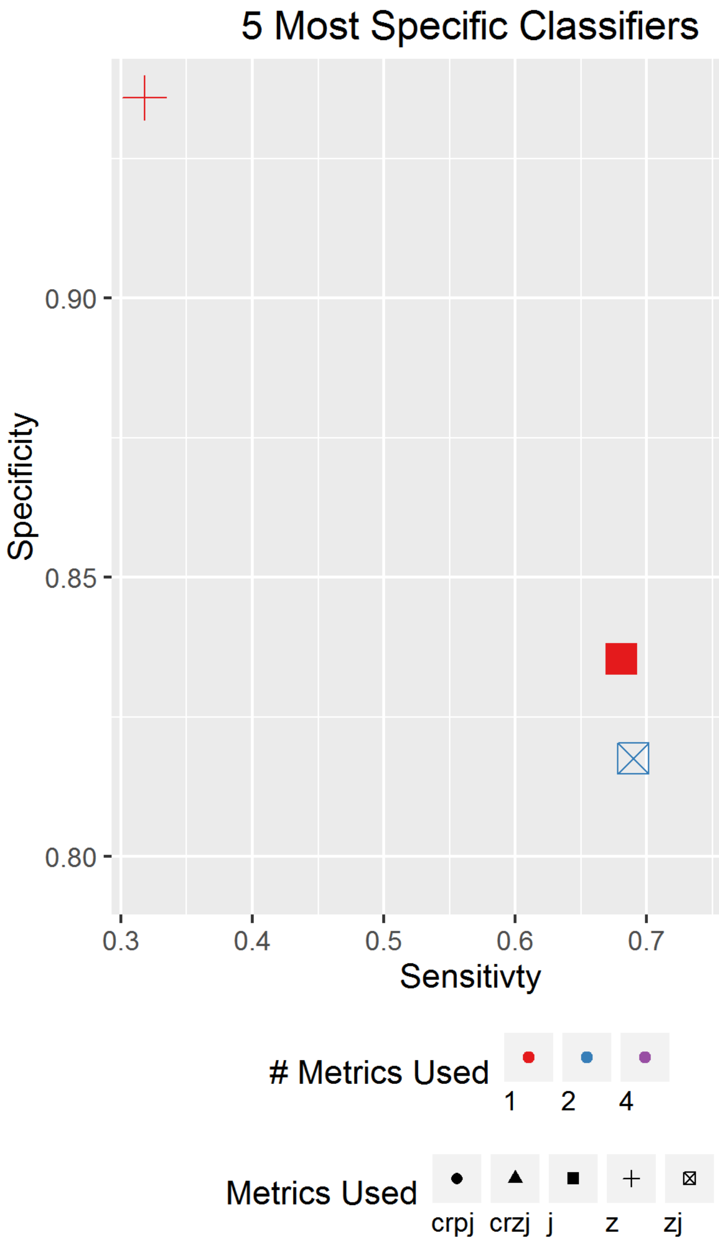

Figure 3.12: Top 5 classifiers based on estimated specificity. Shapes denote the metrics used. The symbols used to denote which metrics a classifier used are c for CADD, z for FindZebra, j for rep_j and p for rep_p

3.2.2 Test Set

The test set was composed of candidate variant lists and clinical reports from 20 individuals from 20 TIDE families. Like the individuals in the training set, the individuals in the test set each had one or more established best candidate variants. 12 individuals had monogenic etiology, seven had digenic etiology and one individual had trigenic etiology. Three individuals were not included in this analysis as their candidate variant lists did not include any of their established best candidate variants (again due to customized minor-allele frequency settings). Two of these had monogenic etiology and one had digenic etiology. A fourth individual's candidate variant list was missing one of its two established best candidate variants. This left a total of 22 established best candidate variants across 17 individuals.

The remaining 17 individuals had between 31 and 482 variant genes in their candidate variant lists with a median of 70 variant genes. The total number of variants ranged between 38 and 697 with a median of 89. Clinician reports contained a median of 13 phenotype terms and ranged between 2 and 26 terms. Clinician reported phenotype terms mapped to a median of 518 CUIDs per case with a minimum of 38 and maximum of 2125.

Method Comparison

In order to test the generalizability of Synverita based variant prioritization, ranks were produced using all 7 Synverita aggregation methods with w_{g*p} weighting. Consistent with the training set, the highest median ranking was achieved by rep_j for the raw data approaches. For the prediction methods avg_p produced a median ranking slightly higher than rep_p with lower minimum and average performance. Of the other methods existing methods, CADD performed the best and FindZebra had the lowest median performance. Interestingly, CADD ranking has the highest minimum of any metric for the test set. This may be due to the fact that cases in the training set were analysed before CADD scores became part of the TIDE BC bioinformatics workflow, whereas the test set cases are more recent. These findings are mostly consistent with the results of the training set (see Figure:3.13).

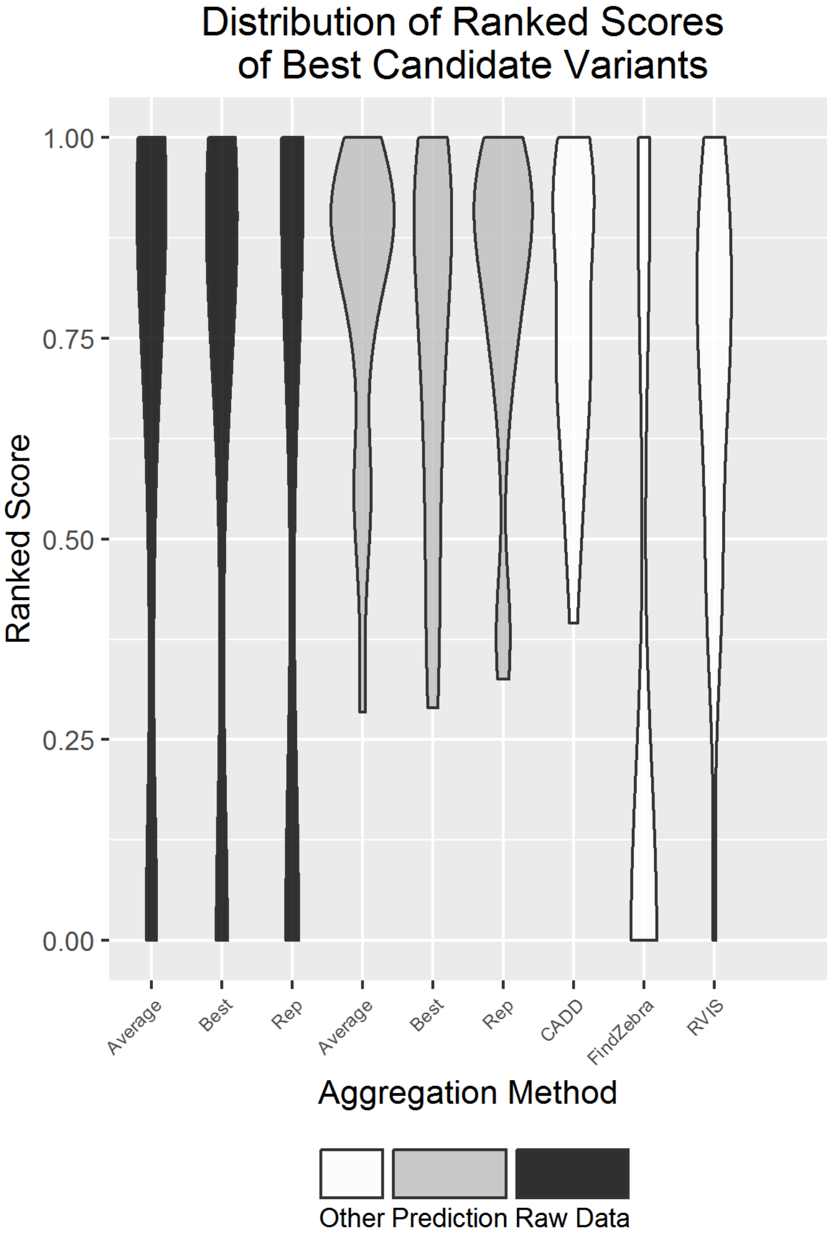

Figure 3.13: Distributions of ranked scores for test set using 9 scoring methods.

Ensembling

To test the generalizability of the classifiers on combinations of metrics in the training set, all test set cases were classified using the 31 random forest classifiers trained in section 3.2.1. The sensitivity of classification did not correlate with number of metrics as strongly as estimated for the training set (see Figure: 3.14). Six classifiers achieved equal sensitivity all of which incorporated rep_p (see Figure: 3.15). The classifier with the highest specificity of these six, $RF_{r,c,p,j}$, incorporated all metrics but FindZebra. $RF_{r,c,p,j}$ was also fifth best with respect to specificity (see Figure: 3.16. $RF_{r,c,p,j}$ ranked probability had better median and mean performance than any individual metric (see Figure: 3.18). $RF_{r,c,p,j}$ ranked best candidate variants in the top 20% of its case 77% of the time (see Figure:3.17. These results suggest that prioritizing variants using Synverita's predictions and raw data is generalizable and its inclusion in variant prioritization is more effective than RVIS and CADD alone.

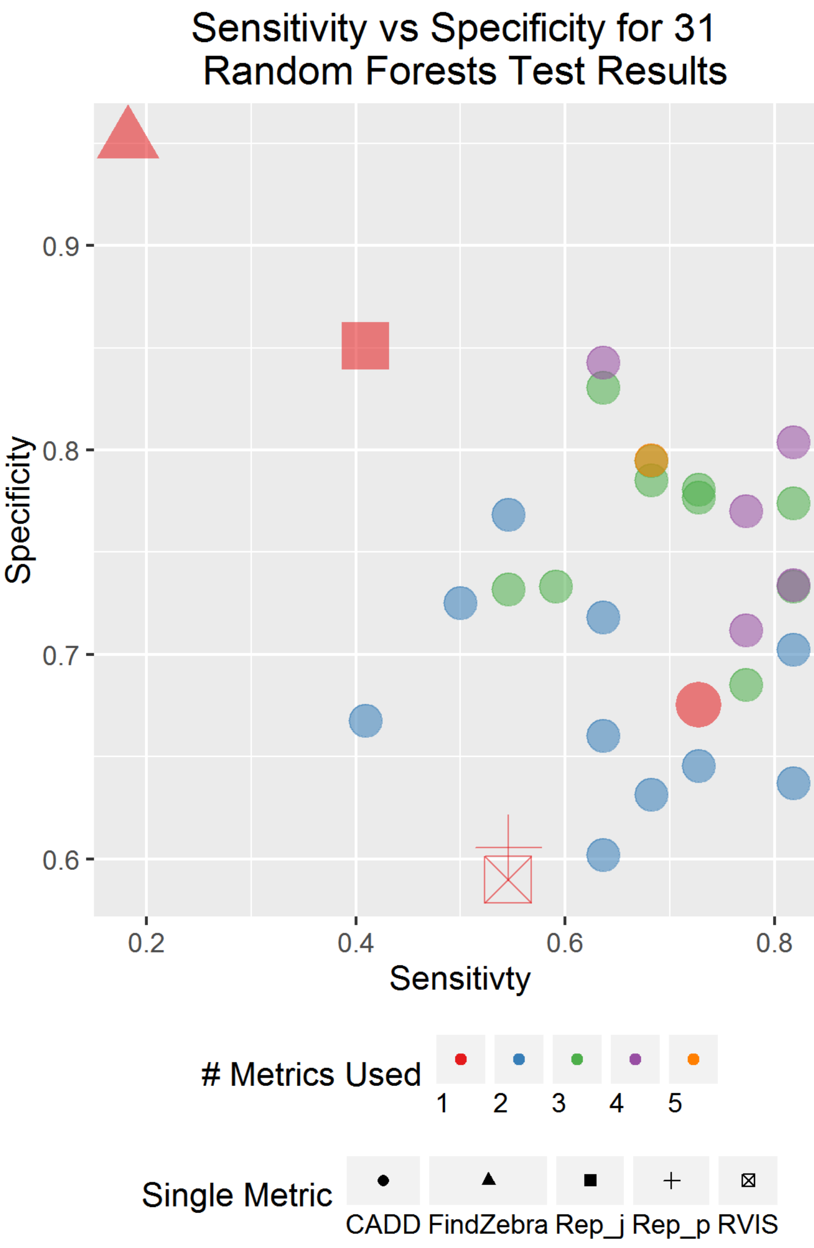

Figure 3.14: Sensitivity and specificity of all classification of test set variants.

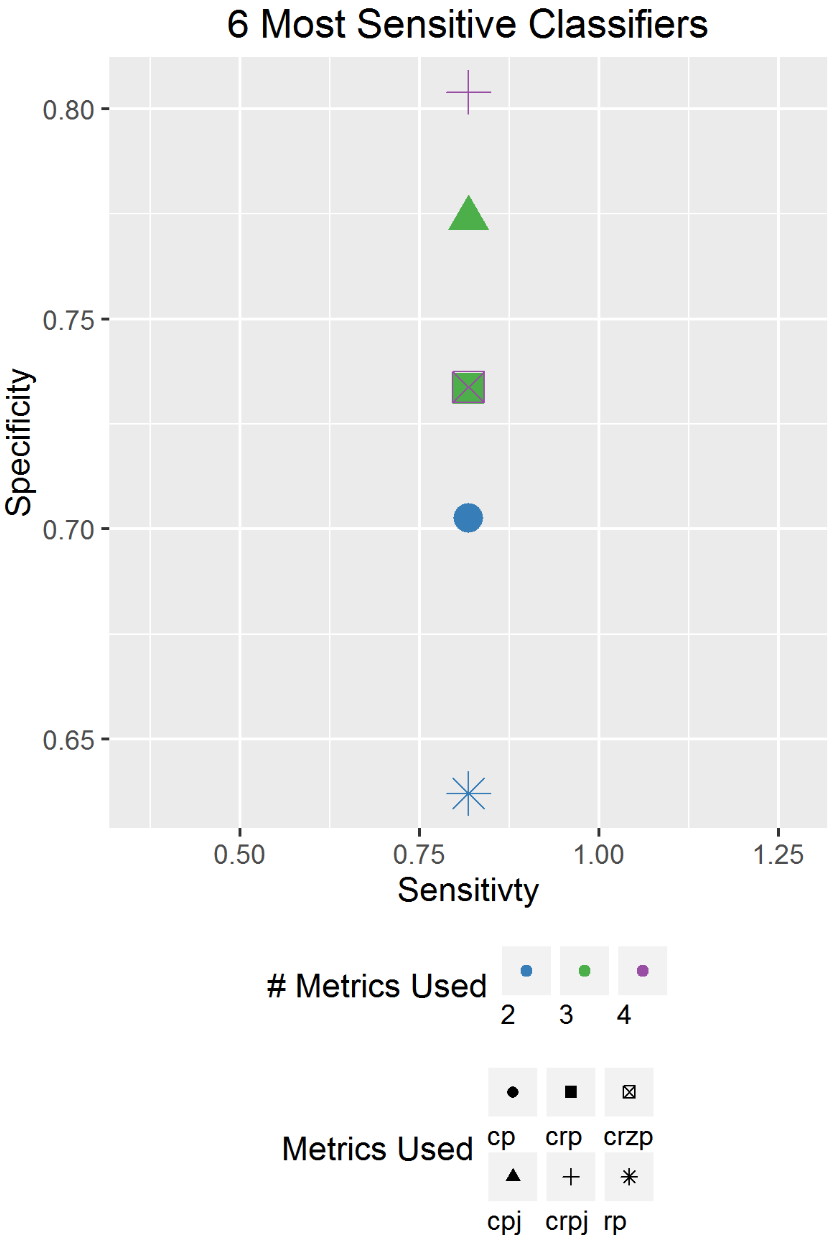

Figure 3.15: Top 6 classifiers based on test set classificiation sensitivity. Shapes denote the metrics used. The symbols used to denote which metrics a classifier used are c for CADD, z for FindZebra, j for rep_j and p for rep_p

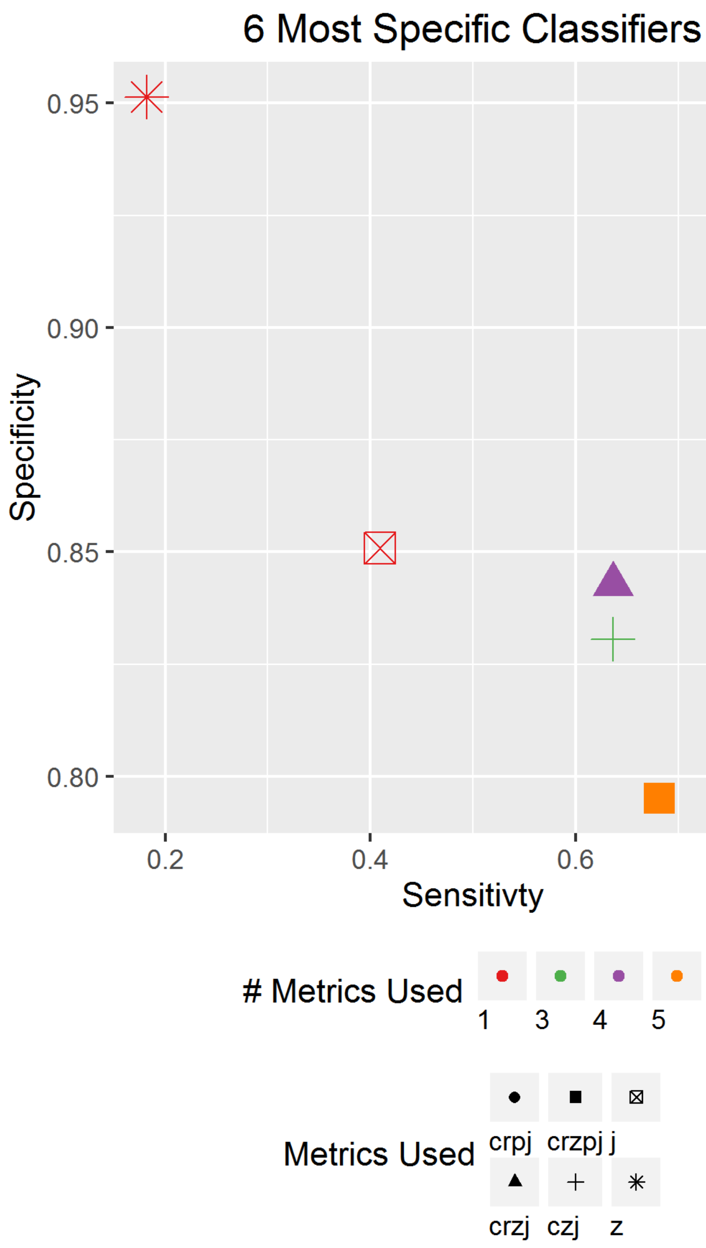

Figure 3.16: Top 6 classifiers based on test set classificiation specificity. Shapes denote the metrics used. The symbols used to denote which metrics a classifier used are c for CADD, z for FindZebra, j for rep_j and p for rep_p

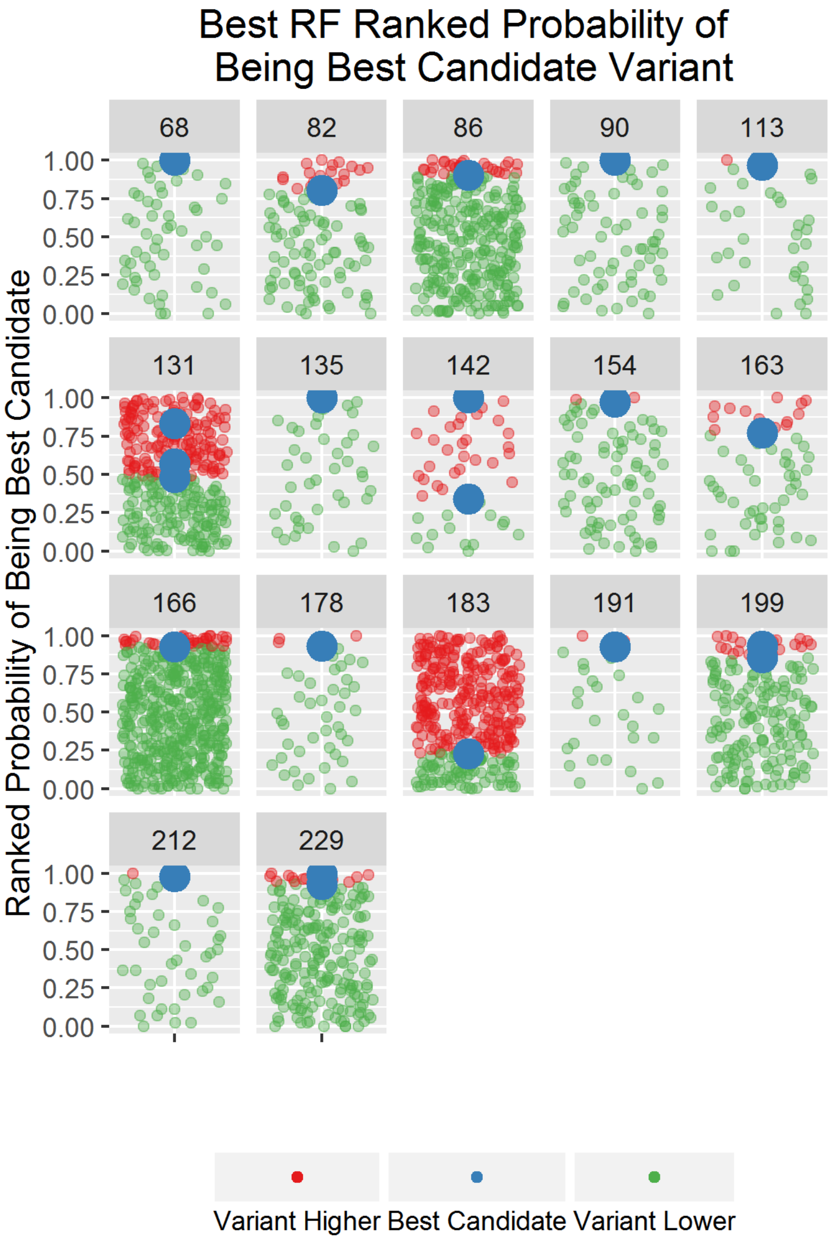

Figure 3.17: Distribution of random forest ranked probability across all 17 indidivudals in the test set. Variant higher are variants ranked higher than their case' lowest ranked best candidate variant. Variant lower are variants ranked lower than their case's lowest ranked best candidate variant.

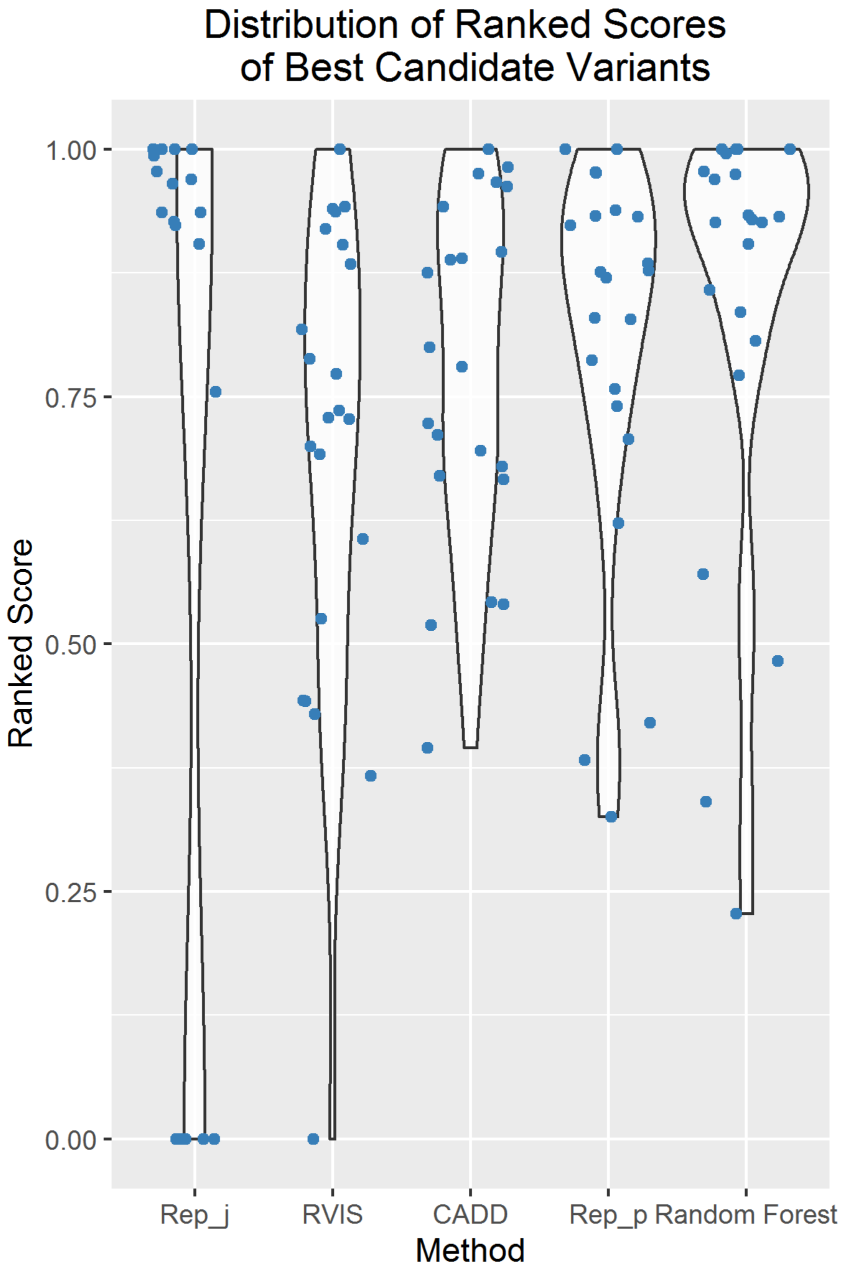

Figure 3.18: Distribution of best candidate variant ranking of four best individual methods and best random forest classifier, $RF_{r,c,p,j}$.

Chapter 4

Discussion

Despite continued advances in next generation sequencing, determining which variant(s) are most likely to cause a particular rare disease phenotype remains a challenge. This is due to a variety of reasons including the inexact relationship between genotype and phenotype and the number of rare diseases with unknown genetic bases. In this study I implemented a text-based approach for genetic variant prioritization, using the Synverita system, and assessed the utility of the approach to identify causal genetic alterations for patients with rare metabolic disorders. The approach performed well relative to widely used ranking methods.

Tools such as RVIS and CADD provide useful measures of potential gene and variant deleteriousness but fail to capture the link between variant genes and prominent clinical features. The above-discussed tools that focus on the link between phenotype and disease each have shortcomings. Beegle and Exomiser have high thresholds for use. Beegle has no backend and performs all text mining on the fly, which imposes a large time penalty. Exomiser imposes a less significant threshold by requiring a patient phenotype to be described in HPO terms. While the HPO continues to gain ground, it is not universally used. This means that in some cases, such as at TIDE BC, Exomiser requires changing existing clinician workflows. Finally, FindZebra provides an exceptionally easy-to-use API and user interface but lacks the sensitivity to prioritize variant genes with respect to disease phenotype in many TIDE BC cases.

Synverita is useful for prioritizing variants in cases of IDD caused by IEM.

Ranking variant genes based on strength of association to disease phenotype is effective using both Synverita's raw data and predictions. Both of these approaches contribute different strengths: raw data ranking is more likely to place a best candidate variant in the top 10% of all variants in a case than any other tested method. Prediction methods provide a floor for performance, with rep_p never ranking an established variant lower than 31% in any case tested in both the training and test sets. Synverita's large vocabulary and corpus and use of predictions provide enhanced sensitivity when compared with FindZebra, granting it potential to enhance diagnostics for rare genetic diseases of unknown etiology.

The strengths of Synverita are complementary to other measures of gene and variant deleteriousness currently used in TIDE BC's bioinformatics analysis pipelines. RVIS, CADD, rep_j and rep_p scores do not show strong correlation with one another, ostensibly measuring different features that correlate with the probability that a variant gene is implicated in disease. The combination of these methods is generalizable. The classifiers trained on the training set and tested on the test set ranked established candidate variant genes ahead of 90% of other variant genes in over half of all cases in the test set.

4.1 Limitations

Synverita's contribution to variant prioritization is tempered by three key limitations. First, the version of Synverita used in this work does not include the CUIDs of the terms in the HPO. These were excluded as described in Lever et al. (submitted) to reduce run time, as these terms frequently appear in Synverita's corpus. The lack of HPO terms in Synverita's corpus prevents us from leveraging HPO's structure. Second, variant prioritization using Synverita currently uses a semi-automated approach to map clinician reported phenotype terms to UMLS CUIDs in Synverita's vocabulary. This semi-automated approach is a barrier to use and may introduce noise due to unpredictability of user (mis)behaviour. Finally, Synverita does not yet have an update schedule or public access protocol, although work is currently underway to establish both.

4.2 Future Work

How terms are mapped to CUIDs is a key aspect of variant prioritization using Synverita. The comparison of different aggregation methods suggests that term mapping may significantly affect how well these methods work. Further work will investigate this. Future versions of Synverita will include all terms in the UMLS metathesaurus. This will allow for HPO-based term selection. Additionally, it will allow us to use existing tools to map phenotype terms to CUIDs, such as MetaMap[5]. These expansions will be compared to and combined with Exomiser. Collaboration with examining physicians will continue along multiple avenues. Performance may be improved by allowing clinicians to directly select the CUIDs and to provide weights to each phenotype term that describes a case. Ongoing collaboration may also see the development of a visual case explorer that builds upon visual phenotype comparison tools such as PhenoBlocks [19]. Perhaps most importantly, future work will use Synverita and ensemble methods on cases with no established best candidate variant.

4.3 Conclusion

Phenotype based variant prioritization using Synverita's raw data and predictions is effective for cases of IDD caused by IEM. It is complementary to measures of potential gene and variant deleteriousness. Classifiers trained using RVIS and CADD produce generalizable results. With modest modifications to Synverita, these methods can feasibly be incorporated into TIDE BC's pipeline and augment variant prioritization for cases of IDD caused by IEM.

Bibliography

[1] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau. Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5):537–544, 2006. ISSN 1087-0156. doi:10.1038/nbt1203. → pages 6

[2] N. Al Hafid and J. Christodoulou. Phenylketonuria: a review of current and future treatments. *Translational pediatrics*, 4(4):304–17, 2015. ISSN 2224-4344. doi:10.3978/j.issn.2224-4336.2015.10.07. URL http://www.ncbi.nlm.nih.gov/pubmed/26835392$\delimiter"026E30F$nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4728993. → pages 2

[3] M. Alfadhel, K. Al-Thihli, H. Moubayed, W. Eyaid, and M. Al-Jeraisy. Drug treatment of inborn errors of metabolism: a systematic review. *Archives of disease in childhood*, 98(6):454–61, 2013. ISSN 1468-2044. doi:10.1136/archdischild-2012-303131. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3693126{&}tool=pmcentrez{&}rendertype=abstract. → pages 2

[4] C. A. Argmann, S. M. Houten, J. Zhu, and E. E. Schadt. A Next Generation Multiscale View of Inborn Errors of Metabolism. *Cell Metabolism*, 23(1): 13–26, 2016. ISSN 19327420. doi:10.1016/j.cmet.2015.11.012. URL http://dx.doi.org/10.1016/j.cmet.2015.11.012. → pages 1

[5] A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–36, 2010. ISSN 1527-974X. doi:10.1136/jamia.2009.002733. URL http://www.ncbi.nlm.nih.gov/pubmed/20442139$\delimiter"026E30F$nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2995713. → pages 47

[6] P. a. Baird, T. W. Anderson, H. B. Newcombe, and R. B. Lowry. Genetic disorders in children and young adults: a population study. *American journal of human genetics*, 42(5):677–693, 1988. ISSN 0002-9297. → pages 1

[7] C. L. Beaulieu, J. Majewski, J. Schwartzentruber, M. E. Samuels, B. A. Fernandez, F. P. Bernier, M. Brudno, B. Knoppers, J. Marcadier, D. Dyment, S. Adam, D. E. Bulman, S. J. M. Jones, D. Avard, M. T. Nguyen, F. Rousseau, C. Marshall, R. F. Wintle, Y. Shen, S. W. Scherer, J. M. Friedman, J. L. Michaud, and K. M. Boycott. FORGE Canada consortium: Outcomes of a 2-year national rare-disease gene-discovery project. *American Journal of Human Genetics*, 94(6):809–817, 2014. ISSN 15376605. doi:10.1016/j.ajhg.2014.05.003. → pages 2

[8] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001): 267D–270, 2004. ISSN 1362-4962. doi:10.1093/nar/gkh061. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkh061. → pages 6

[9] W. P. Bone, N. L. Washington, O. J. Buske, D. R. Adams, J. Davis, D. Draper, E. D. Flynn, M. Girdea, R. Godfrey, G. Golas, C. Groden, J. Jacobsen, S. Köhler, E. M. J. Lee, A. E. Links, T. C. Markello, C. J. Mungall, M. Nehrebecky, P. N. Robinson, M. Sincan, A. G. Soldatos, C. J. Tifft, C. Toro, H. Trang, E. Valkanas, N. Vasilevsky, C. Wahl, L. A. Wolfe, C. F. Boerkoel, M. Brudno, M. A. Haendel, W. A. Gahl, and D. Smedley. Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genetics in Medicine*, 18(6):608–617, 2016. ISSN 1098-3600. doi:10.1038/gim.2015.137. URL http://www.nature.com/doifinder/10.1038/gim.2015.137. → pages 5

[10] A. Borzutzky, B. Crompton, A. K. Bergmann, S. Giliani, S. Baxi, M. Martin, E. J. Neufeld, and L. D. Notarangelo. Reversible severe combined immunodeficiency phenotype secondary to a mutation of the proton-coupled folate transporter. *Clinical Immunology*, 133(3):287–294, 2009. doi:10.1016/j.clim.2009.08.006.Reversible. → pages 1

[11] K. M. Boycott, M. R. Vanstone, D. E. Bulman, and A. E. MacKenzie. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature reviews. Genetics*, 14(10):681–91, 2013. ISSN 1471-0064. doi:10.1038/nrg3555. URL http://www.nature.com/doifinder/10.1038/nrg3555$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/pubmed/23999272. → pages 1, 2

[12] O. J. Buske, M. Girdea, S. Dumitriu, B. Gallinger, T. Hartley, H. Trang, A. Misyura, T. Friedman, C. Beaulieu, W. P. Bone, A. E. Links, N. L. Washington, M. A. Haendel, P. N. Robinson, C. F. Boerkoel, D. Adams, W. A. Gahl, K. M. Boycott, and M. Brudno. PhenomeCentral: A Portal for Phenotypic and Genotypic Matchmaking of Patients with Rare Genetic Diseases. *Human Mutation*, 36(10):931–940, 2015. ISSN 10981004. doi:10.1002/humu.22851. → pages 5

[13] L. S. CARULLA, G. M. REED, L. M. VAEZ-AZIZI, S.-A. COOPER, R. M. LEAL, M. BERTELLI, C. ADNAMS, S. COORAY, S. DEB, L. A. DIRANI, S. C. GIRIMAJI, G. KATZ, H. KWOK, R. LUCKASSON, R. SIMEONSSON, C. WALSH, K. MUNIR, and S. SAXENA. Intellectual developmental disorders: towards a new name, definition and framework for "mental retardation/intellectual disability" in ICD-11. *World Psychiatry*, 10 (3):175–180, 2011. ISSN 17238617. doi:10.1002/j.2051-5545.2011.tb00045.x. URL http://doi.wiley.com/10.1002/j.2051-5545.2011.tb00045.x. → pages 1

[14] R. Dragusin, P. Petcu, C. Lioma, B. Larsen, H. L. Jørgensen, I. J. Cox, L. K. Hansen, P. Ingwersen, and O. Winther. FindZebra: A search engine for rare diseases. *International Journal of Medical Informatics*, 82(6):528–538, 2013. ISSN 13865056. doi:10.1016/j.ijmedinf.2013.01.005. URL http://dx.doi.org/10.1016/j.ijmedinf.2013.01.005. → pages 6

[15] S. ElShal, L.-C. Tranchevent, A. Sifrim, A. Ardeshirdavani, J. Davis, and Y. Moreau. Beegle: from literature mining to disease-gene discovery. *Nucleic acids research*, 44(2):e18, 2016. ISSN 1362-4962. doi:10.1093/nar/gkv905. URL http://nar.oxfordjournals.org/content/44/2/e18.short?rss=1. → pages 6, 11

[16] J. Ford, F. Makedon, and J. Pearlman. Using Singular Value Decomposition Approximation for Collaborative Filtering. *Seventh IEEE International Conference on E-Commerce Technology (CEC'05)*, pages 257–264, 2005. doi:10.1109/ICECT.2005.102. URL http://ieeexplore.ieee.org/xpl/freeabs{_}all.jsp?arnumber=1524053. → pages 8

[17] M. K. C. from Jed Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, and T. Hunt. *caret: Classification and Regression Training*, 2016. URL

https://CRAN.R-project.org/package=caret. R package version 6.0-72. → pages 15

[18] M. Girdea, S. Dumitriu, M. Fiume, S. Bowdin, K. M. Boycott, S. Chénier, D. Chitayat, H. Faghfoury, M. S. Meyn, P. N. Ray, J. So, D. J. Stavropoulos, and M. Brudno. PhenoTips: Patient phenotyping software for clinical and research use. *Human Mutation*, 34(8):1057–1065, 2013. ISSN 10597794. doi:10.1002/humu.22347. → pages 5

[19] M. Glueck, P. Hamilton, F. Chevalier, S. Breslav, A. Khan, D. Wigdor, and M. Brudno. PhenoBlocks: Phenotype Comparison Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):101–110, 2016. ISSN 10772626. doi:10.1109/TVCG.2015.2467733. → pages 5, 47

[20] M. M. Gottlieb, D. J. Arenillas, S. Maithripala, Z. D. Maurer, M. Tarailo-Graovac, L. Armstrong, M. Patel, C. van Karnebeek, and W. W. Wasserman. GeneYenta: A phenotype-based rare disease case matching tool based on online dating algorithms for the acceleration of exome interpretation. *Human Mutation*, 36(4):432–438, 2015. ISSN 10981004. doi:10.1002/humu.22772. → pages 5

[21] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–315, 2014. ISSN 1546-1718. doi:10.1038/ng.2892. URL http://www.ncbi.nlm.nih.gov/pubmed/24487276. → pages 4

[22] A. Knight and T. Senior. The common problem of rare disease in general practice. *Medical Journal of Australia*, 185(2):2–3, 2006. URL https://www.mja.com.au/system/files/issues/185{_}02{_}170706/kni10328{_}fm.pdf. → pages 1

[23] S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. N. Robinson. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *American Journal of Human Genetics*, 85(4):457–464, 2009. ISSN 00029297. doi:10.1016/j.ajhg.2009.09.003. → pages 5

[24] S. Köhler, S. C. Doelken, C. J. Mungall, S. Bauer, H. V. Firth, I. Bailleul-Forestier, G. C. M. Black, D. L. Brown, M. Brudno, J. Campbell, D. R. Fitzpatrick, J. T. Eppig, A. P. Jackson, K. Freson, M. Girdea, I. Helbig, J. A. Hurst, J. Jähn, L. G. Jackson, A. M. Kelly, D. H. Ledbetter, S. Mansour,

C. L. Martin, C. Moss, A. Mumford, W. H. Ouwehand, S. M. Park, E. R. Riggs, R. H. Scott, S. Sisodiya, S. V. Vooren, R. J. Wapner, A. O. M. Wilkie, C. F. Wright, A. T. Vulto-Van Silfhout, N. D. Leeuw, B. B. A. De Vries, N. L. Washingthon, C. L. Smith, M. Westerfield, P. Schofield, B. J. Ruef, G. V. Gkoutos, M. Haendel, D. Smedley, S. E. Lewis, and P. N. Robinson. The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(D1):1–9, 2014. ISSN 03051048. doi:10.1093/nar/gkt1026. → pages 5

[25] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, and D. G. MacArthur. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536 (7616):285–291, 2016. ISSN 0028-0836. doi:10.1038/nature19057. URL http://www.nature.com/doifinder/10.1038/nature19057. → pages 4

[26] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, 26(16):2069–2070, 2010. ISSN 13674803. doi:10.1093/bioinformatics/btq330. → pages 4

[27] W. J. Meerding, L. Bonneux, J. J. Polder, M. A. Koopmanschap, and P. J. van der Maas. Demographic and epidemiological determinants of healthcare costs in Netherlands: cost of illness study. *BMJ (Clinical research ed.)*, 317 (7151):111–115, 1998. ISSN 0959-8138 (Print). doi:10.1136/bmj.317.7151.111. → pages 1

[28] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási. Disease networks. Uncovering disease-disease

relationships through the incomplete interactome. *Science*, 347(6224): 1257601, 2015. ISSN 1095-9203. doi:10.1126/science.1116608. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1257601$\delimiter"026E30F$npapers3://publication/doi/10.1126/science.1257601. → pages 5

[29] N. A. Miller, E. G. Farrow, M. Gibson, L. K. Willig, G. Twist, B. Yoo, T. Marrs, S. Corder, L. Krivohlavek, A. Walter, J. E. Petrikin, C. J. Saunders, I. Thiffault, S. E. Soden, L. D. Smith, D. L. Dinwiddie, S. Herd, J. A. Cakici, S. Catreux, M. Ruehle, and S. F. Kingsmore. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Medicine*, 2015. ISSN 1756-994X. doi:10.1186/s13073-015-0221-8. URL http://dx.doi.org/10.1186/s13073-015-0221-8. → pages 2

[30] S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, and D. B. Goldstein. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genetics*, 9(8), 2013. ISSN 15537390. doi:10.1371/journal.pgen.1003709. → pages 4

[31] J. Piñero, A. Berenstein, A. Gonzalez-Perez, A. Chernomoretz, and L. I. Furlong. Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing. *Scientific reports*, 6(October 2015): 24570, 2016. ISSN 2045-2322. doi:10.1038/srep24570. URL http://www.ncbi.nlm.nih.gov/pubmed/27080396$\delimiter"026E30F$nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4832203. → pages 5

[32] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/. → pages 22

[33] C. R. Scriver and P. J. Waters. Monogenic traits are not simple: Lessons from phenylketonuria. *Trends in Genetics*, 15(7):267–272, 1999. ISSN 01689525. doi:10.1016/S0168-9525(99)01761-8. → pages 1

[34] M. Shevell. Global Developmental Delay and Mental Retardation or Intellectual Disability: Conceptualization, Evaluation, and Etiology. *Pediatric Clinics of North America*, 55(5):1071–1084, 2008. ISSN 00313955. doi:10.1016/j.pcl.2008.07.010. → pages 1

[35] C. Shyr, M. Tarailo-Graovac, M. Gottlieb, J. J. Y. Lee, C. van Karnebeek, and W. W. Wasserman. FLAGS, frequently mutated genes in public exomes.

BMC medical genomics, 7:64, 2014. ISSN 1755-8794. doi:10.1186/s12920-014-0064-y. URL http://www.biomedcentral.com/1755-8794/7/64. → pages 5

[36] D. Smedley and P. N. Robinson. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Medicine*, 7(1): 81, 2015. ISSN 1756-994X. doi:10.1186/s13073-015-0199-2. URL http://genomemedicine.com/content/7/1/81. → pages 5

[37] D. Smedley, J. O. B. Jacobsen, M. Jäger, S. Köhler, M. Holtgrewe, M. Schubach, E. Siragusa, T. Zemojtel, O. J. Buske, N. L. Washington, W. P. Bone, M. a. Haendel, and P. N. Robinson. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature protocols*, 10(12): 2004–2015, 2015. ISSN 1750-2799. doi:10.1038/nprot.2015.124. URL http://www.ncbi.nlm.nih.gov/pubmed/26562621. → pages 5

[38] S. E. Soden, C. J. Saunders, L. K. Willig, E. G. Farrow, L. D. Smith, J. E. Petrikin, J.-b. Lepichon, N. A. Miller, I. Thiffault, D. L. Dinwiddie, G. Twist, A. Noll, A. Bryce, L. Zellmer, A. M. Atherton, A. T. Abdelmoity, N. Safina, S. S. Nyp, B. Zuccarelli, I. A. Larson, A. Modrcin, S. Herd, M. Creed, Z. Ye, X. Yuan, R. A. Brodsky, and F. Stephen. Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. 6(265), 2015. doi:10.1126/scitranslmed.3010076.Effectiveness. → pages 1, 2

[39] H. Stranneheim, M. Engvall, K. Naess, N. Lesko, P. Larsson, M. Dahlberg, R. Andeer, A. Wredenberg, C. Freyer, M. Barbaro, H. Bruhn, T. Emahazion, M. Magnusson, R. Wibom, R. H. Zetterström, V. Wirta, U. von Döbeln, and A. Wedell. Rapid pulsed whole genome sequencing for comprehensive acute diagnostics of inborn errors of metabolism. *BMC Genomics*, 15(1):1090, 2014. ISSN 1471-2164. doi:10.1186/1471-2164-15-1090. URL http://www.biomedcentral.com/1471-2164/15/1090. → pages 2

[40] D. C. Swinney. Challenges and Hurdles to Business as Usual in Drug Development for Treatment of Rare Diseases. *Clinical Pharmacology & Therapeutics*, 100(4):339–341, 2016. ISSN 00099236. doi:10.1002/cpt.422. URL http://doi.wiley.com/10.1002/cpt.422. → pages 1

[41] M. Tarailo-Graovac, C. Shyr, C. J. Ross, G. A. Horvath, R. Salvarinova, X. C. Ye, L.-H. Zhang, A. P. Bhavsar, J. J. Lee, B. I. Drögemöller, M. Abdelsayed, M. Alfadhel, L. Armstrong, M. R. Baumgartner, P. Burda, M. B. Connolly, J. Cameron, M. Demos, T. Dewan, J. Dionne, A. M. Evans,

J. M. Friedman, I. Garber, S. Lewis, J. Ling, R. Mandal, A. Mattman, M. McKinnon, A. Michoulas, D. Metzger, O. A. Ogunbayo, B. Rakic, J. Rozmus, P. Ruben, B. Sayson, S. Santra, K. R. Schultz, K. Selby, P. Shekel, S. Sirrs, C. Skrypnyk, A. Superti-Furga, S. E. Turvey, M. I. Van Allen, D. Wishart, J. Wu, J. Wu, D. Zafeiriou, L. Kluijtmans, R. A. Wevers, P. Eydoux, A. M. Lehman, H. Vallance, S. Stockler-Ipsiroglu, G. Sinclair, W. W. Wasserman, and C. D. van Karnebeek. Exome Sequencing and the Management of Neurometabolic Disorders. *New England Journal of Medicine*, page NEJMoa1515792, 2016. ISSN 0028-4793. doi:10.1056/NEJMoa1515792. URL http://www.nejm.org/doi/10.1056/NEJMoa1515792. → pages 2, 3, 14, 21

[42] L. C. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo, B. Coessens, B. De Moor, S. Aerts, and Y. Moreau. ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic acids research*, 36(Web Server issue):377–384, 2008. ISSN 13624962. doi:10.1093/nar/gkn325. → pages 6

[43] C. D. M. Van Karnebeek and S. Stockler. Treatable inborn errors of metabolism causing intellectual disability: A systematic literature review. *Molecular Genetics and Metabolism*, 105(3):368–381, 2012. ISSN 10967192. doi:10.1016/j.ymgme.2011.11.191. URL http://dx.doi.org/10.1016/j.ymgme.2011.11.191. → pages 2

[44] C. D. M. Van Karnebeek, M. Shevell, J. Zschocke, J. B. Moeschler, and S. Stockler. The metabolic evaluation of the child with an intellectual developmental disorder: Diagnostic algorithm for identification of treatable causes and new digital resource. *Molecular Genetics and Metabolism*, 111 (4):428–438, 2014. ISSN 10967206. doi:10.1016/j.ymgme.2014.01.011. URL http://dx.doi.org/10.1016/j.ymgme.2014.01.011. → pages 2

[45] Z. Yang and J. R. Bielawski. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, 15(12):496–503, 2000. ISSN 01695347. doi:10.1016/S0169-5347(00)01994-7. → pages 5

Appendix A

Tables

Weighting	Method	Min	1st Quartile	Median	Mean	3rd Quartile	Correlation
None	avg_j	0.00	0.73	0.89	0.79	0.97	0.738
	$best_j$	0.00	0.63	0.92	0.76	0.97	0.649
	for_j	0.00	0.65	0.88	0.77	0.97	0.652
	rep_j	0.00	0.69	0.87	0.72	0.98	0.585
	avg_p	0.03	0.62	0.80	0.74	0.92	0.810
	$best_p$	0.27	0.59	0.78	0.74	0.90	0.939
	rep_p	0.37	0.65	0.83	0.77	0.92	0.785
w_g	avg_j	0.00	0.72	0.90	0.80	0.98	0.731
	$best_j$	0.00	0.62	0.90	0.76	0.97	0.633
	for_j	0.00	0.64	0.86	0.77	0.97	0.636
	rep_j	0.00	0.71	0.95	0.72	0.98	0.575
	avg_p	0.05	0.63	0.80	0.74	0.93	0.793
	$best_p$	0.27	0.59	0.76	0.74	0.92	0.931
	rep_p	0.29	0.66	0.81	0.76	0.92	0.774
w_p	avg_j	0.00	0.73	0.90	0.80	0.98	0.736
	$best_j$	0.00	0.63	0.92	0.76	0.97	0.643
	for_j	0.00	0.62	0.86	0.76	0.97	0.644
	rep_j	0.00	0.71	0.87	0.72	0.99	0.584
	avg_p	0.03	0.63	0.80	0.74	0.92	0.791
	$best_p$	0.27	0.59	0.78	0.74	0.90	0.948
	rep_p	0.29	0.65	0.80	0.76	0.92	0.792
w_{g+p}	avg_j	0.00	0.73	0.90	0.80	0.98	0.736
	$best_j$	0.00	0.63	0.92	0.76	0.97	0.640
	for_j	0.00	0.62	0.86	0.76	0.97	0.643
	rep_j	0.00	0.71	0.87	0.72	0.98	0.580
	avg_p	0.03	0.63	0.80	0.74	0.92	0.795
	$best_p$	0.27	0.59	0.78	0.74	0.90	0.946
	rep_p	0.38	0.65	0.81	0.77	0.92	0.784
w_{g*p}	avg_j	0.00	0.72	0.92	0.80	0.98	0.728
	$best_j$	0.00	0.63	0.89	0.76	0.97	0.628
	for_j	0.00	0.63	0.87	0.76	0.97	0.629
	rep_j	0.00	0.71	0.95	0.73	0.99	0.566
	avg_p	0.05	0.64	0.80	0.74	0.92	0.772
	$best_p$	0.27	0.59	0.76	0.74	0.92	0.931
	rep_p	0.37	0.64	0.84	0.78	0.92	0.758

Table A.1: Ranked scoring results for best candidate variants for all weighting methods. Correlation is Spearman's ρ of all ranked scores of all variants in a case and their ranked number of unique terms that they associate with in Syverita's raw data. Maximum is 1 all cases.