Statistical Ranking Model for Candidate Genes in Rare Genetic Disorders

Presented by: Vishnu Emani

Mentors: Dr. Klaus Schmitz-Abe, Dr. Pankaj Agrawal

Introduction

- The exome (protein coding region) makes up approximately 1% of the human genome
- Studying mutations in this region very important for identifying genetic basis of certain rare disorders
- The advent of whole exome and whole genome sequencing allows us to find the most likely pathogenic mutations much more efficiently and on a greater scale
- Next-gen sequencing has been growing rapidly in the past decade and has led to numerous successful disease-detection pipelines.

Introduction- Genetic Analyses

- In analyzing a candidate variant, a geneticist utilizes a variety of factors
- Categories: genotype and phenotype
- Use these metrics to select a few variants that are most likely to be mutation-causing
- The goal of this project was to use computational approaches to generate a ranked list of most likely mutations

Genotype:
-Allele frequency in control populations
-Resistance of gene to LoF mutation
-Evolutionary conservation

Phenotype:
-Comparison with associated disease
-Specificity of gene phenotype for patient
-Novelty of gene

General Methods

- Create two scores for each mutation (genotype score and phenotype score) using appropriate classification techniques
- Combine these two scores to generate an overall mutation probability score and generate a ranking based upon this score
- Develop a user-friendly website that allows clients to input mutation files and view the returned ranking

Methods- Genotype classification

- To train the model, data was imported from 35 manual classifications of previous patient cases
- Difficulty: data format was one “Correct” variant per family
  - Computer is being trained to think that all of the other variants are unlikely
- Thus, very high number of patient cases are necessary to train our model
- Logistic regression models were trained using a large number of predictors against the pathogenicity.
- Insignificant predictors were removed to ensure the accuracy of the model.
- Testing of the model was performed, as will be outlined in the results section

Methods- Phenotype algorithms

Patient phenotype: P1, P2, P3

Gene 1
Gene 5
Gene 102
Gene 105
Gene 107

Database:

Gene 1	P1, P2
Gene 2	P2, P4, P5
Gene 3	P6, P7
Gene 4	P1, P5, P6
Gene 5	P6, P7
…	…
Gene 102	P1, P2, P8
Gene 103	P2, P9
Gene 104	P2, P8, P10
Gene 105	P2, P10, P11
Gene 106	P12, P13,P14
Gene 107	P15, P16, P17

Methods- Phenotype algorithms

Patient phenotype: P1, P2, P3

Gene 1
Gene 5
Gene 102
Gene 105
Gene 107

1. How much of the patient's phenotype is covered by each gene?
2. How common is the patient's phenotype in the gene database?

Database:

Gene 1	P1, P2
Gene 2	P2, P4, P5
Gene 3	P6, P7
Gene 4	P1, P5, P6
Gene 5	P6, P7
…	…
Gene 102	P1, P2, P8
Gene 103	P2, P9
Gene 104	P2, P8, P10
Gene 105	P2, P10, P11
Gene 106	P12, P13,P14
Gene 107	P15, P16, P17

Phenotype algorithms- Part 2

For each gene, two values:

1. Percent of the patient's phenotype that is covered by each gene (P_{gene})
2. Percent of genes in the database that catch the patient's phenotype(P_{db})

Phenotype algorithms- Part 2

For each gene, two values:

1. Percent of the patient's phenotype that is covered by each gene (P_{gene})
2. Percent of genes in the database that catch the patient's phenotype(P_{db})

Algorithm:

$$P = (1 - P_{db}) * P_{genes}$$

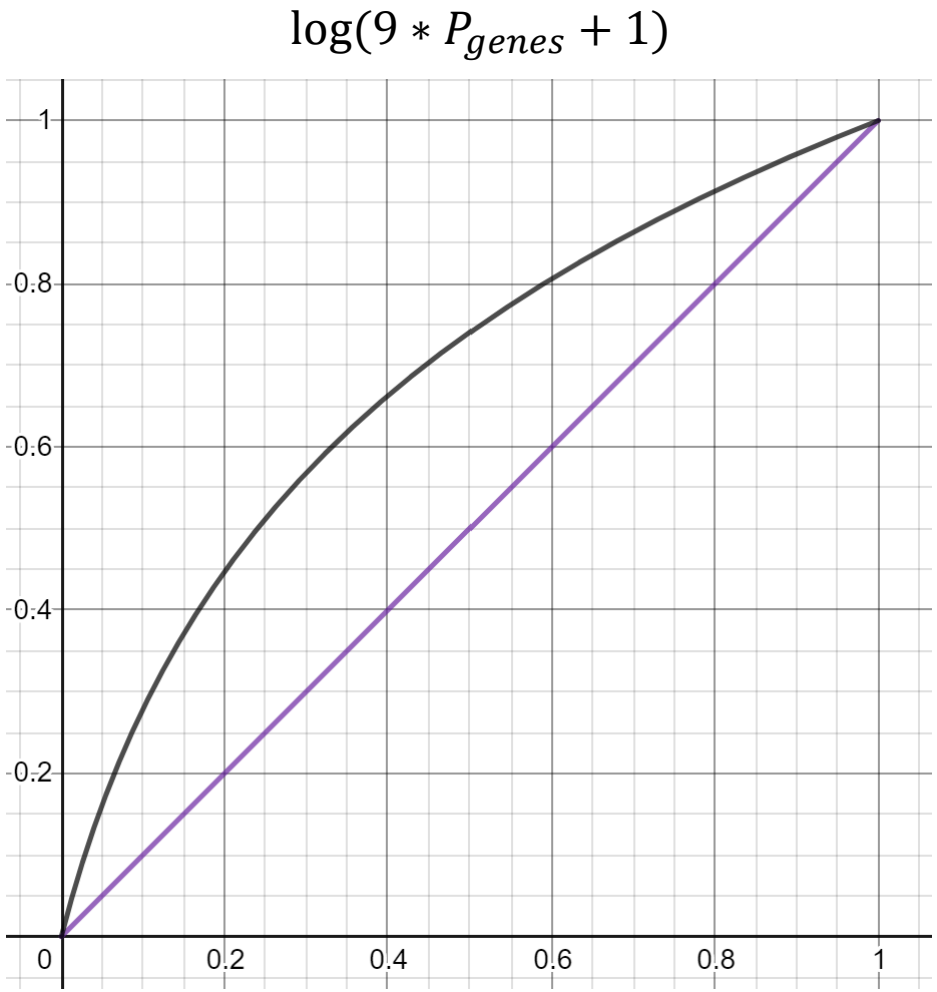
- When $P_{db} = 100\%$, this means no specificity, so $P = 0$
- When $P_{gene} = 0\%$, meaning that a gene has no phenotype hits, $P = 0$
- Issue with this algorithm was that it was too restrictive for P_{gene}: if 50% of the patient's phenotype had overlap, the value could only be a max of 50%, even if 0 other genes caught keywords
    - Used a log scale to help with this issue

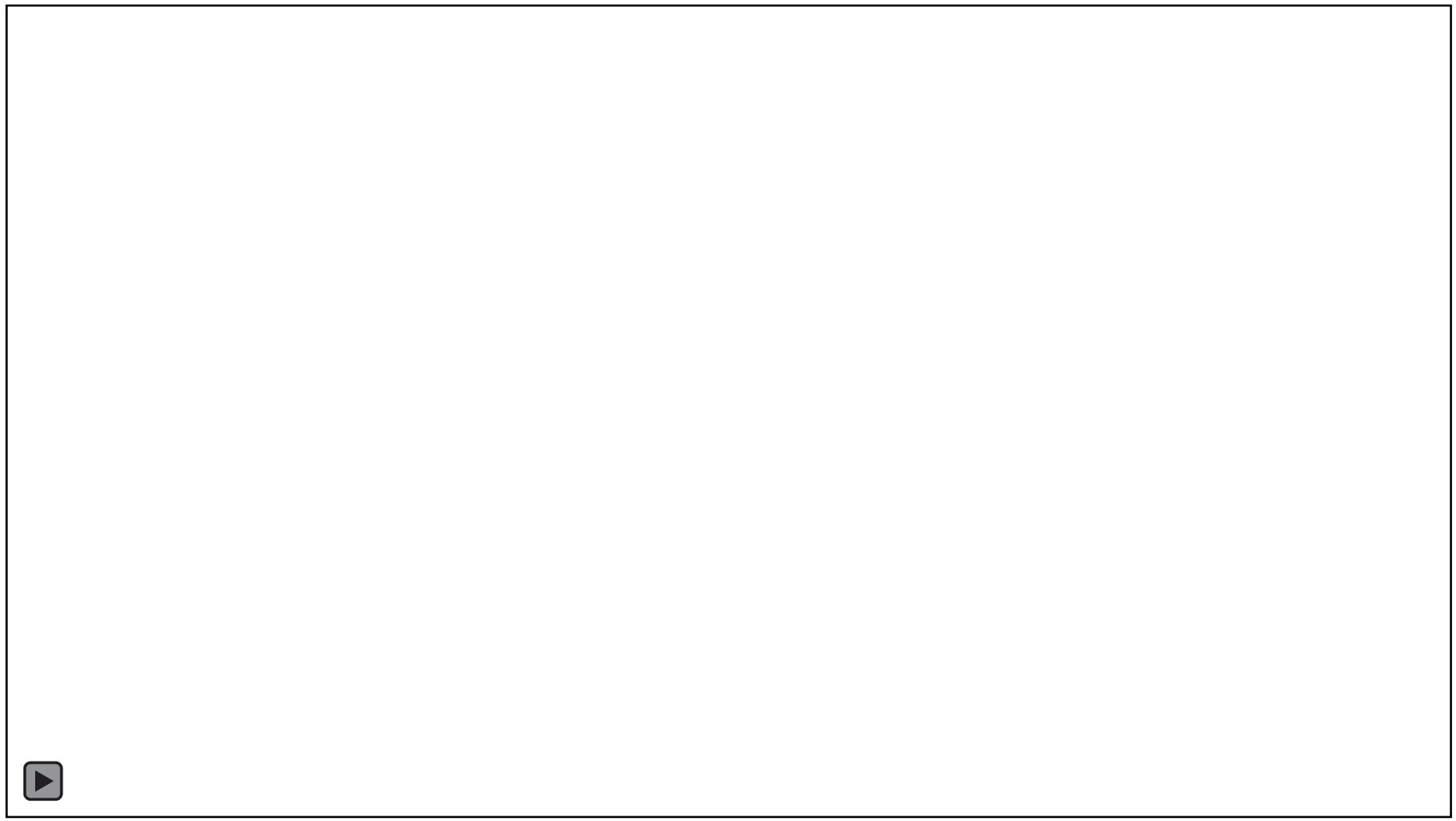
Phenotype algorithms- Part 3

With log scales:

$$P = (1 - P_{db}) * \log(9 * P_{genes} + 1)$$

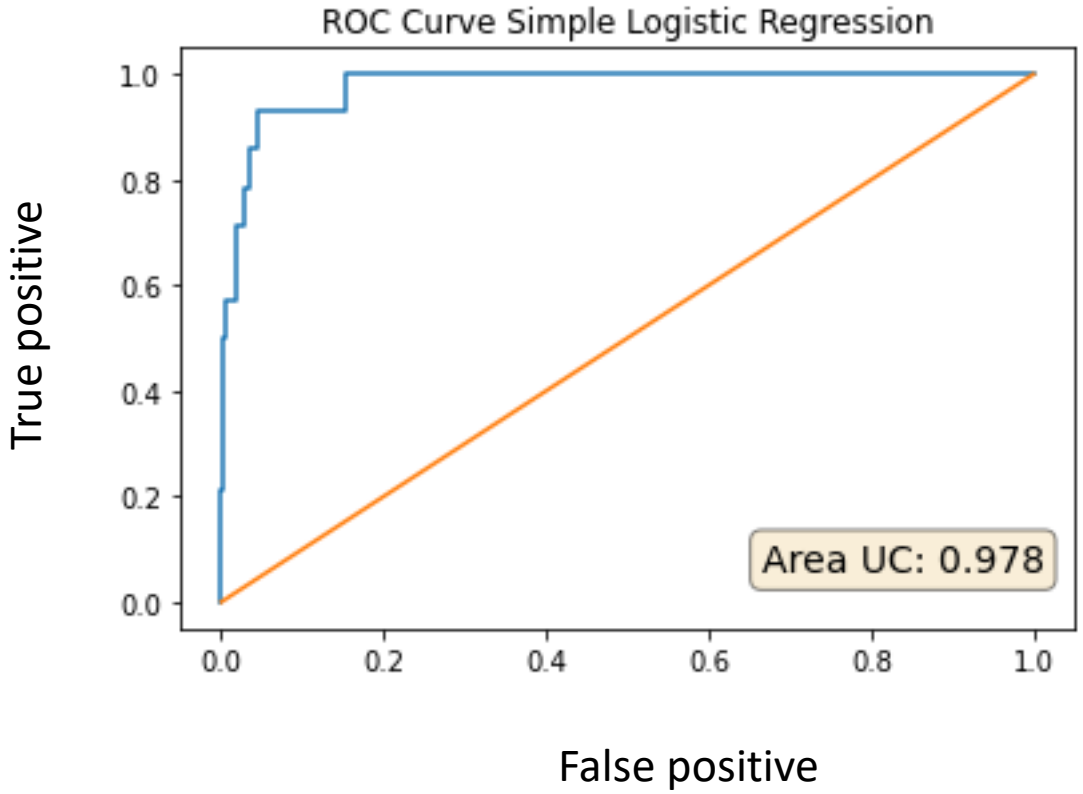
With this function, the endpoints of P_{genes} (i.e. [0,1]) stay the same, while the rest of the graph is distorted up from the initial values





Results/testing our algorithms

- ROC Curves:
  - A graph of the false-positive rate vs. the true-positive rate
  - Change the thresholds and make the graph
  - Take the area under the curve
  - Helps determine how well the model predicts the outcome
  - Perfect model would be box: there exists a threshold such that the true positive rate is 100% and false positive rate is 0%
- After cross-validation was implemented across the 35 sets available, the average AUC score was around 0.89



Conclusion

- More work can still be done to improve the models
- In the future, more data will have to be used to train the model, since the current model was trained with only limited data.
- Nevertheless, computational approaches for genetic analysis of rare disease patients seem to show great promise.
- Furthermore, the phenotype databases available are continuously aggregating more and more data, so our phenotype algorithms will become even more informative as time goes on.
- As next-gen sequencing continues to grow, these tools can continue to provide more effective and efficient ways to perform genetic analyses.

References

- Schmitz-Abe K et al. Unique bioinformatic approach and comprehensive reanalysis improve diagnostic yield of clinical exomes. Eur J Hum Genet. 2019 Sep;27(9):1398-1405. doi: 10.1038/s41431-019-0401-x. Epub 2019 Apr 12. PMID: 30979967; PMCID: PMC6777619.