Phenotypic Driven Approach to Identify Clinically Relevant Pattern in Trio WES Data

Children's Hospital of Philadelphia
Epilepsy Neurogenetics Initiative

Children's Hospital of Philadelphia
Department of Biomedical & Health Informatics

Shiva Ganesan[1,2,3], Peter Galer[1,2,3,4,5], Julie Xian[1,2,3], Sridhar Parthasarathy[1,2,3], Sarah M. Ruggiero[1,3], Stacey Cohen[1,3], Katherine L. Helbig[1,3], Colin A. Ellis[3,5], and Ingo Helbig[1,2,3,5]

1 Division of Neurology, Children's Hospital of Philadelphia; 2 Department of Biomedical and Health Informatics (DBHi), Children's Hospital of Philadelphia; 3 The Epilepsy NeuroGenetics Initiative (ENGIN), Children's Hospital of Philadelphia; 4 Department of Bioengineering, University of Pennsylvania ; 5 Department of Neurology, Perelman School of Medicine, University of Pennsylvania.

Rationale

- Trio-base Whole Exome Sequencing (WES) data are used to diagnose *de novo* variants confidently.
- Some genetic etiologies have a clear phenotypic association, but many demonstrate with a wide spectrum of phenotypic heterogeneity.
- Correlating genetic findings with clinical feature at scale remains a hurdle
- We use computational phenotypes in individuals with WES data to identify relevant phenotypic similarities.

Methods

- Trio WES data in 9,190 individuals were analyzed for *de novo* variants. The most frequent genes are:
  - ***BPTF*** (n=19), ***KCNQ2*** (n=17), ***STXBP1*** (n=17), ***MECP2*** (n=16), ***SCN1A*** (n=15), ***PACS1*** (n=14).
- Human Phenotype Ontology (HPO) terms was used to annotate phenotypic data in 14,270 individuals
- The most common HPO terms were Global developmental delay (HP:0001263, 27%), Delayed speech and language development (HP:0000750, 16%).
- Resnik-mod similarity algorithm was primarily used for this study. We also compared performance with other similarity algorithms.

Results

- We analyzed 103,523 HPO terms in 14,270 individuals, including 4,357 unique terms.
- Of the 280 genes with two or more *de novo* variants, 23 genes had a phenotypic similarity higher than expected:
  - ***AP2M1*** (n=4; $p < 0.0001$), ***DNM1*** (n=6; $p < 0.0001$), ***GABRB3*** (n=6; $p < 0.0001$), ***SCN1A*** (n=15; $p < 0.0001$), ***PACS1*** (n=14; $p = 0.0001$), and ***STXBP1*** (n=17; $p = 0.0006$)
- Decomposition of phenotypic similarity revealed gene-specific signatures including:
  - ***SCN1A*** – Bilalteral tonic-clonic seizure (HP:0002069, $p < 0.001$), Focal clonic seizure (HP:0002266, $p < 0.001$).

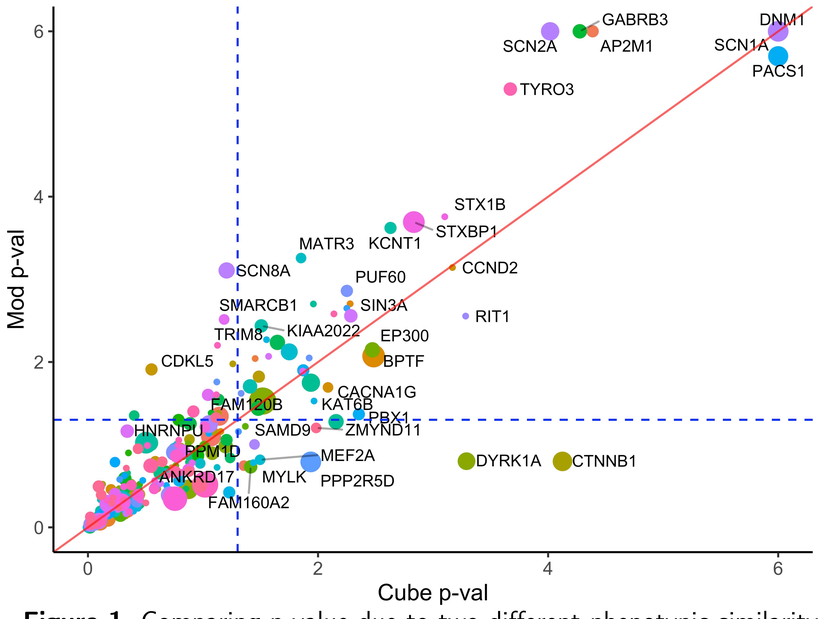

Figure 1. Comparing p-value due to two different phenotypic similarity algorithms. The dotted blue line signifies a score of p=0.05. X-axis represents the Cube algorithm and Y-axis represents Resnik-mod algorithm.

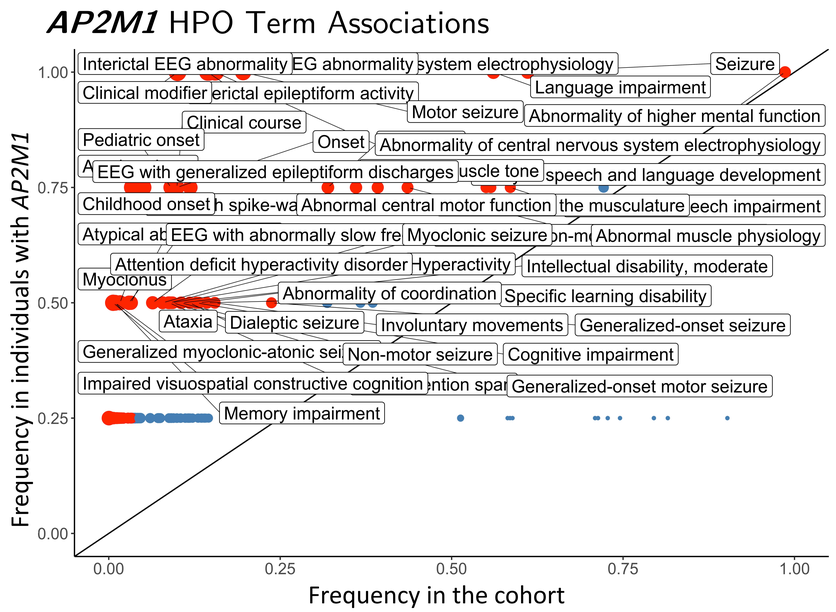

Figure 2. Frequency of HPO terms associated with *AP2M1* (n=4). Color of points indicate significance level of each HPO term with respect to the gene. Red indicates significance of $p < 0.05$.

Conclusion

- Computational phenotyping can be used to generate statistical evidence for disease causation using similarity algorithms.
- HPO driven phenotypic analysis picks up distinct genetic profiles, reflecting the heterogeneity of phenotypic data.
- Phenotypic similarity algorithms can be used to detect disease entities by identifying individuals with overlapping phenotypic features with same rare genetic etiology.
- This approach can aid in gene discovery and gene prediction.