University of Kansas Medical Center

Automated Functional Annotation Pipeline Rapidly Prioritizes Clinically Relevant Genes for Autism Spectrum Disorders

Olivia J. Veatch, PhD[1], Merlin G. Butler, MD, PhD, FFACMG[1], Sarah H. Elsea, PhD, FACMG[2], Beth A. Malow, MD[3], James S. Sutcliffe, PhD[4], Jason H. Moore, PhD, FACMI[5]

[1] Department of Psychiatry and Behavioral Sciences, University of Kansas Medical Center, Kansas City, Kansas; [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; [3]Sleep Disorders Division, Department of Neurology, Vanderbilt University Medical Center, Nashville, TN; [4]Vanderbilt Genetics Institute, Department of Molecular Physiology & Biophysics, Department of Psychiatry, Vanderbilt University Medical Center, Nashville, TN; [5]Department of Biostatistics, Epidemiology, & Informatics, University of Pennsylvania, Philadelphia, Pennsylvania

INTRODUCTION

- There are currently hundreds of genetic risk variants for autism spectrum disorders (ASDs), and the list is continually growing
- To understand how genetic data can be used to inform treatment, it is important to develop efficient ways to sort through association study and other related results to prioritize clinically relevant information
- We developed an automated translational bioinformatics pipeline using databases that incorporate evidence from multiple sources to identify genes that:
  1. Are associated with ASDs and expressed in the brain
  2. Have evidence for convergent biological function
  3. Encode proteins with evidence for functional consequences relevant to NDDs in mice
  4. Have known pathogenic variants
  5. Encode proteins with evidence for direct interactions with known pathogenic proteins
  6. Encode proteins targeted by pharmaceutical compounds

CANDIDATE GENE IDENTIFICATION AND FUNCTIONAL ANNOTATION

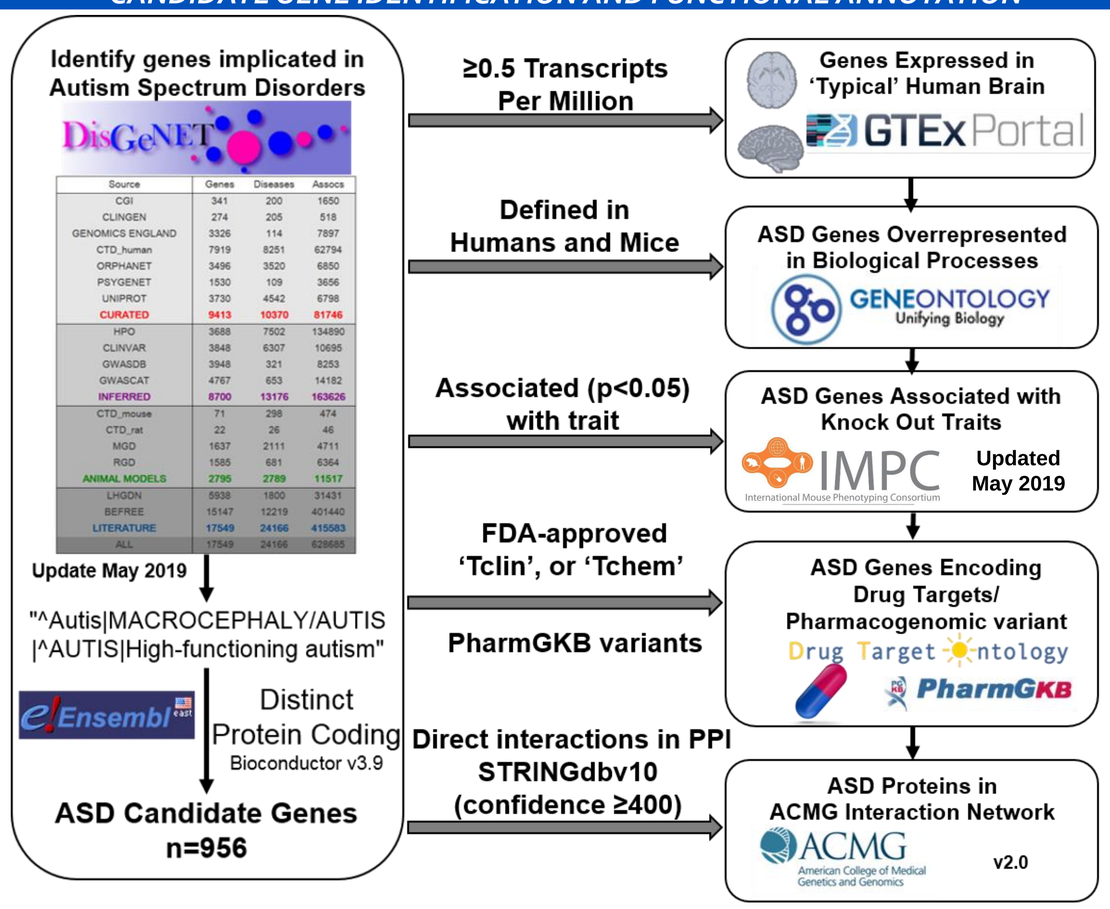

Figure 1. Candidate Gene Identification and Annotation Pipeline. All human protein coding genes with evidence for influencing risk for the most common NDD, autism spectrum disorder (ASD), were identified. The final set of ASD candidate genes was then annotated in the above pipeline using publicly available resources.

To determine if the pipeline was relevant for prioritizing NDD candidate genes, we selected 1,000 random sets of protein coding genes in humans of equal number to the ASD gene set and annotated in above pipeline. Results from annotations of ASD risk genes were compared to results from 1,000 random gene sets.

COMPARISONS BETWEEN ASD RISK GENES (n=861) AND RANDOM GENE SETS

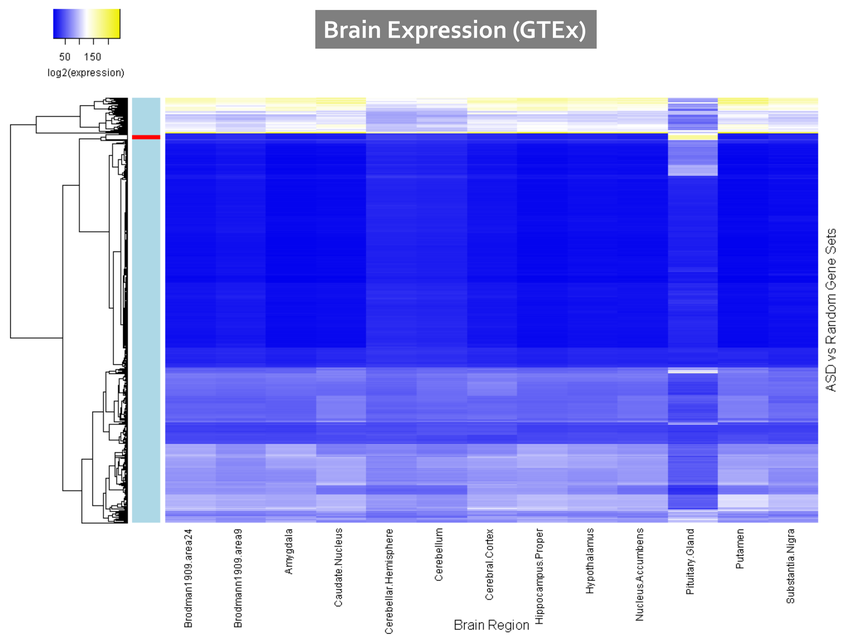

Figure 2. Brain Expression Profiles of ASD Risk Gene Set and Random Gene Sets. Shown are average expression levels for ASD and random gene sets, based on transcripts per million, for each brain region obtained from typically developing individuals in GTEx. Average expression of each gene set in the respective brain region are clustered based on Euclidean measures. The ASD risk gene set is highlighted in red on the y-axis.

COMPARISONS BETWEEN ASD RISK GENES AND RANDOM GENE SETS

Knock-out Mouse Phenotypes (KOMP)

Table 2. Proportions of ASD Genes Associated with Specific Top Level Mammalian Phenotypes.

Associated Mouse Trait	ASD (n)	Random (mean±sd)	X^2 (95%CI)	FDR
Growth phenotype	63	43.59±6.32	8.60 (0.05, 0.08)	1.01E-02
Nervous system phenotype	18	12.16±3.36	2.37 (0.01, 0.03)	1.85E-01
Embryo phenotype	17	18.46±4.18	0.05 (0.01, 0.03)	8.21E-01

Results comparing the number of ASD candidate genes that were associated with a mammalian phenotype reflecting overrepresented Gene Ontology Biological Processes defined in humans, compared to the average number of genes across 1,000 random sets. Chi-square test results are based on the proportion of genes in the ASD list (n=956) to the average proportion of genes across all random lists. Benjamini-Hochberg corrected p-values are included. sd=standard deviation.

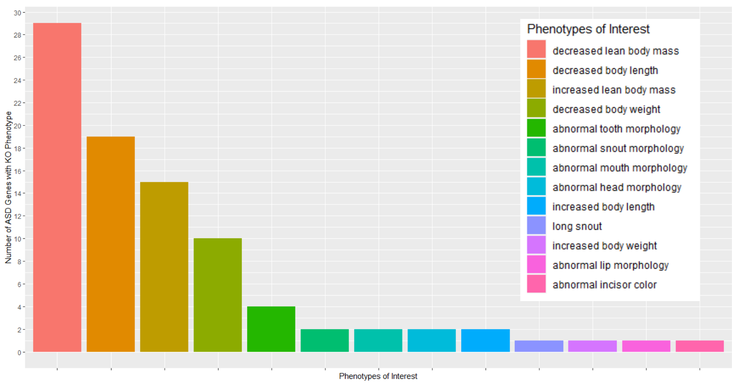

Figure 3. Frequency of ASD-related Genes Associated with Abnormal Postnatal Growth Phenotypes. Shown are the specific traits that were associated more often with knocking out ASD candidate genes in mice when compared to genes from random sets. Traits are presented in descending order based on the number of associated ASD candidate genes in KOs.

Encoding Drug Targets (Pharos)

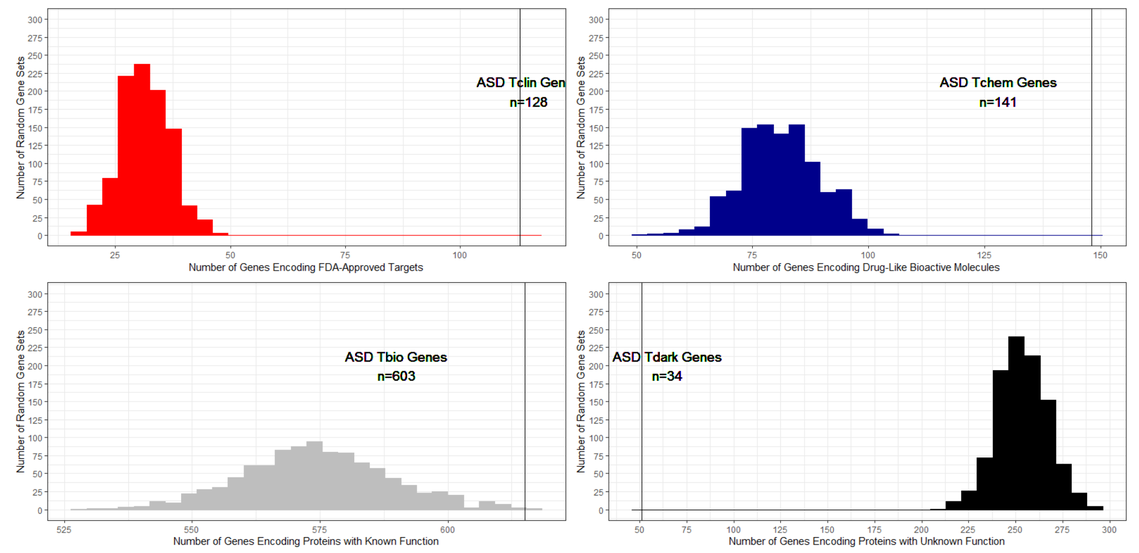

Figure 4. Comparisons between ASD Risk Genes and Random Gene Sets across Drug Development Levels. Shown are the number of ASD risk gene-encoded proteins at all categories of drug development, compared to the distribution of the number of proteins encoded by genes in each random set. Drug development levels include FDA-approved compound targets (Tclin), molecules with known properties similar to approved drug targets (Tchem), proteins with known biological or molecular functions but no known drug target properties (Tbio), proteins with relatively unknown function (Tdark).

Annotated in All Functional Categories

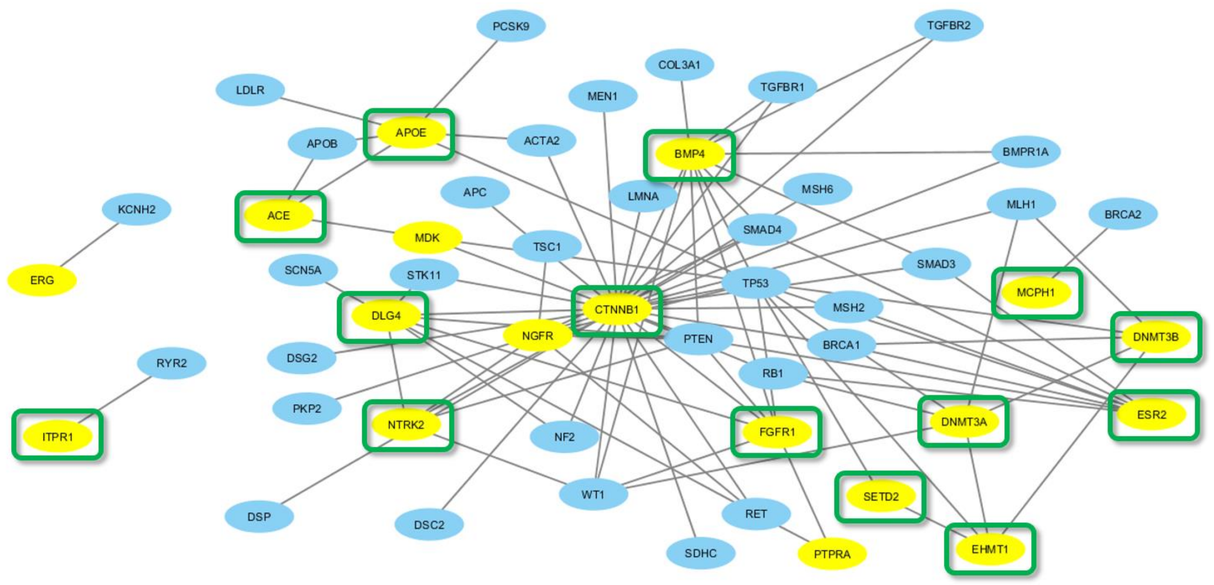

Figure 5. Top ASD Proteins and ACMG Proteins Interaction Network. Direct interactions predicted between proteins encoded by ASD risk genes annotated in all categories and proteins encoded by American College of Medical Genetics (ACMG) recommended actionable genes. Top ASD risk proteins are highlighted in yellow, ACMG-recommended proteins in blue and direct interactions for ASD proteins are indicated in red. ASD risk genes with pathogenic variants in ClinVar are indicated with green boxes.

CONCLUSIONS

- When compared to random gene sets:
  - More ASD risk genes were expressed in the brain and associated with relevant phenotypes in knock-out mice
  - More ASD risk genes encoded targets for FDA-approved drugs or bioactive molecules with drug-like properties, and encoded proteins with unknown function or had pharmacogenomic variants
  - Proteins encoded by ASD genes showed more evidence for direct interactions with proteins encoded by American College of Medical Genetics recommended clinically actionable genes
- Eighteen genes were prioritized based on functional evidence in all categories compared to an average of 8±2 genes each across all random gene sets
- Our automated approach helps to prioritize potentially actionable results from genetic studies, informing future work focused on supporting clinical decisions regarding the benefits of genetic testing for optimizing personalized approaches to treatment on a case-by-case basis

ACKNOWLEDGEMENTS

This work was directly supported by a grant, K01LM012870, from the National Institutes of Health, National Library of Medicine