Detecting the Presence of Autism

Sabrina Starr

Table of Contents

01

Introduction

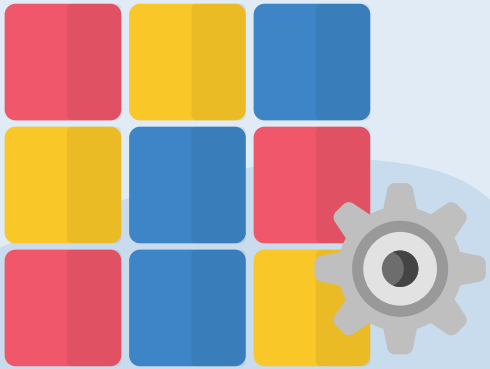
Introduction

NIH describes Autism Spectrum Disorder as being: "A neurological and developmental disorder that affects how people interact with others, communicate, learn, and behave. Although autism can be diagnosed at any age, it is described as a “developmental disorder” because symptoms generally appear in the first two years of life."

02

Background

Background

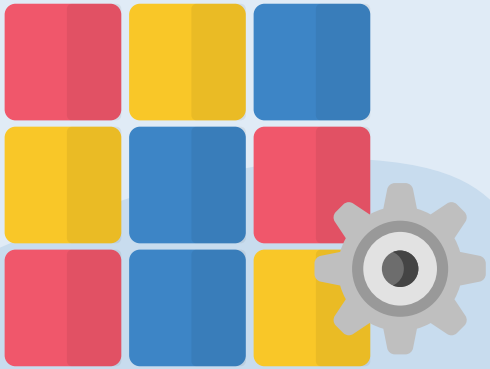


Initially, I decided to make a prediction app for rather or not an individual's offspring may have ASD based off of factors such as family history. This had a few obstacles:

- No way to prove my theory medically after extensive research into medical documentation.
- More sensical to predict the presence of autism in an individual than in their offspring.
- Already had data available to predict solely the present of ASD.

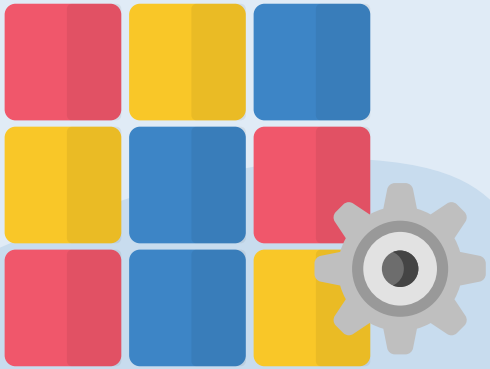
This led to a change in trajectory...

Background Continued ...



- My secondary data became my primary data (SFARI ASD dataset)
- Research was done to determine if this was enough data to predict the presence of autism in an individual– it was

… Background Continued …



Publications and Studies

- A twin study of autism in Denmark
- Methods for Detection of Parent-of-Origin Effects in Genetic Studies of Case-Parents Triads
- Clinical research: Short test finds autism traits in families
- Elevated Polygenic Burden for Autism Spectrum Disorder Is Associated With the Broad Autism Phenotype in Mothers of Individuals With Autism Spectrum Disorder
- Maternal lifestyle and environmental risk factors for autism spectrum disorders
- Molecular mechanisms underlying neurodevelopmental disorders, ADHD and autism
- There Are Two Types of “Social Emotional Agnosia” in Autism

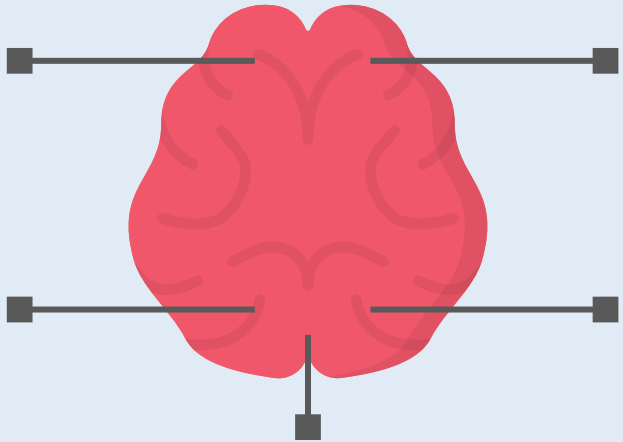
03

XGBoost

About XGBoost

Used very heavily in production environments.

Known for its speed and performance.



XGBClassifier import from the xgboost library-- which is scikit-learn's API for XGBClassification.

Is based on decision trees.

Stands for eXtreme Gradient Boosting.

04

Process

Process

Data

Load the data from SFARI database.

Features

Get features and labels.

Scale

Scale the features (and dummy the features).

Split

Split the dataset.

Build

Build the XGBClassifier.

Review

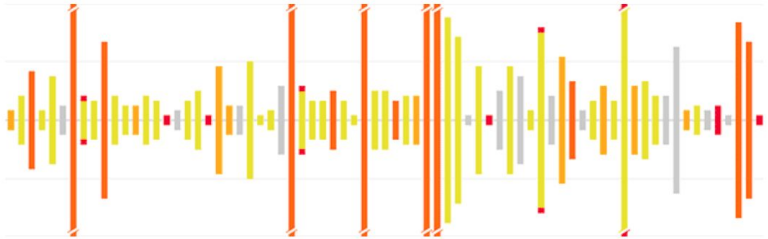
Calculate the model accuracy and EDA with Plotly.

SFARI Gene Database

LATEST NEWS SFARI Gene release notes for Q4 2021 **Learn More**

Human Gene

The Human Gene module provides researchers around the world with instant access to the most up-to-date information on all known human genes associated with autism spectrum disorders (ASD).



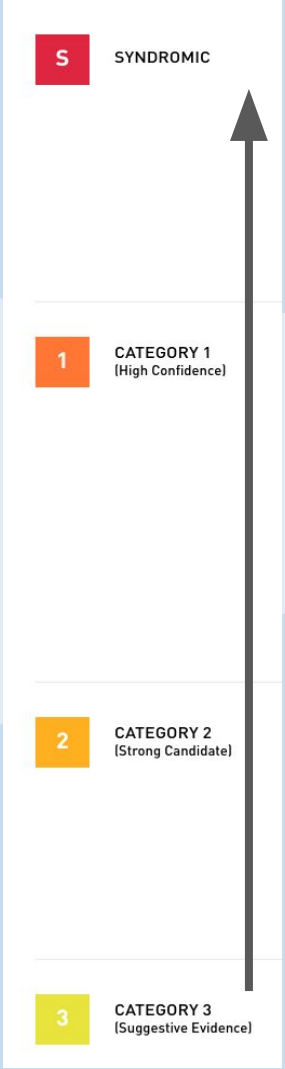
Used Data from SFARI

Only the ‘genes,’ dataset (for humans) was used, which had the following columns:

- Status (dropped)
- Gene Symbol (dropped)
- Gene Name (dropped)
- Ensembl ID (dropped)
- **Chromosome (dummied)**
- **Genetic Category (dummied)**
- **Gene Score (dummied)**
- **Syndromic (dummied, used as feature)**
- Number of Reports (dropped)

Genetic Categories

- **Rare**
  Rare single gene variants, disruptions/mutations, and submicroscopic deletions/duplications directly linked to ASD.
- **Syndromic**
  Genes implicated in syndromes in which a significant subpopulation develops symptoms of autism (examples: Angelman Syndrome, Fragile X Syndrome).
- **Association**
  Small risk-conferring common polymorphisms identified from genetic association studies in idiopathic ASD.
- **Functional**
  Functional candidates not yet genetically linked with ASD.



Gene Scores

Syndromic: The syndromic category includes mutations that are associated with a substantial degree of increased risk and consistently linked to additional characteristics not required for an ASD diagnosis.

Category 1 (High Confidence): Each of these genes has been clearly implicated in ASD.

Category 2 (Strong Candidate): Genes with two reported de novo likely-gene-disrupting mutations.

Category 3 (Suggestive Evidence): Genes with a single reported de novo likely-gene-disrupting mutation.

Other Unused Data from SFARI

- Gene cnvs
- Gene scores
- Mouse genes
- Mouse gene cnvs
- Mouse inbred
- Mouse induced
- Mouse rescue

CNVs

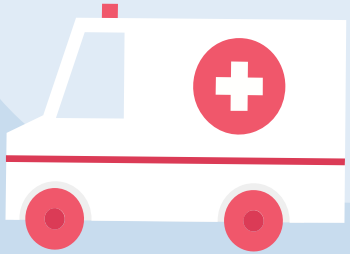
CNVs are segments of DNA, typically greater than 1,000 basepairs in length, that vary in number from person to person. These submicroscopic deletions and duplications are increasingly thought to be involved in the pathogenesis of a wide range of human diseases, including neuropsychiatric disorders such as ASD.

Inbred

Inbred models are studies on animals that exhibit a particular phenotype of interest without carrying a known genetic mutation.

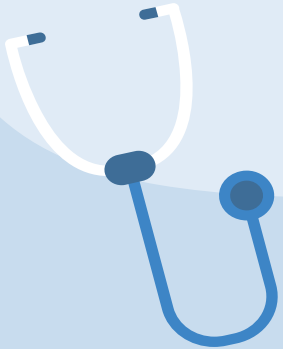
Induced

Induced models are studies on animals in which a particular trait or disorder has been purposefully induced with a biological or chemical agent.



Rescue

Rescue animal models are studies on animals that have received either pharmaceutical, genetic, or cell transplant treatments during the course of the study. These treatments have the effect of ameliorating one or more of the associated phenotypes.



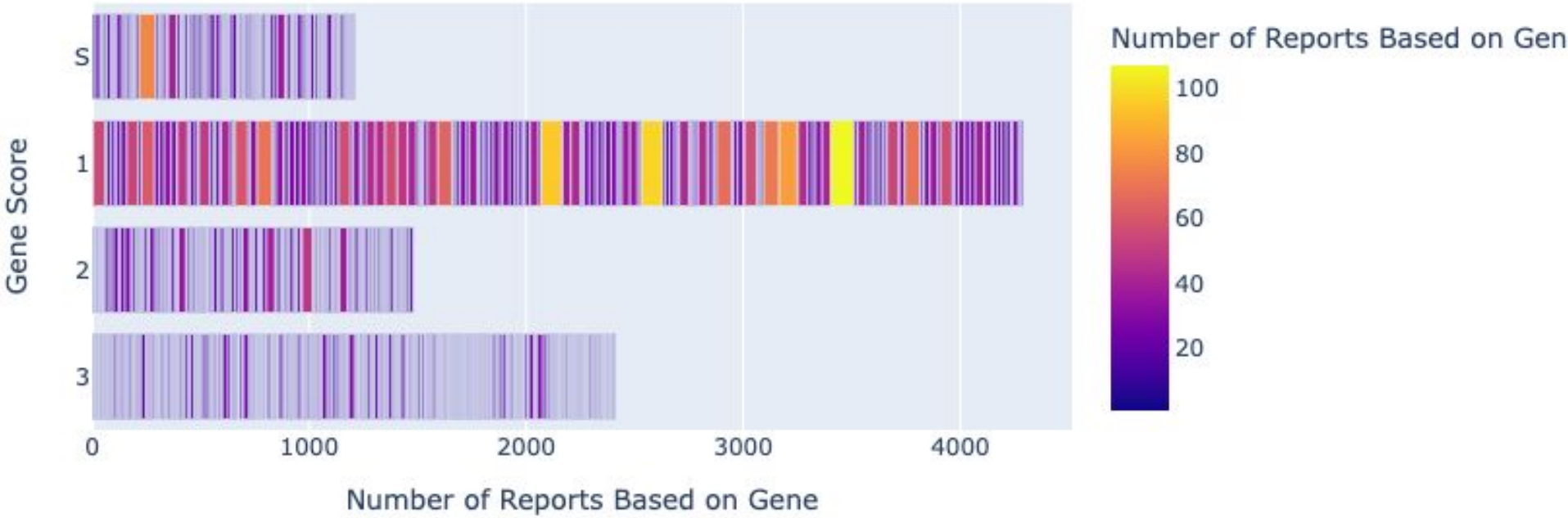
05

Discoveries

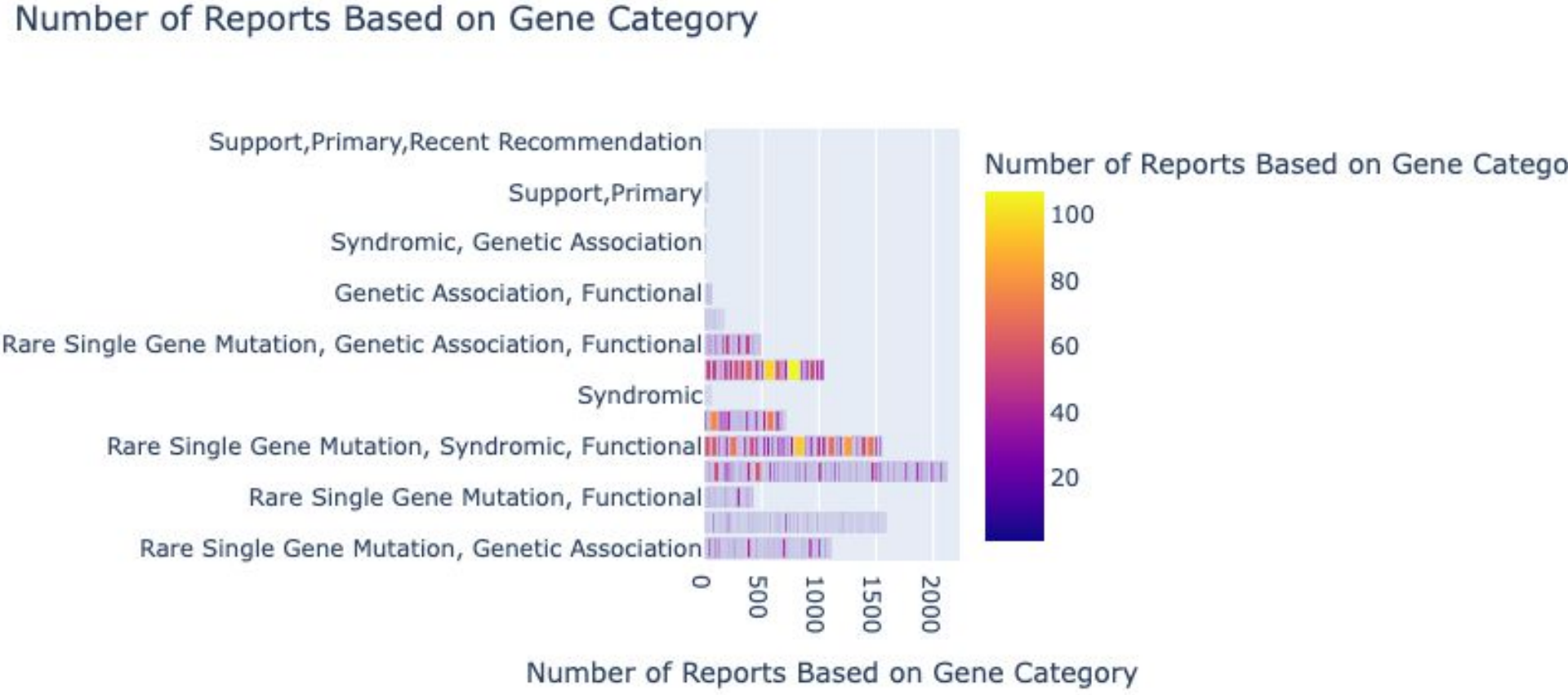
Overall

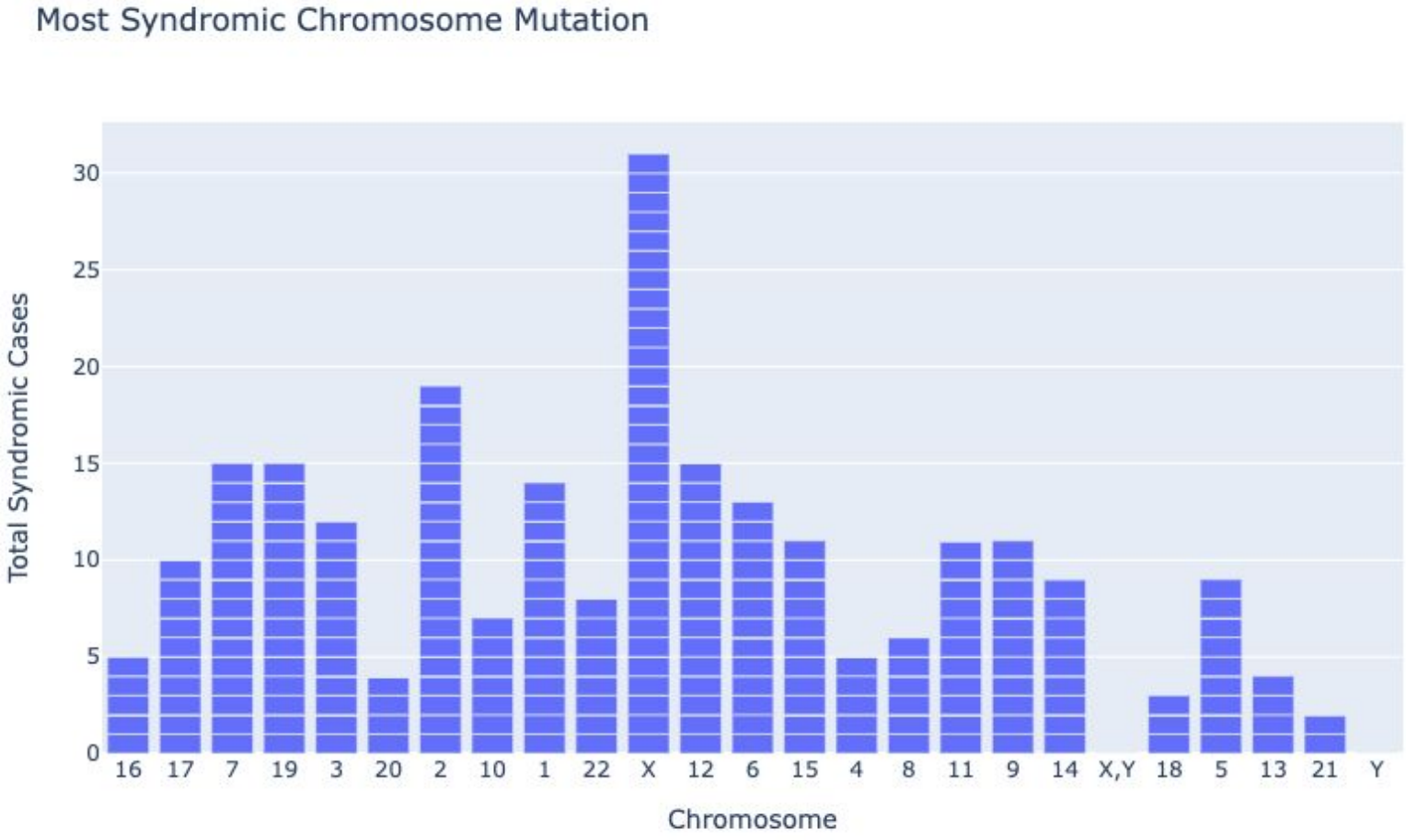
There were 793 syndromic cases, and 239 non-syndromic cases.

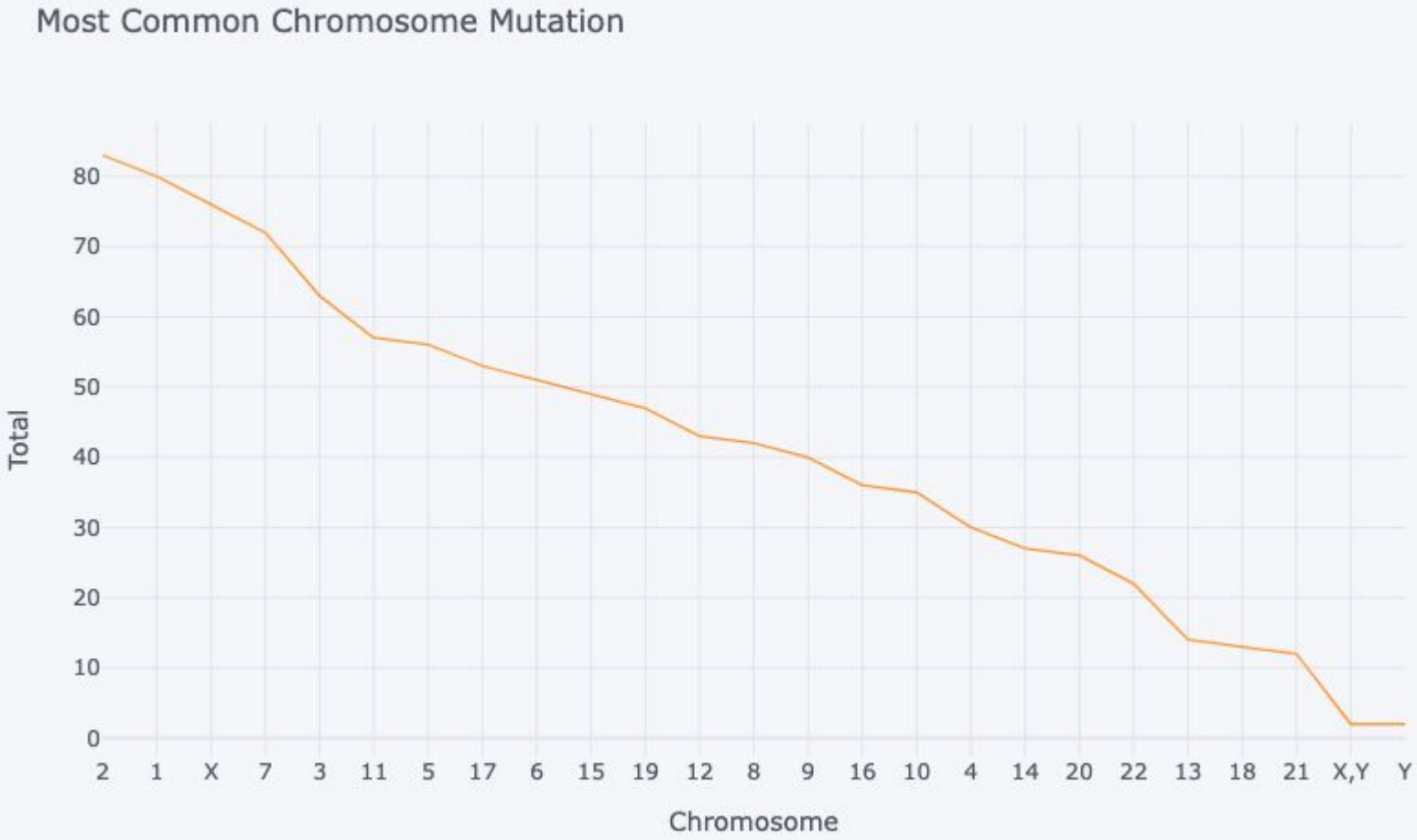
There were 477 score 3 genes, 206 score 1 genes, 198 score 2 genes, and 150 actual syndromic genes.


Number of Reports Based on Gene vs. Gene Score
Gene Score
S
1
2
3
0
1000
2000
3000
4000
Number of Reports Based on Gene
Number of Reports Based on Gen
100
80
60
40
20

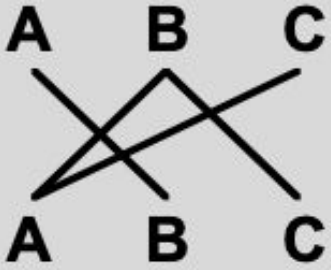

Number of Genes Based on Gene-score
500
400
300
200
100
0
3
1
2
S
Gene-score
Total Number of Gene
400
300
200


Number of Reports Based on Gene Category
Support,Primary,Recent Recommendation
Support,Primary
Syndromic, Genetic Association
Genetic Association, Functional
Rare Single Gene Mutation, Genetic Association, Functional
Syndromic
Rare Single Gene Mutation, Syndromic, Functional
Rare Single Gene Mutation, Functional
Rare Single Gene Mutation, Genetic Association
0
500
1000
1500
2000
Number of Reports Based on Gene Category
Number of Reports Based on Gene Catego
100
80
60
40
20


Most Syndromic Chromosome Mutation
Total Syndromic Cases
0
5
10
15
20
25
30
16
17
7
19
3
20
2
10
1
22
X
12
6
15
4
8
11
9
14
X,Y
18
5
13
21
Y
Chromosome


Most Common Chromosome Mutation
Total
80
70
60
50
40
30
20
10
0
2
1
X
7
3
11
5
17
6
15
19
12
8
9
16
10
4
14
20
22
13
18
21
X,Y
Y
Chromosome

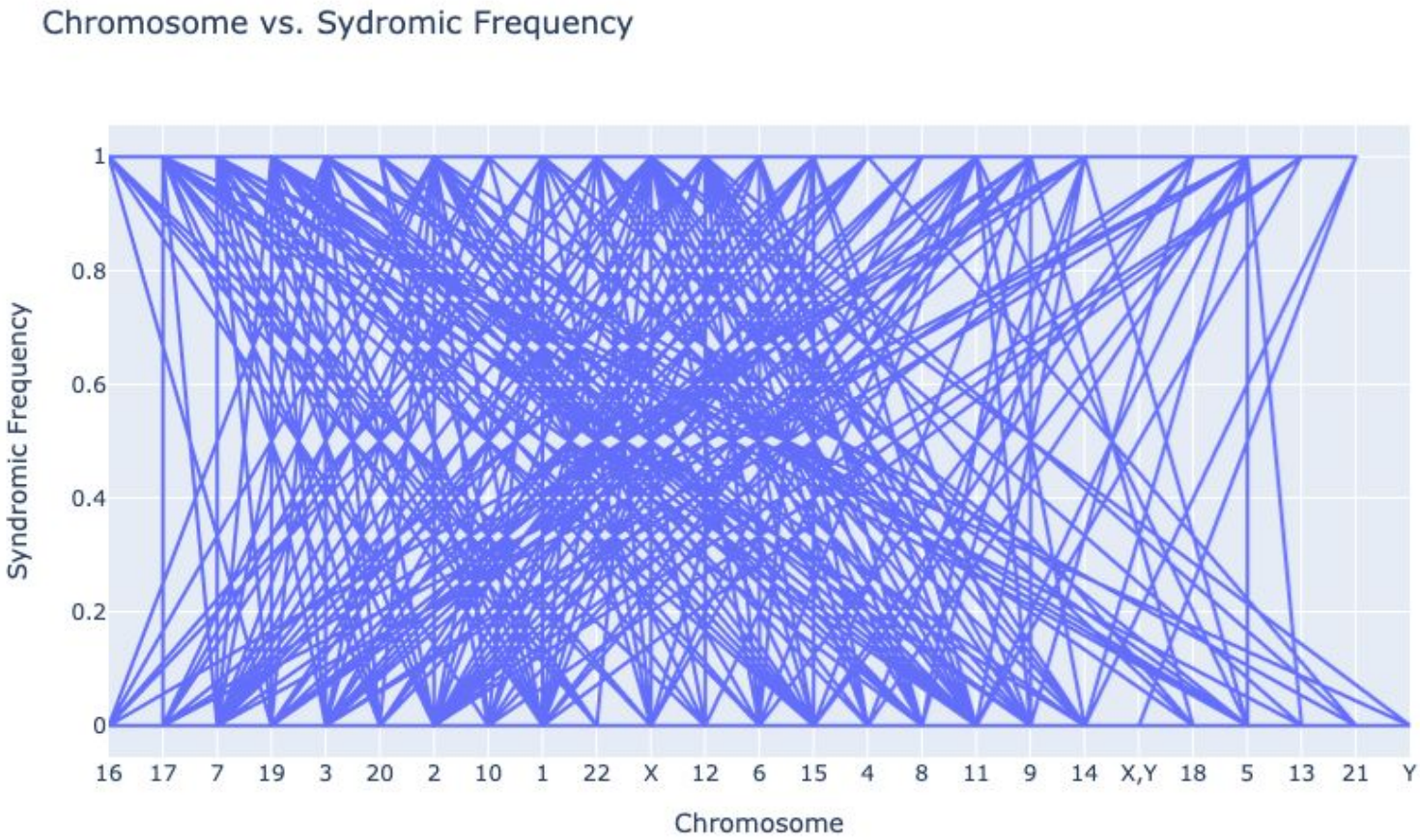
Wiring and rules



inputs	2	1	1
rule	4	2	2

Basis for rules:

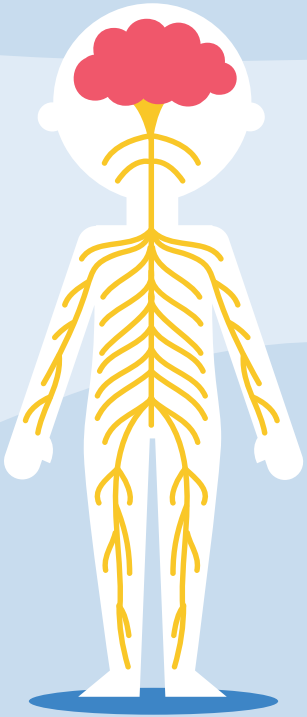
1. A activates B
2. B activates A and C
3. C inhibits A


Chromosome vs. Sydromic Frequency
Syndromic Frequency
1
0.8
0.6
0.4
0.2
0
16 17 7 19 3 20 2 10 1 22 X 12 6 15 4 8 11 9 14 X,Y 18 5 13 21 Y
Chromosome

06

Closing

Discovery Summary



The overall most reported gene score is gene score number 1 (high confidence).

The most common gene score in those with ASD based on this dataset is gene score 3 (suggestive evidence).

The main gene mutated in individuals with **syndromic** ASD tends to be the X gene.

There are multiple chromosomes with multiple syndromic chromosome frequencies.

The most common **overall** mutated gene in the dataset is gene number 2.

The rare, syndromic, functional gene category had the highest number of reports.

Future Exploration

23andMe

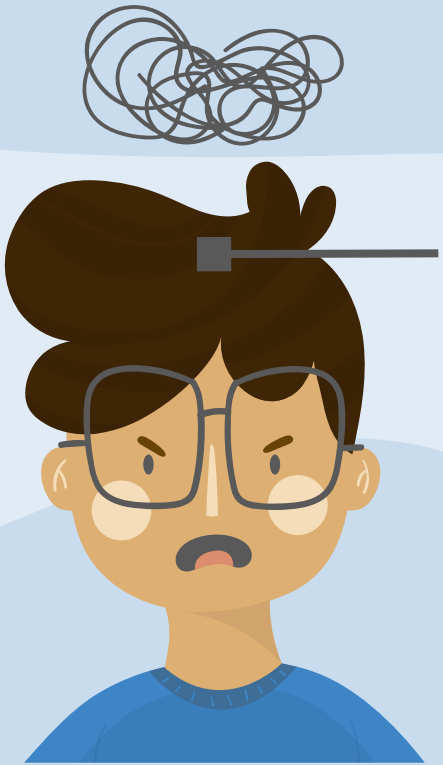
Building an app that could use gene data from 23andMe to tell someone rather or not they have ASD.

Ancestry.com

Building an app that could use gene data from Ancestry.com to tell someone rather or not they have ASD.

Further EDA

Answering deeper questions about Autism Spectrum Disorder that came to surface.



Sources

https://pubmed.ncbi.nlm.nih.gov/27516006/
https://www.kaggle.com/jayachandruk/autism-patients-mutated-gene-dataset
https://www.kaggle.com/mdsakiburhasan/latino-mothers-and-their-newborn/metadata
https://pubmed.ncbi.nlm.nih.gov/2745591/
https://theisaacs22.wordpress.com/2018/06/30/there-are-two-types-of-social-emotiona
gnosia-in-autism/
file:///Users/breezehomefolder/Downloads/cdc_113924_DS1.pdf
https://www.spectrumnews.org/news/clinical-research-short-test-finds-autism-traits-in-
milies/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3997376/pdf/dyt282.pdf
https://www.biologicalpsychiatryjournal.com/article/S0006-3223(20)31911-9/fulltext#supp
mentaryMaterial
https://data-flair.training/blogs/data-science-project-ideas/
https://data-flair.training/blogs/python-machine-learning-project-detecting-parkinson-
ease/
https://gene.sfari.org/user-guide/glossary/
https://gene.sfari.org/about-gene-scoring/
https://psb.stanford.edu/previous/psb99/genetutorial.pdf
https://www.sciencedirect.com/science/article/pii/S0002929707637471
https://www.genome.gov/about-genomics/fact-sheets/X-Chromosome-facts#:~:text=Th
20X%20chromosome%20determines%20your,a%20certain%20breed%20of%20cat.

Thank you!
Questions?

LinkedIn:

https://www.linkedin.com/in/s-starr/


slidesgo