SIMONS FOUNDATION

Large Genomic Data Sets in Autism Research: SFARI DOMA Practices

Natalia Volfovsky

November 16, 2017

SFARI

SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE

https://www.sfari.org

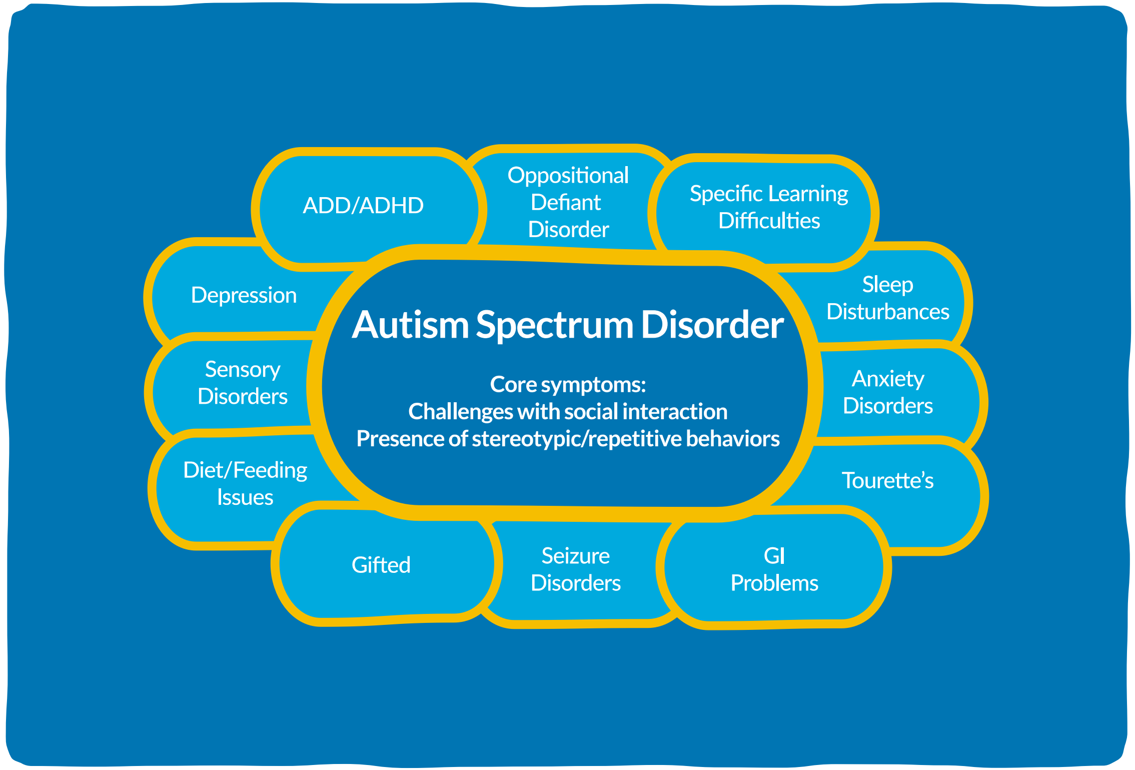
Mission: improve the understanding, diagnosis and treatment of autism spectrum disorders (ASD) by funding innovative research of the highest quality and relevance

Identify risk factors whether genetic, environmental or epidemiological.

Use non-human organisms to understand how these risk factors alter brain function and animal behavior.

Promote preclinical and clinical investigations to improve autism diagnosis & therapy.

Autism spectrum disorder(ASD) is extremely heterogeneous

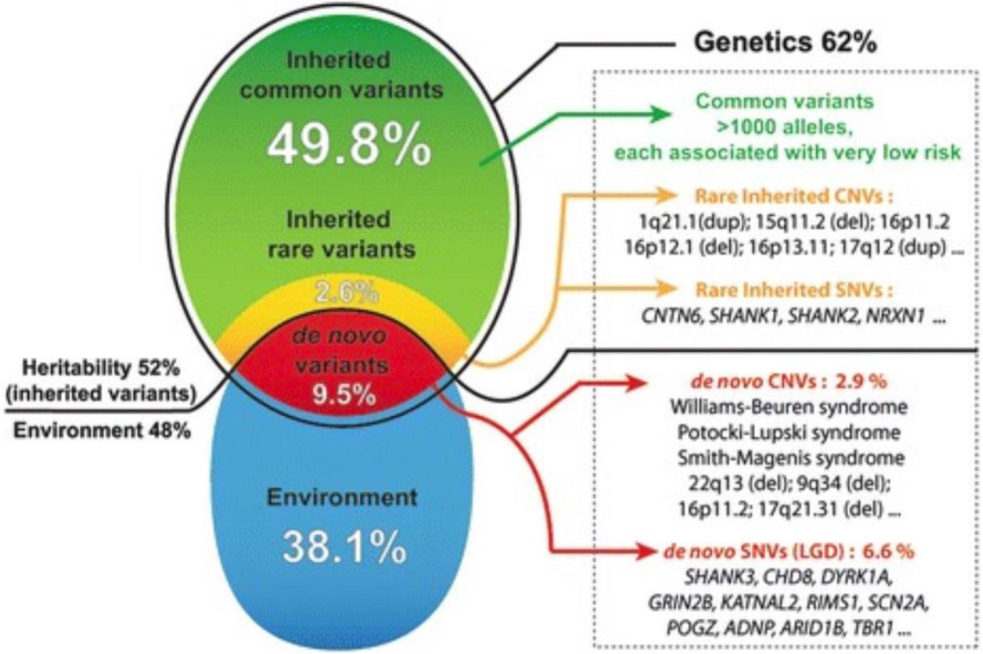


Autism Spectrum Disorder

Core symptoms:
Challenges with social interaction
Presence of stereotypic/repetitive behaviors

Pam Feliciano, 2017

Autism risk is complex



Common variants
>1000 alleles,
each associated with very low risk

Rare Inherited CNVs :
1q21.1(dup); 15q11.2 (del); 16p11.2
16p12.1 (del); 16p13.11; 17q12 (dup) ...

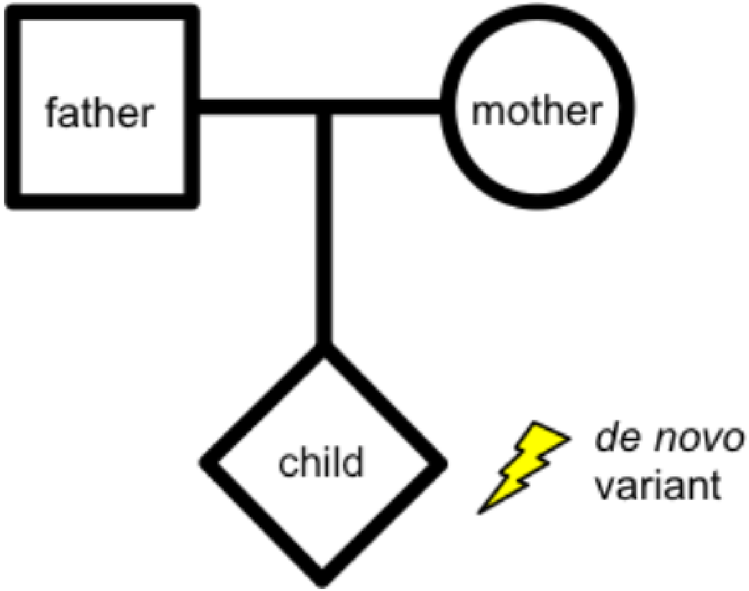
Rare Inherited SNVs :
CNTN6, SHANK1, SHANK2, NRXN1 ...

de novo CNVs : 2.9 %
Williams-Beuren syndrome
Potocki-Lupski syndrome
Smith-Magenis syndrome
22q13 (del); 9q34 (del);
16p11.2; 17q21.31 (del) ...

de novo SNVs (LGD) : 6.6 %
SHANK3, CHD8, DYRK1A,
GRIN2B, KATNAL2, RIMS1, SCN2A,
POGZ, ADNP, ARID1B, TBR1 ...

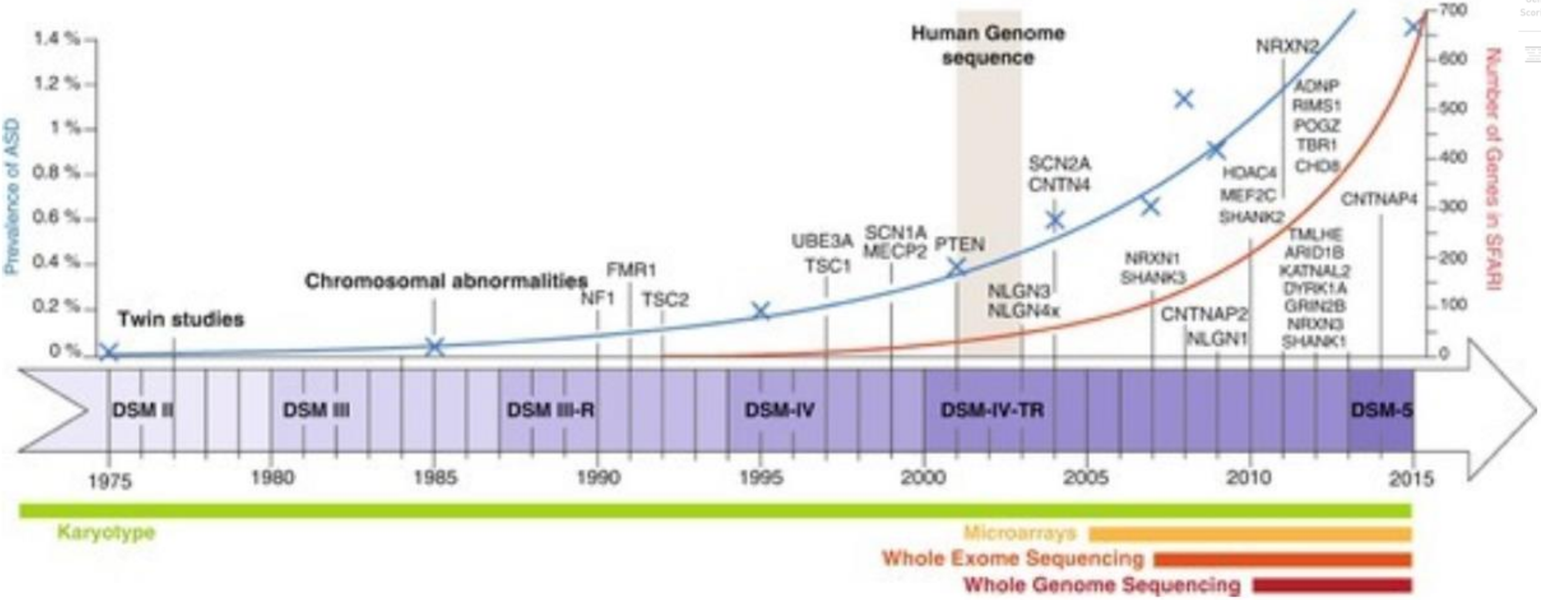
Huguet, Benabou, and Bourgeron, 2016

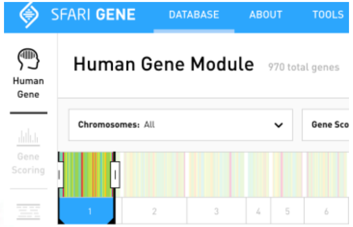
De Novo Mutations



- Genotype is present in neither parent, but usually heterozygous in the child
- Spontaneous genetic mutations

The history of the genetics of autism





SFARI GENE DATABASE ABOUT TOOLS

Human Gene Module 970 total genes

Human Gene

Gene Scoring

Chromosomes: All

Gene Sco

gene.sfari.org

Huguet, Benabou, and Bourgeron, 2016

SIMONS FOUNDATION

SFARI Research Cohorts

https://www.sfari.org

Simons Simplex Collection (SSC)

~10,000 individuals

SIMONS VIP

Simons Variation in Individuals Project

~ 1,500 individuals

SPARK

Igniting autism research
Improving lives

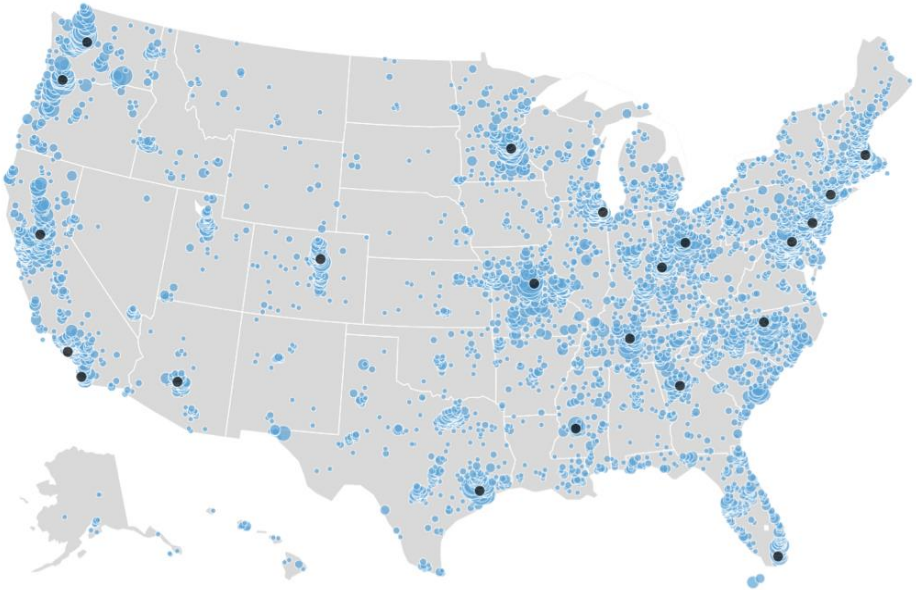
Simons Foundation Powering Autism Research through Knowledge

SPARK
Igniting autism research
Improving lives

https://sparkforautism.org

Recruit, engage, and retain 50
individuals with ASD
and their biological family
members to:

- identify causes of ASD
- enable genotype-driven res
- find better treatments to im
lives



Individuals with autism
n= 28,501

Family members
n= 52,070

Usable saliva samples
n= 29,374

Complete trios + DNA
n = 4,294 trios

Unaffected sibs + DNA n= 2,086

Pam Feliciano, October 2017

SIMONS FOUNDATION

Danish Neonatal Screening Biobank

The DNSB houses blood spots from all individuals born in Denmark Since 1981 (more than 2 million individuals). These samples can be linked to first-degree and other relatives, and there are now at least 15,000 diagnosed cases of autism in the biobank, as well as 15,000 cases of ADHD and 25,000 cohort-matched controls.

Exome sequencing of blood spot-derived DNA:
collaboration between Mark Daly (Broad Institute) and Preben Mortensen (Aarhus University, Denmark).

STANLEY CENTER
FOR PSYCHIATRIC RESEARCH
AT BROAD INSTITUTE

iPSYCH

LUNDBECK FOUNDATION

SIMONS FOUNDATION

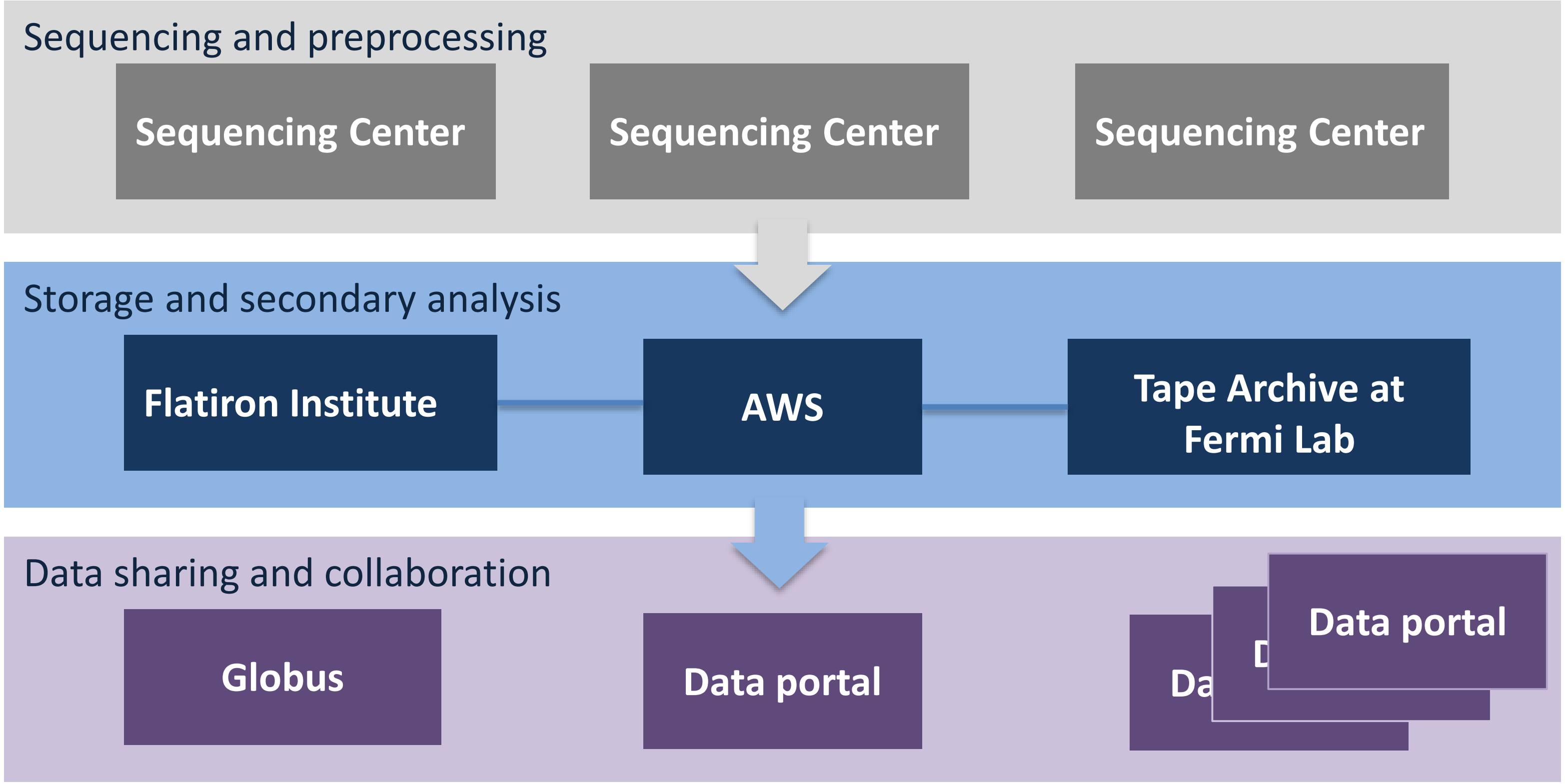
Data Sharing Policy

SFARI aims to support reproducible scientific research of the highest quality. This can be greatly facilitated by making all raw data, analysis methods, and, when applicable, computer code, available to the research community as quickly and transparently as possible.

Data generated from collections is considered “community resources” and, as such, all these data, including, but not limited to, alignment files and variant calls, will be made available to the entire research community, pre-publication, in real-time, as they are produced.

Besides being subject to any limitations by consent and IRB protocol, a publication embargo will also be enforced for 6 months after the defined project is complete while foundation-sponsored analyses are underway.

SFARI Data Overview



Sequencing and preprocessing

Sequencing Center

Sequencing Center

Sequencing Center

Storage and secondary analysis

Flatiron Institute

AWS

Tape Archive at Fermi Lab

Data sharing and collaboration

Globus

Data portal

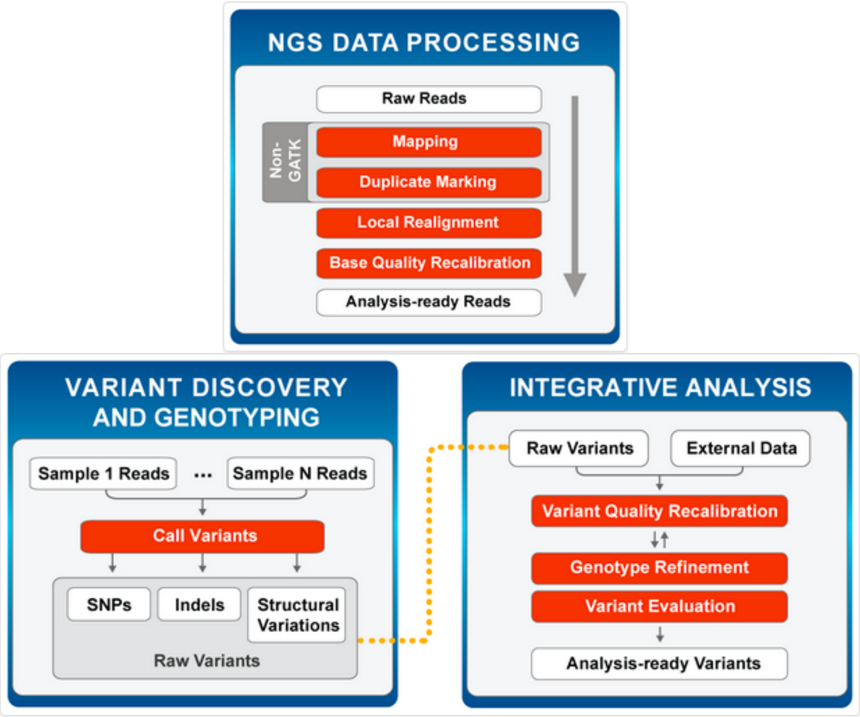
Da

D

Data portal

Data Analysis

Genomic Analysis Workflows



NGS DATA PROCESSING

Raw Reads

Non-GATK

Mapping

Duplicate Marking

Local Realignment

Base Quality Recalibration

Analysis-ready Reads

VARIANT DISCOVERY AND GENOTYPING

Sample 1 Reads ... Sample N Reads

Call Variants

SNPs

Indels

Structural Variations

Raw Variants

INTEGRATIVE ANALYSIS

Raw Variants

External Data

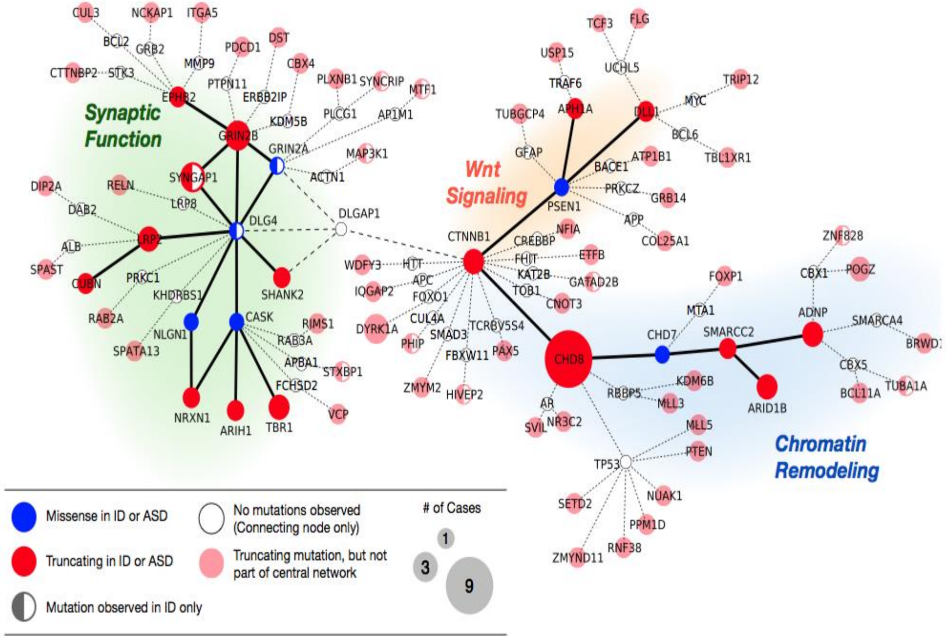
Variant Quality Recalibration

Genotype Refinement

Variant Evaluation

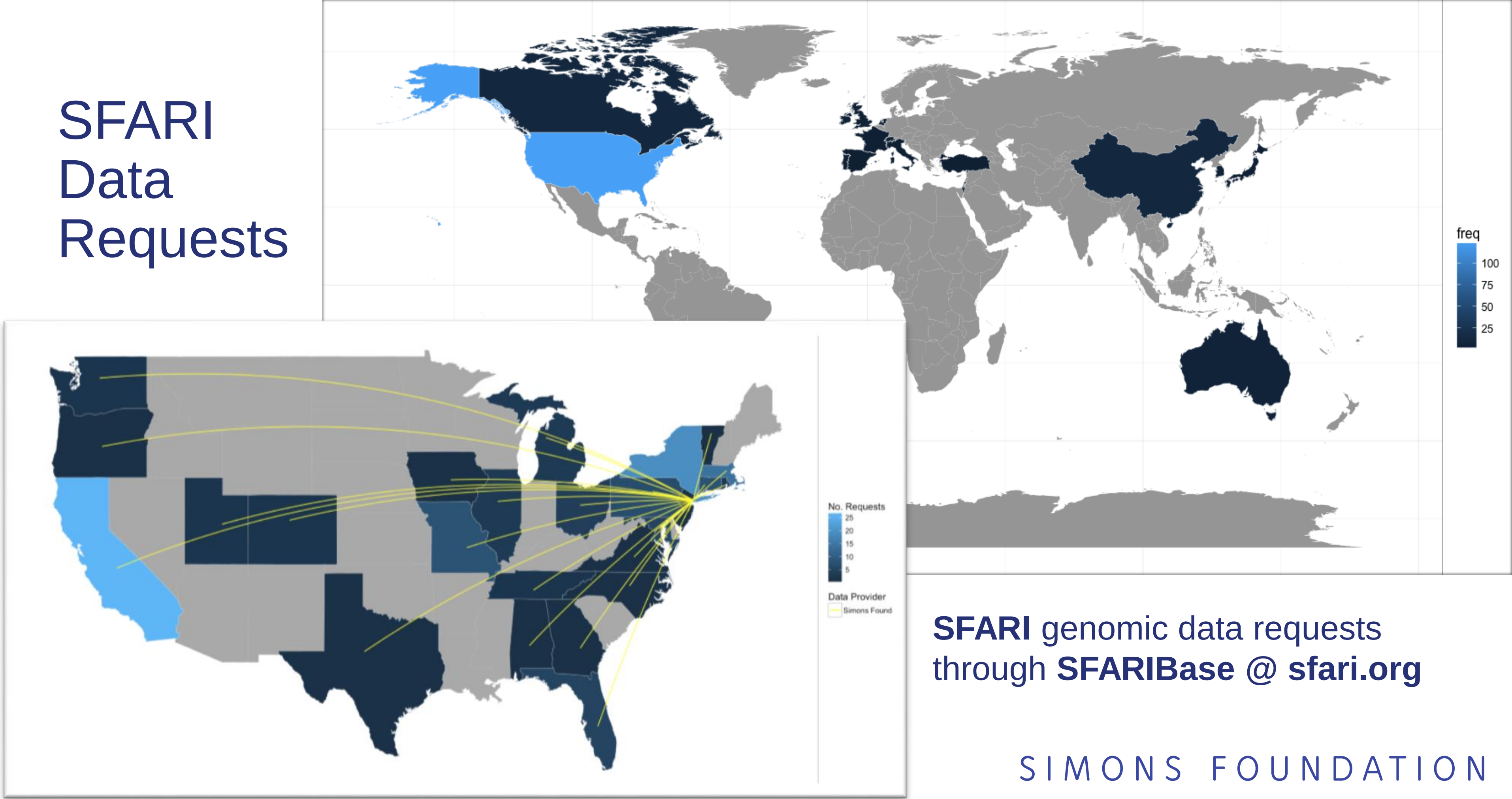
Analysis-ready Variants

Interpretation of Genomic Results



SIMONS FOUNDATION

SFARI Data Requests



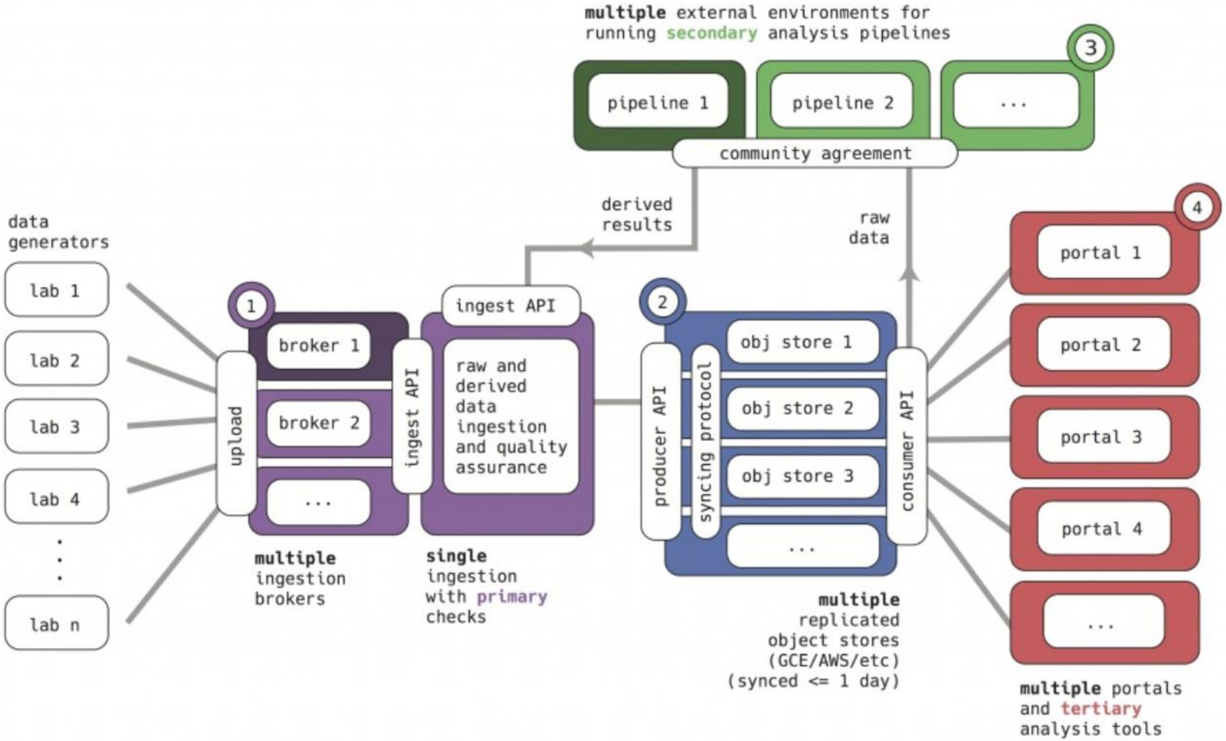
SFARI genomic data requests through **SFARIBase @ sfari.org**

SIMONS FOUNDATION

Challenges of Data Sharing

- Data management
- Collaboration utilities
- Users' ability to deploy their own analysis tools
- Ability to harmonize data from different projects
- Ability to harmonize data from different archives of the same project
    - AWS, Google, local servers
- Price, speed
    - storage, data transfers, computin
- Support of international collaborations and ability to deal with data governance across countries, including local restrictions

Human Cell Atlas (HCA) Data Coordination Portal

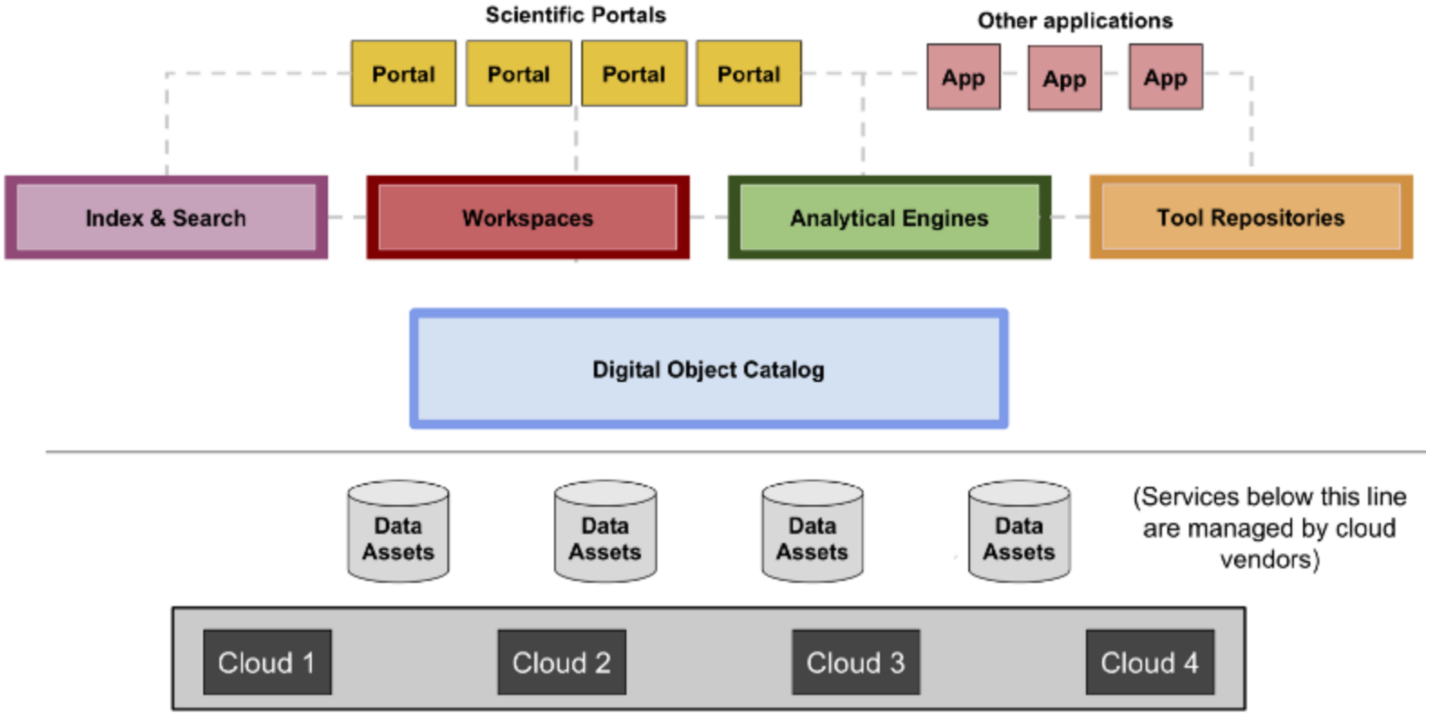


multiple external environments for running secondary analysis pipelines
3
pipeline 1
pipeline 2
...
community agreement
derived results
raw data
data generators
lab 1
lab 2
lab 3
lab 4
lab n
upload
1
broker 1
broker 2
...
ingest API
ingest API
raw and derived data ingestion and quality assurance
2
producer API
syncing protocol
obj store 1
obj store 2
obj store 3
...
consumer API
4
portal 1
portal 2
portal 3
portal 4
...
multiple ingestion brokers
single ingestion with primary checks
multiple replicated object stores (GCE/AWS/etc) (synced <= 1 day)
multiple portals and tertiary analysis tools

HUMAN CELL ATLAS

SIMONS FOUNDATION

Global Alliance for Genomics and Health (GA4GH): Data Biosphere



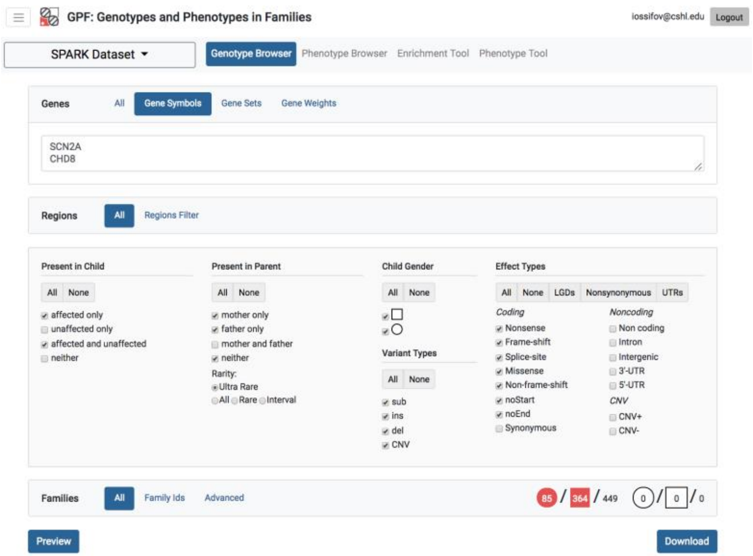
Genomics data portals

- Support
  - **collaboration**
  - **data sharing and**
  - **reproducibility** in scientific research
- Technology
  - **API access**
  - **docker containers and**
  - **standardized workflow description languages**
- Leverage experiences of scientific data management communities
- Invest in **the implementations of science specific standards**

Genome and Phenotype Tool (GPF)

https://gpf.sfari.org/

Query



GPF: Genotypes and Phenotypes in Families

iossifov@cshl.edu Logout

SPARK Dataset ▾ | Genotype Browser | Phenotype Browser | Enrichment Tool | Phenotype Tool

Genes All | Gene Symbols | Gene Sets | Gene Weights

SCN2A
CHD8

Regions All | Regions Filter

Present in Child All | None

- ☑ affected only
- ☐ unaffected only
- ☑ affected and unaffected
- ☐ neither

Present in Parent All | None

- ☑ mother only
- ☑ father only
- ☐ mother and father
- ☑ neither

Rarity:
● Ultra Rare
○ All ○ Rare ○ Interval

Child Gender All | None

- ☑ □
- ☑ ○

Variant Types All | None

- ☑ sub
- ☑ ins
- ☑ del
- ☑ CNV

Effect Types All | None | LGDs | Nonsynonymous | UTRs

Coding

- ☑ Nonsense
- ☑ Frame-shift
- ☑ Splice-site
- ☑ Missense
- ☑ Non-frame-shift
- ☑ noStart
- ☑ noEnd
- ☐ Synonymous

Noncoding

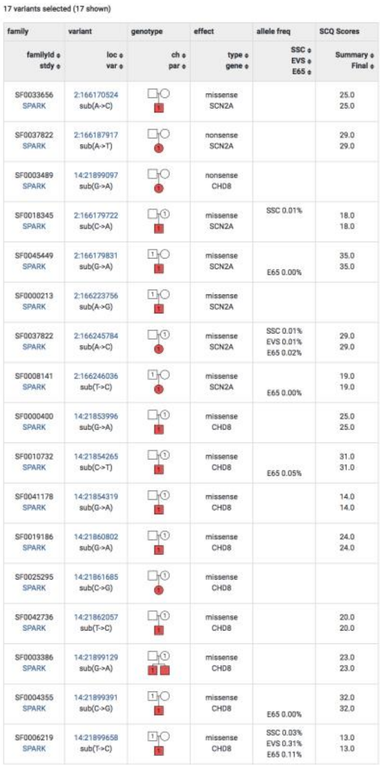
- ☐ Non coding
- ☐ Intron
- ☐ Intergenic
- ☐ 3'-UTR
- ☐ 5'-UTR

CNV

- ☐ CNV+
- ☐ CNV-

Families All | Family Ids | Advanced 85 / 364 / 449 0 / 0 / 0

Preview Download



17 variants selected (17 shown)

family	variant	genotype	effect	allele freq	SCQ Scores
familyId ⬍ stdy ⬍	loc ⬍ var ⬍	ch ⬍ par ⬍	type ⬍ gene ⬍	SSC ⬍ EVS ⬍ E65 ⬍	Summary ⬍ Final ⬍
SF0033656 SPARK	2:166170524 sub(A->C)		missense SCN2A		25.0 25.0
SF0037822 SPARK	2:166187917 sub(A->T)		nonsense SCN2A		29.0 29.0
SF0003489 SPARK	14:21899097 sub(G->A)		nonsense CHD8		
SF0018345 SPARK	2:166179722 sub(C->A)		missense SCN2A	SSC 0.01%	18.0 18.0
SF0045449 SPARK	2:166179831 sub(G->A)		missense SCN2A	E65 0.00%	35.0 35.0
SF0000213 SPARK	2:166223756 sub(A->G)		missense SCN2A		
SF0037822 SPARK	2:166245784 sub(A->C)		missense SCN2A	SSC 0.01% EVS 0.01% E65 0.02%	29.0 29.0
SF0008141 SPARK	2:166246036 sub(T->C)		missense SCN2A	E65 0.00%	19.0 19.0
SF0000400 SPARK	14:21853996 sub(G->A)		missense CHD8		25.0 25.0
SF0010732 SPARK	14:21854265 sub(C->T)		missense CHD8	E65 0.05%	31.0 31.0
SF0041178 SPARK	14:21854319 sub(G->A)		missense CHD8		14.0 14.0
SF0019186 SPARK	14:21860802 sub(G->A)		missense CHD8		24.0 24.0
SF0025295 SPARK	14:21861685 sub(C->G)		missense CHD8		
SF0042736 SPARK	14:21862057 sub(T->C)		missense CHD8		20.0 20.0
SF0003386 SPARK	14:21899129 sub(G->A)		missense CHD8		23.0 23.0
SF0004355 SPARK	14:21899391 sub(C->G)		missense CHD8	E65 0.00%	32.0 32.0
SF0006219 SPARK	14:21899658 sub(T->C)		missense CHD8	SSC 0.03% EVS 0.31% E65 0.11%	13.0 13.0

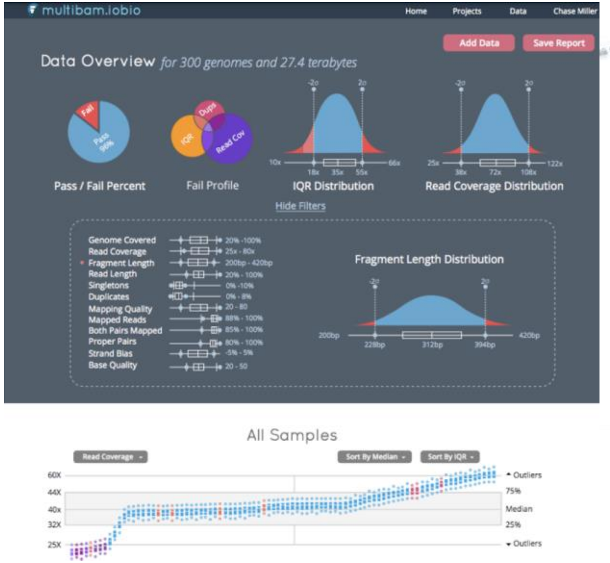
Features:

- SPARK Content:
  - De novo variants
  - Rare and common transmitted variants
  - Phenotypic data
- Integration:
  - SSC exome and whole-genome
  - VIP
- Query: Interface
  - By gene
  - By set of genes (i.e. by pathway)
  - By predicted variant effect (i.e. LGDs)
  - By variant frequency
  - By transmission pattern (i.e. de novo or transmitted from mother)
  - By phenotypic properties (i.e. affected children with SCQ score larger than 20)
- Analysis tools:
  - Enrichment of de novo variant in a given gene set.

SIMONS FOUNDATION

SFARI-IOBIO Data Federation Project

bam.iobio: http://bam.iobio.io
vcf.iobio: http://vcf.iobio.io
gene.iobio: http://gene.iobio.io



Data Overview *for 300 genomes and 27.4 terabytes*

Pass / Fail Percent

Fail Profile

IQR Distribution

Read Coverage Distribution

Hide Filters

Genome Covered	20% -100%
Read Coverage	25x - 80x
Fragment Length	200bp - 420bp
Read Length	20% - 100%
Singletons	0% -10%
Duplicates	0% - 8%
Mapping Quality	20 - 80
Mapped Reads	88% - 100%
Both Pairs Mapped	85% - 100%
Proper Pairs	80% - 100%
Strand Bias	-5% - 5%
Base Quality	20 - 50

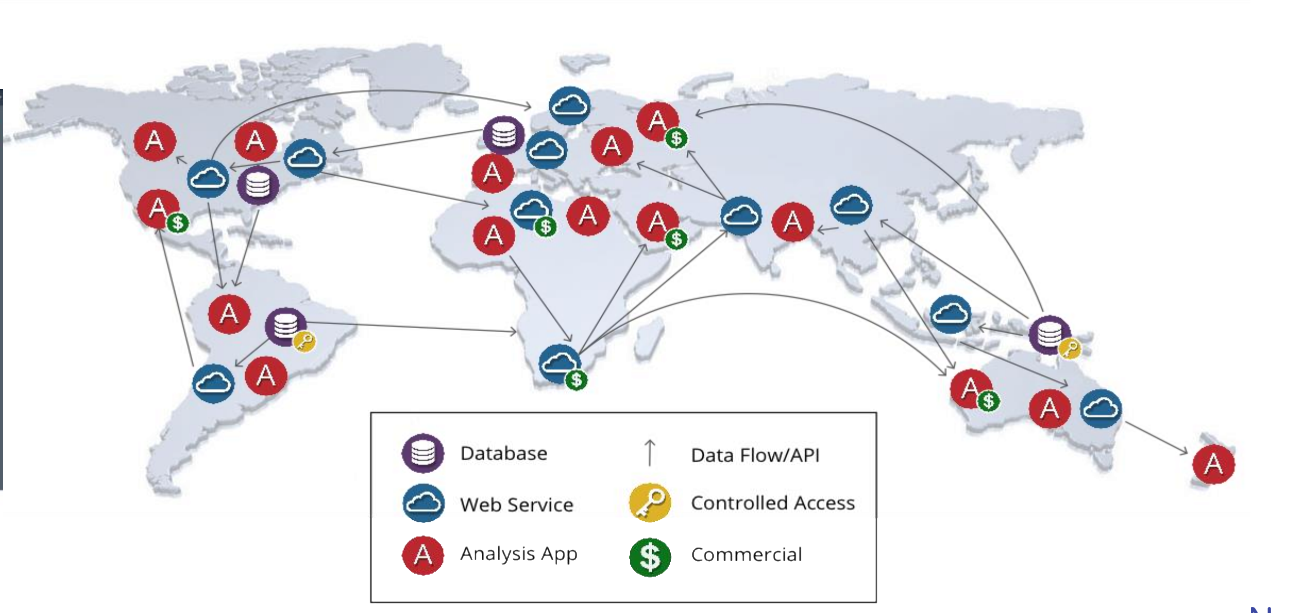
Fragment Length Distribution

All Samples

Data Hub
register + find databases
hub.iobio/data

Service Hub
register + publish + find services
hub.iobio/services

App Hub
register + find apps
hub.iobio/apps



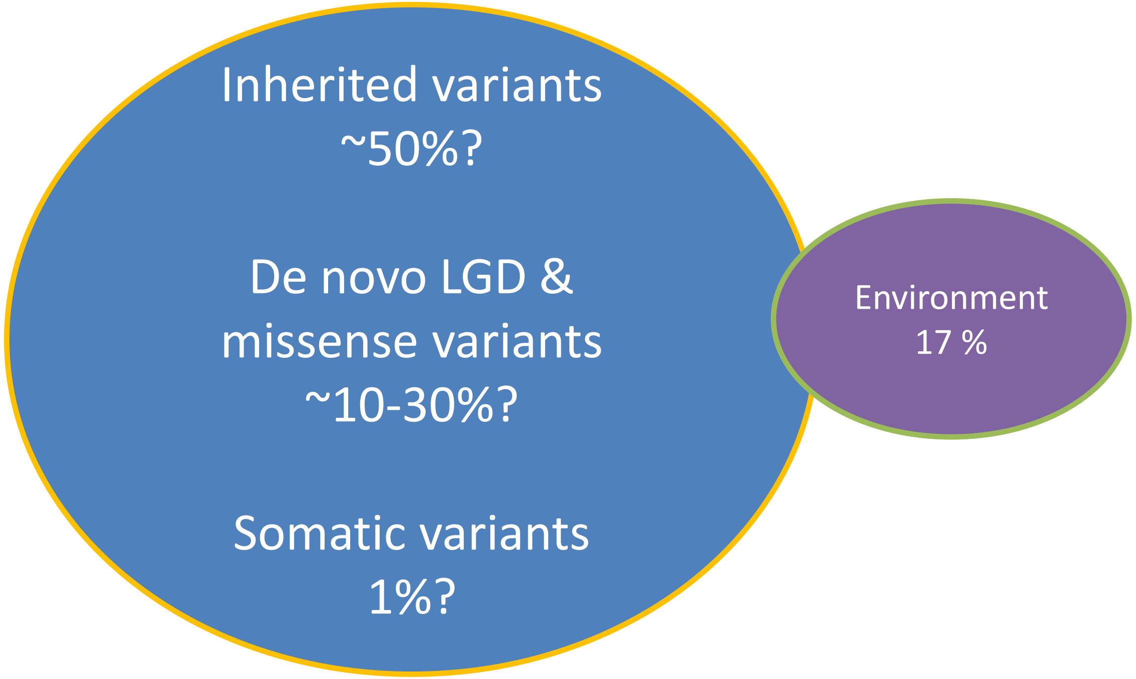
Database
Web Service
Analysis App
Data Flow/API
Controlled Access
Commercial

Acknowledgements

- SF Informatics Team
- SCC Flatiron Institute
- SFARI Science Team
- SPARK Team
- New York Genome Center(NYGC), Baylor College of Medicine Sequencing Center
- Iossifov Lab at NYGC/CSHL
- Marth Lab at University of Utah and IOBIO –Frameshift team

Thank you!

Autism risk is complex



Inherited variants
~50%?

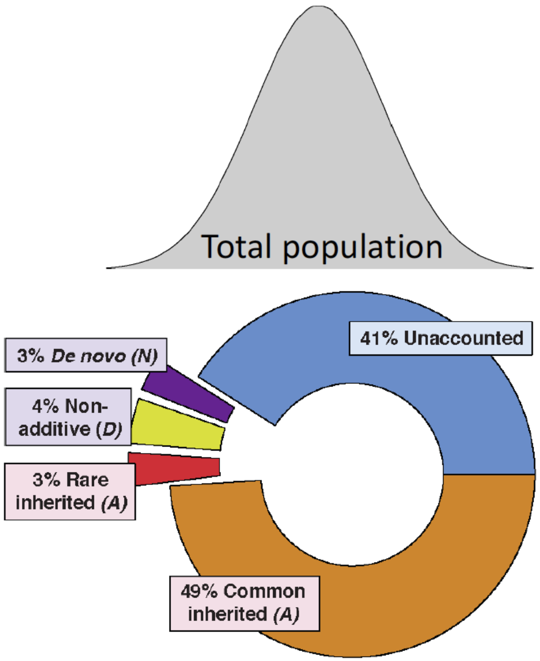
De novo LGD &
missense variants
~10-30%?

Somatic variants
1%?

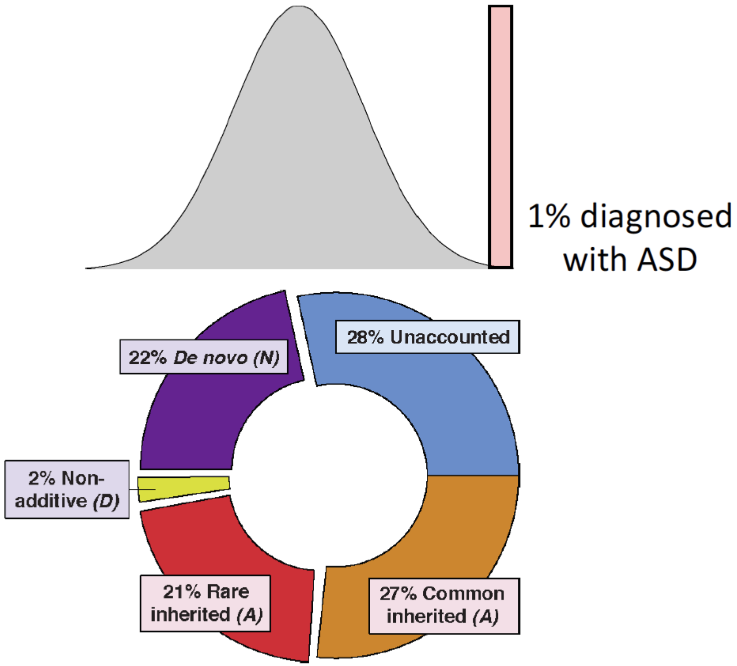
Environment
17 %

Pam Feliciano, 2017

Distribution of genetic risk in autism



Gauglet at al, 2014



(from Stephan Sanders)