Today’s Attendance URL and QR code:

https://tinyurl.com/AttendanceFeb20

Reminder: Questions

- Type your questions into the “Q&A” box
- Monitored by committee members
  - Will verbally answer as many as time allows (at the end)
  - May receive a typed answer through the Q&A box from a committee member
  - Unanswered questions will be saved for consideration during the March 12 Q&A webinar

gnomAD-SV

A Population Reference for CNV Interpretation

Ryan L. Collins
Ph.D. Candidate, Talkowski Lab
Broad Institute, Massachusetts General Hospital, and Harvard Medical School

Contact Info

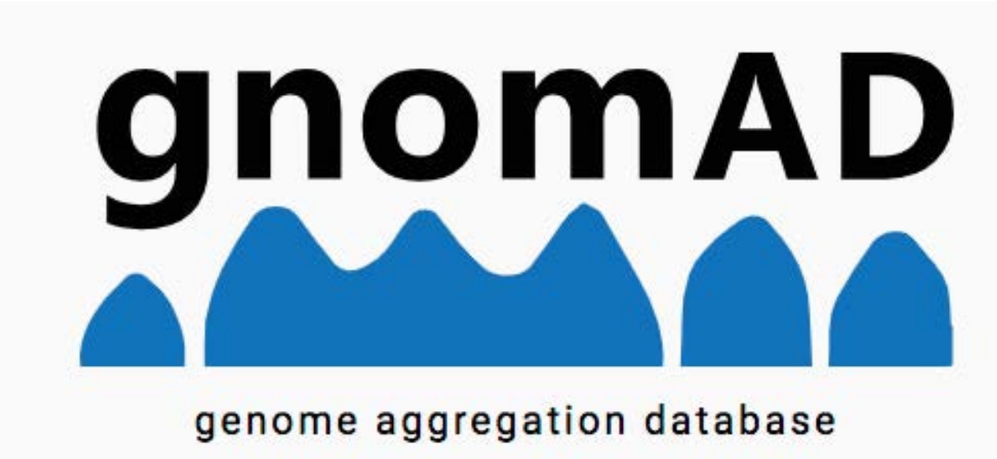
rlcollins@g.harvard.edu ryanlcollins.com talkowski.mgh.harvard.edu @ryanlcollins13

ClinGen CNV Technical Standards Webinar

February 20, 2020

WHAT IS GNOMAD-SV?

Structural variation (SV) catalog built from genome sequencing of 14,891 individuals



- Part of the Genome Aggregation Database (gnomAD)

https://gnomad.broadinstitute.org

- Integrated into existing gnomAD resources (browser, downloads)

See our preprint for more info

Last updated: October 2019

An open resource of structural variation for medical and population genetics

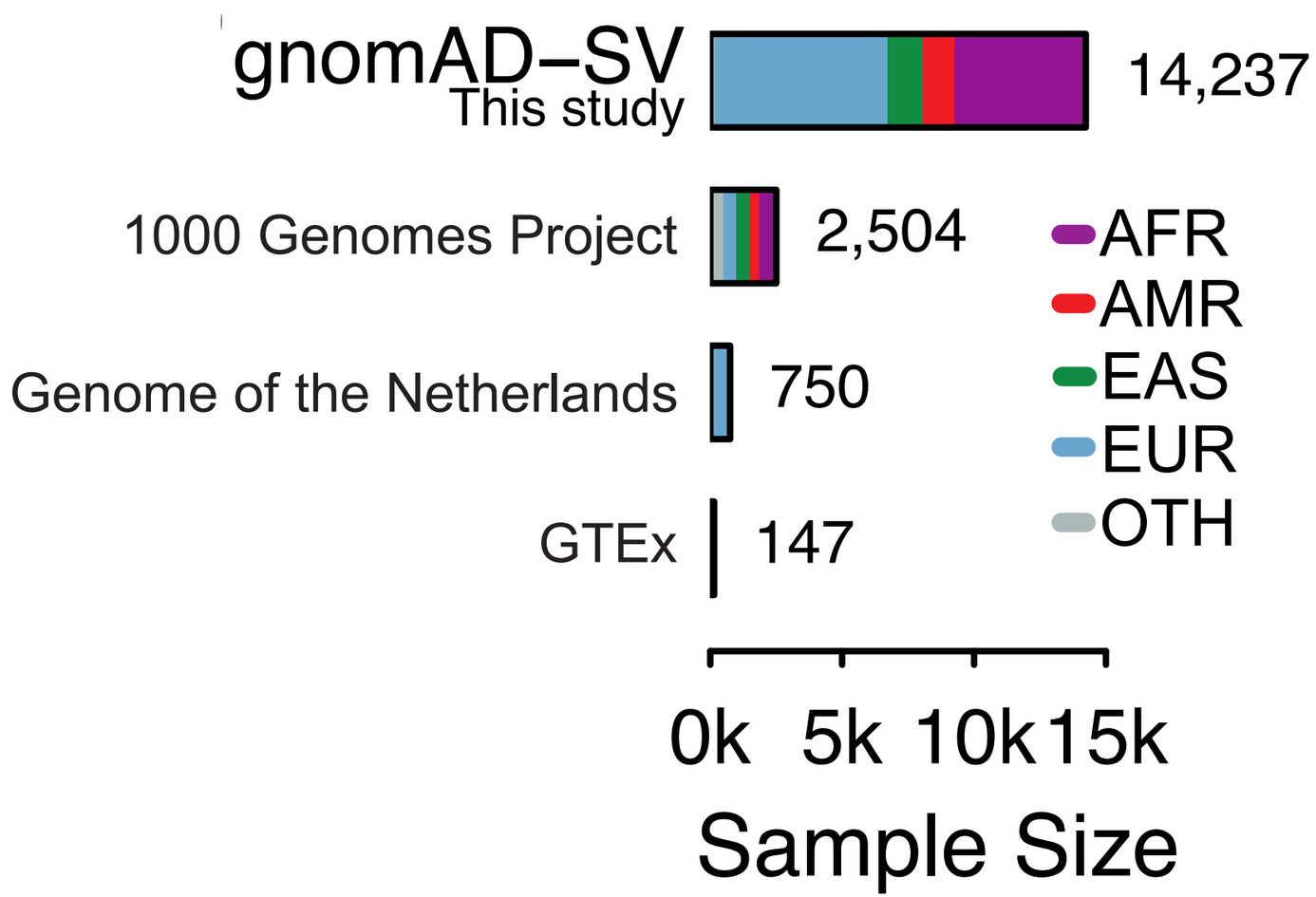
Ryan L. Collins, Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C. Francioli, Amit V. Khera, Chelsea Lowther, Laura D. Gauthier, Harold Wang, Nicholas A. Watts, Matthew Solomonson, Anne O'Donnell-Luria, Alexander Baumann, Ruchi Munshi, Mark Walker, Christopher Whelan, Yongqing Huang, Ted Brookings, Ted Sharpe, Matthew R. Stone, Elise Valkanas, Jack Fu, Grace Tiao, Kristen M. Laricchia, Valentin Ruano-Rubio, Christine Stevens, Namrata Gupta, Lauren Margolin, Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, Kent D. Taylor, Henry J. Lin, Stephen S. Rich, Wendy Post, Yii-Der Ida Chen, Jerome I. Rotter, Chad Nusbaum, Anthony Philippakis, Eric Lander, Stacey Gabriel, Benjamin M. Neale, Sekar Kathiresan, Mark J. Daly, Eric Banks, Daniel G. MacArthur, Michael E. Talkowski

doi: https://doi.org/10.1101/578674

https://broad.io/gnomad_sv

Current gnomAD-SV Release

Illumina WGS to 32X average coverage on 14,891 samples; majority (54%) non-European.



Samples, at a glance:

- 14,891 total samples processed
- Aggregated from population genetic and common & complex disease association studies
- 14,237 samples passed all QC
- 12,653 samples remained after pruning first-degree relatives
- **10,847 samples** appropriately consented for data **release via browser**

PHENOTYPES IN GNOMAD-SV

Depleted for dominant & early-onset diseases

Sample exclusion criteria for phenotypes:

- Pediatric/early-onset diseases
- First-degree relative of individuals with Mendelian dominant disease

Cohorts with known phenotypes in gnomAD-SV:

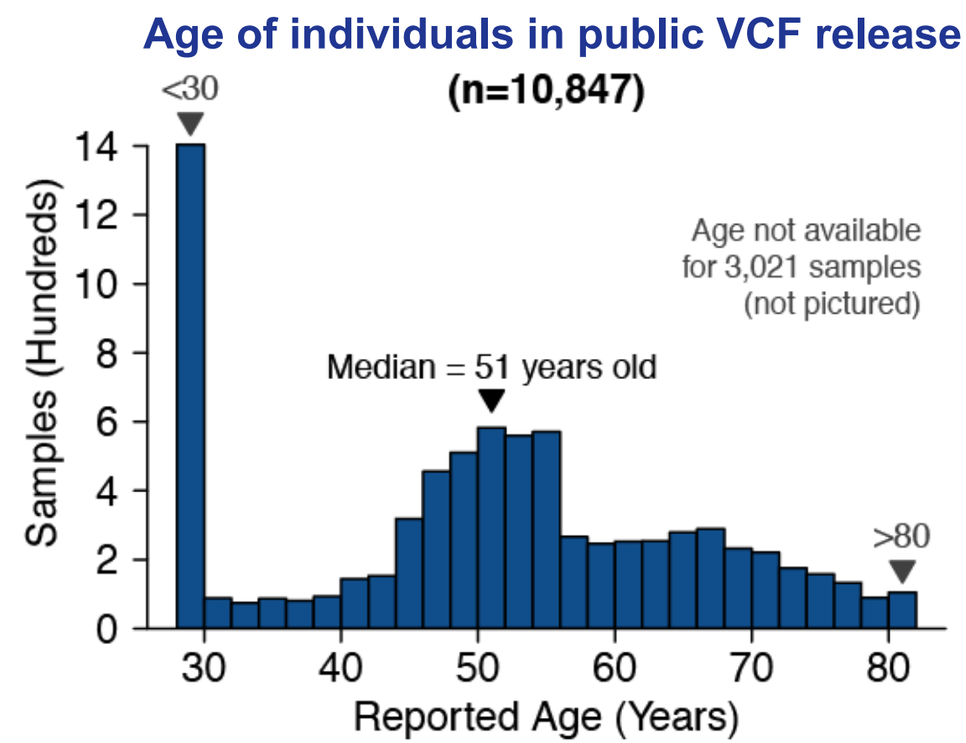
- Cardiovascular disease (CAD, hypertension)
- Adult-onset neuropsychiatric disorders

We provide the following cohort subsets:

- ***Non-neuro***: all known neurological diseases excluded
- ***Controls only***: only includes individuals labeled as a control in their contributing study

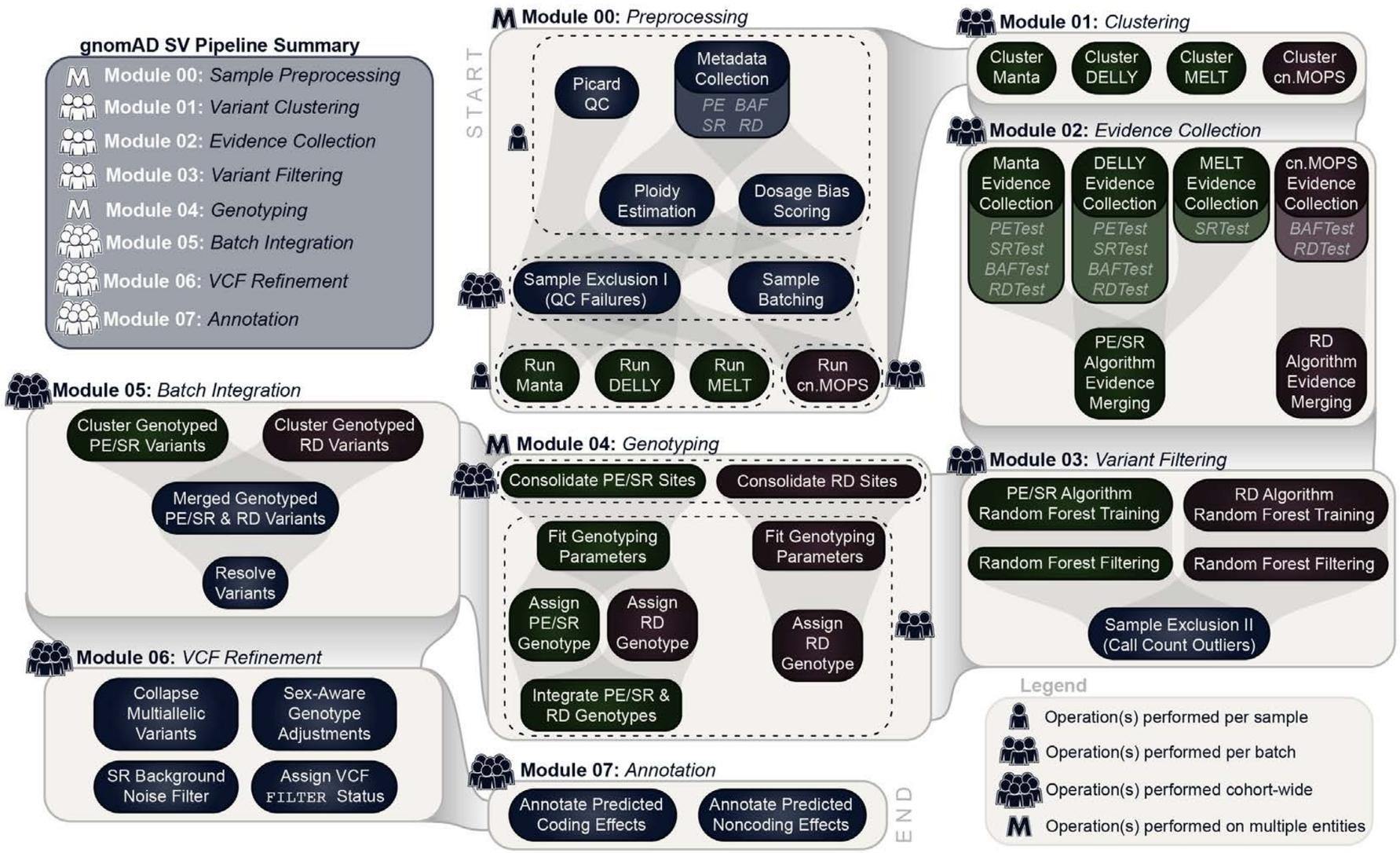
A list of contributing projects can be found here:

https://gnomad.broadinstitute.org/about



SV Discovery Methods

GATK-SV is a cloud-based pipeline for large scale SV discovery from WGS



Key Features

Maximizes sensitivity with multiple orthogonal algorithms

Integrates evidence across four modes

Boosts specificity with machine learning variant filtering

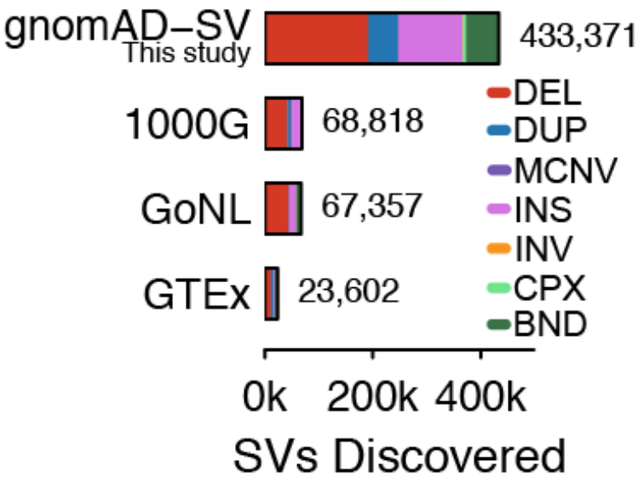
Uniform **genotyping**, variant **resolution**, and **annotation** across all samples

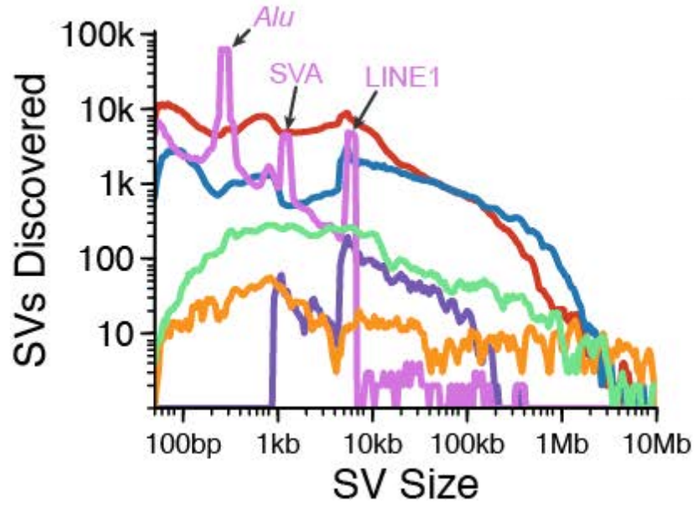
Overview of gnomAD-SV Callset

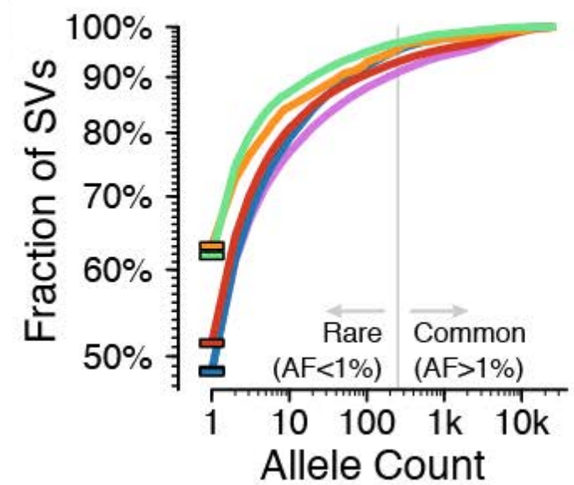
Discovered **433k SVs** in 7 mutational classes

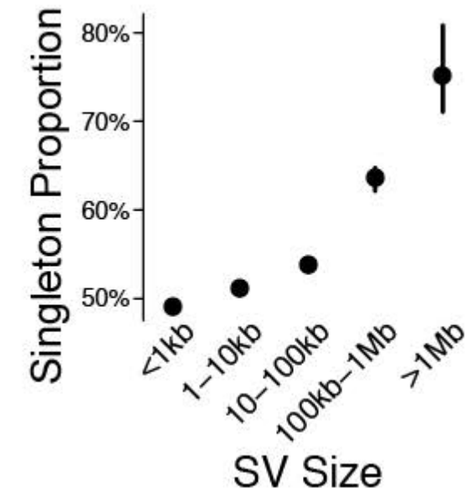
Most SVs are small & rare

Frequency inversely correlated w/ size

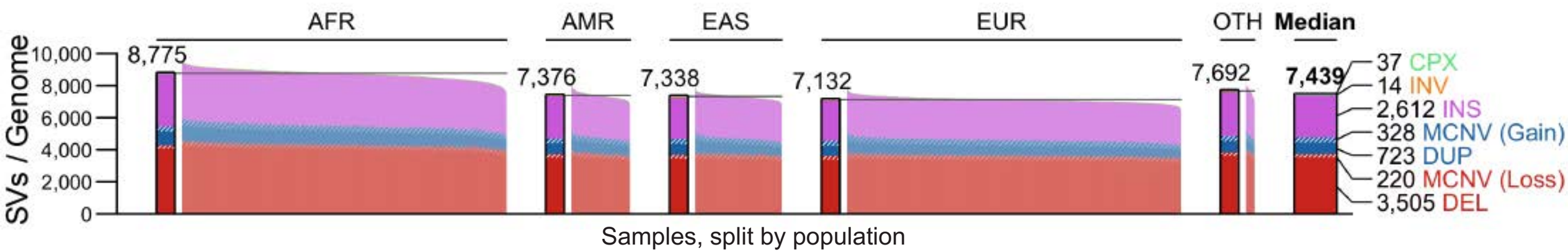






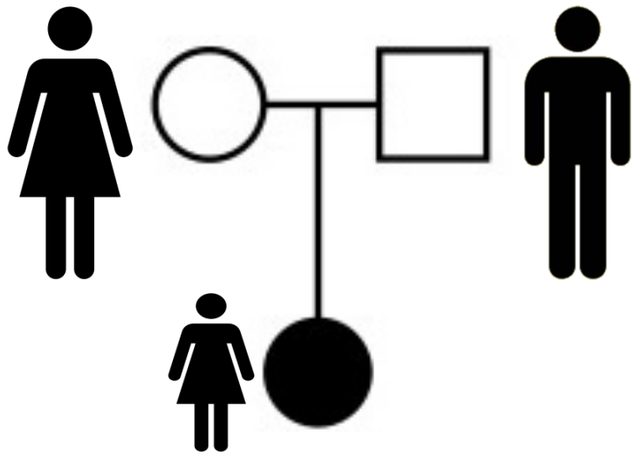


The average genome harbors **7.5k** SVs

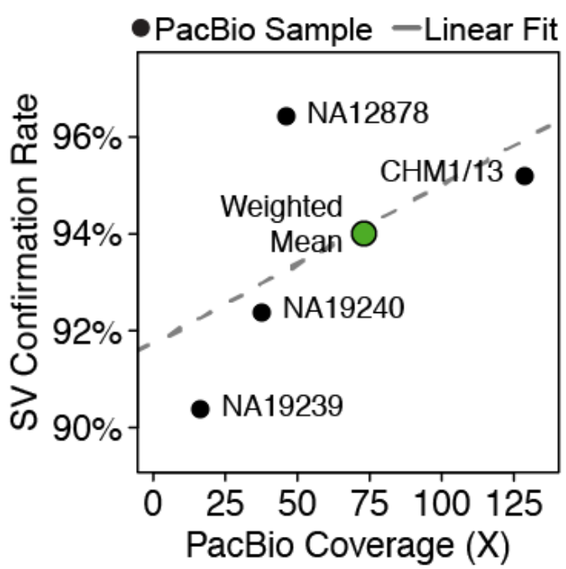


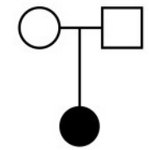
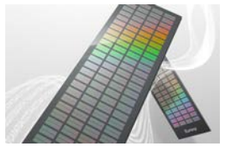
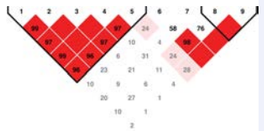
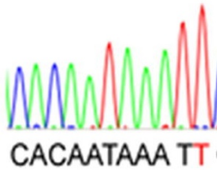
QUALITY ASSESSMENT

3% *de novo* rate from **970 trios**



94% confirmation rate of ~20k SVs by **PacBio**

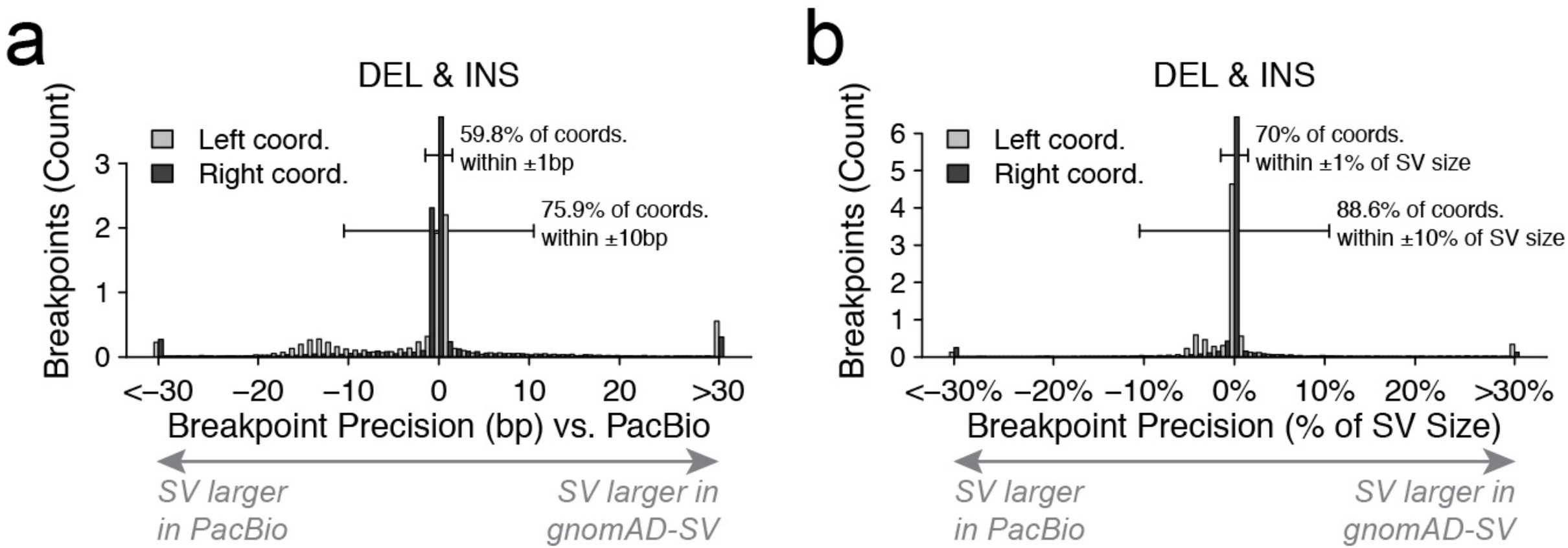


	Approach	Samples	SVs	Result
	Inheritance analysis	2,910	8,512 per family	**3.8%** Mendelian violation rate
	Microarray comparison	1,893	2,524	**97.1%** sensitivity
$p^2 + 2pq + q^2$	Hardy-Weinberg Equilibrium	12,653	321,140	**85.8%** in HWE
	SV vs. SNV linkage disequilibrium	5,353	23,597	R^2 = **0.85**
	PCR/Sanger & ddPCR validation	2,076	177	**97%** validation rate of *de novo* SV predictions
	Allele frequency comparisons vs. 1000G Project	12,653 & 2,504	37,907	R^2 = **0.72**
	Long-read confirmation	4	4,829 per sample	**94.0%** PPV

Breakpoint Precision

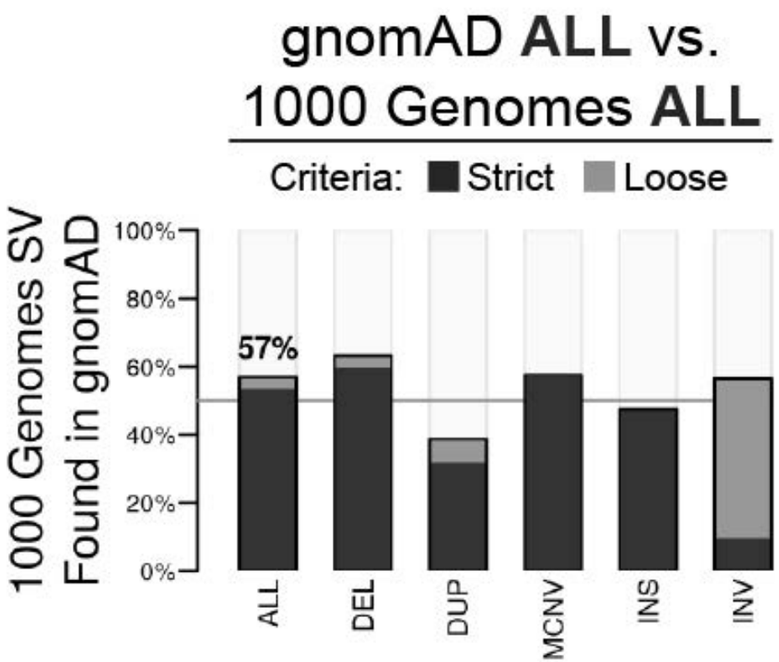
Comparison to published long-read WGS SV calls for two samples also present in gnomAD-SV

75.6% of breakpoints accurate within ±10bp

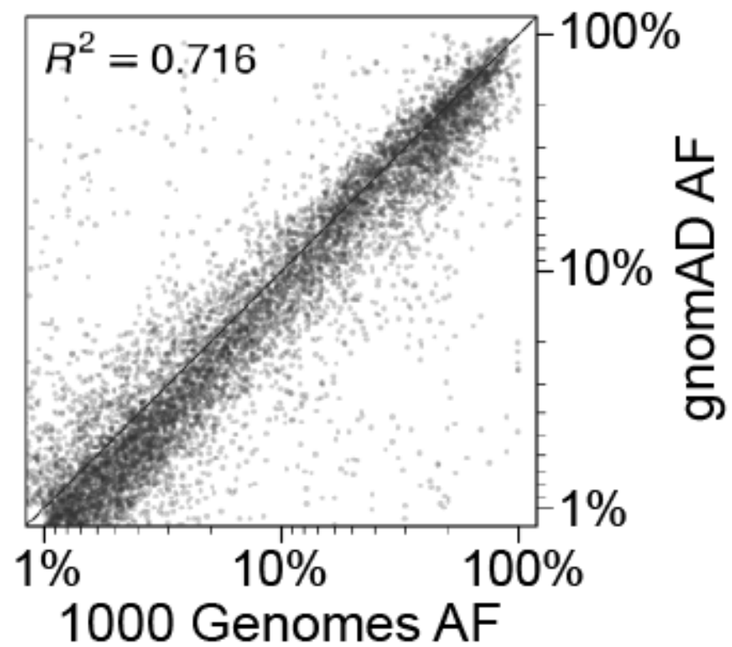


Comparisons to the 1000 Genomes Project

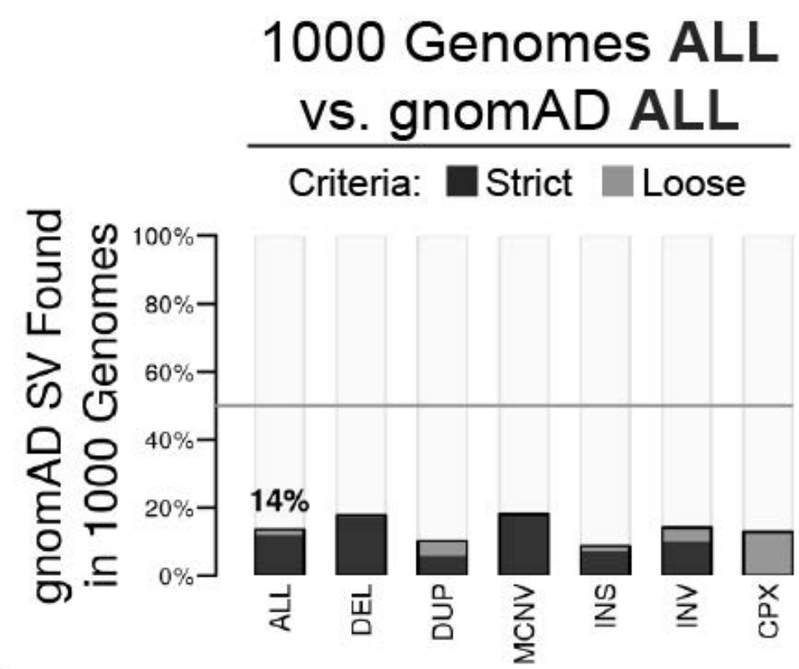
gnomAD-SV captures **57%** of all SVs in 1000G



Allele frequencies are strongly correlated



1000G captures **14%** of all SVs in gnomAD-SV



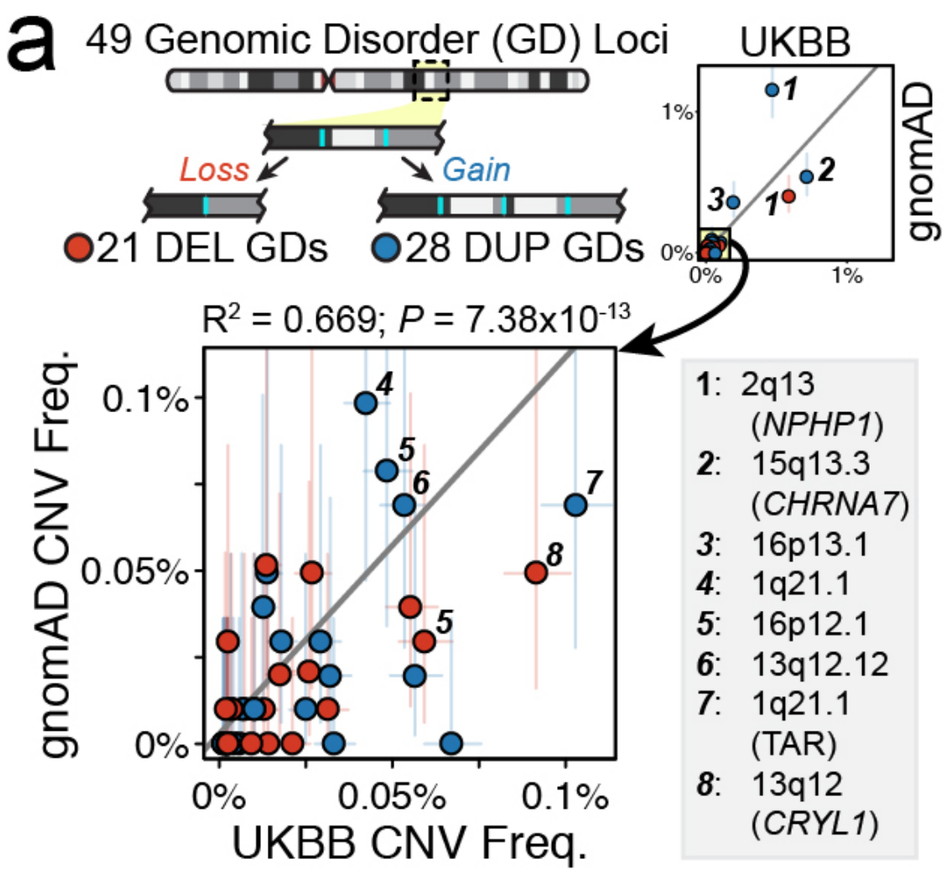
See Sudmant *et al., Nature*, 2015 for description of 1000 Genomes phase 3 SV callset

Annotated Gene Consequences

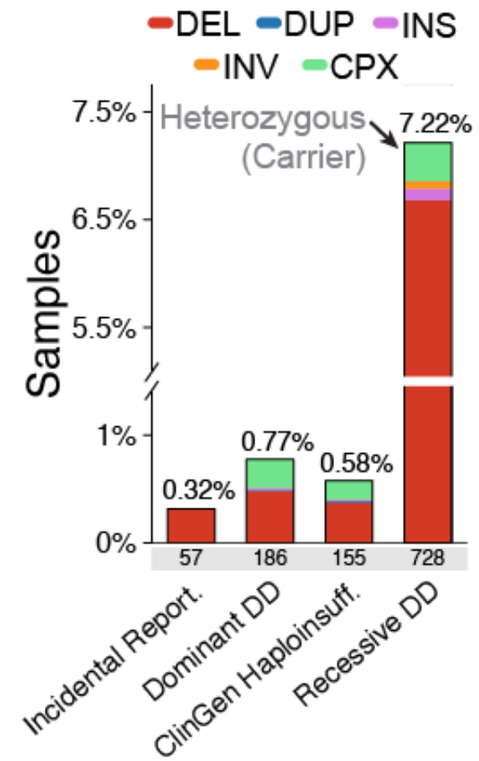
	Loss of Function (In manuscript: pLOF) (In VCF: LOF)	Intragenic Exonic Duplication (In manuscript: IED) (In VCF: DUP_LOF)	Whole-Gene Copy Gain (In manuscript: CG) (In VCF: COPY_GAIN)	Whole-Gene Inversion (In manuscript: INV) (In VCF: INV_SPAN)
DEL	Any overlap with any exon	Not assigned	Not assigned	Not assigned
DUP	Both breakpoints in exons of the same gene	Overlaps at least one exon and both breakpoints within gene	Both breakpoints outside gene and duplication spans whole gene	Not assigned
INS	Inserted into any exon	Not assigned	Not assigned	Not assigned
INV	Exactly one breakpoint in gene, or both breakpoints in same gene and inversion spans any exon	Not assigned	Not assigned	Both breakpoints outside gene and inversion spans whole gene
CPX	Any combination of SV intervals such that at least one meets loss of function criteria	Any combination of SV intervals such that at least one meets internal exon duplication criteria	Any combination of SV intervals such that at least one meets duplicated gene criteria	Any combination of SV intervals such that at least one meets inverted gene criteria
CTX	Any breakpoint in gene	Not assigned	Not assigned	Not assigned

MEDICALLY RELEVANT BENCHMARKS IN GNOMAD-SV

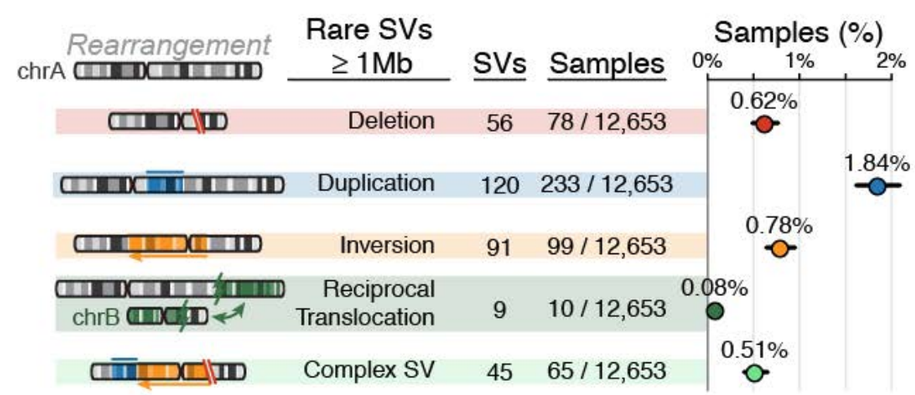
Genomic disorder CNV carrier frequencies match microarray-based estimates in UK BioBank

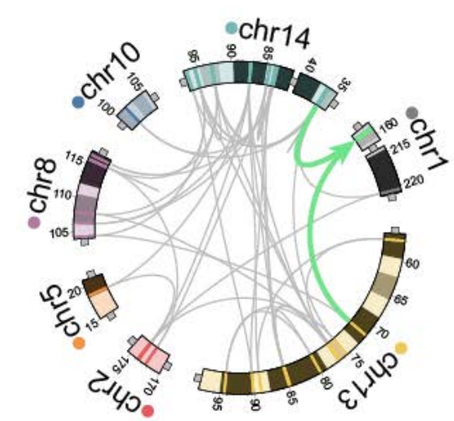


~1 in 200 genomes carry very rare pLoF SVs in dominant disease genes



3.8% of genomes carry a rare SV ≥1Mb

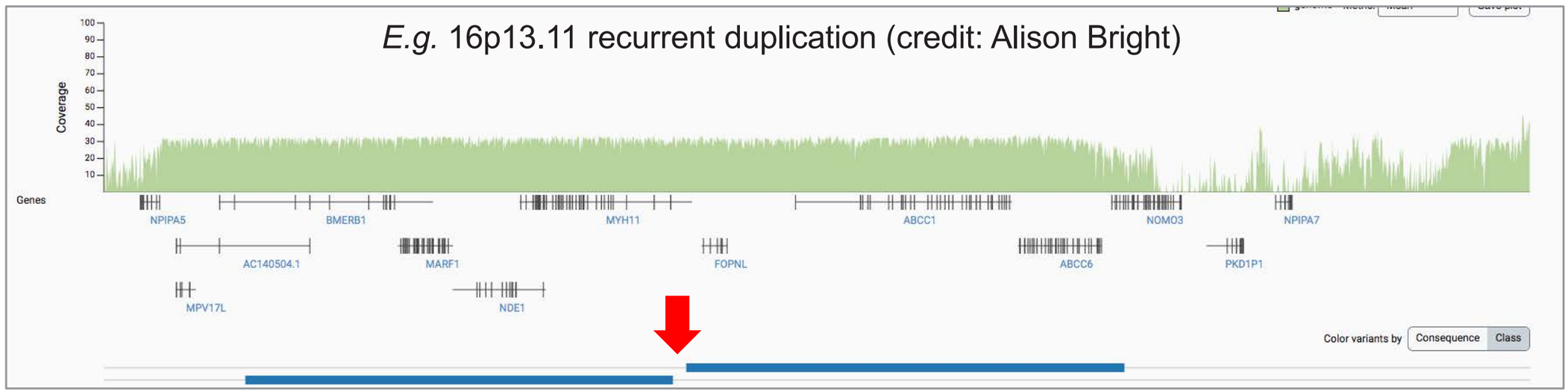




LIMITATIONS OF GNOMAD-SV CALLSET

1. Reduced sensitivity for small (<300bp) duplications and insertions

2. Large (>200kb) CNVs can sometimes be fragmented
   - This seems to be a infrequent occurrence, but can happen at isolated loci



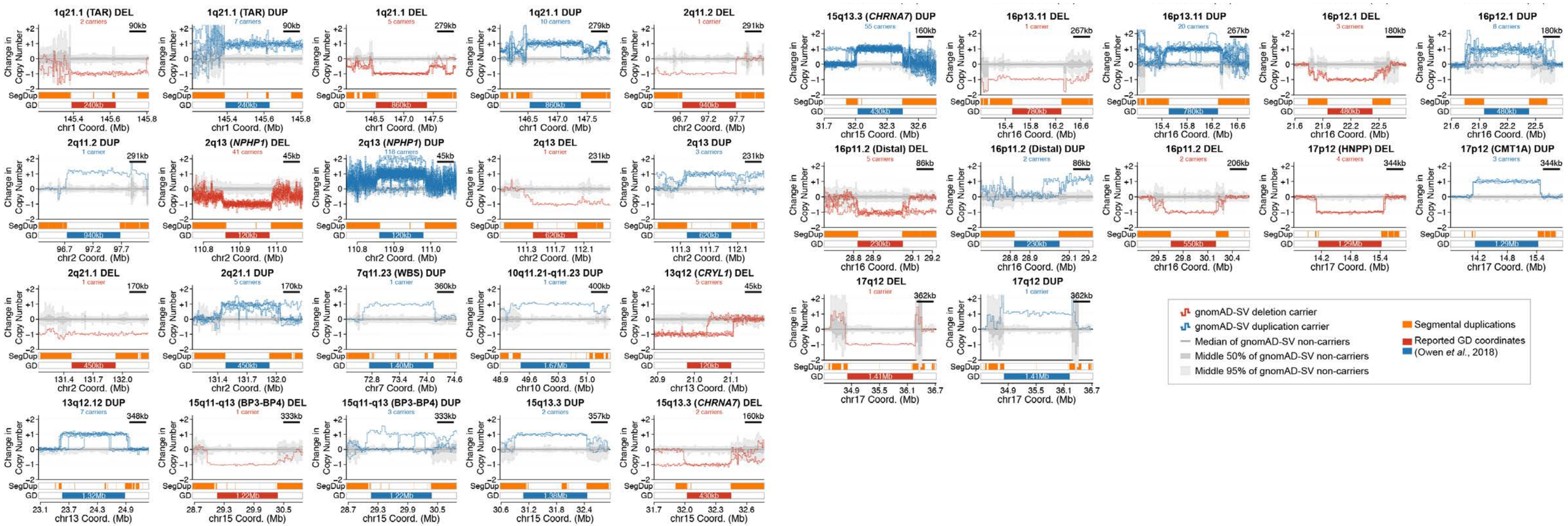
These two duplication calls appear in the same 20 individuals, and should have been merged

3. Limited coverage of Asian, Latino, and Oceanic populations
   - This will be dramatically improved in next version of gnomAD-SV
   - Same comment for sub-populations provided for gnomAD SNVs & indels

Contiguity of Known Genomic Disorder Regions

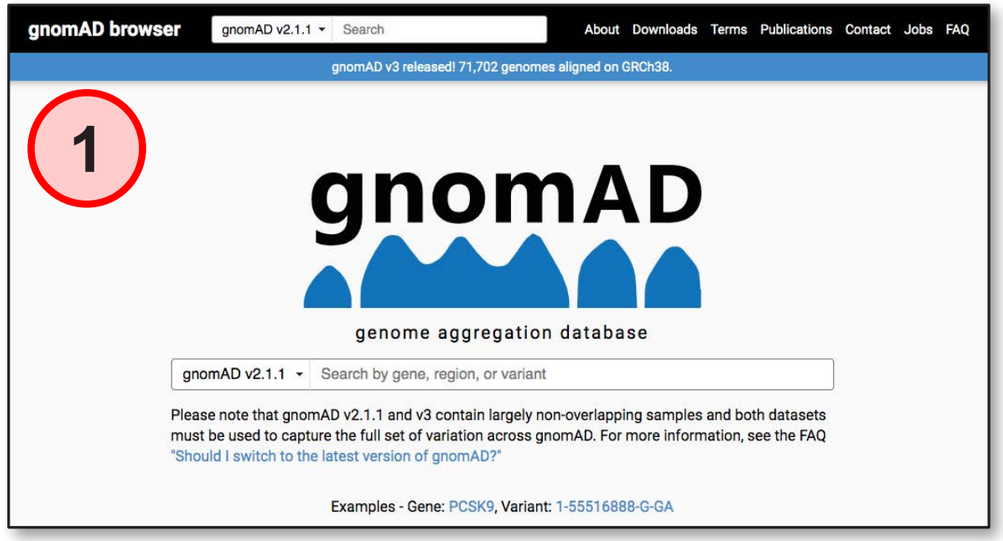
CNV fragmentation & misgenotyping at NAHR-mediated regions is an exception, not the rule

Clear and consistent read-depth support for predicted CNV carriers at known genomic disorder loci

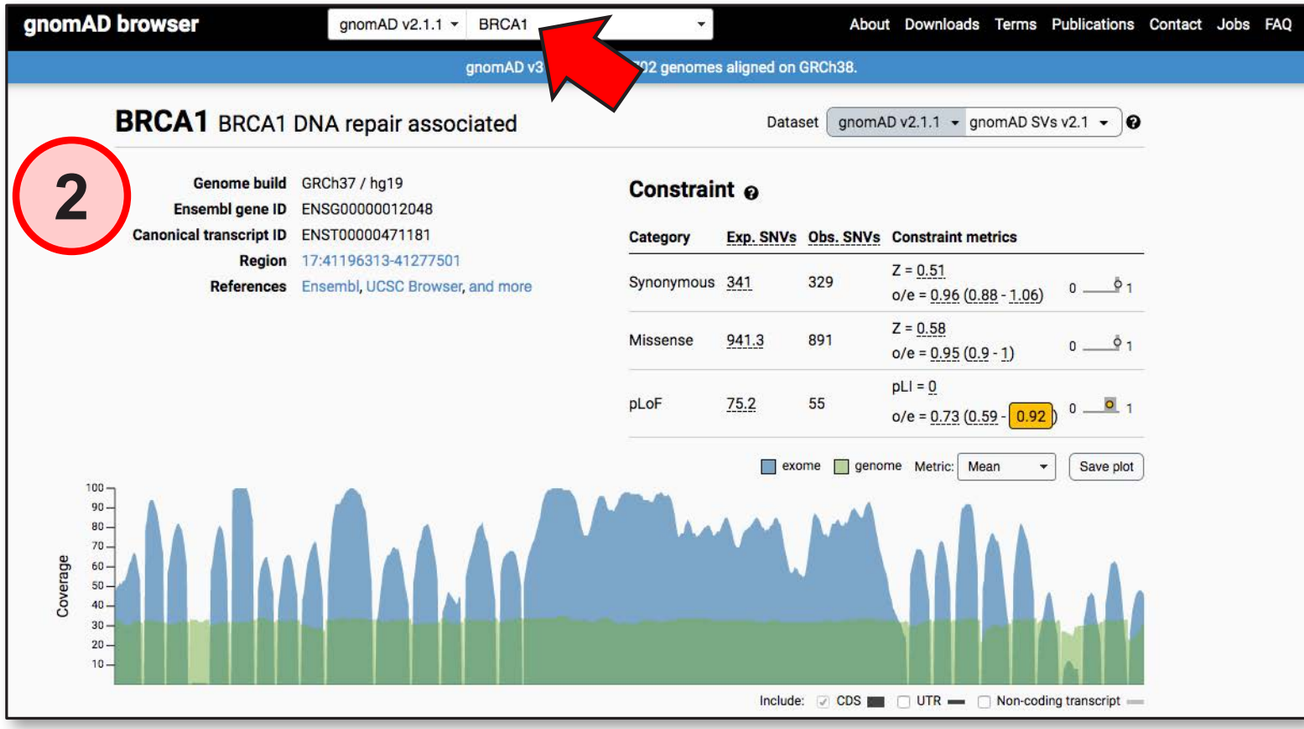


How to Browse gnomAD-SV

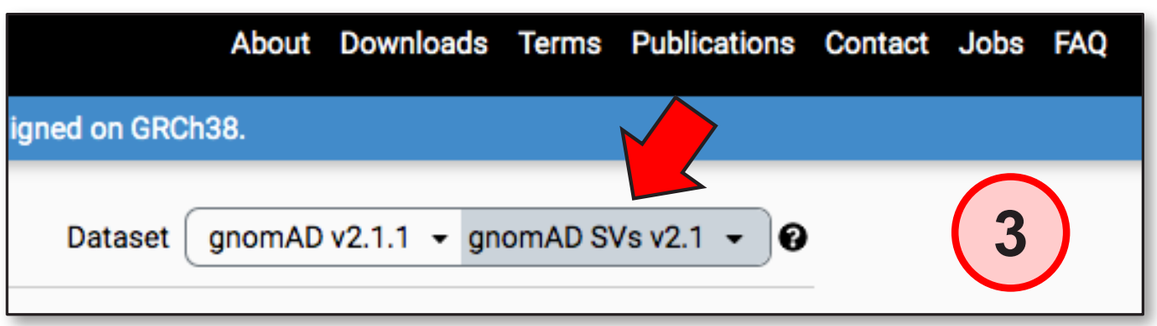
1. Navigate to https://gnomad.broadinstitute.org



2. Search for your gene or locus of interest



3. Activate the SV dataset

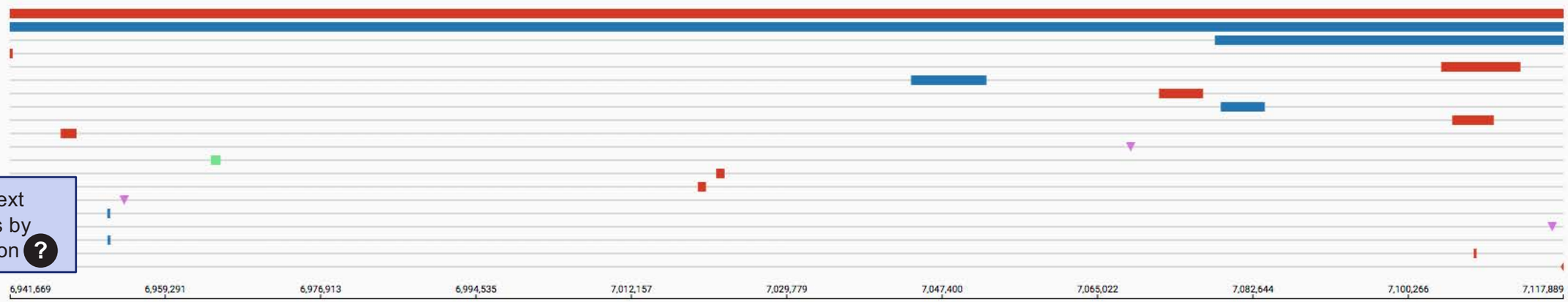


Locus/Gene View

Toggle color code (consequence vs. SV class)

Variants displayed in genomic coordinate space

Color variants by Consequence Class



Help text popups by clicking on ?

Consequences

pLoF only | Int. exon duplication only | Copy gain only | Other only

Filter options

Show low-quality variants

Include filtered variants

Search

Search variant table

Classes

DEL only | DUP only | MCNV only | INS only | INV only | CPX only | OTH only

Click to download table

Export variants to CSV

Click on column header to sort table by those values

Table of SVs with (some) metadata

Variant ID	Source	Consequence	Class	Position	Size	Allele Count	Allele Number	Allele Frequency	Number of Homozygotes
DEL_18_163883	G	loss of function	deletion	6839499 - 8715000	1.88 Mb	1	21642	4.6e-5	0
DUP_18_45227	G	copy gain	duplication	6935549 - 8087897	1.15 Mb	2	21694	9.2e-5	0
DUP_18_45233	G	partial duplication	duplication	7078339 - 7578544	500 kb	6	21692	2.77e-4	0
DEL_18_163886	G	utr	deletion	6909700 - 6942000	32.3 kb	1	21556	4.6e-5	0
DEL_18_163903	G	intronic	deletion	7104000 - 7113000	9.00 kb	6	21000	2.86e-4	0
DUP_18_45232	G	int. exon duplication	duplication	7043872 - 7052442	8.57 kb	1	20614	4.9e-5	0
DEL_18_163898	G	intronic	deletion	7072000 - 7077000	5.00 kb	1	19736	5.1e-5	0
DUP_18_45234	G	int. exon duplication	duplication	7079000 - 7084000	5.00 kb	1	21624	4.6e-5	0
DEL_18_163904	G	intronic	deletion	7105256 - 7109965	4.71 kb	2	21694	9.2e-5	0
DEL_18_163894	G	loss of function	deletion	6947438 - 6949241	1.80 kb	1	21694	4.6e-5	0
INS_18_104832	G	intronic	insertion	7068787	1.24 kb	1	21694	4.6e-5	0
CPX_18_4762	G	loss of function	complex	6964469 - 6965587	1.12 kb	1	21694	4.6e-5	0

Click on a variant ID to be taken to variant view (next slide)

VARIANT VIEW

Structural variant: DUP_1_3816

Dataset gnomAD SVs v2.1

Toggle for “controls only” and “non-neuro” subsets

Variant metadata listed here

Filter	Pass
Allele Count	3679
Allele Number	21424
Allele Frequency	0.1717
Quality score	999
Position	1:248620000-
Size	27,000 bp
Class	duplication
Evidence	Normalized B-allele frequency, Read depth
Algorithms	Depth

Link back to locus view

References

- UCSC

Links for more information and/or reporting issues

Report

- Report this variant
- Request additional information

Consequences

This variant has consequences in 1 gene.

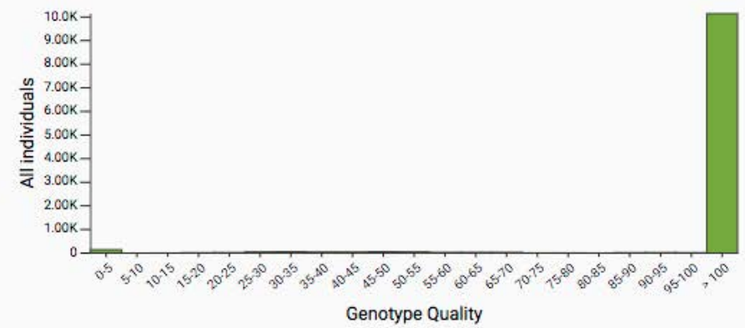
copy gain

- OR2T3

Help text popups by clicking on ?

All annotated gene consequences will be listed here

Genotype Quality Metrics



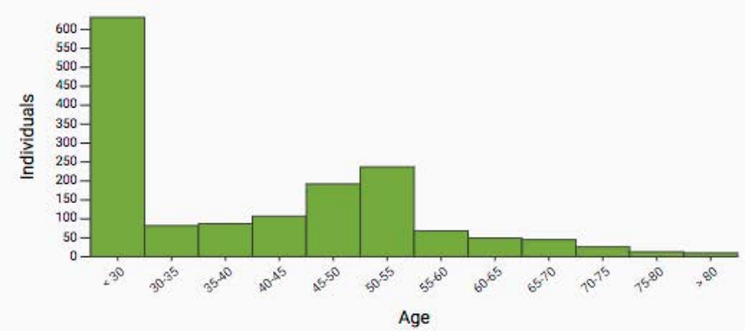
All | Variant Carriers

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
African	2249	9408	75	0.2391
East Asian	348	2416	10	0.1440
Other	25	190	0	0.1316
European	867	7588	6	0.1143
Latino	190	1822	3	0.1043
Male	1913	10906	55	0.1754
Female	1762	10474	39	0.1682
Total	**3679**	**21424**	**94**	**0.1717**

Click each population to stratify by male/female

Age Distribution



Heterozygous Variant Carriers

How to Download gnomAD-SV

https://gnomad.broadinstitute.org/downloads

gnomAD browser | gnomAD v2.1.1 | Search | About | Downloads | Terms | Publications | Contact | Jobs | FAQ

gnomAD v3 released! 71,702 genomes aligned on GRCh38.

Downloads

gnomAD is available for download in Hail Table (.ht) and VCF formats.

Files can be browsed and downloaded in parallel using gsutil. After installing gsutil, start browsing with `gsutil ls gs://gnomad-public/release`.

To work efficiently with gnomAD, we recommend using Hail and our Hail utilities for gnomAD.

gnomAD v2 | gnomAD v2 liftover | gnomAD v3 | ExAC

Update, March 6, 2019: gnomAD 2.1.1 released (README.txt)

Summary

- Variants
- Coverage
- Constraint
- Multi-nucleotide variants (MNVs)
- Proportion expressed across transcripts (pext)
- Structural variants
- Resources

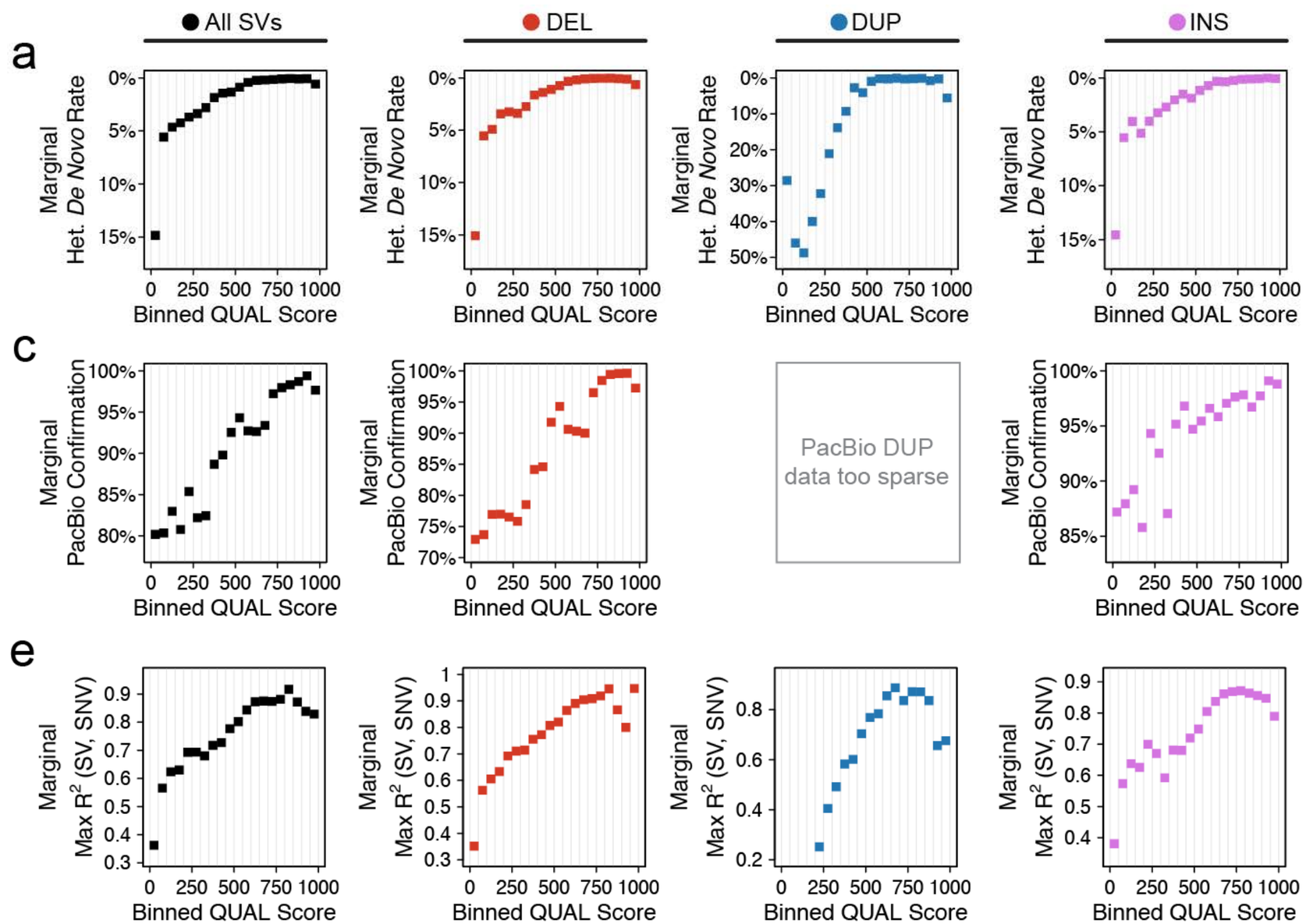
Structural variants

- SV 2.1 sites VCF (.tbi)
- SV 2.1 sites BED (.tbi)
- SV 2.1 (controls) sites VCF (.tbi)
- SV 2.1 (controls) sites BED (.tbi)
- SV 2.1 (non-neuro) sites VCF (.tbi)
- SV 2.1 (non-neuro) sites BED (.tbi)

Recommendations for Filtering

- QUAL scores
- Overlap with repetitive sequences (seg. dups, RepeatMasker, etc.)
- Don't rely *only* on gnomAD-SV (*i.e.*, also check DGV & other datasets)

Quality scores track with benchmarking metrics



gnomAD-SV FAQs

- **Is there an hg38 version of gnomAD-SV?**
  - The next version of gnomAD-SV will be native to hg38 (expected release: end of 2020)
  - For now, NCBI dbVar maintains an hg38-liftover version (accession nstd166)

- **Is individual-level genetic data available for gnomAD-SV?**
  - Unfortunately, no. Given the broad variety of studies contributing data to gnomAD, we are unable to release anything beyond site-frequency data

- **Is phenotype data available for gnomAD-SV?**
  - Other than age and inferred sex, no (same reasons as for individual-level genetic data)

- **Is GATK-SV publicly available?**
  - Yes, but the current implementation is undocumented and restricted to Google Cloud
  - Broad Data Sciences Platform will be releasing a fully documented & production-quality version of GATK-SV soon (mid-2020)

- **Will you be releasing constraint scores for SVs?**
  - Eventually. This is a major research focus for us, but current sample sizes are inadequate

CENTER FOR GENOMIC MEDICINE

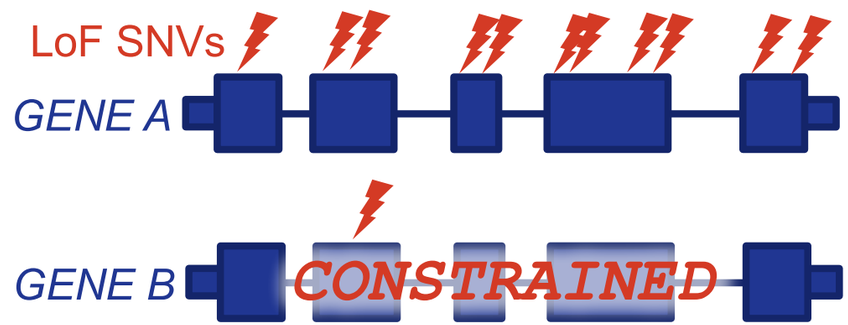
Part II

CONSTRAINT & SVs

Two Definitions of Constraint

Mutational constraint

Genes **depleted** for LoF **SNVs** in gnomAD



Konrad Karczewski

Kaitlin Samocha

pLI

- Probability of LoF intolerance
- ~Binary yes/no classification
- ***Does***: indicate whether a gene is significantly LoF constrained
- ***Does not***: measure strength of constraint per gene
- *Example: longer genes will have higher pLI scores than smaller genes, but this is a statistical property, not a biological one*

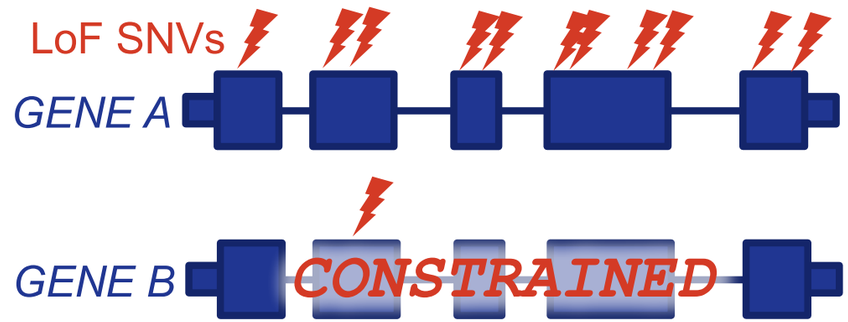
Observed/Expected

- o/e: ratio of observed LoFs vs. expected LoFs
- LOEUF: upper bound of 90% confidence interval for o/e (more conservative)
- ***Does***: measure strength of constraint per gene
- ***Does not***: indicate whether a gene is significantly constrained
- *Example: MEF2C, a known autosomal dominant disease gene, has pLI=0.02 but o/e=0.32 (clearly depleted for LoFs, but not significantly so given small gene size)*

Interpreting SVs that Alter Genes

Mutational constraint

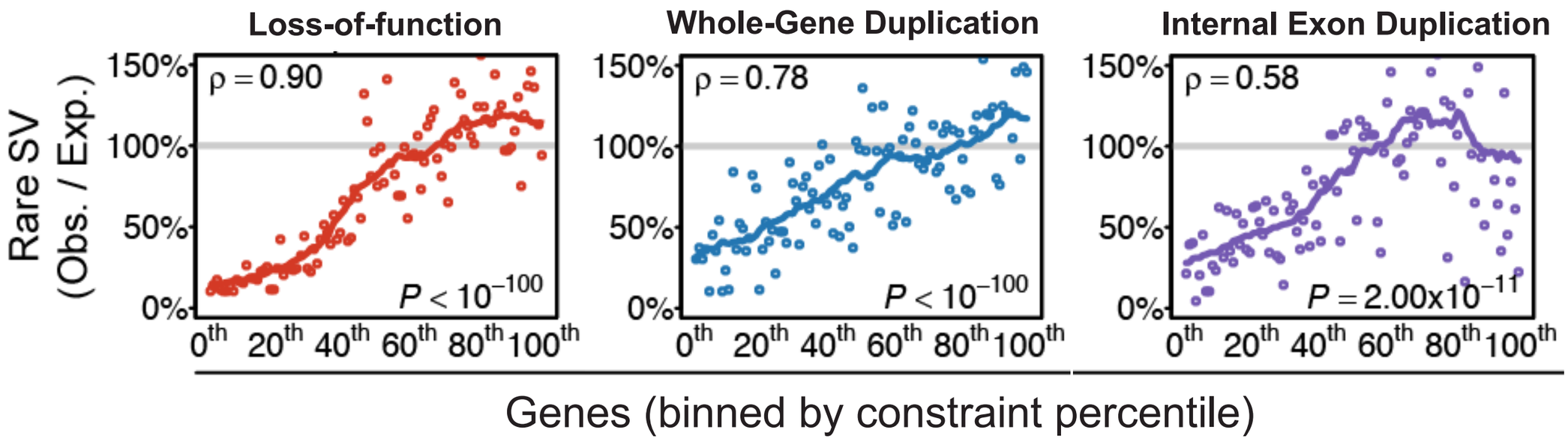
Genes **depleted** for LoF **SNVs** in gnomAD



Konrad Karczewski

Kaitlin Samocha

Constrained genes don't tolerate deletions... or duplications



Constraint marks a **general intolerance** to all SVs

Including **noncoding SVs** (see gnomAD-SV preprint for details)

Acknowledgements

Harrison Brand

Michael Talkowski

Daniel MacArthur

Broad-SV

Laura Gauthier
Chris Whelan
Mark Walker
Ted Brookings
Ted Sharpe
Steve Huang
Samuel Lee
Andrey Smirnov
Valentin Ruano-Rubio
Alex Baumann
Ruchi Munshi

Talkowski Lab

Xuefang Zhao
Harold Wang
Chelsea Lowther
Jack Fu
Elise Valkanas
Matt Stone

gnomAD

Jessica Alföldi
Konrad Karczewski
Laurent Francioli
Eric Banks
Mark Daly
Ben Neale
Sekar Kathiresan
Nick Watts
Matt Solomonson
Amit Khera
Namrata Gupta

Contact info:

rlcollins@g.harvard.edu
ryanlcollins.com
@ryanlcollins13

Preprint on *bioRxiv*

https://broad.io/gnomad_sv

New Results

4 comments

An open resource of structural variation for medical and population genetics

Ryan L. Collins, Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Amit V. Khera, Laurent C. Francioli, Laura D. Gauthier, Harold Wang, Nicholas A. Watts, Matthew Solomonson, Anne O'Donnell-Luria, Alexander Baumann, Ruchi Munshi, Chelsea Lowther, Mark Walker, Christopher Whelan, Yongqing Huang, Ted Brookings, Ted Sharpe, Matthew R. Stone, Elise Valkanas, Jack Fu, Grace Tiao, Kristen M. Laricchia, Christine Stevens, Namrata Gupta, Lauren Margolin, The Genome Aggregation Database (gnomAD) Production Team, The gnomAD Consortium, John A. Spertus, Kent D. Taylor, Henry J. Lin, Stephen S. Rich, Wendy Post, Yii-Der Ida Chen, Jerome I. Rotter, Chad Nusbaum, Anthony Philippakis, Eric Lander, Stacey Gabriel, Benjamin M. Neale, Sekar Kathiresan, Mark J. Daly, Eric Banks, Daniel G. MacArthur, Michael E. Talkowski

doi: https://doi.org/10.1101/578674

This article is a preprint and has not been certified by peer

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

GRFP NSF GRADUATE RESEARCH FELLOWSHIP PROGRAM

NIH

SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

BROAD INSTITUTE

MGH 1811

CENTER FOR GENOMIC MEDICINE

HARVARD MEDICAL SCHOOL