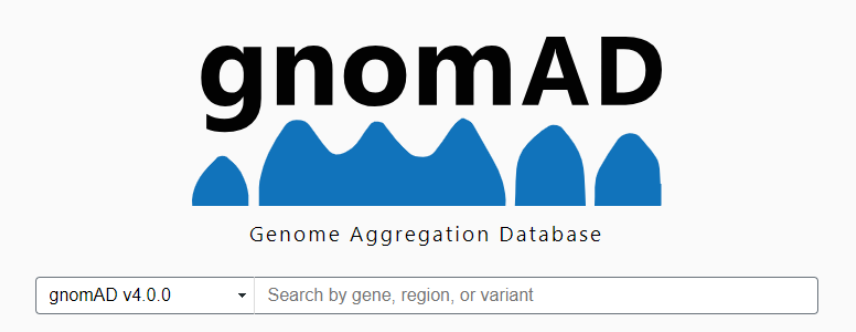
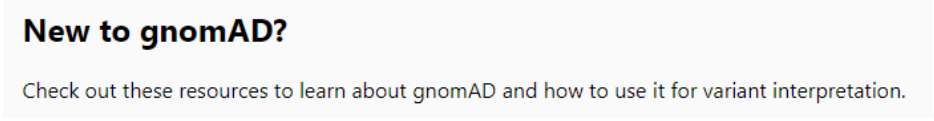
gnomAD: The Genome Aggregation Database

- A database of exomes and genomes that is widely used during the analysis of constitutional variants
- Launched in 2014 as ExAC (i.e., Exome Aggregation Consortium), consisting of exomes and genomes compiled by a team of investigators
    - https://gnomad.broadinstitute.org/about
- Subsequent versions of the database were released to the public as more exomes and genomes were collected
- gnomAD is the successor of ExAC
- Proper use of the gnomAD database requires addressing the following questions:
    - Where does the data in gnomAD come from?
    - Which version of the database should you use?
    - What are the limitations of gnomAD?

The gnomAD Browser: https://gnomad.broadinstitute.org/



- Choose which version of the database you want to search from the dropdown, and then search by gene, region, or variant
- Information about using gnomAD can be found on their website below the search bar:

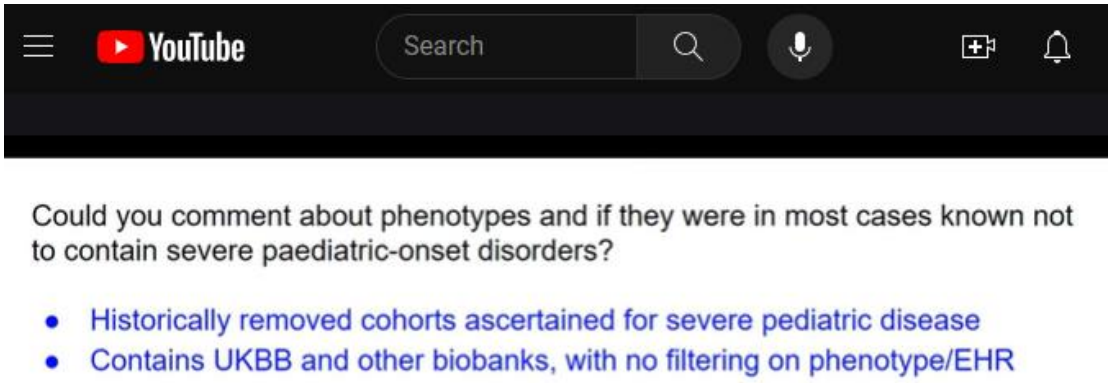


Using gnomAD - tips and tricks (video)

Six lessons for variant interpretation

What are the differences between the gnomAD dataset versions?

- ***All gnomAD dataset versions are designed to be depleted for severe pediatric disease. Cohorts that had ascertainment for known severe pediatric disease have been removed from all gnomAD versions. However, individuals with disease can still be present in any version of gnomAD. See an explanation by the gnomAD Production Team in the video link below (15 min 10 sec time point):***
  - https://youtu.be/NeTj4BwLwpQ?si=QTMfpqXPEdaQVdQq&t=15m10s



- Comparison of the gnomAD versions:

gnomAD dataset	Reference	Exomes	Genomes	Total	Phenotype subsets	Strength
version 2	GRCh37	125,748	15,708	141,456	Yes	Gene constraint metrics
version 3	GRCh38	0	76,156	76,156	Yes	Mito and intron variants
version 4	GRCh38	730,947	76,215	807,162	No	Large size

- **Version 4 details**
  - Mapped against GRCh38/hg38
  - Contains 730,947 exomes and 76,215 genomes
  - Nearly 5x larger than v2 and v3 combined
  - Contains all genomes from gnomAD v3 and most exomes/genomes from v2
  - Includes biobank data for which no phenotype data is available for individuals

- Because phenotype data is not available for many biobank samples, specific subsets are not available for v4
- The largest biobank contribution is from UK Biobank, which collected data from individuals aged 40-69 (variants causing severe early-onset disease are not expected to be present)
- UK Biobank is known to be depleted for individuals with intellectual disability, autism, and severe early-onset psychiatric disorders
- The composition of gnomAD v4 regarding known and unknown phenotypes is provided on their website's Stats page:
  - https://gnomad.broadinstitute.org/stats

Phenotypes	Case	Control	Unknown	Total	% of cases out of all v4 exomes
Alzheimer's disease	2,594	665	1,632	4,890	0.35%
Atrial Fibrillation	4,398	3,546	38,289	46,233	0.60%
Biobank or control dataset*	-	24,016	447,750	471,766	N/A
Bipolar disorder	19,284	16,383	80	35,747	2.64%
Cardiac arrhythmia	458	-	-	458	0.06%
Coronary heart disease	1,557	-	-	1,557	0.21%
Inflammatory bowel disease spectrum and related disorders^	35,008	11,928	280	47,217	4.79%
Myocardial infarction	11,900	369	-	12,269	1.63%
Neurodevelopmental**	-	132	-	143	N/A
Non-specific cardiovascular disease	1,888	11,376	15,000	28,264	0.26%
Schizophrenia spectrum and related disorders	30,278	17,689	39	47,994	4.14%
Type 2 Diabetes	17,506	13,096	3,807	34,409	2.39%
Grand Total	**124,871**	**99,200**	**506,877**	**730,947**	**17.08%**

* This category includes: GTEx, 1KG, UKBB, and the Qatar Genome Project, as well as the FinnGen and MGB biobank samples when no phenotype was specified
^ includes diseases like Crohn's disease, irritable bowel syndrome, interstitial cystitis, ulcerative colitis
** Neurodevelopmental controls are unaffected parents of children with confirmed or suspected de novo cause of their neurodevelopmental disorder

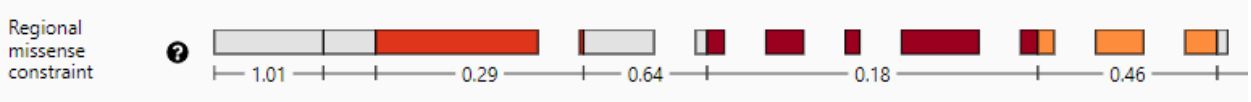
What version of gnomAD should I use?

The ideal version of gnomAD depends on the analysis you are performing:

- **Version 2 is smaller than v4.0 but has several strengths and additional tools:**
  - Gene constraint metrics such as pLI, LOEUF, and z-score are more established in gnomAD v2.1.1 (v4 metrics are in "beta stage")

- Phenotype data was used in v2 to remove individuals known to have severe early onset disease, when possible
- Several tools in v2 are not yet available in v4:
  - **Mean Pext track**: This track shows expression of exons across transcripts, using data from Genotype-Tissue Expression (GTEx) project

  - **Regional missense constraint track**: This track shows regions of the gene with less missense variation than expected, indicating regions where missense variants may be deleterious
    - https://www.biorxiv.org/content/10.1101/148353v1



  - **Variant co-occurrence**: predicts if 2 variants are in cis or in trans
    - https://gnomad.broadinstitute.org/variant-cooccurrence

- **Version 4 is a powerful tool due to its large size:**
  - Version 4 is 5x bigger than v2 and v3 combined
  - Includes 2.9x more non-European individuals than previous versions
  - Like previous versions, v4 is depleted for individuals with early-onset disease
    - version 4 can be a powerful tool for evaluating rare VUS

References

- The gnomAD Browser:
  - https://gnomad.broadinstitute.org/
  - https://gnomad.broadinstitute.org/about
  - https://gnomad.broadinstitute.org/stats
  - https://gnomad.broadinstitute.org/help

- ClinGen Biocurator Working Group videos featuring the gnomAD Production Team:
  - Video one: https://youtu.be/1_VXzxF3V98?si=cQuD6_utjJ2V4ALg
  - Video two: https://youtu.be/NeTj4BwLwpQ?si=QTMfpqXPEdaQVdQq

Author Information and Affiliations

Casey Brewer, PhD, FACMG

Cincinnati Children's Hospital

Division of Human Genetics

Date Created: 5/6/2024