MSSNG - Data Overview

The MSSNG project makes data available to approved researchers with the goal of improving our understanding of Autism Spectrum Disorder (ASD). A publication describing the MSSNG resource and its associated data (titled "**Genomic architecture of autism from comprehensive whole-genome sequence annotation")**, can be found here:

https://pubmed.ncbi.nlm.nih.gov/36368308

Data

Overview

Data are available for 13790 individuals (13837 genome samples[1]), including:

- 6002 affected individuals (4761 males, 1241 females)
- 7466 unaffected individuals (3647 males, 3819 females)
- 252 autism-related affected individuals (121 males, 131 females)
- 70 unknown affection individuals (31 males, 39 females)

[1]A few individuals were sequenced more than once.

Types of data

The following types of data for these individuals are available:

- Sample/subject data
- Aligned reads
- Variants

Sample/subject data

Sample/subject data are divided into three tables: `subject`, `measures`, and `subject_sample`. For an overview of the measures available, please see the publication associated with MSSNG, specifically supplementary table S1D.

Aligned reads

In this MSSNG database release, alignments in CRAM file format are available for 12,104 samples sequenced on Illumina platforms and aligned to the GRCh38 human reference assembly. For further information about the alignment pipeline for MSSNG Illumina samples,

please read this document. For Complete Genomics samples, information about liftover and post-processing of variants can be found in this document.

CRAM and VCF files are available to researchers by following the Process for Researchers to Access MSSNG CRAM and VCF files.

Variants

An individual sample's variants can be found in BigQuery tables. Variants are available for all 13,837 Illumina and Complete Genomics genome samples. For further information about the variant calling pipeline for MSSNG Illumina samples, please read this document. Complete Genomics variant calls are processed using a custom pipeline to liftover calls generated by Complete Genomics to GRCh38. More information can be found here.

Copy Number Variants (CNVs)

For samples sequenced on Illumina platforms, copy number variants (CNVs) were detected using ERDS (Zhu et al, 2012) and CNVnator (Abyzov et al, 2011) as previously described (Trost et al, 2018).

Annotations

Illumina and Complete Genomics variant annotations were generated using an ANNOVAR-based custom pipeline, following ANNOVAR priority rules to report variants. For further information on the databases/data sources and versions used to derive annotations, please read this document.

Putative *de novo* variants

De novo variants are those called in probands but not in the parents' genomes. *De novo* variants are available for families for which both parents were sequenced (4795/6002 affected).

Sanger-validated variants

For the variants in the `variants_sanger` table, Sanger validation results are available; positive as well as negative results are reported. Only a small fraction of variants have undergone Sanger validation.

MSSNG data locations

MSSNG data are hosted on Google Cloud Platform services as follows:

	Google BigQuery	Google Cloud Storage
Aligned reads		X
Called variants	X	X
Sample/subject data	X	

Access

The following describes ways to access the MSSNG data stored in the Google Cloud:

- To get started, you can access the MSSNG researcher portal at https://research.mss.ng
- If you would like to issue custom queries against the MSSNG BigQuery tables, then you will need to create a Google Cloud Project.
- If you would like to download the CRAM and VCF files, please refer to the document on the process for researchers to access MSSNG CRAM and VCF files.

References

Zhu, M. et al. Using ERDS to infer copy-number variants in high-coverage genomes. American Journal of Human Genetics 91:408–421 (2012).

Abyzov, A. et al. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research 21:974–984 (2011).

Trost, B. et al. A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. American Journal of Human Genetics 4:142-155 (2018).