MSSNG - Researcher README

The MSSNG project makes data available to trusted researchers with the goal of improving our understanding of Autism Spectrum Disorder (ASD). An associated publication can be found at https://www.ncbi.nlm.nih.gov/pubmed/28263302

The purpose of this document is to provide an overview of the available data and associated tools, as well as basic examples of using the tools to access the data.

Notable updates to the data or portal can be found in the CHANGELOG.

Table of Contents

Data

Overview

Data are available for 11312 individuals (11359 genome samples[1]), including:

- 5102 affected individuals (4074 males, 1028 females)
- 6079 unaffected individuals (3033 males, 3046 females)
- 131 autism-related affected individuals (61 males, 70 females)

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

Individuals typically belong to family trios (two parents and one affected child) or quads (two parents and two affected children). A few other family structures are also present. A total of 4258 families are available.

Family members	Families	Individuals
1	1149	1149
2	154	308
3	2128	6384
4	697	2788
5	106	530
6	17	102
7	6	42
9	1	9

This provides, in summary:

Genome Samples	Individuals	Affected	Autism-related affected	Unaffected	Sequencing Technology
1738	1732	729	1	1002	Complete Genomics
8830	8828	4129	111	4588	Illumina HiSeqX
747	747	244	13	490	Illumina HiSeq2000
44	44	15	6	23	Illumina HiSeq2500

A summary of the DNA source of the samples is as follows:

DNA Source	Genome Samples
Blood	9529
Cell line	359
Saliva	7
White blood cell	381
Unknown	1083

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`. These tables are available as BigQuery tables (`idyllic-analyst-574:db6_release` data-set); see the MSSNG Data Locations sub-section and the Examples section for how to access and query BigQuery tables.

subject

The `subject` table provides basic information about each individual, such as sex, date of birth, and whether they are affected:

Field	Description
INDEXID	Unique identifier for the individual
FATHERID	Identifier of the individual's father
MOTHERID	Identifier of the individual's mother
AFFECTION	"1" if unaffected, "2" if affected, "0" if autism-related affected
SEX	"M" (male), "F" (female)

FAMILYID	Family identifier
FAMILYTYPE	“SPX” (simplex), “MPX” (multiplex)
DOB	Date of birth; yyyy-mm-dd. (if information available). Day set to “01” for anonymization.

Subjects may have multiple samples, and each will be referenced as a separate row in the `subject_sample` table.

subject_sample

The `subject_sample` table provides metadata about all genome samples available in the MSSNG database. `SUBMITTEDID` is the genome sample identifier that you should use to join subject/sample data to the variant data ‘`call.name`’ field.

Field	Description
SUBMITTEDID	Unique identifier for the genome sample. Note that while this value is usually the same as the INDEXID, that is not always the case. This corresponds to ‘call.name’ in the variant tables.
INDEXID	Unique identifier of the individual found in the `subject` table
DNASOURCE	biological sample type used as DNA source: "Blood" (fresh blood), “White blood cell” (frozen as opposed to fresh white blood cells), "Cell line" (lymphoblastoid cell line), “Saliva”
PLATFORM	sequencing platform: "Illumina HiSeq", (HiSeq2000), “Illumina HiSeq2500”, “Illumina HiSeqX”, "Complete Genomics" (different pipeline versions)
NIMHID	NIMH identifier
RUDCRID	Rutgers repository identifier
COMMENTS	Any specific comments regarding a sample
SOFTWARE_VERSION	For Complete Genomics samples only, the software version used to analyse sample
PREDICTED_ANCESTRY	Predicted ancestry of sample. Consensus of computationally derived predictions from two tools

father_SUBMITTEDID	Unique identifier of fathers genome sample, if it exists in the dataset and has been sequenced on the same platform as the genome sample
mother_SUBMITTEDID	Unique identifier of mothers genome sample, if it exists in the dataset and has been sequenced on the same platform as the genome sample

measures

Psychometric test results are typically available only for affected individuals and use established scales. Subjects are identified by `INDEXID`. Test results are linked to the date at which the test(s) were run (TESTDATE). For a subset of subjects, measurements for the same test performed on different dates are available and need to be collapsed if used for analysis. Please see this spreadsheet for a more detailed explanation on the measures available.

measures: table with 4 columns organized in tidy format (many records per subject)

measures

Field	Description
INDEXID	Unique identifier for the individual
CODE	Identifier for the type of test
TESTDATE	Date in which the test was administered
MEASURE	Test result

Aligned Reads

In this MSSNG database release, alignments in CRAM file format are available for 9,621 samples sequenced on Illumina platforms and aligned to 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, please see 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.

The `clustered_compact_variants_ilmn` table contains jointly genotyped variants for all samples sequenced on Illumina platforms. The `clustered_compact_variants_cg` table contains individually genotyped variants for all samples sequenced by Complete Genomics. In order to improve query performance and data organization, BigQuery's table clustering capabilities have been employed. Column descriptions are available by clicking on the table in BigQuery, and viewing the schema.

For representing insertions and deletions we follow the VCF convention of capturing the first reference base before the insertion or deletion within the variant locus (e.g. reference_bases: AT, alternate_bases: A, for a deletion of T).

Variants are available for all 11,359 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.

MSSNG Portal

When performing a variant query within the MSSNG Portal website, a number of columns are viewable, some by default and some can be viewed by modifying the column visibility. The following is a description of each field.

Field	Description
Sample	Sample name
Sequencing platform	NGS platform; Illumina or Complete Genomics
Sex	Gender of the sample
Family ID	Family ID
Sanger Validated	Confirmed by Sanger sequencing

Sanger Inheritance	Inheritance by Sanger sequencing
Chr	Chromosome (autosomes 1-22 and sex chromosomes X, Y)
Start	Start position (0-positional system)
End	End position
Reference allele, Alternate allele	The reference allele and alternative allele(s) observed for this variant and represented in forward strand. For insertions, the **Alternate allele** includes the inserted sequence as well as the base preceding the insertion. For deletions, the **Alternate allele** allele is the base before the deletion
Zygosity	Heterozygous, homozygous, hemizygous
Genotype	Genotype, represented as Reference allele, Alternate allele or Alternate allele, Alternate allele or Alternate allele, del/ins/sub
De Novo	'High-confidence' if this variant a rare de novo variant
Inheritance	string in the format - ,0:0,1:0,1:ref-alt|mat-pat 0,0:0,1:0,1- ":" separated list of maternal (0,0: homozygous reference), paternal (0,1: heterozygous) and child (0,1: heterozygous) genotypes ref-alt|mat-pat - child is heterozygous, where "ref" is maternally inherited and "alt" is paternally inherited
FILTER	Filter status: PASS if this position has passed all filters
Read depth	Depth of coverage

Allelic depth	Reference and alternate allele counts (comma separated)
Genotype quality	Variant Confidence or Quality by Depth represents the Phred-scaled confidence that the genotype assignment is correct
Call.EHQ*	Complete Genomics, calibrated haplotype quality based on equal allele fraction assumption
Call.HQ	Complete Genomics, haplotype quality
Max frequency 1000 Genomes	Maximum allele frequency from 1000 Genome data-set. Data set includes five populations (caucasian-european, latin americans, east-asian, south-asians, african-american)
Max frequency GnomAD genome	Maximum allele frequency from Genome Aggregation Database(gnomAD) for whole-genome data
Max frequency GnomAD exome	Maximum allele frequency from Genome Aggregation Database(gnomAD) for exome data
Max frequency ExAC	Maximum allele frequency from ExAC 65000 data set
Max Frequency (MSSNG)	Maximum parental frequency observed among different MSSNG platforms
Max Frequency	Maximum frequency from all frequency data sets (ie, 1000g, ExAC, gnomAD and Complete Genomics)
Minimum Reference Fraction (MSSNG)	Fraction of parental samples with homozygous reference calls
RefSeq ID	Combined Annovar output on coding sequence mapping and effect; composed

	of: (a) for coding exonic changes (typeseq “exonic”): gene official symbol, RefSeq transcript isoform ID, position in the coding sequence, amino acid change; (b) for core splice site changes (typeseq "exonic"): gene official symbol, RefSeq transcript isoform ID, exon number, coding sequence position and change
Typeseq priority	Type of sequence overlapped, with respect to known genes/transcripts and their coding / noncoding status: (a) “exonic” represents coding exons, (b) “exonic:splicing” represents the beginning/end of coding exons which may also affect splicing, (c) “splicing” represents core splicing site (2 bp on the intron side of intron-exon and exon-intron junctions), (d) "ncRNA_exonic" represents exons of non-coding RNA genes, (e) "ncRNA_splicing" represents core splicing sites of non-coding RNA genes , (f) “UTR5” represents 5’ untranslated region, (g) “UTR3” represents 3’ unstranslated region, (h) “upstream” represents 1kb ubstream of TSS, (i) “downstream” represents 1kb downstream of TSS and (j) “intergenic” represents intergenic regions (beyond upstream/downstream threshold(1kb)). For variants with multiple sequence overlaps (eg, exonic for one transcript and intronic for other), we used annovar prioritization scheme to prioritize them. (http://annovar.openbioinformatics.org/en/latest/user-guide/gene/).
Effect priority	Type of effect on the coding sequence: (a) “synonymous SNV”, (b) "nonsynonymous

	SNV", (c) "stopgain SNV", (d) "frameshift deletion", (e) "frameshift insertion", (f) "frameshift substitution", (g) "nonframeshift deletion", (h) "nonframeshift insertion", (i) "nonframeshift substitution" , (j) “stoploss SNV”. Prioritized effect is selected for variants with multiple effects (http://annovar.openbioinformatics.org/en/latest/user-guide/gene/).
Gene Symbol	Official gene symbol
Entrez Id	Entrez-gene id
OMIM Phenotype	Omim disorder/disease description when available for the corresponding omim gene accession
CGD_disease	The Clinical Genomics Database is compiled by curators and maintained by the NHGRI (National Human Genome Research Institute); for every gene in the database, the CGD provides a list of one or more genetic disorders and a mode of inheritance; this field reports the genetic disorder(s)
CGD_inheritance	The Clinical Genomics Database is compiled by curators and maintained by the NHGRI (National Human Genome Research Institute); for every gene in the database, the CGD provides a list of one or more genetic disorders and a mode of inheritance; this field reports the mode of inheritance (AD, AR, AD/AR, XL, more complex modes); since the CGD mode of inheritance is directly added by a curator and it's tied to specific genetic disorder(s), it could be considered more accurate than the

	mode of inheritance for top-level HPO phenotypes
Comment	Annotation database issues (ie, ambiguous liftover or incomplete ORF for gene transcript)
LOF observed/expected (oe) metric - CI upper bound	GnomAD per-gene constraint score for LOF
Missense observed/expected (oe) metric - CI upper bound	GnomAD per-gene constraint score for missense
Probability of being loss-of-function intolerant	GnomAD pLI score
Probability of being intolerant of homozygous but not heterozygous LOF variants	GnomAD pRec score
Clinvar significance	Overall ClinVar significance code; “pathogenic” is the code of interest for rare disorders (https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/)
Clinvar Significance Simple	Expected values 0,1 or -1; 0 = no current value of pathogenic; 1 = at least one record submitted with pathogenic/likely pathogenic; -1 = no values for clinical significance at all for this variant or set of variants. Used for the "included" variants that are only in ClinVar because they are included in a haplotype or genotype with an interpretation
Effect - Impact	Pre-computed effects and impacts for the variant. This was computed using various annotation factors such as sequence overlap, coding effect, prediction scores etc.

Affection	0 = Autism related affected, 1 = unaffected, 2 = affected

Annotations

The `annotations_ilmn` (Illumina) and `annotations_cg` (Complete Genomics) variant annotations are 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. Column descriptions are available for both tables, by clicking on either one in BigQuery and viewing the schema.

See the example section for how to join this table to the variant table.

Putative De novo Variants

De novo variants are called in probands but not present in the parent's genomes. De novo variants are available for families where both parents were sequenced (3881/5114 affected). The `variants_de_novo` table lists all putative de-novo variants. The variants with annotations can also be viewed as a spreadsheet.

Field	Description
id	as in `annotation` table
reference_name	as in `variant` table
start	as in `variant` table
end	as in `variant` table
reference_bases	as in `variant` table
alternate_bases	as in `variant` table
PLATFORM	as in `subject_sample` table
COMMENT	Any comments about this variant
SUBMITTEDID	as in `subject_sample` table

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 minority of variants have undergone Sanger validation.

Field	Description
id	as in `annotations_*` table
reference_name	as in `clustered_compact_variants*` table
start	as in `clustered_compact_variants*` table
end	as in `clustered_compact_variants*` table
reference_bases	as in `clustered_compact_variants*` table
alternate_bases	as in `clustered_compact_variants*` table
Sanger_validated	possible values: YES (variant found), NO (variant not found), NULL
Sanger_inheritance	possible values: de novo, maternal inherited, paternal inherited, not maternal, inherited, no variant, variant present, NULL
PLATFORM	as in `subject_sample` table
SUBMITTEDID	as in `subject_sample` table

MSSNG Data Locations

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

	Google BigQuery	Google Cloud Storage
Aligned Reads		X
Called Variants	X	X
Sample/subject Data	X	

Google BigQuery is a service designed for storing generic structured data and allowing for querying over massive datasets in seconds. Google BigQuery supports a SQL-like query

language which can be accessed via the BigQuery web-based interface, command line tool, or programmatic API. This allows access to data from any data analysis tool (such as Python) that supports the Google BigQuery API.

Google Cloud Storage is a repository for storing and sharing files. Cloud Storage supports storing and retrieving files using a web-based interface, command-line tool, or programmatic API. This allows file-based access to data from any data analysis tool (such as Python).

Some data is also available as a direct download via the MSSNG Portal. This includes complete subject/sample information as well as annotated de novo variants.

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. See the instructions below for getting started using BigQuery.
- 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.

Once you have created your own Google Cloud project, you can try some of the examples in the next section.

BigQuery Examples

Subject/sample data

Subject/sample and variant data is stored in Google BigQuery. Genomics data in BigQuery is most commonly accessed through the BigQuery web interface. Subject/sample data is also available for download from the MSSNG Portal.

BigQuery web interface

The BigQuery web interface can be used for issuing ad hoc queries over the genomic variant data and subject/sample data.

Setup

The following steps demonstrate accessing the MSSNG subject/sample data.

1. Go to https://bigquery.cloud.google.com
2. Set the active project
    a. If the project name in the left-hand navigation is **MSSNG Portal**, then it must be changed:
        i. Click on the drop down icon beside MSSNG Portal in the left-hand navigation pane.
        ii. Pick *'Switch to project'* in the menu, and then select your Google Cloud Project from the list.
3. Add the MSSNG project to the list of available datasets
    a. Click on the drop down icon beside your project name in the left-hand navigation pane.
    b. Pick *'Switch to project'* in the menu, and then select *'Display Project'*.
    c. In the Add Project dialog, enter the Project ID "`idyllic-analyst-574`"

In the left-hand navigation pane, you should see listed MSSNG project:

- `idyllic-analyst-574`

If you click on the `idyllic-analyst-574` project, it should expand to show the dataset:

- `db6_release`

If you click on the `db6_release` dataset it should expand to show (among many others), the tables:

    - `annotations_cg`
    - `annotations_ilmn`
    - `clustered_compact_variants_cg`
    - `clustered_compact_variants_ilmn`
    - `subject`
    - `subject_sample`
    - `variants_de_novo`

- variants_sanger

Subject/Sample Data Example

Your first example query will be on the subject table. Clicking on the `subject` table will open the New Query pane on the right hand side.

In the New Query text area enter the query:

```
#standardSQL
SELECT
  sex,
  COUNT(INDEXID) AS count
FROM
  `idyllic-analyst-574.db6_release.subject`
GROUP BY
  sex
ORDER BY
  sex
```

Clicking on the Run Query button should generate results in a few seconds which looks like:

Row	SEX	count	
1	F	4144	
2	M	7168	

To see the number of autism affected individuals, change the query to:

```
#standardSQL
SELECT
  sex,
  COUNT (INDEXID) AS count
FROM
  `idyllic-analyst-574.db6_release.subject`
WHERE
  affection = '2'
GROUP BY
  sex
ORDER BY
  sex
```

(note that: `AFFECTION = '2'` means: autism affected)

Clicking on the Run Query button should generate results in a few seconds which looks like:

Row	gender	count	
1	F	1028	
2	M	4074	

Genomic Variants Examples

Genomic variants are stored in the `db6_release.clustered_compact_variants_cg` and `db6_release.clustered_compact_variants_ilmn` tables (described above). This table uses some features of Google BigQuery not commonly seen in relational databases (which you may already be familiar with), namely Array fields.

Each record in the `clustered_compact_variants_*` tables describes a variant which has been called at least once within the set of samples. Within the variant record is a `call` field, which contains a reference to all calls of this variant.

The schema for the `clustered_compact_variants_*` tables can be found by:

1. Selecting one of the `clustered_compact_variants_*` tables in the left hand pane of the BigQuery interface
2. The Schema button in the right hand pane should be selected by default.

A button for Table Details should also be displayed. Select this to view information such as the size of the table and number of rows. To see a sampling of the data, select the Preview button.

Many example queries which can be used on the `clustered_compact_variants_*` tables can be found here.

To build your own, more sophisticated queries, see the BigQuery Query Reference.