# 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 individuals | Autism-related affected individuals | Unaffected individuals | 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. |

**subject_sample**

The `subject_sample` table provides metadata about all genome samples available in the MSSNG database. Subjects may have multiple samples, and each will be referenced as a separate row in the `subject_sample` table. 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 the father's 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 the mother's genome sample, if it exists in the dataset and has been sequenced on the same platform as the genome sample |
|---|---|
| sample_QC | "ok": sample passes all QC;"QC_FAIL": sample failed multiple QC procedures; "CNV_QC_FAIL": sample failed CNV QC; "SNP_DN_QC_FAIL": sample failed small-variant QC; "exclude from downstream analyses": miscellaneous problems with 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 on which the tests 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. The table is organized in tidy format (many records per subject)

| Field | Description |
|---|---|
| INDEXID | Unique identifier for the individual |
| CODE | Identifier for the type of test |
| TESTDATE | Date on 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 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.

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; for example, a deletion of T would be represented by reference_bases = AT and alternate_bases = A.

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](#).

When performing a variant query within the MSSNG Portal website, a number of columns are viewable, some by default and some by modifying the column visibility. The following is a description of each field that is available in the variant query tab of the MSSNG Portal website. These can be toggled by using the "Show/hide columns" feature.

| Displayed by default? | Field | Description |
|---|---|---|
| yes | Sample | Sample name |
| yes | Sequencing platform | NGS platform; Illumina or Complete Genomics |
| yes | Sex | Gender of the sample |
| no | Family ID | Family ID |

| | | |
|---|---|---|
| no | Sanger Validated | Whether confirmed by Sanger sequencing (Yes/No) |
| no | Sanger Inheritance | Inheritance by Sanger sequencing |
| yes | Chr | Chromosome (autosomes 1-22 and sex chromosomes X, Y) |
| yes | Start | Start position (0-positional system) |
| yes | End | End position |
| yes | Reference allele, Alternate allele | The reference allele and alternate 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 is the base before the deletion. |
| yes | Zygosity | Heterozygous, homozygous, hemizygous |
| yes | Genotype | Genotype, represented as "reference allele, alternate allele" or "alternate allele, alternate allele" or "alternate allele, del/ins/sub" |
| yes | De Novo | 'High-confidence' if this variant is a rare *de novo* variant |
| yes | Inheritance | string in the format "0,0:0,1:0,1:ref-alt\|mat-pat", where "0,0:0,1:0,1" is a colon-separated list of maternal (0,0: homozygous reference), paternal (0,1: heterozygous) and child (0,1: heterozygous) genotypes, and "ref-alt\|mat-pat" indicates the inheritance (in this case, the child is heterozygous, where "ref" is maternally inherited and "alt" is paternally inherited). |

| no | FILTER | Filter status: PASS if this position has passed all filters |
|---|---|---|
| yes | Read depth | Depth of coverage |
| yes | Allelic depth | Reference and alternate allele counts (comma separated) |
| yes | Genotype quality | Variant Confidence or Quality by Depth represents the Phred-scaled confidence that the genotype assignment is correct |
| no | Call.EHQ* | Complete Genomics, calibrated haplotype quality based on equal allele fraction assumption |
| no | Call.HQ | Complete Genomics, haplotype quality |
| yes | Max frequency 1000 Genomes | Maximum allele frequency from 1000 Genome dataset. Dataset includes five populations (caucasian-european, latin americans, east-asian, south-asians, african-american) |
| yes | Max frequency GnomAD genome | Maximum filtering allele frequency from the Genome Aggregation Database (gnomAD) for whole-genome data |
| no | Max frequency GnomAD exome | Maximum filtering allele frequency from Genome Aggregation Database (gnomAD) for exome data |
| no | Max Frequency | Maximum of filtering allele frequencies from both gnomAD datasets and allele frequencies from Complete Genomics |
| yes | RefSeq ID | Combined ANNOVAR output on coding sequence mapping and effect; composed of: (a) for coding exonic changes (typeseq "exonic"): official gene symbol, RefSeq transcript isoform ID, position in the coding sequence, amino acid change; (b) for core |

| | | |
|---|---|---|
| | | splice site changes (typeseq "exonic"): official gene symbol, RefSeq transcript isoform ID, exon number, coding sequence position and change |
| no | 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 exonic and splicing overlaps (for multi-transcript genes), (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' untranslated region, (h) "upstream" represents 1kb upstream of TSS, (i) "downstream" represents 1kb downstream of TSS, and (j) "intergenic" represents intergenic regions beyond the upstream/downstream threshold (1kb). For variants with multiple sequence overlaps (for example, exonic for one transcript and intronic for other), we used the ANNOVAR prioritization scheme to prioritize them. (http://annovar.openbioinformatics.org/en/latest/user-guide/gene/). |
| no | 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/). |
| yes | Gene Symbol | Official gene symbol |
| no | Entrez Id | Entrez-gene ID |

| | | |
|---|---|---|
| no | OMIM Phenotype | OMIM disorder/disease description when available for the corresponding OMIM gene accession |
| no | CGD_disease | The Clinical Genomics Database (CGD) is compiled by curators and maintained by the National Human Genome Research Institute (NHGRI); 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) |
| no | CGD_inheritance | This field reports the CGD mode of inheritance (AD, AR, AD/AR, XL, more complex modes). Since the CGD mode of inheritance is directly added by a curator and is tied to specific genetic disorder(s), it could be considered more accurate than the mode of inheritance for top-level HPO phenotypes. |
| no | Comment | Annotation database issues (for example, ambiguous liftover or incomplete ORF for gene transcript) |
| no | LOF O/E metric - CI upper bound | GnomAD per-gene constraint score for LOF |
| no | Missense O/E metric - CI upper bound | GnomAD per-gene constraint score for missense |
| no | pLI | GnomAD pLI score |
| no | pRec | GnomAD pRec score |
| yes | ClinVar significance | Overall ClinVar significance code; "pathogenic" is the code of interest for rare disorders (https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/) |
| yes | 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 |

| | | |
|---|---|---|
| yes | 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. |
| no | Affection | 0 = Autism-related affected, 1 = unaffected, 2 = affected |
| no | dbSNP ID | dbSNP identifier |
| no | Min Ref Fraction ilm (MSSNG) | min of MSSNG reference proportions for Illumina |
| no | Min Ref Fraction cgi (MSSNG) | min of MSSNG reference proportions for CGI |
| no | Max int frequency ilm (MSSNG) | max of MSSNG Internal frequencies for Illumina |
| no | Max int frequency cgi (MSSNG) | max of MSSNG Internal frequencies for CGI |
| yes | Sample QC | "ok": sample passes all QC;"QC_FAIL": sample failed multiple QC procedures; "CNV_QC_FAIL": sample failed CNV QC; "SNP_DN_QC_FAIL": sample failed small-variant QC; "exclude from downstream analyses": miscellaneous problems with sample |

## 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). For CNVnator, calls for which > 50% of the reads in the CNV region had zero mapping quality were removed (q0 filter), except for in homozygous autosomal deletions or X-linked deletions in males (with normalized read depth (NRD) < 0.03). For

samples sequenced by Complete Genomics (CG), CNV calls were used as provided, with all CNVs being >= 2 kb. Quality and frequency of CNVs were defined as follows:

· **High-quality CNVs** for Illumina samples were defined as those that are >=1 kb, are detected by both ERDS and CNVnator with >= 50% reciprocal overlap, and have <= 70% overlap with telomeres, centromeres, and segmental duplications ("Unclean genome overlap" column below). High-quality CNVs for Complete Genomics samples were defined as those with <= 70% overlap with telomeres, centromeres, and segmental duplications.

· **Rare CNVs** were defined as those detected at <=1% frequency in the QC-passing parental samples in MSSNG according to ERDS, CNVnator, and Complete Genomics, as well as <=1% frequency according to both ERDS and CNVnator in QC-passing samples from 2,504 unrelated individuals from the 1000 Genomes Project sequenced on the Illumina NovaSeq platform.

A CNV was tagged as "high quality rare" (see "High quality rare" column below) if it was both high-quality and rare according to the criteria defined above. The table below describes the columns found when querying the CNV table, as well as noting column visibility. Column visibility can be modified under "Show/hide columns" in the MSSNG Portal website.

| Displayed by default? | Column name (BigQuery) | Column name (UI) | Description |
|---|---|---|---|
| yes | sample | Sample | sample identifier |
| yes | chr | Chr | chromosome |
| yes | start | Start | start position |
| yes | end | End | end position |
| yes | CNV_type | CNV type | CNV type (DEL or DUP) |
| yes | size | Size | size of CNV in base pairs |

| yes | overlap | Overlap | overlap between CNV calls (Illumina samples only; format is overlap of ERDS call with CNVnator call \| overlap of CNVnator call with ERDS calls) |
|---|---|---|---|
| no | copy_number | Copy number | estimated copy number (for Illumina, the copy number reported by ERDS is given) |
| no | GC_content_percent | GC content (%) | % GC content |
| no | cytoband | Cytoband | cytoband |
| yes | gene_symbol | Gene symbol | official gene symbols with transcript overlap |
| no | gene_entrez_id | Gene ID | Entrez gene IDs with transcript overlap |
| yes | exon_symbol | Exon symbol | official gene symbols with exon overlap |
| no | exon_entrez_id | Exon ID | Entrez gene IDs with exon overlap |
| no | CDS_symbol | CDS symbol | Official gene symbols with CDS overlap |
| no | CDS_entrez_id | CDS ID | Entrez gene IDs with CDS overlap |
| yes | gnomAD_pLI | gnomAD pLI | gnomAD pLI value |
| no | gnomAD_oe_lof_upper | gnomAD o/e LoF upper | upper bound of gnomAD observed/expected ratio for loss of function variants |
| no | RepeatMasker_percent_overlap | RepeatMasker overlap (%) | % overlap with RepeatMasker repeats |
| yes | unclean_genome_percent_overlap | Unclean genome overlap (%) | % overlap with segmental duplications, centromeres, and telomeres |
| no | MPO_nervous_system | MPO nervous system | mouse phenotype ontology terms related to nervous system phenotypes |
| no | HPO_nervous_system | HPO nervous system | human phenotype ontology terms related to nervous system phenotypes |

| | | | |
|---|---|---|---|
| no | CGD_disease_inheritance | CGD disease inheritance | gene symbols with transcript overlap in the Clinical Genomic Database |
| yes | OMIM_morbid_map | OMIM Morbid Map | OMIM Morbid Map |
| no | ISCA_region | ISCA region | genomic disease region from the International Standards for Cytogenomic Arrays database |
| no | CNV_ISCA_percent_overlap | CNV ISCA overlap (%) | % length of CNV overlapped by International Standards for Cytogenomic Arrays region |
| no | DECIPHER_region | DECIPHER region | genomic disease region from the DECIPHER database |
| no | CNV_DECIPHER_percent_overlap | CNV DECIPHER overlap (%) | % length of CNV overlapped by DECIPHER region |
| no | DGV_N_studies | DGV N studies | number of studies in which there is an overlapping CNV in the Database of Genomic Variants (DGV) (50% reciprocal overlap) |
| no | DGV_percent_freq_subjects_all_studies | DGV freq subjects all studies (%) | DGV frequency, any study (50% reciprocal overlap) |
| no | DGV_percent_freq_subjects_coverage_studies | DGV freq subjects coverage studies (%) | DGV frequency, only studies with coverage (50% reciprocal overlap) |
| no | DGV_percent_freq | DGV freq (%) | DGV frequency (50% reciprocal overlap) |
| yes | CG_percent_freq_MSSNG_parents | CG freq parents (%) | percentage of MSSNG parents sequenced on the Complete Genomics platform in which the CNV is detected (50% reciprocal overlap) |
| yes | CNVnator_percent_freq_MSSNG_parents_HiSeqX | CNVN freq parents HiSeqX (%) | percentage of MSSNG parents sequenced on Illumina HiSeq X in which the CNV is detected by CNVnator (50% reciprocal overlap) |

| | | | |
|---|---|---|---|
| no | CNVnator_percent_f req_MSSNG_ parents_HiSeq2000 | CNVN freq parents HiSeq (%) | percentage of MSSNG parents sequenced on Illumina HiSeq 2000/2500 in which the CNV is detected by CNVnator (50% reciprocal overlap) |
| yes | ERDS_percent_freq _MSSNG_ parents_HiSeqX | ERDS freq parents HiSeqX (%) | percentage of MSSNG parents sequenced on Illumina HiSeq X in which the CNV is detected by ERDS (50% reciprocal overlap) |
| no | ERDS_percent_freq _MSSNG_ parents_HiSeq2000 | ERDS freq parents HiSeq (%) | percentage of MSSNG parents sequenced on Illumina HiSeq 2000/2500 in which the CNV is detected by ERDS (50% reciprocal overlap) |
| no | ERDS_percent_freq _1000G | ERDS freq 1000G (%) | percentage of 1000 Genomes Project samples in which the CNV is detected by ERDS (50% reciprocal overlap) |
| no | CNVnator_percent_f req_1000G | CNVN freq 1000G (%) | percentage of 1000 Genomes Project samples in which the CNV is detected by CNVnator (50% reciprocal overlap) |
| yes | putative_inheritance | Putative inheritance | putative inheritance - possible values are "Maternal", "Paternal", "NA" (no parents available), "Inherited_Ambiguous" (both parents have the CNV), "One_parent_sequenced" (only one parent available), "Ambiguous" (inheritance is ambiguous), and "P_denovo" (putative *de novo* variant) |
| yes | high_quality_rare | High quality rare | "High quality rare" if the CNV meets the criteria for this described above, or "-" otherwise. |
| yes | sample_QC | Sample QC | "ok": sample passes all QC;"QC_FAIL": sample failed multiple QC procedures; "CNV_QC_FAIL": sample failed CNV QC; "SNP_DN_QC_FAIL": sample |

| | | | failed small-variant QC; "exclude": miscellaneous problems with sample |
|---|---|---|---|
| no | FAMILYID | Family ID | family identifier |
| yes | platform | Platform | sequencing platform (Complete Genomics, Illumina HiSeq 2000, Illumina HiSeq 2500, or Illumina HiSeq X) |

**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 those called in probands but not in the parents' genomes. *De novo* variants are available for families for which 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 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 | |

Google BigQuery is a service designed for storing generic structured data and allowing for querying over massive datasets in seconds. BigQuery supports standard SQL queries, 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.

Some data are also available as a direct download via the MSSNG Portal.  These include 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 are stored in Google BigQuery.  Genomics data in BigQuery is most commonly accessed through the [BigQuery web interface](#). Subject/sample data are 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://console.cloud.google.com/bigquery](https://console.cloud.google.com/bigquery)

2. 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`

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 that 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 that 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](#).

# 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).