MSSNG Portal Quickstart

***Note:** Please read the Researcher ReadMe document before reading this document, as the ReadMe provides a background and understanding of the data content, structure and how to access it.

Searching for and analysing variants

Variant queries

Use the simple search (one box query) function to search, for example:

- All high-impact variants in a specific gene or small number of genes
- A specific variant, codon or amino acid change in all subjects
- Any variant modifying a specific genomic position, codon or amino acid in all subjects
- A specific known SNP (according to dbSNP ID) in all subjects
- All variants within a specific genomic interval

By default, the simple search returns “PASS” variants found in affected individuals that are predicted to be damaging (high-impact) and have a frequency of <= 0.01 in large scale sequencing projects (ExAC, 1000 Genomes and GnomAD (exomes and genomes)).
Note: The simple variant search queries the BigQuery table [idyllic-analyst-574.db6_release.portal_use_only_rare_feature_variants]

Use the advanced query to, for example:

- Select the type of variants returned (stop-gain, frameshift, missense, synonymous etc.)
- Restrict the variants returned to *de novo* variants
- Select the variants returned according to the predicted impact or population frequency
- Select the affection or relation of subjects included in the variant search
- Search for variants using an uploaded BED file

Note: The variant search queries the BigQuery table [idyllic-analyst-574.db6_release.portal_use_only_feature_variants]

Sample/Family Queries

The sample/family query uses a similar format to the variant query for samples .
Use the family query to search, for example:

- Putative de novo variants
- Rare variants

By default, the simple sample search returns “PASS” variants found in affected individuals that are predicted to be damaging and have a frequency of <= 0.05 in large scale sequencing projects (ExAC, 1000 Genomes and GnomAD (exomes or genomes)).
These settings can be modified using the advanced family search.
Note: The sample search queries the BigQuery table [idyllic-analyst-574.db6_release.portal_use_only_rare_feature_variants]

The results of variant or sample queries can be viewed in the portal with up to 500 variants returned or the full results table can be downloaded as .tsv files. A description of

column headings for results returned from the variant or sample searches can be found under “Results returned by the Portal” section of this document or in the Researcher ReadMe document.

BigQuery tables
If a search returns more than 500 variants, a warning message is displayed on the results page –
“**This query has exceeded the maximum of 500 variants, the first 500 are being returned.**”

The full set of variants can be retrieved using the Google BigQuery interface.
The webpage with results for any query has a “Toggle” button which when clicked shows the SQL query that was executed for the search. The SQL query can be copied and pasted into the BigQuery interface to obtain the full result set.
The URL to access the BigQuery interface is:
https://bigquery.cloud.google.com/table/idyllic-analyst-574:db6_release.portal_use_only_feature_variants

User authentication is required for access.

Variant quality guidelines
For samples sequenced on Illumina platforms, the best quality variants have the following attributes:

- GQ ≥ 99 for heterozygous variants or GQ ≥ 25 for homozygous variants.
- Depth ≥ 10
- Allelic ratio between 0.3 and 0.7 for heterozygous variants (NOTE: high quality variants with allelic ratio ≤ 0.3 could be mosaic) and ≥0.9 for homozygous variants. Allelic ratio can be calculated from the Allelic Depth.

Read pileups for selected variants may be inspected using the read viewer for visual confirmation.

For samples sequenced by Complete Genomics:
The basic filter is GQ ≥ 40 for heterozygous variants and GQ ≥ 20 for homozygous variants. For de novo variants, the recommended quality scores are:

- GQ ≥ 175
- Depth ≥ 10
- Allelic ratio between 0.3 and 0.7 (NOTE: high quality variant with allelic ratio ≤ 0.3 could be mosaic) and ≥ 0.9 for homozygous variants. Allelic ratio can be calculated from the Allelic Depth.

No read pileups available for Complete Genomics samples.

Additional filtering based on MSSNG parental frequency is recommended – (Max frequency MSSNG < 0.01)

Other type of searches

Gene query
Use the gene query to search for information for a chosen gene compiled from public databases:

Database	URL
Entrez	https://www.ncbi.nlm.nih.gov/gene
RefSeq Gene	https://www.ncbi.nlm.nih.gov/refseq/
Clinical Genomics Database (CGD)	https://research.nhgri.nih.gov/CGD/
Human Phenotype Ontology (HPO)	http://human-phenotype-ontology.github.io/
OMIM	https://www.omim.org/
GO	http://www.geneontology.org/

Phenotypes
Use the phenotypes query for phenotypic information or for direct access to the read viewer for a specific subject.

Definitions

PASS variants

The variant passes applied filters - for samples sequenced by Illumina - VQSLOD value of the variant meets the specified lod threshold. For samples sequenced by Complete Genomics, the variant is either tagged "PASS" or "VQHIGH”

Damaging variants

Variants are classified into high, medium or low categories according to the following criteria:

Category	Criteria for classification
High	LOF (stop gain, frameshift or splice site variants) , non-synonymous variants predicted to be deleterious or highly conserved by four or more methods (Sift, PolyPhen2, Mutation Assessor, Mutation Taster, CADD, PhyloP), splice site variants with SPIDEX spx_dpsi score > 10 or < -10 or highly conserved non-coding bases
Medium	non-synonymous variants predicted to be deleterious or highly conserved by two or more methods (Sift, PolyPhen2, Mutation Assessor, Mutation Taster, CADD, PhyloP), splice site variants with SPIDEX spx_dpsi score > 2.5 or < -2.5 or conserved non-coding bases (using relaxed cutoffs)
Low	All other non-synonymous variants, splice site variants with Spidex spx_dpsi score > 0 or < 0, all other non-coding variants and variants in UTRs.

Prediction methods used to classify missense variants:
SIFT
Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. Genome Res. 11, 863–874 (2001).

PolyPhen2
Adzhubei, I.A. et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249 (2010).

Mutation Assessor

Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 39, e118 (2011).

Mutation Taster
Schwarz, J.M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods 7, 575–576 (2010).

CADD
Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–315 (2014).

SPIDEX
Xiong, H.Y et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science. 347(6218):1254806 (2015).

Results returned by the Portal:

Column definitions

Column heading	Description
Sample	Subject ID
Sex	Sex of subject
Chr	Chromosome
Start	Start coordinate of variant (hg38; zero-base nomenclature)
End	End coordinate of variant
Reference allele	Reference allele
Alternate allele	Alternate allele observed
Genotype	Genotype observed in subject. Alleles separated by a comma with reference allele first
Zygosity	Zygosity of observed variant
Effect - Impact	Predicted impact of observed variant. High, medium and low indicate variants that are predicted to be damaging to RNA or protein function.
Max frequency 1000 Genomes	Maximum frequency observed among populations in 1000 Genomes project
Max frequency ExAC	Maximum frequency observed among populations in ExAc project
max frequency GnomAD (Exomes)	Maximum frequency observed among populations in GnomAD (Exomes) project
Max frequency GnomAD (WGS)	Maximum frequency observed among populations in GnomAD (WGS) project
Max frequency	Maximum of all of the above
Max frequency MSSNG	Maximum parental frequency observed among different MSSNG platforms
Inheritance	Observed inheritance of variant where available. "Unknown" indicated that parental genotypes are not

	available.
De Novo	Variants predicted to be high quality de novo variants indicated as "High-confidence"
Filter	Sequencing platform based quality filter - "PASS" indicates that the variant passed the threshold. NOTE: this is a low stringency filter
Read depth	Read depth at variant position
Allelic depth	Allelic depth (reads supporting the reference allele, reads supporting the alternate allele)
Sequencing platform	Sequencing platform of subject
RefSeq ID	Effect of variant on RefSeq transcripts predicted by Annovar
Gene Symbol	Gene impacted by variant (for variants in coding, splice site or untranslated region sequence)
Entrez Id	ID of gene impacted by variant in Entrez database
Sanger Validated	Variants confirmed by Sanger sequencing indicated by "true"
Sanger Inheritance	Inheritance pattern determined by Sanger sequencing
Affection	Affected status of subject
Family ID	Family ID of subject
Genotype quality	Genotype quality of variant. The higher the value, the greater the confidence in the variant's being true.
dbSNP ID	ID of variants know in dbSNP
Clinvar significance	Clinical significance of variant according to ClinVar
Clinvar significance simple	1 if one of the submission tagged the variant as pathogenic or likely pathogenic
CGD Disease	Disease associated with gene impacted by variant in Clinical Genomics Database (where available)
CGD Inheritance	inheritance pattern of disease assigned to gene impacted by variant in Clinical Genomics Database (where available)
OMIM Phenotype	Phenotype associated with gene impacted by variant in OMIM (where available)
Typeseq priority	Type of sequence overlapped, with respect to known genes/transcripts and their coding / noncoding status
Effect priority	Type of effect on the coding sequence
Call.HQ	Complete Genomics, haplotype quality
Call.EHQ	Complete Genomics, calibrated haplotype quality based on equal fraction assumption
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
Minimum Reference Fraction (MSSNG)	Fraction of parental samples with homozygous reference calls
references:	
1000 genomes project	Genomes Project, C., et al. A global reference for human genetic variation. Nature 526, 68-74 (2015)
ExAc	Lek, M., et al. Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv (2016).
GnomAD	http://gnomad.broadinstitute.org
Entrez	https://www.ncbi.nlm.nih.gov/gene
Clinical Genomics Database	https://research.nhgri.nih.gov/CGD/
OMIM	https://www.omim.org/
ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/