Genomics England Research

One of the goals of The 100,000 Genomes Project from Genomics England is to enable new medical research. Researchers will study how best to use genomics in healthcare and how best to interpret the data to help patients. The causes, diagnosis and treatment of disease will also be investigated. This is currently the largest national sequencing project of its kind in the world.

To achieve this goal Genomics England set up a *Research* environment for researchers and clinicians. OpenCGA, CellBase and IVA from OpenCB were installed as data platform. We loaded **64,078 whole genomes** in OpenCGA, in total **about 1 billion unique variants** were loaded and indexed in OpenCGA Variant Storage, and all the metadata and clinical data for samples and patients were loaded in OpenCGA Catalog. **OpenCGA was able to load and index about 6,000 samples a day**, executing the variant annotation and computing different cohort stats for the all the data run in less than a week. In summary, all data was loaded, indexed, annotated and stats calculated in less than 2 weeks. Genomic variants were annotated using CellBase and the IVA front-end was installed for researchers and clinicians to analyse and visualise the data. In this document you can find a full report of about the loading and analysis of the 64,078 genomes.

Table of Contents:

Genomic and Clinical Data

Clinical data and genomic variants of **64,078 genomes** were loaded and indexed in OpenCGA. In total we loaded more than **30,000 VCF files accounting for about 40TB** of compressed disk space. Data was organised in four different datasets depending on the genome assembly *(GRCh37* or *GRCh38)* and the type of study *(germline* or *somatic)*, and this was organised in OpenCGA in three different *Projects* and four *Studies:*

Project	Study ID and Name	Samples	VCF Files	VCF File Type	Samples /File	Variants
GRCh37 Germline	**RD37** Rare Disease GRCh37	12,142	5,329	Multi sample	2.28	**298,763,059**
GRCh38 Germline	**RD38** Rare Disease GRCh38	33,180	16,591	Multi sample	2.00	**437,740,498**
	CG38 Cancer Germline GRCh38	9,167	9,167	Single sample	1.00	**286,136,051**
GRCh38 Somatic	**CS38** Cancer Somatic GRCh38	9,589	9,589	Somatic	1.00	**398,402,166**
Total		**64,078**	**40,676**			**1,421,041,774**

OpenCGA Catalog stores all the metadata and clinical data of **files, samples**, **individuals** and **cohorts**. Rare Disease studies also include pedigree metadata by defining **families**. Also, a **Clinical Analysis** were defined for each family. Several Variable Sets have been defined to store GEL custom data for all these entities.

Platform

For the Research environment we have used **OpenCGA v1.4** using the new Hadoop Variant Storage that use **Apache HBase** as back-end because of the huge amount of data and analysis needed. We have also used **CellBase v4.6** for the variant annotation. Finally we set up a **IVA v1.0** web-based variant analysis tool.

The **Hadoop cluster** consists of about 30 nodes running **Hortonworks HDP 2.6.5** (which comes with **HBase 1.1.2**) and a LSF queue for loading all the VCF files, see this table for more detail:

Node	Nodes	Cores	Memory (GB)	Storage (TB)
Hadoop Master	5	28	216	7.2 (6x1.2)
Hadoop Worker	30	28	216	7.2 (6x1.2)

LSF Loading Queue	10	12	364	Isilon storage

Genomic Data Load

In order to improve the **loading performance,** we set up a small LSF queue of ten computing nodes. This configuration allowed us to load multiple files at the same time. We configured LSF to load up to 6 VCF files per node resulting in 60 files being loaded in HBase in parallel without any incidence, by doing this we observed a **50x in loading throughput**. This resulted in an average of 125 VCF files loaded per hour in studies RD37 and RD38, which is about **2 files per minute**. In the study CG38 the performance was 240 VCF files per hour or about **4 files per minute**.

Rare Disease Loading Performance

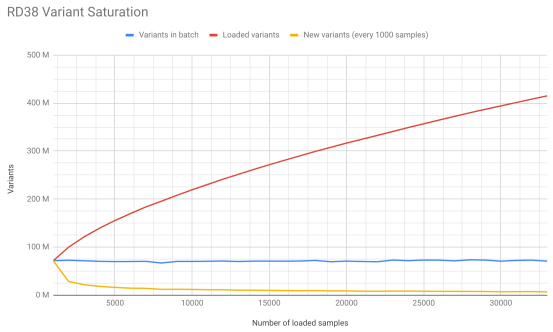
The files from Rare Disease studies (RD38 & RD37) contain 2 samples per file on average. This results in larger files, increasing the loading time compared with single-sample files. As mentioned above the loading performance was about 125 files per hour or 3,000 files per day. In terms of number of samples it is about **250 samples per hour or 6,000 samples a day**.

The loading performance always depend on the number of variants and concurrent files being loaded, the performance was quite stable during the load and performance degradation was observed as can be seen here:

Concurrent files loaded	60
Average files loaded per hour	125.72
Load time per file	00:28:38

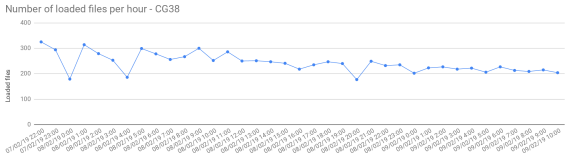
Saturation Study

As part of the data loading process we decided to study the number of unique variants added in each batch of 500 samples. We generated this saturation plot for RD38:



Cancer Loading Performance

The files from Cancer Germline studies (CG38) contain one sample per file. Compared with the Rare Disease, these files are smaller in size, therefore, as expected the file load was almost 2x faster. As mentioned above, the loading performance was about 240 genomes per hour or 5,800 files per day. In terms of number of samples it is about **5,800 samples a day**, which is consistent with Rare Disease performance.



Concurrent files loaded	60
Average files loaded per hour	242.05
Load time per file	00:14:52

Analysis Benchmark

In this section you can find information about the performance of main variant storage operations and most common queries and clinical analysis. Please, for data loading performance information go to section **Genomic Data Load** above.

Variant Storage Operations

Variant Storage operations take care of preparing the data for executing queries and analysis. There are two main operations: **Variant Annotation** and **Cohort Stats Calculation**.

Variant Annotation

This operation uses the **CellBase** to annotate each unique variant in the database, this annotation include consequence types, population frequencies, conservation scores clinical info, ... and will be typically used for variant queries and clinical analysis. Variant annotation of the 585 million unique variants of project **GRCh38 Germline** took about 3 days, **about 200 million variants were annotated per day**.

Cohort Stats Calculation

Cohort Stats are used for filtering variants in a similar way as the population frequencies. A set of cohorts were defined in each study.

- **ALL** with all samples in the study
- **PARENTS** with all parents in the study (only for Rare Disease studies)
- **UNAFF_PARENTS** with all unaffected parents in the study (only for Rare Disease studies)

Pre-computing stats for different cohort and ten of thousands of samples is a high-performance operation that run in **less than 2 hours** for each study.

Query and Aggregation Stats

To study the performance we used **RD38** which the largest study with 438 million variants and 33,000 samples. We first run some queries to the aggregated data filtering by variant annotation and cohort stats. We were interested in the different index performance so we limit the results to be returned the first 10 variants excluding the genotypic data of the 33,000 samples, by doing this we remove the effect of reading from disk or transferring data through the network which is very variable across different clusters. For queries using patient data go to the next section. Here you can find some of the common queries executed.

Filters	Results	Total Results	Time (sec)
consequence type = LoF + missense_variant	10	3704626	0.189
consequence type = LoF + missense_variant **biotype** = protein_coding	10	3576472	0.260
panel = with 200 genes	10	3882902	0.299
gene = BMPR2	10	37244	0.344

gene = BMPR2 **consequence type** = LoF	10	189	0.443
type = INDEL	10	79597426	0.802
type = INDEL **biotype** = protein_coding	10	38799454	0.358
type = INDEL **biotype** = protein_coding **consequence type** = LoF + missense_variant	10	240443	0.556
consequence type = LoF + missense_variant **population frequency** = 1000G ALL < 0.005	10	3157533	5.96

As can be observed most queries run below 1 second, you can combine as many filters as wanted.

Clinical Analysis

We also use here **RD38** which is the largest study. Clinical queries, or sample queries, enforces queries to return variants of a specific set of samples. These queries can use all the filters from the general queries. The result here also includes a **pathogenic prediction** for each variant, which determines possible conditions associated to the variant.

Filters	Results	Total Results	Time (sec)
mode of inheritance = recessive **filter** = PASS	10	211787	0.420
mode of inheritance = recessive **consequence type** = LoF + missense_variant	10	710	1.95
mode of inheritance = recessive **consequence type** = LoF + missense_variant **filter** = PASS	10	656	1.92
mode of inheritance = recessive **consequence type** = LoF + missense_variant **filter** = PASS **panel** = with 58 genes	10	7	2.11
***de novo* Analysis** **filter** = PASS **consequence type** = LoF + missense_variant	24	24	0.680
Compound Heterozygous Analysis **filter** = PASS **biotype** = protein_coding **consequence type** = LoF + missense_variant	417	417	10.930

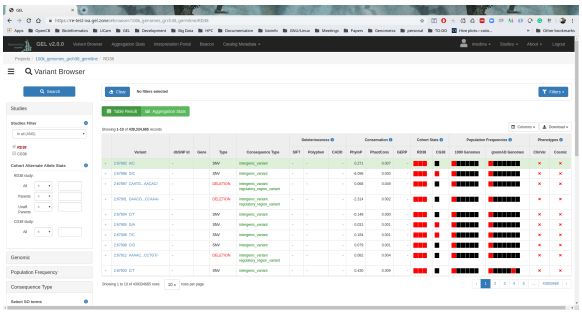
As it can be observed most of the family clinical analysis run in less than 2 seconds in the largest study with 33,000 samples.

User Interfaces

Several user interfaces have been developed to query and analyse data from OpenCGA: IVA web-based tool, Python and R clients, and a command line.

IVA

IVA v1.0.3 was installed to provide a friendly web-based analysis tool to browse variants and execute clinical analysis.



Command line

You can also query variants efficiently using the command line built in. Performance depends on the number of samples fetched and the RPC used *(REST* or *gRPC),* in the best scenario you can fetch few thousands variants per second. You can see a simple example here producing a VCF file:

```
imedina@ivory:~/appl/opencga/build[develop]$ ./bin/opencga.sh variant
query --gene BRCA2 --include-sample none --limit 10 --study RD37
##fileformat=VCFv4.2
##...
##...
##...
##...
##...
##contig=<ID=22,length=51304566>
##contig=<ID=MT,length=16569>
##contig=<ID=X,length=155270560>
##contig=<ID=Y,length=59373566>
##reference=/genomes/resources/genomeref/Illumina/Homo_sapiens/Ensembl
/GRCh37/Sequence/WholeGenomeFasta/genome.fa
#CHROM        POS        ID        REF        ALT        QUAL
FILTER        INFO
13        32884632        .        C        T        .        .
AC=1;AF=0.0000412;AN=1;
CT=T|ZAR1L|ENSG00000189167|ENST00000533490|protein_coding|intron_variant|||
||&T|ZAR1L|ENSG00000189167|ENST00000345108|protein_coding|intron_variant|||
||&T|BRCA2|ENSG00000139618|ENST00000380152|protein_coding|upstream_gene_var
iant|||||&T|BRCA2|ENSG00000139618|ENST00000544455|protein_coding|upstream_g
ene_variant|||||&T|||||regulatory_region_variant|||||&T|||||TF_binding_site
_variant|||||;PARENTS_AF=0;UNAFF_PARENTS_AF=0
13        32884636        .        T        A        .        .
AC=1;AF=0.0000412;AN=1;
CT=A|ZAR1L|ENSG00000189167|ENST00000533490|protein_coding|intron_variant|||
||&A|ZAR1L|ENSG00000189167|ENST00000345108|protein_coding|intron_variant|||
||&A|BRCA2|ENSG00000139618|ENST00000380152|protein_coding|upstream_gene_var
iant|||||&A|BRCA2|ENSG00000139618|ENST00000544455|protein_coding|upstream_g
ene_variant|||||&A|||||regulatory_region_variant|||||&A|||||TF_binding_site
_variant|||||;PARENTS_AF=0.0000843;UNAFF_PARENTS_AF=0.0000899
13        32884640        .        G        A        .        .
AC=2;AF=0.0000824;AN=2;
CT=A|ZAR1L|ENSG00000189167|ENST00000533490|protein_coding|intron_variant|||
||&A|ZAR1L|ENSG00000189167|ENST00000345108|protein_coding|intron_variant|||
||&A|BRCA2|ENSG00000139618|ENST00000380152|protein_coding|upstream_gene_var
iant|||||&A|BRCA2|ENSG00000139618|ENST00000544455|protein_coding|upstream_g
ene_variant|||||&A|||||regulatory_region_variant|||||&A|||||TF_binding_site
_variant|||||;PARENTS_AF=0.0000843;UNAFF_PARENTS_AF=0.0000899
13        32884647        rs142690293        T        C        .
.        AC=13;AF=0.0005355;AN=13;
CT=C|ZAR1L|ENSG00000189167|ENST00000533490|protein_coding|intron_variant|||
||&C|ZAR1L|ENSG00000189167|ENST00000345108|protein_coding|intron_variant|||
||&C|BRCA2|ENSG00000139618|ENST00000380152|protein_coding|upstream_gene_var
iant|||||&C|BRCA2|ENSG00000139618|ENST00000544455|protein_coding|upstream_g
```

```
ene_variant|||||&C|BRCA2|ENSG00000139618|ENST00000530893|protein_coding|ups
tream_gene_variant|||||&C|||||regulatory_region_variant|||||&C|||||TF_bindi
ng_site_variant|||||;PARENTS_AF=0.00059;POPFREQ=GNOMAD_GENOMES_ALL:
0.0029706|GNOMAD_GENOMES_EAS:0.0568603|GNOMAD_GENOMES_MALE:0.003972
|GNOMAD_GENOMES_FEMALE:0.0017329|1kG_phase3_ALL:0.0087859|1kG_phase3_CHS:
0.052381|1kG_phase3_JPT:0.0240385|1kG_phase3_CDX:0.0752688|1kG_phase3_KHV:
0.0353535|1kG_phase3_EAS:0.0436508|1kG_phase3_CHB:0.0339806|UK10K_ALL:
0.0001322|UK10K_ALSPAC:0.0002595;UNAFF_PARENTS_AF=0.0005392
13        32884653        rs206110        G        T        .
.        AC=14729;AF=0.6066809;AN=14729;
CT=T|ZAR1L|ENSG00000189167|ENST00000533490|protein_coding|intron_variant|||
||&T|ZAR1L|ENSG00000189167|ENST00000345108|protein_coding|intron_variant|||
||&T|BRCA2|ENSG00000139618|ENST00000380152|protein_coding|upstream_gene_var
iant|||||&T|BRCA2|ENSG00000139618|ENST00000544455|protein_coding|upstream_g
ene_variant|||||&T|BRCA2|ENSG00000139618|ENST00000530893|protein_coding|ups
tream_gene_variant|||||&T|||||regulatory_region_variant|||||&T|||||TF_bindi
ng_site_variant|||||;PARENTS_AF=0.6065408;POPFREQ=GNOMAD_GENOMES_ALL:
0.5944614|GNOMAD_GENOMES_OTH:0.5784114|GNOMAD_GENOMES_EAS:0.7806692
|GNOMAD_GENOMES_AMR:0.7494034|GNOMAD_GENOMES_ASJ:0.5298013
|GNOMAD_GENOMES_FIN:0.6099656|GNOMAD_GENOMES_NFE:0.5957703
|GNOMAD_GENOMES_AFR:0.5403784|GNOMAD_GENOMES_MALE:0.6001878
|GNOMAD_GENOMES_FEMALE:0.5873913|GONL_ALL:0.5621243|1kG_phase3_MXL:
0.7578125|1kG_phase3_ALL:0.6259984|1kG_phase3_SAS:0.6237219|1kG_phase3_CLM:
0.6382979|1kG_phase3_ITU:0.6029412|1kG_phase3_AFR:0.4962178|1kG_phase3_CHS:
0.7761905|1kG_phase3_JPT:0.7548077|1kG_phase3_YRI:0.4907407|1kG_phase3_PJL:
0.6302083|1kG_phase3_GWD:0.4336283|1kG_phase3_STU:0.6323529|1kG_phase3_GBR:
0.5934066|1kG_phase3_CDX:0.7473118|1kG_phase3_KHV:0.6818182|1kG_phase3_IBS:
0.682243|1kG_phase3_BEB:0.6046512|1kG_phase3_ACB:0.5416667|1kG_phase3_ESN:
0.530303|1kG_phase3_LWK:0.489899|1kG_phase3_EUR:0.6182902|1kG_phase3_ASW:
0.5409836|1kG_phase3_AMR:0.7262248|1kG_phase3_MSL:0.4705882|1kG_phase3_GIH:
0.6456311|1kG_phase3_FIN:0.5606061|1kG_phase3_TSI:0.6401869|1kG_phase3_PUR:
0.6586539|1kG_phase3_CEU:0.6060606|1kG_phase3_PEL:0.8823529|1kG_phase3_EAS:
0.7371032|1kG_phase3_CHB:0.723301|MGP_ALL:0.6086142|UK10K_ALL:0.6011637
|UK10K_TWINSUK_NODUP:0.606883|UK10K_ALSPAC:0.5949663|UK10K_TWINSUK:
0.6076052;UNAFF_PARENTS_AF=0.6061287
13        32884664        rs532184055        C        T        .
.        AC=35;AF=0.0014416;AN=35;
CT=T|ZAR1L|ENSG00000189167|ENST00000533490|protein_coding|intron_variant|||
||&T|ZAR1L|ENSG00000189167|ENST00000345108|protein_coding|intron_variant|||
||&T|BRCA2|ENSG00000139618|ENST00000380152|protein_coding|upstream_gene_var
iant|||||&T|BRCA2|ENSG00000139618|ENST00000544455|protein_coding|upstream_g
ene_variant|||||&T|BRCA2|ENSG00000139618|ENST00000530893|protein_coding|ups
tream_gene_variant|||||&T|||||regulatory_region_variant|||||&T|||||TF_bindi
ng_site_variant|||||;PARENTS_AF=0.0014329;POPFREQ=GNOMAD_GENOMES_ALL:
0.000678|GNOMAD_GENOMES_AMR:0.0023866|GNOMAD_GENOMES_ASJ:0.0066225
|GNOMAD_GENOMES_NFE:0.0010657|GNOMAD_GENOMES_AFR:0.0001146
|GNOMAD_GENOMES_MALE:0.000701|GNOMAD_GENOMES_FEMALE:0.0006496|GONL_ALL:
0.001002|1kG_phase3_ALL:0.0003994|1kG_phase3_AFR:0.0007564|1kG_phase3_GBR:
0.0054945|1kG_phase3_ACB:0.0052083|1kG_phase3_EUR:0.000994|MGP_ALL:0.011236
|UK10K_ALL:0.0018514|UK10K_TWINSUK_NODUP:0.001399|UK10K_ALSPAC:0.0023352
|UK10K_TWINSUK:0.0013484;UNAFF_PARENTS_AF=0.0014378
13        32884665        rs774189070        C        G        .
.        AC=14;AF=0.0005767;AN=14;
CT=G|ZAR1L|ENSG00000189167|ENST00000533490|protein_coding|intron_variant|||
||&G|ZAR1L|ENSG00000189167|ENST00000345108|protein_coding|intron_variant|||
||&G|BRCA2|ENSG00000139618|ENST00000380152|protein_coding|upstream_gene_var
iant|||||&G|BRCA2|ENSG00000139618|ENST00000544455|protein_coding|upstream_g
ene_variant|||||&G|BRCA2|ENSG00000139618|ENST00000530893|protein_coding|ups
tream_gene_variant|||||&G|||||regulatory_region_variant|||||&G|||||TF_bindi
ng_site_variant|||||;PARENTS_AF=0.0005057;POPFREQ=GNOMAD_GENOMES_ALL:
0.0000323|GNOMAD_GENOMES_AFR:0.0001147|GNOMAD_GENOMES_MALE:0.0000584
|UK10K_ALL:0.0002645|UK10K_TWINSUK_NODUP:0.0002798|UK10K_ALSPAC:0.0002595
|UK10K_TWINSUK:0.0002697;UNAFF_PARENTS_AF=0.0004493
13        32884667        rs539248433        A        C        .
.        AC=4;AF=0.0001648;AN=4;
CT=C|ZAR1L|ENSG00000189167|ENST00000533490|protein_coding|intron_variant|||
||&C|ZAR1L|ENSG00000189167|ENST00000345108|protein_coding|intron_variant|||
||&C|BRCA2|ENSG00000139618|ENST00000380152|protein_coding|upstream_gene_var
iant|||||&C|BRCA2|ENSG00000139618|ENST00000544455|protein_coding|upstream_g
ene_variant|||||&C|BRCA2|ENSG00000139618|ENST00000530893|protein_coding|ups
tream_gene_variant|||||&C|||||regulatory_region_variant|||||&C|||||TF_bindi
```

```
ng_site_variant|||||;PARENTS_AF=0.0000843;POPFREQ=GNOMAD_GENOMES_ALL:
0.0006782|GNOMAD_GENOMES_OTH:0.0010183|GNOMAD_GENOMES_NFE:0.0000666
|GNOMAD_GENOMES_AFR:0.0021774|GNOMAD_GENOMES_MALE:0.0007598
|GNOMAD_GENOMES_FEMALE:0.0005775|1kG_phase3_ALL:0.0007987|1kG_phase3_AFR:
0.0030257|1kG_phase3_GWD:0.0132743|1kG_phase3_MSL:0.0058824;
UNAFF_PARENTS_AF=0.0000899
13        32884672        .        NAGAC        N        .        .
AC=1;AF=0.0000412;AN=1;
CT=|ZAR1L|ENSG00000189167|ENST00000533490|protein_coding|intron_variant||||
|&|ZAR1L|ENSG00000189167|ENST00000345108|protein_coding|intron_variant|||||
&|BRCA2|ENSG00000139618|ENST00000380152|protein_coding|upstream_gene_varian
t|||||&|BRCA2|ENSG00000139618|ENST00000544455|protein_coding|upstream_gene_
variant|||||&|BRCA2|ENSG00000139618|ENST00000530893|protein_coding|upstream
_gene_variant|||||&|||||regulatory_region_variant|||||&|||||TF_binding_site
_variant|||||;PARENTS_AF=0.0000843;UNAFF_PARENTS_AF=0.0000899
13        32884685        .        G        A        .        .
AC=1;AF=0.0000412;AN=1;
CT=A|ZAR1L|ENSG00000189167|ENST00000533490|protein_coding|intron_variant|||
||&A|ZAR1L|ENSG00000189167|ENST00000345108|protein_coding|intron_variant|||
||&A|BRCA2|ENSG00000139618|ENST00000380152|protein_coding|upstream_gene_var
iant|||||&A|BRCA2|ENSG00000139618|ENST00000544455|protein_coding|upstream_g
ene_variant|||||&A|BRCA2|ENSG00000139618|ENST00000530893|protein_coding|ups
tream_gene_variant|||||&A|||||regulatory_region_variant|||||&A|||||TF_bindi
ng_site_variant|||||;PARENTS_AF=0;UNAFF_PARENTS_AF=0
```

Support

OpenCB team is setting up **Zetta Genomics**, a start-up to offer support, consultancy and custom feature development. We have partnered with Microsoft Azure to ensure OpenCB Suite runs efficiently in **Microsoft Azure** cloud. We are running a proof-of-concept at the moment with GEL data to benchmark and test Azure.

Acknowledgements

We would like to thank Genomics England very much for their support and for trusting in OpenCGA and the rest of OpenCB Suite for this amazing release. In particular, we would like to thank Augusto Rendon, Anna Need, Carolyn Tregidgo, Frank Nankivell and Chris Odhams for their support, test and valuable feedback.