Aggregated gVCF dataset v1_20190228

Brief description

This is a set of multi-sample gVCF files containing germline genomic data from 59464 participants from Release 5.1. The file contains samples from both the rare disease and the cancer programs, but only genomes on build GRCh38 were included. All included samples have passed a set of basic QC metrics

- cross-contamination <5%
- mapping rate >75%
- mean sample coverage >20
- insert size <250).

These QC metrics are given in the LabKey table `aggregate_gvcf_sample_stats`.

The aggregated dataset is split into 1009 pieces for easier handling, due to its large size. No variant QC filters were applied in the dataset, but the VCF filter was set to PASS for variants which passed GQ, DP, missingess, allelic imbalance, and Mendel error filters. We recommend only using variants that have PASS in the filter column in your analyses. The data set alongside with a more detailed description is stored here: `/gel_data_resources/main_programme/aggregated_illumina_gvcf/GRCH38/20190228/`

Pre-aggregation QC

Sample QC

Basic QC metrics are provided in the LabKey table `aggregate_gvcf_sample_stats`. Overview of initial QC is given below. This QC was conducted prior to the data aggregation. Individual sample QC data was retrieved from Genomics England openCGA data base. Most metrics are bam file statisitics provided from Illumina or the Genomics England WGS data processing pipeline BERTHA.

	cut off value	failed samples	samples
Insert size medium, bp	250	3	60124
Chimeric DNA fragments, %	2%	0	60124
Cross-contamination	5%	261	59863
genetic vs reported sex	fail	140	59723
Genome-wide coverage mean	20	1	59722
Mapped Reads	75%	32	59690

sample swaps (tumour)		17	59673
samples failed in 30k merge		204	59469
samples failed 60k merge		5	59464

Genetic vs reported sex checking is done as part of the Genomics England Bioinformatics interpretation pipeline, and has been completed in approximately 50% of the aggregated data set so far.

gVCF aggregation process

The genome was broken down into 355 chunks for easier handling due to big file sizes. To cut the genome into smaller regions we used a file made available by Broad, where region boundaries are defined by a high fraction of missing basecalls (N). This way there is a low risk of losing indels that could span two regions in the aggregation process. We use Illumina's agg software to produce multi-sample VCFs. (https://github.com/Illumina/agg). Agg handles multiallelic nucleotide polymorphisms (MNPs) in the following manner:

"We decompose MNPs (Multiallelic Nucleotide polymorphisms) and perform basic left-shifting/trimming of indels (taken from `bcftools norm` implementation). We apply the complex substitution decomposition routine implemented in [vt] (http://genome.sph.umich.edu/wiki/Vt), but as noted on their webpage, there is no unique solution for this problem. Multiallelic Variants are written as one allele per line in the VCF file."

Aggregated dataset

The dataset contains 677,003,512 variants (SNPs and short INDELs) for 59464 participants. We calculate the following QC metrics for each variant and give them as an INFO field for each site:

1. INFO/missing: Missingness per site (using VCFtools version 0.1.15)
2. INFO/meanDP: Mean coverage per site (using VCFtools version 0.1.15)
3. INFO/meanGQ: Mean GQ per site (using BCFtools version 1.9)

4. INFO/ AB_ratio: Allelic imbalance: Fraction of heterozygotes alleles that passed a binomial test for allelic depth; ABratio = (the number of sites that passed AB test / the number of heterozygote alleles). The binomial test was calculated in BCFtools version 1.9; Binomial test P-value threshold = 0.01
5. INFO/inbreed: Inbreeding coefficient: The F-statistic of expected heterozygosity was calculated with the following formula: F=1-(observed frequency(heterozygote)) / expected frequency(heterozygote)). The expected and observed heterozygote frequencies were retrieved with plink1.9.
6. INFO/MErrRatio: The fraction of genotypes with Mendelian errors. The ratio is calculated as the Mendel Errors/allele count in trio probands. The number of alleles with Mendel Errors was calculated with plink (PLINK version 1.9; function --mendel summaries-only). Validated trios are given in the labkey table aggregate_gvcf_sample_stats. Monomorphic alleles in trio probands were set to missing (assigned a dot), whereas 0 indicates 0 Mendel Errors. We counted the minor alleles for the probands (using VCFtools version 0.1.15).
7. INFO/MErr: Genotypes with Mendelian errors. This is the same as point 6, but reports the total number of Mendelian errors per site instead of the fraction of Mendelian errors relative to the MAC in trio probands.

Additionally, we set the VCF FILTER field to PASS if [all of the following are true?]

- Missingness < 5%
- Coverage >= 10
- GQ >=15
- Fraction of variants passed allelic imbalance test >= 0.25

Example filter fields for sites that fail one of the checks are detailed below:

- FILTER = callrate if a site has missingness > 5%
- FILTER = coverage if coverage is < 10
- FILTER = GQ if GQ < 15
- FILTER = ABratio if < 25% of heterozygote genotypes at a specific site fail the allelic imbalance test
- FILTER = callrate:coverage:GQ:ABratio if a site fails all checks

We recommend using only variants that have a PASS filter.

Final dataset

Files are stored here:
`/gel_data_resources/main_programme/aggregated_illumina_gvcf/GRCH38/20190228/data/`

Large chunks were broken down further after aggregation, to even out waiting times when working the data set. This resulted in 1009 chunks.
Files are named in the following format:
60k_GRCH38_germline_mergedgVCF_chr[start-pos]_[end-pos].bcf
where CHR, STARTPOS, ENDPOS are the chromosome, starting and ending positions of the variants contained in the respective chunk.
Specific file names and regions in bed format can be found in this tab delimited file;
`/gel_data_resources/main_programme/aggregated_illumina_gvcf/GRCH38/20190228/docs/gvcf_aggregation_regions_and_names`

Example usage

To view variants that PASS or pass Missingness and DP but fail GQ and ABratio for a specific region:

```
bcftools view -H -r chr17:43044295-43125483 -f
PASS,ABratio,GQ,GQ:ABratio
60k_GRCH38_germline_mergedgVCF_chr17_42478369_46311094.bcf |
less -S
```

This will include all sites where the Filter field is PASS, ABratio,GQ or GQ:ABratio.

Extract some information from the files:

```
bcftools view -r chr17:43044295-43125483 -f
PASS,ABratio,GQ,GQ:ABratio
60k_GRCH38_germline_mergedgVCF_chr17_42478369_46311094.bcf |
bcftools query -f '%CHROM %POS %REF %ALT %FILTER\n'  >
BRCA1_variants
```

Filter levels are: `PASS, callrate, coverage, GQ, ABratio, callrate:coverage, callrate:GQ,coverage:GQ,callrate:ABratio, coverage:ABratio,GQ:ABratio, callrate:coverage:ABratio, callrate:coverage:GQ, callrate:GQ:ABratio, coverage:GQ:ABratio, callrate:coverage:GQ:ABratio`