







DATA NOTE

# Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project [version 1; peer review: 2 not approved]

Ernesto Lowy-Gallego <sup>1</sup>, Susan Fairley <sup>1</sup>, Xiangqun Zheng-Bradley<sup>1</sup>,  
Magali Ruffier<sup>1</sup>, Laura Clarke <sup>1</sup>, Paul Flicek <sup>1</sup>,  
the 1000 Genomes Project Consortium

<sup>1</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

**V1** **First published:** 11 Mar 2019, 4:50 (<https://doi.org/10.12688/wellcomeopenres.15126.1>)  
**Latest published:** 11 Mar 2019, 4:50 (<https://doi.org/10.12688/wellcomeopenres.15126.1>)

## Abstract


We present biallelic SNVs called from 2,548 samples across 26 populations from the 1000 Genomes Project, called directly on GRCh38. We believe this will be a useful reference resource for those using GRCh38, representing an improvement over the “lift-overs” of the 1000 Genomes Project data that have been available to date and providing a resource necessary for the full adoption of GRCh38 by the community. Here, we describe how the call set was created and provide benchmarking data describing how our call set compares to that produced by the final phase of the 1000 Genomes Project on GRCh37.



## Keywords

Genomics, population genetics, variant calling, single nucleotide variation, variant discovery

## Open Peer Review

Reviewer Status  

	Invited Reviewers	
	1	2
<b>version 1</b> published 11 Mar 2019	 report	 report

- 1 **Deanna M. Church** , Inscripta, Inc., Boulder, USA
- 2 **Augusto Rendon** , Genomics England, London, UK  
University of Cambridge, Cambridge, UK

Any reports and responses or comments on the article can be found at the end of the article.

**Corresponding author:** Paul Flicek ([flicek@ebi.ac.uk](mailto:flicek@ebi.ac.uk))

**Author roles:** **Lowy-Gallego E:** Data Curation, Formal Analysis, Methodology, Software, Validation, Writing – Original Draft Preparation; **Fairley S:** Methodology, Project Administration, Supervision, Writing – Review & Editing; **Zheng-Bradley X:** Data Curation, Software; **Ruffier M:** Supervision; **Clarke L:** Conceptualization, Funding Acquisition, Project Administration, Supervision; **Flicek P:** Conceptualization, Funding Acquisition, Supervision;

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was completed thanks to the funding from the Wellcome Trust (grant number 104947) and the European Molecular Biology Laboratory.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2019 Lowy-Gallego E *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Lowy-Gallego E, Fairley S, Zheng-Bradley X *et al.* **Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project [version 1; peer review: 2 not approved]** Wellcome Open Research 2019, 4:50 (<https://doi.org/10.12688/wellcomeopenres.15126.1>)

**First published:** 11 Mar 2019, 4:50 (<https://doi.org/10.12688/wellcomeopenres.15126.1>)

## Introduction

The 1000 Genomes Project started in 2008 with the aim of producing a deep catalogue of human genomic variation and, for this, more than 2600 samples from 26 different populations were sequenced. The project completed its final phase (named phase three) in 2015, with the release of more than 85 million variants of various types and phased haplotypes for those variants<sup>1</sup>. This data has been widely used by the scientific community for genotype imputation and many other applications<sup>2</sup>. The strategy adopted by the project consisted of sequencing samples using whole genome sequencing (WGS) and whole exome sequencing (WES), and the alignment of that sequence data to a version of the GRCh37 human reference genome, which included decoy sequences for optimal read mapping.

While the 1000 Genomes Project was based on GRCh37, the latest version of the human reference assembly (GRCh38) was released by the Genome Reference Consortium (GRC) in 2013 and is the current best representation of the human genome available [PMID: 28396521]. This updated assembly has the following fundamental improvements with respect to its predecessors:

- corrects thousands of small sequencing artifacts and misassembled regions in addition to filling or reducing more than 100 gaps
- includes synthetic centromeric sequences that previously were represented in the reference by gaps of three million base pairs
- improves the diversity of the reference by including new alternate sequences, to address the fact that some genomic regions are highly variable

By improving the reference genome, GRCh38 improves the foundation for calling variation by providing both a more accurate and more diverse representation of the genome, thereby enabling better read mapping and reducing opportunities for erroneous variation calls.

To make full use of GRCh38, there has been a need for widely used genomic reference data sets, like the 1000 Genomes data, to be made available on the assembly, so that pipelines and analyses that rely on such additional reference materials can use GRCh38 and benefit from its improvements.

[dbSNP](#) have facilitated the use of the 1000 Genomes variation data on GRCh38 by “lifting-over” the calls, using a method relying on an alignment created between GRCh37 and GRCh38. The alignment is then used to determine equivalent locations between the two assemblies, allowing variation data to be “lifted-over”. Files from dbSNP are reformatted into a standard VCF by the [European Variation Archive](#) (EVA) and shared as part of our resources through the [1000 Genomes FTP site](#) and also via the Ensembl genome browser [PMID: 30407521].

Lift-over approaches, however, have several limitations. 1) Necessarily, they rely on an equivalent region existing in the new genome, so new sequence in the improved assembly is effectively excluded. 2) Reliable transfer requires a good mapping between the assemblies, covering not just a given variation but the context that was used to make that call—it may be possible to “lift-over” a SNP where the data supporting the original call would not lift-over. Where the context of a call alters, the data becomes less reliable. 3) While lift-overs can give an approximation of the variant sites on the new assembly the results will differ from calling directly on the new assembly, the latter taking advantage of the increased representation of genomic sequence, assembly corrections and making calls from the underlying read data in context. With the above in mind, and given that the 1000 Genomes data is a heavily used resource, we decided to create a new call set from alignments of the original 1000 Genomes read data to GRCh38.

The first step was alignment of the 1000 Genomes sequence data to the GRCh38 as previously described<sup>3</sup>. These alignments were taken as the starting point in creating the variation calls described in this data note.

To generate calls from the 1000 Genomes data, we adopted a multi-caller approach, aiming to produce a similar quality reference call set to that produced by the 1000 Genomes Project while using a simpler methodology, reflecting both practical considerations and the improved understanding of the process developed by the 1000 Genomes Project itself.

Four supporting call sets were created, using different callers and combinations of the exome and WGS sequence data.

For the final call set, biallelic SNVs (single nucleotide variants) only were selected from the four supporting call sets. These represent the major part of the SNVs present in the human genome.

The inclusion of only biallelic SNVs generates a data set useful for many purposes, while enabling more streamlined data processing than is possible when handling indels and multi-allelic variants. We are thus able to share what we believe to be a useful data set while planning to revisit our supporting call sets and, in future, produce updated call sets including a broader spectrum of variation.

## Methods

### Input data

The methods used for sample collection, library construction, and sequencing are described in the previous 1000 Genomes Project publications<sup>1,4,5</sup>. The read data used for this analysis followed the same criteria as the final phase of the 1000 Genomes Project, namely only sequence data generated by Illumina sequencing and only reads longer than 70 bp (WGS) and 68 bp (WES). This data was aligned to GRCh38 as previously described<sup>3</sup>. The complete list of the whole genome and whole exome sequencing alignment files used as the input for generating the callsets can be found on our FTP site at [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\\_collections/1000\\_genomes\\_project/1000genomes.low\\_coverage.GRCh38DH.alignment.index](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/1000genomes.low_coverage.GRCh38DH.alignment.index) and at [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\\_collections/1000\\_genomes\\_project/1000genomes.exome.GRCh38DH.alignment.index](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/1000genomes.exome.GRCh38DH.alignment.index).

### Reference genome

We used the full GRCh38 reference, including ALT contigs, decoy and EBV sequences (accession [GCA\\_000001405](https://www.ncbi.nlm.nih.gov/nuccore/GCA_000001405)). In addition, more than 500 HLA sequences compiled by Heng Li from the IMGT/HLA database provided by the Immuno Polymorphism Database (IPD) [PMID: 27899604] are included. The reference genome can be accessed at [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38\\_reference\\_genome/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/).

### Ethical considerations

Information concerning ethical approval and the informed consent procedure for the 1000 Genomes project can be found [here](#).

### Quality control of the alignment files

We adopted a similar quality control process to that used in the final phase of the 1000 Genomes Project. [Chk\\_indel\\_rg](#) was applied to discard alignment files with an unbalanced ratio of short insertions and deletions (greater than 5). [Picard CollectWgsMetrics](#) was used with the whole genome files and those with mean non-duplicated aligned coverage level  $\leq 2x$  were discarded. In the case of the exome files, we used [Picard CollectHsMetrics](#) using the exome target coordinates that can be found at [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\\_collections/1000\\_genomes\\_project/working/20190125\\_coords\\_exon\\_target/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/working/20190125_coords_exon_target/), and we kept the files having more than 70% of the target regions covered by 20x or greater of sequence reads.

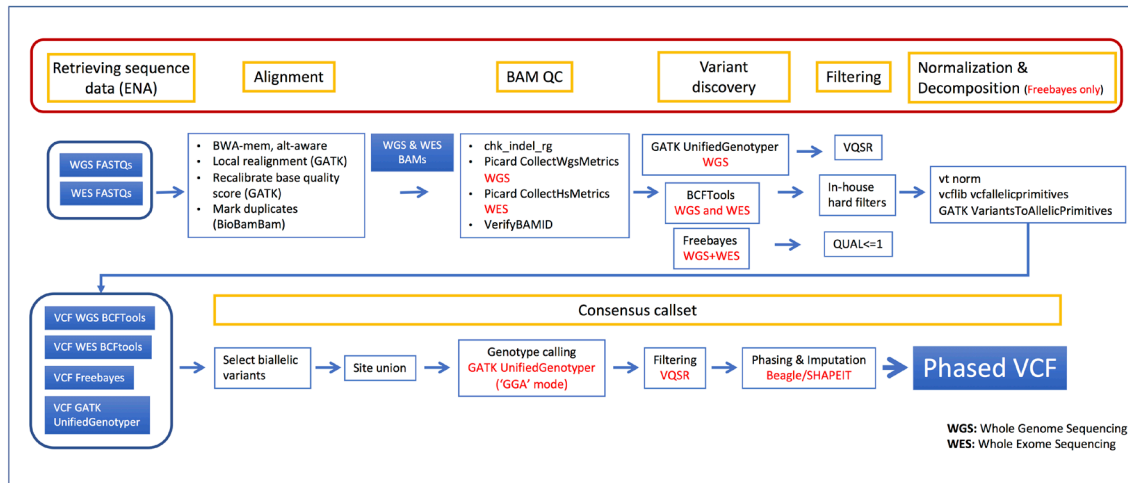
In addition, [VerifyBAMID](#)<sup>6</sup> was used to assess sample contamination and sample mix-ups and the following cutoffs were used:

- `free_mix > 0.03` and `chip_mix > 0.02` for whole genome files
- `free_mix > 0.035` and `chip_mix > 0.02` for exome files

Only files passing the quality assessment were used in subsequent variant calling.

### Variant discovery

A total of 2,659 WGS and 2,498 WES BAMs were generated corresponding to 2,698 samples<sup>3</sup> were used for variant identification. [Figure 1](#) details the analysis of the alignment files with three established methods ([BCFtools](#) version 1.3.1-220-g9f38991, [Freebayes](#)<sup>7</sup> version v1.0.2-58-g054b257 and [GATK UnifiedGenotyper](#)<sup>8</sup> version 3.5-0-g36282e4). BCFtools was used to analyse WGS and WES files in two independent runs, GATK UnifiedGenotyper



**Figure 1.** Schematic representation of our approach illustrating the entire process from the alignment files previously generated to the generation of the four supporting callsets and finally to the production of the final phased consensus callset. VCF, variant call format; WGS, whole-genome sequencing; WES, whole-exome sequencing; VQSR, variant quality score recalibration.

was used only with WGS files and Freebayes was used to analyse everything together (WGS+WES). The following command lines were used for each of the methods:

- BCFtools with the WGS files:

```
bcftools mpileup -E -a DP -a SP -a AD -P ILLUMINA \
  -pm3 -F0.2 -C50 -d 700000 \
  -f $ref.fa $file.bam | bcftools call -mv -O z \
  --ploidy GRCh38 -S $samples.ped -o $out.vcf.gz
```

- GATK UnifiedGenotyper with the WGS files:

```
java -Xmx6g -jar GenomeAnalysisTK.jar \
  -T UnifiedGenotyper \
  -R $ref.fa \
  -I $file.bam \
  -o $out.vcf.gz \
  -dcov 250 \
  -stand_emit_conf 10 \
  -glm both \
  --genotyping_mode GENOTYPE_GIVEN_ALLELES \
  --dbsnp ALL_20141222.dbSNP142_human_GRCh38.snps.vcf.gz \
  -stand_call_conf 10
```

- BCFtools with the WES files:

```
bcftools mpileup -E -a DP -a SP -a AD -P ILLUMINA \
  -pm3 -F0.2 -C50 -d 1400000 \
  -f $ref.fa $file.bam | bcftools call -mv -O z \
  --ploidy GRCh38 -S $samples.ped -o $out.vcf.gz
```

- Freebayes with the WGS+WES files:

```
freebayes --genotyping-max-iterations 10 \
  --min-alternate-count 3 \
  --max-coverage 2000000 \
  --min-mapping-quality 1 \
  --min-alternate-qsum 50 \
  --min-base-quality 3 \
  -f $ref.fa \
  -b $file.bam | bgzip -c > $out.vcf.gz
```

## Variant filtering

Our variant discovery pipeline produced four initial call sets as described above. To create the final call set, we discarded the variants falling in the centromeres, as these are regions of low complexity that hinder variant calling. Variants on the chromosome Y or in regions of the chromosome X not corresponding to the pseudoautosomal regions (PAR) were also discarded due to the ploidy settings used in this work. Additionally, the initial call sets contained spurious variants filtered using different methods and parameters depending on the call set:

**GATK UnifiedGenotyper call set.** We used the VariantScoreRecalibration (VQSR)<sup>8</sup> method following the GATK best practices and GATK training call sets. The combination of commands and parameters we used were different depending on the the variant type being analysed. For SNPs we used GATK VariantRecalibrator and ApplyRecalibration as follows:

```
java -jar GenomeAnalysisTK.jar \
  -T VariantRecalibrator \
  -R $ref.fa \
  -input $file.vcf.gz \ -resource:hapmap,known=false,training=true,truth=true
  ,prior=15.0 hapmap_3.3.hg38.vcf.gz \ -resource:omni,known=false,training=tru
  e,truth=true,prior=12.0 1000G_omni2.5.hg38.vcf.gz \ -resource:1000G,known=fa
  lse,training=true,truth=false,prior=10.0 1000G_phase1.snps.high_confidence.
  hg38.vcf.gz \ -resource:dbsnp,known=true,training=false,truth=false,prior=2.
  0 dbsnp_146.hg38.vcf.gz \
  -an DP \
  -an QD \
  -an FS \
  -an SOR \
  -an MQ \
  -an MQRankSum \
  -an ReadPosRankSum \
  -an InbreedingCoeff \
  -mode SNP \
  -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 98.0 -tranche 97.0 -
  tranche 96.0 -tranche 95.0 -tranche 92.0 -tranche 90.0 -tranche 85.0 -
  tranche 80.0 -tranche 75.0 -tranche 70.0 -tranche 65.0 -tranche 60.0 -
  tranche 55.0 -tranche 50.0 \
  -recalFile recalibrate_SNP.recal \
  -tranchesFile recalibrate_SNP.tranches \
  -rscriptFile recalibrate_SNP_plots.R
```

And:

```
java -jar GenomeAnalysisTK.jar
  -T ApplyRecalibration \
  -R $ref.fa \
  -input $file.vcf.gz \
  -mode SNP \
  --ts_filter_level 99.9 \
  -recalFile recalibrate_SNP.recal \
  -tranchesFile recalibrate_SNP.tranches | bgzip -c > recalibrated_snps_raw_
  indels.vcf.gz
```

And for INDELs we used:

```
java -jar GenomeAnalysisTK.jar \
  -T VariantRecalibrator \
  -R $ref.fa \
  -input recalibrated_snps_raw_indels.vcf.gz \ -resource:mills,known=false,tra
  ining=true,truth=true,prior=12.0 Mills_and_1000G_gold_standard.indels.hg38.
  vcf.gz \ -resource:dbsnp,known=true,training=false,truth=false,prior=2.0
  dbsnp_146.hg38.vcf.gz \
  -an QD \
  -an DP \
  -an FS \
```

```

-an SOR \
-an ReadPosRankSum \
-an MQRankSum \
-an InbreedingCoeff \
-mode INDEL \
-tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 98.0 -tranche 97.0 -
tranche 96.0 -tranche 95.0 -tranche 92.0 -tranche 90.0 -tranche 85.0 -
tranche 80.0 -tranche 75.0 -tranche 70.0 -tranche 65.0 -tranche 60.0 -
tranche 55.0 -tranche 50.0 \
-recalFile recalibrate_INDEL.recal \
-tranchesFile recalibrate_INDEL.tranches \
-rscriptFile recalibrate_INDEL_plots.R \
--maxGaussians 4

```

And:

```

java -jar GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-R $ref.fa \
-input recalibrated_snps_raw_indels.vcf \
-mode INDEL \
--ts_filter_level 80.0 \
-recalFile recalibrate_INDEL.recal \
-tranchesFile recalibrate_INDEL.tranches | bgzip -c > recalibrated_variants.
vcf.gz

```

**BCFTools call sets.** We compared the distribution of the values for different variant annotations in the set of true positive and false positive sites and established the set of variant annotations and cutoff values used in the filtering. We considered true positives the sites identified in our call set for genome NA12878 that were also present in the gold-standard call set generated for the same sample by [Genome in a Bottle](#) (GIAB). GIAB's calls for NA12878 are the result of an effort to integrate data generated by 13 different sequencing technologies and analysis methods<sup>9</sup>. Sites that were present in our call sets and absent in GIAB were considered false positive sites. [Table 1](#) and [Table 2](#) show the variant annotations and cutoff values used for the SNPs and INDELS with the low coverage data and [Table 3](#) and [Table 4](#) show the annotations and cutoff values used for the exome data with the SNPs and INDELS respectively. These cutoff values were applied using the following command:

- SNPs from the low coverage data:

```

bcftools filter -s GIABFILTER \
-e'INFO/DP>24304 | MQ<34 | MQ0F>0.049737 | HOB>0.1643732 | SGB>2347.043 |
SGB<-64440.286 | QUAL<20' \
$file.snps.vcf.gz \
-o $out.snps.filtered.vcf.gz -O z

```
- INDELS from the low coverage data:

```

bcftools filter -s GIABFILTER \
-e'INFO/DP>23758 | MQ<41 | MQ0F>0.009913696 | HOB>0.20265508 | SGB>2143.8876
| SGB<-29513.557 | IDV>51 | IMF<0.387097 | QUAL<20' $file.indels.vcf.gz -o
$out.indels.filtered.vcf.gz -O z

```
- SNPs from the exome data:

```

bcftools filter -sGIABFILTER \
-e'INFO/DP>656519 | MQ<38 | MQ0F> 0.0146629| HOB>0.1536016 | SGB>57489.21
| SGB < -226326.93| QUAL<20' $file.snps.vcf.gz \
-o $out.snps.filtered.vcf.gz -O z

```
- INDELS from the exome data:

```

bcftools filter -sGIABFILTER \
-e'MQ<45 | MQ0F>0.002034686| HOB> 0.269603| SGB>53165.5 | SGB<-85919.729 |
IMF<0.3323922 | QUAL<20' $file.indels.vcf.gz \
-o $out.indels.filtered.vcf.gz -O z

```

**Table 1. Variant annotations and cutoff values used for SNPs identified using the low coverage data.**

Annotation	Description	Cutoff value
INFO/DP	Raw read depth	>24,304
INFO/MQ	Average mapping quality	<34
INFO/MQ0F	Fraction of MQ0 reads (smaller is better)	>0.049737
INFO/HOB	Bias in the number of HOMs number (smaller is better)	>0.1643732
INFO/SGB	Segregation based metric	>2347.043
INFO/SGB	Segregation based metric	<-64440.286
QUAL	Variant quality	<20

**Table 2. Variant annotations and cutoff values used for INDELS identified using the low coverage data.**

Annotation	Description	Cutoff value
INFO/DP	Raw read depth	>23,758
INFO/MQ	Average mapping quality	<41
INFO/MQ0F	Fraction of MQ0 reads (smaller is better)	>0.009913696
INFO/HOB	Bias in the number of HOMs number (smaller is better)	>0.20265508
INFO/SGB	Segregation based metric	>2143.8876
INFO/SGB	Segregation based metric	<-29513.557
INFO/IDV	Maximum number of reads supporting an indel	>51
INFO/IMF	Maximum fraction of reads supporting an indel	<0.387097
QUAL	Variant quality	<20

**Table 3. Variant annotations and cutoff values used for SNPs identified using the exome data.**

Annotation	Description	Cutoff value
INFO/DP	Raw read depth	>656,519
INFO/MQ	Average mapping quality	<38
INFO/MQ0F	Fraction of MQ0 reads (smaller is better)	>0.0146629
INFO/HOB	Bias in the number of HOMs number (smaller is better)	>0.1536016
INFO/SGB	Segregation based metric	>57489.21
INFO/SGB	Segregation based metric	<-226326.93
QUAL	Variant quality	<20

**Table 4. Variant annotations and cutoff values used for INDELS identified using the exome data.**

Annotation	Description	Cutoff value
INFO/MQ	Average mapping quality	<45
INFO/MQ0F	Fraction of MQ0 reads (smaller is better)	>0.002034686
INFO/HOB	Bias in the number of HOMs number (smaller is better)	>0.269603
INFO/SGB	Segregation based metric	>53165.5
INFO/SGB	Segregation based metric	<-85919.729
INFO/IMF	Maximum fraction of reads supporting an indel	<0.3323922
QUAL	Variant quality	<20



**Freebayes call set.** We used a simple hard filter that discarded variants having a QUAL value less than or equal to 1. This filter was applied using the following command:

```
bcftools filter -sQUALFILTER -e'QUAL<1' $file.vcf.gz \
-o $file.filtered.vcf.gz -O z
```

#### Generation of the consensus call set

First, each call set was normalized using a combination of `vt normalize`<sup>10</sup> (version 0.5) and `vcfib vcfallelicprimitives` (version v1.0.0-rc1). This procedure was necessary because sometimes the different variant callers describe the same variant in a different way, which makes comparison difficult and affects the integration of the different initial call sets. Additionally, GATK `VariantsToAllelicPrimitives` was used to decompose the multi-nucleotide polymorphisms (MNPs) that were present in the Freebayes call set.

Finally, and in order to take advantage of the strengths of each method used for the variant identification, we generated a consensus call set by the union of the biallelic sites from each call set and by the calculation of the genotype likelihoods for each site using GATK `UnifiedGenotyper` in 'genotype\_given\_alleles' (GGA) mode using the following command line:

```
java -jar GenomeAnalysisTK.jar \
-T UnifiedGenotyper \
-R $ref.fa \
-I input.$chr:$start-$end.bam \
-glm SNP \
--intervals $chr:$start-$end \
--intervals integrated.biallelic.sites.vcf.gz \
--output_mode EMIT_ALL_SITES \
--alleles integrated.biallelic.sites.vcf.gz \
--interval_set_rule INTERSECTION \
--genotyping_mode GENOTYPE_GIVEN_ALLELES \
--max_deletion_fraction 1.5
```

Where `$chr:$start-$end` is the genomic chunk that is being analysed and `integrated.biallelic.sites.vcf.gz` is the VCF containing the union of the biallelic sites for which the genotype likelihoods will be calculated.

We then filtered the spurious variants resulting of the union of the sites using Variant Quality Score Recalibration (VQSR) and the same parameters and training call sets that were described above used for filtering the supporting call set generated using GATK `UnifiedGenotyper`. GATK `ApplyRecalibrator` was used with the same `--ts_filter_level` value of 99.9 used for SNPs.

**Phasing and imputation of the consensus call set.** The VCF file containing the genotype likelihoods obtained following the procedure described in previous section was divided into single chromosome VCF files that were further divided into genomic chunks containing 2,100 sites of which 600 were shared between consecutive chunks. These chunks were processed by in parallel by `Beagle`<sup>11</sup> by using the following command:

```
java -jar beagle.08Jun17.d8b.jar \
chrom=$chr:$start-$end \
gl=$chr.biallelic.GL.vcf.gz \
out=$chr.$start.$end.beagle \
niterations=15
```

Where `$chr.biallelic.GL.vcf.gz` is the VCF file containing the genotype likelihoods.

After processing all the chunks with `Beagle` we obtained an initial set of genotypes and haplotypes used in the next step consisting of phasing the genotype likelihoods onto a highly accurate haplotype scaffold obtained using `SHAPEIT2`<sup>12</sup> (version v2.r837) with microarray genotype data available on the same samples. This scaffold was obtained by leveraging family information and running `SHAPEIT2` in two different independent runs on either the Illumina Omni 2.5 or Affymetrix 6.0 microarray data that was generated as part of the 1000 Genomes Project. `SHAPEIT2` was run using the following settings (`--window 0.5`, `--states 200`, `--burn 10`, `--prune 10`, `--main 50`, `--duohmm`) and SNPs with a missing data rate above 10% and a Mendel error rate above 5% were removed before phasing.

In order to phase the genotype likelihoods obtained from Beagle onto the haplotype scaffold we used SHAPEIT2. Genotypes called by Beagle with a posterior probability greater than 0.995 were fixed as known genotypes and the haplotypes estimated by Beagle were used to initialize the SHAPEIT2 phasing. This phasing was run in chunks of 12,250 sites with 3,500 sites overlapping between consecutive chunks. SHAPEIT2 was run using the following command:

```
shapeit -call \
  --input-gen input.shapeit.$chr.gen.gz input.shapeit.$chr.gen.sample \
  --input-init input.shapeit.$chr.hap.gz input.shapeit.hap.sample \
  --input-scaffold chip.omni.snps.$chr.haps chip.omni.snps.$chr.sample chip.
  affy.snps.$chr.haps chip.affy.snps.$chr.sample \
  --input-map $chr.gmap.gz \
  --input-thr 1 \
  --window 0.1 \
  --states 400 \
  --states-random 200 \
  --burn 0 \
  --run 12 \
  --prune 4 \
  --main 20 \
  --input-from $chunk_start \
  --input-to $chunk_end \
  --output-max out.$chr.$chunk_start.$chunk_end.haps.gz out.$chr.$chunk_
  start.$chunk_end.haps.sample
```

Where `--input-gen` specifies the genotype/GL input data from Beagle, `--input-init` specifies the haplotypes from Beagle, `--input-map` specifies the genetic map used in the estimation, `--input-scaffold` gives the SNP-array derived haplotype scaffold obtained from SHAPEIT2. The genetic map used was downloaded from [https://data.broadinstitute.org/alkesgroup/Eagle/downloads/tables/genetic\\_map\\_hg38\\_withX.txt.gz](https://data.broadinstitute.org/alkesgroup/Eagle/downloads/tables/genetic_map_hg38_withX.txt.gz). Each of the phased chunks resulting from running SHAPEIT2 were joined together using the program `ligateHAPLOTYPES`.

The strategy described here was used in the final phase of the 1000 Genomes Project and has been shown to produce low error rates for genotype calls<sup>13</sup>.

The pipelines used in this work were implemented using the eHive workflow system<sup>14</sup> and modules developed in Perl and Python, which have been packaged for ease of deployment. All the analyses were run in parallel on a high-throughput compute cluster to ensure completion in a reasonable timeframe. Code is publicly available via GitHub (see software availability section)<sup>14-16</sup>.

### Data set validation

To assess our call set and compare it to the released data from the final phase of the 1000 Genomes Project, we utilised resources from GIAB. Our strategy compares our GRCh38 calls for NA12878 with those on the same assembly for that sample from GIAB. In addition, we compared the 1000 Genomes variant calls for NA12878 to the set of calls from GIAB for NA12878 on GRCh37. NA12878 was selected for this due to the availability of high quality call sets. In the sequence data used in generating our call set, NA12878 has lower coverage (4.6x) than the average coverage (6.2x) for the WGS alignment files and has higher coverage (144.1x) than the average coverage (84.9x) for the WES alignment files. We assume that the conclusions derived from NA12878 can be extrapolated to the rest of the samples and are likely to be conservative regarding the accuracy of our calls on GRCh38. Our comparison approach has benefits of both enabling us to benchmark the performance on a given sample with an independently produced gold-standard call set and allowing us to apply the equivalent benchmark to data from the 1000 Genomes Project, which gives a direct indication of how our call set compares to that produced by the 1000 Genomes Project.

In order to validate our data set we used the variants for NA12878 from the multi-sample phased VCF and compared them with the GIAB sites on GRCh38 downloaded from [[ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878\\_HG001/latest](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest)] (version 3.3.2). For comparative purposes we also compared the GRCh37 variants from the final phase of the 1000 Genomes Project (downloaded [here](#)) with the GRCh37 GIAB variants obtained [here](#) (version 3.3.2). Our comparison is restricted to regions of the genomes for which GIAB considers calls to be high confidence and was performed using the Nextflow<sup>17</sup> workflow accessible from the link in the software availability section.

The result of our comparison is shown in [Table 5](#). The average percentage of sites among all the chromosomes identified in our work that were also present in GIAB represents 96.4% of the total GIAB sites. This percentage is comparable to 97.9% resulting from the comparison with the final phase of the 1000 Genomes Project

**Table 5. Site comparison for NA12878 between our call set and Genome in a Bottle (GIAB)-mapped to GRCh38 and between the 1000 Genomes Project phase 3 (P3) call set and GIAB mapped to GRCh37.** Results are shown for each chromosome. *'Shared (TP)'* are the true positive variants identified in the compared call sets. *'giab\_only (FN)'* are the false negative variants identified by GIAB only. *'Thiswork\_only (FP)'* are the false positive variants identified in our call set only.

Dataset	Shared (TP)	%shared (TP)	giab_only (FN)	%giab_only (FN)	Thiswork_only (FP)	%Thiswork_only (FP)	Total (GIAB)	Total thiswork_only
<b>Chr1 (b38)</b>	238,323	96.37	8,965	3.63	1,347	0.56	247,288	239,670
<b>Chr1 (b37)</b>	242,331	98.09	4,707	1.91	1,700	0.70	247,038	244,031
<b>Chr2 (b38)</b>	237,017	96.42	8,791	3.58	1,264	0.53	245,808	238,281
<b>Chr2 (b37)</b>	260,921	98.14	4,942	1.86	1,209	0.46	265,863	262,130
<b>Chr3 (b38)</b>	214,201	96.17	8,520	3.83	1,134	0.53	222,721	215,335
<b>Chr3 (b37)</b>	218,474	97.93	4,608	2.07	926	0.42	223,082	219,400
<b>Chr4 (b38)</b>	188,608	96.00	7,860	4.00	847	0.45	196,468	189,455
<b>Chr4 (b37)</b>	232,888	97.93	4,927	2.07	888	0.38	237,815	233,776
<b>Chr5 (b38)</b>	181,015	96.26	7,031	3.74	865	0.48	188,046	181,880
<b>Chr5 (b37)</b>	193,359	95.48	9,162	4.52	766	0.39	202,521	194,125
<b>Chr6 (b38)</b>	197,830	96.04	8,151	3.96	940	0.47	205,981	198,770
<b>Chr6 (b37)</b>	191,018	98.05	3,801	1.95	844	0.44	194,819	191,862
<b>Chr7 (b38)</b>	166,888	96.54	5,982	3.46	854	0.51	172,870	167,742
<b>Chr7 (b37)</b>	167,924	97.98	3,464	2.02	712	0.42	171,388	168,636
<b>Chr8 (b38)</b>	145,748	96.24	5,700	3.76	678	0.46	151,448	146,426
<b>Chr8 (b37)</b>	171,950	97.76	3,937	2.24	715	0.41	175,887	172,665
<b>Chr9 (b38)</b>	131,987	96.42	4,899	3.58	635	0.48	136,886	132,622
<b>Chr9 (b37)</b>	132,596	97.84	2,924	2.16	581	0.44	135,520	133,177
<b>Chr10 (b38)</b>	153,504	96.55	5,480	3.45	815	0.53	158,984	154,319
<b>Chr10 (b37)</b>	153,080	97.87	3,338	2.13	648	0.42	156,418	153,728
<b>Chr11 (b38)</b>	154,516	95.83	6,720	4.17	775	0.50	161,236	155,291
<b>Chr11 (b37)</b>	155,511	97.86	3,407	2.14	609	0.39	158,918	156,120
<b>Chr12 (b38)</b>	136,457	96.46	5,008	3.54	745	0.54	141,465	137,202
<b>Chr12 (b37)</b>	148,026	98.03	2,972	1.97	676	0.45	150,998	148,702

Dataset	Shared (TP)	%shared (TP)	giab_only (FN)	%giab_only (FN)	Thiswork_only (FP)	%Thiswork_only (FP)	Total (GIAB)	Total thiswork_only
<b>Chr13 (b38)</b>	121,294	96.89	3,889	3.11	560	0.46	125,183	121,854
<b>Chr13 (b37)</b>	122,424	98.08	2,395	1.92	423	0.34	124,819	122,847
<b>Chr14 (b38)</b>	99,613	96.03	4,122	3.97	493	0.49	103,735	100,106
<b>Chr14 (b37)</b>	99,543	97.74	2,300	2.26	434	0.43	101,843	99,977
<b>Chr15 (b38)</b>	85,881	96.59	3,031	3.41	386	0.45	88,912	86,267
<b>Chr15 (b37)</b>	87,224	97.95	1,822	2.05	390	0.45	89,046	87,614
<b>Chr16 (b38)</b>	54,542	96.72	1,850	3.28	282	0.51	56,392	54,824
<b>Chr16 (b37)</b>	92,735	97.92	1,967	2.08	424	0.46	94,702	93,159
<b>Chr17 (b38)</b>	73,765	96.69	2,524	3.31	484	0.65	76,289	74,249
<b>Chr17 (b37)</b>	76,187	98.27	1,341	1.73	441	0.58	77,528	76,628
<b>Chr18 (b38)</b>	73,419	96.89	2,360	3.11	344	0.47	75,779	73,763
<b>Chr18 (b37)</b>	93,004	97.97	1,923	2.03	365	0.39	94,927	93,369
<b>Chr19 (b38)</b>	56,210	95.27	2,788	4.73	461	0.81	58,998	56,671
<b>Chr19 (b37)</b>	59,138	97.93	1,248	2.07	376	0.63	60,386	59,514
<b>Chr20 (b38)</b>	64,786	96.78	2,154	3.22	419	0.64	66,940	65,205
<b>Chr20 (b37)</b>	64,827	97.89	1,400	2.11	275	0.42	66,227	65,102
<b>Chr21 (b38)</b>	42,453	96.96	1,329	3.04	225	0.53	43,782	42,678
<b>Chr21 (b37)</b>	43,941	98.13	836	1.87	178	0.40	44,777	44,119
<b>Chr22 (b38)</b>	33,351	96.81	1,099	3.19	193	0.58	34,450	33,544
<b>Chr22 (b37)</b>	36,132	98.16	678	1.84	207	0.57	36,810	36,339
<b>ChrX (b38)*</b>	109	93.97	7	6.03	2	1.80	116	111
<b>AVG** (b38)</b>	129,609	96.41	4,921	3.59	670	0.53	134,530	130,280
<b>AVG (b37)</b>	138,329	97.86	3,095	2.14	627	0.45	141,424	138,955

\* Only PAR regions

\*\* Not considering chrX for the calculation

(P3). Additionally, the percentage of sites identified in our call set but not in GIAB is 0.5%, which is comparable to the 0.4% obtained in the comparison with 1000 Genomes P3. Taken together, these results demonstrate both the high sensitivity and high specificity of our callset.

### Data availability

The variants resulting from this work are available in the European Variation Archive. Accession number [PRJEB30460](https://www.eva.ac.uk/variation/variant/PRJEB30460).

This call set is also available from the International Genome Sample Resource (IGSR) [PMID: 27638885] at: [http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\\_collections/1000\\_genomes\\_project/release/20181203\\_biallelic\\_SNV/](http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/).

## Software availability

Task	Codebase	Documentation	Licence	DOI	Ref.
<b>eHive (workflow system)</b>	<a href="https://github.com/Ensembl/ensembl-hive">https://github.com/Ensembl/ensembl-hive</a>	<a href="https://ensembl-hive.readthedocs.io/en/version-2.5/">https://ensembl-hive.readthedocs.io/en/version-2.5/</a>	Apache 2.0	NA	14
<b>BAM quality control</b>	<a href="https://github.com/igsr/igsr_analysis/tree/master/PyHive/BamQC">https://github.com/igsr/igsr_analysis/tree/master/PyHive/BamQC</a>	<p><b>WGS BAM QC:</b> <a href="https://igsr-analysis.readthedocs.io/en/latest/workflows/wgs_bamqc_pipeline.html">https://igsr-analysis.readthedocs.io/en/latest/workflows/wgs_bamqc_pipeline.html</a></p> <p><b>WES BAM QC:</b> <a href="https://igsr-analysis.readthedocs.io/en/latest/workflows/wes_bamqc_pipeline.html">https://igsr-analysis.readthedocs.io/en/latest/workflows/wes_bamqc_pipeline.html</a></p>	Apache 2.0	<a href="http://doi.org/10.5281/zenodo.2573911">http://doi.org/10.5281/zenodo.2573911</a>	15
<b>Variant discovery</b>	<a href="https://github.com/EMBL-EBI-GCA/reseqtrack/tree/master/modules/ReSeqTrack/Hive">https://github.com/EMBL-EBI-GCA/reseqtrack/tree/master/modules/ReSeqTrack/Hive</a>	<a href="https://github.com/EMBL-EBI-GCA/reseqtrack/blob/master/docs/variantcalling_pipeline.txt">https://github.com/EMBL-EBI-GCA/reseqtrack/blob/master/docs/variantcalling_pipeline.txt</a>	Apache 2.0	<a href="https://doi.org/10.5281/zenodo.2573969">https://doi.org/10.5281/zenodo.2573969</a>	16
<b>Variant filtering</b>	<a href="https://github.com/igsr/igsr_analysis/tree/master/PyHive/PipeConfig/FILTER">https://github.com/igsr/igsr_analysis/tree/master/PyHive/PipeConfig/FILTER</a>	<p><b>BCFtools WGS variant filtering pipeline:</b> <a href="https://igsr-analysis.readthedocs.io/en/latest/workflows/bcftools_wgs_filtering_pipeline.html">https://igsr-analysis.readthedocs.io/en/latest/workflows/bcftools_wgs_filtering_pipeline.html</a></p> <p><b>BCFtools WES variant filtering pipeline:</b> <a href="https://igsr-analysis.readthedocs.io/en/latest/workflows/bcftools_wes_filtering_pipeline.html">https://igsr-analysis.readthedocs.io/en/latest/workflows/bcftools_wes_filtering_pipeline.html</a></p> <p><b>Freebayes variant filtering pipeline:</b> <a href="https://igsr-analysis.readthedocs.io/en/latest/workflows/freebayes_filtering_pipeline.html">https://igsr-analysis.readthedocs.io/en/latest/workflows/freebayes_filtering_pipeline.html</a></p> <p><b>GATK variant filtering pipeline:</b> <a href="https://igsr-analysis.readthedocs.io/en/latest/workflows/gatk_vc_filtering_pipeline.html">https://igsr-analysis.readthedocs.io/en/latest/workflows/gatk_vc_filtering_pipeline.html</a></p>	Apache 2.0	<a href="http://doi.org/10.5281/zenodo.2573911">http://doi.org/10.5281/zenodo.2573911</a>	15
<b>Variant integration</b>	<a href="https://github.com/igsr/igsr_analysis/tree/master/PyHive/PipeConfig/INTEGRATION/VCFIntegrationGATKUG.pm">https://github.com/igsr/igsr_analysis/tree/master/PyHive/PipeConfig/INTEGRATION/VCFIntegrationGATKUG.pm</a>	<a href="https://igsr-analysis.readthedocs.io/en/latest/workflows/consensus_callset_pipeline.html">https://igsr-analysis.readthedocs.io/en/latest/workflows/consensus_callset_pipeline.html</a>	Apache 2.0	<a href="http://doi.org/10.5281/zenodo.2573911">http://doi.org/10.5281/zenodo.2573911</a>	16
<b>Phasing</b>	<a href="https://github.com/igsr/igsr_analysis/blob/master/PyHive/PipeConfig/INTEGRATION/PHASING.pm">https://github.com/igsr/igsr_analysis/blob/master/PyHive/PipeConfig/INTEGRATION/PHASING.pm</a>	<a href="https://igsr-analysis.readthedocs.io/en/latest/workflows/phasing_pipeline.html">https://igsr-analysis.readthedocs.io/en/latest/workflows/phasing_pipeline.html</a>	Apache 2.0	<a href="http://doi.org/10.5281/zenodo.2573911">http://doi.org/10.5281/zenodo.2573911</a>	16
<b>Benchmarking using Genome in a Bottle</b>	<a href="https://github.com/igsr/igsr_analysis/blob/master/scripts/VCF/QC/compare_with_giab.nf">https://github.com/igsr/igsr_analysis/blob/master/scripts/VCF/QC/compare_with_giab.nf</a>	<a href="https://igsr-analysis.readthedocs.io/en/latest/workflows/compare_with_giab_pipeline.html">https://igsr-analysis.readthedocs.io/en/latest/workflows/compare_with_giab_pipeline.html</a>	Apache 2.0	<a href="http://doi.org/10.5281/zenodo.2573911">http://doi.org/10.5281/zenodo.2573911</a>	16

## Grant information

This work was completed thanks to the funding from the Wellcome Trust (grant number 104947) and the European Molecular Biology Laboratory.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

## Acknowledgements

We would like to thank Petr Danecek (Matthew Hurler Group, Wellcome Sanger Institute), Erik Garrison (Durbin Group, Wellcome Sanger Institute) and Tommy Carstensen (Global Health & Population Science, Department of Medicine, University of Cambridge) for participating in discussions on the methodology used in this work. Shane McCarthy (Department of Genetics, University of Cambridge) for detailed advice and discussion of the project plan. We would also like to thank Zamin Iqbal (Iqbal group, EMBL-EBI) for discussions on the project methodology and outputs. In addition, our thanks go to the Systems Infrastructure team of EMBL-EBI for providing continuous support and maintenance of the computing infrastructure required to complete this work. Finally, we would like to thank Tommy Carstensen for providing the leftover of the array data used for the phasing of the variants identified in this work.

Members of the 1000 Genomes Project Consortium are listed in the Supplementary Note, contained within the [Supplementary Text and Figures](#) of Poznik *et al.*<sup>18</sup>.

## Author information

Xiangqun Zheng-Bradley is currently at 'Illumina Center, Illumina UK Ltd., Cambridge, UK'.

## References

- 1000 Genomes Project Consortium, Auton A, Brooks LD, *et al.*: **A global reference for human genetic variation.** *Nature.* 2015; **526**(7571): 68–74.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zheng-Bradley X, Flicek P: **Applications of the 1000 Genomes Project resources.** *Brief Funct Genomics.* 2017; **16**(3): 163–170.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zheng-Bradley X, Streeeter I, Fairley S, *et al.*: **Alignment of 1000 Genomes Project reads to reference assembly GRCh38.** *Gigascience.* 2017; **6**(7): 1–8.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, *et al.*: **A map of human genome variation from population-scale sequencing.** *Nature.* 2010; **467**(7319): 1061–1073.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 1000 Genomes Project Consortium, Abecasis GR, Auton A, *et al.*: **An integrated map of genetic variation from 1,092 human genomes.** *Nature.* 2012; **491**(7422): 56–65.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Jun G, Flickinger M, Hetrick KN, *et al.*: **Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data.** *Am J Hum Genet.* 2012; **91**(5): 839–848.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing.** *arXiv preprint.* arXiv: 1207.3907 [q-bio.GN]. 2012.  
[Reference Source](#)
- McKenna A, Hanna M, Banks E, *et al.*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res.* 2010; **20**(9): 1297–1303.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zook JM, Chapman B, Wang J, *et al.*: **Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls.** *Nat Biotechnol.* 2014; **32**(3): 246–251.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Tan A, Abecasis GR, Kang HM: **Unified representation of genetic variants.** *Bioinformatics.* 2015; **31**(13): 2202–2204.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.** *Am J Hum Genet.* 2007; **81**(5): 1084–1097.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Delaneau O, Marchini J, Zagury JF: **A linear complexity phasing method for thousands of genomes.** *Nat Methods.* 2011; **9**(2): 179–181.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Delaneau O, Marchini J, 1000 Genomes Project Consortium *et al.*: **Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel.** *Nat Commun.* 2014; **5**: 3934.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Severin J, Beal K, Vilella AJ, *et al.*: **eHive: an artificial intelligence workflow system for genomic analysis.** *BMC Bioinformatics.* 2010; **11**: 240.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lowy E, GabeAldam, Fairley S: **igsr/igsr\_analysis: First release of code (Version v1.0.0).** *Zenodo.* 2019.  
<http://www.doi.org/10.5281/zenodo.2573911>
- istreetreter, Richardson D, HollyZB, *et al.*: **EMBL-EBI-GCA/reseqtrack: zenodo (Version zenodo).** *Zenodo.* 2019.  
<http://www.doi.org/10.5281/zenodo.2573969>
- Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Poznik GD, Xue Y, Mendez FL, *et al.*: **Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences.** *Nat Genet.* 2016; **48**(6): 593–599.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

# Open Peer Review

Current Peer Review Status:  

---

## Version 1

Reviewer Report 10 April 2019

<https://doi.org/10.21956/wellcomeopenres.16504.r35054>

© 2019 Rendon A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Augusto Rendon** 

Genomics England, London, UK

### Summary

The authors present a new call set from the 1000 genomes project. This time the call set is a fresh recall of the data against GRCh38. The call set is only composed of biallelic SNPs (biallelic in this study). Previous variant call sets on GRCh38 had been lifted over from their GRCh37 native counterparts. The study identifies SNPs by first calling variants with several algorithms, creating an union set and then explicitly genotyping these sites across the complete data set. The genotypes are then phased. A comparison with GIAB data for NA12878 is performed to assess the sensitivity and specificity of the call set.

### General observations

A native call set of the 1000 genomes project data on GRCh38 is quite an important effort. These data have many important downstream uses including clinical genomics uses for variant filtering and population genomics studies. Notwithstanding the importance of this data set, I have issues with the data note as it stands. I think the paper simply described what the authors did to generate these data but makes little effort to explain why it was done in this way. The latter is important to gain confidence in the call set.

As a user of the data I would look to be satisfied with the quality of the call set. This is particularly important as several other large scale projects have released allele frequencies of many thousands of deep sequenced whole genomes (e.g. Topmed and Gnomad). In parallel, frameworks for assessing the analytical performance of variant calls have been proposed by GIAB and GA4GH and recently published. Truth sets for several important samples beyond the NA12878 have been released; including the ability to assess to how well phased the data sets was. As explained in the specific comments below, I believe more work should have been done to convince the reader that sufficient due diligence has been committed to this data set.

From the perspective of having a native call set on GRCh38 it is a shame that there is very little to show the potential benefits of this call set on GRCh38. There is no mention [or I really missed it] about how

alternative haplotypes were handled and the implications of having variants in alternative haplotypes. The data release also ignores any non diploid areas of the genome.

From a methodological perspective, the tool chain feels outdated with BCFtools and GATK being at least 2 years old. Thresholds are widely used but little work is done to explain how these thresholds are determined. I assume that the parameters used in the tools themselves, these are standard or perhaps defined in previous iterations of the project. However, the filtering thresholds in Tables 1-4, at least from the reading, sounds like they were plucked out of thin air.

On a more positive note, I would like to commend authors for the great effort placed to make available the code, organising it, document it, and making this data set reproducible.

### Specific observations

1. *\*Tool chain is really outdated\**. Is this because there have been little changes for the specific algorithms used (mpileup and unified genotyper?). I assume that many improvements and bugfixes have appeared in the last two years plus since these versions were released.
2. *\*Demonstrating improvements on GRCh38 and why it is better than the lift over\**. I missed some figure showing how this was an improvement. For example a comparison of the lifted over and the native call set. Are there regions of the genome where they perform different? Was it worth the effort? What about the ALTs, have we now better allele frequencies on these regions? How do they affect the frequencies on the corresponding frequencies on the primary assembly?
3. *\*Variant normalisation\**. This is glossed over in the text and little is said about how it was performed. In my experience this step often introduces difficult tradeoffs. Please expand on this.
4. *\*Benchmarking\**. This area is quite lacking here. I understand that you are only looking at biallelic sites so comparison with a truth set is easier. However, standards to doing this have existed for over a year and recently published. They are based on comparing at the haplotype level and not the site level.
5. *\*Phasing\**. These standards also enable comparing the phasing accuracy. Given the effort placed here to phase the genotypes, it would be helpful to also benchmark the phasing data.
6. *\*Union set\**. I would have loved to see a figure that shows how the various variant callers contributed to the union call set. Is there one that is unnecessary? Is there one that is responsible for many of the false positives?
7. *\*Analytical performance\**. "Taken together, these results demonstrate both the high sensitivity and high specificity of our callset". You seem to have less TP and more FN than in the 37 release. Now a days these numbers do not represent high sensitivity and specificity (at least with 30X genomes). At least a more nuanced discussion is required here. This is strongly linked to the thresholds you chose for various filtering steps.
8. *\*Truth sets\**. It would be very helpful to add further truth sets, for example for the Ashkenazim and Chinese trios.
9. *\*Joint calling\**. The approach here was to do single sample calling, assembling a union set and then genotyping those sites. Could the authors please explain why this approach as opposed to joint calling. I would assume that for low coverage genomes, joint calling may be more powerful as it can leverage information across more samples.

### References

1. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, Gonzalez-Porta M, Eberle MA, Tezak Z, Lababidi S, Truty R, Asimenos G, Funke B, Fleharty M, Chapman BA, Salit M, Zook JM, Global Alliance for Genomics and Health Benchmarking Team: Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol.* 2019. [PubMed Abstract](#) | [Publisher Full Text](#)
2. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean



CY, De La Vega FM, Xiao C, Sherry S, Salit M: An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol.* 2019. [PubMed Abstract](#) | [Publisher Full Text](#)

**Is the rationale for creating the dataset(s) clearly described?**

Yes

**Are the protocols appropriate and is the work technically sound?**

Partly

**Are sufficient details of methods and materials provided to allow replication by others?**

Yes

**Are the datasets clearly presented in a useable and accessible format?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Bioinformatics, clinical genomics

**I have read this submission. I believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.**

Reviewer Report 08 April 2019

<https://doi.org/10.21956/wellcomeopenres.16504.r35051>

© 2019 Church D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Deanna M. Church** 

Inscripta, Inc., Boulder, CO, USA

**Summary**

In the work entitled 'Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project', Lowy-Gallego *et al.* describe their effort to re-analyze the 1000 genomes data on the current GRCh38 assembly. They do not perform a full variant analysis, but rather release a set of biallelic SNVs as a preliminary variant call set. They compare this variant set to the Genome in a Bottle (GIAB) variant calls on the sample NA12878.

**High Level Comments**

It is great to see efforts to update important datasets onto the current human reference assembly, GRCh38. As the authors note, the GRCh38 reference represents a substantial improvement over the GRCh37 reference, but the lack of GRCh38 based annotation has hindered adoption of this version of the reference assembly. The authors go on to discuss why 'lift-over' based approaches are inadequate, which motivated this work. I agree that 'lift-over' based approaches are inadequate, but I find the results

presented in this manuscript unconvincing with respect to this assertion.

The authors spend significant space in the introduction on both explaining the improvements in GRCh38, including the addition of alternate loci, but then put no effort into demonstrating why these are valuable. Additionally, the authors spend time discussing why 'lift-over' based approaches are inadequate, but then do no comparisons to show why their *de novo* approach is an improvement.

While I believe this work is important, I feel the authors fail to make the point of why doing this *de novo* analysis on the GRCh38 reference is important.

## Detailed Comments

**1. Explanation of why 'lift-over' approaches have limitations:** I agree with the statement that 'lift-over' is inadequate. However, the description of this on page 1 is not clear. Statement 1 'they rely on an equivalent region existing in the new genome, so new sequence in the improved assembly is effectively excluded' confounds two points. Regions that are present on the old assembly but not the new one will be excluded from a 'lift-over' approach. Additionally, sequence that is new on the updated reference will also be omitted - but these are two separate cases.

Point 2, relating two alignments also confounds multiple issues. Yes - correct alignments are key to the 'lift-over' approach, but there are two 'bad alignment' cases. The case I **think** the manuscript is referring to is a case where increased diversity in one version of the assembly can confound alignments (that is, sequence change). The other relevant case is the addition of paralogous sequence to one assembly that is missing from the other. This can lead to a locus aligning to a paralogous region rather than the equivalent locus (I have seen examples of this), that can also lead to incorrect 'lift-over'. Point three in this statement is a clear statement, but the authors provide no evidence to actually support this.

**2. The authors only provide biallelic SNPs:** I can see the utility in concentrating on a restricted set of variants, but only if this set of data are actually used to demonstrate the value of *de novo* analysis over 'lift-over', which was not done in this manuscript. On page 3, the authors state that "These represent the major part of the SNVs present in the human genome." but I'd like more hard numbers on this. What percentage of all SNVs do the biallelics represent? What percentage of all variation do they represent?

**3. Page 3, Quality control of alignment files:** Are the steps presented here just the differences from the original protocol? I think this is OK, but it is not clear from reading the manuscript if this is the full set of steps or just the differences.

**4. Variant discovery:** Why did you use the variant calling tools you chose?

**5. Variant Filtering:** The omission of variants from the sex chromosomes seems like a significant omission and limits the use of this dataset.

**6. Data set validation:** I have significant concerns here. I understand why NA12878 was used for some validation. However, my understanding is that the GIAB dataset does not take the alternate loci into account in their variant calling, while this manuscript tries to take advantage of these sequences - how did this impact the comparison? For example, I would predict more conflicts in regions where alt-loci exist in GRCh38. Does this occur?

I am also not convinced that accuracy on NA12878 really translates well to other samples, particularly non-European samples (as NA12878 has a European ancestry). Will the accuracy really extend to non-European samples? Also, my reading of Table 5 is that this dataset performs slightly worse than the

GRCh37 call set. This does not do a lot to convince this reader that the work of doing the re-analysis is worthwhile - and I'm a believer, based on previous work I have been a part of! I have some concerns that this may be due to improvements in the new call set (due to the inclusion of the alts and more complex decoy) but it takes some significant work to track this down. There are examples of this kind of analysis<sup>1</sup>. The authors should also clearly state what fraction of the genome they are able to assess using this method.

**7. Omitted analysis:** The authors discuss the value of the improved reference in the introduction, but then do nothing to show the value of the alternate loci. How many new variants are identified on these loci? How does the inclusion of these sequences change variant calling on the primary? Perhaps most disappointing is that there is no analysis of how the *de novo* is an improvement over 'lift-over' approaches. How do the *de novo* variant calls compare to the 'lift-over' calls? Without this analysis, it is unclear to me that anyone would be convinced that doing the *de novo* call approach is worth the effort.

Lastly, the authors miss the opportunity to do an accuracy comparison by looking at the regions of the reference comprised of the 'ABC' clones. These are fosmid libraries constructed from several of the samples that went into the 1000 genomes project. These provide a great test bed for both looking at variant calls (any call in this region should be heterozygous or hemizygous as the reference sequence represents one valid haplotype in the sample being analyzed) and also allows for the confirmation of the local haplotype assertions.

## References

1. Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A, Fiddes IT, Galvin B, Heaton H, Herschleb J, Hindson C, Holt E, Jabara CB, Jett S, Keivanfar N, Kyriazopoulou-Panagiotopoulou S, Lek M, Lin B, Lowe A, Mahamdallie S, Maheshwari S, Makarewicz T, Marshall J, Meschi F, O'Keefe CJ, Ordonez H, Patel P, Price A, Royall A, Ruark E, Seal S, Schnall-Levin M, Shah P, Stafford D, Williams S, Wu I, Xu AW, Rahman N, MacArthur D, Church DM: Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res.* **29** (4): 635-645 [PubMed Abstract](#) | [Publisher Full Text](#)

**Is the rationale for creating the dataset(s) clearly described?**

Partly

**Are the protocols appropriate and is the work technically sound?**

Partly

**Are sufficient details of methods and materials provided to allow replication by others?**

Yes

**Are the datasets clearly presented in a useable and accessible format?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Genomics, genome assembly, genome annotation, variant calling, genome engineering.

**I have read this submission. I believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.**

