DATA NOTE

Variant calling across 505 openly consented samples from four Gambian populations on GRCh38 [version 1; peer review: 1 approved with reservations]

Ernesto Lowy, Susan Fairley, Paul Flicek

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

Abstract
The International Genome Sample Resource (IGSR) repository was established to maximise the utility of human genetic data derived from openly consented samples within the research community. Here we describe variant detection in 505 samples from four populations in The Gambia, using the GRCh38 reference genome, adding to the range of populations for which this has been done and, importantly, making allele frequencies available. A multi-caller site discovery process was applied along with imputation and phasing to produce a phased biallelic single nucleotide variant (SNV) and insertion/deletion (INDEL) call set. Variation had not previously been explored on the GRCh38 human genome assembly for 387 of the samples. Compared to our previous work with the 1000 Genomes Project data on GRCh38, we identified over nine million novel SNVs and over 870 thousand novel INDELS.

Keywords
Genomics, population genetics, variant calling, single nucleotide variation, variant discovery
Introduction
The 1000 Genomes Project collected samples from self-declared healthy individuals across numerous specific populations, organised into five continental super-populations with the aim of cataloguing common human genetic variation (1000 Genomes Project Consortium et al., 2015). In this work, 100 samples were collected from each of four populations in The Gambia. However, it proved impossible to generate immortalised cell lines from these samples, which would have allowed DNA to be obtained for further studies. As cell lines were required for samples to be used in the 1000 Genomes Project, this inability prevented these 400 samples from being included in the 1000 Genomes Project. A subsequent set of over 100 additional samples were collected from one of the original four populations, the Mandinka. For these samples, cell line creation was successful, and they were included in the 1000 Genomes Project, forming the “Gambian in Western Division, The Gambia - Mandinka” population, and assigned the three-letter code GWD.

Through its sampling strategy (sampling around 100 individuals in a population), the 1000 Genomes Project enabled population-specific allele frequencies to be determined. These data are useful for many studies including as way to assess the potential clinical significance of a variant, as a common variant is unlikely to be pathogenic (Zheng-Bradley & Flicek, 2017).

While extensive, the populations present in the 1000 Genomes Project are not exhaustive and, given the higher levels of genetic variation found in individuals with African ancestry, the loss of the original populations from The Gambia diminished the range of genetic variation captured by the project.

Although cell line creation was not successful for the original samples, efforts were made to analyse the material that was available. The samples formed the basis of the Gambian Genome Variation Project (GGVP) and data was generated. The Kwiatkowski lab at the Wellcome Sanger Institute (WSI) and the Nuffield Department of Medicine, University of Oxford succeeded in creating both array-based genotyping data and whole genome sequencing (WGS) data of around 4x coverage from the samples. These data have been used, in conjunction with whole-genome sequences of individuals from additional groups, to examine genetic susceptibility to malaria (Malaria Genomic Epidemiology Network, 2019), using variants called using GATK (McKenna et al., 2010) on the GRCh37 human genome assembly.

In this work, we aimed to place variant discovery in the GGVP samples on a par with those in the 1000 Genomes Project. We applied a multi-caller site discovery method, with subsequent phasing and imputation across the original ~400 samples and the 118 GWD samples included in the 1000 Genomes Project (hereafter referred to simply as ‘118 GWD samples’). This work used GRCh38, the current and most complete human reference genome, giving the best available foundation for variant calling.

Here, we use a modified version of the pipeline we developed to call variants from the low-coverage WGS and exome data from the original 1000 Genomes Project (Lowy-Gallego et al., 2019). This pipeline is tailored to those data types and array data, and differs from standard, modern production pipelines designed to operate on high-coverage WGS. We describe the use of this pipeline to produce a phased single nucleotide variant (SNV) and insertion/deletion (INDEL) call set for both the GGVP samples and the 118 GWD samples.

Materials and methods
Sequence Data
The sequence data used in this work comes from 505 samples from four different Gambian populations generated by the GGVP and the 100 Genomes Project. All data is available in the European Nucleotide Archive (ENA) (see Table 1).

Alignment of the sequence data
Low-coverage alignments. The GGVP sequence data deposited in ENA was aligned to a copy of the GRCh38 reference genome including the alternative sequences, decoys and human leukocyte antigen (HLA). Alignments were generated using BWA-mem in an alt-aware manner, as described in http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/gambian_genome_variation_project/README_gambian_genome_variation_project.md and as used in previous work (Zheng-Bradley et al., 2017). The resulting alignments are listed on the IGSR (Fairley et al., 2020) FTP site.
The alignments for the 118 samples in the 1000 Genomes Project were generated as described previously (Zheng-Bradley et al., 2017). These alignments are listed on the IGSR FTP site. The alignment process, in addition to using BWA, makes use of additional processing steps, such as realignment around known indels. Full details are in the provided references.

We adopted the alignment filtering approach described in (Lowy-Gallego et al., 2019) (see section ‘Quality control of the alignment files’). Briefly, chk_indel_rg was used to discard alignment files with an unbalanced ratio of short INDELs (greater than 5). Picard CollectWgsMetrics was used to discard the alignment files with a mean non-duplicated aligned coverage level <=2x and VerifyBAMID (Jun et al., 2012) was used to assess sample contamination and mix-ups and the following cutoffs were used:

\[
\text{free\_mix} > 0.03 \text{ and chip\_mix} > 0.02
\]

Only files passing quality control were used in site discovery.

**High-coverage alignments used in genotyping.** High-coverage (30x) sequencing data for the 118 GWD samples has recently been produced and aligned to GRCh38 by the Zody lab at the New York Genome Center (NYGC) (funded by NHGRI Grant 3UM1HG008901-03S1) (Byrská-Bishop et al., 2021). These alignments were submitted to ENA (PRJEB31736) and are listed on the IGSR FTP site. These alignments were used in the genotyping step for the 118 GWD samples.

**Variant calling**

**Site discovery.** We used joint genotyping (where information from all samples is considered during variant calling) and three different variant calling methods. These produced three call sets that were later integrated (see below). Calls were made only on the primary assembly autosomes and in the pseudoautosomal regions (PARs) of chromosome X. The strategy is similar to that described in (Lowy-Gallego et al., 2019) but does differ in some details, mainly in that we have used an automatic machine learning (ML)-based method to filter the BCFTools and Freebayes call sets.

Below, we list the methods and commands used.

**BCFTools**

- BCFTools (v.1.9)

  bcf tools mpileup -E -a DP -a SP -a AD -p ILLUMINA -pm3 -F0.2 -C50 -d 700000 -f $ref.fa $file.bam | bcf tools call -mv -O z --ploidy GRCh38 -S $ samples.ped -o $out.vcf.gz
**GATK UnifiedGenotyper**

- GATK UnifiedGenotyper (v. 3.8-1-0-gf15c1c3ef)

```
java -Xmx6g -jar GenomeAnalysisTK.jar 
  -T UnifiedGenotyper 
  -R $ref.fa 
  -I $file.bam 
  -o $out.vcf.gz 
  -dcov 250 
  -stand_call_conf 10 
  -glm both 
  --dbsnp dbsnp_146.hg38.vcf.gz
```

**FreeBayes**

- Freebayes (v. 1.3.1-dirty)

```
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**

The variant discovery pipeline produced three initial call sets as described above. To create the integrated call set, we discarded variants in the centromeres, as these are regions of low complexity, which hinder variant calling (Saha et al., 2019). Additionally, filtering methods were applied to each of the input call sets. These are described below and vary depending on the method used to generate the specific input call set.

**GATK UnifiedGenotyper callset.** We used the Variant Quality Score Recalibration (VQSR) method (McKenna et al., 2010) following the GATK best practices and GATK training call sets. The combination of commands we used is very similar to the procedure described in (Lowy-Gallego et al., 2019) (see section 'Variant filtering-GATK UnifiedGenotyper call set'), although this time, we ran GATK ApplyRecalibration with --ts_filter_level values of 99.9 and 99.5 for SNV and INDEL, respectively.

**Freebayes and BCFTools call sets.** These two call sets were each filtered twice in separate applications of the process: once to filter SNVs and another to filter INDELs. This used a supervised ML filtering pipeline (Lowy, 2021) to predict a binary outcome: variant or non-variant site. This particular type of problem can be modelled using a logistic regression binary classifier, and more specifically, our pipeline uses the Scikit-learn Python library implementation. The supervised ML model needs to be trained before it can be used for filtering.

The following commands were used to train the model for the FreeBayes callset. We generated a different model for each chromosome.

**FreeBayes SNV training**

As a truth set for training, we used the high-confidence 1000 Genomes Project phase one SNVs from the GATK bundle:

```
extflow run train.nf --vcf ggv.freebayes.snps.vcf.gz --true_cs 1000G_phase1.snps.high_confidence.hg38.$chr.vcf.gz --annotations \%CHROM\tPOS\tQUAL\tINFO/ABP\tINFO/DP\tINFO/DPB\tINFO/DPA\tINFO/EPP\tINFO/GTI\tINFO/MEANALT\tINFO/MQ\tINFO/NUMALT\tINFO/PAO\tINFO/PQA\tINFO/PRO\tINFO/QA\tINFO/R-Cal\tINFO/RPL\tINFO/RPP\tINFO/RPR\tINFO/GTI/SAF\tINFO/SAP\tINFO/SAR\tINFO/SRF\tINFO/SRR

--vt snps --outprefix ggvp.freebayes.snps.$chr
```
**FreeBayes INDEL training**
We used the high-confidence Mills_and_1000G_gold_standard.indels.hg38.vcf.gz from the GATK bundle as the truth set for training:

```
nextflow run train.nf --vcf ggvp.freebayes.indels.vcf.gz --true_cs Mills_and_1000G_gold_standard.indels.hg38.vcf.gz --annotations "%CHROM\tPOS\tQUAL\n\tINFO/AB\tINFO/ABP\tINFO/DP\tINFO/DPB\tINFO/DPR\tINFO/EP\tINFO/GTR\tINFO/MANALT\n\tINFO/MEANALT\tINFO/MQM\tINFO/NUMALT\n\tINFO/PA\tINFO/PQA\tINFO/PRO\tINFO/Qt\INFO/RO\tINFO/SAFE\tINFO/SAR\tINFO/SRF\tINFO/SRR\n' --vt indels --outprefix ggvp_bam.freebayes.indels.$chr
```

Then we applied the models using the following commands:

**FreeBayes SNV filtering**
```
nextflow run apply.nf --vcf ggvp.freebayes.$chr.vcf.gz --model fitted_snps_$chr_logreg_vts.sav --annotations "%CHROM\tPOS\tQUAL\n\tINFO/DP\tINFO/BQB\tINFO/HOB\tINFO/ICB\tINFO/MQ\tINFO/MQ0F\tINFO/MQB\n\tINFO/MQSB\tINFO/RPB\tINFO/SGB\tINFO/VDB\n' --vt snps --chr $chr --cutoff 0.3 --outprefix ggvp_freebayes.$chr.filt
```

**FreeBayes INDEL filtering**
```
nextflow run apply.nf --vcf ggvp.freebayes.$chr.vcf.gz --model fitted_indels_$chr_logreg_vts.sav --annotations "%CHROM\tPOS\tQUAL\n\tINFO/AB\tINFO/ABP\tINFO/DP\tINFO/DPB\tINFO/DPR\tINFO/EP\tINFO/GTR\tINFO/MANALT\n\tINFO/MEANALT\tINFO/MQM\tINFO/NUMALT\n\tINFO/PA\tINFO/PQA\tINFO/PRO\tINFO/Qt\INFO/RO\tINFO/SAFE\tINFO/SAR\tINFO/SRF\tINFO/SRR\n' --vt indels --chr $chr --cutoff 0.2 --outprefix ggvp_freebayes.$chr.filt
```

For the BCFTools call set, we trained the models using the same truth sets (from the GATK bundle) and used the following commands:

**BCFTools SNV training**
```
nextflow run train.nf --vcf ggvp.bcftools.snps.vcf.gz --true_cs 1000G_phase1.snps.high_confidence.hg38.$chr.vcf.gz --annotations "%CHROM\tPOS\tQUAL\n\tINFO/DP\tINFO/BQB\tINFO/HOB\tINFO/ICB\tINFO/MQ\tINFO/MQ0F\tINFO/MQB\n\tINFO/MQSB\tINFO/RPB\tINFO/SGB\tINFO/VDB\n' --vt snps --region $chr --outprefix ggvp.bcftools.$chr.snps
```

**BCFTools INDEL training**
```
nextflow run train.nf --vcf ggvp.bcftools.indels.vcf.gz --true_cs Mills_and_1000G_gold_standard.indels.hg38.vcf.gz --annotations "%CHROM\tPOS\tQUAL\n\tINFO/IDV\tINFO/IMF\tINFO/DP\tINFO/VDB\tINFO/MQSB\tINFO/SGB\tINFO/MQ0F\n\tINFO/ICB\tINFO/HOB\tINFO/MQ\n' --vt indels --region $chr --outprefix ggvp.bcftools.$chr.indels
```

Then we applied the trained model to:

**BCFTools SNV filtering**
```
nextflow run apply.nf --vcf ggvp.bcftools.$chr.vcf.gz --model fitted_snps_$chr_logreg_vts.sav --annotations "%CHROM\tPOS\tQUAL\n\tINFO/DP\tINFO/BQB\tINFO/HOB\tINFO/ICB\tINFO/MQ\tINFO/MQ0F\tINFO/MQB\n\tINFO/MQSB\tINFO/RPB\tINFO/SGB\tINFO/VDB\n' --vt snps --chr $chr --cutoff 0.1 --outprefix ggvp.bcftools.$chr.filt
```
**BCFTools INDEL filtering**

```
nextflow run applym.nf --vcf ggvp.bcftools.$chr.vcf.gz --model fitted_indels_$chr_logreg_vts.sav --annotations '%CHROM\t%POS\t%QUAL\t%INFO/IDV\t%INFO/IMF\t%INFO/DP\t%INFO/VDB\t%INFO/MOSB\t%INFO/MQ0F\t%INFO/ICB\t%INFO/HOB\t%INFO/MQ\n' --vt indels --chr $chr --cutoff 0.1 --outprefix ggvp.bcftools.$chr.filt
```

The three filtered call sets are available on the IGSR FTP site.

**Generating a consensus call set**

To generate a consensus call set, we integrated the call sets described above and the calls for the 118 GWD samples (which were produced in the work described in Lowy-Gallego et al., 2019).

**Normalisation, genotype calling and integration.** The sets were normalised ahead of genotype calling as described previously (Lowy-Gallego et al., 2019) (see section ‘Generating consensus call sets’).

We generated a consensus call set by taking the union of the biallelic sites from each normalised call set and calculated the genotype likelihoods for each site using GATK UnifiedGenotyper in ‘genotype_given_alleles’ (GGA) mode. Alignment files are required for the genotyping step. For the GGVP samples, we used the alignments of the low-coverage data (see above). For the 118 GWD samples, alignments of high-coverage data from NYGC were used (see above).

The GATK command for generating the consensus call was described previously (Lowy-Gallego et al., 2019) (see section ‘Generating consensus call sets’).

**Filtering the consensus call set.** We applied the VQSR method. This was used for SNVs and INDELs independently, with the same parameters and training call sets described previously (Lowy-Gallego et al., 2019) (see section ‘Generating consensus call sets’). GATK ApplyRecalibrator was used with a `--ts_filter_level` value of 99.7 for SNVs and a `--ts_filter_level` value of 99.0 for INDELs.

**Phasing and imputation of the consensus call set.** The VCF file containing the resulting genotype likelihoods was processed using Beagle (v. 4.1) (Browning & Browning, 2007). After this, SHAPEIT2 (version v2.r837) (Delaneau et al., 2011) was used, along with microarray genotype data for the samples (available on the IGSR FTP site) (Malaria Genomic Epidemiology Network, 2019), following the same process described previously (Lowy-Gallego et al., 2019; see section ‘Phasing and imputation of the consensus call set’) to produce the final call set.

**Call set description and quality control**

**Comparison of site discovery in the samples from The Gambia with the GRCh38 phase three recall.** Table 2 and Table 3 compare the SNV and INDEL sites identified in this work, looking solely at populations from The Gambia, with the sites detected in our previous recall of the 1000 Genomes Project phase three data on GRCh38 (Lowy-Gallego et al., 2019). As noted, the 118 GWD samples are present in both data sets and the sites for these samples in the earlier work were included in the consensus call set described here. The previous work spanned 26 populations and 2,548 samples, compared with four populations and 505 samples in this work. As can be seen in Table 2 and Table 3, over nine million SNV sites were identified uniquely in this work and over 870,000 INDELS.

**Comparing calls for the GWD samples in this and previous work.** When recalling the 1000 Genomes Project phase three data realigned to GRCh38, we were able to use the presence of NA12878, for which gold-standard data is available from Genome In A Bottle (GIAB) (Zook et al., 2016), to assist in benchmarking (Lowy-Gallego et al., 2019). This was one of the methods by which the quality of that call set was assessed.

For the samples from The Gambia used in this work, no gold-standard data, similar to GIAB, was available. However, having previously assessed the calls for the 1000 Genomes Project data on GRCh38, we can compare the calls in that call set and the calls in this work for the 118 GWD samples which are present in both sets. As such, we treat the previous calls for these samples as our ‘gold-standard’ for benchmarking purposes.
Table 2. Comparing single nucleotide variant (SNV) sites on each chromosome. ‘Common sites’ are those present in both the call set produced by this work and the recall of the 1000 Genomes Project Phase Three data on GRCh38. ‘Gambian only’ sites are the sites which are only found in this work, calling across all samples in the four populations from The Gambia, from the Gambian Genome Variation Project (GGVP) and the 1000 Genomes Project. ‘GRCh38 P3 only’ are sites that were detected in the recall of the 1000 Genomes Project data but not the work described here. ‘Total’ gives the total number of sites when sites are combined from the two call sets.

| Chrom | Common sites (%) | Gambian only (%) | GRCh38 P3 only (%) | Total |
|-------|-----------------|-----------------|--------------------|-------|
| 1     | 1,843,975       | 28.72           | 730,868            | 34.88 |
| 2     | 2,020,076       | 28.89           | 751,862            | 10.75 |
| 3     | 1,702,151       | 29.52           | 590,392            | 10.24 |
| 4     | 1,702,073       | 30.37           | 577,691            | 10.31 |
| 5     | 1,557,283       | 29.72           | 539,484            | 10.29 |
| 6     | 1,510,229       | 30.44           | 502,619            | 10.13 |
| 7     | 1,398,486       | 29.78           | 547,034            | 11.65 |
| 8     | 1,361,677       | 29.83           | 477,988            | 10.47 |
| 9     | 1,032,803       | 28.75           | 471,369            | 13.12 |
| 10    | 1,195,313       | 29.78           | 446,582            | 11.13 |
| 11    | 1,170,777       | 29.31           | 417,851            | 10.46 |
| 12    | 1,134,595       | 29.54           | 405,191            | 10.55 |
| 13    | 852,855         | 30.34           | 282,675            | 10.06 |
| 14    | 770,340         | 29.29           | 289,487            | 11.01 |
| 15    | 697,407         | 29.02           | 287,575            | 11.97 |
| 16    | 762,432         | 28.23           | 334,020            | 12.37 |
| 17    | 660,316         | 28.26           | 306,758            | 13.13 |
| 18    | 673,782         | 30.06           | 231,092            | 10.31 |
| 19    | 540,795         | 29.39           | 242,071            | 13.15 |
| 20    | 556,601         | 28.97           | 246,673            | 12.84 |
| 21    | 329,200         | 28.87           | 180,724            | 15.85 |
| 22    | 327,783         | 27.70           | 207,106            | 17.50 |
| X     | 40,245          | 32.14           | 28,561             | 22.81 |
| All   | 23,841,194      | 29.43           | 9,095,673          | 11.23 |

Table 3. Comparing insertion/deletion (INDEL) sites on each chromosome. ‘Common sites’ are those present in both the call set produced by this work and the recall of the 1000 Genomes Project Phase Three data on GRCh38. ‘Gambian only’ sites are the sites only which are found in this work, calling across all samples in the four populations from The Gambia, from the Gambian Genome Variation Project (GGVP) and the 1000 Genomes Project. ‘GRCh38 P3 only’ are sites that were detected in the recall of the 1000 Genomes Project data but not the work described here. ‘Total’ gives the total number of sites when sites are combined from the two call sets.

| Chrom | Common sites (%) | Gambian only (%) | GRCh38 P3 only (%) | Total |
|-------|-----------------|-----------------|--------------------|-------|
| 1     | 161,851         | 34.88           | 71,284             | 15.36 |
| 2     | 175,193         | 34.79           | 74,209             | 14.74 |
| 3     | 147,916         | 35.45           | 60,488             | 14.50 |
| 4     | 151,615         | 36.76           | 58,868             | 14.27 |
Our comparison is restricted to calls in high confidence regions from GIAB in both the autosomes and pseudoautosomal regions (PARs). On average 77.9% (standard deviation 12.1%) of the bases of each chromosome are in high confidence regions. The comparison was performed using the Nextflow-based (Di Tommaso et al., 2017) workflow, accessible from the link in the data availability section of this manuscript and in Table 4.

We summarised the result of our comparison by calculating the positive predictive value (PPV), which is the proportion of the identified sites that are ‘true’ variants (i.e. those that correspond to calls in the 1000 Genomes Project, on the GRCh38 GWD call set), where 1 would indicate all ‘true’ variants had been found and 0 that no ‘true’ variants had been detected. The PPV value was calculated for each of the 118 samples for both SNVs and INDELs. Figure 1 shows that the PPV values for SNVs (Figure 1A, median: 0.96) and INDELs (Figure 1B, median: 0.79) exhibit a tightly grouped distribution, without extreme outliers that have a larger mean for SNVs than for INDELs. This was expected as INDEL variants are more difficult to identify.

Contribution of each of the four calling methods to the final integrated call set. To illustrate the contribution of each of the call sets that were integrated to create the final call set, we generated the plots in Figure 2 and Figure 3 for SNVs and INDELs, respectively. Figure 2 shows that the GATK UnifiedGenotyper call set contributed the most to the final SNV integrated call set (32,394,305 variants), followed by the BCFTools call set (26,982,090 variants). Figure 3 shows that in the case of INDELs, the GATK UnifiedGenotyper call set has contributed the most to the final integrated call set (2,417,049 variants), followed by the call set obtained for the 118 GWD samples from the previous work (1,888,782 variants) generated using GATK. ‘Integrated’ is the final INDEL call set generated after integrating the supporting call sets. Vertical bars show the size of the intersection between the call sets. Horizontal bars show the aggregated size of each call set. We used the filtered supporting call sets to generate this plot.

| Chrom | Common sites (%) | Gambian only (%) | GRCh38 P3 only (%) | Total |
|-------|------------------|------------------|--------------------|-------|
| 5     | 135,428          | 35.82            | 54,706             | 49.71 | 378,046 |
| 6     | 136,408          | 36.38            | 54,483             | 49.09 | 374,924 |
| 7     | 119,361          | 35.44            | 51,180             | 49.36 | 336,790 |
| 8     | 109,471          | 35.91            | 44,423             | 49.52 | 304,864 |
| 9     | 83,700           | 34.56            | 38,957             | 49.35 | 242,176 |
| 10    | 100,861          | 35.74            | 42,677             | 49.14 | 282,223 |
| 11    | 98,383           | 35.42            | 40,418             | 50.03 | 277,756 |
| 12    | 100,747          | 35.26            | 43,021             | 49.68 | 285,690 |
| 13    | 78,855           | 36.60            | 31,579             | 48.74 | 215,422 |
| 14    | 68,136           | 35.39            | 28,403             | 49.52 | 164,686 |
| 15    | 59,899           | 34.59            | 27,109             | 49.75 | 173,145 |
| 16    | 55,399           | 33.64            | 27,727             | 49.52 | 164,986 |
| 17    | 57,688           | 34.07            | 28,311             | 49.21 | 169,329 |
| 18    | 59,612           | 36.19            | 23,971             | 49.25 | 164,700 |
| 19    | 47,357           | 35.04            | 23,054             | 47.90 | 135,136 |
| 20    | 45,819           | 35.13            | 20,518             | 49.14 | 130,427 |
| 21    | 29,813           | 36.28            | 14,169             | 46.47 | 82,164 |
| 22    | 27,140           | 33.82            | 15,725             | 46.59 | 80,253 |
| X     | 4,394            | 38.03            | 2,771              | 37.99 | 11,555 |
| All   | 2,055,046        | 35.44            | 878,051            | 49.42 | 5,799,120 |

Our comparison is restricted to calls in high confidence regions from GIAB in both the autosomes and pseudoautosomal regions (PARs). On average 77.9% (standard deviation 12.1%) of the bases of each chromosome are in high confidence regions. The comparison was performed using the Nextflow-based (Di Tommaso et al., 2017) workflow, accessible from the link in the data availability section of this manuscript and in Table 4.
Table 4. Links to code and documentation of packages used for analysis.

| Task | Analysis code | Documentation |
|------|---------------|---------------|
| eHive (workflow system) | https://github.com/Ensembl/ensembl-hive | https://ensembl-hive.readthedocs.io/en/version-2.5/ |
| BAM quality control | https://github.com/igsr/igsr_analysis/tree/master/PyHive/BamQC | WGS BAM QC: https://igsr-analysis.readthedocs.io/en/latest/workflows/wgs_bamqc_pipeline.html WES BAM QC: https://igsr-analysis.readthedocs.io/en/latest/workflows/wes_bamqc_pipeline.html |
| Variant discovery | https://github.com/EMBL-EBI-GCA/reseqtrack/tree/master/modules/ReseqTrack/Hive | https://github.com/EMBL-EBI-GCA/reseqtrack/blob/master/docs/variantcalling_pipeline.txt |
| Variant filtering for Freebayes and BCFTools call sets | https://github.com/elowy01/vcf_filtering | https://github.com/elowy01/vcf_filtering/wiki |
| Variant filtering for GATK UG call set | https://github.com/igsr/igsr_analysis/blob/master/PyHive/PipeConfig/FILTER/VCFilterGATK.pm | GATK variant filtering pipeline: https://igsr-analysis.readthedocs.io/en/latest/workflows/gatk_vc_filtering_pipeline.html |
| Variant integration | https://github.com/igsr/igsr_analysis/blob/master/PyHive/PipeConfig/INTEGRATION/VCFIntegrationGATKUG.pm | https://igsr-analysis.readthedocs.io/en/latest/workflows/consensus_callset_pipeline.html |
| Phasing | https://github.com/igsr/igsr_analysis/blob/master/PyHive/PipeConfig/INTEGRATION/PHASING.pm | https://igsr-analysis.readthedocs.io/en/latest/workflows/phasing_pipeline.html |
| Benchmarking with a true call set | https://github.com/igsr/igsr_analysis/blob/master/PyHive/scripts/VCF/QC/BENCHMARKING_TRUESET/vcf_eval.nf | https://igsr-analysis.readthedocs.io/en/latest/workflows/compare_with_true_pipeline.html |

Figure 1. Boxplot showing the distribution of the positive predictive values (PPVs) for single nucleotide variants (SNVs) (1A) and insertions/deletions (INDELs) (1B) for the 118 Gambian in Western Division (GWD) samples, compared between the variants identified in this work and the ones identified in the analysis from (Lowy-Gallego et al., 2019). PPVs are calculated with the following ratio: ‘number of true variants’ / (‘number of true variants’ + ‘number of false variants’).

Switch error rate comparison with the GWD samples. To gauge phasing accuracy, we estimated the switch error rate (SER). Using a comparison with a known true haplotype block, this measures the rate at which the phase is incorrectly predicted. As an example, if the correct haplotype is 000111|111000 and the predicted haplotype is 00000|111111, then we count one switch error between positions 3 and 4. WhatsHap ‘compare’ (version 0.18)
Figure 2. UpSet plot showing the contribution of each of the four supporting call sets to the final single nucleotide variant (SNV) consensus call set. 'GRCh38 P3 GWD' represents the call set for the 118 Gambian in Western Division (GWD) samples extracted from (Lowy-Gallego et al., 2019). 'FreeBayes' is the call set generated using FreeBayes. 'BCFTools' is the call set generated using BCFTools. 'GATK' is the call set generated using GATK. 'Integrated' is the final SNV call set generated after integrating the supporting call sets. Vertical bars show the size of the intersection between the call sets. Horizontal bars show the size of each of the input call sets. We used the filtered supporting call sets to generate this plot.

(Patterson et al., 2015) was used to estimate the SER in our phased variants for the 118 GWD samples, comparing calls from this and our previous work. The SER was calculated for each autosome using the following command for SNVs:

```
whatshap compare --sample $sample \
   --only-snvs $sample.phased.GWD_1KG.snps.chr$i.vcf.gz \
   $sample.phased.GWD_GGVP.snps.chr$i.vcf.gz
```

For INDELs, the following command was used:

```
whatshap compare --sample $sample \
   $sample.phased.GWD_1KG.indels.chr$i.vcf.gz \
   $sample.phased.GWD_GGVP.indels.chr$i.vcf.gz
```

Where $sample represents each of the GWD samples and chr$i is the chromosome being compared.

The distribution of the mean SER obtained across all chromosomes in each of the samples is shown in Figure 4A for SNVs and Figure 4B for INDELs, and a plot with the SER for each chromosome in each of the samples in Figure 4C for SNVs and Figure 4D for INDELs.

We observed higher SER on chromosome X. The mean SER for chromosome X is 3.664% for SNV and 8.844% for INDELs, this contrasts with the mean SER, calculated with only the autosomes, which was 1.615% (SD=0.201) for SNVs and 3.011% (SD=0.552) for INDELs. This difference may be due to the limited size of the regions analysed.
Figure 4. A and B: Boxplots showing the mean switch error rate (SER) calculated for each sample for single nucleotide variants (SNVs) (A) and insertions/deletions (INDELs) (B). Each data point (in red) was calculated by averaging the SER for each chromosome in each individual sample. C and D: SER plots per chromosome for SNVs (C) and INDELs (D). Each data point represents the SER obtained for a particular chromosome from an individual sample.
Figure 5. Genotype data for Gambian Genome Variation Project (GGVP) samples at rs74045212, showing summaries for the populations and individual genotypes (see rs74045212 Ensembl). GGVP populations are represented with three-letter codes. GWF: Gambian in Western Division, The Gambia - Fula, GWJ: Gambian in Western Division, The Gambia - Jola, GWW: Gambian in Western Division, The Gambia - Wolof, GWD: Gambian in Western Division, The Gambia - Mandinka.

Lastly, the average SER across all the samples was 1.704% (SD = 0.184) for SNVs and 3.265% (SD = 0.342) for INDELs. For context, this compares with a SER for SNVs for the single sample NA12878 of 0.71% in our previous work, and 1.54% for the 1000 Genomes Project’s phase three call set on GRCh37 (calculating SER against GIAB gold-standard data). The INDEL SER for NA12878 in those projects was 1.78% in our previous work, and 5.18% for the 1000 Genomes Project on GRCh37.

The pipelines used in this work were implemented using the eHive (Severin et al., 2010) and Nextflow (Di Tommaso et al., 2017) workflow management systems and modules developed in Perl as well as Python, which have been packaged for ease of deployment. All analyses were run in parallel on a high-throughput computing cluster to ensure completion in a reasonable timeframe. Code is publicly available via GitHub as shown in Table 4.

Data availability
Underlying data
EMBL-EBI European Nucleotide Archive: Variant calling on GRCh38 with the Gambian Genome Variation Project (GGVP) samples. Accession number PRJEB36811; https://identifiers.org/ena.embl:PRJEB36811.

EMBL-EBI European Nucleotide Archive: Low coverage sequencing of the Fula from Gambia Western Division GWF. Accession number ERP001420; https://identifiers.org/ena.embl:ERP001420.

EMBL-EBI European Nucleotide Archive: Low coverage sequencing of the Jola from Gambia Western Division GWJ. Accession number ERP001781; https://identifiers.org/ena.embl:ERP001781.

EMBL-EBI European Nucleotide Archive: Low coverage sequencing of the Wolof from Gambia Western Division GWW. Accession number ERP002150; https://identifiers.org/ena.embl:ERP002150.

EMBL-EBI European Nucleotide Archive: Low coverage sequencing of the Mandinka from Gambia Western Division GWD. Accession number ERP002385; https://identifiers.org/ena.embl:ERP002385.

EMBL-EBI European Nucleotide Archive: Whole genome sequencing of (GWD) Gambian in Western Division, The Gambia HapMap population. Accession number SRP001518; https://identifiers.org/ena.embl:SRP001518.
Frequency distributions and genotypes are available in the Ensembl genome browser (Howe et al., 2021) (Figure 5).

EMBL-EBI International Genome Sample Resource: All intermediate and other data https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/gambian_genome_variation_project/

Extended data
Analysis code available from: https://github.com/igsr/igsr_analysis
Archived code at the time of publication: http://doi.org/10.5281/zenodo.5081940 (Lowy et al., 2021)
License: Apache-2.0

Code for the vcf filtering is available from: https://github.com/elowy01/vcf_filtering/
Archived code for vcf filtering at the time of publication: http://doi.org/10.5281/zenodo.5085159
License: Apache-2.0

Code used for the variant calling available from: https://github.com/EMBL-EBI-GCA/reseqtrack
Archived code for variant calling at the time of publication: http://doi.org/10.5281/zenodo.2573969
License: Apache-2.0

Acknowledgements
We wish to thank the Kwiatkowski lab for their support in this work and assistance with provision of data and metadata, especially Kirk Rockett and Gavin Band. We would also like to acknowledge the contribution of Gabriel Aldam in lifting over the array data to GRCh38 and Astrid Gall and Sarah Hunt for helpful comments on this manuscript.

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

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(1): 1084-97.

Byr ska-Bishop M, Evi nus US, Zhao X, et al.: High Coverage Whole Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios. bioRxiv. 2021.

Delaneau O, Marchini J, Zagury JF: A Linear Complexity Phasing Method for Thousands of Genomes. Nat Methods. 2011; 9(2): 179-81.
PubMed Abstract | Publisher Full Text

Di Tommaso P, Chatzou M, Flo den EW, et al.: Nextflow Enables Reproducible Computational Workflows. Nat Biotechnol. 2017; 35(4): 316-19.

Fairley S, Lowy-Gallo ge E, Perry E, et al.: The International Genome Sample Resource (IGSR) Collection of Open Human Genomic Variation Resources. Nucleic Acids Res. 2020; 48(D1): D941–47.
PubMed Abstract | Publisher Full Text | Free Full Text

Howe KL, Achuthan P, Allen J, et al.: Ensembl 2021, Nucleic Acids Res. 2021; 49(D1): D884-91.
PubMed Abstract | Publisher Full Text | Free Full Text

Istreeter, Richardson D, Holly ZB, et al.: EMBL-EBI-GCA/reseqtrack: zenodo. (Version zenodo). Zenodo. 2019. http://www.doi.org/10.5281/zenodo.2573969

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): 839–48.
PubMed Abstract | Publisher Full Text | Free Full Text

Lowy-Gallo ge 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 2; peer review: 2 approved). Wellcome Open Res. 2019; 4(50): 50.
PubMed Abstract | Publisher Full Text | Free Full Text

Lowy E, GabeAldam, ranjti s: igsr/igsr_analysis: v.1.1.4. (Version v1.1.4). Zenodo. 2021. http://www.doi.org/10.5281/zenodo.5081940

Malaria Genomic Epidemiology Network: Insights into Malaria Susceptibility Using Genome-Wide Data on 17,000 Individuals from Africa, Asia and Oceania. Nat Commun. 2019; 10(1): 5782.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Patterson M, Marschall T, Pisanti N, et al.: WhatsApp: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. J Comput Biol. 2015; 22(6): 498-509.
PubMed Abstract | Publisher Full Text

Saha AK, Mou ard M, Kaplan MH, et al.: The Genomic Landscape of Centromeres in Cancers. Sci Rep. 2019; 9(1): 11259.
PubMed Abstract | Publisher Full Text | Free Full Text

Severin J, Beal K, Vilella AJ, et al.: ehive: An Artificial Intelligence
Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 07 December 2021

https://doi.org/10.21956/wellcomeopenres.18773.r47343

© 2021 Piccolo S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Stephen R. Piccolo
Department of Biology, Brigham Young University, Provo, UT, USA

The paper summarizes a new dataset containing DNA variant data for a population of 505 individuals. Some of these individuals had been included in the 1000 Genomes Project, whereas most of them had not due to issues with creating immortalized cell lines for those individuals. This paper describes sequencing and analysis of all of these individuals. Although sequencing coverage is low, the computational pipeline used to process the data seems appropriate, is openly available, and well documented. The data are available as FASTQ files from the European Nucleotide Archive, and the called variants files are openly available. I was able to access all of the files except the pages that were linked at the beginning of the "Data Availability" section. These links all begin with https://identifiers.org. The authors should kindly fix those before I offer my full approval.

The manuscript is well written and thorough. I am happy to see this work published so that other researchers will be able to access this dataset and use it for additional projects.

One additional comment. Toward the beginning of the paper, it says, "The strategy is similar to that described in (Lowy-Gallego et al., 2019) but does differ in some details, mainly in that we have used an automatic machine learning (ML)-based method to filter the BCFTools and Freebayes call sets." It is not quite clear at this location of the manuscript that these details will be provided later. If the authors added a brief statement explaining this, it would be helpful.

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

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

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, genomics, applied machine learning, computational reproducibility

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.