# Supplemental information for DePristo et al., “A framework for variation discovery and genotyping using next-generation DNA sequencing data”

## Supplementary Figures and Tables

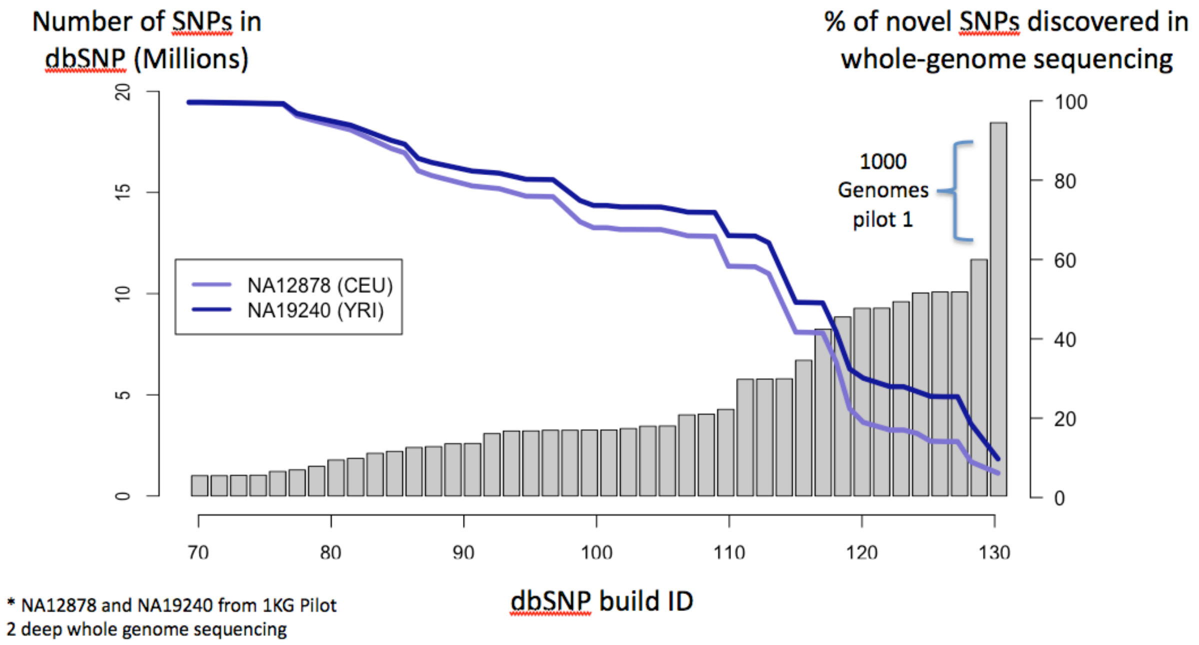

Figure S1: dbSNP rate for NA12878 and NA19240 from dbSNP build id 70 to 129 and then including additionally the CEU and YRI 1000 Genomes low-pass call sets

Table S1: Impact of local realignment in the HiSeq data set at sites in 1000 Genomes for which there is a homozygous indel for NA12878

| Homozygous non-reference indel sites in NA12878[a] | Homozygous indel sites with realigned reads | Spanning reads in raw alignments that do not contain the indel | Total # of realigned reads at homozygous non-reference indel sites[b] |
|---|---|---|---|
| 124,568 | 116,259<br>(90% of all such sites) | 1,448,627<br>(14.7% of all spanning reads) | 1,083,125<br>(74.8% of reads without the indel) |

a) According to 1000 Genomes CEU Trio indel call set

b) 25% of the reads spanning indels do not contain the indel but match perfectly the reference sequence and so no realignment is necessary

Table S2: Number of regions, reads, and sites with significant mismatches affected by local realignment

| Sequencing data set | Regions with realigned reads | Total size of affected regions (Mb) | Reads that were realigned | Sites with significant mismatch removal |
|---|---|---|---|---|
| HiSeq | 947,765 | 21.3 | 6,621,462 | 1,749,840 |
| Low-pass | 2,627,318 | 57.3 | 16,586,650 | 991,532 |
| Exome | 49,170 | 1.1 | 149,449 | 106,182 |

Sites were considered to have a significant number of mismatches if over 15% of the bases (HiSeq and Exome data sets) or at least 2 bases (low-pass data) at the site mismatch the reference.

Table S3: Performance of hard-filtering and variant quality score recalibration

| | Site discovery | | | | | | Comparison to NA12878 variants | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | No. of SNPs | | | | Ti/Tv | | HM3 concordance | | 1KG concordance | |
| Call set | All | Known | Novel | dbSNP% | Known | Novel | NR sensitivity | NRD rate | NR sensitivity | NRD rate |
| HiSeq | | | | | | | | | | |
| Recalibrated, MSA raw calls | 4.18M | 3.45M | 722K | 82.71 | 2.06 | 1.57 | 99.72 | 0.09 | 99.48 | 0.19 |
| Hard filtered | 3.53M | 3.19M | 351K | 90.07 | 2.10 | 1.97 | 99.33 | 0.07 | 98.51 | 0.14 |
| Variant recalibrated | **3.58M** | **3.22M** | **362K** | **89.89** | **2.15** | **2.05** | **99.05** | **0.07** | **97.28** | **0.10** |
| Low-pass* | | | | | | | | | | |
| Recalibrated, MSA raw calls | 13.4M | 6.5M | 6.9M | 48.77 | 2.05 | 1.13 | 83.97 | 20.34 | 80.45 | 22.53 |
| Direct Ti/Tv optimized | 7.42M | 5.05M | 2.37M | 67.98 | 2.10 | 1.82 | 82.79 | 20.23 | 78.14 | 22.17 |
| Variant recalibrated call set | **7.3M** | **5.8M** | **1.5M** | **79.7** | **2.18** | **2.05** | 83.02 | 20.26 | 76.99 | 22.01 |
| Exome capture | | | | | | | | | | |
| Recalibrated, MSA raw calls | 18.5K | 16.8K | 1.7K | 90.77 | 3.20 | 1.61 | 99.07 | 0.08 | 99.13 | 0.11 |
| Hard filtered | 15.9K | 15.1K | 807 | 94.93 | 3.38 | 2.74 | 97.24 | 0.08 | 97.10 | 0.11 |
| Variant recalibrated | **17.2K** | **16.2K** | **1039** | **93.96** | **3.27** | **2.57** | **98.49** | **0.08** | **98.38** | **0.11** |

* HM3 sensitivity and NRD rate only includes NA12878

Table S4: Comparison of NA12878 exome calls to reprocessing with Crossbow

| | GATK | Crossbow | Intersection | Unique to GATK | Unique to Crossbow |
|---|---|---|---|---|---|
| No. known SNPs | 16152 | 16086 | 15194 | 958 | 892 |
| No. novel SNPs | 1039 | 1806 | 799 | 240 | 1007 |
| Known SNPs Ti/Tv | 3.27 | 3.26 | 3.35 | 2.36 | 2.17 |
| Novel SNPs Ti/Tv | 2.57 | 1.94 | 2.77 | 2.04 | 1.50 |
| HM3 NR sensitivity | 98.5% | 95.9% | 95.3% | 3.2% | 0.6% |
| 1000G Trio NR sensitivity | 98.4% | 95.4% | 94.7% | 3.7% | 0.7% |
| Percent of calls not in 1000G trio | 18.3% | 23.9% | 15.5% | 55.8% | 94.8% |
| HiSeq NR sensitivity | 94.1% | 90.9% | 88.8% | 5.3% | 2.1% |
| Percent of calls not in HiSeq | 3.5% | 10.4% | 2.1% | 22.0% | 80.5% |
| Percent synonymous variants | 54.2% | 52.8% | 54.7% | 48.0% | 36.7% |
| Percent missense variants | 45.5% | 46.8% | 45.1% | 51.1% | 61.6% |
| Percent nonsense/read-through | 0.3% | 0.4% | 0.3% | 0.9% | 1.7% |

Only includes calls in coding target regions.

Table S5: dbSNP 129 rates for several CEU and YRI samples using multiple sequencing technologies and SNP calling approaches

| Sample | Population | dbSNP 129 rate | Sequencing technologies | SNP caller | Notes |
|---|---|---|---|---|---|
| NA12878 | CEU | 92% | Solexa, SOLiD and 454 | Samtools + GATK | 1000 Genomes official release |
| NA12891 | CEU | 92% | Solexa | Samtools + GATK | 1000 Genomes official release |
| NA12892 | CEU | 92% | Solexa | Samtools + GATK | 1000 Genomes official release |
| NA20431 | CEU | 90% | CG | CG | Complete genomics [1] |
| NA07022 | CEU | 90% | CG | CG | Complete genomics [1] |

Table S6: Genome-wide and exome target Transition / Transversion (Ti/Tv) ratios expectations from published whole genome call sets against Human Genome build 36*.

| | | | | | **WGS** | | | **Exome** | |
|---|---|---|---|---|---|---|---|---|---|
| **Data set** | **Sequencing tech(s)** | **Year** | **SNP caller(s)** | **N sites** | **Known Ti/Tv** | **Novel Ti/Tv** | **N sites** | **Known Ti/Tv** | **Novel Ti/Tv** |
| **Initial resequencing projects** | | | | | | | | | |
| Venter[1] | ABI | 2007 | Celera assembler | 3.0M | 2.10[2] | 1.53[2] | 15.2K | 3.21[2] | 2.54[2] |
| Watson[3] | 454 | 2008 | Wheeler et al. caller | 2.1M | 2.13[2] | 1.49[2] | 12.2K | 3.38[2] | 1.90[2] |
| **Single sample or trio NGS data sets** | | | | | | | | | |
| Complete Genomics NA19240[1] | CGI | 2009 | CGI | 4.1M | 2.14 | 2.09 | 20.2K | 3.42 | 2.98 |
| 1000 Genomes CEU trio | Solexa, SOLiD and 454 | 2010 | glfTrio, GATK | 3.6M | 2.08 | 2.02 | 17.6K | 3.54 | 2.74 |
| 1000 Genomes YRI trio | Solexa, SOLiD and 454 | 2010 | glfTrio, GATK | 4.5M | 2.09 | 2.07 | 25K | 3.51 | 3.18 |
| **Weighted average** | | | | | **2.10** | **2.07** | **-** | **3.49** | **2.98** |
| **1000 Genomes low-coverage call sets** | | | | | | | | | |
| CEU low-pass | Solexa, SOLiD and 454 | 2010 | QCall, Mach, GATK | 7.7M | 2.10 | 1.90 | 45K | 3.43 | 2.77 |
| YRI low-pass | Solexa, SOLiD and 454 | 2010 | QCall, Mach, GATK | 10.6M | 2.11 | 2.00 | 42K | 3.55 | 2.91 |
| **Weighted average** | | | | | **2.10** | **1.96** | | **3.48** | **2.84** |

1) Obtained from http://huref.jcvi.org/
2) Compared to dbSNP build 126, based on Wheeler et al. 2008
3) Obtained from http://jimwatsonsequence.cshl.edu/

Note that the transition / transversion ratio may depend on the properties of human genome reference build. These expected values should be calibrated for each major human genome reference version.

Table S7: Base miscalling confusion matrices by technology

| Illumina (GA&HiSeq) | | | | |
|---|---|---|---|---|
| | A | C | G | T |
| A | N/A | 57.7% | 17.1% | 25.2% |
| C | 34.9% | N/A | 11.3% | 53.9% |
| G | 31.9% | 5.1% | N/A | 63.0% |
| T | 45.8% | 22.1% | 32.0% | N/A |
| SOLiD | | | | |
| | A | C | G | T |
| A | N/A | 18.7% | 42.5% | 38.7% |
| C | 27.0% | N/A | 18.9% | 54.1% |
| G | 61.0% | 15.7% | N/A | 23.2% |
| T | 40.5% | 34.3% | 25.2% | N/A |
| 454 | | | | |
| | A | C | G | T |
| A | N/A | 23.2% | 42.6% | 34.3% |
| C | 19.7% | N/A | 8.4% | 71.9% |
| G | 71.5% | 6.6% | N/A | 21.9% |
| T | 43.8% | 37.8% | 18.5% | N/A |

# Supplementary Notes

## Data generation

**NA12878 HiSeq data**

We sheared 1-3 ug of genomic DNA to a range of 100-700bp using the Covaris E210 instrument. DNA fragments were end-repaired and phosphorylated, followed by adenylation of 3'ends. Standard paired end adaptors were ligated according to the manufacturer's protocol (Illumina). We performed Qiagen min-elute column based cleanups between all enzymatic steps. Adapter ligated fragments were purified with preparatory gel electrophoresis (4% agarose, 85volts, 3 hours) and two bands were excised (500-520bp and 520-540bp) resulting in two libraries per sample with inserts averaging 380bp and 400bp respectively. DNA was extracted from gel bands using Qiagen min-elute columns. The entire volume of final purified fragments was enriched via PCR with Phusion polymerase for 10 cycles.

Libraries were quantified using a Sybr qPCR protocol with specific probes for the ends of the adapters. The qPCR assay measures the quantity of fragments properly adapter ligated that are appropriate for sequencing. Based on the qPCR quantification, libraries were normalized to 2nM and then denatured using 0.1 N NaOH. Cluster amplification of denatured templates occurred according to manufacturer's protocol (Illumina) using cBot reagent plates and HiSeq Flowcells (Illumina cat# PE-401-1001). Sybr Green dye was added to all flowcell lanes to provide a quality control checkpoint after cluster amplification to ensure optimal cluster densities on the flowcells. Flowcells were paired end sequenced with 101 bp reads on HiSeq2000s, using HiSeq Sequencing-by-Synthesis kits (Illumina cat# PE-401-1001) and analyzed with the Illumina v1.8 pipeline. Standard quality control metrics including error rates, % passing filter reads, and total Gb produced were used to characterize process performance prior to downstream analysis. The final data set includes 16 lanes totaling ~64x coverage of the genome.

**NA12878 whole exome hybridc capture data**

We sheared 1-3 ug of genomic DNA to a range of 100-300bp using the Covaris E210 instrument. DNA fragments were end-repaired and phosphorylated, followed by adenylation of 3'ends. Standard paired end adaptors were ligated according to the manufacturer's protocol (Illumina). We performed Ampure Bead-based cleanups between all enzymatic steps using the Bravo liquid handling platform (Agilent). The entire volume of final library fragments was enriched via PCR with Pfu (Agilent) polymerase for 6 cycles.

Libraries were quantified using an automated picogreen fluorescent assay compared against a standard curve of known samples and normalized to 25ng/ul prior to hybridization.

Hybridization of 500ng of library with 500ng of biotin-linked RNA (Agilent), designed specifically to the desired exome targets, was incubated at 65°C for 72 hours. Capture of resulting DNA-RNA duplexes was performed by the addition of streptavidin M280 beads (Invitrogen). Multiple washes at high stringency removed off-target material and any non-hybridized fragments. Desired fragments were PCR amplified directly off of beads using primers specific to the universal library sequences on the ends of captured fragments. The resulting captured libraries were quantified using picogreen in addition to a Sybr qPCR protocol (KAPA biosystems) with specific probes for the ends of the adapters.

Based on the qPCR quantification, libraries were normalized to 2nM and then denatured using 0.1 N NaOH. Cluster amplification of denatured templates occurred according to manufacturer's protocol (Illumina) using V2 Chemistry and V2 Flowcells (1.4mm channel width). Sybr Green dye was added to all flowcell lanes to provide a quality control checkpoint after cluster amplification to ensure optimal cluster densities on the flowcells. Flowcells were sequenced on Genome Analyzer II's, using V3 Sequencing-by-Synthesis kits and analyzed with the Illumina v1.3.4 pipeline. Standard quality control metrics including error rates, % passing filter reads, and total Gb produced were used to characterize process performance prior to downstream analysis.

### <u>NA12878 and 60 sample CEPH low-pass from 1000 Genomes</u>

For the low-pass analysis, we used the publically available sequencing data from 61 individuals in the pilot phase of the 1000 Genomes Project. All 60 samples from the Caucasian (CEU) population of the low-pass wing were used; these individuals were sequenced to approximately 4x average coverage genome-wide on a variety of sequencing platforms: Illumina/GA, 454, and SOLiD. Additionally, we used the Illumina sequencing data of the CEU daughter (NA12878) from the high-pass trio wing of the 1000 Genomes pilot; as she was sequenced to over 30x coverage genome-wide, we downsampled her data to an average coverage of 4x. The downsampling was achieved by using only twenty-five read groups (lanes) from the high-pass data (ERR001751-ERR001775). The CEU low-pass data set includes 963 lanes of Solexa single- and paired-end reads, 95 lanes of SOLiD reads, and 906 lanes of 454 lanes comprising a total of ~600 Gb of sequence with on average 145x, 81x, and 42x coverage per platform, respectively.

Duplicate removal was performed using samtools rmdupse. An initial version of the GATK quality score recalibration tool was applied by the DCC to the GA and 454 BAMs; subsequent improvements to this tool include separately calibrating the first and second reads of a pair, reference-bias correction for SOLiD reads, as well as storing the original base qualities, which were not retained in the 1000 Genomes released BAMs and so could not be recalibated here. Consequently, local multiple-sequence realignment was performed after quality score recalibration on all 61 individuals simultaneously.

## Base miscalling confusion matrices

In order to account for biased miscalling for all three platforms, we tabulated reference and miscalled bases on chromsome 1 in the original 1000 Genomes CEU sample NA12878 sequenced with Illumina/GA, SOLiD, and 454 reads as part of the trio wing of the pilot project. Only loci containing exactly one non-reference base, at least 20x depth, where all bases had base qualities > Q20 and all reads had mapping qualities > Q30, and at least a Q50 homozygous reference genotype according to the genotyping algorithm caller presented below using the unified miscalling model (where Pr{B true | b miscall} = 1/3) were considered. Such sites exhibit systematic miscalling biases of each instrument, from which we can calculate:

$$\Pr\{B_{true} \mid b_{miscalled}\} = \frac{\text{count}(b_{miscall} \mid B_{true})}{\sum_{X \neq B_{true}} \text{count}(b_{miscall} \mid X)}$$

These base miscalling confusion matrices for Solexa, SOLiD, and 454 are given in Table S3. Because they depend on particularities of the base calling algorithm applied for each technology which do change over time, these confusion matrices should be recalculated periodically.

## Evaluating the quality of detected variation

We obtained HapMap3.2 consensus genotypes from hapmap.org, dbSNP 129 mapped to HG18 from the UCSC genome browser, and 1000 Genomes trio-aware SNP and indels calls from the 1000 Genomes DCC. Note that the 1000 Genomes CEU Trio call set is the intersection of SNP sites called by the GATK SNP caller (presented here, with hard-filtering as defined below) and a trio-aware extension to samtools (http://genome.sph.umich.edu/wiki/GlfTrio) using three sequencing technologies to ~120x total depth. Genotypes for NA12878 and her parents were derived from the trio-aware caller and not the GATK caller.

## Detecting reads from duplicate molecules

One variable and potentially large source of variation miscalls in NGS is non-independent sampling of DNA molecules during sequencing, such as occurs with repeated sampling of molecules that are molecular duplicates of one another. The PCR amplification steps involved in the majority of NGS library construction techniques can introduce significant biases due to preferential amplification of shorter molecules, molecules without extreme GC composition, etc. which will cause the sequencing to be a non-random sampling of the source genome. This is particularly problematic if any single molecule experiences a PCR error early in amplification as this error is propagated and sampled many times during sequencing.

To correct this problem we have developed an algorithm to detect and mark molecules that are probable duplicates of one another. This algorithm is simplified by the assumption that it is unlikely to sample the same exact molecule more than once from the source genome given true random sampling. Given current NGS protocols it is clear that this does in fact occur, but at an acceptably low rate:

De-duplication rate penalties by sequencing design strategy

| Sequencing Application | Average #Molecules in Library | Read Length | Average #Molecules Sampled | Molecules Sampled > 1 times |
|---|---|---|---|---|
| 30X Whole Genome | 5bn | 2x101b | ~450m | 4.4%* |
| 4X Whole Genome | 5bn | 2x101b | ~60m | 0.6% |
| 100X Whole Exome | 500m | 2x76b | ~20m | 2.0% |

* Note that typically multiple independent libraries are created to see to such depth, thereby reducing the penalty of overmarking ‘duplicate’ molecules

Our duplicate marking algorithm relies on sequencing reads having been mapped to the genome to identify reads and read pairs that share the same start positions on the genome and mark these as duplicates. Concretely, the steps taken to identify duplicate molecules are:

1. For each molecule, or cluster, sequenced identify the putative genomic position and strand for the 5’-most (with respect to the read) bases of each read originating from the molecule. The mapping positions reported by the aligner are then adjusted for any soft or hard clipping to determine the most

likely location of the 5'-most base whether or not that base has been mapped to the genome. The 5'-most bases are used as it is expected that these are the bases flanked by the sequencing adapters, which in turn are used as universal amplification sites during library construction.

2. Identify molecules where paired-reads have been performed and both reads have been mapped to the genome, and group these molecules by the genomic position and strand computed in step #1.
3. In groups of size > 1, mark reads from all except one molecule as duplicates.
4. Identify molecules with only a single mapped read such that the mapped read's position and strand are identical to one end of a molecule with mapped paired-end reads and mark these molecules as duplicates also.
5. Group the remaining molecules with only a single mapped read by genomic position and strand and, similar to step #3, mark as duplicates all but one molecule in each group.

Within a group of duplicate molecules a simple heuristic is used to determine which molecule and hence which reads to retain. The base quality scores of each read are summed, ignoring those bases that are below Q15, and the read with the highest sum of quality scores is retained.

This algorithm is implemented in the program called MarkDuplicates in the Picard suite of tools (http://picard.sourceforge.net/). The program reads a SAM or BAM file as input and produces as output a SAM or BAM file with duplicate records retained but flagged.

## Multi-sample SNP discovery and genotyping

The mathematical formulation of the multi-sample genotyping algorithm is given in Box 1. The genotyping algorithm makes a significant assumption by not directly modeling sequencing and alignment errors in the Bayesian framework, namely that any read present at a site actually belongs there. In actuality though, a high enough percentage of reads are misaligned as to affect the accuracy of the calling. To that end, we instituted several filters that must be passed in order for a given base to be used in the genotype likelihoods calculation: its phred-scaled base quality must be at least Q20, the mapping quality of its read must be at least 20, the read and its mate pair (for paired-end reads) must lie on the same chromosome, fewer than 10% of the bases on the read are permitted to mismatch the reference in a 20bp window on either side of the given base, and the read cannot be flagged as failing vendor quality filters or as originating from a duplicate molecule. Any base that passes all of these filters has an extremely high probability of actually originating at the given position.

The discovery algorithm is comprised of two phases: calculating the per-sample genotype likelihoods and then using a heuristic grid search to determine the most likely alternate allele frequency and genotype conformation over all samples. In the first phase, we use the bases at the position in question belonging to a given sample to calculate the likelihoods of the potential diploid genotypes for that sample (Box 1). As a heuristic improvement in the complexity of the problem we make the assumption that any given site is at most bi-allelic, choosing the most likely alternate allele based on the total sum of base qualities for each of the three possible non-reference bases. While tri-allelic sites, though rare, certainly do exist in diploid organisms, this assumption enables us to calculate only three likelihoods per sample (representing the possibility of the sample's being homozygous reference, heterozygous, or homozygous variant) instead of all ten possible diploid genotypes, which vastly improves the running time of the algorithm. We note that while this assumption may affect the genotype assignments at truly tri-allelic sites, it should not affect our ability to discover those sites.

Whereas the initial phase of the algorithm is run per sample, the second stage combines the genotype likelihoods over all samples in order to determine the most likely alternate allele frequency in the cohort. The likelihood for a given set of genotype assignments at a given frequency is simply the product of the genotype likelihoods for each sample given that sample's assigned genotype (Box 1). We then apply a population genetic prior to the allele frequency likelihoods based on θ, the population specific heterozygosity, and choose the most likely allele frequency and associated genotype assignments. The variant quality score for a polymorphic call is given as -10*$\log_{10}$(probability that the site is actually monomorphic).

In theory, for each possible alternate allele frequency, we would need to calculate the likelihood of each possible conformation of genotypes in the cohort under that frequency; however, as the number of possible genotype assignments is exponential in the number of samples, this calculation becomes intractable for larger cohorts. In practice, we have found that an excellent estimate of the most likely genotype assignment for a given allele frequency, *f*, is the most likely assignment for *f-1* with another alternate allele added to one of the samples. By using a best-first search algorithm, calculating the likelihood of *f* is linear in the number of samples, as we can simply iterate over each sample and determine, using the per-sample genotype likelihoods, the most likely recipient of the new allele given the conformation at *f-1*. The fact that this greedy best-first algorithm converges once it hits a local maximum allows us to employ a further significant optimization: we can terminate the search for the most likely allele frequency early whenever the likelihood for a given frequency is significantly lower than the maximum likelihood previously observed for any non-zero frequency.

We use indel calls in a post-process filtering step after SNP genotyping to remove those SNPs that are overly close to indels [2,3]; these SNPs are usually an artifact of the partial discovery of an indel and are almost entirely false positives. We identify a site as potentially containing an indel if the indel occurs in a large enough fraction of the reads at a site (30% for high-pass data and 3% for low-pass) and is consistent within the reads (i.e. there is just a single consensus indel seen in the reads). Any SNPs called within 10bp on either end of these indels are then marked as filtered out of the callset. Note that these heuristic indel calls suffice to avoid SNP artifacts around misaligned indels, regardless of whether these indels result from machine artifacts or are truly segregating in the sample(s). Sites found to overlap the indel mask, and SNPs found in clusters (three or more within a ten-base-pair window) were filtered out of the call set [3].

## Variant Quality Score Recalibration

For the Dirichlet prior distribution over the mixing coefficients $\vec{\pi}$:

$$\Pr\{\vec{\pi}\} = Dir(\vec{\pi} \mid \alpha_0) = C(\alpha_0)\prod_{k=1}^{K}\pi_k^{\alpha_0 - 1}$$

where Dirichlet parameter $\alpha_0$ is chosen to be $1\text{x}10^{-4}$. For the Gaussian-Wishart prior over the mean and precision ($\Lambda^{-1} = \Sigma$) of each Gaussian in the mixture:

$$\Pr\{\vec{\mu},\Lambda\} = \Pr\{\vec{\mu} \mid \Lambda\}\Pr\{\Lambda\} = N(\vec{\mu}_k \mid \vec{m}_0,(\beta_0\Lambda_k)^{-1})W(\Lambda_k \mid W_0,\upsilon_0)$$

we use a shrinkage parameter $\beta_0$ of $1\text{x}10^{-4}$ and $\upsilon_0$ is given by the degrees of freedom which for this model is the number of covariates + 2 = 6.

Only a subset of known variants are used for clustering in order to avoid training on poorly determined annotations for variants with little sequencing data. In particular, only variants satisfying the following criteria are used to train the Gaussian mixture model:

Parameters for variant quality score recalibration

| | HiSeq | Exome | Low-pass |
|---|---|---|---|
| Max. number of Gaussians to learn | 16 | 8 | 6 |
| Min. variant quality score for training | 300 | 2800 | 1000 |
| Max standard deviation from mean annotation value for inclusion | 3.5 | 3.5 | 4.5 |
| Max. percent of reads at a variant to be including in training | 10% | 10% | 10% |

## Standard hard filters

For many projects, including our contributions to all three wings of the 1000 Genomes Project, we used a variety of hard filtering and optimization approaches to select high quality call sets from the raw calls. Here we list, for completeness, the application of our standard hard-filters for deep coverage and the predecessor to the variant recalibator, a Ti/Tv-based standard-bias optimizer [4], applied to the CEPH 61 sample set. Although it is possible to produce a reasonable quality call set using these approaches, the adaptive error modeling used by the variational Bayes recalibrator is able to better identify true positive variation with limited subjective intervention. In Table S9 we contrast the performance of the hard filtering approach to the variant recalibator, and the improved specificity achievable with the recalibrator is clear. As with the variant recalibator, even these hard filters become increasingly selective with additional samples, so that multiple deep read sets analyzed together provide a better trade-off in sensitivity and specificity.

For deep whole genomes, we filter out any SNPs matching the following criteria:

- The SNP cluster and proximity to indels filter as in the main analysis, or
- Greater than 10% of aligned reads at a site have mapping quality 0 (MAPQ0) among at least 40 reads, or
- SB > -0.1, or
- Depth of coverage above 120

For deep whole exomes, we filter out any SNPs matching the following criteria:

- The SNP cluster and proximity to indels filter as in the main analysis, or
- Greater than 10% of aligned reads at a site have mapping quality 0 (MAPQ0) among at least 40 reads, or
- SB > -0.1, or
- Quality over depth (QD) < 5, or
- HRun > 3

## Additional annotation details

For SLOD: for each site, the procedure resulted in an estimate of allele frequencies in the population, an estimate of genotype likelihoods and posteriors for each member of the population, a LOD in favor of a site being variant, and an SB value measuring the strand bias in the non-reference allele. Under the null hypothesis, for a site detected as a variant the true non-reference allele frequency in the forward direction equals the true non-reference allele frequency in the reverse direction. Under the alternative hypothesis, where the site is not a variant but rather error prone, the non-reference allele frequency in the forward direction equals the estimated allele frequency and the non-reference allele frequency in the reverse direction equals to zero or vice-versa. SB is simply the log of the ratio of likelihood densities computed for the best supporting alternative hypothesis versus that of the null hypothesis.

The HaplotypeScore annotation associated with a SNP call at position POS is calculated by first (1) determining the two more prevalent 21 bp haplotypes around POS and then (2) calculating the probability of each read covering POS being sampled from either of these haplotypes:

1. Each read is enqueued into a priority queue with the priority being the sum of base quality scores within the 21 bp window (POS +/- 10 bp on each side). The set of prefect matching reads is set to [2].
2. The read with the greatest sum of quality scores is taken from the queue. If it matches exactly any of the putative haplotypes, it is added to that haplotype's read set. Bases that are present in the new read but not in its haplotype are appended to the haplotype. If no exact match is available, a new haplotype is created with just this read in its read set. This process continues until the priority queue is exhausted and all reads have been placed into exact haplotype match sets.
3. We then construct consensus haplotypes with bases and quality scores for the two haplotypes with the greatest sum of base quality scores across all reads in their read sets.

4. Finally, the HaplotypeScore equation is determined for reads spanning POS against these two most common haplotypes.

The following figure depicts the intuition behind the Haplotype Score annotation:

Two segregating haplotypes

| | |
|---|---|
| Read1 | AAGC**T**CG |
| Read2 | AAGC**A**CGA |
| Read3 | AAGC**T**CGAT |
| Read4 | AAGC**A**CGAT |
| Read5 | AGC**A**CGAT |
| Read6 | GC**T**CGAT |

A/T polymorphic explains the reads well, so has a low haplotype score

Three segregating haplotypes

| | |
|---|---|
| Read1 | AA**G**C**T**CG |
| Read2 | AA**G**C**A**CGA |
| Read3 | AA**C**C**T**CGAT |
| Read4 | AA**C**C**A**CGAT |
| Read5 | A**G**C**A**CGAT |
| Read6 | **C**C**T**CGAT |

Inconsistent variation between A/T and C/G bases likely due to mapping artifacts, so has a high haplotype score

## Likelihood-based genotype refinement with imputation

Beagle [5] was used to refine genotypes in the low-pass 61-sample data set using likelihoods obtained during multi-sample SNP calling with default parameters and no external reference panel. In sites where no likelihoods were available for a sample due to lack of coverage, a model of uniform likelihoods was adopted. Genotypes were updated to those with the greatest posterior probabilities according to Beagle.

Specifically, given a set of variant sites obtained by the methods described in the preceding sections, the final step in variant discovery is the refinement and improvement of the obtained sample genotypes. In order to carry out this sample genotype improvement, the imputation software package Beagle 3.2 was used [5] on the low-pass CEU population of 60 individuals, augmented by NA12878 downsampled to 4x coverage, after such data set was filtered using the Variant Recalibration procedure described above. For this application no reference panel was used, so that Beagle was only used to infer missing genotypes and to provide posterior genotype probabilities. The GATK was used to write the input to Beagle in its required format, using the genotype likelihoods from the SNP caller in Phase 2. In samples for which genotypes were missing (for example, at sites with no coverage), a uniform likelihood model was used (i.e. the likelihoods for a site being homozygous reference, heterozygous, or homozygous variant were set to an equal value of 1/3). After Beagle was run, the GATK was used again to parse the resulting output genotypes and posterior probabilities in order to produce an updated call set.

Two partitions of the data sets have proven useful for analyzing genotype imputation accuracy: a partition either by the non-reference allele frequency (AF) or the sequencing depth at each site. In both cases, NRD rates are computed for each bin. Figures 5(c) and 5(d) show the resulting NRD rates before and after imputation for each partition, as well as the global NRD rate from the whole data set as computed above. As Figure 5(d) illustrates, genotyping accuracy increases significantly at all points, and the increase is especially dramatic in low-depth sites. Even at sites where there was no sequencing coverage, we are able to recover genotypes with 20.9% NRD rate just by using imputation. With just one read, the NRD rate decreases from 58.8% to 7.6%. At the given target depth of 4x, the NRD rate improves from 8.5% to 3.0% with imputation. Importantly, the improvement in genotyping accuracy with 6 or more reads is marginal, which is the reason why imputation was only applied to the low-pass data and not to deep coverage data sets.

## Comparison with Crossbow

To compare our results to existing data processing tools for next-generation DNA sequencing, we applied the Crossbow package [6] to the NA12878 whole exome sequence data. Specifically, the original machine output fastq files were aligned with bowtie [7] (-M 2, paired-end mode), and SNPs called with SoapSNP [8] (using the binomial probability calculation for greater accuracy). We considered only those sites identified as having QUAL score >= 20 and P >= 0.01 for the rank sum test to determine if two allele of a possible HET call have the same sequencing quality. The results of this comparison are summarized in Table S11.

## Data availability

These data sets are available on the Genome Sequencing and Analysis website:

http://www.broadinstitute.org/gsa/wiki/

## References

1. Drmanac, R. *et al.* Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. in *Science* Vol. 327 78-81 (2010).
2. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. in *Nature* Vol. 456 53-9 (2008).
3. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. in *Genome Research* Vol. 18 1851-1858 (2008).
4. The 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. in *Nature* (2010).
5. Browning, B.L. & Yu, Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. in *Am J Hum Genet* Vol. 85 847-61 (2009).
6. Langmead, B., Schatz, M.C., Lin, J., Pop, M. & Salzberg, S.L. Searching for SNPs with cloud computing. in *Genome Biol* Vol. 10 R134 (2009).
7. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. in *Genome Biol* Vol. 10 R25 (2009).
8. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. in *Genome Research* Vol. 19 1124-1132 (2009).