nature genetics

A framework for variation discovery and genotyping using next-generation DNA sequencing data

Mark A DePristo[1], Eric Banks[1], Ryan Poplin[1], Kiran V Garimella[1], Jared R Maguire[1], Christopher Hartl[1], Anthony A Philippakis[1–3], Guillermo del Angel[1], Manuel A Rivas[1,4], Matt Hanna[1], Aaron McKenna[1], Tim J Fennell[1], Andrew M Kernytsky[1], Andrey Y Sivachenko[1], Kristian Cibulskis[1], Stacey B Gabriel[1], David Altshuler[1,3,4] & Mark J Daly[1,3,4]

Recent advances in sequencing technology make it possible to comprehensively catalog genetic variation in population samples, creating a foundation for understanding human disease, ancestry and evolution. The amounts of raw data produced are prodigious, and many computational steps are required to translate this output into high-quality variant calls. We present a unified analytic framework to discover and genotype variation among multiple samples simultaneously that achieves sensitive and specific results across five sequencing technologies and three distinct, canonical experimental designs. Our process includes (i) initial read mapping; (ii) local realignment around indels; (iii) base quality score recalibration; (iv) SNP discovery and genotyping to find all potential variants; and (v) machine learning to separate true segregating variation from machine artifacts common to next-generation sequencing technologies. We here discuss the application of these tools, instantiated in the Genome Analysis Toolkit, to deep whole-genome, whole-exome capture and multi-sample low-pass (~4×) 1000 Genomes Project datasets.

Recent advances in next-generation sequencing (NGS) technology now provide the first cost-effective approach to large-scale resequencing of human samples for medical and population genetics. Projects such as the 1000 Genomes Project[1] (1KG), The Cancer Genome Atlas and numerous large medically focused exome sequencing projects[2] are underway in an attempt to elucidate the full spectrum of human genetic diversity[1] and the complete genetic architecture of human disease. The ability to examine the entire genome in an unbiased way will make possible comprehensive searches for standing variation in common disease and mutations underlying linkages in Mendelian disease[3], as well as spontaneously arising variation for which no gene-mapping shortcuts are available (for example, somatic mutations in cancer[4–6] and *de novo* mutations[7] (Conrad, D.F. *et al*. unpublished data) in autism and schizophrenia).

Many capabilities are required to obtain a complete and accurate record of the variation from NGS from sequencing data. Mapping reads to the reference genome[8–11] is a first critical computational challenge whose cost necessitates that each read be aligned independently, guaranteeing that many reads spanning indels will be misaligned. The per-base quality scores, which convey the probability that the called base in the read is the true sequenced base[12], are quite inaccurate and co-vary with features like sequencing technology, machine cycle and sequence context[13–15]. These misaligned reads and inaccurate quality scores propagate into SNP discovery and genotyping, a general problem that becomes acute in projects with multiple sequencing technologies generated by many centers using rapidly evolving experimental processing pipelines, such as the 1000 Genomes Project.

Given well-mapped, aligned and calibrated reads, resolving even simple SNPs, let alone more complex variation such as multinucleotide substitutions, insertions and deletions, inversions, rearrangements and copy number variation, requires sensitive and specific statistical models[8–11,15–25]. Separating true variation from machine artifacts as a result of the high rate and context-specific nature of sequencing errors is the outstanding challenge in NGS analysis. Previous approaches have relied on filtering SNP calls that have characteristics outside of their normal ranges, such as those occurring at sites with too much coverage[17,19], or by requiring non-reference bases to occur on at least three reads in both synthesis orientations[20]. Though effective, such hard filters are frustratingly difficult to develop, require parameterization for each new dataset and are necessarily either restrictive (high specificity, as in the 1000 Genomes Project) or tolerant (high sensitivity, used in Mendelian disease studies, with concomitantly more false positives). Moreover, all of these challenges must be addressed within the context of a proliferation of sequencing technology platforms and study designs (for example, whole-genome shotgun, exome capture sequencing and multiple samples sequenced at shallow coverage), a point not tackled in previous work.

Here we present a single framework and the associated tools capable of discovering high-quality variation and genotyping individual samples using diverse sequencing machines and experimental designs (**Fig. 1**).

[1] Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [2] Brigham and Women's Hospital, Boston, Massachusetts, USA. [3] Harvard Medical School, Boston, Massachusetts, USA. [4] Center for Human Genetic Research, Massachusetts General Hospital, Richard B. Simches Research Center, Boston, Massachusetts, USA. Correspondence should be addressed to M.A.D. (depristo@broadinstitute.org).

Received 27 August 2010; accepted 17 March 2011; published online 10 April 2011; doi:10.1038/ng.806

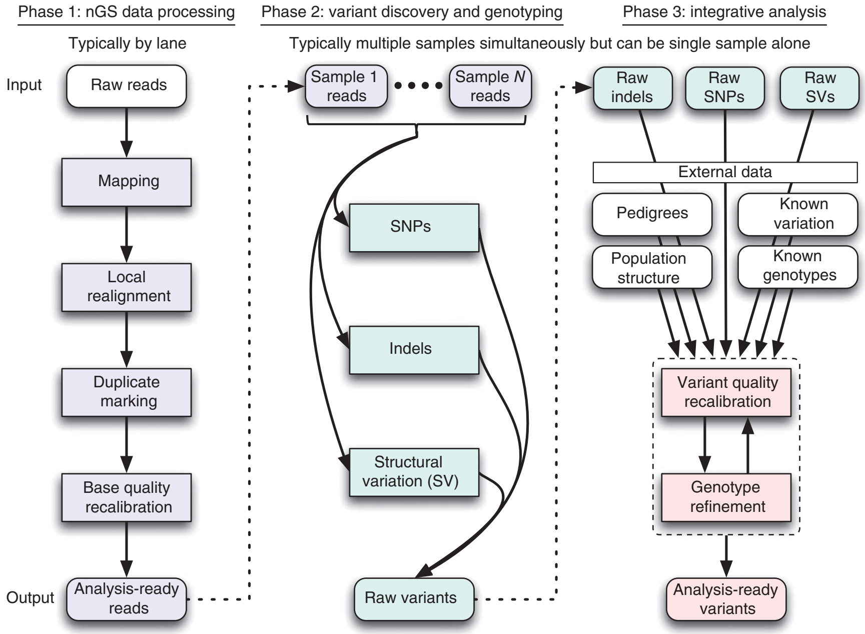


Figure 1 Framework for variation discovery and genotyping from next-generation DNA sequencing. See text for a detailed description.

We present several new methods addressing the challenges listed above in local realignment, base quality recalibration, multi-sample SNP calling and adaptive error modeling, which we apply to three prototypical NGS datasets (**Table 1**). In each dataset, we included CEPH individual NA12878 to show the consistency of results for this individual across all three datasets.

RESULTS

Below we describe a three-part conceptual framework (**Fig. 1**).

- Phase 1: raw read data with platform-dependent biases were transformed into a single, generic representation with well-calibrated base error estimates, mapped to their correct genomic origin and aligned consistently with respect to one another. Mapping algorithms placed reads with an initial alignment on the reference genome, either generated in, or converted to, the technology-independent SAM reference file format[24]. Next, molecular duplicates were eliminated (**Supplementary Note**), initial alignments were refined by local realignment and then an empirically accurate per-base error model was determined.
- Phase 2: the analysis-ready SAM/BAM files were analyzed to discover all sites with statistical evidence for an alternate allele present among the samples including SNPs, short indels and copy number variations (CNVs). CNV discovery and genotyping methods, though part of this conceptual framework, are described elsewhere[25].
- Phase 3: technical covariates, known sites of variation, genotypes for individuals, linkage disequilibrium (LD), and family and population structure were integrated with the raw variant calls from phase 2 to separate true polymorphic sites from machine artifacts, and at these sites, high-quality genotypes were determined for all samples.

All components after initial mapping and duplicate marking were instantiated in the Genome Analysis Toolkit (GATK)[26].

Applying the analysis pipeline to HiSeq

Of the 2.83 billion non-N bases in the autosomal regions and chromosome X of the human reference genome, 2.72 billion bases (~96%) had sufficient coverage to call variants in the 101-bp paired-ended HiSeq data (**Table 1**). Even though the HiSeq reads were aligned with the gap-enabled BWA[10], more than 15% of the reads that span known homozygous indels in NA12878 were misaligned (**Supplementary Table 1**). Realignment corrected 6.6 million of 2.4 billion total reads in 950,000 regions covering 21 Mb in the HiSeq data, eliminating 1.8 million loci with substantial accumulation of mismatching bases (**Supplementary Table 2**). The initial data-processing steps (phase 1) eliminated ~300,000 SNP calls, which is more than one fifth of the raw new calls, with quality metrics consistent with more than 90% of these SNPs being false positives (**Table 2**).

The initial 4.2 million confidently called non-reference sites included 99.7% and 99.5% of the HapMap3 and 1KG Trio sites, respectively, genotyped as non-reference in NA12878; at these variant sites, the sequencing and genotyping calls were concordant 99.9% of the time (**Table 2**). Variant quality score recalibration of these initial calls identified a tranche of SNPs with estimated false discovery rate (FDR) of <1%, containing 3.2 million known variants and 362,000 new variants, a 90% dbSNP rate, and transition/transversion (Ti/Tv) ratios of 2.15 and 2.05, respectively, consistent with our genome-wide expectations (**Online Methods**). Although the variant recalibrator removed ~595,000 total variants with a Ti/Tv ratio of ~1.2, it retained 99% of the HapMap3 and 97.3% of the 1KG Trio non-reference sites. The discordant sites have 100 times higher genotype discrepancy rates, suggesting that the sites themselves may be problematic. Almost all of the variants in the 1% tranche are already present in the even higher stringency 0.1% FDR tranche, and analysis of the 10% FDR tranche suggests that some more variants could be obtained, but at the cost of many more false positives.

Table 1 Next-generation DNA sequencing datasets analyzed

	HiSeq	Exome	Low-pass
Samples	NA12878	NA12878	NA12878 + 60 unrelated CEPH individuals
Sequencing technologies	Whole genome shotgun; Illumina HiSequation (2000)[17]	Agilent exome hybrid capture[31,32]; Illumina GenomeAnalyzer[17]	Whole genome shotgun; Illumina GenomeAnalyzer[17]; Life/SOLiD[33]; Roche/454 (ref. 19)
Coverage per sample	~60×	~150×; 93% of bases at >20× coverage	~4×
Read architecture	101 bp paired end	76/101 bp paired end	25, 36, 51, 76, ~250 (454) bp single and paired ends
Targeted area	2.85 Gb of autosomes and chr. X	28 Mb	2.85 Gb of autosomes and chr. X
Data set source	New, generated for this article	New, generated for this article	1000 Genomes Project
Aligner(s)	BWA[10]	MAQ[9]	MAQ[10]; Corona Lite; SSAHA[12]

Chr., chromosome.

Table 2 Raw to recalibrated, imputed SNP calls HiSeq, Exome and 61 sample low-pass datasets

	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										
Raw reads, all calls	4.43M	3.49M	941K	78.77	2.05	1.29	99.74	0.10	99.57	0.20
Unique to raw read calls	263K	37K	226K	13.95	1.37	0.70	0.02	37.97	0.09	12.64
Unique to +recal/+MSA calls	9.8K	1.8K	8.0K	18.08	1.38	1.39	0.00	18.18	0.00	9.93
+recal/+MSA, all calls	4.18M	3.45M	722K	82.71	2.06	1.57	99.72	0.09	99.48	0.19
Filtered by variant recalibration	595K	235K	360K	39.44	1.19	1.21	0.67	3.00	2.2	4.31
Final call set	**3.58M**	**3.22M**	**362K**	**89.89**	**2.15**	**2.05**	**99.05**	**0.07**	**97.28**	**0.10**
Low pass										
Raw reads, all calls	13.4M	6.5M	6.9M	48.77	2.05	1.13	83.97	20.34	80.45	22.53
Unique to raw read calls	670K	32K	638K	4.74	1.19	0.67	0.01	49.21	0.02	52.57
Unique to +recal/+MSA calls	45K	2.5K	42K	5.62	0.94	0.68	0.00	N/A	0.00	38.89
+recal/+MSA, all calls	12.8M	6.5M	6.3M	50.92	2.06	1.18	83.97	20.33	80.43	22.52
Filtered by variant recalibration	5.5M	706K	4.8M	12.84	1.31	1.01	0.95	26.54	3.44	32.91
Variant recalibrated call set	**7.3M**	**5.8M**	**1.5M**	**79.7**	**2.18**	**2.05**	**Itemized below**			
Sample variant calls for NA12878 only										
Variant recalibrated NGS reads only	2.44M	2.30M	140K	94.28	2.15	2.06	83.02	20.26	76.99	22.01
Recalibrated with Beagle imputation	3.20M	3.01M	191K	94.03	2.18	2.09	96.72	3.32	91.21	3.35
Exome capture										
Raw reads, all calls	18.9K	16.8K	2.1K	88.83	3.20	1.16	99.10	0.09	99.12	0.12
Unique to raw read calls	483	39	444	8.07	2.55	0.31	0.04	25.00	0.03	33.33
Unique to +recal/+MSA calls	81	40	41	49.38	3.44	1.73	0.01	0.00	0.04	16.67
+recal/+MSA, all calls	18.5K	16.8K	1.7K	90.77	3.20	1.61	99.07	0.08	99.13	0.11
Filtered by variant recalibration	1,274	609	665	47.8	1.85	0.84	0.59	N/A	0.76	N/A
Final call set	**17.2K**	**16.2K**	**1,039**	**93.96**	**3.27**	**2.57**	**98.49**	**0.08**	**98.38**	**0.11**

Part one of each section summarizes the impact of local realignment and base quality recalibration by comparing SNP calls on reads with raw quality scores and alignments to those made on the realigned, recalibrated reads. M, million; K, thousand.

Applying the analysis pipeline to 28-Mb exome capture

The raw data processing tools here eliminated ~450 new call sites from the raw call set, representing more than 20% of all the new calls, with a Ti/Tv of 0.30—fully consistent with all being false positives—and adding several sites present in HapMap3 and the 1KG Trio. The raw whole-exome data-call set, at ~150× coverage (**Table 1**), includes >99% of both the HapMap3 and 1KG Trio non-reference sites within the 28-Mb exome target region, with >99.8% genotype concordance at these sites. As with the HiSeq data, even with recalibration and local realignment, the Ti/Tv ratio of the new sites in the initial SNP calls indicates that more than 50% of these calls are false positives. Variant quality score recalibration, using only ~5,400 SNPs for training, identified a high-quality subset of calls that captured >98% of the HapMap3 and 1KG Trio sites in the target regions. The value of the tranches was more pronounced in the whole exome (**Fig. 4d**), where 900 of the 1,039 new calls come from tranches with FDRs under 1%, despite needing to reach into the 10% FDR tranche to include most true positive SNPs.

The HiSeq whole genome shotgun (WGS) and exome capture datasets differed drastically in their sequencing protocols (WGS versus hybrid capture), the sequencing machines (HiSeq versus Genome Analyzer) and the initial alignment tools (BWA[10] versus MAQ[9]). Nevertheless, the exome call set is remarkably consistent the subset of calls from HiSeq that overlap the target regions of the hybrid capture protocol. Ninety-four percent of the HiSeq calls were also called in the final exome set sliced at 10% FDR (data not shown), and at these sites, the non-reference discrepancy rate was extremely low (<0.4%). Mapping differences between the aligners used for HiSeq (BWA) and exome (MAQ) datasets accounted for vast the majority of these discordant calls, with the remainder of the differences being because of limited coverage in the exome and only a small minority of sites being because of differential SNP calling or variant quality score recalibration. Overall, despite the technical differences in the capture and sequencing protocols of the HiSeq and exome datasets, the data processing pipeline presented here uncovered a remarkably consistent set of SNPs in exomes with excellent genotyping accuracy.

Applying the analysis pipeline to low-pass (4×) sequencing

Multi-sample low-pass resequencing poses a major challenge for variant discovery and genotyping because there is so little evidence at any particular locus in the genome for any given sample (**Table 1**). Consequently, it is in precisely this situation, where there is little signal from true SNPs, that our data processing tools are most valuable, as can be seen from the progression of call sets in **Table 2**. Local realignment and base quality recalibration eliminated ~650,000 false-positive SNPs among 13 million sites, 4 times more sites than in the HiSeq dataset, with an aggregate Ti/Tv of 0.7. The initial low-pass CEU set includes over 13 million called sites among all individuals, of which nearly 7 million are new. NA12878 herself has 2.9 million variants, of which 430,000 are new. The 4× average coverage limits the sensitivity and concordance of this call set, with only 84% and 80% of HapMap3 and 1KG Trio sites, respectively, assigned a non-reference genotype in the NA12878 sample, both with a ~20% non-reference discrepancy (NRD) rate.

The variant quality recalibrator identified from the 13 million potential variants ~6 million known and 1.5 million new sites in tranches with 0.1% to 10% FDR. **Figure 5a** highlights several key features of the data: the allele frequency distribution of these calls closely matched the population genetics expectation, and the vast majority of HapMap3 and 1000 Genomes Project official CEU call sites were recovered, with the proportion nearing 100% for more common variant sites (**Fig. 5a**). Although we selected a 0.1% FDR

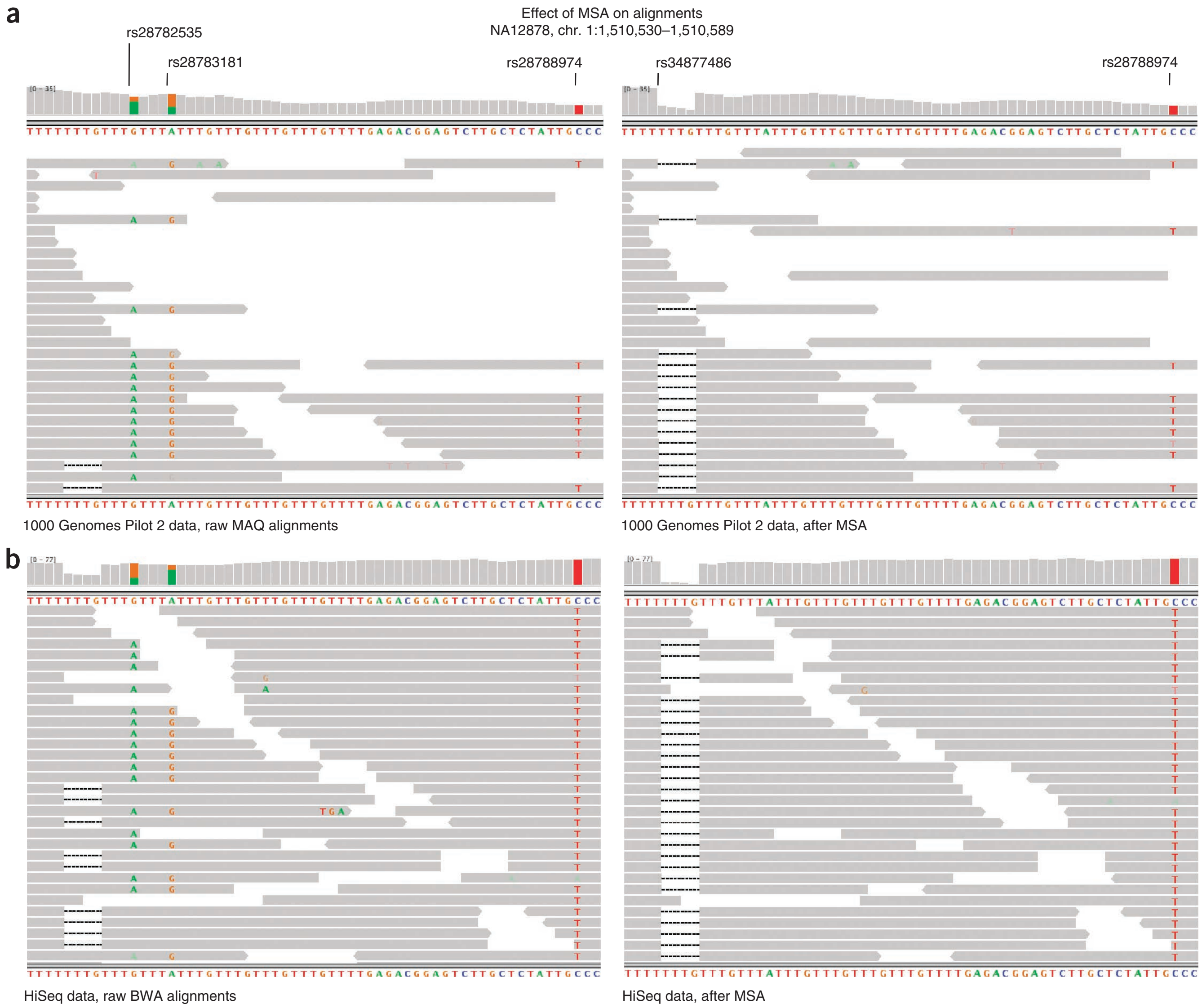

a
Effect of MSA on alignments
NA12878, chr. 1:1,510,530–1,510,589
rs28782535
rs28783181
rs28788974
rs34877486
rs28788974
[0 - 35]
TTTTTTTGTTTGTTTATTTGTTTGTTTGTTTGTTTTGAGACGGAGTCTTGCTCTATTGCCC
1000 Genomes Pilot 2 data, raw MAQ alignments
1000 Genomes Pilot 2 data, after MSA
b
[0 - 77]
HiSeq data, raw BWA alignments
HiSeq data, after MSA

Figure 2 Integrative genomics viewer (IGV) visualization of alignments in region chr. 1: 1,510,530–1,510,589 from the Trio NA12878 Illumina reads from the 1000 Genomes Project (**a**) and NA12878 HiSeq reads before (left) and after (right) multiple sequence realignment (**b**). Reads are depicted as arrows oriented by increasing machine cycle; highlighted bases indicate mismatches to the reference: green, A; orange, G; red, T dashes, deleted bases a coverage histogram per base is shown above the reads. Both the 4-bp indel (rs34877486) and the C/T polymorphism (rs2878874) are present in dbSNP, as are the artifactual A/G polymorphisms (rs28782535 and rs28783181) resulting from the mis-modeled indel, indicating that these sites are common misalignment errors.

tranche for analysis here, which contains the bulk of HapMap3, 1KG Trio and HiSeq sites, there are another ~700,000 true sites that can be found in the 1% and 10% FDR tranches, albeit among many more false positives. This highest-quality tranche includes nearly all variants observed more than five times in the samples and 1.4 million new variants, with the SNPs in the tranches at 1% and 10% FDR generally occupying the lower alternate allele frequency range (**Fig. 5b**). The overall picture is clear: calling multiple samples simultaneously, even with only a handful of reads spanning a SNP for any given sample, enables one to detect the vast majority of common variant sites present in the cohort with a high degree of sensitivity.

Although the bulk properties of the 61-sample call set were good, we expected the low-pass 4× design to limit variation discovery and genotyping in each sample relative to deep resequencing. In the 61-sample call set, we discovered ~80% of the non-reference sites in NA12878 according to the HapMap3, 1KG Trio and HiSeq call sets (**Table 2**). The ~20% of the missed variant sites from these three datasets had little to no coverage in the NA12878 sample in the low-pass data and, therefore, could not be assigned a genotype using only the NGS data, a general limitation of the low-pass sequencing strategy (**Table 2** and **Fig. 5c**,**d**). The multi-sample discovery design, however, affords us the opportunity to apply imputation to refine and recover genotypes at sites with little or no sequencing data. Applying genotype-likelihood–based imputation with Beagle[27] to the 61-sample call set recovered an additional 15–20% of the non-reference sites in NA12878 that had insufficient coverage in the sequencing data (**Table 2**) as well as vastly improving genotyping accuracy (**Fig. 5c**,**d**).

We further characterized the quality of our low-pass call set as a function of the number of samples included during the discovery

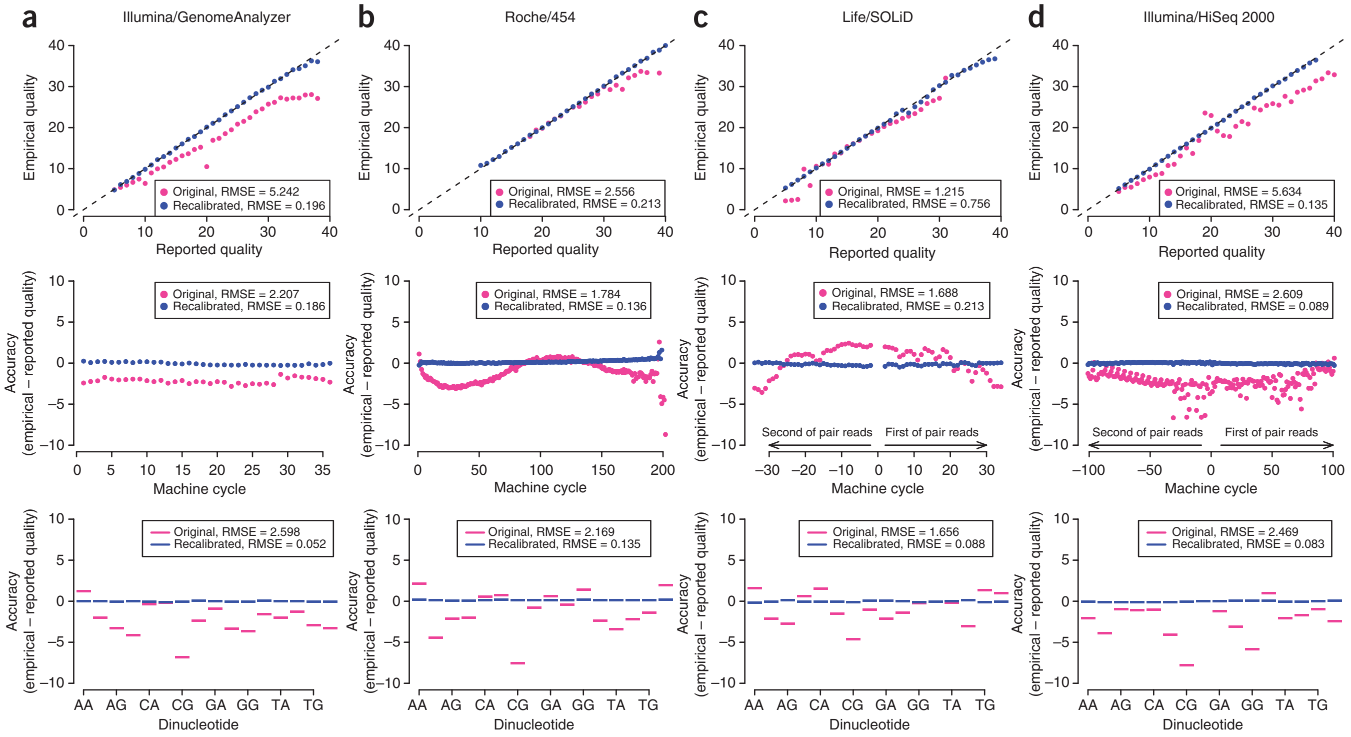


Figure 3 Raw (pink) and recalibrated (blue) base quality scores for NGS paired-end read sets of NA12878 of Illumina/GA (**a**), Roche/454 (**b**) and Life/SOLiD (**c**) lanes from the 1000 Genomes Project and Illumina/HiSeq (**d**). For each technology, the top panel shows reported base quality scores compared to the empirical estimates (Online Methods); the middle panel shows the difference between the average reported and empirical quality score for each machine cycle, with positive and negative cycle values given for the first and second read in the pair, respectively; and the bottom panel shows the difference between reported and empirical quality scores for each of the 16 genomic dinucleotide contexts. For example, the AG context occurs at all sites in a read where G is the current nucleotide and A is the preceding one in the read. Root-mean-square errors (RMSE) are given for the pre- and post-recalibration curves.

process in addition to NA12878 herself. Increasing the number of samples in the cohort rapidly improved both the sensitivity and specificity of the call set. As evidence mounts with more samples that a particular site is polymorphic, our confidence in the call increases and the site is more likely to be called (**Fig. 6a**). Distinguishing true positive variants from sequencing and data processing artifacts is more difficult with few samples and, consequently, low aggregated coverage; adding more reads allows the error covariates to identify sites as errors using the variant recalibrator (**Fig. 6b**,**c**).

The combination of multi-sample SNP calling, variant quality recalibration using error covariates and imputation allows one to achieve

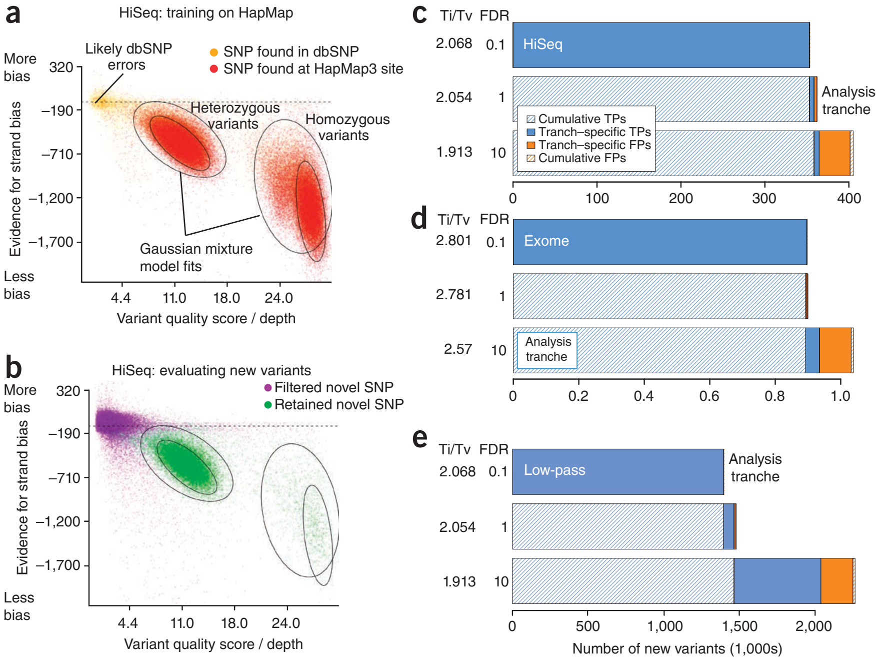
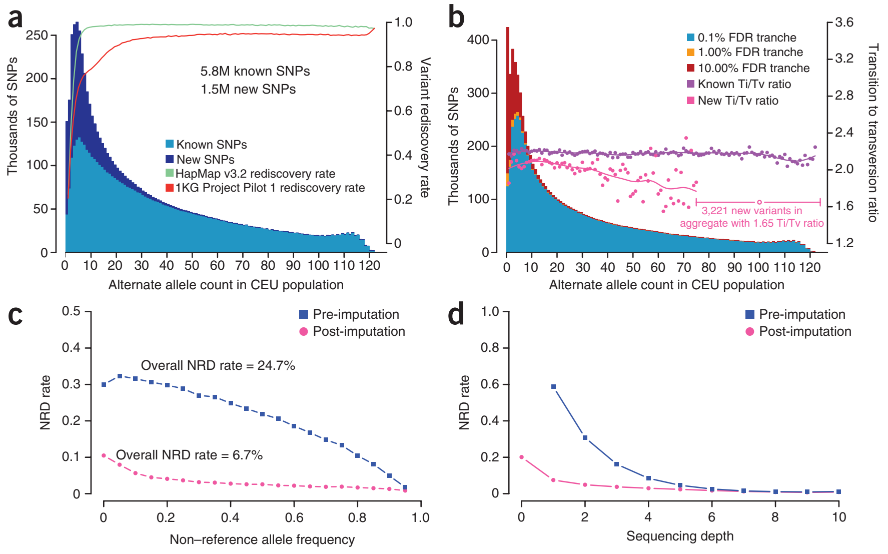


Figure 4 Results of variant quality recalibration on HiSeq, exome and low-pass data sets. (**a**) Relationship in the HiSeq call set between strand bias and quality by depth for genomic locations in HapMap3 (red) and dbSNP (orange) used for training the variant quality score recalibrator (left), (**b**) and the same annotations applied to differentiate likely true positive (green) from false positive (purple) new SNPs. (**c**–**e**) Quality tranches in the recalibrated HiSeq (**c**), exome (**d**) and low-pass CEU (**e**) calls beginning with (top) the highest quality but smallest call set with an estimated false positive rate among new SNP calls of <1/1000 to a more comprehensive call set (bottom) that includes effectively all true positives in the raw call set along with more false positive calls for a cumulative false positive rate of 10%. Each successive call set contains within it the previous tranche's true- and false-positive calls (shaded bars) as well as tranche-specific calls of both classes (solid bars). The tranche selected for further analyses here is indicated.

Figure 5 Variation discovered among 60 individuals from the CEPH population from the 1000 Genomes Project pilot phase plus low-pass NA12878. (**a**) Discovered SNPs by non-reference allele count in the 61 CEPH cohort, colored by known (light blue) and new (dark blue) variation, along with non-reference sensitivity to CEU HapMap3 and 1000 Genomes Project low-pass variants. (**b**) Quality and certainty of discovered SNPs by non-reference allele count. The histogram depicts the certainty of called variation broken out into 0.1%, 1% and 10% new FDR tranches. The Ti/Tv ratio is shown for known and new variation for each allele count, aggregating the new calls with allele count >74 because of their limited numbers. (**c**,**d**) Genotyping accuracy for NA12878 from reads alone (blue squares) and following genotype-likelihood based imputation (pink circles) called in the 61 sample call set as assessed by the NRD rate to HiSeq genotypes as a function of allele count (**c**) and sequencing depth (**d**).



a high-quality call set, both in aggregate and per sample, with very little data. The aggregated 61-sample set at 4× coverage includes only four times as much sequencing data as the HiSeq data, yet we discovered 3.2 million polymorphic sites in NA12878, which includes 97%, 91% and 87% of the variants in the HapMap3, 1000 Genomes Project Trio and HiSeq call sets, respectively, while also finding ~5 million additional variants among the 60 other samples.

Hard filtering versus variant quality score recalibration

Supplementary Table 3 lists the quality of call sets derived using our previous filtering approaches on all three datasets relative to the adaptive recalibrator described here. In all cases, the adaptive approach outperformed the manually optimized hard filtering previously developed for this calling system for the 1000 Genomes Project pilot data. This highlights two important points: first, that a principled integration of all covariates (which may have a complex correlation structure) should and does outperform single manually defined thresholds on covariates independently, with the added benefit of not requiring human intervention; and second, that an accurate ranking of discovered putative variants by

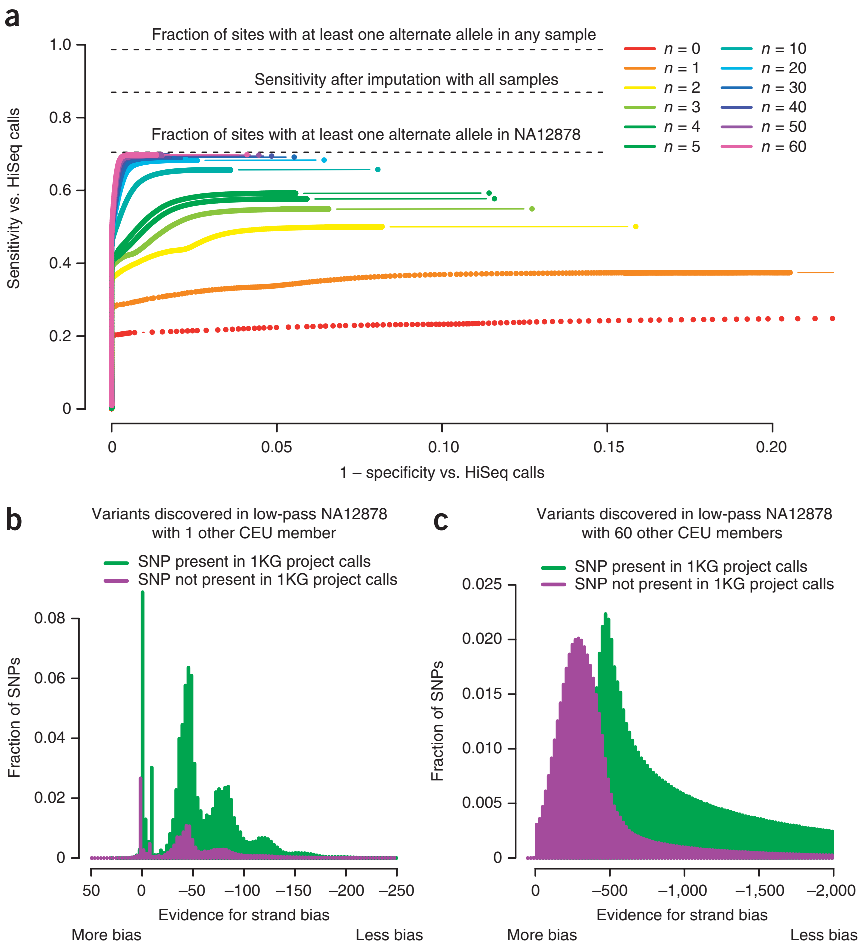


Figure 6 Sensitivity and specificity of multi-sample discovery of variation in NA12878 with increasing cohort size for low-pass NA12878 read sets processed with *N* additional CEPH samples. (**a**) Receiver operating characteristic (ROC) curves for SNP calls relating specificity and sensitivity to discover non-reference sites from the NA12878 HiSeq call set. The maximum callable sensitivity, 66%, is the percent of sites from the HiSeq call set where at least one read carries the alternate allele in the low-pass data for NA12878; it reflects both differences in the sequencing technologies (36–76-bp GAII for the low-pass NA12878 sample compared to 101-bp HiSeq) as well as the vagaries of sampling at 4× coverage. Because most of these missed sites are common and are consequently called in the other samples, imputation recovers ~50% of these sites. (**b**,**c**) Increasing power to identify strand-biased, likely false positive SNP calls with additional samples. Histograms of the strand bias annotation at raw variant calls discovered in the low-pass CEU data using NA12878 at 4× combined with one other CEU individual (**b**) and with 60 other individuals (**c**) stratified into sites present (green) and not (purple) in the 1000 Genomes Project CEU trio.

the probability that each represents a true site permits the definition of tranches for specificity or sensitivity (**Fig. 4c–e**) as appropriate to the needs of the specific project. Although the most permissive tranche includes almost all sites that have any chance of being true polymorphisms—critical for projects looking for single large-effect mutations—the vast majority of true polymorphisms are present in the highest quality tranche of data (data not shown).

Comparison of this calling pipeline to Crossbow

To calibrate the additional value of the tools described here, we contrasted our results with SNPs called on our raw NA12878 exome data using Crossbow[28], a package combining bowtie, a gapless read mapping tool based on the Burrows-Wheeler transformation[29], and SoapSNP for SNP detection[15]. We chose to perform this analysis on the exome data because its wide range of read depths and complex error modes make SNP calling a challenge, especially given the small number of new variants (~1,000 per sample) expected in this 28-Mb target. In **Supplementary Table 4**, the high-level results of the GATK and Crossbow calling pipelines are compared and contrasted. Key metrics such as the number of new SNP calls, their Ti/Tv ratio, the number of calls not seen in either the 1000 Genomes Project trio or the HiSeq data and the high nonsense and read-through rates indicate that the Crossbow call set has lower specificity than the GATK pipeline. This was true even after we applied an aggressive P value threshold ($P < 0.01$) for the base quality rank sum test[15] to filter false-positive variants, which reduced the sensitivity of the HM3, 1000 Genomes Project and the HiSeq call sets by >3%. The intersection set between GATK and Crossbow is more specific but less sensitive than the calls unique to each pipeline (**Table 1**), a clear sign that despite the advances presented here, a lot of work remains to be done in perfecting calling in datasets like single sample exome capture. Although the value of the data processing and error modeling presented here is also clear, applying local realignment and base quality score recalibration (using publicly available, easy-to-use modules in the GATK) are likely to improve the results of the Crossbow pipeline.

DISCUSSION

The inaccuracy and covariation patterns differ strikingly between sequencing technologies (**Fig. 3**), which, if uncorrected, can propagate into downstream analyses. Accurately recalibrated base quality scores eliminate these sequencer-specific biases (**Fig. 3**) and enable integration of data generated from multiple systems. Although developed for early NGS datasets like those from the 1000 Genomes Project pilot, the impact of recalibration is still substantial even for data emerging today on newer sequencers like the HiSequation (2000). Together with local realignment, these two data processing methods eliminated millions of mostly false positive variants while preserving nearly all true variable sites, such as those in HapMap3 and 1KG Trio (**Table 2**). In single sample datasets, such as HiSeq and exome, without realignment and recalibration, these false variants account for more than a fifth of all of the new calls.

Even with very deep coverage, the naïve Bayesian model for SNP calling results in an initial call set with a surprisingly large number of false-positive calls. Although we expected 3.3 million known and 330,000 new non-reference sites in a single European sample sequenced genome wide, the initial HiSeq call set contains 3.5 million known and 800,000 new calls. The excessive number of variable sites, and the low Ti/Tv ratio in particular among the new calls, implies that ~600,000 of these variants are likely errors resulting from stochastic and systemic sequencing and alignment errors. The same calculations suggest that a similar fraction of the initial exome calls are likely false positives, and more than 80% of the initial new low-pass SNP calls are likely errors. The adaptive error modeling developed here enabled us to identify these false-positive variants based on their dissimilarity to known variants, despite error rates of 50–80% among the new variants.

In each step of the pipeline, the improvements derive from the correction of systematic errors made in base calling or read mapping. By characterizing the specific NGS machine error processes and capturing our certainty, or lack thereof, that a putative variant is truly present in the sample or population, we delivered not a single concrete call set but a continuum from confident to less reliable variant calls for use as appropriate to the specific needs of downstream analysis. Mendelian disease projects can select a more sensitive set of calls with a higher error rate to avoid missing that single, high-impact variant, whereas community resource projects like the 1000 Genomes Project can place a high premium on specificity.

The division between SNP discovery and preliminary genotyping and genotype refinement (columns 2 and 3 of **Fig. 1**) avoids embedding in the discovery phase assumptions about population structure, sample relationships and the LD relationships between variants. Consequently, our calling approach applies equally well to population samples in Hardy-Weinberg equilibrium like mother-father-child trios or interbreeding families suffering from Mendelian disorders. Critically, our framework produces highly sensitive and specific variation calls without the use of LD and so can be applied in situations where LD information is unavailable or weak (many organisms) or would confound analytic goals such as studying LD patterns themselves or comparing Neanderthals and modern humans[30]. Where appropriate, however, imputation can be applied to great value, as we demonstrated in the 61-sample CEU low-pass call set.

The analysis results presented here clearly indicate that even with our best current approaches we are still far from obtaining a complete and accurate picture of genetic variation of all types in even a single sample. Even with the HiSeq 10-bp paired-end reads, nearly 4% (~100 Mb) of the potentially callable genome is considered poorly mapped (**Supplementary Note**), and analysis of variants within these regions requires care. Nearly two thirds of the differences between the HiSeq and exome call sets can be attributed to different read mappings between BWA and MAQ.

The challenge of obtaining accurate variant calls from NGS data is substantial. We have developed an analysis framework for NGS data that achieves consistent and accurate results from a wide array of experimental design options including diverse sequencing machinery and distinct sequencing approaches. We have introduced here an integrated approach to data processing and variation discovery from NGS data that is designed to meet these specifications. Using data generated both at the Broad Institute and throughout the 1000 Genomes Project, we have shown that the introduction of improved calibration of base quality scores, local realignment to accommodate indels, the simultaneous evaluation of multiple samples from a population, and finally, an assessment of the likelihood that an identified variable site is a true biological DNA variant greatly improves the sensitivity and specificity of variant discovery from NGS data. The impending arrival of yet more NGS technologies makes even more important modular, extensible frameworks like ours that produce high-quality variant and genotype calls despite distinct error modes of multiple technologies for many experimental designs.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

Many thanks to our colleagues in Medical and Population Genetics and Cancer Informatics and the 1000 Genomes Project who encouraged and supported us during the development of the Genome Analysis Toolkit and associated tools. This work was supported by grants from the National Human Genome Research Institute, including the Large Scale Sequencing and Analysis of Genomes grant (54 HG003067) and the Joint SNP and CNV calling in 1000 Genomes sequence data grant (U01 HG005208). We would also like to thank our excellent anonymous reviewers for their thoughtful comments.

AUTHOR CONTRIBUTIONS

M.A.D., E.B., R.P., K.V.G., J.R.M., C.H., A.A.P., G.d.A., M.A.R., T.J.F., A.Y.S. and K.C. conceived of, implemented and performed analytic approaches. M.A.D., E.B., R.P., K.V.G., G.d.A., A.M.K. and M.J.D. wrote the manuscript. M.A.D., M.H. and A.M. developed Picard and GATK infrastructure underlying the tools implemented here. M.A.D., S.B.G., D.A. and M.J.D. lead the team.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. The 1000 Genomes Project Consortium. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
2. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
3. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2009).
4. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477 (2010).
5. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2009).
6. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
7. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
8. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
9. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
10. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
11. Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
12. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
13. Brockman, W. *et al.* Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.* **18**, 763–770 (2008).
14. Li, M., Nordborg, M. & Li, L.M. Adjust quality scores from alignment and improve sequencing accuracy. *Nucleic Acids Res.* **32**, 5183–5191 (2004).
15. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132 (2009).
16. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
17. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
18. Koboldt, D., Chen, K., Wylie, T. & Larson, D. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).
19. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
20. Mokry, M. *et al.* Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. *Nucleic Acids Res.* **38**, e116 (2010).
21. Shen, Y. *et al.* A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res.* **20**, 273–280 (2010).
22. Hoberman, R. *et al.* A probabilistic approach for SNP discovery in high-throughput human resequencing data. *Genome Res.* **19**, 1542–1552 (2009).
23. Malhis, N. & Jones, S. High quality SNP calling using Illumina data at shallow coverage. *Bioinformatics* **26**, 1029 (2010).
24. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
25. Handsaker, R.E., Korn, J.M., Nemesh, J. & McCarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).
26. McKenna, A.H. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
27. Browning, B.L. & Yu, Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am. J. Hum. Genet.* **85**, 847–861 (2009).
28. Langmead, B., Schatz, M.C., Lin, J., Pop, M. & Salzberg, S.L. Searching for SNPs with cloud computing. *Genome Biol.* **10**, R134 (2009).
29. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
30. Green, R.E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
31. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
32. Ng, S., Turner, E., Robertson, P. & Flygare, S. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
33. Mckernan, K.J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).

ONLINE METHODS

Evaluating the quality of SNP calls. *Number of SNP calls and allele frequency.* The number of calls and frequency for multi-sample calling should follow relatively closely the neutral expectation for N individuals for small N:

$$\text{Number of polymorphic sites} \approx L \times \theta \sum_{i=1}^{2N} 1/i$$

where L is the number of confidently called bases and θ is the population-specific heterozygosity, genome wide of ~0.8 × 10^{-3} for CEPH individuals (H. Li, unpublished data). A surplus of variants, especially heterozygous variants for single samples or lower-frequency variants for populations, is a strong indicator of false positives.

dbSNP rate. Most variants are already catalogued in the dbSNP database of human variation. For a single European sample, ~90% of their true variants will appear in dbSNP build 129 (**Supplementary Table 5**), which will reach ~99% following the completion of the 1000 Genomes Project (**Supplementary Fig. 1**). For population-level SNP calls, the aggregate dbSNP rate for the call set decreases as more rare variants are found, which are less frequently found in dbSNP. Nevertheless, the per sample dbSNP rate should remain consistent across individuals. Note that presence in dbSNP is not an absolute confirmation that a variant is true (for example, see **Fig. 2** and **Fig. 4**), but because dbSNP build 129 contains 11.6 million SNP entries (only 0.4% of all genomic positions), relative differences between call sets with high dbSNP rates can be reasonably interpreted as quality differences.

Non-reference sensitivity and non-reference discrepancy (NRD) rate. For single samples, comparison with non-reference genotype calls from microarray chips, such as HapMap3 (~1.3–1.5 million sites), provides a good initial assessment of variant discovery sensitivity. With sufficient coverage, >99% of non-reference sites can generally be discovered. The NRD rate reports the percent of discordant genotype calls at commonly called non-reference sites on the chip and should reach <1% with sufficient coverage. Mathematical definitions of these terms are:

$$\text{NR sensitivity } (E, C) = \frac{|Enr \cap Cnr|}{|Cnr|}$$

$$\text{NRD rate} (E, C) = \frac{|\{i \in Enr \cup Cnr : Ei \neq Ci\}|}{|Enr \cup Cnr|}$$

$$X_i = \text{Number of non-reference alleles for genotype call } i \text{ in call set } X$$

$$X_{nr} = \{i \in X : X_i > 0\}$$

$$E = \text{Call set to be evaluated}$$

$$C = \text{Call set to be compared to}$$

Transition/transversion ratio (Ti/Tv). The Ti/Tv ratio is a critical metric for assessing the specificity of new SNP calls. Inter-species comparisons[34] and previous sequencing projects (**Supplementary Table 6**) agree on a Ti/Tv ratio of ~2.0–2.1 for genome-wide datasets and 3.0–3.3 for exonic variation[35]. The expected values for the Ti/Tv for known and new variants genome wide are 2.10 and 2.07, respectively, and in the exome target are 3.5 and 3.0, respectively. Currently the lower Ti/Tv ratio at new sites than at known sites is because of a combination of residual false positives lowering the Ti/Tv, a relative deficit of transitions due to sequencing context bias, as well as an apparently higher transition ratio at lower frequency variation. These uncertainties should limit the interpretation of minor differences in Ti/Tv ratios (<0.05), especially across sequencing technologies and datasets.

The Ti/Tv ratio for randomly assigned 'variation', such as results from systematic sequencing errors, alignment artifacts and data processing failures will be ~0.5, as there are two transversion mutations for each transition. Given an expected Ti/Tv ratio, as above, and an observed Ti/Tv ratio from a call set, an estimate of the fraction of false positive variants in the call set can be obtained by:

$$FDR_{\text{test}} = \frac{TiTv_{\text{observed}} - 0.5}{TiTv_{\text{expected}} - 0.5}$$

which should be bounded above by 100% (because of Ti/Tv ratios below 0.5) and a minimum false-pisitive rate (here assumed to be 0.1%) when the observed Ti/Tv exceeds the expected value.

Local multiple sequence realignment. We developed a local realignment algorithm that provides a consistent alignment among all reads spanning an indel. The algorithm begins by first identifying regions for realignment where (i) at least one read contains an indel, (ii) there exists a cluster of mismatching bases or (iii) an already known indel segregates at the site (for example, from dbSNP). At each region, haplotypes are constructed from the reference sequence by incorporating any known indels at the site, indels in reads spanning the site or from Smith-Waterman[36] alignment of all reads that do not perfectly match the reference sequence. For each haplotype H_i, reads are aligned without gaps to H_i and scored according to:

$$L(R_j \mid H_i) = \prod_k L(R_{j,k} \mid H_{j,i})$$

$$L(R_{j,k} \mid H_{j,i}) = \begin{cases} 1 - \in j,k \approx 1 & R_{j,k} = H_{j,i} \\ \in j,k & R_{j,k} = H_{j,i} \end{cases}$$

$$L(H_i) = \prod_j L(R_j \mid H_i)$$

where R_j is the jth read, k is the offset in the gapless alignment of R_j and H_i and $\varepsilon_{j,k}$ is the error rate corresponding to the declared quality score for the k^{th} base of read R_j. The haplotype H_i that maximizes $L(H_i)$ is selected as the best alternative haplotype. Next, all reads are realigned against just the best haplotype H_i and the reference (H_0), and each read R_j is assigned to H_i or H_0 depending on whichever maximizes $L(R_j|H)$. The reads are realigned if the log odds ratio of the two-haplotype model is better than the single reference haplotype by at least five log units:

$$\frac{L(H_0, H_i)}{L(H_0)} = \frac{\prod_j \max\left[L(R_j \mid H_i), L(R_j \mid H_0)\right]}{\prod_j L(R_j \mid H_0)}$$

This discretization reflects a tradeoff between accuracy and efficient calculation of the full statistical quantities. Note that this algorithm operates on all reads across all individuals simultaneously, which ensures consistency in the inferred haplotypes among all individuals, a critical property for reliable indel calling and contrastive analyses such as somatic SNP and indel calling. The realigned reads are written to a SAM/BAM file for further analysis. The reads around a homozygous deletion, before and after local realignment, for Genome Analyzer reads from the 1000 Genomes Project and HiSeq, are shown in **Figure 2**.

Base quality score recalibration. We developed a base quality recalibration algorithm that provides empirically accurate base quality scores for each base in every read while also correcting for error covariates like machine cycle and dinucleotide context, as well as supporting platform-specific error covariates like color-space mismatches for SOLiD and flow-cycles for 454 (refs. 13–15,37,38). For each lane, the algorithm first tabulates empirical mismatches to the reference at all loci not known to vary in the population (dbSNP build 129), categorizing the bases by their reported quality score (R), their machine cycle in the read (C) and their dinucleotide context (D). For each category we estimate the empirical quality score:

$$\text{mismatch}\,(R, C, D) = \sum_{r \in R} \sum_{c \in C} \sum_{d \in D} \sum_{br,c,d} br,c,d \neq b\text{ref}$$

$$\text{bases}\,(R, C, D) = \sum_{r \in R} \sum_{c \in C} \sum_{d \in D} |\{b, r, c, d\}|$$

$$\text{Qempirical}(R, C, D) = (\text{mismatch}(R, C, D) + 1) / (\text{bases}(R, C, D) + 1)$$

These covariates are then broken into linearly separable error estimates and the recalibrated quality score Q_{recal} is calculated as:

$$\mathrm{recal}(r,c,d) = Qr + \Delta Q(r) + \Delta\Delta Q(r,c) + \Delta\Delta Q(r,d)$$

$$\Delta Q = Q_{\mathrm{empirical}}(R,C,D) - \left(\sum_{\rho} \varepsilon_r \times N_r\right) / \,\mathrm{bases}(R,C,D)$$

$$\Delta Q(r) = Q_{\mathrm{empirical}}(r,C,D) - Q_r - \Delta Q$$

$$\Delta Q(r,c) = Q_{\mathrm{empirical}}(r,c,D) - (\Delta Q_r + \Delta Q(r))$$

$$\Delta Q(r,d) = Q_{\mathrm{empirical}}(r,C,d) - (\Delta Q_r + \Delta Q(r))$$

where each ΔQ and $\Delta\Delta Q$ are the residual differences between empirical mismatch rates and that implied by the reported quality score for all observations conditioning only on Q_r or on both the covariate and Q_r; Q_r is the base's reported quality score and ε_r is its expected error rate; $b_{r,c,d}$ is a base with specific covariate values, and r, c, d and R, C, D are the sets of all values of reported quality scores, machine cycles and dinucleotide contexts, respectively. The quality score and covariate distributions for four datasets before and after quality score recalibration are shown in **Figure 3**.

Multi-sample SNP calling. We apply a Bayesian algorithm for variant discovery and genotyping that simultaneously estimates the probability that two alleles A, the reference allele, and B, the alternative allele, are segregating in a sample of N individuals and the likelihoods for each of the AA, AB and BB genotypes for each of individual. Given D_i aligned bases at a specific genomic position for individual i, we estimate the genotype likelihoods GT_i of observing the D_i bases for each of AA, AB and BB genotypes according to the following equation:

$$\Pr\{D_i \mid GT_i\} = \prod_j \Pr\{D_{i,j} \mid GT_i\}$$

$$\Pr\{D_i \mid GT_i = \mathrm{AB}\} = \left(\Pr\{D_{i,j} \mid \mathrm{A}\} + \Pr\{D_{i,j} \mid \mathrm{B}\}\right)/2$$

$$\Pr\{D_{i,j} \mid B\} = \left\{\begin{matrix} 1 - \varepsilon_{i,j} \\ \varepsilon_{i,j} \cdot \Pr\{B \text{ is true} \mid D_{i,j} \text{ is miscalled}\} \end{matrix}\right\} D_{i,j} = B, \text{otherwise.}$$

where $\Pr\{D_{i,j} \mid GT_i\}$ is the probability of observing base $D_{i,j}$ under the hypothesized genotype GT_i; $\Pr\{D_{i,j} \mid \mathrm{B}\}$ and $\Pr\{D_{i,j} \mid \mathrm{A}\}$ are the probability of observing base $D_{i,j}$ given that the true base is B or A, respectively; $\varepsilon_{i,j}$ is the probability of a base miscall given the quality score of base $D_{i,j}$; and Pr{B is true | $D_{i,j}$ is miscalled} is the probability of B_{true} being the true chromosomal base given that b is a miscall (**Supplementary Table 7**). As these are raw likelihoods, no prior probabilities are applied.

Let us define $q_i = \{0,1,2\}$ as the number of alternate B alleles carried by individual i, so that $q = \sum_i^N q_i$ is the number of chromosomes carrying the B allele among all individuals. We estimate the probability that $q = X$ as:

$$\Pr\{q = X \mid D\} = \frac{\Pr\{q = X\}\Pr\{D \mid q = X\}}{\sum_Y \Pr\{D \mid q = Y\}}$$

$$\Pr\{q = X\} = \left\{\begin{matrix} \theta / X \\ 1 - \theta \sum_{i=1}^{2N} 1/i \end{matrix}\right\} X > 0 \text{ otherwise.}$$

$$\Pr\{D \mid q = X\} = \sum_{GT \in \Gamma} \prod_i^N \Pr\{D_i \mid GT_i\}$$

$$\Gamma = \left\{GT \text{ where} \sum_i q_i = X\right\}$$

where Γ is the set of all genotype assignments for the N individuals that contain exactly $q = X$ B alleles, $\Pr\{q = X\}$ is the infinite-sites neutral expectation to observe X alternative alleles in $2N$ chromosomes with heterozygosity of θ, and GT_i and D_i are the ith individual's genotype and NGS reads, respectively. The sum over Γ involves potentially evaluating 3^N combinations but can be approximated by a heuristic algorithm like expectation-maximization through the introduction of a Hardy-Weinberg equilibrium assumption, using a greedy combinatorial search algorithm (**Supplementary Note**) or using an exact summation (H. Li, unpublished data). This algorithm emits the probability of a variant segregating at the site at some frequency:

$$\mathrm{QUAL} = -10 \cdot \log_{10}\left[\Pr\{q = 0 \mid D\}\right]$$

represented conventionally by the Phred-scaled confidence, as well as the genotype assignments at the value that maximizes Pr{q | D}. Only sites with QUAL > Q50 for deep coverage or Q10 for shallow coverage, respectively, are considered here as potentially variable sites.

Variant quality score recalibration. Given a set of putative variants along with SNP error covariate annontations, variant quality score recalibration employs a variational Bayes Gaussian mixture model (GMM)[39] to estimate the probability that each variant is a true polymorphism in the samples rather than a sequencer, alignment or data processing artifact. The set of variants $\{v_i\}$ are treated as an n-dimensional point cloud, each variant v_i positioned by its covariate annotation vector, $\vec{v}$. A mixture of Gaussians is fit to the set of likely true variants, here approximated by the variants already present in HapMap3 (**Fig. 4a**). Following training, this mixture model is used to estimate the probability of each variant call being true (**Fig. 4b**), capturing the intuition that variants with similar characteristics as previously known variants are likely to be real, whereas those with unusual characteristics are more likely to be machine or data processing artifacts.

Mathematically, we write the probability of a variant's vector of covariate values as the linear superposition of Gaussians:

$$\Pr\{v_i \mid GMM\} = \sum_{k=1}^{K} \pi k N\left(\bar{v}_i \mid \bar{\mu} k, \Sigma_k\right)$$

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

$$\Pr\{\vec{\mu}, \Lambda\} = N\left(\vec{\mu} \mid \bar{m}_0, (\beta_0 \Lambda_k)^{-1}\right) W(\Lambda_k \mid W_0, v_0)$$

where K is the number of Gaussians in the mixture (GMM), and the last two equations are standard conjugate prior distributions over the parameters $\vec{\pi}$, $\vec{\mu}$ and Σ.

We then use an analog of the expectation-maximization algorithm[39] to learn the optimal parameters for the clusters using only variant calls at sites present in HapMap3. By restricting training to known polymorphic sites, the resulting GMM captures the distribution of covariate parameters for true SNPs. Consequently, we estimate the likelihood of each putative variant v_i being true under the learned GMM as:

$$L(v_i \mid GMM) = \Pr\{v_i\} \Pr\{\bar{v}_i \mid GMM\}$$

$$L(v_i \mid GMM) = (1 - FP_{\mathrm{singleton}})^{AC} \Pr\{\text{novelty of } v_i\} \sum_{k=1}^{K} \pi k N\left(\bar{v}_i \mid \bar{\mu} k, \Sigma_k\right)$$

$$\Pr\{\text{novelty of } v_i\} = \left\{\begin{matrix} 97\% \ v_i \text{ is in HapMap3,} \\ 37\% \ \text{otherwise.} \end{matrix}\right.$$

where $\Pr\{v_i\}$ is the prior expectation that the putative variant v_i is true, $\bar{v}_i$ is the vector of covariate values for v_i, $FP_{\mathrm{singleton}}$ is the false positive rate for singletons (50% here), and AC is the number of chromosomes estimated to carry the variant among all called samples. The prior probability of $\Pr\{v_i\}$ depends on whether it is present in HapMap3 and its frequency in the samples being called, given an estimate of the false positive rate for singletons. This model can be easily extended to include more training data, more prior information and/or more error covariates.

For convenience of presentation and analysis, we partition the raw SNP calls into tranches based on the Ti/Tv ratio of their new variants. For each desired new false discovery rate target (FDR_i), tranche$_i$ is defined as:

$$\text{tranche}_i = \left\{ SNP_j \text{ where } L(SNP_j \mid GMM) > T_i \right\}$$

$$T_i = \text{smallest } X \text{ where } titv\left(\left\{ SNP_j \text{ is novel } \Lambda\ L(SNP_j \mid GMM) > X \right\}\right) > TiTv_i$$

$$TiTv_i = FDR_i * (TiTv_{\text{expected}} - 0.5) + 0.5$$

The first tranche is exceedingly specific but less sensitive, and each subsequent tranche in turn introduces additional true positive calls along with a growing number of false positive calls. More specificity in the learned GMM translates into better-separated tranches, where all true variants have high likelihoods and appear in the lowest FDR tranches and all false ones have low likelihoods and are excluded. Downstream applications can select in a principled way more specific or more sensitive call sets or incorporate directly the recalibrated quality scores to avoid entirely the need to analyze only a fixed subset of calls but rather weight individual variant calls by their probability of being real.

34. Ebersberger, I., Metzler, D., Schwarz, C. & Pääbo, S. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70**, 1490–1497 (2002).
35. Freudenberg-Hua, Y. *et al.* Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the European population. *Genome Res.* **13**, 2271–2276 (2003).
36. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* (Cambridge University Press, Cambridge, UK, 1998).
37. Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).
38. HUGO Consortium. *et al.* Mapping human genetic diversity in Asia. *Science* **326**, 1541–1545 (2009).
39. Bishop, C. *Pattern Recognition and Machine Learning* (Springer, New York, New York, USA, 2006).