Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls

Justin M Zook[1], Brad Chapman[2], Jason Wang[3], David Mittelman[3,4], Oliver Hofmann[2], Winston Hide[2] & Marc Salit[1]

Clinical adoption of human genome sequencing requires methods that output genotypes with known accuracy at millions or billions of positions across a genome. Because of substantial discordance among calls made by existing sequencing methods and algorithms, there is a need for a highly accurate set of genotypes across a genome that can be used as a benchmark. Here we present methods to make high-confidence, single-nucleotide polymorphism (SNP), indel and homozygous reference genotype calls for NA12878, the pilot genome for the Genome in a Bottle Consortium. We minimize bias toward any method by integrating and arbitrating between 14 data sets from five sequencing technologies, seven read mappers and three variant callers. We identify regions for which no confident genotype call could be made, and classify them into different categories based on reasons for uncertainty. Our genotype calls are publicly available on the Genome Comparison and Analytic Testing website to enable real-time benchmarking of any method.

As whole human genome and targeted sequencing start to offer the real potential to inform clinical decisions[1–4], it is becoming critical to assess the accuracy of variant calls and understand biases and sources of error in sequencing and bioinformatics methods. Recent publications have demonstrated hundreds of thousands of differences between variant calls from different whole human genome sequencing methods or different bioinformatics methods[5–11]. To understand these differences, we describe a high-confidence set of genome-wide genotype calls that can be used as a benchmark. We minimize biases toward any sequencing platform or data set by comparing and integrating 11 whole human genome and three exome data sets from five sequencing platforms for HapMap/1000 Genomes CEU female NA12878, which is a prospective reference material (RM) from the National Institute of Standards and Technology (NIST). The recent approval of the first next-generation sequencing instrument by the US Food and Drug Administration highlighted the utility of this candidate NIST reference material in approving the assay for clinical use[12].

NIST, with the Genome in a Bottle Consortium, is developing well-characterized whole-genome reference materials, which will be available to research, commercial and clinical laboratories for sequencing and assessing variant-call accuracy and understanding biases. The creation of whole-genome reference materials requires a best estimate of what is in each tube of DNA reference material, describing potential biases and estimating the confidence of the reported characteristics. To develop these data, we are developing methods to arbitrate between results from multiple sequencing and bioinformatics methods. The resulting arbitrated integrated genotypes can then be used as a benchmark to assess rates of false positives (o r calling a variant at a homozygous reference site), false negatives (or calling homozygous reference at a variant site) and other genotype calling errors (e.g., calling homozygous variant at a heterozygous site).

Current methods for assessing sequencing performance are limited. False-positive rates are typically estimated by confirming a subset of variant calls with an orthogonal technology, which can be effective except in genome contexts that are also difficult for the orthogonal technology[13]. Genome-wide, false-negative rates are much more difficult to estimate because the number of true negatives in the genome is overwhelmingly large (i.e., most bases match the reference assembly). Typically, false-negative rates are estimated using microarray data from the same sample, but microarray sites are not randomly selected, as they only have genotype content with known common SNPs in regions of the genome accessible to the technology.

Therefore, we propose the use of well-characterized whole-genome reference materials to estimate both false-negative and false-positive rates of any sequencing method, as opposed to using one orthogonal method that may have correlated biases in genotyping and a more biased selection of sites. When characterizing the reference material itself, both a low false-negative rate (i.e., calling a high proportion of true variant genotypes, or high sensitivity) and a low false-positive rate (i.e., a high proportion of the called variant genotypes are correct, or high specificity) are important (**Supplementary Table 1**).

Low false-positive and false-negative rates cannot be reliably obtained solely by filtering out variants with low-quality scores because biases in the sequencing and bioinformatics methods are not all included in the variant quality scores. Therefore, several variant

[1]Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA. [2]Bioinformatics Core, Department of Biostatistics, Harvard School of Public Health, Cambridge, Massachusetts, USA. [3]Arpeggi, Inc., Austin, Texas, USA. [4]Virginia Bioinformatics Institute and Department of Biological Sciences, Blacksburg, Virginia, USA. Correspondence should be addressed to J.M.Z. (jzook@nist.gov).

Received 14 December 2013; accepted 27 January 2014; published online 16 February 2014; doi:10.1038/nbt.2835

callers use a variety of characteristics (or annotations) of variants to filter likely false-positive calls from a data set. However, some true variants are usually filtered out when filtering false positives.

Although large data sets such as whole-genome sequencing data sets are challenging to analyze owing to their size, machine learning algorithms can take advantage of the large number of sites across a whole human genome to learn the optimal way to combine annotations and filter genotype errors. For example, the Genome Analysis ToolKit (GATK) includes a Variant Quality Score Recalibration (VQSR) module that uses annotations related to parameters such as strand bias, mapping quality, allele balance, and position inside the read to filter potential errors[14,15]. GATK trains a Gaussian mixture model using suspected true-positive variants to find the optimal way to filter false positives while retaining a specified sensitivity to likely true-positive variants.

Currently, GATK and other variant callers do not effectively use multiple data sets from the same sample to refine genotype calls and find likely false positives and false negatives. Several methods have recently been proposed by the 1000 Genomes Project Consortium to integrate multiple variant call sets, but these methods have not been used to arbitrate between data sets from different sequencing methods on the same genome[16]. Therefore, we have extended GATK's methods to integrate information from multiple publicly available data sets of the same sample and use VQSR to identify possible biased calls and arbitrate between discordant data sets (**Fig. 1** and **Supplementary Fig. 1**). In this way, we can integrate data sets from multiple sequencing technologies and minimize bias toward any particular sequencing technology, similar to co-training in semi-supervised machine learning[17]. Each technology yields erroneous or ambiguous data at particular locations of the genome as a result of systematic sequencing errors (e.g., the GGT motif in Illumina), mapping and alignment errors, or biases in coverage[7]. Although algorithms can filter some systematic sequencing errors, they cannot perfectly distinguish between false positives and true positives. Using multiple platforms that have different systematic sequencing errors can help distinguish between true positives and false positives in any particular sequencing technology. In addition, regions with very low coverage in one platform can have sufficient coverage in a different platform to make a genotype call. The resulting methods, reference materials and integrated genotype calls are a public resource at http://www.genomeinabottle.org/ for anyone to characterize performance of any genome sequencing method.

RESULTS

Arbitrating between data sets that disagree

To develop our high-confidence genotype calls, we used 11 whole-genome and 3 exome data sets from five sequencing platforms and seven mappers (**Table 1**). For the hundreds of thousands of possible SNP sites in the whole genome that differ between sequencing platforms and variant callers, we developed methods to identify biases and arbitrate between differing data sets (Online Methods, **Fig. 1** and **Supplementary Figs. 1**–**3**). Briefly, we first selected all sites that had even a small amount of evidence for a SNP or indel in any data set. Then, we used previously existing and new annotations in the GATK VQSR Gaussian mixture model to identify sites in each data set with characteristics of biases, including systematic sequencing errors (SSEs)[18,19], local alignment difficulties, mapping difficulties or abnormal allele balance. For each site where genotypes in different data sets disagreed, we sequentially filtered data sets with characteristics of systematic sequencing errors, alignment uncertainty and atypical allele balance.

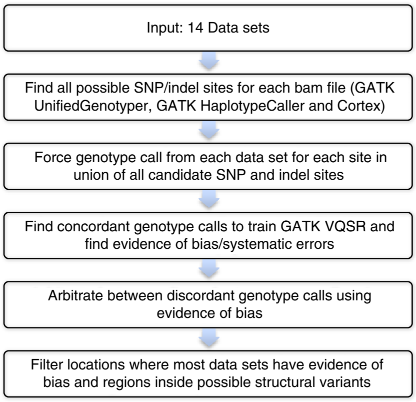
Input: 14 Data sets
Find all possible SNP/indel sites for each bam file (GATK UnifiedGenotyper, GATK HaplotypeCaller and Cortex)
Force genotype call from each data set for each site in union of all candidate SNP and indel sites
Find concordant genotype calls to train GATK VQSR and find evidence of bias/systematic errors
Arbitrate between discordant genotype calls using evidence of bias
Filter locations where most data sets have evidence of bias and regions inside possible structural variants

Figure 1 Integration process used to develop high-confidence genotypes. Description of the process used to develop high-confidence genotype calls by arbitrating differences between multiple data sets from different sequencing platforms, and define regions of the genome that could be confidently genotyped. A more detailed description of our methods and examples of arbitration are in **Supplementary Figures 1–3**.

After arbitration, we filtered as uncertain (i) sites where we could not determine the reason for discordant genotypes, (ii) regions with simple repeats not completely covered by reads from any data set, (iii) regions with known tandem duplications not in the GRCh37 genome reference assembly (which was partly developed from NA12878 fosmid clones and is available at http://humanparalogy.gs.washington.edu/), (iv) regions paralogous to the 1000 Genomes Project "decoy reference," which contains sequences not in the GRCh37 reference genome, (v) regions in the RepeatSeq database, and (vi) regions inside structural variants for NA12878 that have been submitted to dbVar. We provide a file in BED format that specifies the regions in which we can confidently call the genotype. Before filtering structural variants, we are able to call confidently 87.6% of the non-N bases in chromosomes 1–22 and X, including 2,484,884,293 homozygous reference sites, 3,137,725 SNPs and 201,629 indels (**Supplementary Tables 2** and **3**). Excluding structural variants conservatively excludes an additional 10% of the genome, with 2,195,078,292 homozygous reference sites, 2,741,014 SNPs and 174,718 indels remaining. The BED file containing structural variants, as well as some of the other BED files containing uncertain regions, can also be used to help identify sites in an assessed variant call file that may be questionable. We also varied the cut-offs used to differentiate between high-confidence and uncertain variants, and found that they had only a small effect (<0.05% of variants), demonstrating the robustness of our methods that integrate multiple data sets (**Supplementary Discussion**). All variant call files (VCFs) and BED files are publicly available on the Genome in a Bottle ftp site described in the Online Methods.

Different variant representations make comparison difficult

Indels and complex variants (i.e., SNPs and indels less than ~20 bases from each other) are particularly difficult to compare across different variant callers, because they can frequently be represented correctly in multiple ways. Therefore, we first regularized each of the VCFs using the vcfallelicprimitives module in vcflib (https://github.com/ekg/vcflib), so that different representations of the same indel or complex variant were not counted as different variants. Our regularization procedure split adjacent SNPs into individual SNPs, left-aligned indels and regularized representation of homozygous complex variants. However, it could not regularize heterozygous complex variants without phasing information in the VCF, such as individuals that are heterozygous for

Table 1 Description of data sets and their processing to produce bam files for our integration methods

Source[a]	Platform	Mapping algorithm	Coverage	Read length	Genome/exome
1000 Genomes	Illumina GaIIx	BWA	39	44	Genome
1000 Genomes	Illumina GaIIx	BWA	30	54	Exome
1000 Genomes	454	Ssaha2	16	239	Genome
X Prize	Illumina HiSeq	Novoalign	37	100	Genome
X Prize	SOLiD 4	Lifescope	24	40	Genome
Complete Genomics	Complete Genomics	CGTools 2.0	73	33	Genome
Broad	Illumina HiSeq	BWA	68	93	Genome
Broad	Illumina HiSeq	BWA	66	66	Exome
Illumina	Illumina HiSeq	CASAVA	80	100	Genome
Illumina	Illumina HiSeq – PCR-free	BWA	56	99	Genome
Illumina	Illumina HiSeq – PCR-free	BWA	190	99	Genome
Life Technologies	Ion Torrent	tmap	80	237	Exome
Illumina	Illumina HiSeq – PCR-free	BWA-MEM	60	250	Genome
Life Technologies	Ion Torrent	tmap	12	200	Genome

[a] These data and other data sets for NA12878 are available at the Genome in a Bottle ftp site at NCBI (ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/) and are described on a spreadsheet at http://genomeinabottle.org/blog-entry/existing-and-future-na12878-datasets.

the CAGTGA>TCTCT change that is aligned in four different ways in **Figure 2**. Most alignment-based variant callers would output four different VCF files for these four representations. To avoid this problem in our integration process, we represented calls from all data sets in the output format of GATK HaplotypeCaller 2.6-4, which regularizes representation across data sets by performing *de novo* assembly. When comparing calls from other variant callers, we recommend using the vcfallelicprimitives module in vcflib, as well as manually inspecting some discordant sites, to determine whether the calls are actually different representations of the same complex variant.

Integrated variant calls are highly sensitive and specific

Transition/transversion ratio (Ti/Tv) is sometimes used as a metric for accuracy of calls, as the Ti/Tv of mutations is significantly higher than the 0.5 Ti/Tv expected from random sequencing errors. Our integrated calls have a Ti/Tv comparable to that of the other data sets for common variants in the whole genome and exome, but our integrated calls have a higher ratio than that of the other data sets for novel whole-genome variants, which usually indicates a lower error rate (**Table 2**). However, it should be noted that Ti/Tv is limited in its use because the assumption that novel or more difficult variants should have the same ratio as common variants may not be true[20].

We also compared our SNP and indel calls to 'high-quality variants' found in more than one sequencing platform (mostly confirmed using Sanger sequencing) as part of the GeT-RM project (http://www.ncbi.nlm.nih.gov/variation/tools/get-rm/). Our integrated calls correctly genotyped all 427 SNPs and 42 indels. In addition, we compared our calls to Sanger sequencing data from the XPrize, and found that the calls were concordant for all 124 SNPs and 37 indels. Also, we determined that none of our high-confidence variants fell in the 366,618 bases that were covered by high-quality homozygous reference Sanger reads from the GeT-RM project (i.e., there were no false positives in these regions).

In addition, to understand the accuracy of both our SNP and indel calls across larger regions of the genome, we compared our calls to the fosmid calls generated by the XPrize from Illumina and SOLiD sequencing of fosmids covering one allele of ~5% of the genome. Fosmid sequencing is advantageous in that only one allele is measured, so no heterozygous genotypes should exist. However, because only one allele is measured, it can assess both false-positive and false-negative rates of homozygous calls, but it can only assess false-negative rates of heterozygous calls in our integrated calls. Our calls were highly concordant overall, with 76,698 concordant homozygous SNP calls, 58,954 concordant heterozygous SNP calls, 5,061 concordant homozygous indel calls and 5,881 concordant heterozygous indel calls.

To understand which method was correct when our integrated calls and the fosmid calls were discordant, we manually curated alignments from several data sets in the regions around a randomly selected 25% of the discordant variants (**Supplementary Discussion**, **Supplementary Tables 4** and **5**, and **Supplementary Figs. 4–18**). Manual curation of alignments from multiple data sets, aligners and sequencing platforms allowed us to resolve the reasons for all of the differences between our integrated calls and the fosmid calls. For the manually curated variants in our integrated calls and not in the fosmid calls, almost all were false negatives in the fosmid calls owing to miscalled complex variants (**Supplementary Fig. 4**) or overly stringent filtering (**Supplementary Fig. 5**). For variants in the fosmid calls and not in our integrated calls, several reasons were found for the differences, but our integrated calls appeared to be correct except for a few partial complex variant calls. Our analysis suggests that our integrated calls likely contain ~3 partial complex variant calls and between 0 and 1 false-positive or false-negative simple SNP or indel calls per 30 million high-confidence bases, in which our integrated calls contain ~94,500 true-positive SNPs and ~1400 true-positive indels.

Genotyping microarrays are an orthogonal measurement method that has been used to assess the accuracy of sequencing genotype calls at sites interrogated by the microarray[13]. When assessed against microarray genotype calls, our integrated genotype calls are highly sensitive and specific (**Table 2**). We correctly called 564,410 SNPs on the microarray. There were 1,332 SNPs called by the microarray not in

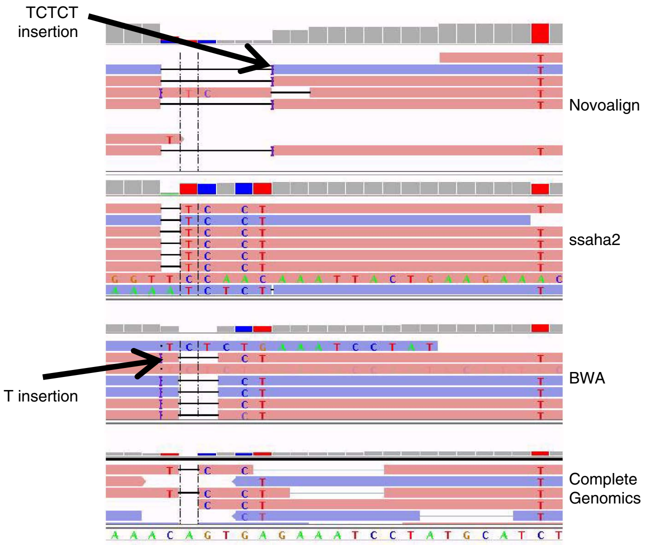

Figure 2 Complex variants have multiple representations. Example of complex variant with four different representations from four different mappers, which can cause data sets to appear to call different variants when in reality they are the same variant. In this case, the six bases CAGTGA are replaced by the five bases TCTCT at location 114841792–114841797 on chromosome 1. The four sets of reads are from Illumina mapped with BWA, 454 mapped with ssaha2, Complete Genomics mapped with CGTools, and Illumina mapped with Novoalign.

Table 2 Performance assessment of two individual callsets and our integrated calls vs. OMNI microarray SNPs and versus our integrated SNPs and indels

		OMNI SNPs with integrated BED file		OMNI SNPs without BED file		Integrated SNPs with BED file		Integrated indels with BED file		Common variants	Novel variants
Data set	Capture	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)	Sensitivity (%)	PR[a] (%)	Sensitivity (%)	PR[a] (%)	Ti/Tv	Ti/Tv
250bp_HC	Genome	99.49	99.97	98.47	99.93	99.90	99.73	99.55	93.11	2.04	1.43
CG	Genome	98.55	99.98	97.11	99.96	97.09	99.27	72.27	89.43	2.10	1.29
Integrated	Genome	99.54	99.98	n/a	n/a	n/a	n/a	n/a	n/a	2.14	1.94
250bp_HC	Exome	99.55	99.98	99.10	99.96	99.90	99.58	100.00	94.60	2.60	1.57
CG	Exome	98.35	99.99	97.64	99.96	99.00	99.04	90.00	85.86	2.71	1.04
Integrated	Exome	99.57	99.98	n/a	n/a	n/a	n/a	n/a	n/a	2.92	1.33[b]

[a]Precision ratio (PR) = true positive/(true positive + false positive). The specificity of all data sets versus our integrated calls is 100.00% owing to the large number of TNs. [b]Our integrated calls only contain 30 novel variants in the exome, so the Ti/Tv has a high uncertainty. 250bp_HC is 250-bp Illumina sequencing mapped with BWA-MEM and called with GATK HaplotypeCaller v2.6. CG is Complete Genomics sequencing from 2010. n/a, not applicable.

our calls, and 527 variants calls in our set that were at positions called homozygous reference in the microarray. We manually inspected 2% of the SNPs specific to the microarray and 4% of the calls specific to our calls. For the manually inspected SNPs specific to the microarray, about half were clearly homozygous reference in all sequencing data sets, without any nearby confounding variants (**Supplementary Fig. 19**). In addition, several sites were adjacent to homopolymers, which were mostly correctly called indels in our high-confidence calls (**Supplementary Figs. 20** and **21**), but several of our calls had incorrect indel lengths (**Supplementary Figs. 22** and **23**). A few sites were also incorrectly called SNPs by the array due to nearby variants (**Supplementary Figs. 24** and **25**). We also manually inspected our calls at locations that were homozygous reference in the microarray, and found that these calls were either true indels (**Supplementary Fig. 26**), had nearby variants that confounded the probe (**Supplementary Fig. 27**), or the probe was designed to detect a SNP that was not present in this sample (**Supplementary Fig. 28**).

Finally, to ensure our variant calling methods are not missing any sites that might be found by other variant callers, we compared our high-confidence genotypes to a callset generated by FreeBayes on Illumina exome sequencing data. There were 208 variants in the FreeBayes variant calls with coverage >20 that we called as high-confidence homozygous reference. We manually inspected a random 10% of these putative variants, and all of them appeared to be likely false positives in FreeBayes owing to systematic sequencing or mapping errors, or sites where FreeBayes called an inaccurate genotype for the correct variant.

Comparison to arrays overestimates sensitivity

Although microarrays can be useful to help understand sequencing performance, they can assess performance only in the regions of the genome accessible to the array (i.e., sequences to which probes can bind and bind uniquely). In addition, microarray genotypes can be confounded by indels (**Supplementary Fig. 26**) and nearby phased variants that are not in the array probes (**Supplementary Fig. 27**). Microarrays tend to contain known common SNPs and avoid genome regions of low complexity. For example, if 'low complexity' is defined as having genome mappability scores[21] <50 for Illumina, SOLiD or Ion Torrent, then only 0.0117% of the microarray sites are in low-complexity regions, compared with 0.7847% of integrated variants. Nevertheless, our integrated calls also ignored many regions of low complexity as 9.8% of uncertain sites are in such regions.

To understand the impact of including lower complexity sites for performance assessment, we explored the use of our integrated genotype calls as a benchmark to assess genotype calls from single data sets, and compared this assessment to an assessment using microarrays (**Table 2**). Many of the sites that were discordant in the microarray were due to errors in the microarray (**Supplementary Figs. 19–28**), so false-positive rates were actually lower when sequencing data sets were assessed against our high-confidence genotypes. Apparent false-negative rates were approximately the same, but many of the apparent false negatives when assessed against the microarray were actually false positives in the microarray.

Using high-confidence calls to understand performance

As an example, we selected a new whole-genome variant call set from the Broad Institute/1000 Genomes Project to show how this set of high-confidence genotype calls can be used to understand and improve performance even for new versions of a sequencing technology (2× 250 paired-end Illumina reads), mapping algorithm (BWA-MEM)[22], and variant caller (GATK v.2.6 HaplotypeCaller)[15]. In addition, we compared an older Complete Genomics callset to see how calls from a completely different pipeline compare. We also assessed performance of several exome data sets on GCAT (http://www.bioplanet.com/gcat) that use 150× Illumina+Novoalign+FreeBayes[23], Illumina+Novoalign+Samtools[24], Illumina+BWA+FreeBayes[25], and 30× Ion Torrent+Tmap+GATK-Haplotype Caller (**Supplementary Figs. 29–36**).

Figure 3 and **Supplementary Figures 31**, **32** and **35** contain receiver operating characteristic (ROC) curves showing how false-positive and true-positive rates change while varying the cutoff for read depth or variant quality score. Variant quality score gives a better ROC curve than does read depth in most cases, probably because sites with very high read depth can actually have higher error rates because of mapping problems. The new 250-bp Illumina whole-genome data analyzed with HaplotypeCaller has a higher accuracy than the older Complete Genomics or any of the exome sequencing data sets for both SNPs and indels. The 150× Illumina exome callsets had a higher accuracy than the 30× Ion Torrent exome callset, particularly for indels. The accuracy for SNPs was much higher than the accuracy for indels in all callsets, which was expected because indels are more difficult to detect than SNPs, especially in homopolymers and low complexity sequence. From the ROC curves, it is apparent that the variant quality score cutoff for the HaplotypeCaller for this data set was probably not optimal, as raising the cutoff could substantially lower the false-positive rate while only minimally increasing the false-negative rate.

Direct observation of alignments around discordant genotypes is often a useful way to understand the reasons behind inaccurate genotype calls. For example, an apparent systematic Illumina sequencing error that was in both the new HaplotypeCaller and UnifiedGenotyper callsets was arbitrated correctly in the integrated callset (**Supplementary Fig. 37**). Many of the differences were due

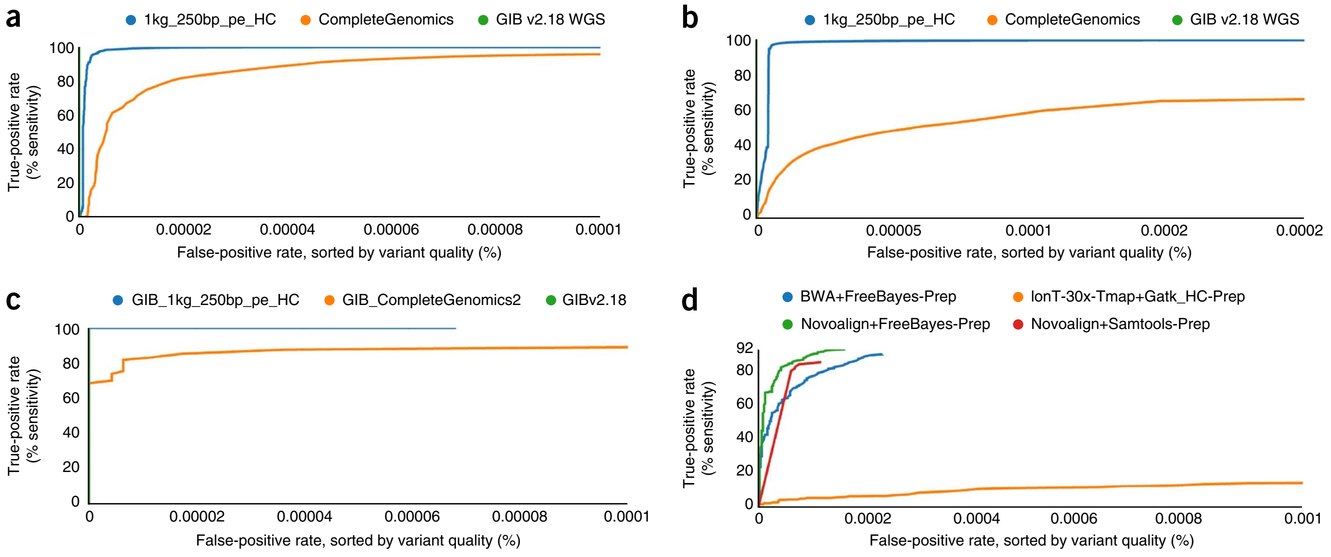

Figure 3 GCAT website can be used to generate performance metrics versus our high-confidence genotypes (GIB v2.18 WGS and GIBv2.18). (**a**–**d**) ROC curves plotting true-positive rate (sensitivity) versus false-positive rate, with variants sorted by variant quality score, for whole-genome SNPs (**a**), whole-genome indels (**b**) and exome indels (**c**,**d**). The assessed variant calls come from Complete Genomics 2.0 (CompleteGenomics and GiB_CompleteGenomics2), 250-bp Illumina mapped with BWA-MEM and called with GATK HaplotypeCaller v2.6 (1kg_250bp_pe_HC and GiB_1kg_250bp_pe_HC), 150× Illumina exome sequencing mapped with BWA and called with FreeBayes (BWA+FreeBayes-Prep), 30× Ion Torrent exome sequencing mapped with Tmap and called with GATK HaplotypeCaller (IonT-30x-Tmap+Gatk_HC-Prep), 150× Illumina exome sequencing mapped with Novoalign and called with FreeBayes (Novoalign+FreeBayes-Prep), and 150× Illumina exome sequencing mapped with Novoalign and called with Samtools (Novoalign+Samtools-Prep).

to difficult regions with low mapping quality, where it was often difficult to determine the correct answer from short-read sequencing (e.g., **Supplementary Fig. 38**).

The variant calls in any data set can also be intersected with our BED files containing different classes of 'difficult regions' of the genome (**Supplementary Table 3**). These comparisons can identify potentially questionable variant calls that should be examined more closely. About 1 million variants called in the 250-bp Illumina HaplotypeCaller VCF are inside NA12878 structural variants reported to dbVar, which is the largest number of variants in any category. Further work will need to be done to determine which structural variants are accurate, but variants in these regions could be inspected further. There are also over 200,000 variants called in the 250-bp Illumina HaplotypeCaller VCF in each of several uncertain categories: sites with unresolved conflicting genotypes, known segmental duplications, regions with low coverage or mapping quality and simple repeats. Although many of these variants may be true variants, they could be examined more closely to identify potential false positives.

DISCUSSION

To develop a benchmark whole-genome data set, we have developed methods to integrate sequencing data sets from multiple sequencing technologies to form high-confidence SNP and indel genotype calls. The resulting genotype calls are more sensitive and specific and less biased than any individual data set, because our methods use characteristics of biases associated with systematic sequencing errors, local alignment errors and mapping errors in individual data sets. We also minimize bias toward any individual sequencing platform by requiring that at least five times more data sets agree than disagree. Therefore, even though there are more Illumina data sets, other platforms would have to agree with the Illumina data sets for them to override two data sets that disagreed. In addition, we include an annotation "platforms" in the INFO field in the VCF file that specifies the number of platforms that support a call. Potential platform bias can be minimized even further by selecting only variants that are supported by data from two or more platforms.

The process used to generate any set of benchmark genotype calls can affect the results of performance assessment in multiple ways (**Supplementary Table 1**). (i) If many 'difficult' regions of a genome are excluded from the truth data set (or labeled "uncertain," meaning that they may be downweighted or disregarded in performance assessment), any assessed data sets will have lower apparent false-positive and false-negative rates than if the difficult regions were included. Therefore, it is important to recognize that any comparison to our high-confidence genotypes excludes the most difficult variants and regions of the genome (currently ~23% of the genome, including potential structural variants, **Supplementary Table 2**). In addition, almost all of the indels are currently <40 bp in length. (ii) False-positive variant calls in the truth data set could result in an assessed false-negative rate higher than the true false-negative rate if the assessed calls are correct, or in an assessed false-positive rate lower than the true false-positive rate if the assessed calls are also false positives. (iii) False-negative variant calls in the truth data set could result in an assessed false-positive rate higher than the true false-positive rate if the assessed calls are correct, or in an assessed false-negative rate lower than the true false-negative rate if the assessed calls are also false negatives. (iv) Many comparison tools treat heterozygous and homozygous variant genotype calls as equivalent, which enables simple calculations of sensitivity and specificity, but these genotypes can have different phenotypes, so it is often important to assess whether the genotype is accurate, as we do in this work, and not just whether a variant is detected.

In general, for the benchmark calls to be useful for performance assessment, the false-positive rate of the benchmark should be much lower than the false-negative rate of the assessed data set, and the false-negative rate of the benchmark should be much lower than the false-positive rate of the assessed data set. To be confident our benchmark integrated calls are not biased toward any sequencing or bioinformatics method and have sufficiently low false-negative and false-positive rates, we compared our integrated calls to multiple independent methods (microarrays, capillary sequencing, fosmid sequencing, Illumina exome sequencing called with FreeBayes, and new 2× 250-bp long-read Illumina sequencing mapped with a recently developed algorithm BWA-MEM and analyzed with a recent version of GATK).

Although we have shown that our integrated calls have low false-positive and false-negative rates, we recommend that users of our integrated calls examine alignments around a subset of discordant genotype calls, such as using the new GeT-RM browser for NA12878 (http://www.ncbi.nlm.nih.gov/variation/tools/get-rm/). Overall statistics such as sensitivity and specificity are useful, but manual inspection of alignments from multiple data sets (or even a single data set) for a subset of discordant sites is essential for properly understanding any comparison between variant call sets. Manual inspection can help identify discordant representations of the same variant, potential biases in sequencing and bioinformatics methods, and difficult regions of the genome where variant calls may be questionable. In addition to comparing genotype calls in the regions for which we determined we can make confident integrated genotype calls, we also recommend examining variant calls in regions we consider uncertain for different reasons. Examining these difficult regions can help identify variants that may be questionable. We also encourage contacting the authors of this manuscript if any genotypes in our integrated calls are questionable or in error, as this call set will be maintained and refined over time as new sequencing and analysis methods become available.

These methods represent the basis of methods to form high-confidence genotype calls for genomes selected as reference materials by the Genome in a Bottle Consortium. This consortium will develop the reference materials, reference data and reference methods to enable translation of genome sequencing to clinical practice. As we show in this work, high-confidence genotype calls from a well-characterized whole genome are useful for assessing biases and rates of accurate and inaccurate genotype calls in any combination of sequencing and bioinformatics methods. High-confidence genotype calls for publicly available genomes will be particularly useful for performance assessment of rapidly evolving sequencing and bioinformatics methods. This resource is publicly available through the Genome in a Bottle Consortium website and ftp site at NCBI (ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878). The resource will continue to be improved and updated with additional types of variants (e.g., complex variants and structural variants) and with increasingly difficult regions of the genome, incorporating new sequencing data as they are collected.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. European Nucleotide Archive: SRP012400, SRP000547, SRP018096, SRP000032, ERP001229.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Johnson and A. Varadarajan from the Archon Genomics X Prize and EdgeBio for contributing their whole-genome sequencing data from SOLiD and Illumina, Complete Genomics and Life Technologies for providing bam files for NA12878, and the Broad Institute and 1000 Genomes Project for making publicly available bam and VCF files for NA12878. The Illumina exome data on GCAT were given to the Mittelman laboratory by M. Linderman at Icahn Institute of Genomics and Multiscale Biology of the Icahn School of Medicine at Mount Sinai. We thank the US Food and Drug Administration High Performance Computing staff for their support in running the bioinformatics analyses. Harvard School of Public Health contributions were funded by the Archon Genomics X PRIZE. Certain commercial equipment, instruments or materials are identified in this document. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.

AUTHOR CONTRIBUTIONS

J.M.Z., M.S., B.C., O.H. and W.H. conceived the integration methods. J.M.Z. wrote the code for the integration methods and wrote the main manuscript. D.M. and J.W. designed the GCAT platform, implemented comparison to our genotype calls, and generated figures.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
2. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).
3. Jones, D.T.W. *et al.* Dissecting the genomic complexity underlying medulloblastoma. *Nature* **488**, 100–105 (2012).
4. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
5. Boland, J.F. *et al.* The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing. *Hum. Genet.* **132**, 1153–1163 (2013).
6. Rieber, N. *et al.* Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PLoS ONE* **8**, e66621 (2013).
7. Ross, M.G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51 (2013).
8. Lam, H.Y.K. *et al.* Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.* **30**, 78–82 (2012).
9. Reumers, J. *et al.* Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat. Biotechnol.* **30**, 61–68 (2012).
10. Author, A. *The Plasma Proteins: Structure, Function and Genetic Control*, edn. 2 (Academic Press, New York, 1975).
11. O'Rawe, J. *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* **5**, 28 (2013).
12. Collins, F. & Hamburg, M. First FDA authorization for next-generation sequencer. *N. Engl. J. Med.* **369**, 2369–2371 (2013).
13. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
14. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
15. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
16. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
17. Blum, A. & Mitchell, T. in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (eds. P. Bartlett & Y. Mansour) 92–100 (ACM, Madison, Wisconsin, USA, 1998).
18. Meacham, F. *et al.* Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12**, 451 (2011).
19. Zook, J.M., Samarov, D., McDaniel, J., Sen, S.K. & Salit, M. Synthetic spike-in standards improve run-specific systematic error analysis for DNA and RNA sequencing. *PLoS ONE* **7**, e41356 (2012).
20. Tian, D.C. *et al.* Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455**, 105–108 (2008).
21. Lee, H. & Schatz, M.C. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics* **28**, 2097–2105 (2012).
22. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv:1303.3997v2 [q-bio.GN] (2013).
23. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at arXiv:1207.3907v2 [q-bio.GN] (2012).
24. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
25. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

ONLINE METHODS

Data sets. Nine whole-genome and three exome sequencing data sets (see **Table 1** for details about source, platform, mapping algorithm, coverage and aligned read length) were used to form the integrated genotype calls for Coriell DNA sample NA12878. Six whole-genome (two PCR-free) and two exome data sets were from Illumina sequencers, one whole genome from SOLiD sequencers, one whole genome from 454 sequencer, one whole genome from Complete Genomics and one exome from Ion Torrent[26]. Some have bam files publicly available, which were used directly in this work. These data and other data sets for NA12878 are available at the Genome in a Bottle ftp site at NCBI (ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878) and are described on a spreadsheet at http://genomeinabottle.org/blog-entry/existing-and-future-na12878-datasets. In addition, the results of this work (high-confidence variant calls and BED files describing confident regions) are available at ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/NIST along with a README.NIST describing the files and how to use them. The files used in this manuscript are NISTIntegratedCalls_14data sets_131103_HetHomVarPASS_VQSRv2.18_2mindata sets_5minYesNoRatio_all_nouncert_excludesimplerep_excludesegdups_excludedecoy_excludeRepSeqSTRs_noCNVs_vardist.VCF.gz, which contains high-confidence heterozygous and homozygous variant calls, and union13callableMQonlymerged_addcert_nouncert_excludesimplerep_excludesegdups_excludedecoy_excludeRepSeqSTRs_noCNVs_v2.18_2mindata sets_5minYesNoRatio.BED.gz, which contains intervals that can be considered high-confidence homozygous reference (for SNPs and short indels) if there is not a variant in the VCF.

Comparison of variant calls using different methods. To compare variants called using different methods, we first sought to normalize representation of short indels, complex variants and multinucleotide polymorphisms (MNPs) so that the same variant represented in different ways would not be considered discordant. We used the vcfallelicprimitives module in vcflib (https://github.com/ekg/vcflib) to help regularize representation of these variants. Regularization minimizes counting different methods of expressing the same variant (e.g., nearby SNPs/indels) as different variants. Our regularization procedure splits adjacent SNPs into individual SNPs, left-aligns indels and regularizes representation of homozygous complex variants. However, it cannot regularize heterozygous complex variants without phasing information in the VCF, such as individuals that are heterozygous for the CAGTGA>TCTCT change that is aligned in four different ways in **Figure 2**. Regularizing heterozygous complex variants without phasing information is not generally possible because they could be phased in multiple ways. All other shell (Sun Grid Engine) and perl scripts written for this work and the resulting BED file are publicly available at https://github.com/jzook/genome-data-integration.

Obtaining high-confidence genotypes for training VQSR. To reduce the number of sites that need to be processed, we first used GATK (v. 2.6-4) UnifiedGenotyper and HaplotypeCaller with a low variant quality score threshold of 2 to find all possible SNP and indel sites in each data set except Complete Genomics. For Complete Genomics, we used their unfiltered set of SNP and indel calls from CGTools 2.0. In addition, we included sites called by Cortex *de novo* assembly method for the ~40× Illumina PCR-free data set. The union of these sites from all data sets served as our set of possible SNP sites for downstream processing.

As each data set did not make a genotype call at every possible SNP and indel site, we forced GATK UnifiedGenotyper to call genotypes for each data set individually at all of the possible SNP sites (GATK_…_UG_recall_…sh). In addition, we forced GATK HaplotypeCaller to perform local *de novo* assembly around all candidate indels and complex variants for each data set individually (GATK_…_haplo_recall…sh). We then combined the UG and HC calls, giving preference to HC within 20 bp of an HC indel with a PL>20. We used the genotype likelihoods (PL in VCF file) to determine which sites had genotypes confidently assigned across multiple data sets. We used the minimum nonzero PL (PLdiff), which is the Phred-scaled ratio of the likelihoods of the most likely genotype to the next most likely genotype (similar to the Most Probable Genotype described previously[27]). In addition, we divided PLdiff by the depth of coverage (PLdiff/DP) as a surrogate for allele balance because PLdiff should increase linearly with coverage in the absence of bias. For a heterozygous variant site to be used to train VQSR, we required that PLdiff>20 for at least two data sets, the net PLdiff for all data sets > 100, the net PLdiff/DP for all data sets > 3.4, fewer than 15% of the data sets had PLdiff>20 for a different genotype, fewer than 30% of the data sets have >20% of the reads with mapping quality zero, and fewer than two data sets have a nearby indel called by HaplotypeCaller but do not call this variant. For a homozygous variant site to be used to train VQSR, we required that PLdiff>20 for at least two data sets, the net PLdiff for all data sets > 80, the net PLdiff/DP for all data sets > 0.8, fewer than 25% of the data sets had PLdiff>20 for a different genotype, and fewer than two data sets have a nearby indel called by HaplotypeCaller but do not call this variant. These requirements were specified to select generally concordant sites with reasonable coverage and allele balances near 0, 0.5 or 1 for training VQSR.

These highly concordant heterozygous and homozygous variant genotypes were used independently to train the VQSR Gaussian Mixture Model separately for each data set for heterozygous and homozygous (variant and reference) genotypes. Unlike the normal VQSR process, we train on heterozygous and homozygous genotypes independently because they could have different distributions of annotations and different characteristics of bias. We fit only a single Gaussian distribution to each annotation because most of the annotations have approximately Gaussian distributions. Thus, additional Gaussians often fit noise in the data, and the model frequently does not converge when attempting to fit more than one Gaussian. We fit VQSR Gaussian mixture models for annotations associated with alignment problems, mapping problems, systematic sequencing errors and unusual allele balance, using the shell and perl scripts *RunVcfCombineUGHaplo_FDA_131103.sh, VcfCombineUGHaplo_v0.3.pl, VcfHighConfUGHaploMulti_HomJoint_1.3_FDA.pl, GATK_VQSR_…_131103.sh, and runVariantRecal…_131103.pl*. The annotations used for systematic sequencing errors, alignment bias, mapping bias and abnormal allele balance for homozygous and heterozygous genotypes are listed in **Supplementary Table 6**. For each genomic position, the VQSR Gaussian mixture model outputs a tranche ranging from 0 to 100, with higher numbers indicating it has more unusual characteristics, which may indicate bias. For example, a tranche of 99.9 means that 0.1% of positions have characteristics more extreme than this position.

Arbitration between data sets with conflicting genotypes. For some positions in the genome, data sets have conflicting genotypes. Our approach to arbitrating between conflicting data sets is summarized in **Figure 1** and **Supplementary Figure 1**. We hypothesize that if a data set has unusual annotations associated with bias at a particular genome site, it is less likely to be correct than a data set with typical characteristics at that genome site. For each possible variant site, we first determine if at least two data sets confidently call the same genotype (PLdiff>20) and at least 5× more data sets confidently call this genotype than disagree (i.e., have PLdiff>20 for a different genotype). In addition, when combining all data sets the net PLdiff and PLdiff/DP must exceed the values in **Supplementary Table 7** for the specific genotype and class of variant. Also, if two data sets have a indel called by the HaplotypeCaller within 20 bp and do not call a variant at this position, then it is declared uncertain. If these conditions are not met, then we use the arbitration process. We start filtering the most unusual sites (tranche > 99.5). We first filter possible systematic sequencing errors above this tranche because they are most likely to be biases. Next, we filter possible alignment problems above this tranche. The order of tranche filtering is 99.5, 99, 95 and 90. We filter decreasing tranches until meeting the conditions above for PLdiff and PLdiff/DP.

Some positions in the genome are difficult for all methods, so even if all data sets agree on the genotype there may be significant uncertainty. For example, if a region has one copy in the hg19/GRCh37 reference assembly but two copies in both alleles in NA12878, and one of the copies has a homozygous SNP, it would incorrectly appear as a heterozygous SNP in all data sets. To minimize incorrect genotype calls, we use the VQSR tranches for annotations associated with systematic sequencing errors, alignment problems, mapping problems and atypical allele balance. For homozygous reference genotypes, we require that at least two data sets have an alignment tranche <99. For heterozygous genotypes, we require that at least three data sets have a mapping tranche <99, at least two data sets have a systematic sequencing error tranche <95, at least two data sets have an alignment tranche <95, at least two data sets have a

doi:10.1038/nbt.2835

mapping tranche <95, and at least two data sets have an allele balance tranche <95. For homozygous variant genotypes, we require that at least three data sets have a mapping tranche <99, at least two data sets have an alignment tranche <99, and at least two data sets have an allele balance tranche <99. For sites not considered potential variants, we determine whether they are callable as homozygous reference by using the GATK CallableLoci walker, requiring that at least three data sets have a coverage greater than 5, excluding base quality scores less than 10, and requiring that the fraction of reads with mapping quality <10 is <10% (CallableLoci_…sh). In addition, we remove all regions with known tandem duplications not in the GRCh37 Reference Assembly, and we optionally have a BED file that removes all structural variants for NA12878 reported in dbVar (as of June 12, 2013), and/or long homopolymers and tandem repeats that do not have at least five reads covering them in one of the data sets with 7 bp mapped on either side (created with BedSimpleRepeatBamCov.pl). We depict regions as "callable" using BED files, which is created using the process described above using MakeBedFiles_v2.18_131103.sh, with results and uncertain categories in **Supplementary Tables 2** and **3**. All bases inside the BED file and not in the variant call file can be considered high-confidence homozygous reference and can be used to assess false-positive rates in any sequencing data set.

GCAT performance assessment of data set. To perform the comparisons in GCAT, we first regularized the variants in the VCF files using vcflib vcfallelicprimitives. For the whole-genome comparisons, the variants were also subset with the BED file excluding dbVar structural variants. For the whole exome comparisons, the variants were subset with both the BED file excluding dbVar structural variants and the target exome BED file from the manufacturer (Ion Torrent TargetSeq_hg19 http://ioncommunity.lifetechnologies.com/docs/DOC-2817 and Illumina exome ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/exome_pull_down_targets/20130108.exome.targets.bed). ROC curves were generated by sorting the variants by coverage or variant quality score and calculating true-positive rate and false-positive rate as variants with decreasing coverage or variant quality score are added.

26. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
27. Ajay, S.S., Parker, S.C.J., Abaan, H.O., Fajardo, K.V.F. & Margulies, E.H. Accurate and comprehensive sequencing of personal genomes. *Genome Res.* **21**, 1498–1505 (2011).

doi:10.1038/nbt.2835