nature biotechnology

ANALYSIS

https://doi.org/10.1038/s41587-019-0054-x

Corrected: Author Correction

Best practices for benchmarking germline small-variant calls in human genomes

Peter Krusche[1], Len Trigg[2], Paul C. Boutros[3], Christopher E. Mason[4,5,6,7], Francisco M. De La Vega[8], Benjamin L. Moore[1], Mar Gonzalez-Porta[1], Michael A. Eberle[9], Zivana Tezak[10], Samir Lababidi[11], Rebecca Truty[12], George Asimenos[13], Birgit Funke[14], Mark Fleharty[15], Brad A. Chapman[16], Marc Salit[17,20], Justin M. Zook[18,20]* and the Global Alliance for Genomics and Health Benchmarking Team[19]

Standardized benchmarking approaches are required to assess the accuracy of variants called from sequence data. Although variant-calling tools and the metrics used to assess their performance continue to improve, important challenges remain. Here, as part of the Global Alliance for Genomics and Health (GA4GH), we present a benchmarking framework for variant calling. We provide guidance on how to match variant calls with different representations, define standard performance metrics, and stratify performance by variant type and genome context. We describe limitations of high-confidence calls and regions that can be used as truth sets (for example, single-nucleotide variant concordance of two methods is 99.7% inside versus 76.5% outside high-confidence regions). Our web-based app enables comparison of variant calls against truth sets to obtain a standardized performance report. Our approach has been piloted in the PrecisionFDA variant-calling challenges to identify the best-in-class variant-calling methods within high-confidence regions. Finally, we recommend a set of best practices for using our tools and evaluating the results.

Advances in sequencing technologies have enabled the generation of large-scale genome, exome, and targeted-sequencing data for both research and clinical diagnostic purposes[1,2]. These technologies output a list of variant calls and their genotypes, often in variant-call format (VCF)[3], and benchmarking the accuracy of these calls is an essential part of analytical validation. This is especially important for clinical laboratories developing sequencing-based tests for medical care. Robust, sophisticated, and standardized benchmarking methods are therefore critical for the development, optimization, and comparison of sequencing, mapping, and variant-calling tools.

Efforts such as the Genome in a Bottle Consortium (GIAB) and Platinum Genomes have developed small variant truth sets, against which variant calls can be compared, for several well-characterized human genomes from publicly available cell lines and DNA[4–7]. Another truth set was recently developed from a 'synthetic-diploid' mixture of two haploid hydatiform-mole cell lines not currently available in a public repository[8]. A framework for benchmarking non-complex small variant calls (that is, simple, isolated single nucleotide variations (SNVs), insertions, or deletions) in the exome has been developed as a web-based tool, called GCAT[9]. However, comparing variant calls from any particular sequencing pipeline to a truth set is not a trivial task. First, variants may be represented in multiple ways in the commonly used VCF[10–13] and when comparing VCF files record by record, many of the observed differences can simply be different representations of the same variant. Second, definitions for performance metrics such as true positive (TP), false positive (FP), and false negative (FN), which are key for the interpretation of the benchmarking results, are not yet standardized. Finally, performance can vary across variant types and genomic regions, which inevitably increases the complexity of benchmarking results.

Two key performance metrics for assessing variant calling accuracy are sensitivity—the ability to detect variants that are known to be present or the absence of FNs—which we refer to as 'recall', and specificity—the ability to correctly identify the absence of variants or the absence of FPs—for which we use the similar metric 'precision'[14]. The robustness of these metrics is particularly important for genome-sequencing applications in which novel sequence variants may be identified. Early professional guidelines call for the use of samples with and without known pathogenic variants to determine sensitivity and specificity, but this approach cannot predict performance for novel variants. To predict performance for novel variants, it is important to maximize the number and variety of variants that can be compared to a 'gold standard' to establish statistical confidence values for different types of variants and genome contexts, which can then be extrapolated to all sequenced bases[15–18]. Guidelines were recently published for validating clinical bioinformatics assays[19] and highlighted not only the utility of reference materials for

[1]Illumina Cambridge Ltd, Little Chesterford, UK. [2]Real Time Genomics, Hamilton, New Zealand. [3]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [4]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [5]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. [6]The Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA. [7]The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA. [8]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA. [9]Illumina Inc., San Diego, CA, USA. [10]Center for Devices and Radiological Health, FDA, Silver Spring, MD, USA. [11]Office of Health Informatics, Office of the Commissioner, FDA, Silver Spring, MD, USA. [12]Invitae, San Francisco, CA, USA. [13]DNAnexus, San Francisco, CA, USA. [14]Veritas Genetics, Danvers, MA, USA. [15]Broad Institute, Cambridge, MA, USA. [16]Bioinformatics Core, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [17]Joint Initiative for Metrology in Biology, Stanford University, Stanford, CA, USA. [18]Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA. [19]The members of the GA4GH Benchmarking Team are the same as the author list. [20]These authors contributed equally: Marc Salit, Justin M. Zook. *e-mail: jzook@nist.gov

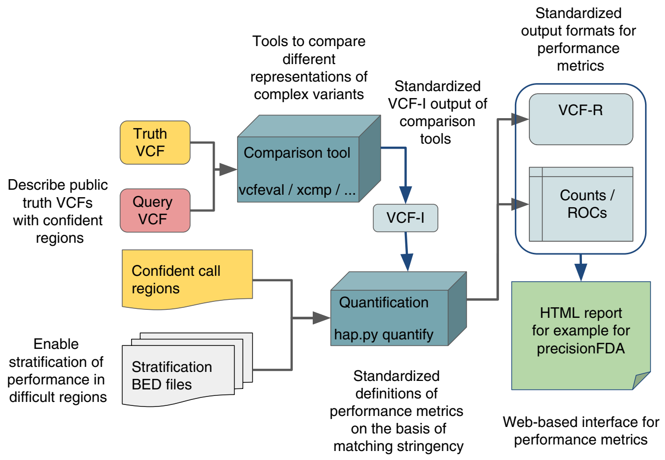
Standardized output formats for performance metrics
Tools to compare different representations of complex variants
Standardized VCF-I output of comparison tools
Truth VCF
Comparison tool
vcfeval / xcmp / ...
VCF-R
Describe public truth VCFs with confident regions
Query VCF
VCF-I
Counts / ROCs
Confident call regions
Quantification
hap.py quantify
HTML report for example for precisionFDA
Enable stratification of performance in difficult regions
Stratification BED files
Standardized definitions of performance metrics on the basis of matching stringency
Web-based interface for performance metrics

Fig. 1 | The GA4GH Benchmarking Team's reference implementation of a comparison framework, annotated with free-floating text describing the team's innovations. The framework takes in a truth VCF, query VCF, confident call regions for the truth and/or query, and optionally BED files to stratify performance by genome context. A standardized intermediate output (VCF-I) from the comparison engines allows them to be interchanged and for TP, FP, and FN to be quantified and output in a standardized report, including a VCF with all TPs, FPs, and FNs and stratifications as annotations (VCF-R) and receiver operating characteristic (ROC) curves.

benchmarking variant calls, but also the importance of stratifying performance by variant type and genome context.

The aim of the Global Alliance for Genomics and Health (GA4GH) Benchmarking Team was to bring together participants from research institutes, technology companies, government agencies, and clinical laboratories, to standardize variant-calling benchmarking. Here we describe the available reference materials and tools to benchmark variant calls, and provide best practices for using these resources and interpreting benchmarking results. We standardized the variant benchmarking process in several ways (Fig. 1). We reconciled methods (for example, rtgtools vcfeval[10] and hap.py xcmp[20]) developed to compare callsets to assess the accuracy of variant and genotype calls independent of different representations of the same variant. We also represented primary performance metrics in the most commonly used binary classification form (that is, TP, FP, FN, and statistics derived from these) and standardized the calculation of performance metrics to make them more easily comparable across methods (for example, by hap.py quantify[20]). Finally, we provide a framework whereby performance metrics can be stratified by variant type and genome context.

We discuss the technical challenges of, and present a solution to, accurately comparing VCF files and provide a strategy for benchmarking variant calls against available truth sets (for example, GIAB or Platinum Genomes). Overall, we develop a framework for standardizing and addressing some of the major challenges associated with benchmarking variant calls and summarize best practices for benchmarking in Supplementary Table 1.

Results

Comparing variant calls to truth sets. We provide a strategy to benchmark single-sample query VCFs against existing truth sets. The inputs to this comparison are a truth callset (in VCF format), and a set of confident regions (in BED format) for the truth set. The confident regions indicate the locations of the genome where, when comparing to the truth callset, variants that do not match the truth callset should be FPs and variants missed in the truth callset should be FNs. Furthermore, our inputs include a query callset in VCF format, a reference genome sequence FASTA file (for example, GRCh37 or GRCh38), and optionally stratification regions to break out variant-calling performance in particular regions of the genome (for example, repetitive regions) or to restrict comparisons to a genomic subset (for example, exons/regions captured by targeted sequencing).

Variant representation. The primary challenge with comparing two VCF files is handling complex variant representations correctly. In a VCF file, we describe two haplotype sequences by means of REF/ALT pairs and genotypes. These variant calls do not always uniquely represent the same haplotype sequences. Alignments are not always unique even when using a fixed set of gap and substitution scores; different variant-calling methods may produce different variant representations. Although some of these differences can be handled using pre-processing of VCF files (for example, variant trimming and left-shifting), others cannot be fixed easily. As a result we cannot compare VCF files accurately by comparing VCF records and genotypes directly. Approaches were developed to standardize indel representation by means of left-shifting and trimming the indel alleles[21,22]. These methods determine the left-most and right-most positions at which a particular indel could be represented in a VCF file (Fig. 2a). These methods work well when considering each VCF record independently. However, when multiple VCF records are used to represent a complex haplotype, normalization methods can cause errors and more sophisticated comparison methods are required (Fig. 2b–d). Different types of variant-representation challenges are detailed in Supplementary Note 1. When benchmarking, these variant-representation differences can also give rise to different notions of giving partial credit for variant calls, such as when a method calls only one SNV in a multi-nucleotide variant (defined as multiple, adjacent SNVs)[23]. Our tools attempt to give partial credit when possible, and we generally recommend using vcfeval as the comparator to provide the most partial matches. Each of the comparison tools is described in Supplementary Note 2.

Matching stringencies and defining performance metrics. Owing to the inherent complexity of the human genome, and the challenge that genotype comparisons do not cleanly fall in a binary classification model, TP, FP, and FN can be defined in different ways. Our reference implementation for benchmarking uses a tiered definition of variant matches, a standardized VCF format for outputting matched variant calls (see Supplementary Note 3), and a common counting and stratification tool for variant type and genome context (see Supplementary Note 4). We consider three types of variant matches from most to least stringent: (1) 'genotype match', for which only sites with matching alleles and genotypes are counted as TPs, (2) 'allele match', for which any site with matching alleles is counted as TP, even if genotypes differ, and (3) 'local match', for which any site in the query with a nearby truth variant is counted as a TP, even if alleles and genotypes differ. 'Genotype match' is used by our current tools to calculate TP, FP, and FN.

In Table 1, we enumerate the types of matches that are clear TP, FP, and FN as well as various kinds of partial matches that may be considered TP, FP, and/or FN depending on the matching stringency, and how they are counted by our tools. Our tools calculate TP, FP, and FN requiring the genotype to match, but output additional statistics related to how many of the FPs and FNs are allele matches (FP.GT) or local matches (FP.AL). Table 2 includes examples of how different truth and query genotypes are counted as TP, FP, FN, FP.GT, and FP.AL. Note that we have chosen not to include true negatives (or consequently specificity) in our standardized definitions, because these metrics tend to be challenging to interpret (Methods)

We have implemented the comparison methods and metrics discussed above in hap.py; a standardized report that can be generated from the tabular output of the benchmarking workflow[20]

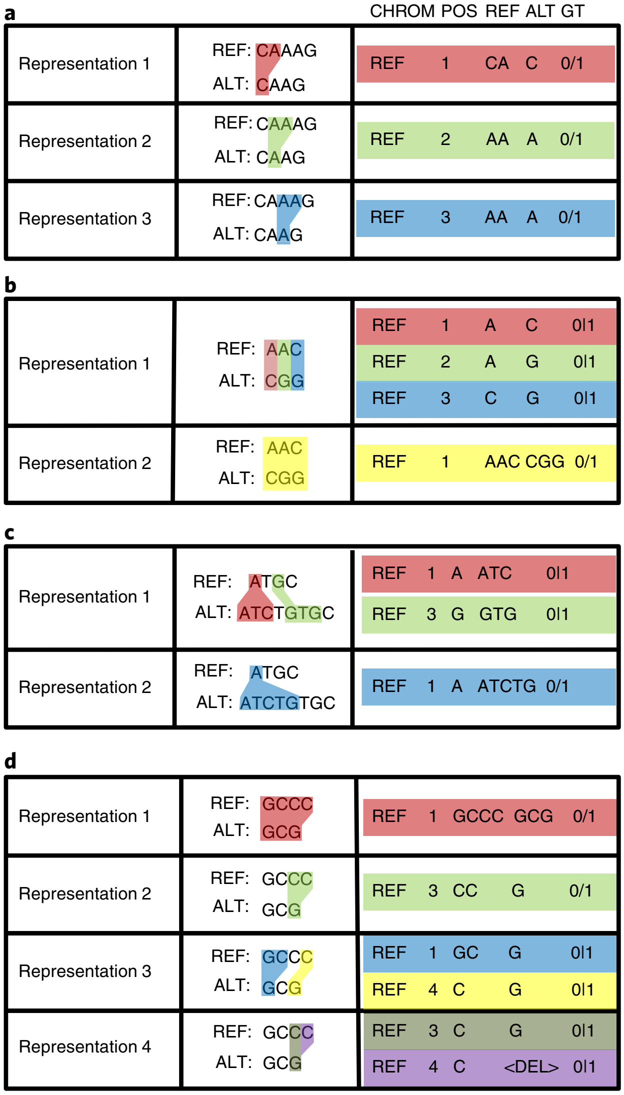

Fig. 2 | Four examples of cases in which variants can be represented in multiple forms in VCF format. a, Three representations of a deletion in a homopolymer. **b**, An MNP can be represented as three SNVs or one larger substitution. **c**, The insertion can be represented as one 4-bp insertion or two 2-bp insertions. **d**, Four different representations of a complex variant. Note that representations include phasing information in these examples in which it is necessary to unambiguously describe the variant. If phasing was not described for these variants, it would be impossible to normalize their representations, but our sophisticated variant-comparison tools can determine that they could describe the same two haplotypes.

(see example in Supplementary Fig. 1). Definitions and formulas for all performance metrics, including derived metrics such as precision and recall, are detailed in the Methods and Supplementary Table 2

Benchmark callsets. Benchmarking of variant calls requires a specific genome and an associated set of calls that represent the 'right answers' for that genome. Such callsets have the property that they can be used as truth to accurately identify FPs and FNs. That is, when comparing calls from any sequencing method to this set of calls, at least half (and ideally more) of the putative FPs and FNs should be errors in the method being assessed. Because it is treated as the truth, this benchmark set will be referred to in this manuscript as the truth set, but other terms used for this include the 'gold-standard' set, the 'high-confidence' set, the 'reference callset,' or 'benchmarking data.'

We describe three sources of benchmark callsets in detail in the Methods. (1) GIAB is an ongoing public–private–academic consortium hosted by the National Institute of Standards and Technology (NIST) to perform authoritative characterization of a small number of broadly consented and disseminated human genomes. Currently, five human genomes are available as NIST Reference Materials with benchmark small variant and reference calls for approximately 90% of GRCh37 and GRCh38[5,7,24]. (2) Platinum Genomes has also created a benchmarking dataset for small variants (SNVs and indels) using the 17-member pedigree (1,463) from Coriell Cell Repositories that includes the GIAB pilot sample NA12878/HG001[6]. This pedigree includes 11 children of the parents (NA12877 and NA12878), producing a fully phased dataset that confirms the accuracy of variant calls through genetic-inheritance patterns. A new draft merged GIAB–Platinum Genomes benchmark set for NA12878/HG001 is described in Supplementary Note 5 and Supplementary Fig. 2. (3) A new 'synthetic-diploid' benchmark callset was created from long read assemblies of the CHM1 and CHM13 haploid cell lines, to benchmark small variant calls in regions difficult to analyze with short reads or in diploid genomes, which are currently excluded from the GIAB and Platinum Genomes high-confidence regions[8]. A current limitation is that CHM1 and CHM13 cell lines are not available in a public repository. It is important to understand how

Table 1 | Contingency table describing the GA4GH definitions of TP, FP, FN, FP.AL, FP.GT, and unknown (UNK)

		Truth				
	Genotype	Ref/ref	Ref/var1	Var1/var2	Var1/var1	Outside bed
Query	Ref/ref	–	FN	FN	FN	–
	Ref/var1	FP	TP	**FP.GT**	**FP.GT**	UNK
	Ref/var2	–	**FP.AL**	**FP.GT**	**FP.AL**	–
	Ref/var3	–	–	**FP.AL**	–	–
	Var1/var2	FP	**FP.GT**	TP	**FP.GT**	UNK
	Var1/var3	–	–	**FP.GT**	–	–
	Var2/var3	–	**FP.AL**	**FP.GT**	**FP.AL**	–
	Var3/var4	–	–	**FP.AL**	–	–
	Var1/var1	FP	**FP.GT**	**FP.GT**	TP	UNK
	Var2/var2	–	**FP.AL**	**FP.GT**	**FP.AL**	–
	Var3/var3	–	–	**FP.AL**	–	–

Matches counted as FP.GT and FP.AL are additionally counted as both FP and FN, since our tool's default matching stringency requires genotypes to match. Query variants outside the truth bed file are counted as UNK. Boxes with dashes are not possible when comparing two VCFs. Matches counted as FP.GT and FP.AL (bold) are additionally counted as both FP and FN, since our tool's default matching stringency requires genotypes to match.

Table 2 | Examples of several combinations of truth and query SNV genotypes and how they are counted as TP, FP, FN, FP.GT, and FP.AL

REF	Truth	Query	Counted as
A	C/C	C/C	1 TP
A	A/A	C/C	1 FP
A	C/C	A/A	1 FN
A	C/C	A/C	1 FP, 1 FN, 1 FP.GT
A	C/C	G/G	1 FP, 1 FN, 1 FP.AL
A	C/G	C/C	1 FP, 1 FN, 1 FP.GT

each truth set was constructed and what biases and limitations it may have. The Methods summarize some limitations of the current truth sets, though these truth sets are likely to be improved in future versions as technologies improve

Lessons from PrecisionFDA challenges. The PrecisionFDA team held two challenges in 2016, with participants publicly submitting results from various mapping/variant-calling pipelines (https://precision.fda.gov/challenges/). While both challenges asked participants to analyze short read whole-genome sequencing datasets, the first 'consistency' challenge used a sample with high-confidence calls already available (HG001/NA12878) and the second 'truth' challenge used a sample without high-confidence calls yet available (HG002 from GIAB, made available by GIAB upon the close of the challenge).

Note that both the truth sets and the comparison methodology in the truth challenge were newly introduced and under active development. The challenge results available on PrecisionFDA should be considered only as an initial evaluation, with the rich dataset resulting from the challenge inviting further exploration. Critical evaluation also calls for manual curation of a subset of FPs and FNs to ensure they are actually FPs and FNs and to understand their cause.

It is important to recognize that our benchmarking metrics indicate performance for the 'easier' variants and regions of the genome, so that precision and recall estimates are higher than if more difficult variants and regions were included. It is likely that some methods will perform worse than other methods for easier variants but perform better for harder variants (for example, methods using a graph reference or de novo assembly may do better calling in regions not assessed, like the major histocompatibility complex or large insertions, while not performing as well for easier variants because the methods are less mature). To gain insight into variant-call variability inside and outside the GIAB high-confidence regions for HG001, we compared two pipelines (DeepVariant and GATK) that had very high accuracy in the PrecisionFDA Challenge inside the high-confidence regions. In the high-confidence regions, when comparing these pipelines to each other (https://precision.fda.gov/jobs/job-FJpqBP80F3YyfJG02bQzPJBj, link immediately accessible by requesting an account), they agreed on 99.7% of SNVs and 98.7% of indels. Outside the high-confidence regions (https://precision.fda.gov/jobs/job-FJpqJF80F3YyXqz6Kv8Q1BQK), they agreed with each other on only 76.5% of SNVs and 78.7% of indels. In tandem repeats longer than 200 bp (lowcmp_AllRepeats_gt200bp_gt95identity_merged_slop5) and regions with 250 bp mapping ambiguity (map_l250_m0_e0) outside the high-confidence regions, they agreed on at most 60% of SNVs and indels. We expect this effect to be even greater when comparing short-read methods to new methods on the basis of long, single-molecule-sequencing reads.

Because larger changes such as structural variants are also generally excluded from the current high-confidence regions, it is important to recognize that large variants can contribute substantially to the number of bases that any individual is different from the reference. This might not be obvious when performing event-driven benchmarking. For example, the synthetic-diploid variant calls have about 26,000 insertions and deletions of at least 50 bp in size. While this number is small relative to the numbers of SNVs and small indels, the total number of bases inserted and deleted is more than 10^7, greater than the total number of bases changed by the more numerous variants of less than 50 bp in size.

Interestingly, stringency of matching can also significantly influence performance metrics. For example, Table 3 shows how the number of FP indels for the assembly-based fermikit submission is much higher than the Real Time Genomics (RTG) submission when counting genotype errors as FPs, but the number of FPs is lower for fermikit when matching only the allele or performing distance-based matching. Exact genotype matching is most useful for many applications, for example, when interpreting variation in a clinical context, and this is why we developed sophisticated comparison tools to perform this matching. However, there are at least two use cases in which less stringent matching is useful in addition to genotype matching. (1) During methods development and optimization, it is often useful to know if most of the FPs are due to genotype or allele mismatch, so that the developer can focus on improving genotype and/or allele calling rather than on discovery. (2) If a laboratory will manually curate or confirm all detected variants, then calling the wrong variant near the true variant can be better than calling no variant if the true variant is determined upon curation or confirmation. Additional information about relative strengths and weaknesses of the pipelines could also be gained through stratification, as discussed in the next section.

Table 3 | Matching stringency can affect relative performance of algorithms

Callset	GT match FPs	Allele match FPs	Local match
mlin-fermikit	14,514	391	264
ltrigg-rtg2	947	514	350

Number of FPs for two PrecisionFDA Challenge submissions is shown for different matching stringencies, showing that the fermikit submission has many more FPs if genotype errors are counted as FPs, but that it has fewer FPs if matching only the allele or performing distance-based matching. Note that this is intended to illustrate the importance of matching stringency and is probably not indicative of the performance of these methods with optimized parameters or current versions.

Stratification illuminates challenging regions sequenced with and without PCR amplification. Our team has defined a large number of regions of different genome contexts (for example, GC content and repeats of different sizes and types) to enable users to stratify performance and understand strengths and weaknesses of a particular method. As an example of using stratification, we compare recall and precision in different genome contexts for whole-genome-sequencing assays with and without a PCR-amplification step. Table 4 shows that indel recall and precision are lower when using PCR amplification than when using PCR-free sequencing. Stratification highlights that this difference almost entirely results from PCR-related errors in homopolymers and tandem repeats, since performance is similar when excluding variants that occur within 5 bp of homopolymer sequences longer than 5 bp and tandem repeats longer than 10 bp. Performance in regions with low GC content is similar, but PCR results in lower SNV and indel recall in which GC content is greater than 85%.

Further stratification by type of repeat can illuminate particularly challenging genome contexts. For example, when sorting stratified genome contexts by recall, indels in 51–200-bp AT dinucleotide tandem repeats have substantially lower recall and precision than all other stratified genome contexts for both PCR and PCR-free results. Also, 86 out of 114 truth indels in 51–200-bp AT dinucleotide tandem repeats are compound heterozygous, and 89% fall outside the high-confidence regions, so our stratification and benchmarking methods help illuminate that these appear to be highly polymorphic and difficult variants to characterize. As discussed above, recall and precision are likely to be even lower outside the high-confidence regions in these stratified regions.

Discussion

The GA4GH Benchmarking Team has developed a suite of methods to produce standardized performance metrics for benchmarking small germline variant calls. These tools address challenges in standardizing metrics like recall and precision, comparing different representations of variant calls, and stratifying performance by variant

Table 4 | Recall and precision stratified by genomic context (for example, GC content and tandem repeat (TR) type) and variant type for Illumina whole-genome-sequence assays with and without a PCR step compared to GIAB v.3.3.2 truth set in the GIAB high-confidence regions

Genomic context	Type	Recall (with PCR)	Recall (PCR free)	Precision (with PCR)	Precision (PCR free)
All	SNV	98.4 (97.7–99.2)	98.4 (97.6–99.1)	86.0 (83.9–87.9)	86.0 (84.0–88.0)
	Indel	85.9 (83.9–87.9)	97.1 (96.1–98.1)	59.0 (56.7–61.3)	56.3 (54.0–58.6)
Not in homopolymers or TRs	SNV	98.6 (97.9–99.3)	98.5 (97.7–99.2)	87.7 (85.7–89.5)	87.8 (85.9–89.7)
	Indel	98.3 (97.5–99.1)	98.4 (97.6–99.1)	75.4 (73.1–77.7)	75.5 (73.1–77.8)
In homopolymers or TRs	SNV	95.6 (94.3–96.8)	97.2 (96.1–98.2)	61.5 (59.1–63.8)	60.7 (58.3–63.0)
	Indel	78.2 (75.9–80.4)	96.4 (95.2–97.5)	50.3 (48.2–52.5)	48.4 (46.3–50.6)
GC content > 85%	SNV	84.7 (80.6–88.7)	94.4 (91.5–96.7)	50.4 (46–54.7)	49.2 (45.2–53.4)
	Indel	73.2 (64.8–81)	97.3 (93.7–99.6)	27.9 (22.9–33)	27.0 (22.8–31.4)
All 51–200-bp dinucleotide TRs	Indel	45.3 (42.4–48.2)	80.9 (78.6–83.2)	22.3 (20.6–24)	27.6 (26.1–29.1)
51–200-bp AT dinucleotide TRs	Indel	12.3 (6.9–18.7)	28.1 (20.4–36.7)	4.3 (2.4–6.8)	6.8 (4.7–9.2)

95% credible intervals in parentheses are calculated using happyCompare.

type and genome context. We have developed a set of best practices for benchmarking variant calls to help users avoid common pitfalls and misinterpretations of performance metrics, summarized in Supplementary Table 1.

Continually evolving benchmark data and methods are an important part of improving variant-calling performance. The performance metrics from current truth sets are useful for assessing the variants that methods detect similar to variant types and genome contexts present in the truth sets. However, some important types of variants probably have much lower accuracy owing to size and/or repetitive genome context and are outside the current truth set high-confidence regions[25]. New technologies will enable the characterization of increasingly difficult variants and genomic regions, which will require improved benchmarks and may lead to discovery of new variants of clinical and research interest that are currently challenging to detect. Simultaneously, characterization of reference materials may be improved through ongoing work by groups like GIAB.

Opportunities for further development of reference materials and benchmarking tools include structural variants and somatic variants. In addition to the genotype, allele, and local matching stringencies we describe for small variants, comparison tools for structural variants will need to consider stringencies for breakpoint matching, size predictions, and inserted-sequence predictions. Assessment of somatic variants also introduces distinct challenges compared with germline variants (for example, assessing accuracy of variant allele frequency). The ICGC-TCGA DREAM Somatic Mutation Calling (SMC) global consortium has been benchmarking both individual somatic variants and subclonal variation[26].

Moving forward, groups will also need to modify benchmarking strategies to address changes in the way the human genome itself is represented. Today the most common way of representing the human genome involves a set of linear chromosomes (for example, the most common usage of GRCh37). There are key advantages to nonlinear, graph representations of the genome, including ability to characterize large variants and complex variants[27–29]. The GRCh38 build of the human genome takes a step towards this by developing new ALT loci, which provide multiple distinct versions of specific complex regions of the genome[30]. These ALT loci are not fully used by most aligners or the benchmarking tools we describe, and their impact on benchmarking studies is largely unexplored and may require a variety of samples with differing ALT alleles. It is likely that the core representation of the genome will continue to evolve over time, and benchmarking tools will need to adapt to these changes. This work provides a framework for developing benchmarking tools capable of addressing new challenges in variant calling and other high-throughput measurement challenges.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability, and associated accession codes are available at https://doi.org/10.1038/s41587-019-0054-x.

Received: 23 May 2018; Accepted: 10 January 2019;
Published online: 11 March 2019

References

1. Yang, Y. et al. Molecular findings among patients referred for clinical whole-exome sequencing. *J. Am. Med. Assoc.* **312**, 1870–1879 (2014).
2. Xue, Y., Ankala, A., Wilcox, W. R. & Hegde, M. R. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. *Genet. Med.* **17**, 444–451 (2015).
3. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
4. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
5. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
6. Eberle, M. A. et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* **27**, 157–164 (2017).
7. Zook, J. et al. An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-019-0074-6 (2019).
8. Li, H. et al. New synthetic-diploid benchmark for accurate variant calling evaluation. Preprint at bioRxiv https://doi.org/10.1101/223297 (2017).
9. Highnam, G. et al. An analytical framework for optimizing variant discovery from personal genomes. *Nat. Commun.* **6**, 6275 (2015).
10. Cleary, J. G. et al. Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. Preprint at bioRxiv https://doi.org/10.1101/023754 (2015).
11. Sun, C. & Medvedev, P. VarMatch: robust matching of small variant datasets using flexible scoring schemes. *Bioinformatics* **33**, 1301–1308 (2017).
12. Talwalkar, A. et al. SMaSH: a benchmarking toolkit for human genome variant calling. *Bioinformatics* **30**, 2787–2795 (2014).
13. *The Variant Call Format Specification* https://samtools.github.io/hts-specs/VCFv4.3.pdf (2017).
14. Chen, B. et al. *Good Laboratory Practices for Molecular Genetic Testing for Heritable Diseases and Conditions* (Centers for Disease Control and Prevention, 2009).
15. Mattocks, C. J. et al. A standardized framework for the validation and verification of clinical molecular genetic tests. *Eur. J. Hum. Genet.* **18**, 1276–1288 (2010).

16. Gargis, A. S. et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat. Biotechnol.* **30**, 1033–1036 (2012).
17. Rehm, H. L. et al. ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med.* **15**, 733–747 (2013).
18. Aziz, N. et al. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch. Pathol. Lab. Med.* **139**, 481–493 (2015).
19. Roy, S. et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the association for molecular pathology and the college of american pathologists. *J. Mol. Diagn.* **20**, 4–27 (2018).
20. Krusche, P. Haplotype comparison tools / hap.py. http://github.com/illumina/hap.py (2018).
21. Hasan, M. S., Wu, X., Watson, L. T., Li, Z. & Zhang, L. UPS-indel: a universal positioning system for indels. Preprint at bioRxiv https://doi.org/10.1101/133553 (2017).
22. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202–2204 (2015).
23. Kaplanis, J. et al. Exome-wide assessment of the functional impact and pathogenicity of multi-nucleotide mutations. Preprint at bioRxiv https://doi.org/10.1101/258723 (2018).
24. Ball, M. P. et al. A public resource facilitating clinical use of genomes. *Proc. Natl Acad. Sci. USA* **109**, 11920–11927 (2012).
25. Lincoln, S. E. et al. An interlaboratory study of complex variant detection. Preprint at bioRxiv https://doi.org/10.1101/218529 (2017).
26. Ewing, A. D. et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* **12**, 623–630 (2015).
27. Novak, A. M. et al. Genome graphs. Preprint at bioRxiv https://doi.org/10.1101/101378 (2017).
28. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **27**, 665–676 (2017).
29. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
30. Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).

Acknowledgements

We thank GA4GH, especially S. Keenan, D. Lloyd, and R. Nag, for their support in hosting and organizing the Benchmarking Team. We thank the many contributors to Benchmarking Team and GIAB discussions over the past few years, especially D. Church, S. Lincoln, H. Li, A. Talwalker, K. Jacobs, and B. O'Fallon. Certain commercial equipment, instruments, or materials are identified to specify adequate experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the NIST or the Food and Drug Administration, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose.

Author contributions

P.K., L.T., P.C.B., C.E.M., F.M.d.l.V., M.A.E., R.T., B.F., M.F., M.S., and J.M.Z. wrote the manuscript. P.K., L.T., F.M.d.l.V., B.L.M., and M.G.-P. designed and implemented the benchmarking tools. Z.T., S.L., G.A., and J.M.Z. designed and/or analyzed results from the PrecisionFDA Challenges. P.K., L.T., G.A., B.A.C., M.S., and J.M.Z. designed the project. All authors contributed to GA4GH Benchmarking Team discussions about this work.

Competing interests

P.K., B.L.M., M.G., and M.A.E. are employees of, and/or hold stock in, Illumina. R.T. is an employee of, and holds stock in, Invitae. G.A. is an employee of DNAnexus. B.F. is an employee of Veritas Genetics and holds leadership positions in AMP, CLSI, CAP, and ClinGen. L.T. is an employee of Real Time Genomics. C.E.M. is a founder of Onegevity Health and Biotia, Inc.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0054-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.M.Z.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Variant representation. A variety of approaches have been recently developed to address the challenges in variant representation[10–12,20]. RTG developed the comparison tool vcfeval, which introduced the idea of comparing variants at the level of the genomic haplotypes that the variants represent as a way to overcome the problems associated with comparing complex variants, in which alternative yet equivalent variant representations can confound direct comparison methods[10]. Variant 'normalization' tools help to represent variants in a standardized way (for example, by left-shifting indels in repeats), but they demonstrated that 'variant normalization' approaches alone were not able to reconcile different representations of many complex variants. By contrast, global optimization permits evaluation of alternative representations that minimize the number of discrepancies between truth and test sets caused by differences in representations of the same variant. Similarly, VarMatch was developed to resolve alternative representations of complex variants, with additional ability to tune the matching parameters depending on the application[11]. Finally, hap.py includes a comparison tool xcmp to perform haplotype-based comparison of complex variants in addition to sophisticated functionality to stratify variant calls by type or region[20]. We use the hap.py framework with the vcfeval comparison tool in this work. More details about each comparison tool are in Supplementary Note 2.

Variant counting. The GA4GH Benchmarking Team developed consensus definitions and recommendations for expressing performance metrics for small germline-variant calls. Assessing the performance of variant callers does not easily lend itself to the typical binary-classification performance-assessment model of simply determining true and false 'positives' and 'negatives'. Several characteristics of the genome do not fit well in a binary classification model:

(1) More than two possible genotypes exist at any given location. For SNVs (if ignoring phasing), any location can have one of 10 different true genotypes (that is, A/A, A/C, A/G, A/T, C/C, C/G and all other combinations). For indels and complex variants, an infinite number of possible genotypes exists (for example, any length of insertion).

(2) A number of variant callers distinguish between 'no-calls' and homozygous reference calls at some genome positions or regions. Some variant callers even output partial no-calls, calling one allele but not the other. 'No-calls' at a true variant site could be treated as FNs or be excluded from counting.

(3) In addition to the challenges comparing different representations of complex variants (that is, nearby SNVs and/or indels) discussed above, there are challenges in standardizing counting of these variants. Complex variants can be treated as a single positive event or as multiple distinct SNV and indel events when counting the number of TP, FP, and FN variants. In addition, only part of a complex variant may be called, which poses challenges in defining TP, FP, and FN.

(4) Methods for assessing accuracy of phasing have not been fully developed or standardized, but accurate phasing can be critical, particularly when multiple heterozygous variants exist in a small region (for example, complex variants). Supplementary Fig. 3 shows two examples of nearby SNVs in coding regions that could be misinterpreted when not considering phasing.

Matching stringencies. Owing to the inherent complexity of the human genome, TPs, FPs, and FNs can be defined in different ways. Our reference implementation for benchmarking uses a tiered definition of variant matches, a standardized VCF format for outputting matched variant calls, and a common counting and stratification tool (see Supplementary Note 3).

We consider the following types of variant matches from most to least stringent, with 'Genotype match' being used by our current tools to calculate TPs, FPs, and FNs:

- Genotype match: variant sets in truth and query are considered TPs when their unphased genotypes and alleles can be phased to produce a matching pair of haplotype sequences for a diploid genome. Each truth (and query) variant may be replayed onto one of two truth (or query) haplotypes. A maximal subset of variants that is replayed to produce matching haplotype sequences forms the TP variants, query variants outside this set are FPs, truth variants outside this set are FNs. The method only considers haploid or diploid samples but could be extended to higher ploidy. Enumerating the possible assignments for haplotype generation is computationally expensive. vcfeval solves this problem using global optimization methods supplemented with heuristic pruning. Genotype match statistics are the default TP, FP, and FN output by our tools. Genotype matching has been implemented in the hap.py tool xcmp and in vcfeval.
- Allele match: truth and query alleles are counted as TP_AM if they contain any of the same (trimmed and left-shifted) alleles. This method is more specific than local matching (for example, repeat expansions must be called with the correct length to get an allele match), but could also be susceptible to spurious mismatches when truth and query variant alleles are decomposed differently. Genotype mismatches (FP.GT in Table 1) are considered TPs in this matching method. We indicate allele matches in scenarios in which variants can be matched when ignoring the genotype. Allele match statistics (TP_AM, FP_AM, and FN_AM) can be calculated from the GA4GH outputs (which require genotypes to match): TP_AM = QUERY.TP + FP.GT, FP_AM = QUERY.FP-FP.GT, and FN_AM = TRUTH.FN-FP.GT. Allele matching has been implemented in the hap.py tool scmp-somatic and in vcfeval with the --squash-ploidy option.
- Note that vcfeval --squash-ploidy and scmp-somatic differ. scmp-somatic checks if the VCF records give the same alleles after normalization and trimming. This will match alleles that overlap on the reference as long as they can be matched directly after left-shifting and trimming. When comparing somatic variant calls, this is probably the best option since technically, every variant could be on a different (low-frequency) haplotype. vcfeval --squash-ploidy does haplotype-based comparison but assumes all variants are homozygous and there is only one haplotype. This will match different representations unless they overlap on the reference (which is also possible using xcmp via the force-gt command line option in hap.py which changes the GTs before comparing).
- Local match: truth and query variants are counted as TP_LM if their reference span intervals are closer than a pre-defined local matching distance, that is, all yellow categories in Table 1 are considered TPs, including 'F' matches that are within a specified number of basepairs. This approach has previously been implemented[8,20]. An advantage of this matching method is that it is robust towards representational differences. A drawback for many applications is that it does not measure allele or genotype accuracy. We use local matches as the lowest tier of matching to label variants which are close-by but cannot be matched with other methods. Local match statistics (TP_LM, FP_LM, and FN_LM) can be calculated from the GA4GH outputs (which require genotypes to match): TP_LM = QUERY.TP + FP.GT + FP.AL, FP_LM = QUERY.FP-FP.GT-FP.AL, and FN_LM = TRUTH.FN-FP.GT-FP.AL. If only local matching is required, this has been implemented in the hap.py tool scmp-distancebased.
- Phased-genotype match: a fourth, more stringent matching, which is not yet fully implemented in the GA4GH framework, requires phasing information to match. When VCF files specify phasing information, we can compare on a haplotype level: variants will only be matched if they produce matching haplotype sequences under phasing constraints. Both vcfeval and the xcmp method of hap.py support phased matching when both callsets include variants that are globally phased (that is, specify a paternal and maternal haplotype for each chromosome). To our knowledge, no current comparison method supports phasesets and local phasing to compare variants. Moreover, assessing phasing requires us to consider not only phasing-variant accuracy, but also completeness of phasing coverage. In our current methods we do not implement phased-genotype matching beyond the basic support provided by vcfeval and xcmp.

Defining TPs, FPs, and FNs. In Table 1, we enumerate the types of matches that are clear TPs, FPs, and FNs as well as various kinds of partial matches that may be considered TPs, FPs, and/or FNs depending on the matching stringency, and how they are counted by our tools. Our tools calculate TPs, FPs, and FNs requiring the genotype to match, but output additional statistics related to how many of the FPs and FNs are allele matches (FP.GT) or local matches (FP.AL). Note that we have chosen not to include true negatives (or consequently specificity) in our standardized definitions. This is due to the challenge in defining the number of true negatives, particularly around complex variants. In addition, precision is often a more useful metric than specificity owing to the very large proportion of true negative positions in the genome.

Another key question is how to count both matching and mismatching variant calls when they are differently represented in the truth dataset and a query. When representing MNPs as multiple SNVs, we may count one variant call for each SNV, or only one call in total for the MNP record. Similar considerations apply to counting complex records. We approach variant counting as follows:

- We count the truth and query VCF files separately. A set of truth records may be represented by a different set of query records.
- To get comparable recall, we count both TPs and FNs in their truth representation. When comparing different variant-calling results to the same truth set, these counts will be based on the same variant representation.
- Precision is assessed using the query representation of variants. We give a relative precision to the number of truth variants in query representation. If a variant caller is consistent about the way it represents variants, this approach mitigates counting-related performance differences.
- We implement a 'partial credit' mode in which we trim, left-shift and decompose all query variant calls before comparison. This resolves the MNV versus SNV comparison issues and also simplifies the variant types we use for stratification; rather than having a category of complex variant calls, which has results that are difficult to interpret, we account for every atomic indel and SNV call independently.
- Variants are stratified into a canonical set of types and subtypes (see Supplementary Note 4).
- When stratification regions are applied, we match variants by their trimmed reference span. If any part of a deletion overlaps the stratification region, it

is counted as part of that stratum. Insertions receive special treatment by requiring both the base before and the base after to be captured. Importantly, this stratification is performed after comparison to deal appropriately with representation issues.

Note that we have chosen not to include true negatives (or consequently specificity) in our standardized definitions, owing to two key reasons. First, true negatives are challenging to define in whole-genome sequencing, as they will heavily depend on the variant type being evaluated. While SNV true negatives could be defined as one true negative per homozygous reference position, indels and complex variants complicate the definition of true negatives, particularly when in repetitive regions and when the two haplotypes have different indels or complex variants. Second, given the large proportion of homozygous reference positions in the genome, specificity generally is very close to one even for methods with many FPs, which can be misleading.

Benchmarking metrics report. To reconcile the comparison methods and metrics discussed above into a simple summary, we have implemented in hap.py a standardized report that can be generated from the tabular output of the benchmarking workflow[20]. This report displays the metrics we believe are most important in an accessible fashion (Tier 1 metrics), while also allowing to examine the data in more detail (Tier 2 metrics). An example for the metrics and plots displayed in such a report is shown in Supplementary Fig. 1.

From the TP, FP, and FN counts defined in Table 1, we calculate:

METRIC.PRECISION = QUERY.TP / (QUERY.TP + QUERY.FP), METRIC.RECALL = TRUTH.TP / (TRUTH.TP + TRUTH.FN)

We use the count of TPs on the basis of the query representation (QUERY.TP) to calculate precision, and we use the count of TPs on the basis of the truth representation (TRUTH.TP) to calculate recall, to account best for cases in which the truth may tend to split a complex variant into multiple variants and the query may combine them into a single variant, or vice versa. Definitions and formulas for all performance metrics are detailed in Supplementary Table 2

An alternative to precision is false positive rate (FPR) = FP per megabase (ref. [8]). It can easily be obtained from GA4GH/hap.py extended csv by taking FP/$1\times10^6\times$Subset.Size (or Subset.IS_CONF.Size, the number of confident bases in each stratification region). Precision approximates the probability that a given query call is true, whereas FPR approximates the probability of making a spurious call. Note that we do not define 'true negatives' or 'specificity' because these are not cleanly applicable to genome sequencing. For example, there are an infinite number of possible indels in the genome, so there are an infinite number of true negatives for any assay.

In addition, the GA4GH Benchmarking framework is able to produce precision–recall curves, which are graphical plots that illustrate the performance of a variant quality score of a test callset as its discrimination threshold is varied, compared to the reference callset (see Supplementary Fig. 1b). The curve is created by plotting the precision against the recall at various quality score threshold settings. Commonly used quality scores include QUAL, GQ (genotype quality), DP (depth of coverage), and machine-learning derived scores such as VQSLOD and AVR. Because some methods use multiple annotations for filtering, precision–recall curves can be generated for a particular quality score either before or after removing filtered sites. Examining the precision–recall curves for various callsets has two main advantages. Firstly, it allows the user to consider how accuracy is affected through the precision–recall trade-off. Secondly, different callsets may have effectively selected different precision–recall trade-off criteria, so simply comparing full callset metrics may reflect more about the different trade-off points than the callsets themselves at some shared trade-off criteria.

Benchmark callsets. All of the truth sets described below use multiple callers and/or technologies to alleviate bias towards any particular method, and our variant comparison tools are designed to minimize biases against callers that represent variants differently from the truth sets. Nevertheless, subtle biases are introduced in terms of which variants and regions are included in the truth and which are excluded from the high-confidence regions. For example, the current GIAB and Platinum Genomes calls are constructed from short reads and exclude regions difficult to access with short reads, and the current synthetic diploid calls exclude 1-bp indels owing to the relatively high indel error rate of long reads. Accurate reporting of truth set versions, variant types, fraction not assessed, and numbers and size of the confidence regions are also key to the successful interpretation of benchmarking results.

GIAB. The GIAB is a public–private–academic consortium hosted by the NIST to perform authoritative characterization of a small number of human genomes to be used as benchmarks. GIAB published a benchmark set of small variant and reference calls for its pilot genome, NA12878, which characterized a high-confidence genotype for approximately 78% of the bases with sequence information (that is, bases that are not an 'N') in the human genome reference sequence (version GRCh37)[4]. Since this publication, GIAB has further developed integration methods to be more reproducible, comprehensive, and accurate, and has incorporated new technologies and analysis methods. The new integration process has been used to form benchmark small variant and reference calls for approximately 90% of GRCh37 and GRCh38 for NA12878, as well as a mother–father–son trio of Ashkenazi Jewish ancestry and the son in a trio of Chinese ancestry from the Personal Genome Project (v.3.3.2 available at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/)[5,7,24]. The five GIAB-characterized genomes are available as NIST Reference Materials (RMs 8391, 8392, 8393, and 8398), which are extracted DNA from a single, homogenized, large growth of cells for each genome. These samples are also all available as cell lines and DNA from the Coriell Institute for Medical Research. The Personal Genome Project samples are also consented for commercial redistribution[24], and several derived products are commercially available, including FFPE-preserved and in vitro mutated cell lines, or with DNA spike-ins with particular variants of clinical interest. GIAB is continuing to improve the characterization of these genomes to characterize increasingly difficult variants and regions with high-confidence.

Platinum Genomes. In addition to the benchmarking data produced by the GIAB, Platinum Genomes has also created a benchmarking dataset for small variants (SNVs and indels) using the 17-member pedigree (1,463) from Coriell Cell Repositories that includes the GIAB pilot sample NA12878/HG001[6]. Every sample of this pedigree was sequenced to approximately ×50 depth on an Illumina HiSeq2000 system. Variant calls were made from this data using different combinations of aligners and variant callers. This pedigree includes 11 children of the parents (NA12877 and NA12878), producing a fully phased dataset that allows validation of the accuracy of variant calls through genetic-inheritance patterns. The HiSeq2000 sequence data used to create these benchmarking calls can be obtained from the Database of Genotypes and Phenotypes (https://www.ncbi.nlm.nih.gov/gap) under accession number phs001224.v1.p1. Additionally, the sequence data for six of the members of this pedigree are released through the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) under accession number ERP001960. The DNA and cell lines for all samples are available from the Coriell Institute for Medical Research, and DNA from a single, homogeneous batch of NA12878 is also available as NIST Reference Material 8398

Merged Platinum Genomes and GIAB. Since the two resources mentioned above constitute two different methods for generating truth callsets for NA12878, we have merged these into a single and more comprehensive dataset. Such a hybrid truth set can leverage the strengths of each input, namely the diversity of technologies used as input to GIAB and the robust validation-by-inheritance methodology employed by Platinum Genomes.

As a first pass, we have compared the callsets in NA12878 and identified the intersection as well as the ones unique to each (Supplementary Fig. 2). Next, starting from the union, we have used a modified version of the *k*-mer validation algorithm described in ref. [6] to validate the merged calls (Supplementary Note 5). This hybrid benchmark callset includes more total variants than either input set (67,000–333,000 additional SNVs and 85,000–90,000 additional indels), allowing us to assess more of the calls made by any sequencing pipeline without a loss in precision.

This new benchmarking set represents the first step towards a more comprehensive callset that includes both easy to characterize variants and those that occur in difficult parts of the genome. Despite this significant advance, there remain areas for continued improvement, such as adjudication between conflicting calls and the merging of confident regions. We will continue to develop this integration method to further expand the breadth of coverage of this hybrid truth set resource.

Currently, neither Platinum Genomes nor GIAB makes high-confidence calls on chromosome Y or the mitochondrial genome. In addition, GIAB currently has chromosome X calls only for females, but Platinum Genomes has haploid chromosome X calls for the male NA12877 as well. hap.py has an optional preprocessing step to guess male or female from the truth VCF. For male samples it converts haploid 1 GT calls on chromosomes X or Y to 1/1 so that they get compared correctly by xcmp. For vcfeval, haploid 1 GT calls are treated as the same as 1/1, so this conversion is not necessary.

Synthetic diploid. A new synthetic-diploid benchmark callset was created from long-read assemblies of the CHM1 and CHM13 haploid cell lines, to benchmark small variant calls in regions difficult to analyze with short reads or in diploid genomes, which are currently excluded from the GIAB and Platinum Genomes high-confidence regions[8]. Because it is based on long reads, performance metrics are probably less biased toward any short-read sequencing technology or informatics method, and it enables benchmarking in regions that are difficult to map with short reads. However, because it currently contains some errors that were not corrected in the long reads, it requires a less stringent benchmarking methodology similar to the local-match method described below. It also excludes 1 bp indels from performance assessment since long-read assemblies contain 1-bp indel errors, and indels of greater than 50 bp because these are not analyzed. Therefore, it is currently not as useful for assessing accuracy of genotypes or accuracy of the exact sequence change predicted in the REF and ALT fields. When using GA4GH tools requiring genotypes to match, the majority of FPs and FNs may not be errors in the query callset, though work is underway to improve this. Nevertheless, it is likely to be complementary to the GA4GH benchmarking strategy by enabling users to assess accuracy in more difficult regions that GIAB and Platinum Genomes currently exclude from their high-confidence regions. In particular, because the truth set was not developed from short reads, and errors in the truth may be different from errors in short reads, it may better assess relative performance between short-read-based methods, particularly in more difficult genomic regions. A current limitation is that CHM1 and CHM13 cell lines are not available in a public repository.

PrecisionFDA challenges. The PrecisionFDA team held two challenges in 2016, with participants publicly submitting results from various-mapping and variant-calling pipelines. Although both challenges asked participants to analyze short-read whole-genome sequencing datasets, the first challenge used a sample with high-confidence calls already available (HG001/NA12878) and the second challenge used a sample without high-confidence calls yet available (HG002 from GIAB, made available by GIAB upon the close of the challenge).

In the first, 'consistency' challenge, ×30 Illumina whole-genome sequencing of the HG001/NA12878 sample was provided from two different sequencing sites, and the VCF file results from 17 participants were assessed for reproducibility and accuracy against the GIAB v.2.19 Benchmark VCF. It is possible to generate reproducible results without much variability but substantial differences from the truth. Additionally, the pipelines that generated the variant calls could be tuned to HG001, which, in many situations, was used to train or optimize pipelines.

Therefore, in the second, 'truth' challenge, participants were asked to use their pipelines with ×50 Illumina whole-genome sequencing to predict variants from the (at the time) unknown reference sample HG0002/NA24385. Challenge results were compared using two benchmarking comparator tools, RTG Tools vcfeval for the consistency challenge, and vcfeval + hap.py comparison for the truth challenge (more information at https://precision.fda.gov/challenges/). There were 35 entries in the truth challenge and the responses were submitted and ranked according to precision and recall for SNVs and indels versus the GIAB v.3.3.2 high-confidence calls for each genome (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/). This was the first time vcfeval + hap.py GA4GH comparison methodology was applied at scale across the large number of entries submitted by pipeline developers. It helped highlight the utility of the tools and the need for further development and careful interpretation of results. Partly on the basis of feedback from the challenges, an improved benchmarking app 'GA4GH Benchmarking' uploaded by user peter.krusche is now available on PrecisionFDA, and PrecisionFDA plans to integrate this app into their default variant comparison framework.

Statistical methods for credible intervals. We calculated credible intervals for performance metrics within a single replicate using happyCompare (https://github.com/Illumina/happyCompare). In brief, we model variant counts in each genomic subset using a Bayesian β-binomial model, which allows us to obtain credible intervals that account for the total number observations in each locus (that is, larger intervals when the number of observations is small). For example, for recall our model can be formulated as follows:

$$S_i \sim \mathrm{Binom}(N_i, p_i)$$

$$p_i \sim \mathrm{Beta}(\alpha_{0i}, \beta_{0i})$$

where S_i is TRUTH.TP in subset i; N_i is TRUTH.TOTAL in subset i; p_i is recall for subset i; and α_{0i} and β_{0i} are model hyperparameters ($\alpha_{0i} = \beta_{0i} = 0.5$ (Jeffrey's prior)).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

All code for benchmarking developed for this manuscript are linked to from the GA4GH Benchmarking Team GitHub repository at https://github.com/ga4gh/benchmarking-tools. The hap.py benchmarking toolkit is available at https://github.com/Illumina/hap.py.

Data availability

Raw sequence data used in the PrecisionFDA Truth Challenge were previously deposited in the NCBI SRA with the accession codes SRX847862 to SRX848317. Benchmark calls from GIAB used in the PrecisionFDA challenges and in the examples in Tables 3 and 4 are available at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/. VCFs submitted to the PrecisionFDA challenge and benchmarking results are available at https://precision.fda.gov/, where browse access is granted immediately upon requesting account.

natureresearch

Corresponding author(s): Justin M Zook

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☒	□	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	□	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software used for data collection.
Data analysis	All code for benchmarking developed for this manuscript are linked to from the GA4GH Benchmarking Team GitHub repository at https://github.com/ga4gh/benchmarking-tools. The hap.py benchmarking toolkit is available at https://github.com/Illumina/hap.py.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequence data used in the PrecisionFDA Truth Challenge were previously deposited in the NCBI SRA with the accession codes SRX847862 to SRX848317. Benchmark calls from Genome in a Bottle used in the PrecisionFDA challenges and in the examples in Table 3 and Table 4 are available at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/. VCFs submitted to the PrecisionFDA Challenge and benchmarking results are available at https://precision.fda.gov/, where browse access is granted immediately upon requesting account.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Benchmarking results for the millions of variants in the single sample HG002 included as examples
Data exclusions	No data excluded.
Replication	Benchmarking results for the millions of variants in the single sample HG002 included as examples
Randomization	Benchmarking results for the millions of variants in the single sample HG002 included as examples
Blinding	Benchmarking results for the millions of variants in the single sample HG002 included as examples. HG002 high-confidence calls were blinded to PrecisionFDA Challenge participants.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Unique biological materials
☒	☐	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☒	☐	Human research participants

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Coriell NIGMS Cell Line Repository (GM24385, GM24149, GM24143, GM24631, GM12878)
Authentication	Whole genome sequencing and variant calling was performed on all specimens
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used