2018

Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects

Allison A. Regier
Washington University School of Medicine in St. Louis

David E. Larson
Washington University School of Medicine in St. Louis

Ira M. Hall
Washington University School of Medicine in St. Louis

et al

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-06159-4 OPEN

Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects

Allison A. Regier[1], Yossi Farjoun[2], David E. Larson[1], Olga Krasheninina[3], Hyun Min Kang[4], Daniel P. Howrigan[2], Bo-Juen Chen[5,11], Manisha Kher[5], Eric Banks[2], Darren C. Ames[6], Adam C. English[7], Heng Li[2], Jinchuan Xing[8], Yeting Zhang[8], Tara Matise[8], Goncalo R. Abecasis[4], Will Salerno[3], Michael C. Zody[5], Benjamin M. Neale[9,10] & Ira M. Hall[1]

Hundreds of thousands of human whole genome sequencing (WGS) datasets will be generated over the next few years. These data are more valuable in aggregate: joint analysis of genomes from many sources increases sample size and statistical power. A central challenge for joint analysis is that different WGS data processing pipelines cause substantial differences in variant calling in combined datasets, necessitating computationally expensive reprocessing. This approach is no longer tenable given the scale of current studies and data volumes. Here, we define WGS data processing standards that allow different groups to produce functionally equivalent (FE) results, yet still innovate on data processing pipelines. We present initial FE pipelines developed at five genome centers and show that they yield similar variant calling results and produce significantly less variability than sequencing replicates. This work alleviates a key technical bottleneck for genome aggregation and helps lay the foundation for community-wide human genetics studies.

[1] McDonnell Genome Institute, Washington University School of Medicine, St. Louis 63108 MO, USA. [2] Broad Institute of MIT and Harvard, Cambridge 02142 MA, USA. [3] Human Genome Sequencing Center, Baylor College of Medicine, Houston 77030 TX, USA. [4] Department of Biostatistics, University of Michigan, Ann Arbor 48109 MI, USA. [5] New York Genome Center, New York 10013 NY, USA. [6] DNAnexus Inc, Mountain View 94040 CA, USA. [7] Spiral Genetics, Seattle 98104 WA, USA. [8] Department of Genetics, Rutgers University, Piscataway 08854 NJ, USA. [9] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge 02142 MA, USA. [10] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston 02114 MA, USA. [11] Present address: Google, New York 10011 NY, USA. These authors contributed equally: Yossi Farjoun, David E. Larson, Olga Krasheninina, Hyun Min Kang, Daniel P. Howrigan. These authors jointly supervised this work: Goncalo R. Abecasis, Will Salerno, Michael C. Zody, Benjamin M. Neale, Ira M. Hall. Correspondence and requests for materials should be addressed to I.M.H. (email: ihall@wustl.edu)

Over the past few years, a wave of large-scale WGS-based human genetics studies have been launched by various institutes and funding programs worldwide[1–4] aimed at elucidating the genetic basis of a variety of human traits. These projects will generate hundreds of thousands of publicly available deep (>20×) WGS datasets from diverse human populations. Indeed, at the time of writing, >150,000 human genomes have already been sequenced by three NIH programs: NHGRI Centers for Common Disease Genomics[5] (CCDG), NHLBI Trans-Omics for Precision Medicine (TOPMed), and NIMH Whole Genome Sequencing in Psychiatric Disorders[6] (WGSPD). Systematic aggregation and co-analysis of these (and other) genomic datasets will enable increasingly well-powered studies of human traits, population history and genome evolution, and will provide population-scale reference databases that expand upon the groundbreaking efforts of the 1000 Genomes Project[7,8], Haplotype Reference Consortium[9], ExAC[10], and GnomAD[11].

Our ability as a field to harness these collective data to their full analytic potential depends on the availability of high quality variant calls from large populations of individuals. Accurate population-scale variant calling in turn requires joint analysis of all constituent raw data, where different batches have been aligned and processed systematically using compatible methods. Genome aggregation efforts are stymied by the distributed nature of human genetics research, where different groups routinely employ different alignment, data processing, and variant calling methods. Prior exome/genome aggregation efforts have been forced to obtain raw sequence data and re-perform upstream read alignment and data processing steps prior to joint variant calling due to concerns about batch effects introduced by trivial incompatibilities in processing pipelines[10,11]. These upstream steps are computationally expensive—representing as much as ~70% of the cost of basic per-sample WGS data analysis—and having to rerun them is inefficient (Supplementary Fig. 1). This computational burden will be increasingly difficult to bear as data volumes grow over coming years.

To help alleviate this burden and enable future genome aggregation efforts, we have forged a collaboration of major U.S. genome sequencing centers and NIH programs, and collaboratively defined data processing and file format standards to guide ongoing and future sequencing studies. Our approach focuses on the harmonization of upstream steps prior to variant calling, thus reducing trivial variability in core pipeline components while promoting the application of diverse and complementary variant calling methods—an area of much ongoing innovation. The guiding principle is the concept of functional equivalence (FE). We define FE to be a shared property of two pipelines that can be run independently on the same raw WGS data to produce two output files that, upon analysis by the same variant caller(s), produce virtually indistinguishable genome variation maps. A key question, of course, is where to draw the FE threshold. There is no one answer; at minimum, we advise that data processing pipelines should introduce much less variability in a single DNA sample than independent WGS replicates of DNA from the same individual.

Here, we present initial FE pipelines developed at five genome centers and show that they yield similar variant calling results—including single nucleotide (SNV), insertion/deletion (indel) and structural variation (SV)—and produce significantly less variability than sequencing replicates. Residual inter-pipeline variability is concentrated at low quality sites and repetitive genomic regions prone to stochastic effects. This work will enable data sharing and genome aggregation at an unprecedented scale.

Results

Functional equivalence standard. Towards this goal, we defined a set of required and optional data processing steps and file format standards (Fig. 1; see GitHub page[12] for details). We focus here on WGS data analysis, but these guidelines are equally suitable for exome sequencing. These standards are founded in extensive prior work in the area of read alignment[13], sequence data analysis[8,14–18] and compression[14,19], and more broadly in WGS analysis best practices employed at our collective institutes, and worldwide. Notable features of the data processing standard include alignment with BWA-MEM[13], adoption of a standard GRCh38 reference genome with alternate loci[7,20], and improved duplicate marking. File format standards include a 4-bin base quality scheme, CRAM compression[19] and restricted tag usage, which in combination reduced file size >3-fold (from 54 to 17 Gb for a 30× WGS and from 38 to 12 Gb for a 20× WGS). This in turn reduces data storage costs and increases transfer speeds, facilitating data access and sharing.

FE pipelines show less variability than data replicates. We implemented initial versions of these pipelines at each of the five participating centers, including the four CCDGs as well as the TOPMed Informatics Resource Core, and serially tested and modified them based on alignment statistics (Supplementary Table 1) and variant calling results from a 14-genome test set, with data contributed from each center (see Methods). In order to isolate the effects of alignment and read processing on variant calling, we used fixed variant calling software and parameters: GATK[21] for single nucleotide variants (SNVs) and small insertion/deletion (indel) variants, and LUMPY[22] for structural variants (SVs). These 14 datasets have diverse ancestry and are composed of well-studied samples from the 1000 Genomes Project[7], including four independently-sequenced replicates of NA12878 (CEPH) and two replicates of NA19238 (Yoruban). We tested pairwise variability in SNV, indel and SV callsets generated separately from each of the five pipelines, before and after harmonization, as compared to variability between WGS data replicates (Fig. 2). As expected, pipelines used by centers prior to harmonization effort exhibit strong levels of variability, especially among SV callsets. Much of the variability can be attributed to the use of different incarnations of the GRCh37 reference sequence pre-harmonization, underscoring the importance of including a single reference as part of the standard. Most importantly, variability between harmonized pipelines (mean 0.4%, 1.8%, and 1.1% discordant for SNVs, indels, and SVs, respectively) is an order of magnitude lower than between replicate WGS datasets (mean 7.1, 24.0, and 39.9% discordant). Note that absolute levels of discordance are somewhat high in this analysis because we performed per-sample variant calling and included all genomic regions, with minimal variant filtering. All pipelines show similar levels of sensitivity and accuracy based on Genome in a Bottle (GiaB) calls for NA12878[23], although one center is systematically slightly more sensitive and less precise, likely due to a slightly different base quality score recalibration (BQSR) model (Supplementary Fig. 2). The working group decided that this difference was within acceptable limits for applications of the combined data.

Pipeline validation with Mendelian inheritance. We next applied the final pipeline versions to an independent set of 100 genomes comprising 8 trios from the 1000 Genomes Project[7,8] and 19 quads from the Simons Simplex Collection[24], and generated separate 100-genome GATK and LUMPY callsets using data from each of the five pipelines. Considering all five callsets in aggregate, the vast majority of GATK variants (97.2%) are identified in data from all five pipelines, with only 1.74% unique to a single pipeline and 1.02% in various minor subsets. Mean pairwise SNV concordance rates are in the range of 99.0–99.9%

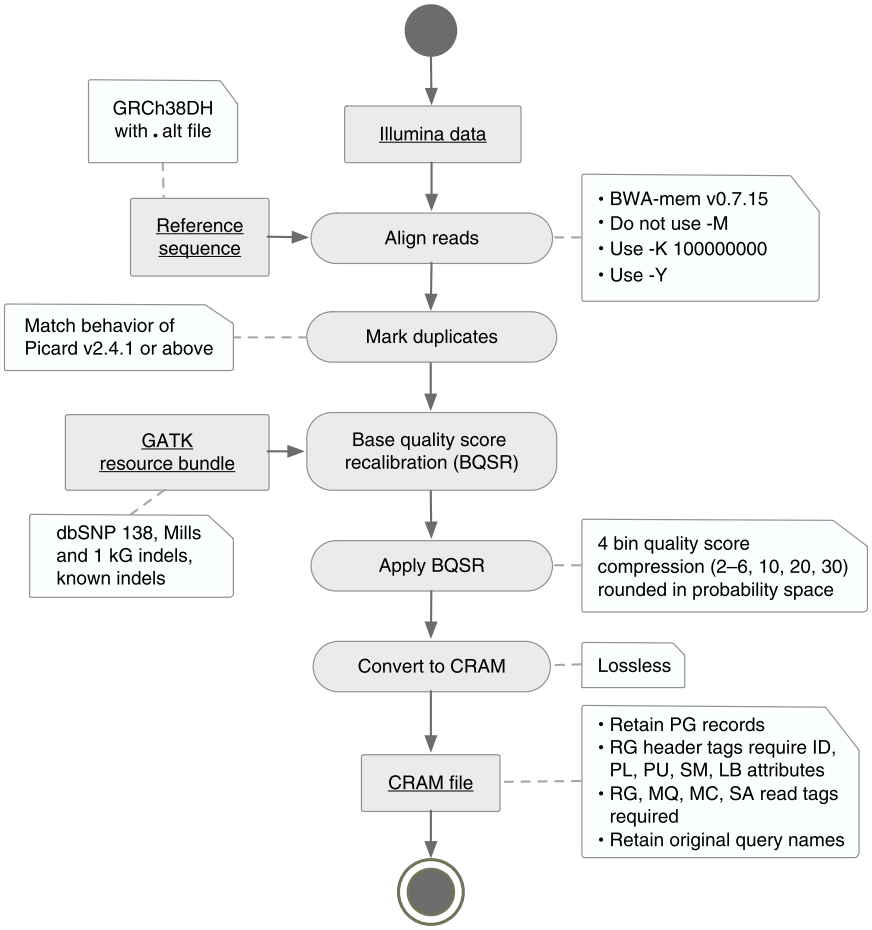

Fig. 1 Highlights of functional equivalence standard. We defined a series of required and allowed processing steps that provide flexibility in pipeline implementation while keeping variation between pipelines at a minimum. Reads must be aligned to a specific reference genome using a minimum version of the BWA-MEM aligner. Algorithms for marking duplicates and recalibrating base quality scores are more flexible and vary somewhat between centers. Compression of quality scores into four bins saves storage and file transfer costs, while maintaining acceptable accuracy and sensitivity

over all sites and comparisons, and Mendelian error rates are ~0.3% at concordant sites, and ~22–24% at discordant sites (Fig. 3). Indel and SV concordance rates are lower—as expected given that these variants are more difficult to map and genotype precisely. Pairwise SNV concordance rates are substantially higher in GiaB high confidence genomic regions comprised predominantly of unique sequence (SNV concordance: 99.7–99.9%; 72% of genome) than in difficult-to-assess regions laden with segmental duplications and high copy repeats (SNV concordance: 92–99%; 8.5% of genome; see Methods). Indeed, 58% of discordant SNV calls are found in the 8.5% most difficult to analyze subset of the genome. Furthermore, the mean quality score of discordant SNV sites are only 0.5% as high as the mean score of concordant SNV sites (16.4% for indels and 90.0% for SVs) (Supplementary Fig. 3). This suggests that many discordant sites are either false positive calls or represent sites that are difficult to measure robustly with current methods. Differences between pipelines are roughly symmetric, with all pipelines achieving similarly low levels of performance at discordant sites, as based on pairwise discordance rates and Mendelian error rates (Supplementary Fig. 4), further suggesting that most discordant calls are due to stochastic effects at sites with borderline levels of evidence. We note that there are some center-specific sources of variability due to residual differences in BQSR models and alignment filtering methods, but that these affect only a trivial fraction of variant calls.

Discussion

Here, we have described a simple yet effective approach for harmonizing data processing pipelines through the concept of functional equivalence. This work resolves a key source of batch effects in sequencing data from different genome centers, and thus alleviates a bottleneck for data sharing and collaborative analysis within and among large-scale human genetics studies. Our approach also facilitates accurate comparison to variant databases; researchers that want to analyze their sample(s) against major datasets such as gnomAD, TOPMed, or CCDG should adopt these standards in order to avoid artifacts caused by non-FE sample processing. The standard is intended to be a living document, and maintaining it in a source control repository provides a natural mechanism for versioning. The standard should be updated to include new data types (e.g., long-reads), file formats and tools, as they become widely adopted in the genomics field and deserving of best-practices status. Additionally, our framework for evaluating FE can be directly used to validate improvements (e.g., new alignment software) and quantify backwards-compatibility with older data. Of course, other challenges remain, such as batch effects from library preparation and sequencing[25], and persistent regulatory hurdles. Nevertheless, we envision that it will be possible to robustly generate increasingly large genome variation maps and shared annotation resources from these and other programs over the next few years, from diverse groups and analysis methods. Ultimately, we hope that

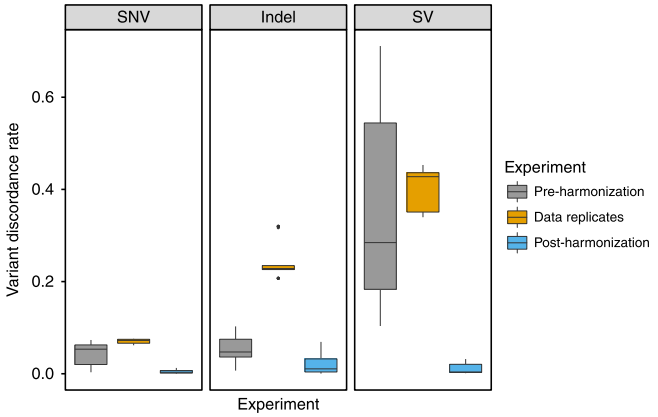

Fig. 2 Pairwise variant discordance rates were calculated between pipelines from each of five centers (pre-harmonization and post-harmonization) as well as between independent sequencing replicates of the same individuals processed by the same pipeline (data replicates). From left, single nucleotide (SNV) and small insertion/deletion (indel) variants were detected with GATK, and structural variants (SV) with LUMPY. The pre-harmonization and post-harmonization comparisons include 14 independently sequenced samples. The data replicate comparisons include four replicates of NA12878 and two replicates of NA19238. Note that the extremely high levels of discordance for SVs pre-harmonization are largely due to variable use of decoy sequences in the reference genomes used by the different centers. The center line is the median, the upper and lower hinges are the first and third quartiles, and the whiskers extend to the largest/smallest values no further than 1.5 * inter-quartile range from the hinge

international efforts such as Global Alliance for Genomics & Health (GA4GH)[26] will adopt and extend these guidelines to help integrate research and medical genomes worldwide.

Methods

Dataset selection. For initial testing, we selected 14 whole genome sequencing datasets based on the following criteria: (1) they include samples of diverse ancestry, including CEPH (NA12878, NA12891, NA12892), Yoruban (NA19238), Luhya (NA19431), and Mexican (NA19648); (2) they were sequenced at multiple different genome centers to deep coverage (>20×) using Illumina HiSeq X technology; (3) they include replicates of multiple samples, including 2 of NA19238 (Yoruban) and 4 of NA12878 (CEPH); (4) they include the extremely well-studied NA12878 genome, for which much ancillary data exists, and (5) they were open access, readily accessible and shareable among the consortium sites. For subsequent characterization of the finalized pipelines, we selected an independent set of 100 samples composed of 8 open-access trios of diverse ancestry from the 1000 Genomes project—including CEPH (NA12878, NA12891, and NA12892), Yoruban (NA19238, NA19239, and NA19240), Southern Han Chinese (HG00512, HG00513, and HG00514), Puerto Rican (HG00731, HG00732, and HG00733), Colombian (HG01350, HG01351, HG01352), Vietnamese (HG02059, HG02060, and HG02061), Gambian (HG02816, HG02817, and HG02818), and Caucasian (NA24143, NA24149, and NA24385)—and 19 quads from the Simons Simplex Collection[24]. The SSC samples were approved for sequencing by the local institutional review board (IRB) at the New York Genome Center (Biomedical Research Alliance of New York [BRANY] IRB File # 17-08-26-385). All relevant ethical regulations were followed.

Downsampling data replicates. To eliminate coverage differences as a contributor to variation between sequencing replicates of the same sample (four replicates of NA12878 and two replicates of NA19238), the data replicates were downsampled to match the lowest coverage sample. To obtain initial coverage, all replicates were aligned to a build 37 reference using speedseq[16] (v 0.1.0). Mean coverage for each BAM file was calculated using the Picard CollectWgsMetrics tool (v2.4.1). For each sample, a downsampling ratio was calculated using the lowest coverage as the numerator and the sample's coverage as the denominator. This ratio was used as the PROBABILITY parameter for the Picard DownsampleSam tool, along with RANDOM_SEED = 1 and STRATEGY = ConstantMemory. The resulting BAM was converted to FASTQ using the script bamtofastq.py from the speedseq repository.

Alignment and data processing pipelines. The pre-harmonization pipeline from the McDonnell Genome Institute at the Washington University School of Medicine aligns reads to the GRCh37-lite reference using speedseq (v0.1.0)[16]. This includes alignment using bwa (v0.7.10-r789)[13], duplicate marking using samblaster (v0.1.22)[15], and sorting using sambamba (v0.5.4)[18].

The post-harmonization pipeline from the McDonnell Genome Institute at the Washington University School of Medicine aligns each read group separately to the GRCh38 reference using bwa-mem (v0.7.15-r1140) with the parameters "-K 100000000 -p -Y". MC and MQ tags are added using samblaster (v0.1.24) with the parameters "-a --addMateTags". Read group BAM files are merged together with "samtools merge" (v1.3.1-2). The resulting file is name-sorted with "sambamba sort -n" (v0.6.4). Duplicates are marked using Picard MarkDuplicates (v2.4.1) with the parameter "ASSUME_SORT_ORDER = queryname", then the results are coordinate sorted using "sambamba sort". A base quality recalibration table is generated using GATK BaseRecalibrator (v3.6) with knownSites files (dbSNP138, Mills and 1 kg indels, and known indels) from the GATK resource bundle (https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0) and parameters "--preserve_qscores_less_than 6 -dfrac .1 -nct 4 -L chr1 -L chr2 -L chr3 -L chr4 -L chr5 -L chr6 -L chr7 -L chr8 -L chr9 -L chr10 -L chr11 -L chr12 -L chr13 -L chr14 -L chr15 -L chr16 -L chr17 -L chr18 -L chr19 -L chr20 -L chr21 -L chr22". The base recalibration table is applied using GATK PrintReads with the parameters "-preserveQ 6 -BQSR "${bqsrt}" -SQQ 10 -SQQ 20 -SQQ 30 --disable_indel_quals". Finally, the output is converted to CRAM using "samtools view".

The pre-harmonization pipeline from the Broad Institute at Harvard and MIT contains the following steps:

-Align with bwa-mem v0.7.7-r441: bwa mem –M –t 10 –p GRCh37.fasta

-Merge aligned bam with the original unaligned bam and sort with Picard 2.8.3: MergeBamAlignment ADD_MATE_CIGAR = true ALIGNER_PROPER_PAIR = false UNMAP_CONTAMINANT_READS = false SORT_ORDER = coordinate

- Mark duplicates with Picard 2.8.3: MarkDuplicates

- Find target indels to fix with GATK 3.4-g3c929b0: CreateRealignerTargets –known dbSnp.138.vcf –known mills.vcf –known 1000genome.vcf

-Fix indel alignments with GATK 3.4-g3c929b0: –known dbSnp.138.vcf –known mills.vcf –known 1000genome.vcf

- Create recalibration table using GATK 3.4-g3c929b0: RecalibrateBaseQuality –knownSites dbSnp.138.vcf using –known dbSnp.138.vcf –known mills.vcf –known 1000genome.vcf

- Apply base recalibration using GATK 3.4-g3c929b0: PrintReads –disable_indel_quals –emit_original_quals

The post-harmonization pipeline from the Broad Institute at Harvard and MIT contains the following steps:

- Align with bwa-mem 0.7.15.r1140: bwa mem -K 100000000 -p -v 3 -t 16 –Y GRCh38.fasta

-Merge aligned bam with the original unaligned bam with Picard 2.16.0: MergeBamAlignment EXPECTED_ORIENTATIONS = FR ATTRIBUTES_TO_RETAIN = X0 ATTRIBUTES_TO_REMOVE = NM ATTRIBUTES_TO_REMOVE = MD REFERENCE_SEQUENCE = ${ref_fasta} PAIRED_RUN = true SORT_ORDER = "unsorted CLIP_ADAPTERS = false MAX_INSERTIONS_OR_DELETIONS = -1 PRIMARY_ALIGNMENT_STRATEGY = MostDistant UNMAPPED_READ_STRATEGY = COPY_TO_TAG ALIGNER_PROPER_PAIR_FLAGS = true UNMAP_CONTAMINANT_READS = true ADD_PG_TAG_TO_READS = false

- Mark duplicates with Picard 2.16.0: MarkDuplicates ASSUME_SORT_ORDER = "queryname"

- Sort with Picard 2.16.0: SortSam SortOrder = coordinate

- Create BQSR table using GATK 4.beta.5: BaseRecalibrator –knownSites dbSnp.138.vcf using –known dbSnp.138.vcf –known mills.vcf –known 1000genome.vcf

- Apply recalibration using GATK 4.beta.5: ApplyBQSR -SQQ 10 -SQQ 20 -SQQ 30

- Convert output to cram with SamTools v 1.3.1: samtools view -C -T GRCh38.fasta

In the HGSC pre-harmonized WGS protocol (https://github.com/HGSC-NGSI/HgV_Protocol_Descriptions/blob/master/hgv_resequencing.md), reads are mapped to the GRCh37d reference with bwa-mem (v0.7.12), samtools (v1.3) fixmate, sorting and duplicate marking with sambamba (v0.5.9), base recalibration and realignment with GATK (v3.4.0), and the quality scores are binned and tags removed with bamUtil squeeze (v1.0.13). Multiplexed samples follow the same steps up through sorting and duplicate marking, resulting in sequencing-event BAMs. The BAMs are merged and duplicates marked using sambamba (v0.5.9), followed by the recalibration, realignment and binning described above.

The HGSC harmonized WGS protocol (https://github.com/HGSC-NGSI/HgV_Protocol_Descriptions/blob/master/hgv_ccdg_resequencing.md) aligns each read group to the GRCh38 reference using bwa-mem (0.7.15) with the parameters "-K 100000000 -Y". MC and MQ tags are added using samblaster (v0.1.24) with the parameters "-a --addMateTags". The resulting file is name-sorted with "sambamba sort -n" (v0.6.4). Duplicates are marked using Picard MarkDuplicates (v2.4.1) with the parameter "ASSUME_SORT_ORDER = queryname", then the results are coordinate-sorted using "sambamba sort". For multiplexed samples, these

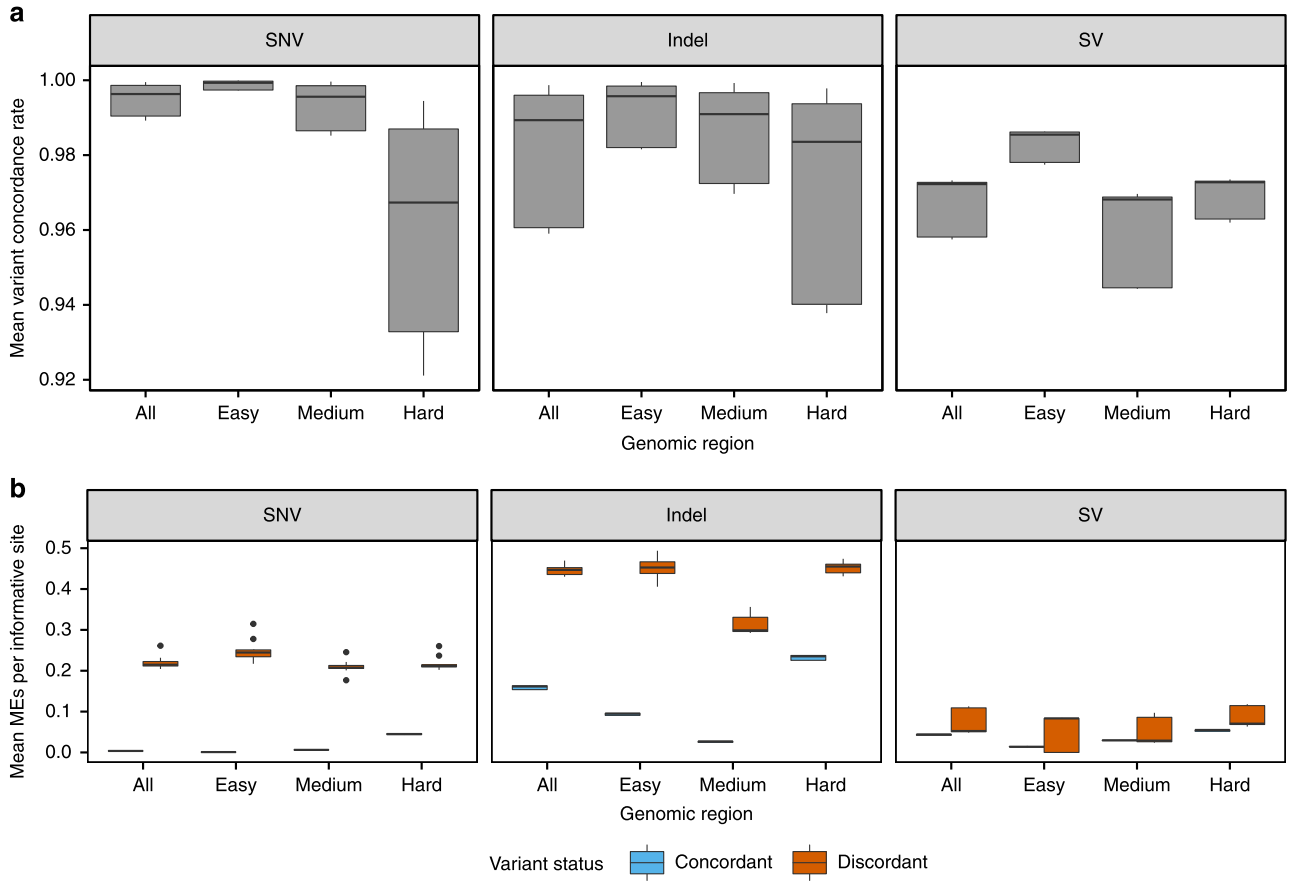

Fig. 3 Variant concordance and Mendelian error (ME) rates were calculated for different variant classes and genomic regions using 100 samples, including 8 trios from the 1000 Genomes Project and 19 quads from the Simons Simplex Collection. **a** Variant concordance rates were calculated from pairwise comparisons across five pipelines for 100 samples. **b** Mendelian error rates were calculated using informative sites in 44 parent-offspring trios, for variants classified as concordant and discordant in pairwise comparisons between five pipelines. The center line is the median, the upper and lower hinges are the first and third quartiles, and the whiskers extend to the largest/smallest values no further than 1.5 * inter-quartile range from the hinge

sequence-event BAMs are then merged with sambamba (v0.6.4) merge, name sorted, duplicate marked and coordinate-sorted with the same tools as above. A base quality recalibration table is generated using GATK BaseRecalibrator (v3.6) with knownSites files (dbSNP138, Mills and 1 kg indels, and known indels) from the GATK resource bundle (https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0) and parameters "--preserve_qscores_less_than 6 -dfrac .1 -nct 4 -L chr1 -L chr2 -L chr3 -L chr4 -L chr5 -L chr6 -L chr7 -L chr8 -L chr9 -L chr10 -L chr11 -L chr12 -L chr13 -L chr14 -L chr15 -L chr16 -L chr17 -L chr18 -L chr19 -L chr20 -L chr21 -L chr22". The base recalibration table is applied using GATK PrintReads with the parameters "-preserveQ 6 -BQSR "${bqsrt}" -SQQ 10 -SQQ 20 -SQQ 30 --disable_indel_quals". Finally, the output is converted to CRAM using 'samtools view'.

The pre-harmonization pipeline from the New York Genome Center aligns each read group separately to the Thousand Genomes version of build 37 reference sequence using bwa mem –M (v0.7.8). The aligned files are merged using Picard MergeSamFiles (v1.83), and duplicates are marked using Picard MarkDuplicates (v1.83). Indel realignment and base quality recalibration are both performed using the GATK (v3.4-0) commands RealignerTargetCreator, IndelRealigner, BaseRecalibrator, and PrintReads.

The post-harmonization pipeline from the New York Genome Center aligns each read group separately to the GRCh38 reference using bwa-mem (v0.7.15) with the parameters "-Y -K 100000000'. Picard (v2.4.1) FixMateInformation is run with the parameter 'FixMateInformation = TRUE". Read group BAM files are merged together with Picard MergeSamFiles (v2.4.1) and the parameter "SORT_ORDER = queryname". Duplicates are marked using Picard MarkDuplicates (v2.4.1), then the results are coordinate sorted using Picard SortSam (v2.4.1) with the parameter "SORT_ORDER = coordinate". A base quality recalibration table is generated using GATK BaseRecalibrator (v3.5) with knownSites files (dbSNP138, Mills and 1 kg indels, and known indels) from the GATK resource bundle (https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0) and parameters "--preserve_qscores_less_than 6 -L grch38.autosomes. intervals". The base recalibration table is applied using GATK PrintReads with the parameters "-preserveQ 6 -SQQ 10 -SQQ 20 -SQQ 30". Finally, the output is converted to CRAM using "samtools view -C" (v1.3.1).

The pre-harmonization pipeline from the TOPMED Informatics Resource Center at the University of Michigan aligns reads using default options in the GotCloud alignment pipeline[17] available at https://github.com/statgen/gotcloud. It aligns the sequence reads to GRCh37 reference with decoy sequences used in 1000 Genomes. The raw sequence was aligned using bwa mem (v0.7.13-r1126)[13], and sorted by samtools (v1.3.1). The duplicate marking and base quality recalibration were performed jointly using bamUtil dedup [ref—same as GotCloud] (v1.0.14).

The post-harmonization pipeline procedure from the TOPMED Informatics Resource Center at the University of Michigan (described in https://github.com/statgen/docker-alignment) first aligns each read group to the GRCh38 reference using bwa-mem (v0.7.15-r1140) with the parameters "-K 100000000 -Y -R [read_group_id]". To add MC and MQ tags, samblaster (v0.1.24) was used with the parameters "-a --addMateTags". Each BAM file corresponding to a read group is sorted by genomic coordinate using "samtools sort" (v1.3.1), and merged together using "samtools merge" (v1.3.1). Duplicate marking and base quality recalibration were performed jointly using bamUtil dedup_lowmem (v1.0.14) with parameters "--allReadNames –binCustom –binQualS 0:2,3:3,4:4,5:5,6:6,7:10,13:20,23:30,33:40 --recab --refFile [reference_fasta_file] --dbsnp [dbsnp_b142_vcf_file] --in [input_bam] –out -.ubam" and the piped output (in uncompressed BAM format) is converted into a CRAM file using samtools view.

Calculation of alignment statistics. A total of 184 alignment statistics were generated for all standardized CRAM files from each center with AlignStats software. Results include metrics for both the entire CRAM file and for the subset of read-pairs with at least one read mapping to the autosome or sex chromosomes. We examined all metrics across the five CRAMs for each of the 15 samples to ensure that any differences were consistent with the various options allowed in the functional equivalence specification. Supplementary Table 1 provides examples of

these metrics, and full description of all metrics can be found online (https://github.com/jfarek/alignstats).

Variant calling for the 14-sample analysis. SNPs and indels were called for each center's CRAM/BAM files using GATK[21] version 3.5-0-g36282e4 HaplotypeCaller with the following parameters:

-rf BadCigar
--genotyping_mode DISCOVERY
--standard_min_confidence_threshold_for_calling 30
--standard_min_confidence_threshold_for_emitting 0

For the pre-standardization files, the 1000 genomes phase 3 reference sequence from the GATK reference bundle ftp://ftp.broadinstitute.org/pub/svtoolkit/reference_metadata_bundles/1000G_phase3_25Jan2015.tar.gz was used. For the post-standardization files, the 1000 Genomes Project version of GRCh38DH (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/) was used.

Structural variants (SVs) were called for each center's CRAM/BAM files using lumpy[22] and svtools (https://github.com/hall-lab/svtools). First, split reads and reads with discordant insert sizes or orientations were extracted from the CRAM/BAM files using extract-sv-reads in the docker image halllab/extract-sv-reads@sha256:192090f72afaeaaafa104d50890b2fc23935c8dc98988a9b5c80dd-f4ec50f70c using the following parameters:

--input-threads 4 -e –r

Next, SV calls were made using lumpyexpress (https://github.com/arq5x/lumpy-sv) from the docker image halllab/lumpy@sha256:59ce7551307a54087e57d5cec89b17511d910d1fe9-fa3651c12357f0594dcb07 with the -P parameter as well as -x to exclude regions contained in the BED file exclude.cnvnator_100bp.GRCh38.20170403.bed (exclude.cnvnator_100bp.112015.bed for pre-standardization samples). Both exclude files are available in https://github.com/hall-lab/speedseq/tree/master/annotations

Finally, the SV calls were genotyped using svtyper from the docker image halllab/svtyper@sha256:21d757e77dfc52fddeab94acd66b09a561771a7803f9581b8c-ca3467ab7ff94a

Defining genomic regions. The reference genome sequence is not uniformly amenable to analysis—some regions with high amounts of repetitive sequence are difficult to align and prone to misleading analyses, while other regions comprised of mostly unique sequence can be more confidently interpreted. To gain a better understanding of how pipeline concordance differs by region, we divided the reference sequence into three broad categories. The easy genomic regions consist of the GiaB gold standard high confidence regions, lifted over to build 38. The hard regions consist of centromeres (https://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/data/38/Modeled_regions_for_GRCh38.tsv), microsatellite repeats (satellite entries from http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.out.gz), low complexity regions (https://github.com/lh3/varcmp/raw/master/scripts/LCR-hs38.bed.gz), and windows determined to have high copy number (more than 12 copies per genome across 409 samples). Any regions overlapping GiaB high confidence regions are removed from the set of hard regions. All remaining regions are classified as medium.

Cross-center variant comparisons for the 14-sample analysis. The VCF files produced by GATK for both the pre- and post-standardization experiments were compared using hap.py from the docker image pkrusche/hap.py:v0.3.9 using the --preprocess-truth parameter.

The four data replicates of NA12878 were compared to the NA12878 gold standards in the regions defined by to obtain sensitivity and precision measurements. The post-standardization VCFs were first lifted over to GRCh37 using the Picard LiftoverVcf tool (v2.9.0) and the chain files hg38ToHg19.over.chain.gz and hg19ToGRCh37.over.chain.gz downloaded from here: http://crossmap.sourceforge.net/#chain-file. To reduce artifacts from the liftover that negatively impacted sensitivity, the gold standard files were lifted over to the build 38 reference and back to build 37, excluding any variants that didn't lift over in both directions.

Values for sensitivity (METRIC.Recall) and precision (METRIC.Precision) were parsed out of the *.summary.csv file produced by hap.py for each comparison, using only variants with the PASS filter value set.

The downsampled data replicates of NA12878 and NA19238 aligned by the same center were compared to each other in a pairwise fashion. Pairwise comparisons between centers were performed for each non-downsampled aligned file. The variant discordance rates between pairs were calculated using the true positive, true negative, and false positive counts from the *.extended.csv output file from hap.py (TRUTH.FN + QUERY.FP)/(TRUTH.TP + TRUTH.FN + QUERY.FP). The rates reported are only for PASS variants but across the whole genome.

The VCF files of SVs produced by lumpy and svtyper were converted to BEDPE using the command "svtools vcftobedpe" from the docker container halllab/svtools@sha256:f2f3f9c788beb613bc26c858f897694cd6eaab450880c370bf0ef81d85bf8d45 The coordinates are padded with 1 bp on each side to be compatible with bedtools pairtopair. The pairwise comparisons are performed using the bedtools pairtopair command (version 2.23.0), then summarized using a python script (compare_single_sample_based_on_strand.py in https://github.com/CCDG/Pipeline-Standardization). The variant discordance rates between pairs are calculated with the following formula: (discordant + 0-only + 1-only + discordant_discordant_type)/(match + discordant + match_discordant_type + discordant_discordant_type + 0-only + 1-only).

Variant calling for 100-sample analysis. SNPs and indels were called using the GATK best practices pipeline, including per-sample variant discovery using HaplotypeCaller with the following parameters:

"-ERC GVCF -GQB 5 -GQB 20 -GQB 60 -variant_index_type LINEAR -variant_index_parameter 128000". Next, GVCFs from all 100 samples were merged with GATK CombineGVCFs. Genotypes were refined with GATK GenotypeGVCFs with the following parameters: "-stand_call_conf 30 -stand_emit_conf 0". Variants with no genotyped allele in any sample are removed with the GATK command SelectVariants and the parameter "--removeUnusedAlternates", and variant lines where the only remaining allele is a symbolic deletion (*:DEL) are also removed using grep.

SVs were called using the svtools best practices pipeline (https://github.com/hall-lab/svtools/blob/master/Tutorial.md). First, per-sample SV calls were generated with extract-sv-reads, lumpyexpress, and svtyper using the same versions and parameters as the 14 sample analysis. Next, the calls were merged into 100-sample callsets for each pipeline using the following sequence of commands and parameters from the docker container halllab/svtools@sha256:f2f3f9c788beb613bc26c858f897694cd6eaab450880c370bf0ef81d85bf8d45

svtools lsort
svtools lmerge -f 20
create_coordinates

The merged calls were then re-genotyped for each sample using the previous svtyper command. Copy number histograms were generated for each sample using the command cnvnator_wrapper.py with window size 100 (-w 100) in the docker container halllab/cnvnator@sha256:c41e9ce51183fc388ef39484cbb218f7ec2351876e5eda18b709d82b7e8af3a2. Each SV call was annotated with its copy number from the histogram file using the command "svtools copynumber" in that same docker container with the parameters "-w 100 -c coordinates". Finally, the per-sample genotyped and annotated VCFs were merged back together and refined with the following sequence of commands in the svtools docker container:

svtools vcfpaste
svtools afreq
svtools vcftobedpe
svtools bedpesort
svtools prune -s -d 100 -e "AF"
svtools bedpetovcf
svtools classify -a repeatMasker.recent.lt200millidiv.LINE_SINE_SVA.GRCh38.sorted.bed.gz -m large_sample

Cross-center variant comparisons for the 100-sample analysis. The VCF of SNPs and indels was split into per-sample VCFs using the command "bcftools view" with the following parameters: "-a -c 1:nref". Additionally, any remaining variant lines with only the symbolic allele (*) remaining were removed. Pairwise comparisons between the same sample processed by different pipelines were performed using hap.py using the same commands as the 14 sample analysis. Variant concordance rates per sample were calculated using results from the extended.csv output file produced by hap.py the following formula: TRUTH.TP/(TRUTH.TP + TRUTH.FN + QUERY.FP). The reported statistics were calculated using all variants genome-wide except those that were marked LowQual by GATK. No VQSR-based filtering was used. Figure 3a reports the mean rates across all 100 samples for each pairwise comparison of pipelines.

The per-pipeline SV VCFs were converted to BEDPE using the command "svtools vcftobedpe" in the docker container halllab/svtools@sha256:f2f3f9c788beb613bc26c858f897694cd6eaab450880c370bf0ef81d85bf8d45. The variants were compared using bedtools pairtopair as in the 14 sample analysis. Next they were classified into hard, medium, and easy genomic regions by intersecting each breakpoint with BED files describing the regions using "bedtools pairtobed". Variants were classified by the most difficult region that either of their breakpoints overlapped (see compare_round3_by_region.sh in https://github.com/CCDG/Pipeline-Standardization). Then, the variants were extracted and annotated in per-sample BEDPE files with the script compare_based_on_strand_output_bedpe.py (in https://github.com/CCDG/Pipeline-Standardization). The BEDPE files were converted to VCF using "svtools bedpetovcf" and sorted using "svtools vcfsort". The number of shared and pipeline-unique variants were counted using "bcftools query" (version 1.6) to extract the genomic region and concordance status of each variant, then summarized with 'bedtools groupby' (v2.23.0). The rates of shared variants per sample were calculated using the output of this file with the following formula: match/(match + 0-only + 1-only).

Mendelian error (ME) rate calculation. SNPs and indels that were classified by hap.py into categories (shared between pipelines, or unique to one pipeline) were further characterized by looking at the ME rate for each of the offspring in the trios/quads. For each offspring in the sample set, the parents and offspring sample VCFs output by hap.py were merged together using "bcftools merge --force-samples" (v1.3), and the genotypes from the first pipeline in the pair were extracted. Any variants with missing genotypes or uniformly homozygous genotypes were excluded using "bcftools view -g ^miss" and "bcftools view -g het". A custom python script (classify_mie.py in https://github.com/CCDG/Pipeline-Standardization) was used to classify each variant as uninformative, informative with no Mendelian error, or informative with Mendelian error. Total informative error and non-error sites in each genomic region were counted for shared sites and unique sites separately, and ME rate was calculated by dividing the number of ME sites by the total number of informative sites. A similar calculation was performed for the per-sample SV VCFs produced by the SV concordance calculations. Fig. 3b and Supplementary Fig. 4 report the mean ME rate across 44 offspring-parent trios for each pairwise pipeline comparison.

Variant quality evaluation. To evaluate possible causes of remaining differences between pipelines, we extracted variant quality scores for each variant type and summarized them by concordance status in each pairwise pipeline comparison across 100 samples. For SNPs and indels, the QUAL field was extracted along with the concordance annotation from the per-sample hap.py comparison VCFs using "bcftools query" (version 1.6). The median QUAL score for each category was reported using "bedtools groupby". For SVs, MSQ (mean sample quality) is a more informative measure of variant quality, so this field was extracted and summarized in a similar way.

Cost calculations. To calculate the fraction of per-sample pipeline cost attributed to upstream steps, the Broad Institute production tables were queried for total workflow cost and HaplotypeCaller cost. The upstream cost was calculated as the difference between the two. All successful pipeline runs that didn't use call caching from October 31, 2017 to May 9, 2018 were included, totaling 13,704 pipeline runs on 13,295 distinct samples.

Code availability. All custom scripts used for the analysis are available under an MIT license at https://github.com/CCDG/Pipeline-Standardization/tree/master/scripts.

Data availability

The 14 input WGS data sets (10 original data sets and 4 downsampled data sets) used in the initial development of the pipeline are available in the SRA under the BioProject PRJNA393319. Files in unaligned BAM format as well as CRAM as aligned by all five centers are available via the Download tab on the RunBrowser pages (for testing additional pipelines for functional equivalence). The WGS data from 19 Simon Simplex Collection quad families (accession SFARI_SSC_WGS_P, family codes 11026, 11063, 11069, 11505, 11671, 12083, 12121, 12202, 12261, 12405, 12480, 13226, 13540, 13556, 13567, 13888, 13996, 14497, 14509) are available upon approved application from SFARI Base. The WGS data from the 8 trios are available in the SRA under the BioProject PRJNA477862.

Received: 21 May 2018 Accepted: 16 August 2018
Published online: 02 October 2018

References

1. Consortium, U. K. et al. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
2. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).
3. Caulfield, M. et al. The 100,000 Genomes Project Protocol. *figshare* https://doi.org/10.6084/m9.figshare.4530893.v2 (2017).
4. Alliance Aviesan. *Genomic Medicine France 2025* (Aviesan, 2017).
5. Felsenfeld, A. Centers for Common Disease Genomics. *National Human Genome Research Institute*, https://www.genome.gov/27563570 (2016).
6. Sanders, S. J. et al. Whole genome sequencing in psychiatric disorders: the WGSPD consortium. *Nat. Neurosci.* **20**, 1661–1668 (2017).
7. 1000 Genomes Project Consortium. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
8. 1000 Genomes Project Consortium. et al. An integrated map of genetic variation from 1092 human genomes. *Nature* **491**, 56–65 (2012).
9. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
10. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
11. Karczewski, K. J. & Francioli, L. The Genome Aggregation Database (gnomAD). *MacArthur Lab*, https://macarthurlab.org/2017/02/27/the-genome-aggregation-database-gnomad/ (2017).
12. Regier, A. P. Pipeline-Standardization. *GitHub*, https://github.com/CCDG/Pipeline-Standardization/blob/master/PipelineStandard.md (2017).
13. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at http://arxiv.org/abs/1303.3997 (2013).
14. Li, H. et al. The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
15. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
16. Chiang, C. et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).
17. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* **25**, 918–925 (2015).
18. Tarasov, A. et al. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032-2034 (2015).
19. Hsi-Yang Fritz, M., Leinonen, R., Cochrane, G. & Birney, E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* **21**, 734–740 (2011).
20. Church, D. M. et al. Extending reference assembly models. *Genome Biol.* **16**, 13 (2015).
21. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
22. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
23. 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).
24. Turner, T. N. et al. Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710–722 (2017).
25. Leek, J. T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
26. Global Alliance for, G. & Health. Genomics. A federated ecosystem for sharing genomic, clinical data. Science **352**, 1278–1280 (2016).

Acknowledgements

We thank NHGRI and NHLBI program staff for supporting this effort, and Jose M. Soto for calculating pipeline costs. This work was funded by NHGRI CCDG awards to Washington University in St. Louis (UM1 HG008853), Broad Institute of MIT and Harvard (UM1 HG008895), Baylor College of Medicine (UM1 HG008898), and the New York Genome Center (UM1 HG008901), the NHGRI GSP coordinating center (U24 HG008956), and an NHLBI TOPMed Informatics Research Center award to the University of Michigan (3R01HL-117626-02S1) as well as grants to B.N. (U01 HG00908, R01 MH107649), H.K (1 R21 HL133758-01, 1 U01 HL137182-01) and G.A. (4 R01 HL117626-04). The following DNA samples were obtained from the NHGRI Sample Repository for Human Genetic Research at the Coriell Institute for Medical Research: NA12878, NA12891, NA12892, NA19238, NA19431, NA19648, HG00512, HG00513, HG00514, HG00731, HG00732, HG00733, NA19239, NA19240, HG01350, HG01351, HG01352, HG02059, HG02060, HG02061, HG02816, HG02817, HG02818, NA24143, NA24149, NA24385.

Author contributions

I.H., B.N. and G.A. conceived the approach and designed the study with M.Z. and W.S. Y.F., D.L., H.K., A.R., D.H., T.M., W.S., M.Z. and I.H. developed and tested the FE standard. Y.F., D.L., A.R., O.K., H.K., B.C., M.K., E.B., D.A., A.E., G.A., W.S., M.Z. and I.H. developed the center-specific pipelines. A.R., Y.F., D.L., O.K., H.K., D.H., B.C., M.K., E.B., J.X. and Y.Z. performed the data analysis. J.X., Y.Z. and T.M. reviewed and improved the standards document and supported the logistics of data sharing. H.L. provided important updates to software and intellectual guidance. I.H. and A.R. led manuscript preparation with contributions from E.B., B.N., W.S., O.K., D.H., M.Z., H.K., Y.F., G.A. and D.L.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-06159-4.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

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