SpeedSeq: ultra-fast personal genome analysis and interpretation

Colby Chiang[1,2], Ryan M Layer[3,4], Gregory G Faust[2], Michael R Lindberg[2], David B Rose[2], Erik P Garrison[5], Gabor T Marth[3,4], Aaron R Quinlan[3,4] & Ira M Hall[1,6]

SpeedSeq is an open-source genome analysis platform that accomplishes alignment, variant detection and functional annotation of a 50× human genome in 13 h on a low-cost server and alleviates a bioinformatics bottleneck that typically demands weeks of computation with extensive hands-on expert involvement. SpeedSeq offers performance competitive with or superior to current methods for detecting germline and somatic single-nucleotide variants, structural variants, insertions and deletions, and it includes novel functionality for streamlined interpretation.

Technical advances in second-generation DNA sequencing technologies have reduced both the cost and the time required to generate whole-genome sequencing (WGS) data, thereby creating opportunities in healthcare and academic research to survey the human genome with unprecedented depth and scope. However, bottlenecks in computational processing and variant interpretation have hindered the adoption of these technologies for time-sensitive and large-scale projects. A standard pipeline using the Burrow-Wheeler Aligner (BWA)[1], the Genome Analysis Toolkit (GATK)[2], the Sequence Alignment-Map tools (SAMtools)[3] and the Picard set of tools requires 60–70 h to process a 50× human genome from raw sequence data to variant calls on a 32-thread server (**Supplementary Note 1**). Furthermore, distinguishing pathogenic from benign mutations is a labor-intensive process that can take hours or days of manual curation per patient[4].

SpeedSeq is an open-source software suite designed for rapid whole-genome variant detection and interpretation (https://github.com/hall-lab/speedseq and **Supplementary Software**). Its modular architecture and universal formats confer adaptability to a variety of experimental designs and compatibility with standard industry software (**Fig. 1a**). It achieves superior processing efficiency through rapid duplicate marking with SAMBLASTER[5], through balanced parallelization of external applications and by executing nondependent pipeline components simultaneously. SpeedSeq translates raw 50× WGS data into prioritized single-nucleotide variants (SNVs), short insertions and deletions (indels) and structural variants (SVs) in 13 h on a single 32-thread server with 128 GB of RAM and a current cost of <$10,000. Moreover, our accelerated implementations show little to no difference in results compared to the original software (**Supplementary Note 1**). This represents, at a minimum, a several-fold speed increase over current practices using typical computing resources.

We assessed the accuracy of SpeedSeq's SNV and indel calls against the Genome in a Bottle Consortium (GIAB) truth set derived from the well-studied human sample NA12878 (2,803,144 SNVs and 364,031 indels)[6]. SpeedSeq achieved sensitivities of 99.9% and 89.9% for germline SNVs and indels, respectively, with acceptably low false discovery rates (FDRs) (0.4% and 1.1%, respectively) (**Fig. 1b,c**). These detection rates exceeded those of GATK's UnifiedGenotyper (GATK-UG) tool (SNVs: 99.7%, indels: 89.0%) with similar FDRs (SNVs: 0.5%, indels: 1.0%). The GATK HaplotypeCaller tool (GATK-HC) showed superior indel detection sensitivity (SNVs: 99.8%, indels: 95.7%) with lower FDRs for both variant types (SNVs: 0.2%, indels: 0.6%). SpeedSeq's implementation of FreeBayes therefore exhibits comparable, albeit slightly inferior, performance to GATK-HC when tested on the GIAB call set[7]. However, the GIAB truth set is biased toward GATK because it was primarily derived from GATK-based analyses. We therefore assessed SpeedSeq's performance against an unbiased truth set of 689,788 SNVs at 2,177,040 sites (Illumina Omni 2.5) in which SpeedSeq attained the highest sensitivity at the minor expense of specificity as compared to results obtained with GATK-UG or GATK-HC (**Supplementary Fig. 1**). Miscalled variants were enriched in repetitive regions of the genome and in regions adjacent to assembly gaps (**Supplementary Note 2** and **Supplementary Table 1**). SpeedSeq also supports joint multi-sample variant calling and *de novo* germline mutation detection in families (**Supplementary Note 3**), which is crucial for clinical applications such as rapid diagnosis in newborns[8].

Cancer genome analysis is a common WGS application in research and clinical environments, and it can be a time-sensitive component of patient care. To emulate a WGS data set from a heterogeneous tumor-normal pair, we defined NA12877 as the 'normal' sample and pooled raw data from his 11 children in equal proportions to generate a single 50× 'tumor' sample. The 875,206 SNVs present in the mother (NA12878) but absent from the father (NA12877) were defined as somatic mutations, with variant-allele frequencies (VAFs) ranging from 0.05 to 0.5 (**Supplementary Fig. 2a**). Using this evaluation paradigm, we

[1]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, Missouri, USA. [2]Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, Virginia, USA. [3]Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah, USA. [4]Utah Science Technology and Research (USTAR) Center for Genetic Discovery, University of Utah School of Medicine, Salt Lake City, Utah, USA. [5]Wellcome Trust Sanger Institute, Hinxton, UK. [6]Department of Medicine, Washington University School of Medicine, St. Louis, Missouri, USA. Correspondence should be addressed to I.M.H. (ihall@genome.wustl.edu).

RECEIVED 12 NOVEMBER 2014; ACCEPTED 28 MAY 2015; PUBLISHED ONLINE 10 AUGUST 2015; DOI:10.1038/NMETH.3505

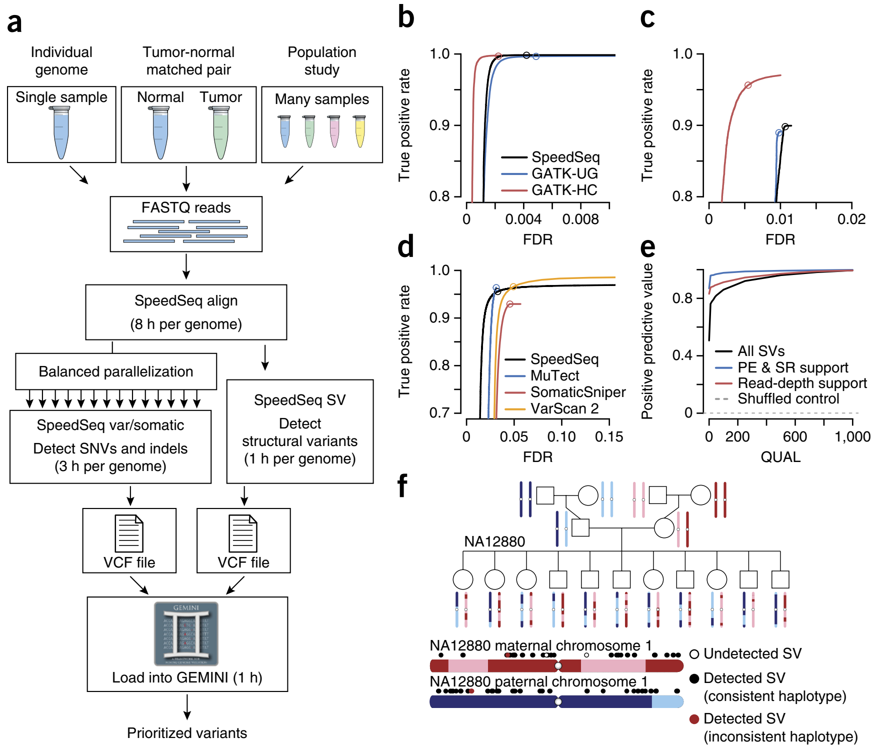


Figure 1 | SpeedSeq workflow. (**a**) SpeedSeq converts raw reads into prioritized variants in 13 h for a 50× human data set. Var/somatic, SpeedSeq var and SpeedSeq somatic software modules for germline and somatic SNP-indel calling, respectively. (**b**,**c**) Germline SNV (*N* = 2,803,144) and indel (*N* = 364,031) receiver-operating characteristic (ROC) curves (**c**) over the GIAB truth set for SpeedSeq, GATK-UG and GATK-HC. (**d**) Somatic SNV detection ROC curves for a simulated 50× tumor-normal pair using SpeedSeq and three other tools (*N* = 875,206). Open circles in **b**–**d** denote the data points reported in the main text. (**e**) SpeedSeq's SV detection performance by quality score (QUAL) of all SVs (black), those with split-read and paired-end support (blue) and those with read-depth support from CNVnator (red), as validated by either PacBio or Moleculo long reads or 1KGP. (**f**) Schematic of haplotype-based SV validation showing undetected (open circles), consistently segregating (black circles) and inconsistently segregating (red circles) SVs through the CEPH 1463 pedigree.

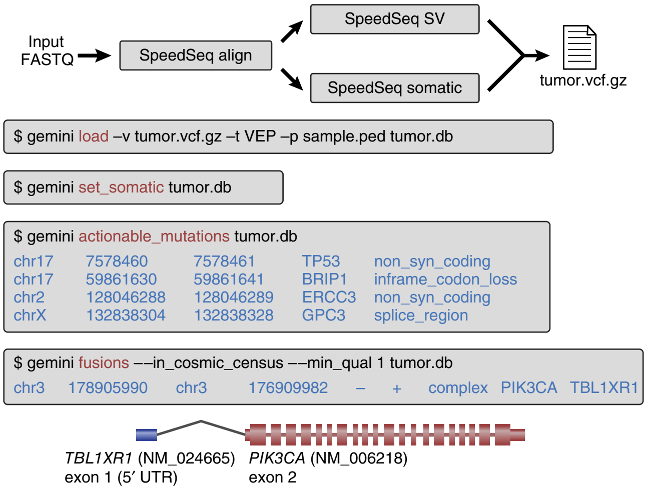
compared SpeedSeq's performance to three other leading somatic-variant calling tools: MuTect[9], SomaticSniper[10] and VarScan 2 (ref. 11). SpeedSeq recalled 96.6% of the somatic variants in the 'tumor' with an FDR of 3.3%, outperforming SomaticSniper in both sensitivity and specificity and delivering competitive performance against MuTect and VarScan 2 (**Fig. 1d** and **Supplementary Fig. 2b,c**).

To test SpeedSeq's performance on real cancer data, we obtained WGS reads (50× tumor, 30× normal) from five tumor-normal pairs with validated somatic mutations, as ascertained by deep exome sequencing, from The Cancer Genome Atlas (TCGA). SpeedSeq recalled 96.4% of the 2,746 orthogonally validated mutations across all five data sets, including 98.8% of mutations in genes that have been causally implicated in cancer[12] (**Supplementary Table 2**).

Ascertainment of structural variants (copy number variants (CNVs), balanced rearrangements and mobile element insertions) is a critical component of comprehensive genome analysis. SV detection poses two key technical challenges. First, SVs are extremely difficult to detect reliably[13]. Second, functional interpretation of SVs requires specialized logic because of their variable size and diverse configurations and because SV breakpoints are often mapped imprecisely. As a result of these challenges, few established genome-analysis pipelines attempt to rigorously detect and interpret SVs.

SpeedSeq achieves comprehensive SV analysis with a suite of three complementary tools that are sensitive to a range of SV signals. At its core is LUMPY, a state-of-the-art breakpoint-detection tool that integrates split-read and discordant paired-end data[14]. Next, a custom parallelized implementation of CNVnator uses read-depth analysis to detect CNVs that may be invisible to LUMPY due to unmappable or repetitive sequences at their breakpoints[15]. Finally, SpeedSeq genotypes SVs with SVTyper, a novel Bayesian likelihood algorithm that can operate on copy-neutral events (such as inversions and translocations) and CNVs (Online Methods). This step produces SV genotypes that are crucial for meaningful variant interpretation and provides quantitative estimates of breakpoint allele frequencies that allow inference of the fraction of tumor cells that carry a particular variant.

Measuring SV detection performance on real data is difficult because of the lack of established truth sets. If we accept the 1000 Genomes Project (1KGP) deletion call set for NA12878 as ground truth[16,17], then SpeedSeq achieves a sensitivity of



```
$ gemini load –v tumor.vcf.gz –t VEP –p sample.ped tumor.db
```

```
$ gemini set_somatic tumor.db
```

```
$ gemini actionable_mutations tumor.db
chr17   7578460     7578461     TP53    non_syn_coding
chr17   59861630    59861641    BRIP1   inframe_codon_loss
chr2    128046288   128046289   ERCC3   non_syn_coding
chrX    132838304   132838328   GPC3    splice_region
```

```
$ gemini fusions ––in_cosmic_census ––min_qual 1 tumor.db
chr3   178905990    chr3     176909982    –    +    complex  PIK3CA  TBL1XR1
```

Figure 2 | Case study in a tumor-normal pair. A SpeedSeq workflow demonstrating the seven succinct commands required to process a tumor-normal pair (TCGA-E2-A14P) from raw FASTQ reads to clinically actionable somatic mutations with predicted damaging consequences. In this tumor, SpeedSeq detected a previously reported somatic gene fusion product between exon 1 of *TBL1XR1* and exon 2 of *PIK3CA*[20].

61.9% (2,089/3,376) and a positive predictive value of 60.8% (2,089/3,438) for detecting deletions, which is consistent with our recent comparative performance tests for LUMPY[14] and by inference shows that SpeedSeq achieves state-of-the-art SV detection relative to other tools. However, this test probably underestimates absolute performance because the 1KGP call set has known false positives and negatives. We therefore developed a composite strategy in which SVs in NA12878 could be validated either by overlap with split-read mapping of deep (30×) long-read data from PacBio and Illumina Moleculo platforms or by overlap with 1KGP. On the basis of this hybrid approach, SVs with quality scores of 100 or greater showed a positive predictive value of 86.0% (2,823/3,282) (**Fig. 1e** and **Supplementary Fig. 3**). Virtually none of these SVs are likely to have been validated by random chance, as 100 permutations of the call set resulted in a validation rate of 0.073% (± 6.1 × 10^{-3}, 95% confidence interval). Moreover, SVTyper's quality scores provide a tunable parameter for refining call sets to a desired confidence threshold. By requiring both paired-end and split-read support, users can generate an extremely high-confidence call set of 1,663 SVs with a 97.8% validation rate.

As an independent measure of SV detection and genotyping performance, we developed a haplotype-based test that exploits the structure of the CEPH 1463 pedigree. First, we phased the pedigree by SNV transmission to produce haplotype lineage maps that allowed us to attribute an average of 63.0% of the mappable genome of each F_2 individual to a particular founding grandparent (**Fig. 1f**). Next, we performed joint SV detection on the pedigree to generate 1,722 high-confidence autosomal SVs that could be assigned to a founding grandparent by transmission; this resulted in a truth set of 8,397 predicted SV observations across the 11 grandchildren with known genotypes. SpeedSeq showed a detection sensitivity of 90.2% (7,578/8,397) for these predicted SVs, encompassing 1,660 of the 1,722 unique variants (**Supplementary Table 3**). Among the SVs that were detected, SVTyper reported the correct genotype at 96.6% (6,845/7,083) of the heterozygous variants and 72.3% (358/495) of the homozygous variants. Moreover, the high specificity of this call set is apparent from the infrequency of Mendelian violations (5.0%) and the consistent cosegregation of SVs with SNV-based haplotypes (93.8%) (**Supplementary Table 4**).

Results from SpeedSeq seamlessly integrate into the GEMINI (GEnome MINIng) variant-interpretation framework, which annotates calls with information from external databases including dbSNP, ENCODE, ClinVar, CADD, ESP and ExAC for efficient filtering with command-line queries or a graphical browser interface[18]. In concert with SpeedSeq, we have made numerous enhancements to GEMINI, particularly in handling structural variants and interpreting somatic mutations. Users can rapidly prioritize somatic mutations through queries on two newly added databases: the COSMIC catalog of somatic mutations in cancer[12] and DGIdb, the Drug-Gene Interaction database[19]. In addition, GEMINI can now identify both structural variants that alter gene dosage or interrupt transcripts and putative somatic gene fusions that affect COSMIC cancer genes.

Finally, to provide an example of a typical cancer analysis interpretation, we performed somatic-variant calling on the tumor-normal pair of an invasive breast carcinoma from TCGA that carries a known gene fusion[20]. With four concise commands and less than an hour of computation, we loaded the variant call format (VCF) file into GEMINI, filtered variant calls for high confidence, clinically informative somatic mutations and predicted gene fusion events (**Fig. 2**). These analyses demonstrate the ease with which high-impact somatic point mutations and genomic rearrangements can be identified using the SpeedSeq framework.

METHODS

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

Accession codes. European Nucleotide Archive (ENA): ERP001960.

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

ACKNOWLEDGMENTS

The authors thank A. Abyzov for helpful discussions about CNVnator. This work was supported by US National Institutes of Health (NIH) training grant T32 GM007267 (C.C.), NIH NHGRI grant R01HG006693 (A.R.Q.), and NIH NHGRI center grant U54 HG003079, NIH New Innovator Award DP2OD006493-01 and a Burroughs Wellcome Fund Career Award (I.M.H.).

AUTHOR CONTRIBUTIONS

C.C. wrote SpeedSeq and analyzed the data. R.M.L. advised on LUMPY implementation, G.G.F. contributed SAMBLASTER features, M.R.L. assisted with cloud implementation and D.B.R. parallelized CNVnator. E.P.G. and G.T.M. advised on implementing FreeBayes. A.R.Q. contributed GEMINI features and advised on software design. I.M.H. conceived and designed the study. C.C. and I.M.H. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Li, H. Preprint at http://arxiv.org/abs/1303.3997v2 (2013).
2. DePristo, M.A. *et al. Nat. Genet.* **43**, 491–498 (2011).
3. Li, H. *et al. Bioinformatics* **25**, 2078–2079 (2009).
4. Dewey, F.E. *et al. J. Am. Med. Assoc.* **311**, 1035–1045 (2014).
5. Faust, G.G. & Hall, I.M. *Bioinformatics* **30**, 2503–2505 (2014).
6. Zook, J.M. *et al. Nat. Biotechnol.* **32**, 246–251 (2014).
7. Garrison, E. & Marth, G. Preprint at http://arxiv.org/abs/1207.3907 (2012).
8. Kingsmore, S.F. & Saunders, C.J. *Sci. Transl. Med.* **3**, 87ps23 (2011).
9. Cibulskis, K. *et al. Nat. Biotechnol.* **31**, 213–219 (2013).
10. Larson, D.E. *et al. Bioinformatics* **28**, 311–317 (2012).
11. Koboldt, D.C. *et al. Genome Res.* **22**, 568–576 (2012).
12. Futreal, P.A. *et al. Nat. Rev. Cancer* **4**, 177–183 (2004).
13. Alkan, C., Coe, B.P. & Eichler, E.E. *Nat. Rev. Genet.* **12**, 363–376 (2011).
14. Layer, R.M., Chiang, C., Quinlan, A.R. & Hall, I.M. *Genome Biol.* **15**, R84 (2014).
15. Abyzov, A., Urban, A.E., Snyder, M. & Gerstein, M. *Genome Res.* **21**, 974–984 (2011).
16. 1000 Genomes Project Consortium. *et al. Nature* **467**, 1061–1073 (2010).
17. 1000 Genomes Project Consortium. *et al. Nature* **491**, 56–65 (2012).
18. Paila, U., Chapman, B.A., Kirchner, R. & Quinlan, A.R. *PLoS Comput. Biol.* **9**, e1003153 (2013).
19. Griffith, M. *et al. Nat. Methods* **10**, 1209–1210 (2013).
20. Stransky, N., Cerami, E., Schalm, S., Kim, J.L. & Lengauer, C. *Nat. Commun.* **5**, 4846 (2014).

ONLINE METHODS

Software availability. The SpeedSeq v0.0.3a source code, documentation and example data files are available in **Supplementary Software**, as well as at https://github.com/hall-lab/speedseq.

Hardware. All timings reported herein were performed on a single machine with 128 GB RAM and two Intel Xeon E5-2670 processors, each with 16 threads.

Data. We benchmarked SpeedSeq's processing time using the NA12878 genome from the Illumina Platinum Genomes data set (European Nucleotide Archive: ERP001960), which comprises 50× WGS data sets for each of the 17 members of the three-generation CEPH 1463 pedigree (**Supplementary Fig. 4**).

WGS data from five matched tumor-normal pairs and their orthogonally validated somatic mutations were obtained from The Cancer Genome Atlas (TCGA). These included three colorectal tumors (TCGA-A6-6141, TCGA-CA-6718 and TCGA-D5-6540), one ovarian tumor (TCGA-13-0751) and one breast tumor (TCGA-B6-A0I6). Raw FASTQ reads were down-sampled to 50× coverage in the tumor and 30× coverage in the normal sample. Samples were processed with SpeedSeq for alignment, somatic mutations and structural variants using default parameters and then loaded into GEMINI for variant interpretation. We also analyzed WGS data from a tumor-normal pair (63× tumor, 49× normal coverage) of a patient with an invasive breast carcinoma (TCGA-E2-A14P) containing a previously reported gene fusion between *TBL1XR1* and *PIK3CA*[20].

FASTQ alignment and BAM processing. SpeedSeq aligns paired-end FASTQ files to the human GRCh37 reference genome with BWA-MEM 0.7.8 (ref. 1) using the "-M" flag to mark shorter alignments as secondary. Aligned reads are streamed directly into SAMBLASTER[5], which seizes idle CPU cycles that are periodically liberated each time BWA reads a FASTQ data chunk into the buffer. Marking duplicates on the presorted BAM file allows simultaneous extraction of discordant read pairs and split-read alignments, followed by rapid sorting and BAM compression with Sambamba[21].

SNV and indel detection strategy. SpeedSeq runs FreeBayes version 0.9.16 with "–min-repeat-entropy 1" and "–experimental-gls" parameters for germline variant calling[7]. To increase specificity, SpeedSeq also requires at least one read on both the left and the right to support the variant allele. For somatic variant detection, SpeedSeq uses parameters tuned to increase sensitivity over low-frequency variants (–pooled-discrete –genotype-qualities –min-alternate-fraction 0.05 –min-alternate-count 2 –min-repeat-entropy 1) and reports a somatic score (SSC) to estimate the confidence of each variant. The SSC is the sum of the log odds ratios for the tumor (LOD_T) and normal (LOD_N) samples, which is based on the genotype likelihood probabilities from FreeBayes (P_T and P_N for tumor and normal genotype probabilities, respectively). The SSC is the preferred tuning parameter as it is robust to sequencing depth by design; however, the minimum alternate fraction and minimum alternate count can also be adjusted on the SpeedSeq command line.

$$LOD_T = \log \frac{P_T(alternate)}{P_T(reference)} \tag{1}$$

$$LOD_N = \log \frac{P_N(alternate)}{P_N(reference)} \tag{2}$$

$$SSC = LOD_T + LOD_N \tag{3}$$

SpeedSeq's implementation of FreeBayes is parallelized over 34,123 windowed regions of the GRCh37 genome using GNU Parallel[22]. We generated these regions, which average 84 kb in length, by partitioning the genome into bins of approximately equal numbers of reads based upon the aggregate coverage depth of all 17 members of the CEPH 1463 family pedigree and excluding high-depth sequences (**Supplementary Note 4**, **Supplementary Table 5** and **Supplementary Fig. 5**). This binning scheme balances the computational load over the FreeBayes instances by allocating processors depending on the quantity of expected input data. It is 13.3-fold faster than the single-threaded version and 34.9% faster than naive parallelization over each chromosome (**Supplementary Note 1**).

Structural variation detection and genotyping strategy. SpeedSeq runs LUMPY with "-msw 4 -tt 0 min_clip 20 min_non_overlap 101 min_mapping_threshold 20 discordant_z 5 back_distance 10" and weights of 1 for both paired-end and split-read evidence. SpeedSeq's implementation of CNVnator parallelizes the genome by chromosome and performs copy number segmentation with a window size of 100 bp.

SVTyper is a maximum-likelihood Bayesian classification algorithm that infers an underlying genotype at each SV. Alignments at SV breakpoints either support the alternate allele with discordant or split-reads or support the reference allele with concordant reads or read-pairs that span the breakpoint. The ratio and quantity of these observations allow probabilistic inference of genotype likelihood. Under the assumption of diploidy, the set of possible genotypes at any locus is G = {reference, heterozygous, homozygous}. We defined the function S, where $S(g)$ is the prior probability of observing a variant read in a single trial given a genotype g at any locus. These priors were set to 0.1, 0.4 and 0.8 for reference, heterozygous and homozygous deletions, respectively. Assuming a random sampling of reads, the number of observed alternate (A) and reference (R) reads (scaled by mapping quality, $10^{(\text{-mapq}/10)}$) will follow a binomial distribution $B(A+R, S(g'))$, where $g' \in G$ is the true underlying genotype. Using Bayes's theorem we can derive the conditional probability of each underlying genotype state from the observed read counts (equation (4)), assuming an a priori probability $P(g)$ of 1/3 for each genotype. Finally, we calculate $\hat{g}$ as the inferred genotype for the variant. Since the algorithm only interrogates SVs in the VCF file that have passed LUMPY filters as nonreference, it reports the more likely genotype of heterozygous or homozygous alternate states.

$$S(g) = \begin{cases} 0.1 & \text{if } g = \textit{homozygous referernce} \\ 0.4 & \text{if } g = \textit{heterozygous} \\ 0.8 & \text{if } g = \textit{homozygous alternate} \end{cases} \tag{4}$$

$$P(A,R \mid g) = \binom{A+R}{A} \cdot S(g)^A \cdot (1 - S(g))^R \tag{5}$$

$$P(g\,|\,A,R)=\frac{P(A,R\,|\,g)\cdot P(g)}{P(A,R)}=\frac{P(A,R\,|\,g)\cdot P(g)}{\sum_{g\in G}P(A,R\,|\,g)\cdot P(g)} \quad (6)$$

$$\hat{g}=\arg\max_{g\in G}P(g\,|\,A,R) \quad (7)$$

SNV and indel evaluation. We compared SpeedSeq's germline SNV- and indel-variant calling against two independent truth sets for NA12878, one derived from the Genome in a Bottle (GIAB) NA12878 gold standard calls and the other based on Omni microarray data from the 1000 Genomes Project (1KGP). The GIAB 2.17 truth set contained 2,803,144 SNVs and 364,031 indels within highly confident regions (excluding segmental duplications, simple repeats, decoy sequence and CNVs) and spanned 2.2 Gb (77.6% of the mappable genome) for which nonvariant sites could be confidently considered homozygous reference. The Omni microarray truth set contained 2,177,040 informative SNVs, of which 689,788 were nonreference in NA12878, excluding markers within 50 bp of known indels.

We aligned the NA12878 raw reads from the Illumina Platinum data with SpeedSeq and then called germline SNVs and indels using SpeedSeq's default parameters. To evaluate SpeedSeq's performance against other standard tools, we also processed the aligned BAM files according to the Genome Analysis Toolkit (version 3.2-2-gec30cee) best practices workflow, including realignment around indels, base recalibration, and variant calling with Unified Genotyper (GATK-UG) and Haplotype Caller (GATK-HC). Variant quality score recalibration was performed on the GATK results using a passing tranche filter of <99%. We normalized and compared variant calls according to the GIAB protocol with vcfallelicprimatives, GATK's LeftAlignAndTrimVariants and VcfComparator[2,6]. We filtered variants for sensitivity and FDR against the GIAB truth set using a minimum quality score of 100 for GATK tools and 1 for SpeedSeq (open circles, **Fig. 1b**,**c**).

To evaluate performance in detecting somatic variants, we generated a simulated tumor-normal matched pair from the CEPH 1463 family Illumina Platinum data. The 'tumor' data set was an equal mixture of all 11 members of the F_2 generation, downsampled to 50× coverage and aligned with SpeedSeq. The father of the F_2 generation (NA12877) represented the 50×-matched normal sample. For inclusion in the somatic SNV truth set, we required a variant to be diallelic and autosomal in the NA12878 GIAB truth set, and called nonreference in NA12878 and reference in NA12877 (refs. 6,23) by Real Time Genomics (RTG). Additionally, variants were disqualified from the truth set if they violated Mendelian inheritance patterns. These criteria resulted in a set of 875,206 high-confidence SNVs covering 77.6% of the mappable genome. The truth set of variants in the chimeric tumor followed the expected binomial pattern of inheritance in her children, with a peak at 0.5 VAF from homozygous SNVs in NA12878 (**Supplementary Fig. 2a**).

We processed the simulated tumor data with SpeedSeq, MuTect 1.1.4, SomaticSniper and VarScan 2 using parameters designed to target variants as low as 5% variant-allele fraction. Receiver-operating characteristic (ROC) curves were generated by varying SSC for SpeedSeq, SomaticSniper and VarScan 2. For MuTect, which does not produce a single quality score for somatic variants, we varied the t_lod_fstar value to construct the ROC curve.

Structural variant evaluation. We constructed the 1KGP truth set by integrating deletions from the Pilot and Phase 1 call sets[16,17]. For long-read validation of SV breakpoints, we obtained 30× PacBio (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20131209_na12878_pacbio) data from 1KGP. We realigned the PacBio data with BWA-MEM 0.7.10 using the -x pacbio flag for consistency with the Moleculo alignments. Validations were performed according to previously published methods[14]. Custom scripts for this analysis are available at https://github.com/hall-lab/long-read-validation. To construct haplotype maps of the CEPH 1463 F_2 genomes, we called SNVs with SpeedSeq on the entire 17-member pedigree and phased SNVs by transmission at polymorphic sites in the parents. We smoothed the chromosomes for contiguous blocks of inheritance by selecting informative bases where 95% of each run of 101 SNVs reported a consistent parent of origin. We then merged regions that shared inheritance and that were within 100 kb of each other. This allowed us to trace an average of 1.8 Gb (63.4%) of each F_2 chromosome back to a particular grandparent, encapsulating meiotic crossovers that occurred in the F_1 germline (**Fig. 1f**). We then used SpeedSeq to jointly call structural variants on the entire pedigree, filtering for deletions that had at least seven pieces of support in at least one member of the pedigree, that had legal Mendelian transmission and whose origin could be unambiguously attributed to a single grandparent. Variants for which the founding grandparent by SV transmission agreed with the founding grandparent by SNV phasing were considered to be concordant, with strong supporting evidence for their authenticity. To test whether the 1,722 informative SVs were representative of the data set as a whole, and not of misleadingly high quality due to their ascertainment criteria, we assessed their validation rate as described above using the 1KGP call set and long-read sequencing (**Supplementary Table 4**). The 1,722 informative SVs had a similar validation rate as the remaining 6,734 SVs, suggesting that they are representative of overall call-set quality.

21. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J. & Prins, P. *Bioinformatics* doi: 10.1093/bioinformatics/btv098 (2015).
22. Tange, O. *The USENIX Magazine* **36**, 42–47 (2011).
23. Cleary, J.G. *et al. J. Comput. Biol.* **21**, 405–419 (2014).