Need for speed in accurate whole-genome data analysis: GENALICE MAP challenges BWA/GATK more than PEMapper/PECaller and Isaac

Michel Plüss[a,b,1], Anna M. Kopps[a,1], Irene Keller[c], Janine Meienberg[a], Sylvan M. Caspar[a], Nicolo Dubacher[a], Rémy Bruggmann[d], Manfred Vogel[b], and Gabor Matyas[a,e,2]

In the current high-throughput genomics era, efficient and accurate analysis of large-scale whole-genome sequencing (WGS) data constitutes a computational bottleneck. Johnston et al. (1) introduce the PEMapper/PECaller software package for short-read WGS alignment and variant calling, promising faster analyses with reduced output file sizes and "nearly identical (or better)" variant calling accuracy compared with the de facto standard Burrows–Wheeler aligner/Genome Analysis Toolkit (BWA/GATK) best-practices pipeline (2). However, we cannot confirm this promised BWA/GATK-like accuracy of PEMapper/PECaller, and there are other pipelines offering ultrafast WGS data analyses with small disk footprints, as we show in this correspondence.

To assess sensitivity/recall, precision, computation time, and disk footprint of four corresponding pipelines, we performed alignment and variant calling for the reference short-read WGS data of NA12878 and the Ashkenazim trio (3, 4). The four pipelines included the downloadable PEMapper/PECaller (1) and BWA/GATK (2) as well as the commercially available Isaac (5) and GENALICE MAP (genalice.com) software packages (versions and settings specified in Fig. 1). To largely reduce systematic errors and alignment artifacts, we limited our benchmarking of whole-genome variant calling to the coding part of the high-confidence BED file of GIAB 3.3 (https://github.com/genome-in-a-bottle), excluding exons with mappability <1, differences between GRCh37 and GRCh38, and/or common copy number variations (CNVs) (6).

In our benchmarking, PEMapper/PECaller was, although powerful, neither the fastest pipeline (Fig. 1) nor as sensitive in variant calling as BWA/GATK (Fig. 2*A*). Indeed, PEMapper/PECaller resulted in the highest number of false-negative calls (Fig. 2*A*), making it less suitable for clinical sequencing. As expected, BWA/GATK showed the highest sensitivity but fell behind the other three pipelines regarding run time and disk footprint. GENALICE MAP showed sensitivity comparable to BWA/GATK (Fig. 2*A*) but with a 112× faster total run time and a 45× lower disk footprint (Fig. 1). In precision, only minor differences were observed among pipelines, except for the PEMapper/PECaller population calling and the GENALICE MAP single-sample calling pipelines, which performed with the lowest and with distinctly lower precision, respectively, using downloaded FASTQ files (Fig. 2*B*). The difference between downloaded and our in-house data was pronounced in the sensitivity of the PEMapper/PECaller single-sample pipeline as well (Fig. 2*A*), suggesting considerable influence of input sequencing reads on PEMapper/PECaller.

However, although the here-applied reference datasets may have been used for pipeline optimization, there are no alternative/unbiased whole-genome truth sets available for benchmarking. Moreover, PEMapper/PECaller does not output BAM files, which are particularly useful in clinical sequencing for evaluating called variants and in CNV detection. Regarding run time, BWA/GATK might soon catch up with PEMapper/PECaller if the upcoming GATK version 4.0 is indeed 5× faster as announced or might even be faster if accelerated by the DRAGEN platform (edicogenome.com) or compressive methods such as CORA (7). Impressively, GENALICE MAP has already achieved ultrarapid speed and superior low disk footprint with BWA/GATK-like sensitivity, thus enabling efficient (re)analyses of ever-increasing amounts of WGS data.

[a]Center for Cardiovascular Genetics and Gene Diagnostics, Foundation for People with Rare Diseases, CH-8952 Schlieren-Zurich, Switzerland; [b]Institute of 4D Technologies, University of Applied Sciences and Arts Northwestern Switzerland, CH-5210 Windisch, Switzerland; [c]Department for BioMedical Research, University of Berne, CH-3008 Berne, Switzerland; [d]Interfaculty Bioinformatics Unit and Swiss Institute of Bioinformatics, University of Berne, CH-3012 Berne, Switzerland; and [e]Zurich Center for Integrative Human Physiology, University of Zurich, CH-8057 Zurich, Switzerland

Author contributions: M.P., A.M.K., I.K., J.M., and G.M. designed research; M.P., A.M.K., I.K., J.M., S.M.C., and N.D. performed research; and M.P., A.M.K., I.K., J.M., S.M.C., N.D., R.B., M.V., and G.M. wrote the paper.

The authors declare no conflict of interest.

[1] M.P. and A.M.K. contributed equally to this work.

[2] To whom correspondence should be addressed. Email: matyas@genetikzentrum.ch.

www.pnas.org/cgi/doi/10.1073/pnas.1713830114

Acknowledgments

We thank David J. Cutler and H. Richard Johnston, and Bas Tolhuis, Jack Findhammer, and Johannes Karten for help with PEMapper/PECaller and GENALICE MAP, respectively. This work was supported by the Blumenau-Léonie Hartmann-Stiftung, Gebauer Stiftung, Gemeinnützige Stiftung der ehemaligen Sparkasse Limmattal, and Wohlfahrtsstiftung des Vereines Zürcher Brockenhaus.

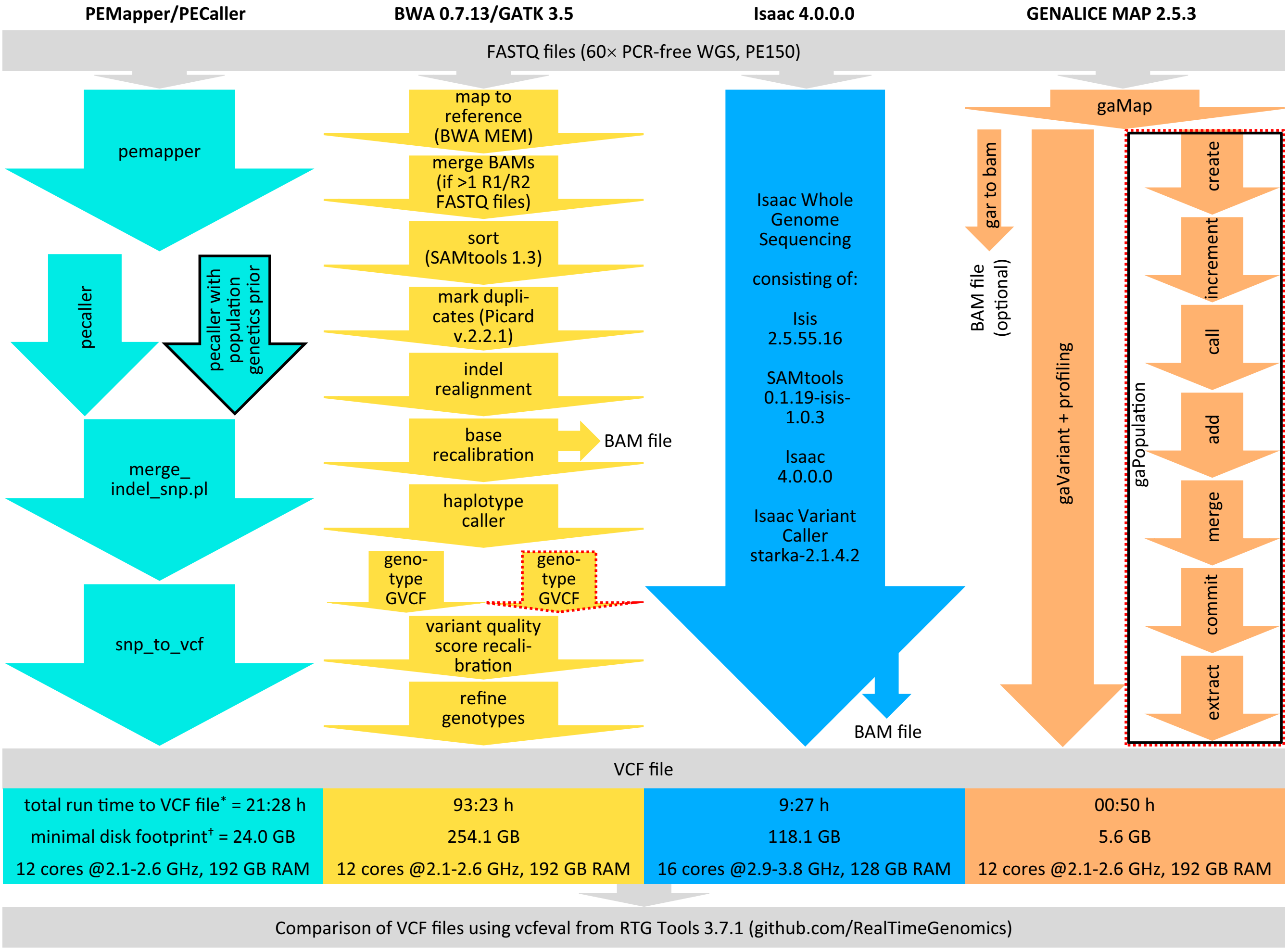
PEMapper/PECaller
BWA 0.7.13/GATK 3.5
Isaac 4.0.0.0
GENALICE MAP 2.5.3
FASTQ files (60× PCR-free WGS, PE150)
pemapper
pecaller
pecaller with population genetics prior
merge_ indel_snp.pl
snp_to_vcf
map to reference (BWA MEM)
merge BAMs (if >1 R1/R2 FASTQ files)
sort (SAMtools 1.3)
mark dupli-cates (Picard v.2.2.1)
indel realignment
base recalibration
BAM file
haplotype caller
geno-type GVCF
geno-type GVCF
variant quality score recali-bration
refine genotypes
Isaac Whole Genome Sequencing
consisting of:
Isis 2.5.55.16
SAMtools 0.1.19-isis-1.0.3
Isaac 4.0.0.0
Isaac Variant Caller starka-2.1.4.2
BAM file
gaMap
gar to bam
BAM file (optional)
gaVariant + profiling
gaPopulation
create
increment
call
add
merge
commit
extract
VCF file
total run time to VCF file* = 21:28 h
minimal disk footprint† = 24.0 GB
12 cores @2.1-2.6 GHz, 192 GB RAM
93:23 h
254.1 GB
12 cores @2.1-2.6 GHz, 192 GB RAM
9:27 h
118.1 GB
16 cores @2.9-3.8 GHz, 128 GB RAM
00:50 h
5.6 GB
12 cores @2.1-2.6 GHz, 192 GB RAM
Comparison of VCF files using vcfeval from RTG Tools 3.7.1 (github.com/RealTimeGenomics)

Fig. 1. Stepwise description, run time, disk footprint, and hardware specifications for the four investigated read mapping and variant calling pipelines. Solid black and dotted red outlines indicate population calling and trio analysis options, respectively. We mapped reads to the GRCh37-like reference genome hs37d5 (8), except for the Isaac pipeline running on BaseSpace Onsite not supporting custom reference genomes, where GRCh37 was used. Notably, hs37d5 contains noncanonical bases which PEMapper/PECaller (downloaded March 29, 2017) was unable to interpret and which were therefore replaced with Ns for this pipeline. Run times shown are for single-sample analyses of the downloaded NA12878 Genome in a Bottle (GIAB) data (legend of Fig. 2) (*). Minimal disk footprints for variant calling (†) were assessed, and thus for GENALICE MAP the size of the optional BAM file was not counted. Analysis parameters: PEMapper/PECaller according to ref. 1; BWA/GATK 3.5 best practices; Isaac default; GENALICE MAP best practices except for max_cigar_complexity = 18, max_context_call_density = 3, and min_map_quality = 1.

Plüss et al. PNAS | October 3, 2017 | vol. 114 | no. 40 | E8321

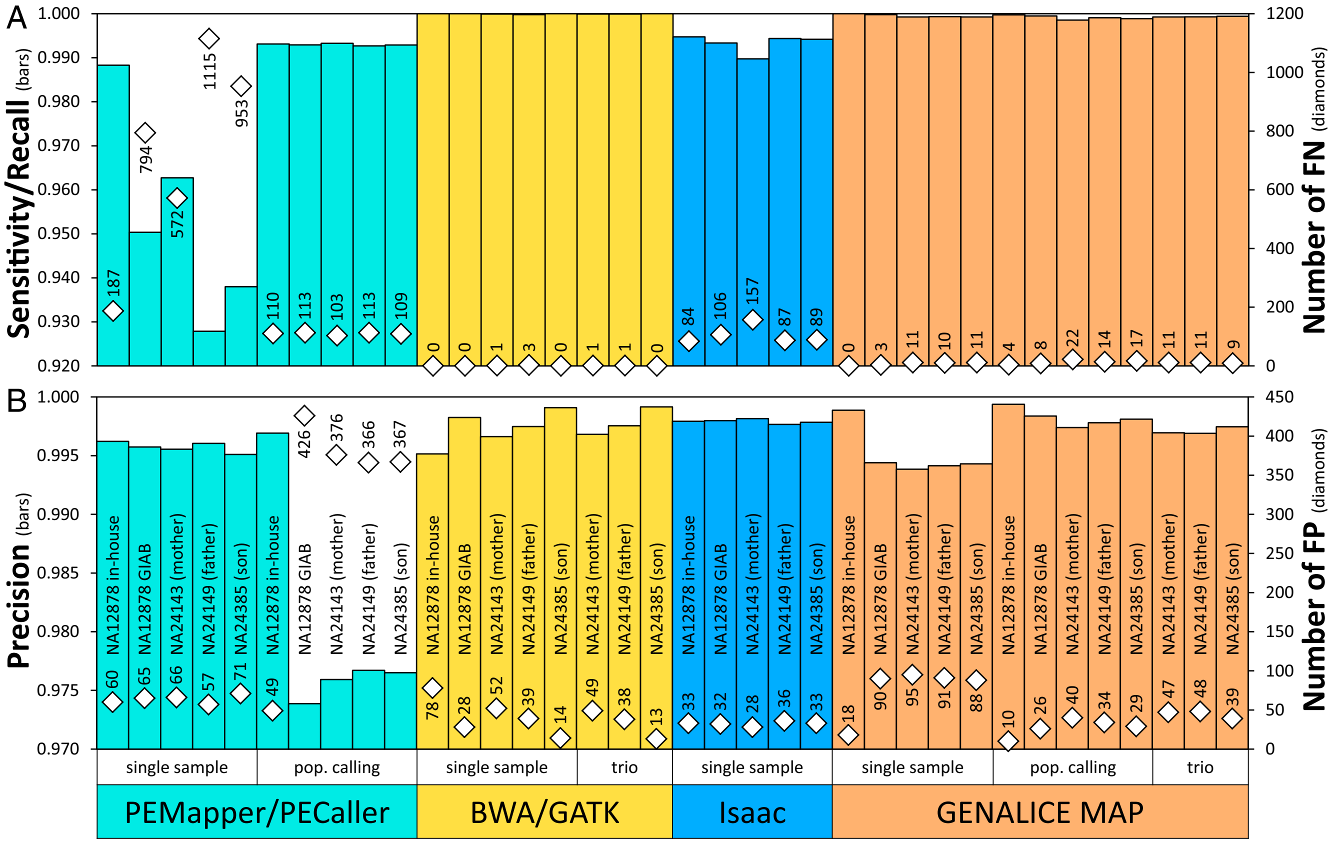

Fig. 2. Variant (SNP + indel) calling performance of the four investigated pipelines in single-sample analyses as well as population (pop.) calling and trio analyses. (*A*) Sensitivity/Recall [TP/[TP+FN]; TP = true-positive and FN = false-negative calls] and number of FN. (*B*) Precision [TP/[TP+FP]; FP = false-positive calls] and number of FP. WGS (Illumina HiSeq 2500, PE150, PCR-free) FASTQ files for NA12878 (NA12878 GIAB) and the Ashkenazim trio (NA24143, NA24149, and NA24385) were downloaded (ftp-trace.ncbi.nlm.nih.gov/giab, 300×) (4) and downsampled to ∼60× . In addition, we analyzed our in-house NA12878 WGS data (NA12878 in-house) sequenced at ∼60× (Illumina X Ten, PE150, PCR-free) (6). For population calling, the focal sample was analyzed together with 96 additional WGS datasets (sequenced like "NA12878 in-house") from our Caucasian (Swiss) patient cohort. The number of National Institute of Standards and Technology–GIAB high-confidence benchmarking TP calls were 15,990 (NA12878), 15,345 (NA24143), 15,458 (NA24149), and 15,366 (NA24385).

1 Johnston HR, et al.; International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome (2017) PEMapper and PECaller provide a simplified approach to whole-genome sequencing. *Proc Natl Acad Sci USA* 114:E1923–E1932.

2 Van der Auwera GA, et al. (2013) From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 11:11.10.1–11.10.33.

3 Zook JM, et al. (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 32:246–251.

4 Zook JM, et al. (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* 3:160025.

5 Raczy C, et al. (2013) Isaac: Ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* 29:2041–2043.

6 Meienberg J, Bruggmann R, Oexle K, Matyas G (2016) Clinical sequencing: Is WGS the better WES? *Hum Genet* 135:359–362.

7 Yorukoglu D, Yu YW, Peng J, Berger B (2016) Compressive mapping for next-generation sequencing. *Nat Biotechnol* 34:374–376.

8 Li H (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30:2843–2851.

E8322 | www.pnas.org/cgi/doi/10.1073/pnas.1713830114 Plüss et al.