

Genome analysis

Benchmarking and improving the performance of variant-calling pipelines with RecallME

Gianluca Vozza^{1,2}, Emanuele Bonetti^{1,2}, Giulia Tini ¹, Valentina Favalli ³, Gianmaria Frigè ¹, Gabriele Bucci ⁴, Simona De Summa⁵, Mario Zanfardino⁶, Francesco Zapelloni ³, Luca Mazzarella ^{1,*}

¹Department of Experimental Oncology, European Institute of Oncology IRCCS, Milan, Italy

²Department of Oncology and Hematology-Oncology, Università degli Studi di Milano, Milan, Italy

³4bases SA, Manno, Ticino, Switzerland

⁴Center for Omics Sciences, IRCCS Ospedale San Raffaele, 20132 Milano, Italy

⁵Molecular Diagnostics and Pharmacogenetics Unit, IRCCS Istituto Tumori, “Giovanni Paolo II”, Bari, Italy

⁶IRCCS Synlab SDN, 80143, Naples, Italy

*Corresponding author. Department of Experimental Oncology, European Institute of Oncology IRCCS, Via Adamello 16, 20139, Milan, Italy.

E-mail: luca.mazzarella@ieo.it (L.M.)

Associate Editor: Tobias Marschall

Abstract

Motivation: The steady increment of Whole Genome/Exome sequencing and the development of novel Next Generation Sequencing-based gene panels requires continuous testing and validation of variant calling (VC) pipelines and the detection of sequencing-related issues to be maintained up-to-date and feasible for the clinical settings. State of the art tools are reliable when used to compute standard performance metrics. However, the need for an automated software to discriminate between bioinformatic and sequencing issues and to optimize VC parameters remains unmet.

Results: The aim of the current work is to present RecallME, a bioinformatic suite that tracks down difficult-to-detect variants as insertions and deletions in highly repetitive regions, thus providing the maximum reachable recall for both single nucleotide variants and small insertion and deletions and to precisely guide the user in the pipeline optimization process.

Availability and implementation: Source code is freely available under MIT license at <https://github.com/mazzalab-ieo/recallme>. RecallME web application is available at <https://translational-oncology-lab.shinyapps.io/recallme/>. To use RecallME, users must obtain a license for ANNOVAR by themselves.

1 Introduction

The analytical performance of novel Next Generation Sequencing (NGS)-based and Third Generation Sequencing (TGS)-based variant calling (VC) pipelines requires comparing the experimental dataset with a “ground truth” dataset of expected variants (Salit and Zook 2019).

A key step in benchmarking is variant harmonization, as discrepancies in variant representation between ground truth and experimental dataset may lead to incorrectly interpreting a variant as missing in the experimental dataset.

Based on initial accuracy quantification, VC pipelines can be further optimized by modifying calling parameters and thresholds for quality metrics. This may be necessary in situations with a different tradeoff between recall and specificity: although large research-oriented studies may prefer removing false positives at the expense of true positives, as this may be compensated by the power afforded by large sample sizes, priorities are inverted in the diagnostic setting, where in general maximal recall is preferred, since the analysis is conducted on individuals, it bears clinical and legal relevance and a false

positive (FP) rate in key variants may be tolerated since these can be orthogonally validated, for instance by Sanger sequencing. In addition, especially for false negative (FN) calls, it is important to understand the source of miscalling, which can be due to two broad families of faults: those occurring during the “wet” part of the workflow (library preparation and sequencing), which can only be corrected by redesigning the analytical assay, and those associated with the bioinformatic pipeline, which can in theory be corrected by modifying the parameters of the “dry” workflow.

Currently, the standard software for pipeline benchmarking is hap.py (Krusche *et al.* 2019) although its “simpler” version som.py is often used across several indications (both for germline and somatic VC) as it simply compares the presence of specific sequences at given positions between the ground truth and query callsets without attempting to match haplotype, which is in general problematic in cancer somatic sequencing. Som.py provides a workflow for variant harmonization and generates accuracy metrics, but suffers from specific limitations: (i) it does not provide information on quality parameters of variants identified as FN and FP; (ii) it has not been

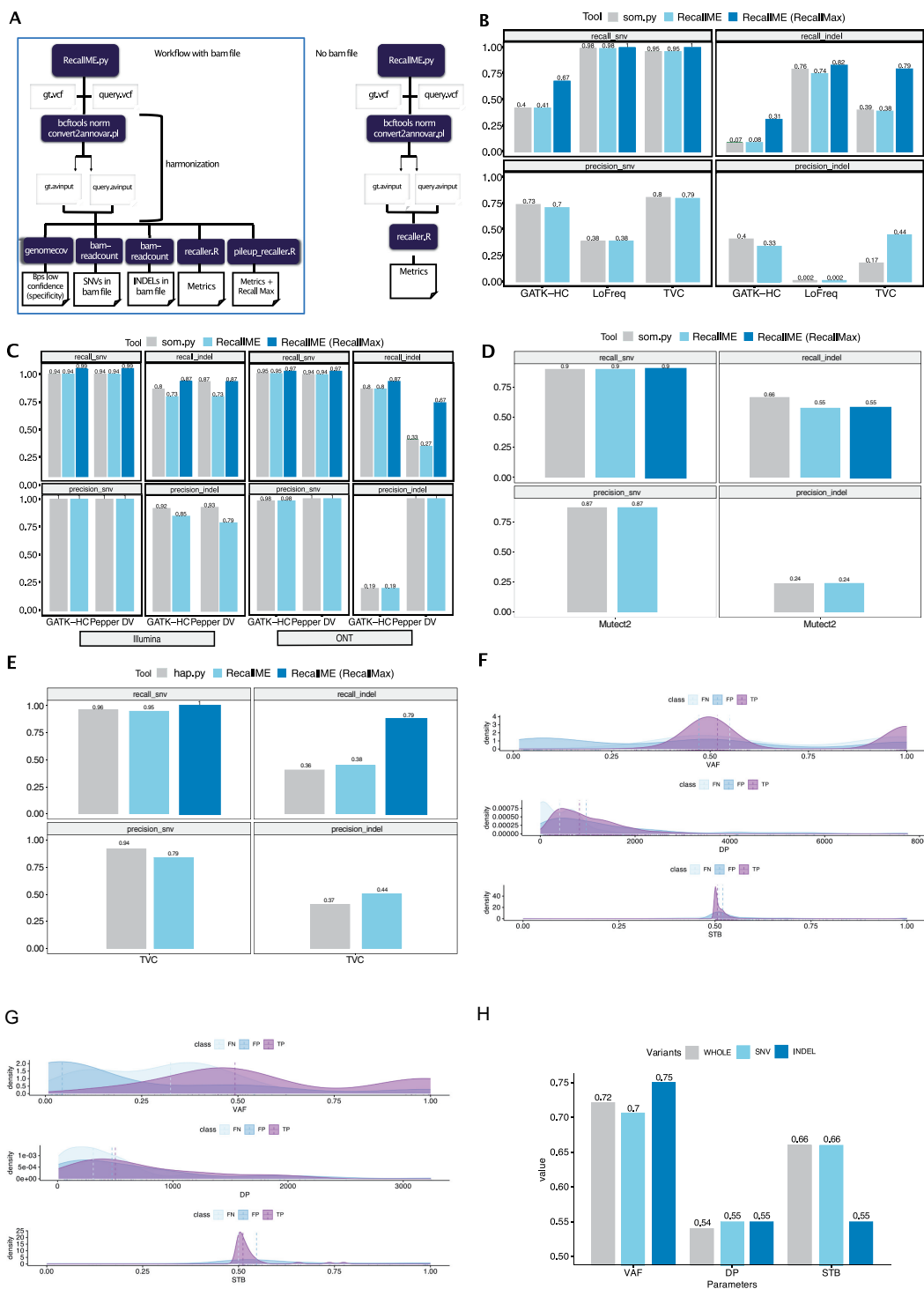


Figure 1. RecalME allows the maximization of the recall in reference samples. (A) Flowchart that shows how RecalME suite works with and without the bam file as input. First, bcftools norm splits the multi-allelic variants, then the query and the ground truth VCF files are converted in annovar inputs to harmonize variant notations. Bedtools genomecov function computes the number of bps that are not considered as high confidence by subtracting bps outside the bed file (to compute the number of true negatives and, consequently, the specificity). Bam-readcount look for SNVs and INDELs within the bam file to check if the recall can be maximized. The recal.R and the pileup_recal.R scripts compute the standard metrics as recall, precision, specificity, F1-score, FDR, and the Recall Max. (B) Barplots showing recall metrics in TVC and LoFreq pipelines (ION technology) and GATK-HC (Illumina-based) computed by som.py and RecalIME (before and after BAM file re-check step) in SNV and INDEL calls in NA12878 sample. (C) Barplots showing recall metrics in GATK-HC and Pepper DV pipelines (Illumina and ONT-based) computed by som.py and RecalIME (before and after BAM file re-check step) in SNV and INDEL calls in HD793 sample (Illumina and ONT). (D) Barplots showing accuracy metrics in Mutect2 pipeline computed by som.py and RecalIME (before and after BAM file re-check step, i.e. RecallMax) in SNV and INDEL calls in SEQC2 somatic dataset. (E) Barplots showing recall metrics in TVC pipeline (ION technology) computed by hap.py and RecalIME (before and after BAM file re-check step) in SNV and INDEL calls in NA12878 dataset. (F) TVC parameters distributions across TPs, FPs and b-FNs in SNV variants. Statistically significant differences have been found in VAF, DP, and STB (Mann–Whitney two-sided test). (G) TVC parameters distributions across TPs, FPs, and b-FNs in INDEL variants. Statistically significant differences have been found in VAF, DP, and STB (Mann–Whitney two-sided test). (H) Accuracies in TVC performances (NA12878) by tuning optimal cutpoints for VAF, DP, and STB for SNVs, INDELs and whole calls (WHOLE).

widely implemented on other VC pipelines, like Ion Torrent-based (ION) sequencing, common in clinical settings (Azzollini *et al.* 2023, Ricci *et al.* 2023, Schnidrig *et al.* 2023), for which the dominant VC algorithm is the Torrent Variant Caller (TVC). ION technology shows heterogeneity in variant annotation and some limitations in detecting insertion and deletion variants (INDELs) that may lead to incorrect estimates of the actual analytical performance (Loman *et al.* 2012, Laehnemann *et al.* 2016, Marine *et al.* 2020). Third-generation sequencing technologies like Oxford Nanopore Technologies are increasingly being used also in the diagnostic setting and thus require appropriate benchmarking tools, as ONT is able to resolve highly repetitive regions (Jain *et al.* 2022) but the detection of small variants remains challenging (Rang *et al.* 2018).

Here, we describe RecallME, a tool designed to standardize variant annotation across multiple callers, rapidly quantify performance metrics for NGS/TGS-based VC pipelines, discriminate between sequencing and bioinformatic errors and guide the user in the pipeline optimization process.

2 Materials and methods

2.1 Overview

RecallME was developed to identify and resolve the following drawbacks in accuracy assessment: (i) variant notation heterogeneity among callers; (ii) deconvolution of multi-allelic sites; (iii) identification of the cause for FN calls, through time-consuming re-check of supporting reads in the BAM file; and (iv) parameter optimization on recall and precision.

Figure 1A shows a standard RecallME workflow:

- i) variant harmonization: complex variants are decomposed with `bcftools norm` (Danecek *et al.* 2021) and harmonized (left-alignment and parsimonious reporting) between the query VCF and the ground truth VCF using the `convert2annovar.pl` function of ANNOVAR (Wang *et al.* 2010). This allows to split multiallelic sites into individual variants, maintaining the associated quality information and VC parameters and retaining variants with nonzero frequencies; the associated annotation is a desirable but nonessential outcome of this step
- ii) accuracy quantification: standard accuracy metrics as recall, precision, false discovery rate, F1-score and specificity through the `bedtools` (Quinlan and Hall 2010) `genomcov` function are computed;
- iii) BAM re-check: FNs are automatically re-evaluated in the BAM file using functions in the `bam-readcount` package (Khanna *et al.* 2022), which outputs a list of variants with variant caller-independent associated parameters as depth (DP), variant allele frequency (VAF) and percentage of strand-bias (STB). If a putative FN is found in the re-check step, we recalculate the recall metric as the maximum theoretical recall attainable (RecallMax, Supplementary Fig. S1) based on the available sequencing results (i.e. by assuming that a variant with supporting reads in the BAM file can be in principle called upon relaxing VC parameters);
- iv) parameter visualization and accuracy simulation: thresholds for VC quality parameters can be fine-tuned through a R Shiny application (`semantic.dashboard` v.0.2.1, available at <https://translational-oncology-lab.shinyapps.io/recallme/>), for instance to identify VC

thresholds that maximize precision or recall, depending on the purpose of the NGS panel. The interactive dashboard reports the dynamically changing lists of True Positives (TPs), FPs and FNs with the related standard metrics and the variants that have not been found within the BAM file. The lists and the performance metrics can be optimized directly within the dashboard by setting new VC parameters thresholds.

Datasets were analyzed with RecallME, `som.py` (v.0.3.15) and `hap.py` (v.0.3.9). Analysis with `som.py` was performed on datasets after variant harmonization through `bcftools norm -m-any` function using as input the query and the ground truth VCF files and the bed file of the panel to limit regions where standard metrics have to be computed and, finally, the FASTA reference files: hg19 for the internal and external datasets. The SEQC2 consortium dataset has been aligned to hg38.

2.2 Definition of benchmarking accuracy metrics

VC benchmarking studies have a potential for ambiguity in the language, since we have in fact three datasets to compare (the ground truth, the VC pipeline output, the benchmarking output) and terms describing accuracy metrics may refer to any of the possible comparisons. This potential ambiguity mostly arises in defining FNs and is related to the cause of miscalling. A FN is a variant that is expected but not observed, but as discussed previously, its absence from the experimental dataset may be: (i) sequencing-related: the variant is not sequenced at all, for instance due to complete lack of coverage or allelic dropout; (ii) bioinformatic: the variant is sequenced, but does not pass the filters imposed by the VC pipeline. It is potentially observed if VC filtering is completely removed; (iii) symbolic: the variant is sequenced and called, but it is represented in a different way than in the ground truth, generating a mismatch.

After resolving type 3 (symbolic) FNs through variant harmonization, the distinction between type 1 and 2 FNs is key to identify points of intervention. To measure type 2 (bioinformatic) error, we introduce a metric, RecallMax, which measures the maximum theoretical sensitivity achievable through changes in VC bioinformatic parameters. RecallMax is defined as

$$\text{RecallMax} = \frac{\text{TPs} + \text{b} - \text{FNs}}{\text{TPs} + \text{FNs}}$$

where `b-FNs` (bioinformatic-FNs) indicates the FNs for which the BAM re-check step has found supporting reads.

We then define the set of sequencing-positive variants:

$$\text{SeqPos} = \text{TPs} + \text{b} - \text{FN} + \text{FPs}$$

which we use as a reference dataset to quantify the impact of individual sequencing quality metrics on variant “callability”, i.e. the accuracy in calling an actually sequenced variant. In SeqPos, we pose $\text{TPs} + \text{b} - \text{FN} = 1$ (i.e. positively callable variants) and $\text{FPs} = 0$ (i.e. variants that should be called as negative). SeqPos is used to build Receiver Operating Curves (ROC) with sequencing quality metrics (e.g. VAF, DP, STB) as explanatory variables. The Area Under the ROC (AUROC) is calculated to measure the predictive power of

each sequencing metric; optimal cutoffs were calculated using the VAF, DP and STB metrics in the cutpointR package.

Statistical analyses were performed the two-sided Mann-Whitney test (with $\alpha = 0.05$) for continuous variables.

2.3 Generation of sequencing datasets

We tested the performance of RecallME on three sets of variants obtained by the standard Genome In A Bottle (GIAB) NA12878 (Zook *et al.* 2019) sample split in two dataset: an internal and an external validation datasets. The internal dataset was composed of a query callset called with GATK Haplotype Caller (GATK-HC) and a ground truth retrieved from the GIAB official ftp. The ground truth set was a publicly available version sequenced with the TruSeq panel from Illumina (https://ftp.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Nebraska_NA12878_HG001_TruSeq_Exome/, BAM file: NIST-hg001-7001-ready.bam; VCF file: NIST-hg001-7001-gatk-haplotype.vcf; BED file: TruSeq_exome_targeted_regions.hg19.bed) which covers 62 286 318 bps for benchmarking the GATK-HC pipeline and a re-sequenced NA12878 (purchased from the Coriell Institute for Medical Research) with an *in-house* designed gene panel based on ION technology (Ampliseq-based panel) for benchmarking the TVC and LoFreq pipelines. Our custom ION panel is designed to cover cancer driver genes for a total of 2 214 732 bps of genomic space, with amplicon size 125–175 bps; a manuscript describing the panel specifications and performance is currently under preparation; details are available upon reasonable request. The overlap between the ION and Illumina panels and the high-confidence region is 1 391 907 bps.

Library preparation was carried out using Ion Ampliseq custom panel (Thermo Fisher Scientific) and 10 ng of DNA were used to prepare libraries using two different primer pools. After 12-cycle PCR amplification products were barcode ligated and purified (Agencourt AMPure XP beads). Libraries were equimolar pooled to 50 pM, amplified and enriched using the Ion Chef system with the Ion 550 Kit-Chef (Thermo Fisher Scientific) and sequenced on 550 chip in a 200 bp run using Ion S5 GeneStudio XL system (Thermo Fisher Scientific). The mean coverage was 995 (SD = 674) for ION (test set), 169 (SD = 138) for Illumina (ground truth).

The external validation was performed on the HD793 sample (Horizon Discovery Ltd., <https://horizondiscovery.com/en/reference-standards/products/brca-germline-i-gdna>), which was sequenced on Illumina and ONT workflows. For both, the DNA library was prepared using the HEVA pro kit (4bases SA). 100 ng of genomic DNA were fragmented using a digestion enzyme and the A-tailed using library preparation kit. Hybridization and capture of target regions were done according to an internal protocol in use at 4bases. Sequencing was done on Illumina ISEC100. The procedure for the ONT sequencing was similar but after the probe captures and the barcoding of sequences were end-repaired and A-tailed and the Nanopore adapters were ligated (nano adapter, 4bases SA). The sample was then purified and loaded on a MinION MK1B sequencer.

For benchmarking on somatic VC, we retrieved data from the SEQC2 consortium (Fang *et al.* 2021). We used the intersection between WGS_IL_1.bowtie.muTect2.vcf dataset (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/analysis/SNVs/vcfs/WGS/WGS_IL_1.bowtie.muTect2.vcf.gz) and the high confidence regions bed

file (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/release/v1.2.1/High-Confidence_Regions_v1.2.bed). The computation of standard metrics has been performed with RecallME and som.py. For the comparison we used a ground truth generated by concatenation of high-confidence_sSNV_in_HC_regions_v1.2.vcf and high-confidence_sINDEL_in_HC_regions_v1.2.vcf for testing both SNV and INDEL calls (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/release/v1.2.1/high-confidence_sINDEL_in_HC_regions_v1.2.1.vcf.gz and https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/release/v1.2.1/high-confidence_sSNV_in_HC_regions_v1.2.1.vcf.gz).

2.4 VC and benchmarking

For the internal test dataset, VC was performed using three different pipelines: GATK HaplotypeCaller v.2.8 (GATK-HC) (McKenna *et al.* 2010, DePristo *et al.* 2011, Van der Auwera *et al.* 2013), Lofreq v.2.1.3.1 (Wilm *et al.* 2012), and TVC, for the latter we kept default parameters.

For the HD739, VC was performed using GATK-HC and Pepper DeepVariant (Pepper DV) (Shafin *et al.* 2021).

Som.py and RecallME were run on the resulting VCFs and recall/precision metrics were extracted and compared separately for SNV and indels.

The analyses were performed on a High-Performance Computing (HPC) cluster (2 frontend machines with 24 cores and 128 GB ram and 12 computing nodes with 28 cores and 128 GB ram).

We then performed an external validation of RecallME in order to assure the reproducibility of our workflow on a different machine (Linux operating system with 8 cores and 32 GB ram).

3 Results

3.1 Comparison of RecallME versus som.py and computation of the RecallMax

Initial software implementation was run on the ION dataset (see Section 2). Figure 1B shows accuracy performances obtained by RecallME and som.py and the computation of the RecallMax metric.

Both som.py and RecallME yielded similar accuracy metrics, although RecallME estimated a markedly higher precision for TVC indels, which we attribute to the two-step variant notation normalization feature implemented in RecallME, which allows the inclusion of the TVC output format among those immediately compatible with RecallME analysis. RecallMax was greater for INDELS (up to 40.5% for TVC) than SNVs (up to 26.5% for GATK-HC).

Further validation of the RecallME workflow was carried out on a variant dataset generated in an independent laboratory on a separate sample with Illumina and ONT-based sequencing.

RecallME was slightly more stringent than som.py in the Illumina dataset, although RecallMax provided virtually identical metrics. On the ONT dataset, performance was very similar, however RecallMax identified a very high rate of b-FNs for PepperDV (Fig. 1C).

Finally, in order to provide a comparison in a dataset especially relevant for somatic VC, we tested RecallME on the SEQC2 dataset. Performances computed by RecallME and som.py are virtually identical except for INDEL recall, in

which som.py provided higher values (0.66) than RecallME (0.55) (Fig. 1D).

3.2 Comparison of RecallME versus hap.py

We tested RecallME and hap.py on the original internal validation dataset TVC-based VC. Recall and precision were quite similar, although INDEL precision was quantified as higher by RecallME (0.44 versus 0.37) (Fig. 1E), perhaps expectedly since exact haplotype matching (more impactful for INDEL calls than SNV) is not required for RecallME to score a match. In addition, RecallMax for SNV was maximal, indicating that were, in fact, all sequenced (recall = 1) and 60 b-FN INDELS can indeed be identified in the BAM. The parameter distribution can guide threshold reset to increase recall.

3.3 RecallME for VC parameter optimization

A key feature of RecallME is the possibility to extract sequencing parameters associated with false calls and simulate changes in accuracy upon relaxing thresholds in these parameters. Distributions of key parameters like VAF, DP, and STB for TPs, b-FNs, and FPs in the ION dataset are shown in Fig. 1F and G. RecallME tests for significant differences in parameter distributions between FPs and TPs and b-FNs and TPs with a two-tailed Mann–Whitney test. We obtained for SNVs: VAF, $p_{\text{SNV}} < 0.0001$ for the comparison TPs versus FPs; DP $p_{\text{SNV}} = 0.01$ for TPs versus FPs and $p_{\text{SNV}} < 0.0001$ for TPs versus b-FNs; STB $p_{\text{SNV}} < 0.0001$ for TPs versus FPs. While for INDELS: VAF, $p_{\text{INDEL}} < 0.0001$ for both comparison TPs versus b-FNs and TPs versus FPs; DP, $p_{\text{INDEL}} < 0.001$ for TPs versus b-FNs; STB, $p_{\text{INDEL}} < 0.01$ for TPs versus FPs. Visualization and testing for these parameters are integrated in a RShiny-based application. It should be noted that some variant callers do not support the QD metric in the parameters. In such cases, RecallME computes an adaptation of the quality by depth provided by TVC ($QD = 4 \frac{\text{Quality}}{\text{DP}}$).

The set of all sequencing-positive variants (SeqPos) that can be theoretically called at the bioinformatic level given the sequencing results can be used to estimate the impact of individual sequencing quality parameters on accuracy (see Section 2) on variant “callability”, i.e. the accuracy in calling an actually sequenced variant. An example of this analysis is shown in Fig. 1H: of three sequencing parameters analyzed (VAF, DP, STB), VAF has the largest impact overall. Importantly, AUC is higher for INDELS than for SNVs. Consequently, estimated optimal cutpoints are different for SNV (0.3) versus INDELS (0.07) (Supplementary Fig. S2), suggesting that using differential VAF filtering SNV versus INDELS may achieve superior diagnostic accuracy.

4 Conclusion

We here present a bioinformatic tool that facilitates the process of accuracy estimation for NGS panels and VC pipelines.

Our approach to variant harmonization is similar to that employed by hap.py/som.py (Krusche *et al.* 2019), with decomposition/reconstruction from complex variants to multiple simpler variants and left-alignment and parsimony. Some differences arise in the handling of complex variants, which result in minor deviations in the metrics computation. In addition, the usage of bcftools norm and ANNOVAR enables RecallME to present the information in a format that is compliant with the format used by bam-readcount, which in turn allows RecallME to identify b-FNs (sequenced but miscalled by the VC

pipeline) and extract associated sequencing metrics, useful to optimize the VC pipeline based on the context-dependent need to prioritize sensitivity or specificity. We believe this feature to be particularly useful especially during the design and validation phases of NGS panels aimed at clinical diagnosis, and indeed this was the initial motivation for development. RecallME is primarily thought for benchmarking cancer somatic sequencing data, but can in principle be applied to germline datasets, which are also more commonly employed as standard references, although haplotype comparison is not featured.

Although it was primarily designed to prevent some specific shortcomings of ION sequencing, it can be useful also for benchmarking the accuracy of other sequencing platforms, including Illumina and ONT, as it facilitates and automatizes tedious steps of fine-tuning the variant-calling workflow. Moreover, similarly to ION sequencing, the detection of INDELS in highly repetitive regions in ONT-based pipelines remain challenging (Sarkozy *et al.* 2017) and RecallME can easily detect noncalled INDELS to improve the maximum achievable recall.

Acknowledgements

G.V. is a former PhD student at the European School of Molecular Medicine (SEMM), Milan, Italy. E.B. is a PhD student at the European School of Molecular Medicine (SEMM), Milan, Italy.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

V.F. and F.Z. are both employed at 4bases Italia s.r.l. G.V. is a former associate of 4bases Italia s.r.l. and currently employed at 4bases Italia s.r.l.

Funding

L.M. was supported by grant GR-2016-02361272 from the Italian Ministry of Health. G.V.’s PhD fellowship was supported by Alliance Against Cancer funds. This work was partially supported by the Italian Ministry of Health with Ricerca Corrente and Ricerca Corrente di Rete (ACCORD) 2020-2022 funds.

Data availability

The data underlying this article are protected by IP and will be shared on reasonable request to the corresponding author.

Cell line/DNA sample NA12878 was obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. A subset of the original internal validation dataset is available under the run_example/ folder within the Github repository and the related bam file is downloadable by running the *run_example.sh* script.

References

Azzollini J, Agnelli L, Conca E *et al.* Prevalence of BRCA homopolymeric indels in an ION torrent-based tumour-to-germline testing workflow in High-Grade ovarian carcinoma. *Sci Rep* 2023;13:7781.

- Danecek P, Bonfield JK, Liddle J *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 2021;10:giab008.
- DePristo MA, Banks E, Poplin R *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43:491–8.
- Fang LT, Zhu B, Zhao Y *et al.*; Somatic Mutation Working Group of Sequencing Quality Control Phase II Consortium. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using Whole-Genome sequencing. *Nat Biotechnol* 2021;39:1151–60.
- Jain C, Rhie A, Hansen N *et al.* A long read mapping method for highly repetitive reference sequences. *Nat Methods* 2022. <https://doi.org/10.1038/s41592-022-01457-8>.
- Khanna A, Larson D, Srivatsan S *et al.* Bam-readcount—rapid generation of basepair-resolution sequence metrics. *JOSS* 2022;7:3722.
- Krusche P, Trigg L, Boutros PC *et al.*; Global Alliance for Genomics and Health Benchmarking Team. Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* 2019;37:555–60.
- Laehnemann D, Borkhardt A, McHardy AC. Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Brief Bioinf* 2016;17:154–79.
- Loman NJ, Misra RV, Dallman TJ *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012;30:434–9.
- Marine RL, Magaña LC, Castro CJ *et al.* Comparison of illumina MiSeq and the ion torrent PGM and S5 platforms for whole-genome sequencing of picornaviruses and caliciviruses. *J Virol Methods* 2020;280:113865.
- McKenna A, Hanna M, Banks E *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–2.
- Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol* 2018;19:90.
- Ricci A, Bergamini E, Scarlini S *et al.* Clinical application of NGS in the diagnosis of iron overload disorders or hyperferritinemia of genetic origin. *Digest Liver Dis* 2023;55:S72.
- Sarkozy P, Jobbágy Á, Antal P. Calling homopolymer stretches from raw nanopore reads by analyzing k-mer dwell times. In: Eskola H, Väisänen O, Viik J, Hyttinen J. (eds) *EMBE & NBC 2017. IFMBE Proceedings*, vol 65, pp. 241–4. Springer, Singapore, 2017. https://doi.org/10.1007/978-981-10-5122-7_61.
- Schnidrig D, Garofoli A, Benjak A *et al.*; SOCIBP Consortium. PipeIT2: a tumor-only somatic variant calling workflow for molecular diagnostic ion torrent sequencing data. *Genomics* 2023;115:110587.
- Shafin K, Pesout T, Chang P-C *et al.* Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat Methods* 2021;18:1322–32.
- Van der Auwera GA, Carneiro MO, Hartl C *et al.* From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinf* 2013;43:11.10.1–10.33.
- Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164. <https://doi.org/10.1093/nar/gkq603>.
- Wilm A, Aw PPK, Bertrand D *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* 2012;40:11189–201.
- Zook JM, McDaniel J, Olson ND *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* 2019;37:561–6.