Genome Analysis

ABEMUS: platform specific and data informed detection of somatic SNVs in cfDNA

Nicola Casiraghi[1,#], Francesco Orlando[1,#], Yari Ciani[1], Jenny Xiang[2,4], Andrea Sboner[2,3], Olivier Elemento[2,3], Gerhardt Attard[5], Himisha Beltran[2,6,7], Francesca Demichelis[1,2,3,*], Alessandro Romanel[1,*]

[1]Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento. Trento, Italy, [2]Caryl and Israel Englander Institute for Precision Medicine, New York Presbyterian Hospital-Weill Cornell Medicine. New York, NY, [3]Institute for Computational Biomedicine, Weill Cornell Medicine. New York, NY, [4]Genomics and Epigenomics Core Facility, Weill Cornell Medicine. New York, NY, [5]University College London Cancer Institute, London, UK, [6]Department of Medical Oncology, Dana Farber Cancer Institute, Boston, MA, [7]Department of Medicine, Division of Hematology and Medical Oncology, Weill Cornell Medicine. New York, NY

co-first author

* co-corresponding author

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The use of liquid biopsies for cancer patients enables the non-invasive tracking of treatment response and tumor dynamics through single or serial blood drawn tests. Next generation sequencing assays allow for the simultaneous interrogation of extended sets of somatic single nucleotide variants (SNVs) in circulating cell free DNA (cfDNA), a mixture of DNA molecules originating both from normal and tumor tissue cells. However, low circulating tumor DNA (ctDNA) fractions together with sequencing background noise and potential tumor heterogeneity challenge the ability to confidently call SNVs.

Results: We present a computational methodology, called Adaptive Base Error Model in Ultra-deep Sequencing data (ABEMUS), which combines platform-specific genetic knowledge and empirical signal to readily detect and quantify somatic SNVs in cfDNA. We tested the capability of our method to analyze data generated using different platforms with distinct sequencing error properties and we compared ABEMUS performances with other popular SNV callers on both synthetic and real cancer patients sequencing data. Results show that ABEMUS performs better in most of the tested conditions proving its reliability in calling low variant allele frequencies somatic SNVs in low ctDNA levels plasma samples.

Availability: ABEMUS is cross-platform and can be installed as R package. The source code is maintained on Github at http://github.com/cibiobcg/abemus and it is also available at CRAN official R repository.

Contact: f.demichelis@unitn.it and alessandro.romanel@unitn.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Liquid biopsy provides an exceptional source of information for the identification and measurement of biomarkers relevant to precision oncology, from diagnosis and prognosis to treatment selection and monitoring of treatment response (Heitzer *et al.*, 2019). Circulating cell

free DNA (cfDNA) carries the genomic characteristics of tumor cell material shed into the bloodstream. In the presence of metastatic disease and/or of multifocal tumors, where single tissue biopsies would fall short in allowing heterogeneity assessment, cfDNA represents an ideal alternative to capture the disease genomic features. Several studies already demonstrated the prognostic value of circulating tumor DNA (ctDNA; the fraction of free DNA released from tumor cells as opposed to normal cells) and the ability to track tumor dynamics through the analysis of genomic lesions detected in the circulation of cancer patients (Annala *et al.*, 2018; Bettegowda *et al.*, 2014; Dawson *et al.*, 2013; Sclafani *et al.*, 2018; Siravegna *et al.*, 2015; Thierry *et al.*, 2014; Tie *et al.*, 2016; Vietsch *et al.*, 2017). One outstanding example of the use of liquid biopsy for the detection of relevant single-nucleotide variant (SNV) is the FDA approved test for *EGFR* exon 21 L858R substitution mutation in metastatic non-small cell lung cancer patient (Kwapisz, 2017), approved June 1st 2016. While highly sensitive technologies as digital PCR can be used for the investigation of SNVs in cfDNA, only next-generation sequencing (NGS) approaches allow for the simultaneous interrogation of large sets of genomic loci and for the discovery of mutations, with yet restricted amount of DNA (10-50ng). In the NGS based cfDNA testing, the perfect trade-off between SNV detection performance and sequencing depth is key. Specifically, low ctDNA fractions together with potential tumor heterogeneity challenge the ability to confidently call SNVs also due to the sequencing background noise. We therefore recognized the need for a benchmarked widely applicable computational method that combines individual's genetic knowledge and empirical signal to readily detect and quantify somatic SNVs in cfDNA also in the presence of low tumor fractions. We set up a computational methodology named ABEMUS (Adaptive Base Error Model in Ultra-deep Sequencing data) to discriminate between true SNVs and artefactual signals by learning locus-specific and data-driven variant allelic fraction thresholds while leveraging platform specific single base resolution information from sequencing assays (**Figure 1**). Performance and results were compared across an array of *in-silico* and real liquid biopsy data (including *in-silico* dilutions) against SNV detection methods commonly used in tumor tissue based studies (Cibulskis *et al.*, 2013; Kim *et al.*, 2018; Koboldt *et al.*, 2012; Larson *et al.*, 2012) or specifically proposed for cfDNA data (Kockan *et al.*, 2017).

2 Materials and Methods

2.1 Plasma and germline sequencing data from cancer patients

To build different ABEMUS platform specific sequencing error reference models and study their properties, we collected germline samples sequencing data profiled using 5 platforms (here intended as the combination of library preparation kit and sequencing machine/chemistry). Specifically, we used both i) whole exome sequencing (WES) data from 40 normal samples sequenced both with NimbleGen (Roche NimbleGen SeqCap Exome v3, 64Mbp covered) (Beltran H, et al, *submitted*) and with HaloPlex (Agilent HaloPlex Exome, 36Mbp covered) kits (Beltran *et al.*, 2016), and ii) custom targeted panel data from three sets of normal samples (N=20, 113 and 3) sequenced via Roche NimbleGen N250 targeted panel, Ion AmpliSeq Targeted Custom Amplicon Panel (Carreira *et al.*, 2014; Romanel *et al.*, 2015) or Illumina True Seq Custom Amplicon and covering 3.2 Mbp, 40 kbp and 106 kbp, respectively (see **Supplementary Table 1**). Additionally, we queried 118 plasma samples from 17 metastatic prostate cancer patients (median number of plasma samples per patient is 5) profiled on an Ion AmpliSeq Targeted Custom Amplicon Panel. The case samples have been previously annotated by tumor content (ctDNA) using CLONET (Prandi *et al.*, 2014) and by manually curated SNVs calls (Carreira *et al.*, 2014).

2.2 Data pre-processing for ABEMUS computations

Pileup data (PILEUP files) were generated using PaCBAM (Valentini *et al.*, 2019) to obtain depth of coverage and allele-specific statistics at each considered locus. Genomic positions with variant allelic fraction greater than zero are available in *.pabs PaCBAM output files. Sequencing reads with read and base qualities ≥ 20 were retained in the pileup computation.

2.3 Global and local estimations of sequencing errors

Given a set of germline samples profiled with the same platform, cumulative PILEUP across all samples for all targeted genomic positions is computed. PILEUP data is used by ABEMUS to build the overall distribution of variant allelic fractions (AFs) observed in the set of germline samples (global sequencing error distribution, GSE) and to compute a locus-specific measure (per-base error measure). The GSE is used to determine a coverage-independent AF threshold (AF_{th}) and a coverage-dependent AF threshold ($AF_{th\_cov\_bin}$). Given a desired level of specificity (user defined, default 0.995), the AF_{th} value is computed as the corresponding quantile of the GSE, while $AF_{th\_cov\_bin}$ values are similarly computed stratified by depth of coverage bins. Formally,

$$AF_{th} = quantile(GSE, s)$$
$$AF_{th\_cov\_bin} = quantile(GSE_{cov\_bin}, s)$$

where s is the desired detection specificity ($0 \leq s \leq 1$), and $GSE_{cov\_bin} \subseteq GSE$ is the subset of AFs in GSE with depth of coverage within a bin of coverage, cov_bin.

The local per-base error measure (*pbem*) is computed as:

$$pbem_x = \frac{\sum_{i=1}^{N} alt_{i\,x}}{\sum_{i=1}^{N} cov_{i\,x}}$$

where x is a genomic locus, alt is the number of sequencing reads supporting an allele different from the reference, cov is the total coverage and N the number of germline samples considered.

2.4 ABEMUS single nucleotide variants calls

Given a plasma sample and pre-computed GSE and $pbem$ estimations, the identification of putative somatic SNVs in the plasma sample is performed through two main sequential filtering steps. First, ABEMUS filters genomic positions using either the pre-computed coverage-independent AF_{th} or -dependent $AF_{th\_cov\_bin}$ thresholds (as determined by user, default coverage-dependent $AF_{th\_cov\_bin}$). The former applies the same threshold across all positions with AF > 0, while the latter applies coverage stratified thresholds to each locus x based on the depth of coverage $cov_x \in cov\_bin$. Last, for each retained locus, ABEMUS tests the plasma sample locus x AF against the corresponding $pbem_x$. Since the $pbem$ is computed on all reads from a panel of germline samples, at a specific locus x, the AF threshold $AF_{th_{pbem}}(x)$ is computed as a function of the $pbem_x$ and of the local plasma sample coverage (cov_x) as follows:

$$AF_{thr_{pbem}}(x) = F(pbem_x, cov_x) * R_{\overline{cov},tsize}$$

This function returns the maximum AF observed among 100,000 experiments modelled as binomial distributions $B(n,p)$ with p corresponding to $pbem_x$ and number of trials n corresponding to the locus coverage cov_x. This value is then rescaled by a factor $R_{\overline{cov},tsize}$ which maximizes ABEMUS precision and recall in plasma samples with global mean coverage equal to $\overline{cov}$ and target size equal to $tsize$.

Further filtering criteria on minimal locus coverage and minimal AF in plasma sample can be applied to reflect a-priori user specific requirements. Additionally, when matched germline sample data is available, filters on minimal locus coverage and maximal AF in matched germline sample can be applied. At each computation step, the list of genomic loci to be processed is reduced (intermediate and final lists are saved). The final list includes the set of putative somatic SNVs for the plasma sample.

2.5 Synthetic BAM files generation, preserving real data features, coverage and sequencing error

To test ABEMUS performance, synthetic BAM files were generated using summary statistics from a collection of human germline samples. Specifically, we considered 50 germline BAM files profiled with Agilent HaloPlex Exome kit (36Mbp covered) at approximately 200x mean depth of coverage (Beltran *et al.*, 2015). Coverage and allele-specific statistics across all captured genomic regions were computed and characterized both at region and base-specific level. In particular, we computed $pbem$ and the probability distribution P_{start}, which for each position x in the panel measures the probability of observing a mapped read with starting position in x. Synthetic BAM files were obtained from synthetic FASTQ files aligned to the human hg19 reference genome using BWA aligner (Li and Durbin, 2009) and were finally processed with SAMtools (Li *et al.*, 2009). Given a number N of required reads of length L and a set S of heterozygous SNPs derived from randomly selected European individuals from the 1,000 Genomes Project, synthetic FASTQ files were created by generating N synthetic reads using the following procedure: 1) select a start alignment position $chr{:}x$ using the probability distribution P_{start}; 2) build the read sequence considering the genomic coordinates $chr{:}x{,}(x+L)$ in the human hg19 reference genome and select an allele with probability 0.5; 3) introduce an error at each read position $chr{:}y \in [chr{:}x{,}chr{:}(x+L)]$ with a probability reflective of $\max(0.002,\ pbem_y)$ where 0.002 is the average background error computed from the original germline data; 4) introduce the alternative base of a SNP $s \in S$ at genomic position $chr{:}y \in [chr{:}x{,}chr{:}(x+L)]$ if $chr{:}y$ corresponds to the genomic position of SNP s. If the synthetic data is intended to represent a case sample, an heterozygous SNV m from a set M of pre-selected heterozygous SNVs is introduced in a read at position $chr{:}y \in [chr{:}x{,}chr{:}(x+L)]$ if $chr{:}y$ corresponds to the genomic position of SNV m; the SNV is introduced with a probability TC, where $TC \in [0,1]$ represents a level of ctDNA. Base quality values in FASTQ files are all set to a pre-defined value Q. Using this procedure, we generated two large datasets of synthetic data, one to optimize ABEMUS performance and one to run comparative performance study with other tools.

2.6 Generation of synthetic data to optimize ABEMUS performance

Using the previously described procedure we generated a set of 50 synthetic germline BAM files and a set of 9 plasma-germline synthetic BAM file pairs reflective of covering 36Mbp (100% of HaloPlex target) at mean coverage of 2000x. Plasma BAM files were generated introducing in each sample a different set of 200 clonal heterozygous SNVs and mimicking a range of ctDNA values, as 80%, 40%, 20%, 15%, 12.5%, 10%, 7.5%, 5% and 2.5%. PILEUP data for these samples was calculated with PaCBAM and used to generate synthetic input data for ABEMUS covering different scenarios of depth of coverage, target size and admixture level. Specifically, starting from those PILEUP data and adopting a sub-sampling procedure, we generated synthetic input data to represent assays with smaller genomic targets (75%, 50%, 25%, 12.5%, 6%, 3%, 1%, 0.5%, 0.1% corresponding to 26.6, 17.7, 8.9, 4.4, 2.1, 1.1, 0.4, 0.2, and 0.04 Mbp of the 36 Mbp HaloPlex target, respectively), each at multiple mean coverages (50%, 25%, 10% of the original coverage corresponding to 1000x, 500x and 200x mean coverage, respectively). Combinations of targets (N=10) and coverage levels (N=4) resulted in an extended collection of 1,600 synthetic germline input data grouped in 32 target-coverage classes and 288 synthetic plasma-germline input data also grouped in 32 target-coverage classes across nine different levels of ctDNA. Case tumor BAM files were generated introducing in each case a different set of 200 clonal heterozygous SNVs except for BAM files covering 0.2Mbp and 0.04Mbp in which sets of 100 clonal heterozygous SNVs were introduced. For all synthetic samples, base qualities were set to 20. Generated synthetic reads length was set to 101bp. This dataset is referred to as **Synthetic Dataset #1.**

2.7 Generation of synthetic data for comparative analyses with published tools

A second set of plasma and matched germline synthetic BAM files was generated to compare ABEMUS performances against published SNV detection tools. Three combinations of depth of coverage and target size were considered: 1) 2000x mean depth of coverage across 1% HaloPlex target; 2) 1000x mean depth of coverage across a 12.5% of HaloPlex target; 3) 200x mean depth of coverage across 100% HaloPlex target. For each scenario we generated 50 synthetic germline BAM files and a set of 9 synthetic plasma-germline samples pairs spanning a range of ctDNA values (80%, 40%, 20%, 15%, 12.5%, 10%, 7.5%, 5% and 2.5%). Case tumor BAM files were generated introducing in each sample a different set of 200 clonal heterozygous SNVs. For each plasma sample, two synthetic BAM files were generated, considering base qualities set to 20 and 30. Generated synthetic reads length was set to 101bp. This dataset is referred to as **Synthetic Dataset #2.**

2.8 *In-silico* dilutions from real cfDNA data for comparative analyses with published tools

To further perform comparative analyses with published tools, we created *in-silico* dilutions from real cfDNA and matched normal data, in order to control for ctDNA. Briefly, from previously published data (Carreira *et al.*, 2014) we selected 41 plasma samples from 8 patients with at least one reported somatic SNV and with plasma mean coverage higher than the matched germline control sample (Buccal Swab). A computational procedure was applied to precisely admix fractions of tumor and germline sequencing reads sampled from original BAM files accordingly to the intended ctDNA while preserving the original mean depth of coverage. Let $ctDNA_{in}$ be the reported ctDNA level of the plasma sample, $ctDNA_{tar}$ be the target ctDNA level, $\overline{cov}_p$ be the mean coverage of the plasma sample and $\overline{cov}_g$ be the mean coverage of the germline sample. Then, synthetic dilutions are obtained by mixing a fraction F_p of plasma data reads and a fraction F_g of germline data reads (when $F_g \leq 1$), with:

$$F_p = \frac{ctDNA_{tar}}{ctDNA_{in}} \text{ and } F_g = (1 - F_p) * \frac{\overline{cov}_p}{\overline{cov}_g}$$

By applying this procedure, a final set of 291 synthetically diluted samples covering a wide range of ctDNA levels (80%, 40%, 20%, 15%, 12.5%, 10%, 7.5%, 5% and 2.5%) was generated. This dataset, which is hence built using a sub-sampling procedure that mixes sequencing reads from real cfDNA and matched control samples, is referred to as **Synthetic Dataset #3.**

2.9 ABEMUS parameters used in study experiments and data availability

ABEMUS parameters applied in study experiments are listed in **Supplementary Table 2** and **3**. The reference error models of the platforms investigated in this study are available at http://github.com/cibiobcg/abemus_models.

3 Results

3.1 ABEMUS summary overview

ABEMUS is a tool specifically designed to detect somatic SNVs from cfDNA data and is implemented as package in the R environment. The identification of somatic SNVs from a plasma sample is performed by ABEMUS using locus-specific and data driven filters that are calculated exploiting pre-computed reference error models (**Figure 1**). For each experimental platform, here intended as the combination of library preparation kit and sequencing machine/chemistry, reference error models that estimate both global and local sequencing error background are built by ABEMUS from a set of germline samples data generated with the same platform. Of note, ABEMUS provides pre-computed reference error models for several experimental platforms. When matched germline sample data is available for a plasma sample, additional filters can be used by ABEMUS to refine the identification of somatic SNVs by further considering private SNPs (e.g. singletons).

As a result, ABEMUS nominates a list of putative somatic SNVs in a format compatible with external tools providing also functional annotations (i.e. Oncotator (Ramos *et al.*, 2015), SnpEff (Cingolani *et al.*, 2012)) together with additional information like the locus strand bias and the genomic context, which altogether can be further used to rank or prioritize the identified SNVs.

3.2 *pbem* is a sequencing platform dependent feature

We tested the hypothesis that sequencing errors, quantified using *pbem*, depend on the experimental platform. To test this hypothesis, we collected a series of data of germline samples profiled using different platforms as reported in **Supplementary Table 1**. We first exploited the 113 germline samples from the 40kb IonTorrent PGM sequencing series and assessed *pbem* for two disjoint subsets of samples (56 and 57) across all targeted genomic loci. The resulting distributions of *pbem* and coverage were comparable and further the *pbem* correlation (Pearson's product-moment correlation, r = 0.72) indicated agreement between the two sets of base level measures (**Figure 2A,** S1 versus S2). Similarly, two subsets of the 36Mbp WES assay (Agilent HaloPlex Exome) sequencing series of 10 germline samples each (**Figure 2A,** S3 versus S4) and to subsets of the 3.2Mbp Roche NimbleGen N250 targeted panel sequencing series (**Figure 2A**, S5 versus S6) demonstrated comparable results. On the contrary, the same procedure but comparing data generated by two platforms (Ion AmpliSeq Targeted Custom Amplicon Panel on IonTorrent PGM and Illumina True Seq Custom Amplicon on Illumina MiSeq) from the same set of normal samples resulted in non-correlated *pbems* series (r = -0.02) on the 7,201 shared bp (**Figure 2B**, S7 versus S8). The same result was obtained from 40 normal samples WES data generated using two kits, the Roche NimbleGen SeqCap Exome v3 and the Agilent HaloPlex Exome (r = -0.03) with 31 Mbp shared positions. These experiments suggest that the background noise of sequencing experiments is locus and platform specific (**Figure 2C**). Indeed, approximately 50% of targeted positions show evidence of errors (*pbem* > 0) only when data are derived from one platform.

3.3 Stability and optimization of global sequencing error estimation *GSE*

To formally investigate the properties of global sequencing error background ABEMUS estimates, we compared the coverage-based AF threshold measures $AF_{th\_cov\_bin}$ computed on synthetic germline data (**Synthetic Dataset #1**) across different mean coverages (N=4; 2000x, 1000x, 500x and 200x), target sizes (N=4; 36 Mbp, 17.7 Mb, 4.4 Mbp and 0.4 Mbp) and detection specificities (0.99, 0.995, 0.999). Overall, although estimations of AF thresholds were relatively stable across different mean coverages and target sizes (**Figure 3A**), especially for strict values of detection specificity, poorly populated coverage bins demonstrated sparse GSE_{cov_bin} distributions (**Figure 3A** and **Figure S1)**. To correct for this bias, we implemented a refined procedure to identify the most suitable coverage-based AF threshold also in those bins that are problematic due to low cardinality. Briefly, assuming that coverage bins stability is function of bins cardinality, we tested the stability of each coverage bin cov_bin by performing sub-sampling analysis on coverage cov_bin', representing the bin having the closest but higher cardinality with respect to cov_bin. Specifically, each coverage bin is first decomposed into subsets N and M containing positions with AFs > 0 and AFs = 0, respectively. Then, coverage bins are sorted by decreasing cardinality of N and starting from the most populated bin of non-zero AFs (and sequentially for each *i-th* coverage bin), k random samplings ($k = 1000$ by default) of $|N_{i+1}|$ and $|M_{i+1}|$ AFs are performed from N_i and M_i, respectively. For each random sub-sample, the resulting GSE'_{cov_bin} (with $GSE'_{cov\_bin} \subseteq GSE_{cov\_bin}$) is used to estimate $AF'_{th\_cov\_bin}$. The variability across the k estimated $AF'_{th\_cov\_bin}$ values is quantified using the coefficient of variation CV. For each *i-th* coverage bin, if $CV_i < th_{CV}$ ($th_{CV} = 0.01$ by default), the cardinality of the coverage bin C_{i+1} is considered reliable for the AF threshold estimation, hence the $AF_{th\_cov\_bin_{i+1}}$ is computed using C_{i+1} AFs. Otherwise, if $CV_i \geq th_{CV}$, the $AF_{th\_cov\_bin_{i+1}}$ is updated as $AF_{th\_cov\_bin_j}$ where $j<i$ is the last coverage bin such as $CV_j < th_{CV}$. If all coverage bins have $CV_i \geq th_{CV}$, then all $AF_{th\_cov\_bin}$ are set to the coverage-independent AF threshold (AF_{th}). As shown in **Figure 3B**, the refined procedure resulted in highly stable AF thresholds, across different combination of coverage mean, target size and detection sensitivity.

3.4 Assessment of scaling factors to maximize ABEMUS performance

Synthetic Dataset #1 was used to identify the best scaling factor R to maximize ABEMUS precision and recall for combinations of coverage and target size at different ctDNA levels. We tested a wide range of R values (N = 71, min = 0.5, max = 8, step 0.01) and evaluated the corresponding F1 scores. For each combination of target size, mean

coverage and ctDNA level, we selected the minimum factor R among those such that $F_R > F_{thr}$, where F_R is the F1 score achieved by ABEMUS using the scaling factor R and F_{thr} a custom threshold. Analyses using $F_{thr} \in \{0.9, 0.92, 0.94, 0.96, 0.98\}$ indicate that the wider the genomic target and the higher the mean coverage, the lower the optimal R required to get a desired F1 score. Conversely, for the same combination of target size and coverage, lower admixtures require a greater R (**Figure 4** and **Figure S2**).

Since the ctDNA level information might not be available upfront for a plasma sample, we also defined the optimal scaling factor R maximizing precision and recall across a set of admixtures only based on target size and coverage. Using a set of thresholds for the F1 score (N = 11, min = 0.9, max = 1, step 0.01), we selected the minimum R generating an F1 score greater than the highest observed threshold in the greater number of ctDNA levels considered (N = 9).

Using these optimization results, ABEMUS enables the selection of the *R* factor that best fits the sample's target size, mean coverage and when available ctDNA level; alternatively, the user can set a preferred scaling factor *R* or disable the scaling factor (*R=1*).

3.5 SNVs detection precision and recall on synthetic data

ABEMUS performances at different target sizes, mean coverages and ctDNA levels were assessed using **Synthetic Dataset #2** and were compared to performances of four tools commonly used in tumor tissue based analysis: SomaticSniper (Larson *et al.*, 2012), MuTect (Cibulskis *et al.*, 2013) (run both in standard mode and with creation and usage of a panel of normals), VarScan2 (Koboldt *et al.*, 2012) and Strelka2 (Kim *et al.*, 2018). All tools were run following developers' instructions reported on the relevant websites. As previously described, background sequencing error in **Synthetic Dataset #2** was introduced using a per-base error model computed from real sequencing data and synthetic reads were generated using two different base quality models. ABEMUS was run by exploiting the optimized scaling factors result of the previous analysis. As shown in **Figure 5** and **Supplementary Table 4**, at the lowest ctDNA level and lowest depth of coverage that we considered, ABEMUS is the only tool reaching an F1 score of 0.1 with a precision above 60%; all other tools demonstrated extremely low F1 score, with Strelka2 being the only one with precision and recall above zero. Increasing the depth of coverage, the performances of all tools increase with ABEMUS being always among the best performing tools for ctDNA level >=10% and outperforming all other tools for ctDNA levels <10%. Of note, performances reported in literature (Narzisi *et al.*, 2018) for the tools used in this comparison are in line with our results.

3.6 Comparison on *in-silico* dilutions of real cfDNA samples

The performances of ABEMUS were further investigated using **Synthetic Dataset #3,** which contains synthetic dilutions we computed from real data generated at high coverage and for a small target (Carreira *et al.*, 2014). ABEMUS was compared with Strelka2 and SomaticSniper - altogether the tools that in the previous analysis achieved reasonable results in a scenario that is similar to the one described by **Synthetic Dataset #3 -** and with SiNVICT, a tool designed for the ultra-sensitive detection of SNVs and InDels in cfDNA samples (Kockan *et al.*, 2017). To measure the performances of the four tools, we used as reference the overall set *S* of SNVs reported in the original study (Carreira *et al.*, 2014) that were manually reviewed and/or experimentally validated through ddPCR (i.e. SNVs in *AR*, *TP53*, *FOXA1* and *PTEN* genes). We defined the Positive Predictive Value (PPV), calculated as the number of SNVs in *S* that are detected over the total number of detected SNVs in *AR*, *TP53*, *FOXA1* and *PTEN* genes, the True Positive Rate (TPR), calculated as the number of SNVs in *S* that are detected over the total number of SNVs in *S*, and the product TPR*PPV. PPV, TPR and TPR*PPV were computed considering the set of calls across the four genes of interest performed by each tool across all set of 291 *in-silico* diluted samples. Although the optimal ABEMUS scaling factor *R* for **Synthetic Dataset #3** was 1.1 (for all synthetic samples), we also tested *R* values around the optimal value, specifically from 0.5 to 1.5. As shown in **Figure 6A**, SomaticSniper obtained the best results in terms of PPV for most ctDNA levels, but failed in terms of TPR and TPR*PPV, indicating very low sensitivity. SiNVICT, instead, obtained reasonable TPR but failed in terms of PPV and TRP*PPV, indicating a potential high fraction of false positives among the detected somatic SNVs. ABEMUS performed better than Strelka2 in terms of PPV for almost all the tested *R* values, with optimal scaling factor $R = 1.1$, demonstrating better PPV than Strelka2 at all ctDNA levels except for the lowest one, where PPV values resulted equal. In terms of TPR values ABEMUS and Strelka2 resulted in similar performances, with better ABEMUS results at lower *R* values. ABEMUS was the best tool in terms of TPR*PPV for most scenarios and for the majority of *R* scaling factors, with optimal scaling factor $R = 1.1$ performing better than Strelka2 in all conditions except for ctDNA level equal to 2.5%, were the two TPR*PPV values resulted equal. Overall, ABEMUS demonstrated the best performances among the majority of tested conditions, especially when pre-computed optimal scaling factor *R* was applied.

3.7 Performances on real cfDNA sequencing data

We finally compared ABEMUS and Strelka2 on a set of serial plasma samples (Carreira *et al.*, 2014). Performances of both tools were tested relying on detection of SNVs annotated in previously relevant studies (Abida *et al.*, 2019; Robinson *et al.*, 2015) or in COSMIC (Forbes *et al.*, 2017); for COSMIC only variants annotated as *confirmed somatic variants* and with primary site *Prostate* were considered. Scaling factors *R* optimized for mean coverage and target size were used. As shown in **Figure 6B** we observed high concordance between ABEMUS and Strelka2, but ABEMUS was able to detect SNVs in positions at low AF were Strelka2 was not. Among the 3 calls performed only by ABEMUS, two were also validated in the original study and present in other samples from the same patient. These two SNVs were identified in patient V4023, the first in *TP53* gene with allelic fraction 0.014 and protein change Cys135* in sample 11-244-B with estimated ctDNA of 13.1%, while the second in gene *FOXA1* with allelic fraction 0.016 and protein change Asp226His in the sample 10-315-B with estimated ctDNA of 15.5%. The remaining SNV identified at low allelic fraction only by ABEMUS was found in another sample from the same patient V4012 by both tools, strongly supporting the validity of the ABEMUS private call.

Considering that optimized scaling factors resulted in $R=1.1$ for all plasma serial samples, we also tested to what extent the knowledge of ctDNA level would have improved ABEMUS calls. Considering ctDNA levels reported in the original study (Carreira *et al.*, 2014), ABEMUS was run again on all plasma samples with results that were overall concordant. ABEMUS was in this case able to identify in patient V4048 a further SNV with an allelic fraction concordant with another SNV captured by both tools in the same sample and in the same gene.

Overall, both ABEMUS and Strelka2 achieved good results but ABEMUS demonstrated increased power in detecting low allelic fraction SNVs in low ctDNA levels plasma samples. In addition, upfront knowledge of

sample's ctDNA levels could be used to further improve detection sensitivity.

4 Discussion

Different approaches have been proposed in the past years to characterize somatic mutations in cfDNA. While methods like optimized quantitative PCR (Taly et al., 2012) or dPCR are highly sensitive (Didelot et al., 2013; Yu et al., 2017), they are limited in the number of mutations to test via multiplexing, while requiring up to 3ng of input DNA. Next-generation sequencing approaches can instead be used to screen a large number of mutations with sensitivity that is limited by background noise and dependent on the sequencing depth. Although recent studies (Mouliere et al., 2018) suggested that fragment size selection might improve somatic SNVs detection sensitivity, highly sub-clonal somatic SNVs due for instance to intra-patient heterogeneity or treatment resistance would remain extremely difficult to detect. In this challenging scenario, tools designed to detect low allelic fraction variants (Carrot-Zhang and Majewski, 2017) or computational pipelines specifically tailored for cfDNA data are necessary. So far, cfDNA specific approaches were either tuned for amplicon based NGS targeted platforms (Kleftogiannis et al., 2019; Pécuchet et al., 2016) or yet partially benchmarked against standard SNVs methods across different scenarios of coverage depth and target size (Kockan et al., 2017), potentially limiting their widespread applicability.

Here we presented a new NGS-based computational method named ABEMUS that uses control samples to build global and local sequencing error reference models that are used to improve the detection of SNVs in cfDNA samples.

We showed that local sequencing error, namely the *per-base error measure*, is platform specific and that hence platform specific sequencing error reference models are needed to effectively discriminate between true SNVs and artefactual signals in the challenging cfDNA scenario. In this respect, ABEMUS provides an automatic approach to build platform specific reference models from NGS control samples.

We showed that ABEMUS sequencing errors reference models are stable across a broad range of depth of coverage and target size scenarios and we optimized, across the same scenarios, the precision and recall of ABEMUS SNVs detection engine.

ABEMUS performances were tested against tools commonly used to identify SNVs in tumor tissue samples and against tools specifically designed for cfDNA samples using synthetic data, cancer patients cfDNA data *in-silico* diluted and cancer patients multi-sample cfDNA data. Overall, we showed that ABEMUS improves the detection of low allelic fraction SNVs in low ctDNA levels plasma samples in scenarios spanning from whole-exome data (tens of Mb) to small targeted panels data (tens of Kb). Of note, a limitation of the current version of ABEMUS is the absence of a module for the detection of InDels. ABEMUS is easy to use, can be applied on any custom or commercial platform or gene panel and can be integrated in any NGS processing and analysis pipeline.

Acknowledgements

We thank the members of the Caryl and Israel Englander Institute for Precision Medicine (WCM) for fruitful discussions and the LaBSSAH-CIBIO Next Generation Sequencing Facility of the University of Trento for input on the True Seq Custom Amplicon assay.

Funding

This work has been supported by Fondazione Cassa di Risparmio Trento e Rovereto (CARITRO; F.D.), National Cancer Institute SPORE P50-CA211024 (H.B., F.D.), Cancer Research UK A13239 (G.A.), Prostate Cancer UK PG12-49 (G.A., F.D.).

Conflict of Interest: none declared.

References

Abida,W. *et al.* (2019) Genomic correlates of clinical outcome in advanced prostate cancer. *Proc Natl Acad Sci U.S.A.*, **116**, 11428–11436.

Annala,M. *et al.* (2018) Circulating Tumor DNA Genomics Correlate with Resistance to Abiraterone and Enzalutamide in Prostate Cancer. *Cancer Discov*, **8**, 444–457.

Beltran,H. *et al.* (2016) Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat Med*, **22**, 298–305.

Beltran,H. *et al.* (2015) Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response. *JAMA Oncol*, **1**, 466.

Bettegowda,C. *et al.* (2014) Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. *Sci Transl Med*, **6**, 224ra24-224ra24.

Carreira,S. *et al.* (2014) Tumor clone dynamics in lethal prostate cancer. *Sci Transl Med*, **6**, 254ra125.

Carrot-Zhang,J. and Majewski,J. (2017) LoLoPicker: detecting low allelic-fraction variants from low-quality cancer samples. *Oncotarget*, **8**, 37032–37040.

Cibulskis,K. *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*, **31**, 213–219.

Cingolani,P. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.

Dawson,S.-J. *et al.* (2013) Analysis of Circulating Tumor DNA to Monitor Metastatic Breast Cancer. *N Engl J Med*, **368**, 1199–1209.

Didelot,A. *et al.* (2013) Multiplex Picoliter-Droplet Digital PCR for Quantitative Assessment of DNA Integrity in Clinical Samples. *Clin Chem*, **59**, 815–823.

Forbes,S.A. *et al.* (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*, **45**, D777–D783.

Heitzer,E. *et al.* (2019) Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nat Rev Genet*, **20**, 71–88.

Kim,S. *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods*, **15**, 591–594.

Kleftogiannis,D. *et al.* (2019) Identification of single nucleotide variants using position-specific error estimation in deep sequencing data. *BMC Med Genomics*, **12**, 115.

Koboldt,D.C. *et al.* (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*, **22**, 568–576.

Kockan,C. *et al.* (2017) SiNVICT: ultra-sensitive detection of single nucleotide variants and indels in circulating tumour DNA. *Bioinformatics*, **33**, 26–34.

Kwapisz,D. (2017) The first liquid biopsy test approved. Is it a new era of mutation testing for non-small cell lung cancer? *Ann Transl Med*, **5**, 46–46.

Larson,D.E. *et al.* (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, **28**, 311–317.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Mouliere,F. *et al.* (2018) Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med*, **10**.

Narzisi,G. *et al.* (2018) Genome-wide somatic variant calling using localized colored de Bruijn graphs. *Commun Biol*, **1**, 20.

Pécuchet,N. *et al.* (2016) Analysis of Base-Position Error Rate of Next-Generation Sequencing to Detect Tumor Mutations in Circulating DNA. *Clin Chem*, **62**, 1492–1503.

Prandi,D. *et al.* (2014) Unraveling the clonal hierarchy of somatic genomic aberrations. *Genome Biol*, **15**, 439.

Ramos,A.H. *et al.* (2015) Oncotator: cancer variant annotation tool. *Hum Mutat*, **36**, E2423-2429.

Robinson,D. *et al.* (2015) Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell*, **162**, 454.

Romanel,A. *et al.* (2015) Plasma *AR* and abiraterone-resistant prostate cancer. *Sci Transl Med*, **7**, 312re10-312re10.

Sclafani,F. *et al.* (2018) KRAS and BRAF mutations in circulating tumour DNA from locally advanced rectal cancer. *Sci Rep*, **8**, 1445.

Siravegna,G. *et al.* (2015) Erratum: Clonal evolution and resistance to EGFR blockade in the blood of colorectal cancer patients. *Nat Med*, **21**, 827–827.

Taly,V. *et al.* (2012) Detecting biomarkers with microdroplet technology. *Trends Mol Med*, **18**, 405–416.

Thierry,A.R. *et al.* (2014) Clinical validation of the detection of KRAS and BRAF mutations from circulating tumor DNA. *Nat Med*, **20**, 430–435.

Tie,J. *et al.* (2016) Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. *Sci Transl Med*, **8**, 346ra92-346ra92.

Valentini,S. *et al.* (2019) PaCBAM: fast and scalable processing of whole exome and targeted sequencing data. *BMC Genomics*, **20**, 1018.

Vietsch,E.E. *et al.* (2017) Circulating cell-free DNA mutation patterns in early and late stage colon and pancreatic cancer. *Cancer Genet*, **218–219**, 39–50.

Yu,Q. *et al.* (2017) Multiplex picoliter-droplet digital PCR for quantitative assessment of EGFR mutations in circulating cell-free DNA derived from advanced non-small cell lung cancer patients. *Mol Med Rep*, **16**, 1157–1166.

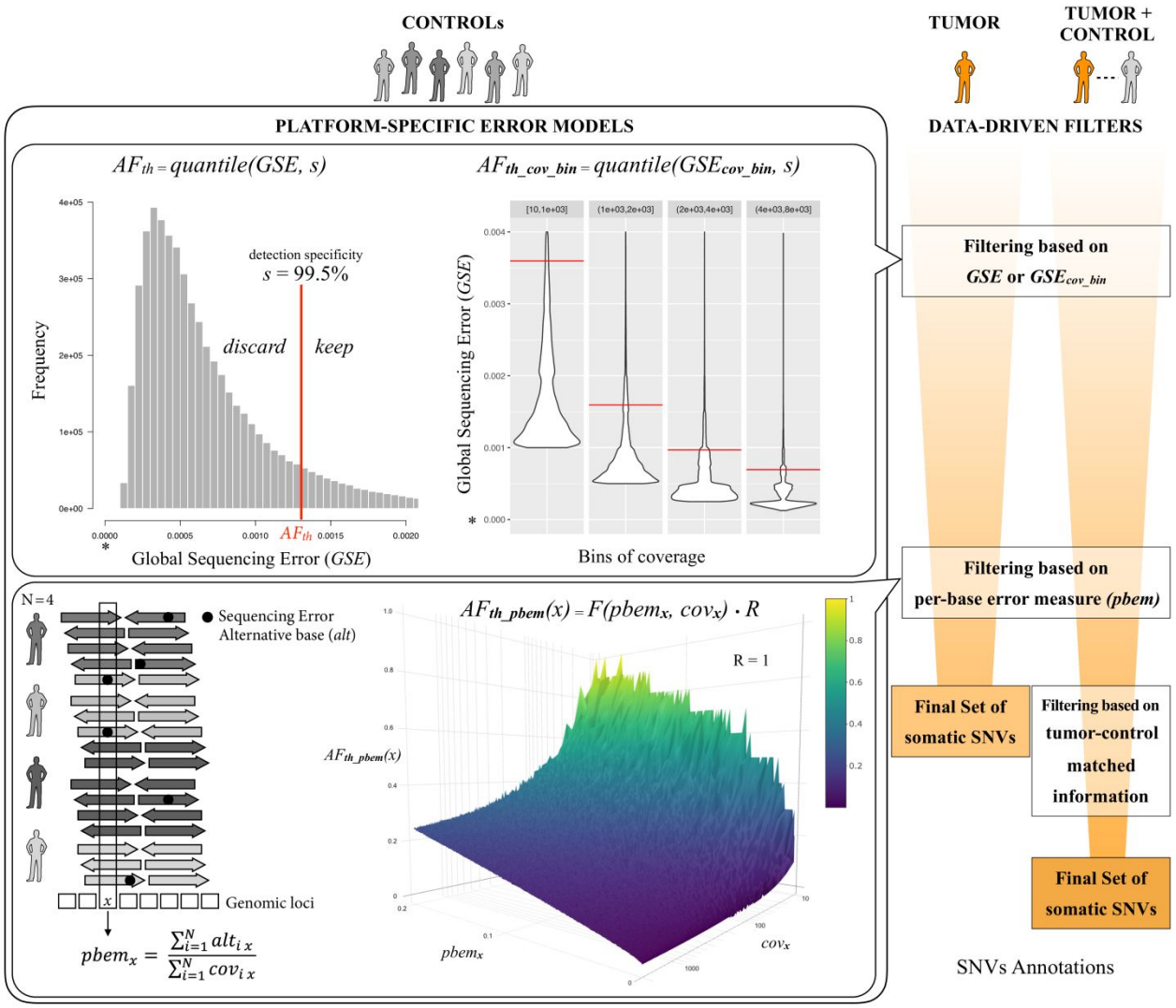

Figure 1. ABEMUS schematic workflow. ABEMUS inputs are pre-processed sequencing data from (matched) control and case samples (e.g. plasma, tumor tissue) profiled using the same platform (library preparation kit and sequencing machine). The computational workflow includes two separate steps. First, control samples are pooled and analyzed to estimate platform-specific error models: i) overall distribution of allelic fractions (AFs) (global sequencing error estimation, GSE) and ii) locus-specific error measure (per-base error measure, *pbem*). For an intended specificity level, the GSE is used to determine coverage-independent (AF_{th}) and coverage-dependent ($AF_{th_{cov_{bin}}}$) AF thresholds (GSEs corresponding to AF=0 not shown although considered in quantile estimations). For each position x, the AF threshold $AF_{th_{pbem}}(x)$ is computed as a function of the observed local $pbem_x$ and is dependent on the locus coverage cov_x and on the assay target size through a rescaling factor R. Second, for each case sample, ABEMUS nominates a set of putative somatic SNVs by filtering all available genomic positions having AF>0 using pre-computed global and local sequencing error estimations.

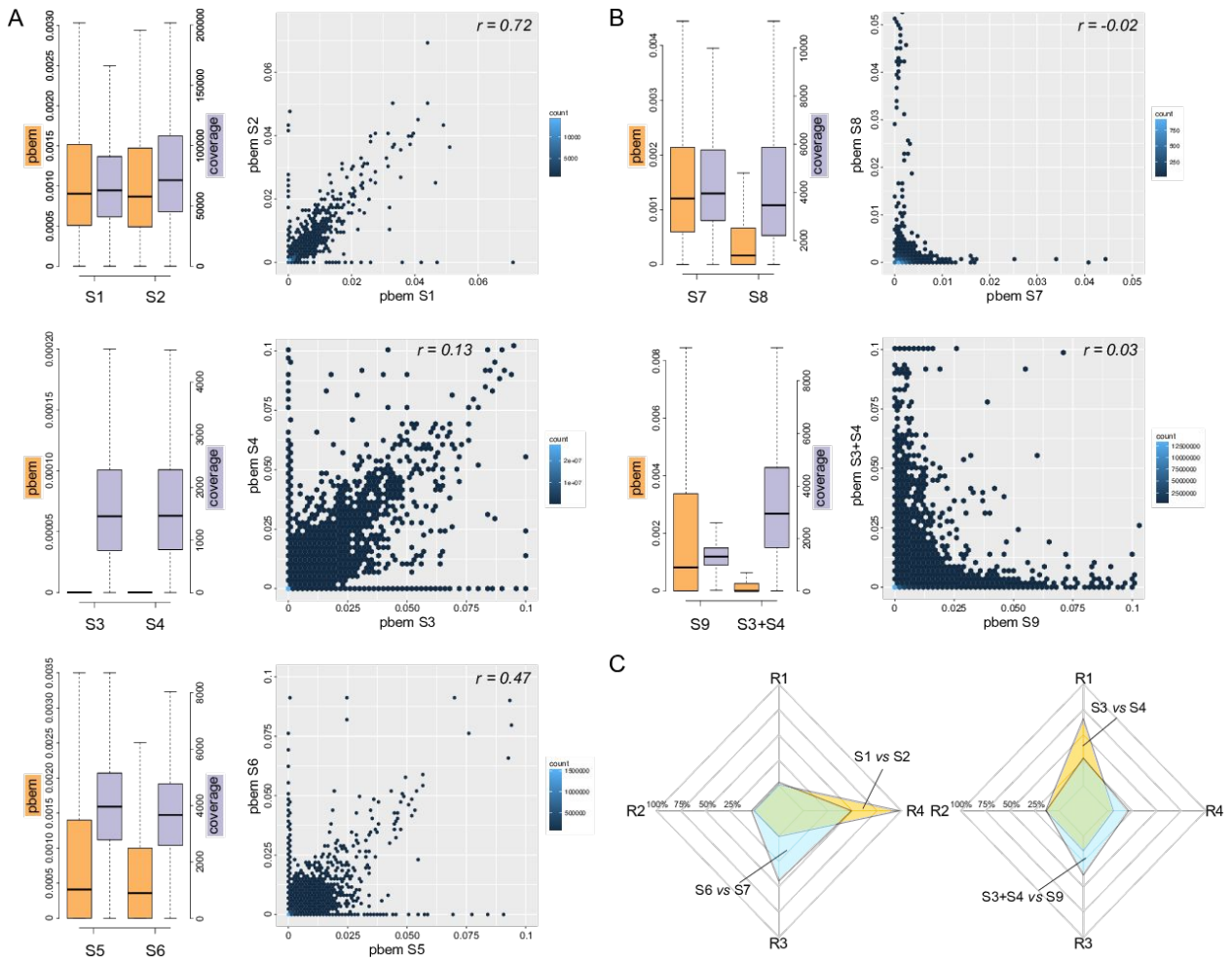

Figure 2. Estimation and comparison of *pbem* within and across platforms. (A) Correlations among *pbem* computed using disjoint sets of control samples sequenced on the same platform. Three platforms are considered. S1 (N=56) and S2 (N=57) are normal samples sequenced using Ion AmpliSeq Targeted Custom Amplicon panel (40kbp; IonTorrent PGM); S3 (N=20) and S4 (N=20) are control samples sequenced using Agilent HaloPlex Exome (36Mbp; Illumina HiSeq2000); S5 (N=10) and S6 (N=10) are control samples sequenced using Roche NimbleGen N250 targeted panel (3.2Mbp; Illumina HiSeq 2000). (B) Correlation among locus specific sequencing error probabilities when computed using same sets of control samples sequenced on different platforms. S7 (N=3) and S8 (N=3) loci shared (7 kbp) between targeted custom Ion AmpliSeq Targeted Custom Amplicon panel and Illumina True Seq Custom Amplicon; S9 (N=40) and S3+S4 (N=40) loci shared (26Mbp) between Roche NimbleGen SeqCap Exome v3 (64Mbp; Illumina HiSeq 2000) and Agilent HaloPlex Exome. (C) Proportion of concordant and discordant *pbems* when comparing samples profiled using the same (yellow polygons) or different platforms (light blue polygons). R1 and R4 axes indicate the proportion of loci characterized by two concordant *pbems* since they are both equal or greater than zero, respectively. R2 and R3 axes indicate the proportion of genomic loci with discordant *pbems*: a genomic locus showing the first *pbem* equal to zero and the second one greater than zero, or viceversa.

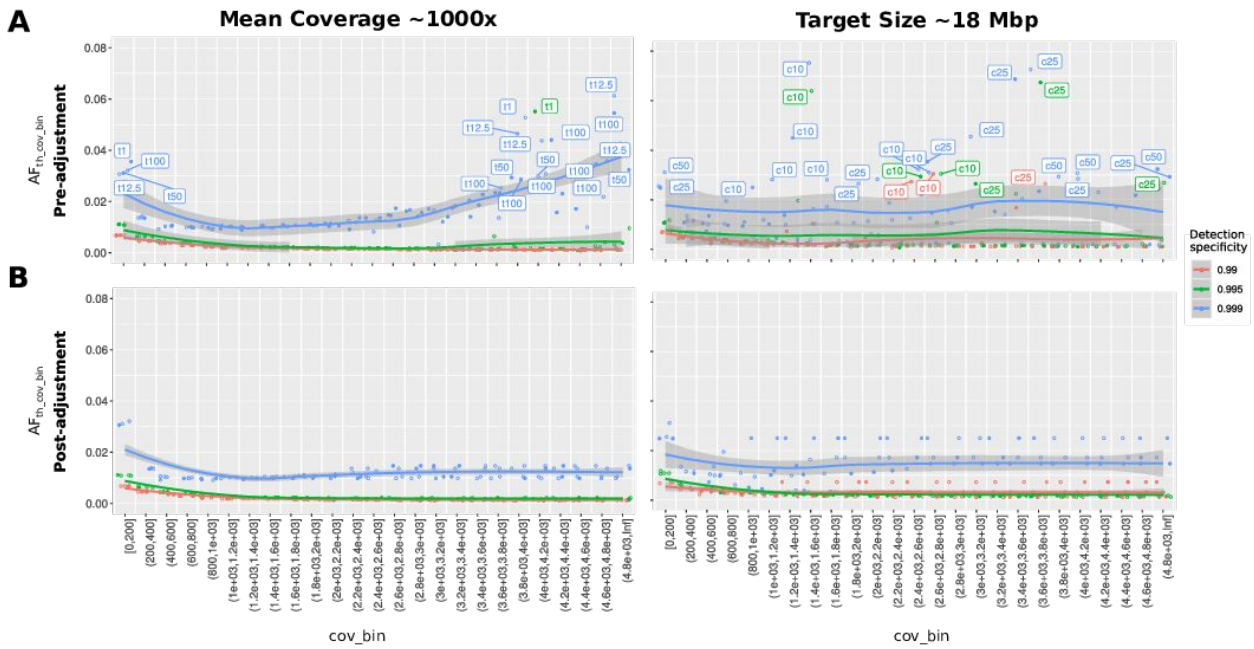

Figure 3. Estimation of coverage-dependent allelic fraction thresholds. Estimation of coverage-dependent allelic fraction thresholds ($AF_{th\_cov\_bin}$) at different detection specificities s across multiple target sizes (left) and depths of coverage (right). Samples with depth of coverage of 1000x (left) and target size of 18Mbp (right) are considered. (**A**) Original $AF_{th\_cov\_bin}$ estimations are affected by high variability in poorly populated coverage bins (cov_bin) (see **Figure S1**). (**B**) $AF_{th\_cov\_bin}$ estimations after ad-hoc procedure is applied resulted in a stable trend. Labels: "tx" denotes the genomic fraction "x" of original HaloPlex panel covered (i.e. t50 indicates that 50% of the base covered by the original HaloPlex panel has been kept); "cx" denotes the fraction "x" of original depth of coverage (i.e. c10 indicates that 10% of the original total number of sequencing reads has been kept).

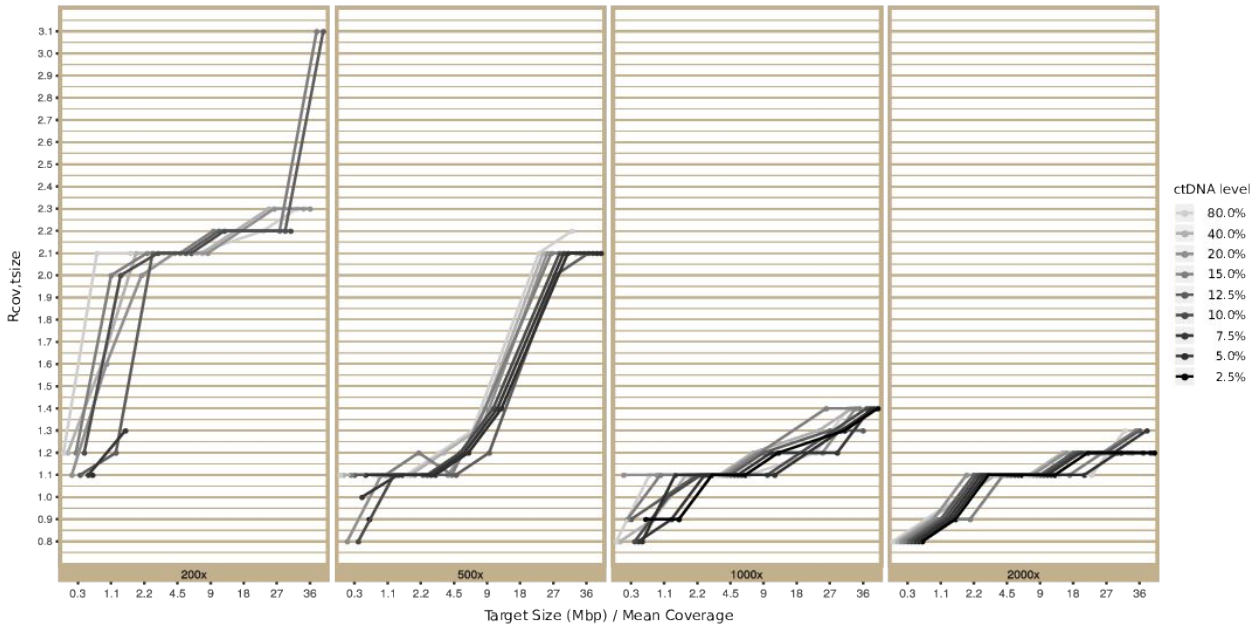

Figure 4. Assessment of the optimal scaling factor *R* to maximize ABEMUS performances for combinations of target size, coverage and ctDNA level. The y-axis reports scaling factors *R* and the x-axis indicates target sizes at four different coverage levels. Dots and lines refer to the ctDNA level tested. Given a combination of coverage, target size and ctDNA level, each dot indicates the optimal *R* scaling factor to be applied to get a F1-score ≥ 0.98. The wider the genomic target and the higher the mean coverage, the lower the optimal R required to get the desired F1-score.

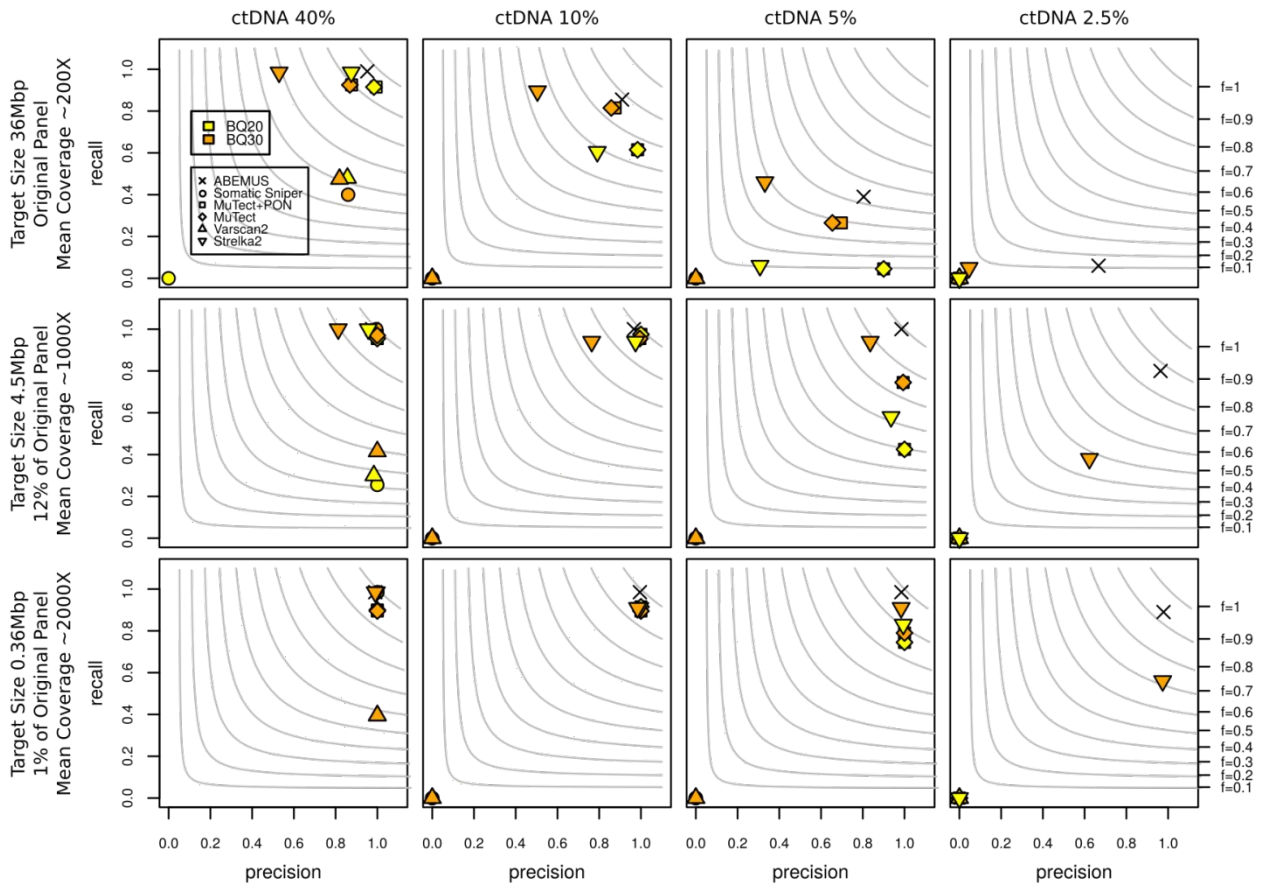

Figure 5. Comparative performance analysis among ABEMUS and other SNV callers based on synthetic data. Precision and recall measures are reported on x and y axes, respectively. Grey curves represent F1-scores as annotated to the right. Shapes represent tools; colors denote base quality ("BQ") of 20 (yellow) and 30 (orange). Decreasing levels of ctDNA (40%, 10%, 5% and 2.5%) are shown from left to right. Top: performances obtained on target size of 36Mbp (100% of original HaloPlex panel) and 200x mean depth of coverage. Middle: performances on target size of 4.5Mbp (12% of original HaloPlex panel) and 1000x mean depth of coverage. Bottom: performances on target size of 0.36Mbp (1% of original HaloPlex panel) and 2000x mean depth of coverage.

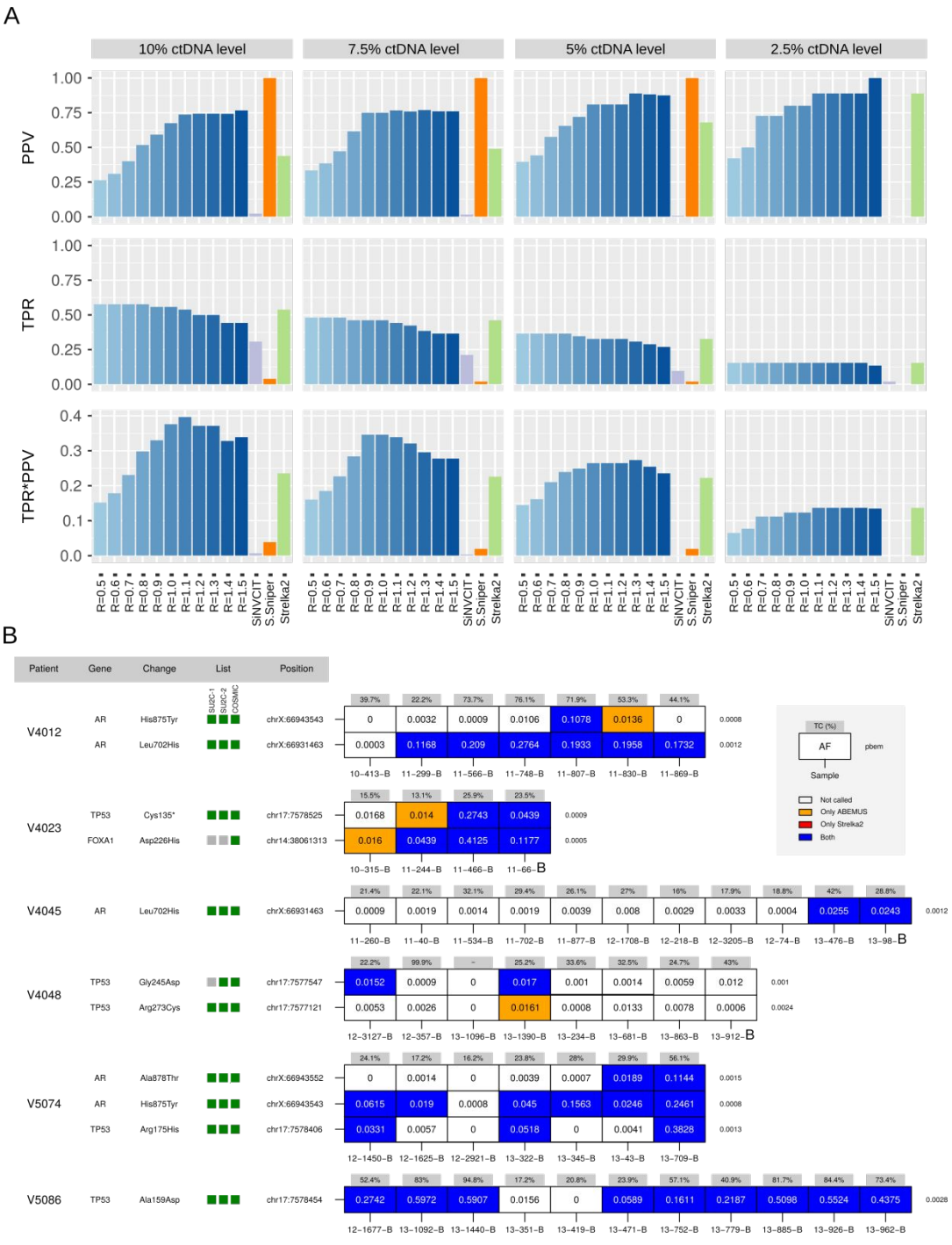

Figure 6. Performance on cancer patient's plasma data. (**A**) Barplot showing performances of ABEMUS applying scaling factors *R* from 0.5 to 1.5 (blue), SiNVICT (purple), SomaticSniper (orange) and Strelka2 (green) on synthetically diluted cancer patient's plasma samples. From top to bottom performances in terms of positive predictive value (PPV), true positive rate (TPR) and the product PPV*TPR are shown on y-axes, respectively. Decreasing levels of ctDNA (10%, 7.5%, 5% and 2.5%) are shown in grey boxes from left to right. (**B**) Overview of ABEMUS and Strelka2 calls on real plasma data. Only genomic positions annotated (green boxes in "List" column) in relevant published studies (Abida *et al.*, 2019; Robinson *et al.*, 2015) or in COSMIC are reported. For each patient, if an annotated genomic position is identified as SNV by ABEMUS or Strelka2 in at least one serial sample, all samples data are shown. TC: tumor content (ctDNA level); AF: allelic fraction; pbem: local per-base error measure.