Identification of single nucleotide variants using position-specific error estimation in deep sequencing data

Dimitrios Kleftogiannis[1], Marco Punta[1], Anuradha Jayaram[2], Shahneen Sandhu[3], Stephen Q. Wong[3], Delila Gasi Tandefelt[4], Vincenza Conteduca[5], Daniel Wetterskog[2], Gerhardt Attard[2,*] and Stefano Lise[1,*]

[1] Centre for Evolution and Cancer, The Institute of Cancer Research, London, UK

[2] UCL Cancer Institute, University College London, London, UK

[3] Peter MacCallum Cancer Centre and University of Melbourne, Melbourne, Victoria, Australia.

[4] Department of Urology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[5] Department of Medical Oncology, Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori (IRST) IRCCS, Meldola, 47014, Italy

*Corresponding author

Abstract

Background

Targeted deep sequencing is a highly effective technology to identify known and novel single nucleotide variants (SNVs) with many applications in translational medicine, disease monitoring and cancer profiling. However, identification of SNVs using deep sequencing data is a challenging computational problem as different sequencing artifacts limit the analytical sensitivity of SNV detection, especially at low variant allele frequencies (VAFs).

Results

To address the problem of relatively high noise levels in amplicon-based deep sequencing data (e.g. with the Ion AmpliSeq technology) in the context of SNV calling, we have developed a new bioinformatics tool called AmpliSolve. AmpliSolve uses a set of normal samples to model position-specific, strand-specific and nucleotide-specific background artifacts (noise), and deploys a Poisson model-based statistical framework for SNV detection. Our tests on both synthetic and real data indicate that AmpliSolve achieves a good trade-off between precision and sensitivity, even at VAF below 5% and as low as 1%. We further validate AmpliSolve by applying it to the detection of SNVs in 96 circulating tumor DNA samples at three clinically relevant genomic positions and compare the results to digital droplet PCR experiments.

Conclusions

AmpliSolve is a new tool for *in-silico* estimation of background noise and for detection of low frequency SNVs in targeted deep sequencing data. Although AmpliSolve has been specifically designed for and tested on amplicon-based libraries sequenced with the Ion Torrent platform it can, in principle, be applied to other sequencing platforms as well. AmpliSolve is freely available at https://github.com/dkleftogi/AmpliSolve.

Keywords

Next generation sequencing (NGS), cancer genomics, variant calling, deep sequencing, targeted sequencing, Ion Torrent, liquid biopsies, error correction

Background

Targeted next-generation sequencing (NGS) is a powerful technology to identify known and novel variants in selected genomic regions of interest [1]. It allows achieving high coverage levels (i.e., higher than 1000x) and, in principle, to confidently identify variants even when they occur at low allele frequencies. This is particularly important in cancer research and has many clinical applications, e.g. in relation to disease management. Typically, tumor cells are heterogeneous consisting of multiple clones and sub-clones the relative abundance of which can change over time depending on several factors, including treatment [2]. Identification of low frequency mutations is clinically relevant, among other reasons, for early diagnosis, disease monitoring and timely detection of the emergence of resistance clones under treatment [3].

Over the past years, it has been established that cancer patients' circulating free DNA (cfDNA) contains tumor-derived DNA fragments (ctDNA) that can be used as an alternative to needle biopsies in clinical settings [4]. However, identifying cancer-specific mutations in liquid biopsy samples is challenging, as the relative proportion of ctDNA in cfDNA can be low, especially at cancer's early stages. There are also several sources of sequencing errors including PCR artifacts, often reaching up to 1% Variant Allele Frequency (VAF), that reduce further the analytical sensitivity for detecting cancer-associated mutations [4]. Error correction techniques can be incorporated into NGS assays enabling ultra-sensitive single nucleotide variant (SNV) detection (VAF ~ 0.1%) but at a significant extra cost [5,6]. Thus, there is a need to reliably detect SNVs in more conventional deep sequencing data.

In silico identification of SNVs from NGS data is a well-studied problem [7,8]. However, the majority of existing variant calling programs have been designed for whole-exome and whole-genome experiments sequenced at coverage of approximately 30x to 100x. At the same time, available variant calling software for targeted deep sequencing experiments have been typically developed for and tested on Illumina data [9].

Compared to Illumina, Ion Torrent sequencing has a higher per base error rate and an associated lower accuracy in identifying mutations [10, 11]. However, it has the advantage of requiring lower amounts of input DNA and it offers both reduced cost and turnaround time. Thus, it is a cost-effective strategy for screening large cohorts of patients and it is particularly suited for point-of-care clinical applications [1], for example in conjunction with the Ion AmpliSeq Cancer Hotspot Panel. Given its translational potential, there is a real need to improve the variant calling workflow and recently a number of methods have been developed to deal specifically with Ion Torrent data [12, 13, 14].

Here we introduce AmpliSolve, a new bioinformatics method to detect SNVs in targeted deep sequencing data. It combines *in-silico* background error estimation with statistical modeling and it is particularly suited to deal with data of comparatively high noise levels, similar to the ones produced by the Ion AmpliSeq library preparation. In order to estimate background noise levels per position, strand and nucleotide substitution, AmpliSolve takes as input deep-sequencing data from a set of normal samples. This information is then fed to a Poisson model for the identification of SNVs. Experimental results using normal samples (self-consistency test), synthetic variants and clinical data sequenced with a custom Ion AmpliSeq gene panel, demonstrate that AmpliSolve achieves a good trade-off between precision and sensitivity, even for VAF values below 5% (and as low as 1%).

Methods

Method overview

AmpliSolve consists of two main programs written in C++: AmpliSolveErrorEstimation and AmpliSolveVariantCalling. AmpliSolveErrorEstimation requires the availability of a set of normal samples processed with the deep sequencing platform and panel of choice. Here, we focus on the Ion Torrent Personal Genome Machine (PGM) and a custom AmpliSeq panel, a technology known to have relatively high rates of sequencing error compared to others. The program uses the normal samples to infer position-specific, nucleotide specific and strand-specific background sequencing error levels (noise) across the targeted regions. Execution of AmpliSolveErrorEstimation is performed only once per panel design. Error estimates are then used as input to the AmpliSolveVariantCalling program for SNVs' detection. The procedures for *in-silico* noise estimation and SNV identification are described below. In Figure 1 we present a graphical overview of the AmpliSolve computational workflow.

In-silico identification of the background sequencing error

Our strategy for estimating background error levels, implemented in the AmpliSolveErrorEstimation program, is based on the assumption that alternative alleles observed at VAF<5% in normal samples are, in the majority of cases, the result of sequencing errors (see Figure S1 for the distribution of non-reference allele frequencies in normal samples showing the separation between heterozygous germline variant and lower frequency ‘noise’ variants). Accordingly, we utilize a set of normal samples to estimate background noise in our custom panel. Notably, we estimate error levels separately for each genomic position, each nucleotide (alternative allele) and each of the two (forward and reverse) strands. Thus, for each genomic position we generate six error estimates (i.e. two each for the three alternative alleles). Error estimates are fed to a Poisson model, which is then used to calculate the p-value of the observed substitutions representing true variants versus them being noise. The detailed implementation is as follows. We first extract “raw” counts for each position, alternative allele and strand from the BAM files [15] of a set of N normal samples using the ASEQ software [16]. We run ASEQ with the quality parameters suggested by the authors of a previous study based on Ion AmpliSeq data [17], namely: minimum base quality = 20, minimum read quality = 20 and minimum read coverage = 20. At every genomic position, we estimate the background error s separately for each alternative allele α and strand (+ or -) by calculating the fraction of reads carrying the alternative allele on a given strand across all normal samples. More specifically we use the following formula:

$$s^{\alpha,+/-} = \frac{Er^{\alpha,+/-}}{Erd^{+/-}} + C \qquad (1)$$

with

$$Er^{\alpha,+/-} = \sum_{i=1}^{N} R_i^{\alpha,+/-} \qquad (1a)$$

and

$$Erd^{+/-} = \sum_{i=1}^{N} RD_i^{+/-} \qquad (1b)$$

We denote with $R_i^{\alpha,+}$ and $R_i^{\alpha,-}$ the number of reads supporting the alternative allele α on the forward and reverse strand, respectively, in normal sample *i*. We denote with RD_i^+ and RD_i^- the total number of reads (read depth) at the genomic position of interest on the forward and reverse strand, respectively, in normal sample i. Summations are taken over all normal samples utilized for the error estimation. C in equation (1) is a constant pseudo-count parameter that is introduced to mitigate the problem of positions in which the alternative allele read count might be underestimated (e.g. due to a relatively low read depth at a given position in the normal samples). Note that if at a given genomic position a normal sample has an alternative allele with VAF > 5%, we do not consider it for the summations in (1a) and (1b) for that position. In fact, a frequency of >5% is more likely to represent either a real variant (i.e. a single nucleotide polymorphism) or a particularly 'noisy' position in the sample and thus we opt to leave it out of the error estimation. For a given position, we additionally discard normal samples that have <100 reads on the forward or on the reverse strand (low coverage). Positions for which 2/3 or more of the normal samples are excluded from the error estimation are considered non-callable. Note that non-callable positions may include cases in which an alternative allele is frequent in the general population or over-represented in the specific set of normal samples used for the error estimation. However, given that AmpliSolve main goal is the identification of somatic mutations this does not constitute a major limitation.

SNV detection using a cumulative Poisson distribution

For every alternative allele substitution on a given + or - strand with non-zero variant read count $k^{+/-}$ in a sample of interest, the AmpliSolveVariantCalling program uses a Poisson model to calculate the probability that $k^{+/-}$ or more variant reads are produced by sequencing errors, i.e. the p-value. At any given position, the calculated p-value is a function of the normal sample-based sequencing error $s^{+/-}$ from previous section and of both the number of reads $k^{+/-}$ supporting the alternative allele on a given strand and of the read depth $K^{+/-}$ for the same strand in the sample of interest. In particular:

$$p - value(variant \mid k^{+/-}, K^{+/-}, s^{+/-}) = 1 - \; e^{-K*s} \sum_{i=0}^{k-1} \frac{(K*s)^i}{i!} \qquad (2)$$

Where, for better readability, on the right side of the equation we have omitted all +/- symbols for k, K and s. We observe that $K*s$ is the expected number of random substitutions for a depth of coverage K or the mean of the Poisson distribution. Note that p-values are not corrected for multiple testing. The strand-specific p-values are finally converted to quality scores using the formula *Q=-10*log10(p-value)*. In its output, for all positions in the panel carrying substitutions with Q score equal to or greater than 5 on both strands, AmpliSolve reports the average Q score between the two strands. Vice versa, positions with no substitutions or with substitutions with associated Q score lower than 5 on one or both strands are not reported in AmpliSolve's output. All reported SNVs are further tested for and potentially assigned one or more of the following warning flags:

a) 'LowQ' if the Q score is lower than 20 in at least one of the two strands.
b) 'LowSupportingReads' if the SNV is supported by less than 5 reads per strand in the tumour samples being analysed.
c) 'AmpliconEdge' if the SNV is located within overlapping amplicon edge regions, which may result in sequencing artifacts.
d) 'StrandBias' if the SNV is associated to a strand-bias. We apply Fisher's exact test to each SNV under the null hypothesis that the number of forward and reverse reads supporting the variant should be proportional to the total number of reads sequenced in the forward and reverse strands respectively. The flag is assigned to substitutions for which the p-value of the Fisher's exact test is lower than 0.05.
e) 'HomoPolymerRegion' if the SNV is located within a homopolymer region using the same criteria as in [18].
f) 'PositionWithHighNoise' if the SNV is supported by more than 5 reads per strand but the associated VAF is lower than the maximum VAF at this position across all normal samples in the training set.

If no warning is issued, AmpliSolve assigns a 'PASS' quality flag to the SNV.

Performance measures

To assess AmpliSolve's success in detecting SNVs, we use a number of performance metrics:

1. Sensitivity or True Positive Rate (TPR) = TP / (TP+FN)
2. Precision or Positive Predictive Value (PPV)= TP/(TP+FP)
3. False Discovery Rate (FDR) = 1-PPV=FP / (FP+TP)

Where, TP is the number of True Positive predictions, FN is the number of False Negative predictions and FP is the number of False Positive predictions.

Clinical data used in this study

For the development and evaluation of AmpliSolve, we have access to an extensive collection of clinical samples from castration-resistant prostate cancer (CRPC) patients, part of which had been already presented in previous publications [17,19,20,21]. The collection comprises 184 germline samples (buccal swabs or saliva) and more than 450 liquid biopsy plasma samples (note that for some patients there are multiple liquid biopsies and a small minority of liquid biopsy samples has no matched normal). In practice for this study, we rely on all the 184 normal samples but only use 96 liquid biopsy samples for which results from digital droplet PCR (ddPCR) assays are available (see below). For 5 additional patients we have access to 10 solid tumour samples from metastatic sites (1, 2 and 3 samples from respectively 1, 3 and 1 patients) and their associated 5 germline samples. For the available samples, we have the following data:

a) For all samples (germlines, liquid biopsies and solid tumors), we have Ion Torrent sequencing data obtained using a custom Ion AmpliSeq panel of 367 amplicons spanning 40,814 genomic positions at around 1000-1500x coverage. The panel targets both intronic and exonic regions in chromosomes 8, 10, 14, 17, 21 and X including commonly aberrated genes such as *PTEN*, *CYP17A1*, *FOXA1*, *TP53*, *SPOP* as well as the androgen receptor (*AR*) gene, which is one of the main drivers of CRPC, and the drug target *CYP17A1*. More details about the sequencing protocol, data processing and additional information about the application of our custom Ion AmpliSeq panel in CRPC diagnostic studies can be found in [17] and [19]. These papers also include a description of a variant caller that we used as starting point for developing AmpliSolve. We call variants in these Ion AmpliSeq data with our program AmpliSolve.

b) For the 10 solid tumor samples with 5 matched germline samples, in addition to Ion Torrent data, we have Illumina Whole Genome Sequencing (WGS) data at around 80-100x (tumor) and 30x (germline) coverage. We call variants in WGS data according to a previously established pipeline [22], which we describe in the next section.

c) For 96 liquid biopsy samples we have results from ddPCR assays to screen 3 clinically relevant SNVs in the *AR* gene. These SNVs have been linked to resistance to targeted therapy in CRPC patients, namely: 2105T>A (p.L702H), 2226G->T (p.W742C) and 2632A>G (p.T878A). ddPCR in the plasma samples was performed using 2-4 ng of DNA, using Life Technologies Custom Taqman snp genotyping assay (AH0JFRC), C_175239649_10 and C_175239651_10, respectively. Following droplet generation (AutoDg, Bio-Rad) and PCR, samples were run on the Bio-Rad QX200 droplet reader and analyzed using the QuantaSoft software.

WGS variant calling pipeline

Illumina Whole Genome Sequencing (WGS) data have been processed using standard tools, such as Skewer [23] for adapter trimming, BWA-MEM [24] for mapping and Picard [25] for duplicate removal. In order to call SNVs we run a previously developed pipeline [22] that utilizes jointly Mutect [26] and Platypus [18] (throughout the manuscript this pipeline is denoted as *MutPlat*). Briefly, we first run Mutect (default parameters) on each paired tumour-normal samples. Then, we use Mutect's calls as priors for Platypus and jointly call variants on all tumors and matched normal samples of a patient.

Germline variants are identified as those variants called in the normal (GT=0/1 or 1/1) with a PASS filter. Additionally, we accept a variant with any flag (e.g. 'badReads') if present in 1000 genomes. For AmpliSolve validation purposes we consider only tumor samples but include both germline and somatic SNVs. By including germline SNVs, in particular, we are able to test a higher number of low VAF mutations than would be possible when considering only somatic mutations. Indeed, while somatic deletions cause loss of some germline SNPs in tumour DNA, germline DNA contamination (i.e. <100% tumor purity) means that these mutations are still present in the tumor samples, albeit at lower VAF. In some cases, depending on tumor purity and depth of sequencing coverage, the detected VAF can even be nominally zero in the tumor.

To call a somatic SNV we require all of the following criteria to be met: i) Platypus filter: PASS, alleleBias, Q20, QD, SC or HapScore, ii) At least 3 reads supporting the variant in the tumor, iii) at least 10 reads covering the position in the germline and no support for the variant in germline (NV=0 and genotype GT= 0/0), iv) SNV not present in the 1000 genomes database

How to run AmpliSolve

AmpliSolve two modules, AmpliSolveErrorEstimation and AmpliSolveVariantCalling, can be downloaded from github (https://github.com/dkleftogi/AmpliSolve). Additional requirements include running versions of the programs Samtools [15], ASEQ [16] and the Boost libraries for C++. Here we provide a brief description of how to run AmpliSolve, however, more detailed information and a number of examples are available on the github page.

For a given amplicon panel, error estimation at each genomic position, for each alternative nucleotide and for each of the two strands requires availability of amplicon-based data from N normal sample files. Although we don't enforce a minimum value for N, values below 10 are likely to give low quality error estimations. In general, we suggest using as many normal samples as possible when training the error matrix for your panel. If normal samples are not available, AmpliSolveErrorEstimation sets 1% error estimation for all positions, nucleotides and strands in the panel. Note that AmpliSolveErrorEstimation does not take bam files as input but rather bam-derived read count files. Read count files can be obtained by running the program ASEQ [16]. Once the read count files have been produced, the user needs to set the value of the C pseudo-count parameter (equation (1)). The choice of C will depend on the trade-off between precision and sensitivity the user is interested in. Users can refer to the benchmarking experiments performed in this paper. In general, values of C between 0.001 and 0.01 should suit most applications.

Once the error matrix has been calculated, it can be fed to the AmpliSolveVariantCalling program together with read count files for the tumour samples again to be produced by running ASEQ. Note that AmpliSolveVariantCalling does not require matched normal-tumour samples for calling SNVs. In fact, AmpliSolveVariantCalling calls all variants it can find in the tumour sample, including germline variants. To separate germline from somatic variants users will need to run AmpliSolveVariantCalling on a matched normal sample and take the difference between the two output files. Command-line syntax for running AmpliSolveErrorEstimation and AmpliSolveVariantCalling is provided on github.

Results and Discussion

Sequencing error estimation, self-consistency test and AmpliSolve FDR

AmpliSolve estimates the background sequencing noise by analyzing the distribution of alternative allele in normal samples. As previously reported, PGM errors tend to be systematic [11] and AmpliSolve assigns a separate error level to each genomic position, each alternative allele and each strand (see Figure S2). These are then utilized to build the Poisson models that are at the core of AmpliSolve SNV calling (Methods). In this section, we study AmpliSolve variant calling performance as a function of two parameters: the pseudo-count *C* (equation (1) in Methods), and the number of normal samples *N* that are used to calculate the error estimations. C is the only user-adjustable parameter of our method. The number of samples N, instead, depends on sample availability for a given panel.

We perform a self-consistency test using sets of normal samples to train our models and other, non-overlapping normal samples for testing them. Given a dataset of N=184 normal samples (Methods), we proceed as follows: 1) we select a number M < N of samples at random and additionally a value *c* of the C parameter; 2) we use the M samples to train our Poisson-models with C=*c*; 3) we use the models obtained in 2) to predict SNVs in the remaining N-M samples; 4) we calculate FDR and TPR by defining as negatives all alternative alleles that have VAF<20% and as positives those for which VAF≥20%. This threshold is chosen empirically based on the distributions of VAFs that we observe in the data (Figure S1); 5) we repeat steps 1) to 4) 50 times for each pair of (M,*c*) values, each time selecting a new set of M samples at random; 6) we calculate median FDR and TPR over the 50 experiments. We perform steps 1) to 6) for all combinations of the following values of M (size of the training set) and *c* (parameter C): M=10, 20, 40, 80, 120 and c=10^{-5}, 5*10^{-5}, 10^{-4}, 5*10^{-4}, 10^{-3}, 2*10^{-3}, 5*10^{-3}, 10^{-2}, 2*10^{-2}. In Figure 2a and 2b, we plot the median FDR for each size of the training set (10-120) as a function of *c*; additionally, for comparison, we plot the median FDR of a method in which we skip the error estimation step and we set instead s=*c* for all positions, nucleotides and strands ('baseline caller') (see equation (1) in Methods for the definition of s). The FDR reported in Figure 2a is calculated by considering an SNV as called by AmpliSolve if and only if it has a Q score higher or equal 20 (i.e. p-value ≤ 0.01), while the FDR reported in Figure 2b is calculated by considering an SNV as called by AmpliSolve if and only if the program assigns a 'PASS' flag to it (that is, if none of the warnings described in the Methods applies). In Figure 2a we see that, for relatively small values of *c*, the training set size N affects the method performance, with more samples providing better error estimation and thus lower FDR. Also, our approach provides an approximately 2 to 4 fold FDR improvement over the baseline caller at all values of *c*≤0.01. For values of *c*>0.01, instead, differences with the baseline caller become negligible. Figure 2b shows that the additional warning flags that we introduce for filtering AmpliSolve's SNV calls (e.g. low number of supporting reads or homopolymer regions, see Methods) have the effect of further improving FDR values and, at the same time, of reducing differences between FDRs obtained when using training sets of different size. All of the above findings suggest that estimating the background noise at each position, for each nucleotide and for each strand is important for reducing the number of FPs arising from noise in Ion AmpliSeq data. If we now consider the median values of the Sensitivity measure (or TPR) across all of the above experiments, we discover that in all instances they are close to 1, irrespective of the value of M and *c*. This close to perfect Sensitivity is not surprising as our definition of positives (VAF≥20%) makes them relatively simple to discriminate from the background noise especially considering the fact that most of them have VAFs that are much higher than 20% (Figure S1). Thus, in order to truly test AmpliSolve Sensitivity, we have to perform a different kind of experiment, which we describe in the next section.

Synthetic variants test for TPR estimation

In order to test the sensitivity of our method at low VAFs (0.5% to 4%), we design the following experiment. We first select two amplicons spanning the *AR* gene (1,017 genomic positions overall); the *AR* gene is chosen because clinically relevant but for this purpose other choices would be equally valid. Then, we use 120 normal samples randomly selected from the full set of 184 described in Methods to estimate the errors at each position in the two amplicons, for each nucleotide and each strand, according to formula (1). Next, we test the method's sensitivity on synthetic variants. For each possible alternative allele at each of the 1,017 amplicon positions, we set read depth to a fixed value *COV* and the number of reads supporting the allele to a value 2*a (a* supporting reads on the forward strand and *a* on the reverse strand). We use COV=800, 1600, 3200, 6400 (values in this range apply to more than 60% of full panel positions with coverage >200, see Figure S3) and for each value of COV we select *a* corresponding to VAFs of 0.5%, 1%, 1.25%, 2%, 3% and 4%. For example for COV=800 we test *a*=2,4,5,8,12,16. We then apply the Poisson models previously trained on the 120 normal samples to predict variants at each position and for each alternative allele and consider only AmpliSolve calls with a 'PASS' quality flag (Methods). We consider all synthetic variants to be positives (thus, no FDR can be calculated in this case) and ask how many of these can be detected by AmpliSolve. We stress that while in each experiment the VAF is by design the same at all positions and for each alternative allele and strand, following estimation from the normal samples the error estimate is position-, alternative allele- and strand-dependent. We calculate the TPR for all combinations of *COV* and *VAF*. We do this for several values of the pseudo-count parameter C in the range of low AmpliSolve FDR as calculated from the self-consistency test in the previous section or the range of main interest for applications (C=0.001, 0.002, 0.005, 0.01, 0.02, see Figure 2b).

Figures 3(a-e) highlight the role of the C parameter as an approximate lower bound for AmpliSolve sensitivity (see equation (1)). Typically, AmpliSolve identifies few or no variants at allele frequencies equal to or lower than C, in the range of tested coverage depth (see, in particular, Figures 3c-e). For example, for C=0.005 no calls are made at VAF=0.5% even at values of COV as high a 6,400. Along the same lines, for values of C equal 0.01 and 0.02, which correspond to FDRs below 1.6% and 0.6%, respectively (Figure 2b), the lowest VAFs that AmpliSolve can detect are above 1% and 2%, respectively. For VAF values above C, on the other hand, sensitivity grows quickly with increasing VAF. For example within the depth of coverage range that we have analyzed, when using C=0.01 and C=0.02 AmpliSolve successfully calls the vast majority of synthetic variants at VAF 2% and 3%, respectively. When we compare the Sensitivity histograms in Figures 3a-e to the FDR curves in Figure 2b, we see that AmpliSolve can reliably predict synthetic SNVs at VAFs as low as 1% while still in a regime of relatively low FDR. Indeed for C=0.002, at an estimated FDR of 6.8% (Figure 2b), AmpliSolve calls most SNVs with 1% allele frequency at depth of coverage >1,600 and most SNVs with allele frequency 0.5% at depth of coverage >3,200. While it will be up to the user to select the best trade-off between FDR and Sensitivity for a specific experiment, it would appear that values of C between 0.001 and 0.01 would likely represent a reasonable compromise between these two performance measures in most applications.

Validating AmpliSolve calls via comparison with WGS Illumina data

For 5 additional CRPC patients, we have access to 10 metastatic solid tumour (for some patients more than one metastasis) and associated normal samples. These were sequenced with both our custom Ion AmpliSeq panel and the Illumina platform as WGS (the latter, with average coverage ~100X) (Methods). We use these 10 samples to provide a validation of AmpliSolve SNV calls in a more realistic set-up with respect to what shown in the previous two sections. For training our AmpliSolve Poisson models, we use the full set of 184 normal samples sequenced with the Ion AmpliSeq technology and set C=0.002.

In the solid tumor samples, when run on the Ion AmpliSeq data AmpliSolve identifies a total of 552 SNVs. For the same set of genomic positions processed by AmpliSolve, our WGS-variant calling pipeline MutPlat (Methods) calls a total of 603 SNVs in the corresponding Illumina data. The list of positions processed by AmpliSolve includes all those covered by our amplicon panel minus the ones for which no background error estimate can be produced (Methods). Almost all SNVs identified in the WGS data are germline (592 out of 603) but some of them have low VAF in the tumor samples because of deletions and loss of heterozygosity (LOH) events in the tumor DNA combined with germline DNA contamination. It is therefore a very valuable test set that includes confidently identified variants at low VAF.

The level of agreement between the two callers (AmpliSolve and MutPlat) on this data is summarised in Figure 4a. Of the 552 SNVs called by AmpliSolve, 525 SNVs are also identified by MutPlat. The remaining 27 are likely false positives although some of them might actually be real somatic variants with very low VAFs (and hence non detectable by a WGS done at 100x). The 78 SNVs additionally identified by MutPlat in the WGS are potential AmpliSolve false negatives. We note however that 49 of them correspond to positions not called because the coverage was below the threshold of 100 reads per strand. Had the coverage been higher, AmpliSolve would likely have identified these variants. Indeed some of these variants are called in metastasis from the same patient where higher coverage at these positions is available. The remaining 29 SNVs appear to be genuine false negatives, most of them (26) filtered out because of the strandBias filter. It is possible that this filter could be improved thus rescuing some of these calls without necessarily affecting AmpliSolve precision.

In Figure 4b and 4c we report a scatter plot of the VAFs in the WGS and AmpliSeq data with colors dots reflecting common calls (purple), WGS-only calls (green) and AmpliSolve-only calls (blue), respectively. Overall there is a good concordance between AmpliSolve and MutPlat calls, even at low VAF (Figure 4c). In particular, AmpliSolve correctly identifies 18 out 21 SNVs with VAF < 5% in the WGS calls. AmpliSolve does call a number of likely false positives at low VAF, however most of them occur at recurrent positions across patients and could therefore likely be identified and discarded at a post variant calling analysis stage

Clinical application using ctDNA samples and ddPCR for validation

One of the most promising clinical applications of ctDNA is profiling of specific mutations associated with tumour progression and resistance to cancer therapies. To evaluate AmpliSolve's usefulness for this important task, we use results from a ddPCR screen on 96 samples from our CRPC patients at three genomic positions within the *AR* gene, which are associated with resistance to targeted therapy (Methods). ddPCR detects 30 variants in total at these positions in a VAF range of 0.1 to 49% (note that in some experiments only the presence or absence of the variant was recorded). Next, we compare AmpliSolve calls at the same positions in the AmpliSeq NGS data for the same samples (predictions made after training AmpliSolve with pseudo-count parameter C=0.002 on 184 normal samples). In Figure 5a we summarize the level of agreement between AmpliSolve and the ddPCR experiments. AmpliSolve correctly calls 19 out of 30 ddPCR variants and predicts variants at two additional positions. If we take the ddPCR experiments as our standard of truth, this translates into ~90% precision (or 10% FDR) and ~63% sensitivity for AmpliSolve at these 3 clinically relevant genomic positions.

For comparison, we additionally predicted variants from the Ion Torrent data at these positions and in these samples using deepSNV, a state-of-the-art method for calling low VAF variants in ctDNA. Like AmpliSolve, deepSNV shows good performance values calling 15 out of 30 ddPCR variants and 3 additional variants. It should be stressed that while here we run deepSNV on Ion Ampliseq data the program is originally designed to work on Illumina data. When looking at AmpliSolve predictions in more details (Figure 5b), we note that all ddPCR positives not called by our program have VAF<1% in the NGS data, while AmpliSolve succeeds in calling all ddPCR positives at higher NGS frequencies including several at VAFs between 1% and 5%. It is also important to note that AmpliSolve correctly predicts 256 out of 258 (99.2%) ddPCR negatives. While the significance of the results presented in this section should not be exaggerated as they refer to only three genomic positions, they are an example of AmpliSolve's potential value in clinically relevant experiments.

Conclusions

In this study, we present AmpliSolve, a new bioinformatics method that combines position-specific, nucleotide-specific and strand-specific background error estimation with statistical modeling for SNV detection in amplicon-based deep sequencing data. AmpliSolve is originally designed for the Ion AmpliSeq platform that is affected by higher error levels compared to, for example, Illumina platforms. Our method is based on the estimation of noise levels from normal samples and uses a Poisson model to calculate the p-value of the detected variant. We assess AmpliSolve's performance with experiments that use normal samples (self-consistency tests) and simulated data (synthetic variants) and via a comparison that utilizes real metastatic samples sequenced with both Ion Torrent and Illumina platforms. In these experiments, AmpliSolve achieves a good balance between precision and sensitivity, even at VAF < 5%. These experiments also suggest possible ways to further improve the method, such as adopting a better strand bias filter and introducing a 'black list' of positions characterized by an unusual noise distribution across samples (e.g. bimodal). Further, we test AmpliSolve in a clinical relevant setting by calling SNVs in 96 liquid biopsy samples at 3 positions that had been additionally screened by ddPCR assay. In this experiment AmpliSolve successfully identifies SNVs at VAF as low as 1% in the NGS data. This opens up interesting possibilities for clinical applications using the Ion Torrent PGM such as, for example, tracking mutations in ctDNA to monitor treatment effectiveness and/or disease relapse. In general we believe that the use of models with position-specific error estimates, as described here, could improve SNV detection especially at low VAF.

List of abbreviations

NGS: Next-Generation Sequencing

SNV: Single Nucleotide Variant

VAF: Variant Allele Frequency

cfDNA: circulating free DNA

ctDNA: circulating tumor DNA

PGM: Personal Genome Machine

CRPC: Castration Resistant Prostate Cancer

WGS: Whole Genome Sequencing

ddPCR: digital droplet PCR

TPR: True Positive Rate

PPV: Positive Predictive Value

FDR: False Discovery Rate

Declarations

Ethics approval and consent to participate

All samples were collected in studies with institutional regulatory board approval and conducted in accordance with the Declaration of Helsinki and the Good Clinical Practice guidelines of the International Conference of Harmonization (REC numbers: 04/Q0801/6 at the Royal Marsden, London, UK, 2192/2013 at the Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori, Meldola, Italy and 15/98 at Peter MacCallum Cancer Centre). Written informed consent was obtained from all patients.

Consent for publication

Patients have consented that next-generation sequencing data with no personal identifiers can be used for publication in an anonymized format.

Availability of data and material

AmpliSolve is freely available at https://github.com/dkleftogi/AmpliSolve. The datasets analysed during the current study is available from the authors on reasonable request.

Competing interests

The authors declare they have no competing interests.

Funding

DK, MP and SL are funded by the Wellcome Trust (105104/Z/14/Z). GA, DW, AJ, VC and DGT were supported by Prostate Cancer UK (PG12-49) and Cancer Research UK (A13239). VC was also funded by a European Society of Medical Oncology Translational Clinical Research Fellowship, AJ by an Irish Health Research Board Clinical Research Fellowship and a Medical Research Council Clinical Research Fellowship, DGT by a European Union Marie Curie Intra-European Postdoctoral Fellowship.

Authors' contributions

DK implemented and tested the AmpliSolve software. DK, MP and SL designed the computational study and analyzed the results. SS, SQW, DW, AJ, VC and DGT were responsible for sample collection; DW, AJ, VC and DGT carried out library preparation and DNA sequencing. DW, AJ, VC and DGT performed the ddPCR assays. SL and GA conceived the study; SL, GA and MP supervised it. DK, MP and SL wrote the paper, DW and GA reviewed it. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the participating men and their families who suffered from metastatic prostate cancer and nonetheless gave the gift of participation so that others might benefit.

References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.

2. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell. 2017;168:613–28.

3. Heitzer E, Perakis S, Geigl JB, Speicher MR. The potential of liquid biopsies for the early detection of cancer. NPJ Precis Oncol. 2017;1:36.

4. Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat Rev Cancer. 2017;17:223–38.

5. Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol. 2016;34:547–55.

6. Mansukhani S, Barber LJ, Kleftogiannis D, Moorcraft SY, Davidson M, Woolston A, et al. Ultra-Sensitive Mutation Detection and Genome-Wide DNA Copy Number Reconstruction by Error-Corrected Circulating Tumor DNA Sequencing. Clin Chem. 2018.

7. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12:443–51.

8. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J. 2018;16:15–24.

9. Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, Moch H, et al. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nat Commun. 2012;3:811.

10. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13:341.

11. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput Biol. 2013;9:e1003031.

12. Kockan C, Hach F, Sarrafi I, Bell RH, McConeghy B, Beja K, et al. SiNVICT: ultra-sensitive detection of single nucleotide variants and indels in circulating tumour DNA. Bioinformatics. 2017;33:26–34.

13. Deshpande A, Lang W, McDowell T, Sivakumar S, Zhang J, Wang J, et al. Strategies for identification of somatic variants using the Ion Torrent deep targeted sequencing platform. BMC Bioinformatics. 2018;19:5.

14. Shin S, Lee H, Son H, Paik S, Kim S. AIRVF: a filtering toolbox for precise variant calling in Ion Torrent sequencing. Bioinformatics. 2018;34:1232–4.

15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

16. Romanel A, Lago S, Prandi D, Sboner A, Demichelis F. ASEQ: fast allele-specific studies from next-generation sequencing data. BMC Med Genomics. 2015;8:9.

17. Romanel A, Gasi Tandefelt D, Conteduca V, Jayaram A, Casiraghi N, Wetterskog D, et al. Plasma AR and abiraterone-resistant prostate cancer. Sci Transl Med. 2015;7:312re10.

18. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, WGS500 Consortium, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014;46:912–8.

19. Carreira S, Romanel A, Goodall J, Grist E, Ferraldeschi R, Miranda S, et al. Tumor clone dynamics in lethal prostate cancer. Sci Transl Med. 2014;6:254ra125.

20. Lawrence MG, Obinata D, Sandhu S, Selth LA, Wong SQ, Porter LH, et al. Patient-derived Models of Abiraterone- and Enzalutamide-resistant Prostate Cancer Reveal Sensitivity to Ribosome-directed Therapy. Eur Urol. 2018;74:562–72.

21. Conteduca V, Wetterskog D, Sharabiani MTA, Grande E, Fernandez-Perez MP, Jayaram A, et al. Androgen receptor gene status in plasma DNA associates with worse outcome on enzalutamide or abiraterone for castration-resistant prostate cancer: a multi-institution correlative biomarker study. Ann Oncol. 2017;28:1508–16.

22. Barry P, Vatsiou A, Spiteri I, Nichol D, Cresswell GD, Acar A, et al. The spatio-temporal evolution of lymph node spread in early breast cancer. Clin Cancer Res. 2018;:clincanres.3374.2018.

23. Jiang H, Lei R, Ding S-W, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics. 2014;15:182.

24. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio]. 2013. http://arxiv.org/abs/1303.3997. Accessed 29 Oct 2018.

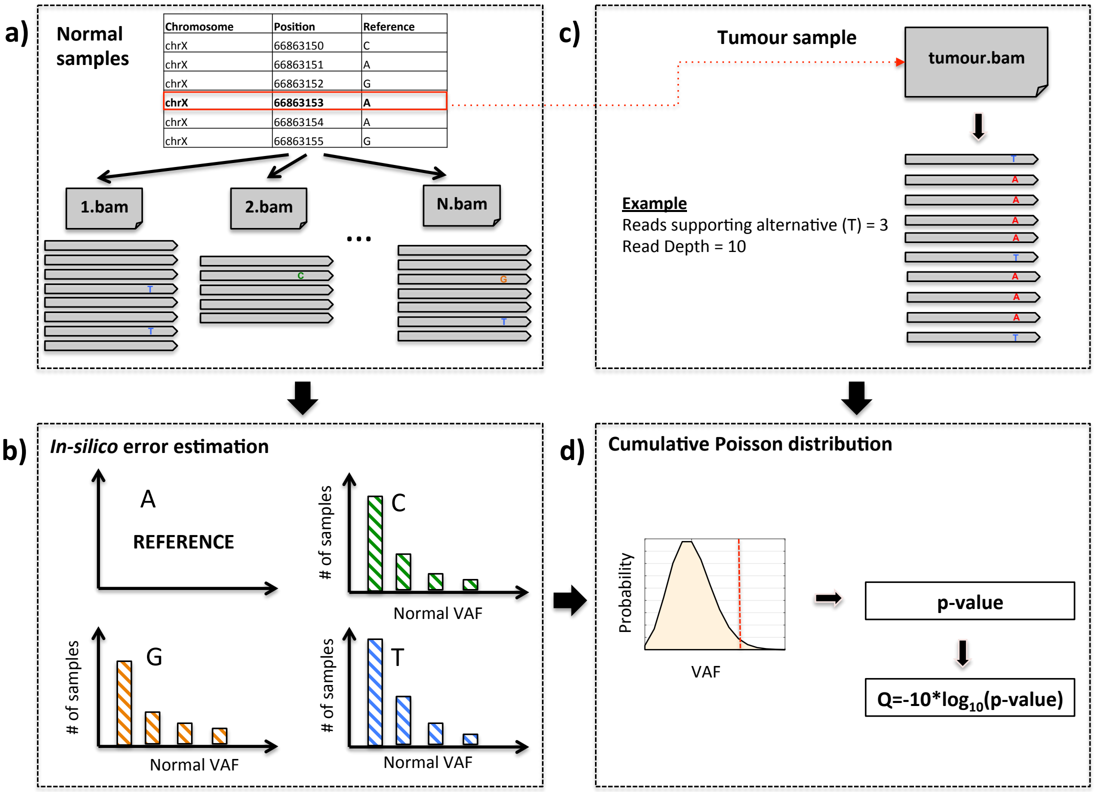
25. Picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. http://broadinstitute.github.io/picard/. Accessed 30 October 2018.

26. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31:213–9.

Figures

Figure 1

Graphical representation of AmpliSolve's workflow for estimating the noise levels and detecting SNVs. The workflow comprises the following steps: a) Screening the available normal samples to identify reads supporting alleles other than the reference. b) Error estimation per position, per nucleotide and per strand for all positions in the gene panel based on the distribution of alternative allele counts in (a); only alternative counts corresponding to VAF<5% are taken into consideration; c) For each genomic position in a tumour sample, the method identifies the total coverage of the position and the number of reads supporting the alternative alleles, if any. d) Given the information from steps b) and c) the method applies a Poisson distribution-based model to compute the p-value that the variant (red line) is real. This p-value is then transformed to a quality score that is used by AmpliSolve together with additional quality criteria to identify SNVs.



Chromosome	Position	Reference
chrX	66863150	C
chrX	66863151	A
chrX	66863152	G
chrX	**66863153**	**A**
chrX	66863154	A
chrX	66863155	G

Figure 2

Assessing AmpliSolve's performance using normal samples. a) Median AmpliSolve FDR (%) as a function of the model pseudo-count parameter, when using different numbers M of normal samples as training set and testing on the remaining normal samples. We consider as TP all normal variants with VAF≥20% and as FP all normal variants with VAF<20% (see Text). We consider all AmpliSolve calls that have Q-score≥20. b) Same as (a) when considering only AmpliSolve calls with a 'PASS' quality flag (see Text).

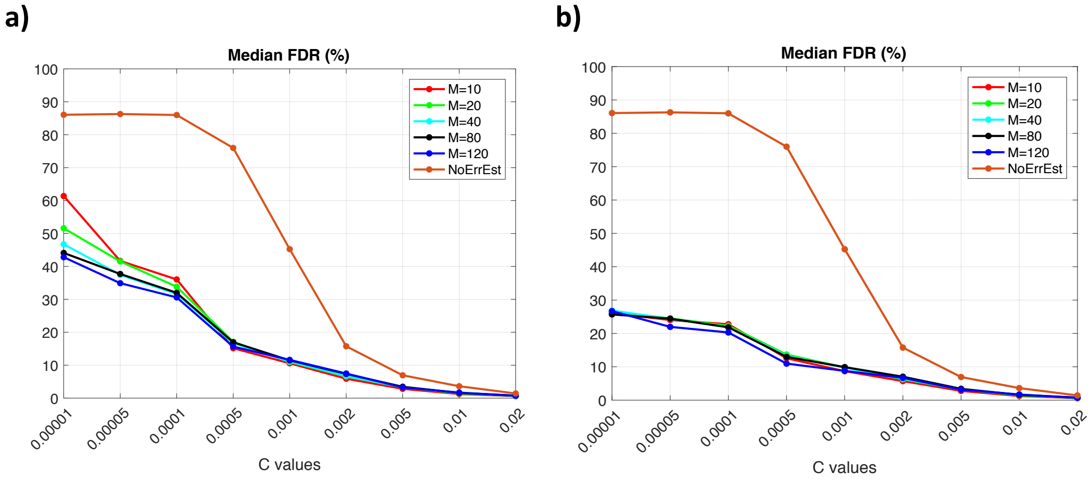


Figure 3

Assessing AmpliSolve's sensitivity using synthetic data. (a-e) AmpliSolve TPR (Sensitivity) values in *in-silico* synthetic variant experiments. We test different combinations of VAF, depth of coverage and C parameter values (see Text).

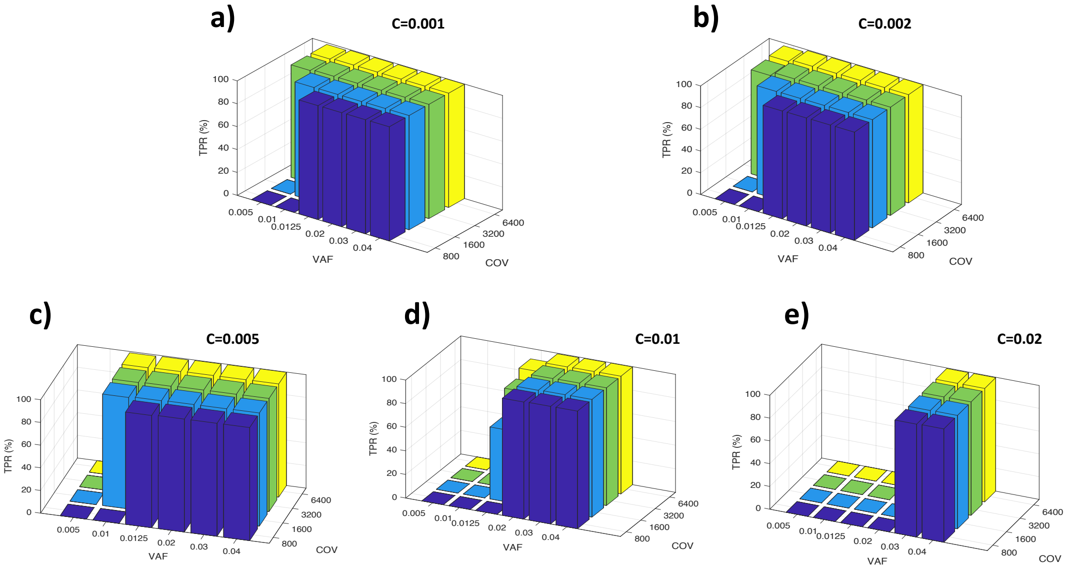


Figure 4

Comparison between Ion Ampliseq AmpliSolve calls and Illumina WGS MutPlat calls a) Venn diagram of mutations on 10 samples sequenced with both Ion Torrent and Illumina platforms and called respectively by AmpliSolve and by MutPlat. Low coverage positions denote mutations excluded by AmpliSolve because poorly covered (<100 reads on at least one strand, 'uncallable' by AmpliSolve). (b) Scatter plot of VAFs in WGS and AmpliSeq data. Note that all the SNVs not called by AmpliSolve (green point) have some support in the data and are reported in its output (hence they have AF > 0) but are filtered out, mostly because of strand bias. (c) Same as (b) but for VAFs<20%. Note that some concordant calls (purple points) have WGS AF=0; these are real germline variants with no support in the tumour (Methods). For the sake of this comparison, both in (b) and in (c) we don't consider the 49 mutations at positions of low coverage in Ion Ampliseq data (see (a)) ('uncallable' for AmpliSolve).

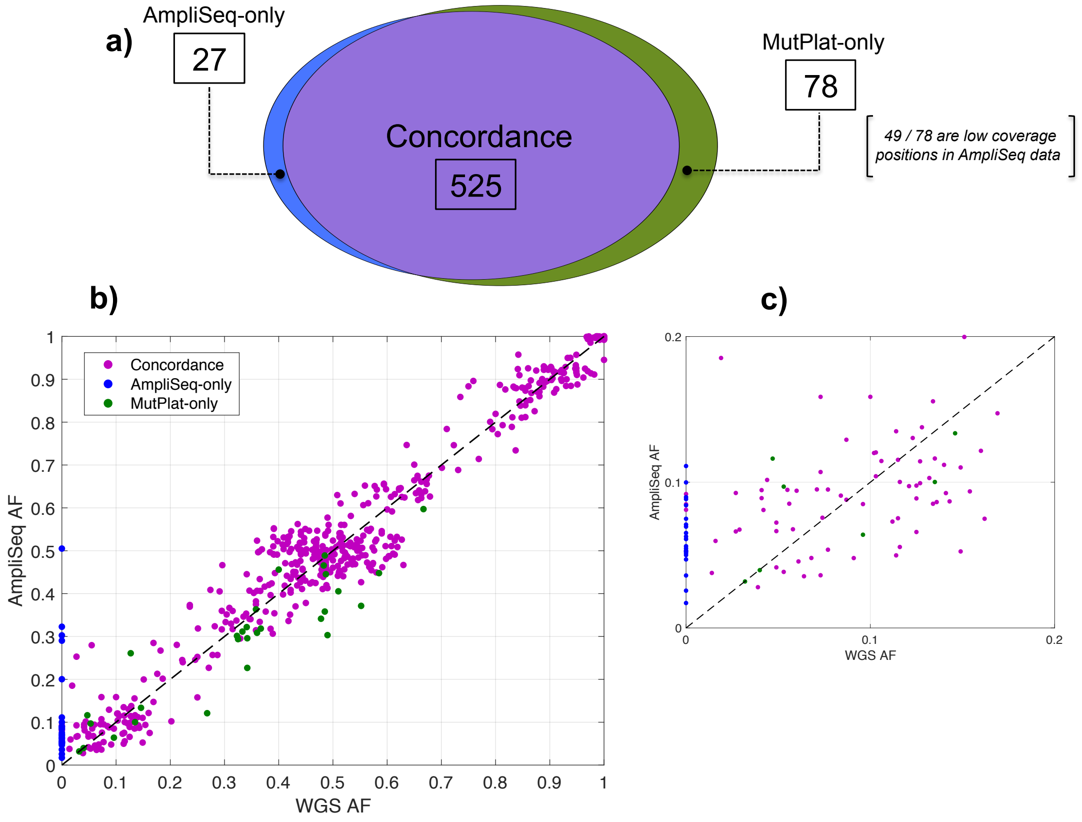
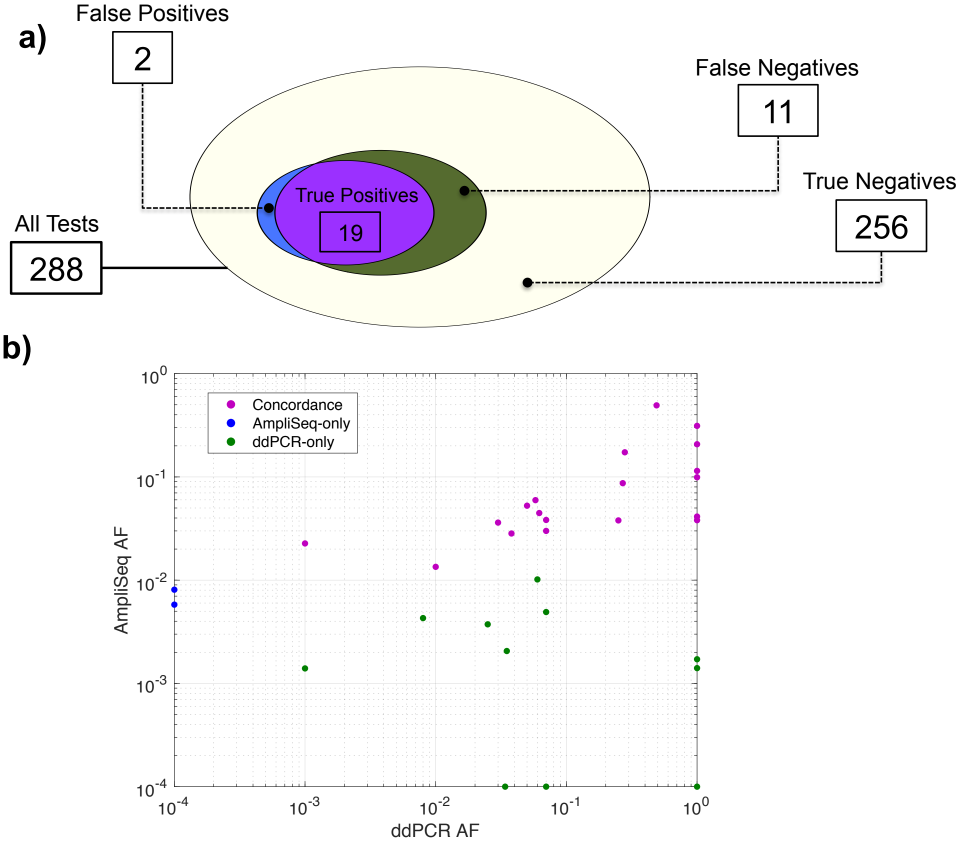


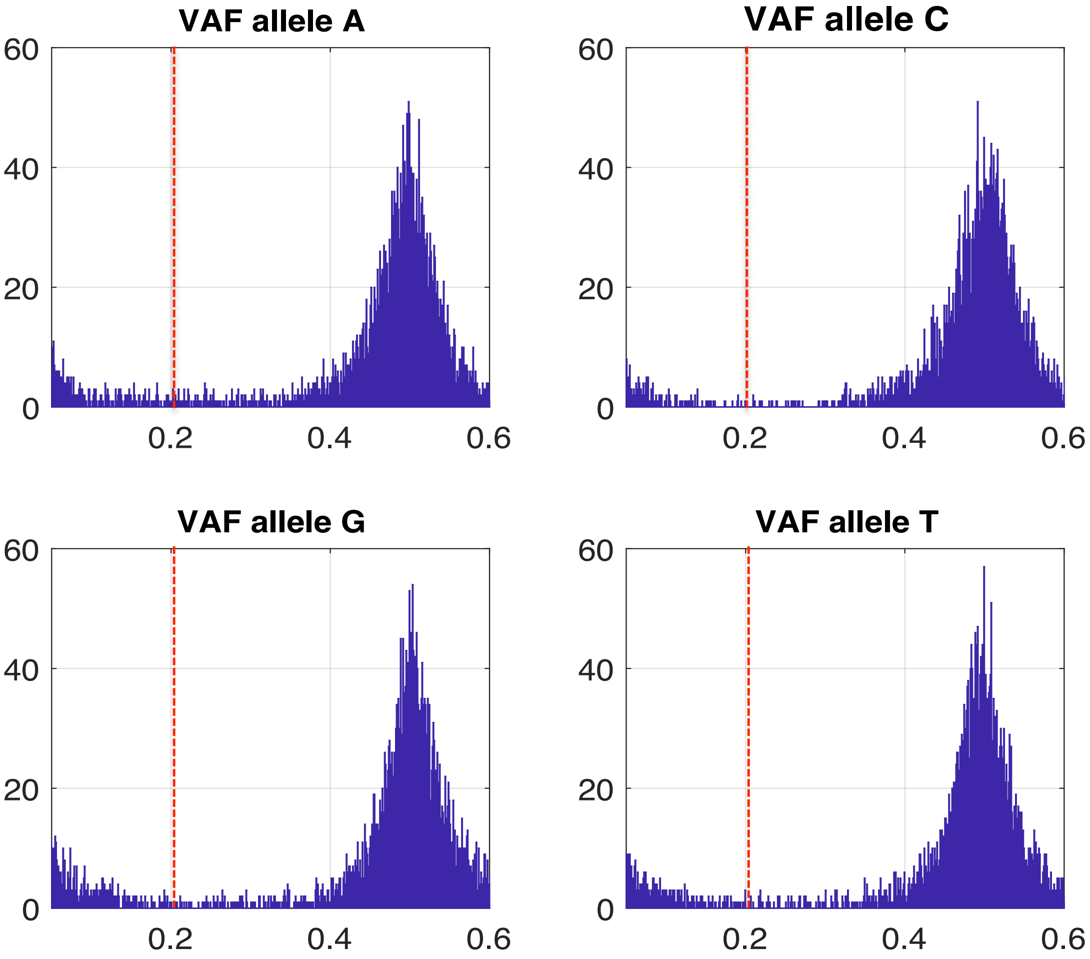
Figure 5

Validating AmpliSolve performance with ddPCR experiments. a) Venn diagram of mutations in 96 samples at 3 positions as determined by AmpliSolve and ddPCR experiments. False positives refer to variants called by AmpliSolve and not detected by ddPCR, false negatives the opposite. In 256 out 288 cases neither AmpliSolve nor ddPCR detect a mutation. (b) Scatter plot of the VAFs in the ddPCR and Ion Torrent data. Most of the SNVs missed by AmpliSolve (green points) have some support in the NGS data but they cannot be distinguished from noise. Because of the log scale, we arbitrarily set AF=$10^{-4}$for negative calls with AF=0. Similarly, we set AF=1 for ddPCR calls for which no allele frequency information is available.

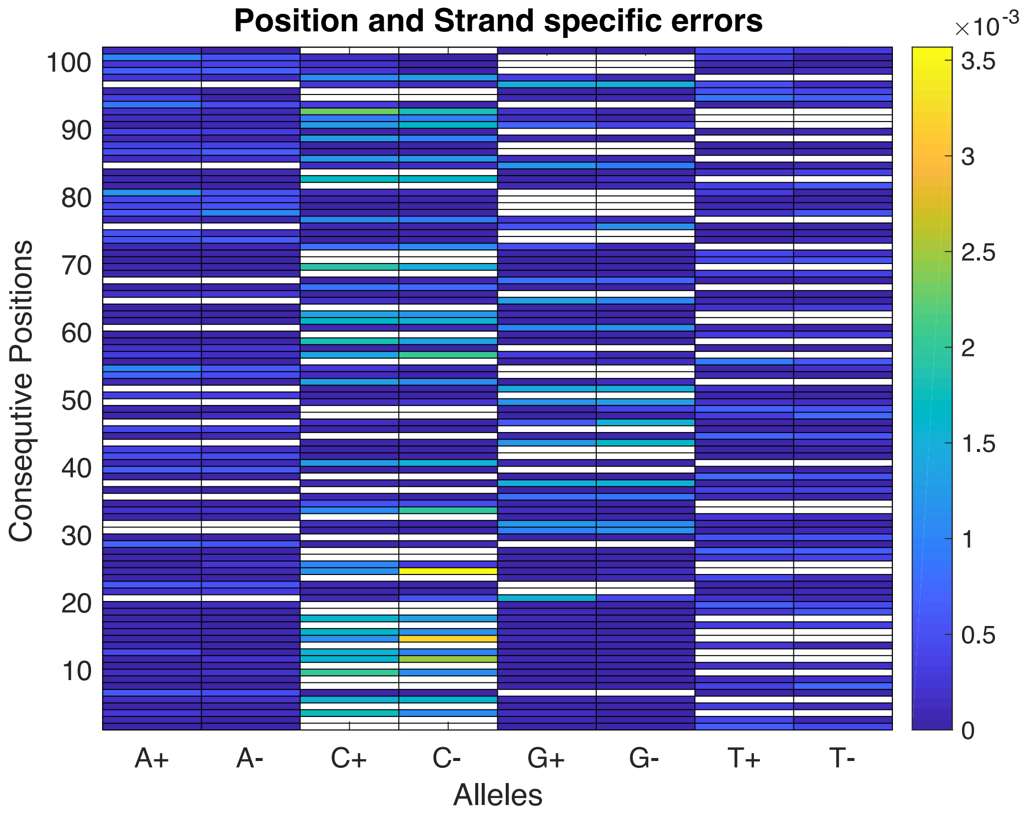


Additional Files

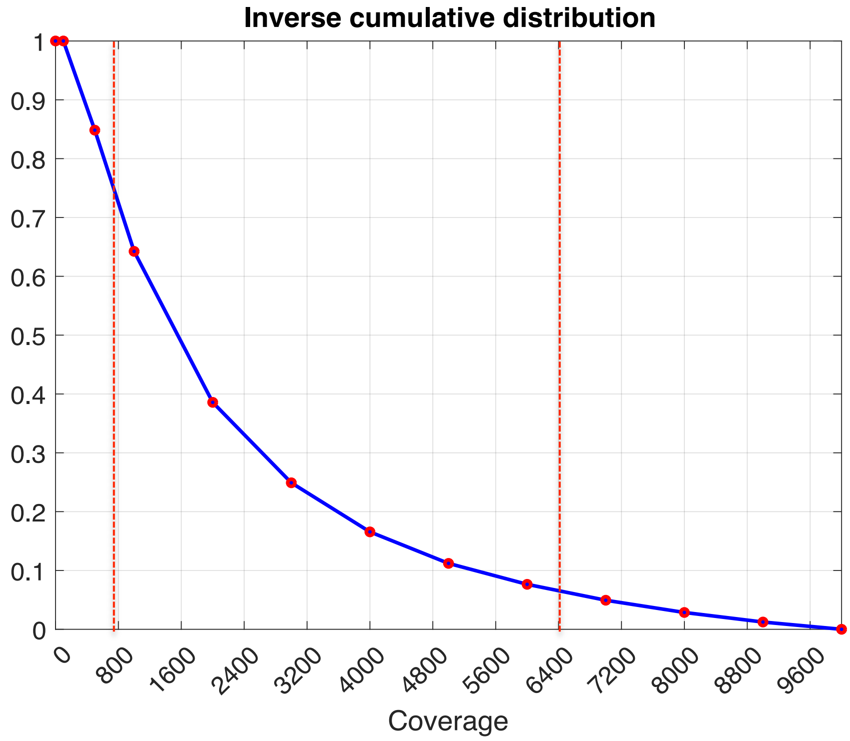
Additional file 1: Figure S1. Variant allele frequency (VAF) distributions for the A, T, C, G nucleotides as calculated from 30 randomly chosen normal samples across our custom AmpliSeq panel. Only VAFs < 60% are displayed. The red lines marks VAF = 20%.



Additional file 2: Figure S2. Position-specific, allele-specific and strand-specific frequency of alternative alleles in 100 consecutive positions in the AR gene.



Additional file 3: Figure S3. Fraction of sites in our custom AmpliSeq panel sequenced at a given coverage or more. The values are calculated over 30 randomly selected ctDNA samples. Note that positions with depth of coverage less than 200 are not considered for calculating the total number of positions. The red lines represent the upper and lower bounds of coverage used in the synthetic variant test.


Inverse cumulative distribution
Coverage