Bioinformatics Advance Access published August 16, 2016

Genome analysis

SiNVICT: Ultra-Sensitive Detection of Single Nucleotide Variants and Indels in Circulating Tumour DNA

Can Kockan [1,2,†], Faraz Hach [1,4†], Iman Sarrafi [1,†], Robert H. Bell [4], Brian McConeghy [4], Kevin Beja [4], Anne Haegert [4], Alexander W. Wyatt [4,5], Stanislav V. Volik [4], Kim N. Chi [4], Colin C. Collins [4,5*], S. Cenk Sahinalp [1,3,4,*]

[1]School of Computing Science, Simon Fraser University, Burnaby, BC, Canada, V5A 1S6
[2]MADD-Gen Graduate Program, Simon Fraser University, Burnaby (BC), Canada, V5A 1S6
[3]School of Informatics and Computing, Indiana University, Bloomington, IN, USA, 47405
[4]Vancouver Prostate Centre, Vancouver, BC, Canada, V6H 3Z6
[5]Dept. of Urologic Sciences, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

[*]To whom correspondence should be addressed.

Associate Editor: Dr. John Hancock

Abstract

Motivation: Successful development and application of precision oncology approaches require robust elucidation of the genomic landscape of a patient's cancer and, ideally, the ability to monitor therapy-induced genomic changes in the tumour in an inexpensive and minimally invasive manner. Thanks to recent advances in sequencing technologies, "liquid biopsy", the sampling of patient's bodily fluids such as blood and urine, is considered as one of the most promising approaches to achieve this goal. In many cancer patients, and especially those with advanced metastatic disease, deep sequencing of circulating cell-free DNA (cfDNA) obtained from patient's blood yields a mixture of reads originating from the normal DNA and from multiple tumour subclones - called circulating tumour DNA or ctDNA. The ctDNA/cfDNA ratio as well as the proportion of ctDNA originating from specific tumour subclones depend on multiple factors, making comprehensive detection of mutations difficult, especially at early stages of cancer. Furthermore, sensitive and accurate detection of SNVs and indels from cfDNA is constrained by several factors such as the sequencing errors and PCR artifacts, and mapping errors related to repeat regions within the genome. In this paper, we introduce SiNVICT, a computational method that increases the sensitivity and specificity of SNV and indel detection at very low variant allele frequencies. SiNVICT has the capability to handle multiple sequencing platforms with different error properties; it minimises false positives resulting from mapping errors and other technology specific artifacts including strand bias and low base quality at read ends. SiNVICT also has the capability to perform time-series analysis, where samples from a patient sequenced at multiple time points are jointly examined to report locations of interest where there is a possibility that certain clones were wiped out by some treatment while some subclones gained selective advantage.

Results: We tested SiNVICT on simulated data as well as prostate cancer cell lines and cfDNA obtained from castration-resistant prostate cancer patients. On both simulated and biological data, SiNVICT was able to detect SNVs and indels with variant allele percentages as low as 0.5%. The lowest amounts of total DNA used for the biological data where SNVs and indels could be detected with very high sensitivity were 2.5ng on the Ion Torrent platform and 10ng on Illumina. With increased sequencing and mapping accuracy, SiNVICT might be utilised in clinical settings, making it possible to track the progress of point mutations and indels that are associated with resistance to cancer therapies and provide patients personalised treatment. We also compared SiNVICT with other popular SNV callers such as MuTect, VarScan2, and Freebayes. Our results show that SiNVICT performs better than these tools in most cases and allows further data exploration such as time-series analysis on cfDNA sequencing data.

Availability: SiNVICT is available at: https://sfu-compbio.github.io/sinvict

Contact: cenk@sfu.ca

1 Introduction

One of the most promising areas of precision oncology is the development of custom targeted therapies tailored for a patient. Successful development and efficient application of such therapies require efficient and inexpensive identification and monitoring of therapy-induced changes in a patient's tumour DNA. Unfortunately, especially in advanced stage cancers, the main cause of cancer's morbidity and mortality is the development of multiple metastatic lesions, often not easily accessible for tissue sampling. For example, in prostate cancer more than 90% of metastases occur in bone and/or deep lymph nodes (Bubendorf *et al.*, 2000). Biopsying such sites is associated with significant morbidity for the patients and thus is not commonly performed.

The existence of circulating cell free DNA (cfDNA) in mammalian blood has been known since 1948 (Mandel, 1948). cfDNA is thought to be released from the dying (necrotic/apoptotic) cells - both normal and tumour, as has been shown in 1994 when mutated RAS gene fragments were detected in the blood of cancer patients - see (Schwarzenbach *et al.*, 2011). The non-specific mechanism of generating cfDNA results in integral representation of all tumour DNA of a patient subject to sampling variability and, possibly, to tumour's access to blood stream. In an earlier study, for example, we observed the presence of multiple mutated forms of AR (androgen receptor) gene in cfDNA of patients with castrate resistant prostate cancer (CRPC) (Azad *et al.*, 2015) that can be best explained by the presence of multiple subpopulations of cancer cells in each patient's body. This integral representation of multiple tumour foci/subclones provides an important advantage to the use of blood plasma as a source of tumour-derived DNA. Unfortunately, the presence of both normal and tumour DNA in a patient's blood poses significant challenges to the analysis of cfDNA sequence data. To make matters worse, tumour DNA is many times derived from multiple subclones and is thus highly heterogeneous. An earlier study we performed on mutations in CRPC patients (Azad *et al.*, 2015) demonstrated that cfDNA comprised an average of 4.7% (IQR[1] 4.5%) of ctDNA, based on the proportion of reads with mutations in AR.

There are several somatic and germline mutation callers that were developed to find single nucleotide variants (SNVs) as well as indels within a given population using WGSS (Whole Genome Shotgun Sequencing), as well as to detect specific variants in a patient's genome through sampling multiple loci from the same patient. Examples include GATK (McKenna *et al.*, 2010), VarScan2 (Koboldt *et al.*, 2012), Freebayes (Garrison and Marth, 2012), Strelka (Saunders *et al.*, 2012), MuTect (Cibulskis *et al.*, 2013), and others. Most of these tools either use a frequentist or Bayesian approach to estimate the probability of a locus being an actual mutation instead of being a false positive caused by noise (due to sequencing or mapping errors). Among them, VarScan2 uses several heuristics to reduce the size of the candidate set and then applies some statistical test like Fisher's Exact on tumour/benign pairs to call somatic mutations. It also provides post-processing capability to enable further filtering based on additional factors such as strand bias. Other tools such as Freebayes, MuTect, and Strelka make use of the prior and posterior probabilities of a location being mutated in a Bayesian context in order to call mutations. Unfortunately, these tools are not designed to work with (i) sequencing data from patients at multiple time points, (ii) very high read depth (e.g. 20k-30k average, up to 90k and possibly more in the future), or, (iii) extremely low dilutions (can be as low as around 0.01% variant allele percentage (Lipson *et al.*, 2014)), or, (iv) samples with high intra-tumour heterogeneity, or, (v) batches of samples that suffer from systematic noise. In addition, ctDNA levels can also be below the analytical sensitivity of existing ctDNA detection approaches in patients with localized disease and in patients that have received therapy (Bettegowda *et al.*, 2014).

In order to address problems mentioned above, we introduce SiNVICT a computational tool that can handle very high read depth and very low dilutions. SiNVICT can be run on a single tumour sample, on a batch of multiple tumour samples, or on multiple samples from a single patient sequenced at multiple time points. This feature allows SiNVICT to process samples from a single patient in multiple cancer stages, as well as a group of different patients that are being sequenced and analyzed at the same time. In cases where these samples have similar disease progression and dilution levels, SiNVICT can make use of the Signal-to-Noise ratio of the batch (explained later) to characterise the systematic noise and try to reduce the number of false positives due to the non-uniformity of noise across the sequenced regions.

We evaluated robustness of SiNVICT on data obtained by two sequencing platforms with distinct error rates (0.1% substitution in Illumina; 1% indel in IonTorrent (Glenn, 2011)), which were applied to the same tumour samples. Our experiments indicate that SiNVICT is highly sensitive to calls on data generated by both sequencing platforms. For example, three previously validated AR (Androgen Receptor gene) mutations (Azad *et al.*, 2015) in a mixture of 22RV1 and 49C cell-lines - which were used as reference in AmpliSeq[2] calibration and Illumina calibration experiments - were detected with almost identical sensitivity by SiNVICT. SiNVICT was also able to detect previously validated mutations successfully from actual cfDNA sequencing data obtained from castrate resistant prostate cancer (CRPC) patients. These findings suggest that SiNVICT might be utilised in the analysis of deep sequencing cancer data obtained from both Ion Torrent and Illumina sequencing technologies.

As importantly, SiNVICT addresses a unique problem and is not comparable to existing popular SNV and indel callers (e.g. GATK) particularly because such tools typically process a fraction of the reads in datasets with high sequencing depth since multiple occurrences of identical reads are marked as PCR duplicates. However, identical reads are to be expected in deep amplicon sequencing and this is not necessarily an artifact of PCR.

2 Methods

As shown in Figure 1, SiNVICT works in three steps: (i) pre-processing of raw input, (ii) SNV and indel discovery, (iii) post-processing and reporting the final calls. Details are given in the relevant subsections below.

2.1 Preprocessing Steps

SiNVICT pre-processes and prepares the raw input file from sequencers for actual detection of SNVs and indels through the following substeps.

Trimming. SiNVICT trims the input reads in order to remove any remaining primers or very low quality bases at the ends of reads. If the reads have already been properly trimmed and quality checked, SiNVICT can skip this substep.

[1] IQR: the interquartile range is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively and IQR = Q3 - Q2 (Speedy Publishing LLC, 2014).

[2] Ion AmpliSeq Targeted Sequencing Technology is a technology offered by Ion Torrent platform for creating custom targeted ultra-deep sequencing libraries

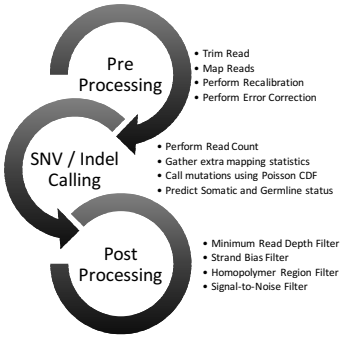

Fig. 1. Overview of SiNVICT data processing pipeline.

Mapping. Once the reads are trimmed, read mapping can be performed using any short read aligner that allows mapping with indels, i.e. mrFAST-fastHASH (Xin *et al.*, 2013), BWA (Li and Durbin, 2009), etc. SiNVICT sorts the mapping output with respect to genomic loci.

Recalibration and error correction of low quality mappings. SiNVICT re-calibrates the "base quality" of low-quality/ambiguous mappings with the goal of improving the mapping accuracy and reducing the noise (errors) introduced by these mappings for downstream analysis. After re-calibration, SiNVICT performs local assembly to do error correction on the bases. The newly corrected reads are then re-aligned to the reference genome per the previous substep. The final product of this substep is thus a set of re-calibrated and error-corrected high quality mappings that will be used for the main part of our method.

We use the tool ABRA (Mose *et al.*, 2014) for recalibration and error correction of initial mappings. For performing calculations on variant allele frequencies we use the bam-readcount tool which provides an interface to samtools pileup. This provides detailed statistics for every single location within the target regions, including the reference base, read depth, variant alleles, base counts for each allele, reads mapped to the forward and reverse strands, average base qualities, average mapping qualities, the distance of the location from the ends of the read as a fraction, etc.

2.2 Somatic Mutation and Indel Discovery Step

The main goals of SiNVICT are to identify mutations (somatic or germline) from read errors and to distinguish potential somatic mutations originating from tumour genomes from allelic variation (due to germline events) in the normal genome. SiNVICT achieves this by calculating for each potential mutation (or indel) locus, the probability of the mutation being real (as a function of the error rate observed for the sequencing platform), as well as that for the observed allelic distribution being a result of a somatic mutation vs a germline mutation, through the use of a Poisson model. In other words, SiNVICT returns the p-value (p) and confidence score ($Q = 10 \cdot log_{10}\, p$) for each potential mutation as well as those for each mutation being somatic.

In order to calculate the p-values, SiNVICT processes the readcount data to obtain the initial set of calls. For that SiNVICT uses (i) N: the total number of reads covering a position, (ii) K: the number of reads that support a mutation for that position, and (iii) r: the average error rate (for each position, determined by the sequencing platform).

Based on this, SiNVICT calculates p_1, the p-value of the mutation as follows (Illumina, 2013).

$$p_1 = P(K|\lambda_1) = e^{-\lambda_1} \sum_{i=0}^{\lfloor K \rfloor} \frac{\lambda_1^i}{i!} \tag{1}$$

The above Poisson cumulative distribution function (CDF) gives the probability that there is an actual mutation at a particular position, if out of N reads covering that position, K reads support a variant allele. Note that given the average error rate r, $\lambda_1 = N \times r$. SiNVICT allows the user to set a threshold for the p-value p_1 implicitly via the confidence score conversion ($Q = 10 \cdot log_{10}\, p_1$). SiNVICT will not report calls with confidence score below the user defined threshold.

Once it has been established that there is an actual mutation at a particular locus, we can again use the Poisson model to calculate p_2, the p-value of the mutation being somatic by setting $\lambda_2 = N/2$.

$$p_2 = P(K|\lambda_2) = e^{-\lambda_2} \sum_{i=0}^{\lfloor K \rfloor} \frac{\lambda_2^i}{i!} \tag{2}$$

In this case, λ_2 is the average number of events per interval in a Poisson distribution and N is the total number of reads covering a location. The null hypothesis here is that the observed mutation is germline. In this case, around half (i.e. $N/2$) of the reads covering this locus are expected to include the mutation and thus λ_2 is set to $N/2$.

This Poisson model has high sensitivity (on both Illumina and Ion Torrent Proton platforms) and can introduce many false positives due to the following. Both Illumina and Proton native mutation callers are designed to run on the mapping data from a single tumour sample, without any consideration for strand bias or the read depth. In addition neither of these callers take into account systematic noise characteristics during the processing of multiple samples. These result in an inflation in the number of mutation candidates, making further downstream analysis virtually intractable. In order to reduce the number of the candidates, we thus apply a number of post-processing steps as described below.

2.3 Postprocessing Steps

SiNVICT applies a number of postprocessing filters to the candidate locations to increase its specificity as follows.

Minimum Read Depth filter. SiNVICT has a filter to discard locations that do not meet the minimum read depth. While SiNVICT is intended to be used with ultra-deep sequencing data, some locations will still have very low coverage due to the limitations of the sequencing technologies. Thus, the read depth is very often non-uniform across the locations. In "low coverage" ($<$ minimum read depth) regions, the sequencing errors can be misinterpreted as SNVs or indels and thus are filtered out.

Strand Bias filter. The strand bias for a genomic location i (see equation 3) is defined as the ratio of the number of reads that are mapped to the forward strand to the total number of reads mapped for that genomic location.

$$\text{StrandBias}_i = \frac{\text{NumReadsForward}_i}{\text{NumReadsTotal}_i} \tag{3}$$

If the potential strand bias is outside of the range $[0.5 - \epsilon, 0.5 + \epsilon]$, then we say that there is a real strand bias in the associated genomic region. Strand bias could lead to both false positives and false negatives. However, most of the regions generated by Illumina sequencing technology have strand bias primarily causing false positives (Guo *et al.*, 2012). In contrast, Ion Torrent technology is known to return only a few regions with real strand bias and for that we only filter regions with extremely high strand bias value. It should be noted that while the strand-bias filtering can usually be more conservative for AmpliSeq (Ion Torrent) technology, due to the level of noise in our calibration experiment, SiNVICT filters a larger number of locations than normally expected.

There is no definitive cut-off for the strand bias values in general but we have obtained good results for $\epsilon = 0.1$. The SNV/indel calls for which there is a real strand bias are declared to be of lower confidence and are filtered out.

Homopolymer Regions. Calling SNVs and indels in homopolymer regions are very challenging because mapping the reads correctly to these regions are very difficult. This source of bias can cause many false positive calls. To eliminate these false positives, for each location that was called (as an SNV or indel) earlier, we check the consecutive 3 bases on both sides of this location and declare it as a lower confidence call (to be filtered out) if either side contains 3 identical bases.

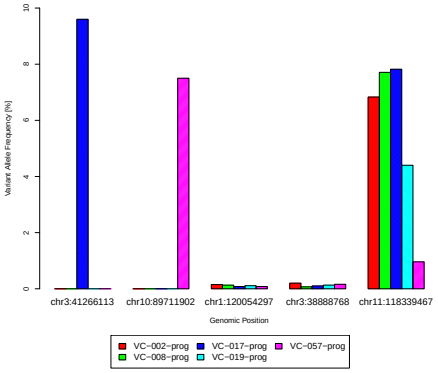

Fig. 2. Calculating the Signal to Noise Ratio (SNR) in multiple samples across several genomic loci. The x-axis depicts 5 loci from a set of samples such that each sample is represented by a distinct colour. For each location on the x-axis, the corresponding value on the y-axis indicates the percentage of the most frequent variant allele. The chance of any one of the loci indicated above having a "uniform" variant allele frequency distribution across 5 unrelated samples is very low and thus there must be location dependent noise in the variant allele frequency estimates. Most current SNV callers will report a potential mutation for one of the samples in locations 1 and 2 while they will report a potential mutation for 4 out of 5 samples in location 5. They will not make any calls for locations 3 and 4 because of their negligible measured variant allele frequencies. For the last position, all samples but one show substantial evidence for a potential mutation when examined individually. The SNR filter utilized by SiNVICT allows such cases to be filtered out under the assumption that SNVs are expected to be unique to a few patients among a batch for all practical purposes.

Signal to Noise Ratio (SNR) in multiple samples. Different regions of the genome can have different noise levels because of the sequencing technology (see Figure 2). Therefore having the average noise information for a particular genomic locus across a number of samples can be very useful in assessing the likelihood of a false positive. As mentioned earlier, SiNVICT is capable of performing analysis on a cohort of samples. In such cases, SiNVICT calculates and stores the average noise level for each location across the samples. Consequently, these average noise values are used to distinguish noisy locations from actual variants, and eventually detect the final set of SNVs and indels more accurately.

For calling a location an SNV (or indel) in noisy regions, we calculate the Signal to Noise Ratio (SNR) as the ratio of mean and standard deviation of major variant allele frequency across the samples within the panel, as per equation 4. The mean μ_i can be calculated as per equation 5 and the standard deviation, σ_i, is given in equation 6; in both equations the sum is taken across n samples in the panel, where i is the current genomic location and j is the current sample.

$$\text{SNR}_i = \frac{\mu_i}{\sigma_i} \tag{4}$$

$$\mu_i = \frac{\sum_{j=1}^{n} \text{VariantAllelePercentage}_j^i}{n} \tag{5}$$

$$\sigma_i^2 = \frac{\sum_{j=1}^{n} (\text{VariantAllelePercentage}_j^i - \mu_i)^2}{n} \tag{6}$$

Each locus with major variant allele percentage $\geq 3\times$ SNR is then declared as a high confidence variant. The remaining loci are filtered out.

SiNVICT applies each one of the 4 filters in the order they are described above, namely, (1) Read Depth, (2) Strand Bias, (3) Homopolymer and (4) SNR. Each genomic locus, "passing" the first k filters and "failing" the $k+1$st is added to the file associated with the filter it fails. Only the genomic loci that pass all filters are considered to be high confidence SNVs (or indels) and are added to the master file.

2.4 Time Series Analysis

SiNVICT also provides the ability to perform time series analysis on cfDNA sequencing data obtained from cancer patients on multiple time points throughout their treatment. The goal here is to provide the user the ability to assess whether specific mutations appear only in specific time points or are present in all time points, sometimes with very low prevalence, e.g. in support of the Big Bang theory of cancer (Sottoriva *et al.*, 2015). SiNVICT achieves this in two steps. (i) Genomic loci that were sequenced successfully in all time points for a patient and assigned a sufficiently high confidence score (e.g. $>= 90$) in at least one of the time points by the Poisson model used by SiNVICT (Equation 1) are chosen; (ii) Among these, only the loci with high read depth (e.g. $>= 1000$) in all samples are used so that only the highest confidence regions are considered for the time series analysis. On each of these loci, rather than relying on the Illumina error rate estimate, SiNVICT calculates the *localized error rate* by using the mean VAF (Variant Allele Frequency) from the 10 neighboring bases (5 on each side). The user can increase the size of the neighborhood (from 10 bases to any user specified number of bases) used to calculate the localized error rate, at the cost of a higher running time. By calculating p-value of the detected variant through the use of a localized error rate (rather than the error rate provided by specs) SiNVICT reduces the position specific and sequence content based biases in sequencing errors (Nakamura *et al.*, 2011). Based on the localized error rate, SiNVICT then recalculates the p-value as $1-(1-err_n)^{1/perc_m}$, where err_n is our error rate estimate and $perc_m$ is the percentage of reads that include the mutation.

3 Results

In order to evaluate our method, we performed the following experiments: (i) we simulated insilico ctDNA/cfDNA with varying dilutions to determine SiNVICT's performance (precision/recall) in SNV detection, (ii) we mixed 22RV1 and 49C prostate cancer cell-lines and sequenced them with Ion Torrent and Illumina technologies to emulate various tumour-normal mixture levels to measure SiNVICT's SNV as well as indel detection performance on a mixture of sequencing data, and finally (iii) we explored the time-series analysis capabilities of SiNVICT on cell-free DNA sequencing data from castration-resistant prostate cancer

patients (Wyatt *et al.*, 2016) [3]. We compared our method to widely used SNV callers: MuTect, VarScan2, and Freebayes. In all the experiments, SiNVICT outperforms Freebayes. Furthermore, our results show that SiNVICT performs better than MuTect and VarScan2 in most cases for ultra-deep sequencing data and allows further data exploration such as time-series analysis on cfDNA sequencing data.

3.1 Simulated Data

3.1.1 SNV calling on simulated data.
We tested all four tools on simulated data obtained from version hg19 - i.e. GRCh37- (Meyer *et al.*, 2012) of the human reference genome. The parameters that are used to run each tool are provided in Supplementary Table 1. We extracted the exons of the AR gene with BEDTools (Quinlan and Hall, 2010), representing the normal tissue, and introduced 18 random SNVs to a copy of the original sequence as the "tumour" tissue. We then used wgsim (part of Samtools (Li *et al.*, 2009)) to simulate ultra-deep sequencing with Illumina MiSeq. We tried to keep the parameters close to the experimentally observed ones (read length = 145, insert size = 175). We obtained an average read depth of $\approx$ 20000 in 7 different tumour-normal mixture levels (50%, 20%, 10%, 5%, 2.5%, 1%, and 0.5% tumour content level).

SiNVICT was highly sensitive on this simulated data set: it was able to detect *all* 18 mutations (as high confidence SNVs) at tumour content levels of 50%, 20%, 10%, 5%, and 2.5%. At tumour content level of 1%, SiNVICT was able to detect 13 of the 18 mutations; at tumour content level of 0.5% it detected 12 of the 18 mutations. With respect to specificity, out of 8938 locations in the corresponding exons, SiNVICT called between 15 and 21 (with an average of 20) locations as high confidence SNVs, resulting in exactly 3 false positives per sample. In all the cases, SiNVICT had a higher precision than MuTect, Freebayes, and VarScan2. In all except one case (tumour content 1%), SiNVICT had a better (or equal) recall than other tools. See Figure 3 and Supplementary Table 2 for details.

3.1.2 Indel calling on simulated data.
We also carried out an experiment to check the precision and recall of all tools for indel calling on simulated data. From the same reference genome used in the previous experiment, we extracted exons 2-5 of the PIK3CA gene and manually added 4 indels (of size 2 each) and generated five samples with different tumour-normal mixtures (50%,20%,10%,5%,1%) with average read depth of $\approx$ 14000, insert size of 150, and read length of 70. SiNVICT and VarScan2 had perfect precision and recall on all of the samples. MuTect only missed two indels at 1% level. However, Freebayes only reported indels on one sample and failed to report anything on the others. See Supplementary Table 3 for details.

3.1.3 SNV calling on simulated data with intra-tumour heterogeneity.
We evaluated all methods on a more challenging dataset by building a sample tumour phylogeny to simulate the effect of intra-tumour heterogeneity. We increased the number of point mutations to 25 and distributed them among 5 clones. Each clone was assigned 5 distinct SNVs out of the 25; similar to (Malikic *et al.*, 2015) we assume that our mutations follow infinite-sites model. This implies that each clone inherits all mutations present in its parent clone. See Figure 5 for the topology of the phylogenetic tree used in this experiment.

We prepared 10 samples, each containing a mixture of normal cells and the above mentioned clones with normal contamination rates of 90%, to 99% - with unit increments.

[3] http://www.ebi.ac.uk/ena/data/view/PRJEB11648, http://www.ebi.ac.uk/ena/data/view/PRJEB11658

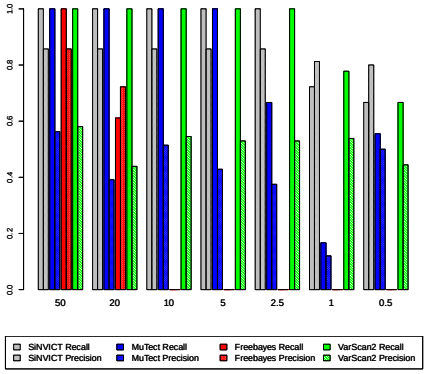

Fig. 3. Precision and Recall of SiNVICT, MuTect, Freebayes, and VarScan2 on simulated data. x-axis represents different samples with different tumour content levels. In each of these simulated samples, there were 18 manually added SNVs. The total number of bases covered per sample was 8938. All of the 18 SNVs were successfully detected in samples at tumour content levels of 50%, 20%, 10%, and 5% by SiNVICT, MuTect, and VarScan2. SiNVICT and VarScan2 also had perfect recall at 2,5% tumour content. In all the cases, SiNVICT had better precision than MuTect, Freebayes, and VarScan2. However in tumour content of 1% VarScan2 had better recall (1 more SNV detected by VarScan2). SiNVICT made relatively less number of false positive calls resulting in its higher precision. Details about the number of calls are provided in Supplementary Table 2.

For this experiment, we selected genomic regions from 5 distinct chromosomes at approximately equal sizes whose total length was 31485 base pairs. We extracted these regions with BEDTools and used wgsim to simulate ultra-deep sequencing with Illumina MiSeq. We kept the parameters close to the experimentally observed ones (read length = 145, insert size = 175). We obtained an average read depth of $\approx$ 20000 in all samples. We observed that the detection abilities of SiNVICT for such a heterogeneous case was adequate for higher prevalence clones up to 97% normal mixed with 3% tumour. Beyond this level, the variant allele percentage for all clones fell below the 0.6% level, which resulted in decreased sensitivity. Note that Freebayes, MuTect, and VarScan2 did not make any calls on this challenging dataset. We believe that this is caused due to the rejection of the SNV sites by the triallelic site filter commonly used by such tools. Because of the extremely high read depth we simulated (giving rise to many reads with errors), as well as the highly clonal structure of the samples, most of these SNV locations show evidence towards more than one possible mutation. SiNVICT only considers the most frequent non-reference base change and ignores the other lower frequency base changes - a likely reason for SiNVICT to make correct calls in contrast to the other tools. For instance, at a sample location with a read depth of 80,000 and a reference base A, one may observe 78,000 reads matching the reference, 1500 reads suggesting an A to T change, and 500 reads suggesting an A to C change. Our results are shown in Figure 4 and Supplementary Table 4 in more detail.

3.2 AmpliSeq and Illumina 22RV1-49C Calibration data

In the above experiment, we assumed that the amount of DNA available for our use is unlimited. Since cfDNA is usually obtained from blood, the amount of DNA available for analysis in reality can be very low, which will introduce sequencing challenges. In addition to the low amount of

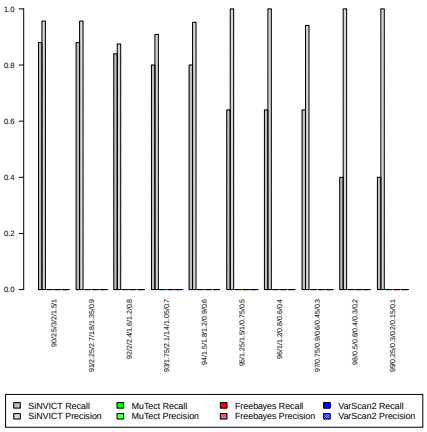

Fig. 4. Precision and Recall of SiNVICT, MuTect, Freebayes, and VarScan2 on simulated data consisting of 5 clones. To each clone, we added 5 SNVs and each subclone inherited additional mutations from their parents as shown in Figure 5 and the number of bases covered was 31485. Detection abilities of SiNVICT for such a heterogeneous case was observed to be adequate for higher prevalence clones up to 97% normal mixed with 3% tumour. Beyond this level, the variant allele percentage for all clones fell below 0.6%, which resulted in a reduction of sensitivity. Freebayes, MuTect, and VarScan2 failed to provide any calls for these simulated data sets. We believe this is caused due to the rejection of the SNV sites by the triallelic site filter commonly used by such tools. Because of the extremely high read depth we simulated, as well as the highly clonal structure of the samples, most of these SNV locations show evidence towards more than one possible mutation. Detailed information about call statistics are provided in Supplementary Table 4.

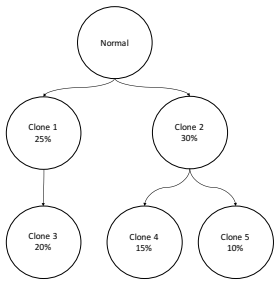

Fig. 5. Sample phylogenetic tree with 5 clones, randomly selected topology and prevalences to simulate the conditions of SNV detection for a very heterogeneous tumour.

DNA, low tumour content can further complicate the analysis of cfDNA data.

In this second experiment, we simulated such real life scenarios. We used a mixture of two cell-lines, 22RV1 and 49C, to simulate normal and tumour tissues respectively. While mixing these cell-lines, we generated samples with varying amounts of DNA (10ng, 5ng, 2.5ng and 1ng for AmpliSeq and 50ng, 25ng, 10ng, and 5ng for Illumina). For each amount, we mixed the cell-lines in different proportions to simulate different dilutions (5:1, 10:1, 20:1 and 50:1 for AmpliSeq and 10:1, 20:1 and 50:1 for Illumina). Finally, we generated 16 and 12 samples by IonTorrent Proton and Illumina MiSeq respectively.

3.2.1 Experimental design for the cell-line data.

We used a mixture of prostate cancer cell lines 22RV1 and a LNCaP derivative 49C (Al Nakouzi *et al.*, 2013) containing different known mutations as an experimental model. We used a 14 gene TSCA gene panel designed using Illumina DesignStudio[4]. We targeted 70,929 bases in total (175bp amplicon size). The targeted genes were APC, CDK12, AR, SPOP, TP53, PTEN, BRCA1, BRCA2, CHEK2, MYC, FOXA1, MED12, HSD3B1, ASXL1. Ampliseq panel targeted 19 genes (AR, TP53, BRCA1, BRCA2, MED12, ASXL1, CTNNB1, OR5A1, PIK3CA, SCN11A, CHD1, KDM6A, SPOP, HSDB3, PTEN, MLL, MYC, CHEK2, FOXA1), had 104.67bp genome footprint and was designed using AmpliSeq Designer[5] applying FFPE parameters (amplicon target range 125-175bp). The sequencing was performed at Vancouver Prostate Centre using MiSeq (KAPA library quantification kit, 25M 2x300bp read kit) and Ion Proton (80M fragments, Ion PI sequencing reagents kit 200 v3, Ion PI chip kit v3) sequencers according to manufacturer's instructions. Each library preparation run included negative (no DNA added) control, in all cases the number of reads from the negative control was negligible ($< 5,000$ reads, compared with $> 1,000,000$ reads for target libraries).

3.2.2 AmpliSeq Calibration Data.

We evaluated the sensitivity of SiNVICT on these 16 samples by examining three previously validated SNVs (H875Y, F877L and T878A) within the AR gene (Azad *et al.*, 2015) that belongs to **only** one of the two cell-lines. H875Y and T878A are homozygous SNVs while F877L is a heterozygous SNV.

SiNVICT successfully detected all three mutations in all dilutions of 10ng, 5ng and 2.5ng. However it failed to detect the heterozygous F877L mutation at 20:1 and 50:1 dilutions of 1ng which had observed allele frequencies of 0.07% and 0.83%, respectively. The failed case at 50:1 dilution is likely due to the very low amount of DNA used and the "tumour" cell-line being highly diluted at this amount.

In summary, SiNVICT was able to detect 46 of the 48 cases (sensitivity of 95.8%). The lowest observed allele frequency for successful detection of a mutation was 1.24% (F877L, 10ng, 50:1). SiNVICT failed to detect 2 cases that fell below 1% observed allele frequency (F877L at 20:1-1ng and F877L at 50:1-1ng). Freebayes only reported the H875Y in all 16 samples and failed to detect other mutations. Freebayes in total detected 16 out 48 (sensitivity of 33.3%) validated calls. VarScan2 failed to call H875Y mutation in dilutions 5:1 and 10:1. Similar to SiNVICT, it failed to call F877 at 20:1-1ng. Unlike SiNVICT, it reported F877L at 50:1-1ng sample. VarScan2 in total reported 39 out of 48 (sensitivity of 81.25%) validated calls. MuTect on this dataset reported 47 out of 48 (sensitivity of 97.92%) cases. MuTect was the only tool to report F877 at 50:1-1ng. See Table 1 for details about the SNV calls and Supplementary Tables 5 and 6 for details about the read statistics.

3.2.3 Illumina Calibration Data.

We evaluated the sensitivity of SiNVICT on calibration dataset generated through Illumina sequencing technology by examining the same validated mutations. As depicted in Table 2, SiNVICT had the best performance on this dataset by being able to detect 33 out of 36 (sensitivity of 91.6%) cases while VarScan2 and Freebayes detected 25 (sensitivity of 69.4%) and 13 (sensitivity of 36.1%) respectively. MuTect was able to detect 28 (sensitivity of 77.77%) out of 36 cases. The highest undetected mutation had around 0.3% observed variant allele frequency. We have provided more details about this experiment in Supplementary Tables 7 and 8.

[4] http://designstudio.illumina.com/truseqca/project/new

[5] https://www.ampliseq.com/protected/startPage.action

Table 1. SNV calling on AmpliSeq calibration data generated from mixtures of 22RV1 and 49C cell lines.

		DNA Amount															
		10ng				5ng				2.5ng				1ng			
Dilution	Mutation	ST[a]	MT[b]	VS[c]	FB[d]	ST[a]	MT[b]	VS[c]	FB[d]	ST[a]	MT[b]	VS[c]	FB[d]	ST[a]	MT[b]	VS[c]	FB[d]
5:1	H875Y	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓
	T878A	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗
	F877L	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗
10:1	H875Y	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓
	T878A	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗
	F877L	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗
20:1	H875Y	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	T878A	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗
	F877L	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗
50:1	H875Y	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	T878A	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗
	F877L	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✗	✓	✓	✗

Three previously validated SNVs on the AR gene for the mixture of the 22RV1-49C cell-lines (H875Y, F877L, and T878A) were used to evaluate the sensitivity of SiNVICT in detection of SNVs in real data. Varying amounts of DNA (10ng, 5ng, 2.5ng and 1ng) were used to prepare each of the samples. For each amount, we mixed the two cell-lines in different proportions to simulate different dilutions (5:1, 10:1, 20:1 and 50:1). Note that, expected allele frequency for F877L is half of that for T878A due to F877L being a heterozygous mutation.
[a] SiNVICT. [b] MuTect. [c] VarScan2. [d] Freebayes.

Table 2. SNV calling on Illumina calibration data generated from mixtures of 22RV1 and 49C cell lines.

		DNA Amount															
		50ng				25ng				10ng				5ng			
Dilution	Mutation	ST[a]	MT[b]	VS[c]	FB[d]	ST[a]	MT[b]	VS[c]	FB[d]	ST[a]	MT[b]	VS[c]	FB[d]	ST[a]	MT[b]	VS[c]	FB[d]
10:1	H875Y	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
	T878A	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	✗
	F877L	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	✗
20:1	H875Y	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓
	T878A	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✓	✓	✗	✗	✗
	F877L	✓	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
50:1	H875Y	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	T878A	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
	F877L	✓	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	✗	✓	✓	✗	✗

[a] SiNVICT. [b] MuTect. [c] VarScan2. [d] Freebayes.

3.3 Cell-Free DNA from castration-resistant prostate cancer patients

We obtained two datasets part of a larger study (Wyatt *et al.*, 2016) to assess the feasibility of using SiNVICT on cfDNA from cancer patients. One of these datasets consisted of cfDNA sequencing data from castration-resistant prostate cancer patients sequenced with the Ion Torrent (AmpliSeq) technology and the other dataset composed of cfDNA sequencing data from metastatic castration-resistant prostate cancer patients sequenced with Illumina MiSeq. The AmpliSeq panel covered several genes whereas the Illumina panel was limited to the AR gene. More details about the sequencing panels can be found at http://www.ebi.ac.uk/ena/data/view/PRJEB11659.

First, we performed a basic validation step to ensure that SiNVICT can perform basic SNV calling in cfDNA using the AmpliSeq dataset with several previously validated mutations by the study. We were able to reproduce all these mutation calls with SiNVICT (see Supplementary Table 9). Given the high VAF of these mutations and the depths of these locations, the aim of this experiment was not to assess the sensitivity of SiNVICT, but to make sure it passed a simple preliminary test for handling cfDNA sequencing data.

We then selected 12 patients from the Illumina dataset belonging to the same study (Wyatt *et al.*, 2016) that were sequenced at all three time points of interest - baseline, on-treatment (12-weeks), and progression - and whose respective samples passed quality checks. These samples were sequenced to obtain DNA from exons 2-8 of the AR gene. Candidate locations suitable for time series analysis were selected based on the methodology described in Section 2.4.

We then plotted the variant allele frequencies for these locations for a patient at the three time points and observed a trend in which one time point

shows an increase in the VAF despite the other two time points showing little evidence of a variant being present (see Supplementary Figure 1).

Based on this observation, we tried to assess whether the drug treatment had eliminated some of the subclones while providing selective advantage to some others that were already present in minuscule amounts before treatment, through recalculating the p-values based on SiNVICT's time series data analysis feature. The recalculated p-values were significantly different than the original p-values (see Supplementary Table 10) implied by the error rate for the (Illumina) sequencing technology, which might be an indicator of a subclone being present at other time points in very low amounts, making it difficult to detect by standard (non-time-series) analysis.

4 Conclusion

SiNVICT is a highly accurate and sensitive tool for detection of SNVs and short indels in circulating tumour DNA at very low variant allele percentages. Mutation detection with high read depth is often difficult due to sequencing errors getting amplified with the Amplicon technology used in most deep sequencing platforms. SiNVICT is capable of filtering mutation calls by several parameters such as the minimum read depth, strand bias, etc. We provide more details on the effect of filters on the experiments performed in Section 3 in Supplementary Table 11. SiNVICT is also highly customisable, allowing the user to fine-tune several parameters to achieve the desired level of sensitivity and specificity. Time-series analysis capabilities of SiNVICT might be utilised to gain insight to how certain drug treatments affect the overall clonal composition for a patient.

Results obtained from experiments on simulated data suggest that at variant allele percentages below 0.5%, even increasing the read depth indefinitely will not help with the calls unless the sequencing errors are reduced. Results obtained from the cell-line experiments might allow us to speculate that 2.5ng for AmpliSeq and 10ng for Illumina are the safest amounts of DNA (among our calibration samples) from which a set of reliable calls can be obtained at all dilutions mentioned before.

Acknowledgement

We thank Salem Malikic for helping us construct sample phylogenetic trees to model intra-tumour heterogeneity and his comments on our simulations. We thank Yen-Yi Lin for providing assistance with the generation of the plots and figures in the manuscript. We also thank Ibrahim Numanagic for his comments and assistance.

Funding: This research is funded by NSERC Create MADD-Gen Training program to Can Kockan, NSERC Discovery Frontiers grant on "Cancer Genome Collaboratory", Genome Canada and Canadian Cancer Society Research Institute Innovation grants to S. Cenk Sahinalp, Terry Fox Research Institute New Frontiers Program (grant #TFF116129) to Colin C. Collins, Canadian Cancer Society (grant #702837) and Prostate Cancer Canada (grant #D2014-13) to Kim N. Chi and is proudly funded in part by the Movember Foundation.

References

Al Nakouzi, N., Wang, C., Jacoby, D., Gleave, M. E., and Zoubeidi, A. (2013). Abstract c89: Galeterone suppresses castration-resistant and enzalutamide-resistant prostate cancer growth in vitro. *Molecular Cancer Therapeutics*, **12**(11 Supplement), C89–C89.

Azad, A. A., Volik, S. V., Wyatt, A. W., Haegert, A., Bihan, S. L., Bell, R. H., Anderson, S. A., McConeghy, B., Shukin, R., Bazov, J., Youngren, J., Paris, P., Thomas, G., Small, E. J., Wang, Y., Gleave, M. E., Collins, C. C., and Chi, K. N. (2015). Androgen receptor gene aberrations in circulating cell-free DNA: Biomarkers of therapeutic resistance in castration-resistant prostate cancer. *Clinical Cancer Research*, **21**(10), 2315–2324.

Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B. R., Wang, H., Luber, B., Alani, R. M., *et al.* (2014). Detection of circulating tumor dna in early-and late-stage human malignancies. *Science translational medicine*, **6**(224), 224ra24–224ra24.

Bubendorf, L., Schöpfer, A., Wagner, U., Sauter, G., Moch, H., Willi, N., Gasser, T. C., and Mihatsch, M. J. (2000). Metastatic patterns of prostate cancer: An autopsy study of 1, 589 patients. *Human Pathology*, **31**(5), 578–583.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*, **31**(3), 213–219.

Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv e-prints*.

Glenn, T. C. (2011). Field guide to next-generation dna sequencers. *Molecular ecology resources*, **11**(5), 759–769.

Guo, Y., Li, J., Li, C.-I., Long, J., Samuels, D. C., and Shyr, Y. (2012). The effect of strand bias in illumina short-read sequencing data. *BMC Genomics*, **13**(1), 666.

Illumina (2013). Amplicon - ds somatic variant caller. Technical report, Illumina Inc., 5200 Illumina Way (formerly 5200 Research Pl) San Diego, CA 92122 USA.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, **22**(3), 568–576.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**(14), 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Lipson, E. J., Velculescu, V. E., Pritchard, T. S., Sausen, M., Pardoll, D. M., Topalian, S. L., and Diaz Jr, L. A. (2014). Circulating tumor dna analysis as a real-time method for monitoring tumor burden in melanoma patients undergoing treatment with immune checkpoint blockade. *J Immunother Cancer*, **2**(1), 42.

Malikic, S., McPherson, A. W., Donmez, N., and Sahinalp, C. S. (2015). Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, **31**(9), 1349–1356.

Mandel, P. (1948). Les acides nucleiques du plasma sanguin chez l'homme. *CR Acad Sci Paris*, **142**, 241–243.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**(9), 1297–1303.

Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., Sloan, C. A., Rosenbloom, K. R., Roe, G., Rhead, B., Raney, B. J., Pohl, A., Malladi, V. S., Li, C. H., Lee, B. T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., Harte, R. A., Haeussler, M., Guruvadoo, L., Goldman, M., Giardine, B. M., Fujita, P. A., Dreszer, T. R., Diekhans, M., Cline, M. S., Clawson, H., Barber, G. P., Haussler, D., and Kent, W. J. (2012). The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Research*, **41**(D1), D64–D69.

Mose, L. E., Wilkerson, M. D., Hayes, D. N., Perou, C. M., and Parker, J. S. (2014). ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics*, **30**(19), 2813–2815.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., *et al.* (2011). Sequence-specific error profile of illumina sequencers. *Nucleic acids research*, page gkr344.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6), 841–842.

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, **28**(14), 1811–1817.

Schwarzenbach, H., Hoon, D. S. B., and Pantel, K. (2011). Cell-free nucleic acids as biomarkers in cancer patients. *Nature Reviews Cancer*, **11**(6), 426–437.

Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D., *et al.* (2015). A big bang model of human colorectal tumor growth. *Nature genetics*, **47**(3), 209–216.

Speedy Publishing LLC (2014). *Statistics Equations And Answers (Speedy Study Guide)*. Speedy Publishing LLC.

Wyatt, A. W., Azad, A. A., Volik, S. V., Annala, M., Beja, K., McConeghy, B., Haegert, A., Warner, E. W., Mo, F., Brahmbhatt, S., *et al.* (2016). Genomic alterations in cell-free dna and enzalutamide resistance in castration-resistant prostate cancer. *JAMA oncology*.

Xin, H., Lee, D., Hormozdiari, F., Yedkar, S., Mutlu, O., and Alkan, C. (2013). Accelerating read mapping with FastHASH. *BMC Genomics*, **14 Suppl 1**, S13.