METHODS FOR THE DETECTION OF SINGLE NUCLEOTIDE VARIANTS AND INDELS FROM CELL-FREE DNA

by

Can Kockan
B.Sc. , Bilkent University, Turkey, 2014

Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Summer 2016

APPROVAL

Name: Can Kockan

Degree: Master of Science

Title of Thesis: METHODS FOR THE DETECTION OF SINGLE NUCLEOTIDE VARIANTS AND INDELS FROM CELL-FREE DNA

Examining Committee:

Dr. Cedric Chauve
Chair

Dr. S. Cenk Sahinalp,
Professor, Senior Supervisor

Dr. Faraz Hach,
Research Associate, Supervisor

Dr. Martin Ester,
Professor, Internal Examiner

Date Approved: August 4th, 2016

Abstract

Successful development and application of precision oncology approaches require robust elucidation of the genomic landscape of a patient's cancer and 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, is considered as one of the most promising approaches to achieve this goal. In many cancer patients, especially those with advanced metastatic disease, deep sequencing of 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 (ctDNA). The ctDNA/cfDNA ratio and 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. We introduce SiNVICT, a computational method for analysis of cfDNA sequencing data.

keywords: Cancer genomics, SNV calling, cell-free DNA

To my dear parents Sevgi and Orhan,
and my brother Umit.

“Common sense is not so common.”

— François-Marie Arouet

Acknowledgments

First and foremost, I would like to thank Dr. S. Cenk Sahinalp and Dr. Faraz Hach for their support and guidance throughout my studies. I also thank Iman Sarrafi for his contributions to this work as well as my development as a computer scientist.

This project would not have been realized without the efforts of Brian McConeghy, Kevin Beja, Anne Haegert, Robert H. Bell, Dr. Alexander W. Wyatt, Dr. Kim N. Chi, Dr. Stanislav V. Volik, and Dr. Colin C. Collins from Vancouver Prostate Centre.

I am grateful to Salem Malikic, Yen-Yi Lin, Ibrahim Numanagic, and all my colleagues and friends from the Lab for Computational Biology at Simon Fraser University for their input and assistance on my work.

Contents

Approval ii

Abstract iii

Dedication iv

Quotation v

Acknowledgments vi

Contents vii

List of Tables ix

List of Figures x

1 Introduction 1

2 Methods 4

2.1 Preprocessing Steps 5

2.2 Somatic Mutation and Indel Discovery Step 5

2.3 Postprocessing Steps 7

2.4 Time Series Analysis 9

3 Results 11

3.1 Simulated Data 11

3.1.1 SNV calling on simulated data. 11

3.1.2 Indel calling on simulated data. 12

3.1.3 SNV calling on simulated data with tumour heterogeneity. 12

3.2 AmpliSeq and Illumina 22RV1-49C Calibration data 14

3.2.1 Experimental design for the cell-line data. 14

3.3 Illumina Calibration Data. 18
3.4 Cell-Free DNA from castration-resistant prostate cancer patients 18

4 Conclusion **20**

Bibliography **21**

Appendix A Appendix:Experiment Details **24**

List of Tables

3.1 SNV calling on AmpliSeq calibration data generated from mixtures of 22RV1 and 49C cell lines. . . . 17
3.2 SNV calling on Illumina calibration data generated from mixtures of 22RV1 and 49C cell lines. . . . 18

A.1 Command-lines and parameters used to execute MuTect, VarScan2, and Freebayes throughout the experiments . . . 25
A.2 Precision and Recall on SNV calling for simulated data . . . 25
A.3 Precision and Recall on indel calling on simulated data . . . 25
A.4 Precision and Recall on SNV calling on simulated data with tumour heterogeneity . . . 26
A.5 AmpliSeq calibration data - expected and observed allele frequencies . . . 27
A.6 AmpliSeq Calibration Experiment - Read Statistics for different variants. . . . 28
A.7 Illumina calibration data - expected and observed allele frequencies . . . 29
A.8 Illumina Calibration Experiment - Read Statistics for different variants. . . . 29
A.9 Validated^{a} somatic mutations detected by SiNVICT in cfDNA samples (AmpliSeq sequencing data) at progression on enzalutamide . . . 30
A.10 Time-Series Analysis for VC-007 . . . 32
A.11 SiNVICT - Effect of filters on the number of calls. . . . 32

List of Figures

2.1 Overview of SiNVICT data processing pipeline. 4

2.2 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. . 8

3.1 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%, 5%, and 2,5% by SiNVICT and VarScan2. Freebayes can successfully detect the SNVs at 50% tumour. In all the cases, SiNVICT had a better precision than MuTect, Freebayes, and VarScan2 however in tumour content of 1% VarScan2 had a better recall (1 more SNV detected by VarScan2). VarScan2 made relatively more number of false positive calls resulting in its lower precision. Details about number of calls are provided in Table A.2. 13

3.2 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 3.3 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. Note that 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. Due to 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. SiNVICT only considers the most frequent non-reference base change and ignores the other lower frequency base changes - this is most likely the reason why SiNVICT makes calls in the first place where other tools do not. For instance, at a sample location with a read depth of 80,000 and a reference base A, one might observe 78,000 reads matching the reference, 1500 reads suggesting an A to T change, and 500 reads suggesting an A to C change. Detailed information about call statistics are provided in Table A.4 15

3.3 Sample phylogenetic tree with 5 clones, randomly selected topology and prevalences to simulate the conditions of SNV detection for a very heterogeneous tumour. 16

A.1 The y-axis shows the variant allele frequencies for the locations given on the x-axis for patient VC-007 from the dataset described in Section 3.3 at the three time points shown in the legend. We 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, 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. 31

Chapter 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 [5]. 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 [15]. 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 [22]. 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 tumours 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) [4] 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 [4] 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

[1] IQR: the interquartile range is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a

somatic and germline mutation callers that have been 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 [16], VarScan2 [10], Freebayes [7], Strelka [21], MuTect [6], 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 –which is increasing in quantity due to the recent interest in liquid biopsy (ii) very high read depth (e.g. 20k-30k average, up to 90k and possibly more in the future with the advances in Deep Amplicon Sequencing and similar sequencing methods), or, (iii) extremely low dilutions (can be as low as around 0.01% variant allele percentage [14]), or, (iv) samples with high tumour heterogeneity, or, (v) batches of samples that suffer from systematic noise.

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 addressed challenges (i) and (ii) through the combination of a Poisson model and a number of postprocessing filters such as the minimum read depth filter. Challenge (v) is addressed through a Signal-to-Noise ratio filter. While the Poisson model and the postprocessing steps utilised by SiNVICT allow significant improvements to overcome challenges (iii) and (iv), these can still be considered as open bioinformatics problems.

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 different 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 in more detail in Methods) 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), which were applied to the same tumour samples. Our experiments indicate that SiNVICT is highly sensitive to calls on data

---

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 [24].

generated by both sequencing platforms. For example, three previously validated AR (Androgen Receptor gene) mutations [4] 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 data sets with high sequencing depth - or multiple occurrences of identical reads as PCR duplicates (for example, in one experiment GATK reduced the depth of coverage from 20K to 300). Identical reads are to be expected in deep amplicon sequencing and this is not necessarily an artifact of PCR.

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

Chapter 2

Methods

As shown in the flowchart in Figure 2.1 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.

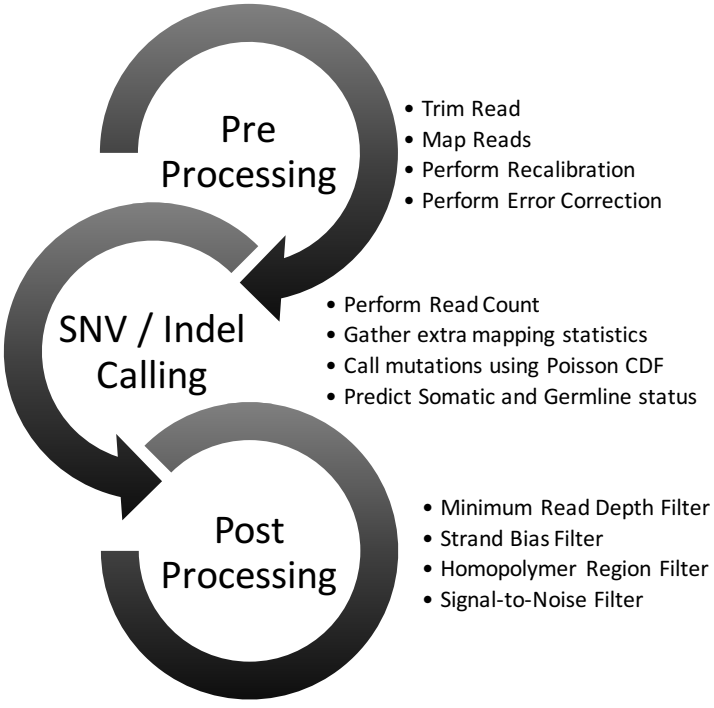

Figure 2.1: Overview of SiNVICT data processing pipeline.

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.

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 [26], BWA [12], 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 sub-step. 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 [18] 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 SNV (or indel without any difference in the model) 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 [9].

$$p_1 = P(K|\lambda_1) = e^{-\lambda_1} \sum_{i=0}^{\lfloor K \rfloor} \frac{\lambda_1^i}{i!} \tag{2.1}$$

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, $\lambda_1 = N \times r$, which gives the expected number of errors. If the number of reads supporting a mutation (K) is significantly greater than this value (λ_1), then there is a greater probability that an actual mutation has occurred at the current genomic position.

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 $\lambda_2 = N/2$.

$$p_2 = P(K|\lambda_2) = e^{-\lambda_2} \sum_{i=0}^{\lfloor K \rfloor} \frac{\lambda_2^i}{i!} \tag{2.2}$$

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 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. SiNVICT provides default values for the parameters of these postprocessing steps based on the observed characteristics of cfDNA sequencing data used in our experiments, but users can change these to values better suited for their own data. The postprocessing steps SiNVICT applies are 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" (user defined) regions, the sequencing errors can be mis-interpreted as SNVs or indels and thus are filtered out.

Strand Bias filter. The strand bias for a genomic location i (see equation 2.3) 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.

$$\mathrm{StrandBias}_i = \frac{\mathrm{NumReadsForward}_i}{\mathrm{NumReadsTotal}_i} \tag{2.3}$$

If the potential strand bias is outside of the range $[0.5 - \epsilon, 0.5 + \epsilon]$ (for $\epsilon < 0.2$), 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 [8]. 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.

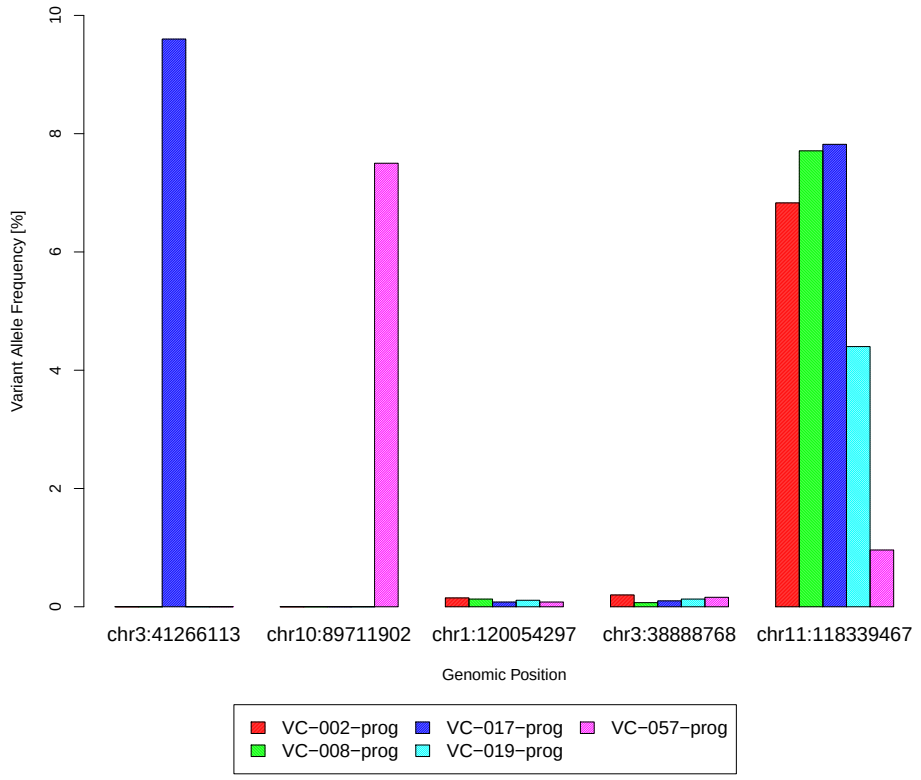

Figure 2.2: Calculating the Signal to Noise Ratio (SNR) in multiple samples across several genomic loci. The x-axis depicts 5 loci from a set of samples such that each sample is represented by a distinct colour. For each location on the x-axis, the corresponding value on the y-axis indicates the percentage 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 2.2). 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 2.4. The mean μ_i can be calculated as per equation 2.5 and the standard deviation, σ_i, is given in equation 2.6; 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.

$$\mathrm{SNR}_i = \frac{\mu_i}{\sigma_i} \tag{2.4}$$

$$\mu_i = \frac{\sum_{j=1}^{n} \mathrm{VariantAllelePercentage}_j^i}{n} \tag{2.5}$$

$$\sigma_i^2 = \frac{\sum_{j=1}^{n} (\mathrm{VariantAllelePercentage}_j^i - \mu_i)^2}{n} \tag{2.6}$$

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 [23]. 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 2.1) 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 [19]. 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.

Chapter 3

Results

In order to evaluate our method, we performed the following experiments: (i) we simulated insilico cfDNA/ctDNA 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 [25] [1]. 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- [17] of the human reference genome. The parameters that are used to run each tool are provided in Table A.1. We extracted the exons of the AR gene with BEDTools [20], 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 [13]) 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

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

= 175). We obtained an average read depth of ~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. Freebayes had high recall and precision for this dataset down to 10% tumour content level while MuTect and VarScan2 kept a consistent level of high recall and precision down to 1%. 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 recall than VarScan2. See Figure 3.1 and Table A.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 ~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 Table A.3 for details.

3.1.3 SNV calling on simulated data with tumour heterogeneity.

We evaluated all methods on a more challenging dataset by building a sample tumour phylogeny to simulate the effect of 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; each clone also inherited all mutations from its parent clone. See Figure 3.3 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.

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 ~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

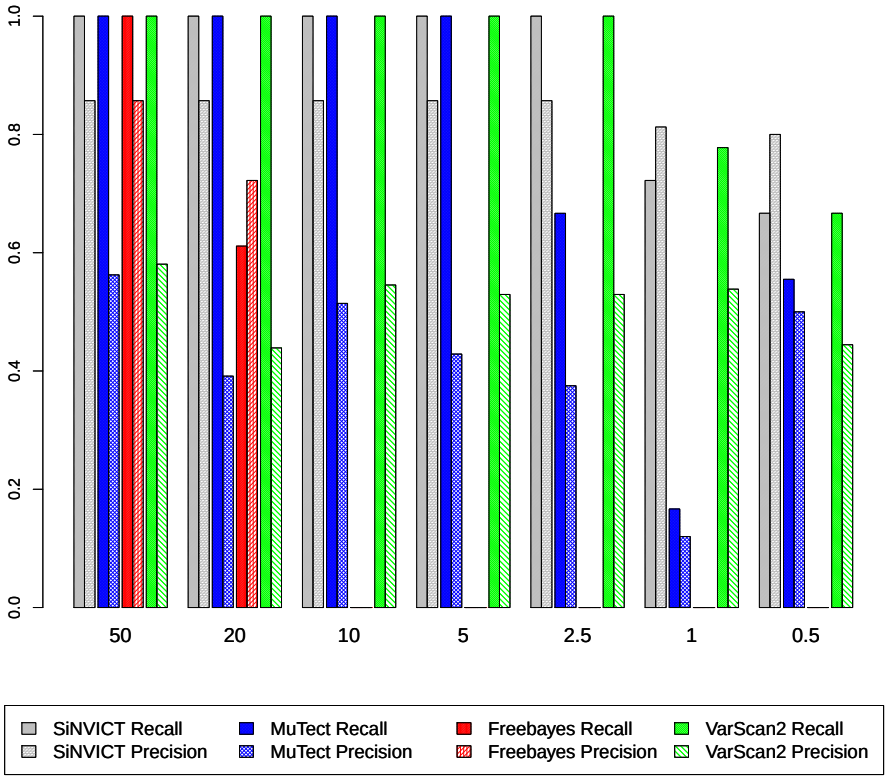

Figure 3.1: Precision and Recall of SiNVICT, MuTect, Freebayes, and VarScan2 on simulated data. x-axis represents different samples with different tumour content levels. In each of these simulated samples, there were 18 manually added SNVs. The total number of bases covered per sample was 8938. All of the 18 SNVs were successfully detected in samples at tumour content levels of 50%, 20%, 10%, 5%, and 2,5% by SiNVICT and VarScan2. Freebayes can successfully detect the SNVs at 50% tumour. In all the cases, SiNVICT had a better precision than MuTect, Freebayes, and VarScan2 however in tumour content of 1% VarScan2 had a better recall (1 more SNV detected by VarScan2). VarScan2 made relatively more number of false positive calls resulting in its lower precision. Details about number of calls are provided in Table A.2.

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. Due to 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. SiNVICT only considers the most frequent non-reference base change and ignores the other lower frequency base changes - this is most likely the reason why SiNVICT makes calls in the first place where other tools do not. For instance, at a sample location with a read depth of 80,000 and a reference base A, one might 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 3.2 and Table A.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 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. For Illumina we did not generate any sequencing data for 5:1 dilutions.

3.2.1 Experimental design for the cell-line data.

We have used a mixture of prostate cancer cell lines 22RV1 and a LNCaP derivative 49C [2] containing different known mutations as an experimental model. We have used a 14 gene TSCA gene panel designed using Illumina DesignStudio (http://designstudio.illumina.com/truseqca/project/new) and 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

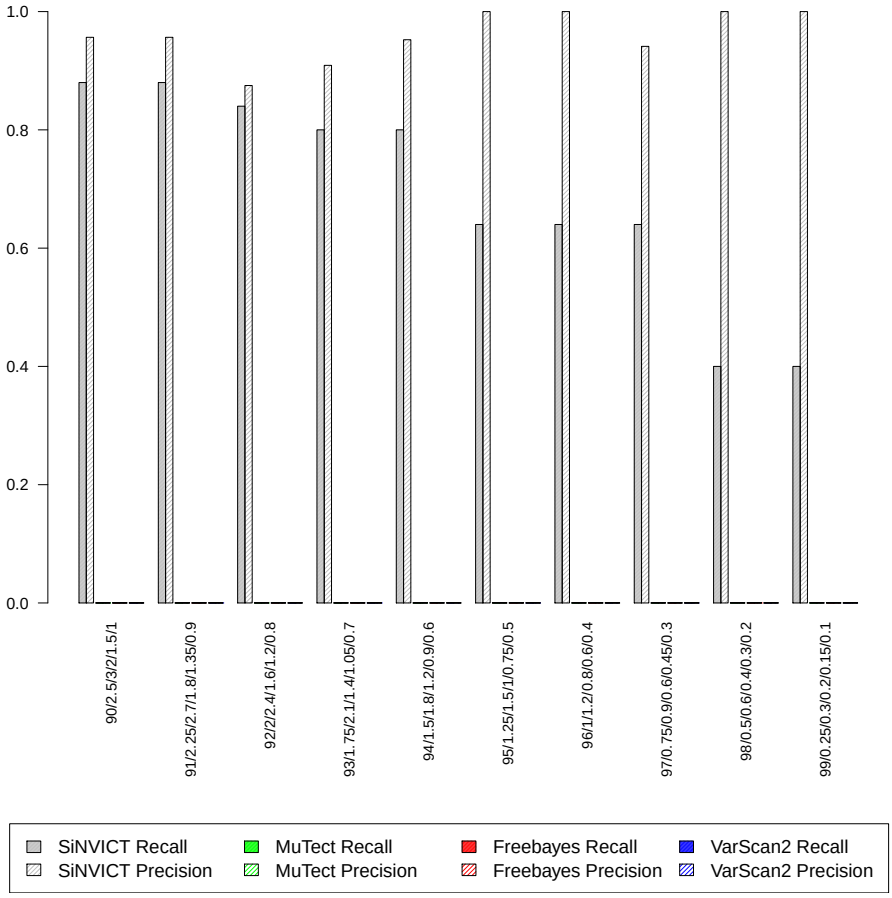

Figure 3.2: Precision and Recall of SiNVICT, MuTect, Freebayes, and VarScan2 on simulated data consisting of 5 clones. To each clone, we added 5 SNVs and each subclone inherited additional mutations from their parents as shown in Figure 3.3 and the number of bases covered was 31485. Detection abilities 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. Note that 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. Due to 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. SiNVICT only considers the most frequent non-reference base change and ignores the other lower frequency base changes - this is most likely the reason why SiNVICT makes calls in the first place where other tools do not. For instance, at a sample location with a read depth of 80,000 and a reference base A, one might observe 78,000 reads matching the reference, 1500 reads suggesting an A to T change, and 500 reads suggesting an A to C change. Detailed information about call statistics are provided in Table A.4

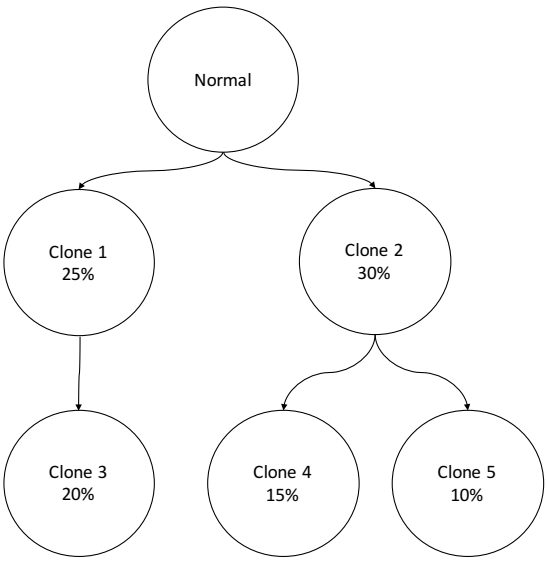

Figure 3.3: Sample phylogenetic tree with 5 clones, randomly selected topology and prevalences to simulate the conditions of SNV detection for a very heterogeneous tumour.

104.67bp genome footprint and was designed using AmpliSeq Designer 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).

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 [4] that belongs to **only** one of the two cell-lines. H875Y and T878A are homozygous SNVs while F877L is a heterozygous SNV. We used F877L and T878A mutations to evaluate the sensitivity of SiNVICT.

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

Table 3.1: 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). The rarest variant which could be successfully detected by SiNVICT was at 1% expected allele frequency. 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.

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 freqency (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 reproted 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 3.1 for details about the SNV calls and Tables A.5 and A.6 for details about the read statistics.

Table 3.2: 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 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 3.2, 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 lowest recorded mutation that was detected by SiNVICT had 1% expected allele frequency while the highest undetected mutation had around 0.3% observed variant allele frequency. We have provided more details about this experiment in Tables A.7 and A.8.

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

We obtained two datasets part of a larger study [25] 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 Table A.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 parent project 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 Figure A.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 Table A.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.

Chapter 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 table A.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 25ng/10ng seem to be the safest amounts of DNA (among our calibration samples) from which a set of reliable calls can be obtained at all dilutions mentioned before. Meanwhile, 5ng seems to be a logical choice for a lower limit before SNV/indel calling deteriorates significantly.

Bibliography

[1] The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Research*, 41(D1):D64–D69, January 2013.

[2] Nader Al Nakouzi, Chris Wang, Douglas Jacoby, Martin E Gleave, and Amina Zoubeidi. Abstract c89: Galeterone suppresses castration-resistant and enzalutamide-resistant prostate cancer growth in vitro. *Molecular Cancer Therapeutics*, 12(11 Supplement):C89–C89, 2013. 14

[3] Can Alkan, Jeffrey M Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O Kitzman, Carl Baker, Maika Malig, Onur Mutlu, S Cenk Sahinalp, Richard A Gibbs, and Evan E Eichler. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics*, 41(10):1061–1067, aug 2009.

[4] A. A. Azad, S. V. Volik, A. W. Wyatt, A. Haegert, S. Le Bihan, R. H. Bell, S. A. Anderson, B. McConeghy, R. Shukin, J. Bazov, J. Youngren, P. Paris, G. Thomas, E. J. Small, Y. Wang, M. E. Gleave, C. C. Collins, and K. N. Chi. 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, feb 2015. 1, 3, 16

[5] Lukas Bubendorf, Alain Schöpfer, Urs Wagner, Guido Sauter, Holger Moch, Niels Willi, Thomas C. Gasser, and Michael J. Mihatsch. Metastatic patterns of prostate cancer: An autopsy study of 1, 589 patients. *Human Pathology*, 31(5):578–583, may 2000. 1

[6] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*, 31(3):213–219, feb 2013. 2

[7] E. Garrison and G. Marth. Haplotype-based variant detection from short-read sequencing. *ArXiv e-prints*, July 2012. 2

[8] Yan Guo, Jiang Li, Chung-I Li, Jirong Long, David C Samuels, and Yu Shyr. The effect of strand bias in illumina short-read sequencing data. *BMC Genomics*, 13(1):666, 2012. 7

[9] Illumina. Amplicon - ds somatic variant caller. Technical report, Illumina Inc., 5200 Illumina Way (formerly 5200 Research Pl) San Diego, CA 92122 USA, 2013. 6

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

[11] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, mar 2012.

[12] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, may 2009. 5

[13] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, jun 2009. 11

[14] Evan J Lipson, Victor E Velculescu, Theresa S Pritchard, Mark Sausen, Drew M Pardoll, Suzanne L Topalian, and Luis A Diaz Jr. 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, 2014. 2

[15] P Mandel. Les acides nucleiques du plasma sanguin chez l'homme. *CR Acad Sci Paris*, 142:241–243, 1948. 1

[16] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, jul 2010. 2

[17] L. R. Meyer, A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, B. J. Raney, A. Pohl, V. S. Malladi, C. H. Li, B. T. Lee, K. Learned, V. Kirkup, F. Hsu, S. Heitner, R. A. Harte, M. Haeussler, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, T. R. Dreszer, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent. The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Research*, 41(D1):D64–D69, nov 2012. 11

[18] L. E. Mose, M. D. Wilkerson, D. N. Hayes, C. M. Perou, and J. S. Parker. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics*, 30(19):2813–2815, jun 2014. 5

[19] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C Linak, Aki Hirai, Hiroki Takahashi, et al. Sequence-specific error profile of illumina sequencers. *Nucleic acids research*, page gkr344, 2011. 10

[20] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, jan 2010. 11

[21] C. T. Saunders, W. S. W. Wong, S. Swamy, J. Becq, L. J. Murray, and R. K. Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817, may 2012. 2

[22] Heidi Schwarzenbach, Dave S. B. Hoon, and Klaus Pantel. Cell-free nucleic acids as biomarkers in cancer patients. *Nature Reviews Cancer*, 11(6):426–437, may 2011. 1

[23] Andrea Sottoriva, Haeyoun Kang, Zhicheng Ma, Trevor A Graham, Matthew P Salomon, Junsong Zhao, Paul Marjoram, Kimberly Siegmund, Michael F Press, Darryl Shibata, et al. A big bang model of human colorectal tumor growth. *Nature genetics*, 47(3):209–216, 2015. 9

[24] Speedy Publishing LLC. *Statistics Equations And Answers (Speedy Study Guide)*. Speedy Publishing LLC, 2014. 2

[25] Alexander W Wyatt, Arun A Azad, Stanislav V Volik, Matti Annala, Kevin Beja, Brian McConeghy, Anne Haegert, Evan W Warner, Fan Mo, Sonal Brahmbhatt, et al. Genomic alterations in cell-free dna and enzalutamide resistance in castration-resistant prostate cancer. *JAMA oncology*, 2016. 11, 18

[26] H. Xin, D. Lee, F. Hormozdiari, S. Yedkar, O. Mutlu, and C. Alkan. Accelerating read mapping with FastHASH. *BMC Genomics*, 14 Suppl 1:S13, 2013. 5

Appendix A

Appendix:Experiment Details

Table A.1 Command-lines and parameters used to execute MuTect, VarScan2, and Freebayes throughout the experiments

Tool	Command-Line	Parameters
SiNVICT	sinvict -t tumor-directory-path -o output-directory-path	–error-rate 0.01 –min-depth 100 –QScoreCutoff 20 –readEndFraction 0.01 –leftStrandBias 0.3 (disabled for AmpliSeq) –rightStrandBias 0.7 (disabled for AmpliSeq)
MuTect	java -Xmx2g -jar mutect.jar	–analysis-type: MuTect –reference-sequence: hg19.fa
Freebayes	freebayes -f reference-genome input-bam $>$ output-vcf	–min-alternate-count: 2 –min-coverage: 0 –standard-filters: -m 30 -q 20 -R 0 -S 0 –min-alternate-total: 1
VarScan2	java -jar VarScan.v2.4.1.jar mpileup2snp input-mpileup –min-var-freq 0.001	Min coverage: 8 Min reads2: 2 Min var freq: 0.001 Min avg qual: 15 P-value thresh: 0.01

Table A.2 Precision and Recall on SNV calling for simulated data

Tumour	SiNVICT					MuTect					Freebayes					VarScan2				
Content (%)	NC[a]	TP[b]	FP[c]	R(%)[d]	P(%)[e]	NC	TP	FP	R(%)	P(%)	NC	TP	FP	R(%)	P(%)	NC	TP	FP	R(%)	P(%)
50	21	18	3	100.00	85.71	32	18	14	100.00	56.25	21	18	3	100.00	85.71	31	18	13	100.00	58.06
20	21	18	3	100.00	85.71	46	18	28	100.00	39.13	11	8	3	61.11	72.72	41	18	14	100.00	43.90
10	21	18	3	100.00	85.71	35	18	17	100.00	51.43	3	0	3	0.00	0.00	33	18	15	100.00	54.54
5	21	18	3	100.00	85.71	42	18	24	100.00	42.86	3	0	3	0.00	0.00	34	18	16	100.00	52.94
2.5	21	18	3	100.00	85.71	32	12	20	66.66	37.50	3	0	3	0.00	0.00	34	18	16	100.00	52.94
1	16	13	3	72.22	81.25	25	3	22	16.66	12.00	3	0	3	0.00	0.00	26	14	12	77.70	53.84
0.5	15	12	3	66.67	80.00	20	1	19	5.55	5.00	3	0	3	0.00	0.00	27	12	15	66.6	44.44

[a] NC: Number of Calls. [b] TP: True Positive. [c] FP: False Positive. [d] R: Recall. [e] P: Precision.
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%, 5%, and 2,5% by SiNVICT and VarScan2. MuTect was able to detect SNVs with high recall down to 2.5% level as well. Freebayes can successfully detect the SNVs at 50% tumour. In all the cases, SiNVICT had a better precision than MuTect, Freebayes, and VarScan2 however in tumour content of 1% VarScan2 had a better recall (1 more SNV detected by VarScan2). VarScan2 made relatively more number of false positive calls resulting in its lower precision.

Table A.3 Precision and Recall on indel calling on simulated data

Tumour	SiNVICT					MuTect					Freebayes					VarScan2				
Content (%)	NC[a]	TP[b]	FP[c]	R(%)[d]	P(%)[e]	NC	TP	FP	R(%)	P(%)	NC	TP	FP	R(%)	P(%)	NC	TP	FP	R(%)	P(%)
50	4	4	0	100.00	100.00	4	4	0	100.00	100.00	4	4	0	100.00	100.00	4	4	0	100.00	100.00
20	4	4	0	100.00	100.00	4	4	0	100.00	100.00	0	0	0	0.00	0.00	4	4	0	100.00	100.00
10	4	4	0	100.00	100.00	4	4	0	100.00	100.00	0	0	0	0.00	0.00	4	4	0	100.00	100.00
5	4	4	0	100.00	100.00	4	4	0	100.00	100.00	0	0	0	0.00	0.00	4	4	0	100.00	100.00
1	4	4	0	100.00	100.00	2	2	0	50.00	100.00	0	0	0	0.00	0.00	4	4	0	100.00	100.00

[a] NC: Number of Calls. [b] TP: True Positive. [c] FP: False Positive. [d] R: Recall. [e] P: Precision.
In each of these simulated samples, there were 4 manually added indels of size 2. The average read depth was 14000 with insert size of 150 and read length of 70. Freebayes detected all 4 indels successfully at 50% tumour content level without any false positives and did not make any calls on any other tumour content levels. All 4 indels were successfully detected by SiNVICT and VarScan2 in all samples with no false positives. MuTect only missed two indels at 1% level.

Table A.4 Precision and Recall on SNV calling on simulated data with tumour heterogeneity

Tumour	SiNVICT					MuTect					Freebayes					VarScan2				
Content (%)	NC[a]	TP[b]	FP[c]	R(%)[d]	P(%)[e]	NC	TP	FP	R(%)	P(%)	NC	TP	FP	R(%)	P(%)	NC	TP	FP	R(%)	P(%)
90/2.5/3/2/1.5/1	23	22	1	88.00	95.65	195	0	195	0.00	0.00	0	0	0	0.00	0.00	0	0	0	0.00	0.00
91/2.25/2.7/1.8/1.35/0.9	23	22	1	88.00	95.65	169	0	169	0.00	0.00	0	0	0	0.00	0.00	0	0	0	0.00	0.00
92/2/2.4/1.6/1.2/0.8	24	21	3	84.00	87.50	188	0	188	0.00	0.00	0	0	0	0.00	0.00	0	0	0	0.00	0.00
93/1.75/2.1/1.4/1.05/0.7	22	20	2	80.00	90.91	194	0	194	0.00	0.00	0	0	0	0.00	0.00	0	0	0	0.00	0.00
94/1.5/1.8/1.2/0.9/0.6	21	20	1	80.00	95.24	211	0	211	0.00	0.00	0	0	0	0.00	0.00	0	0	0	0.00	0.00
95/1.25/1.5/1/0.75/0.5	16	16	0	64.00	100.00	208	0	208	0.00	0.00	0	0	0	0.00	0.00	0	0	0	0.00	0.00
96/1/1.2/0.8/0.6/0.4	16	16	0	64.00	100.00	189	0	189	0.00	0.00	0	0	0	0.00	0.00	0	0	0	0.00	0.00
97/0.75/0.9/0.6/0.45/0.3	17	16	1	64.00	94.12	175	0	175	0.00	0.00	0	0	0	0.00	0.00	0	0	0	0.00	0.00
98/0.5/0.6/0.4/0.3/0.2	10	10	0	40.00	100.00	191	0	191	0.00	0.00	0	0	0	0.00	0.00	0	0	0	0.00	0.00
99/0.25/0.3/0.2/0.15/0.1	10	10	0	40.00	100.00	175	0	175	0.00	0.00	0	0	0	0.00	0.00	0	0	0	0.00	0.00

[a] NC: Number of Calls. [b] TP: True Positive. [c] FP: False Positive. [d] R: Recall. [e] P: Precision.
To each clone, we added 5 SNVs and each subclone inherited additional mutations from their parents. 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. Note that Freebayes, MuTect, and VarScan2 failed to provide any calls for these simulated data sets.

Table A.5 AmpliSeq calibration data - expected and observed allele frequencies

			DNA Amount			
			10ng	5ng	2.5ng	1ng
Dilution	Mutation	EAF[a] (%)	OAF[b](%)	OAF(%)	OAF(%)	OAF(%)
5:1	H875Y	80.0	67.6	73.2	62.6	74.1
	T878A	20.0	23.3	17.1	26.9	22.2
	F877L	10.0	11.6	9.0	14.8	11.3
10:1	H875Y	90.0	63.2	75.4	83.2	71.8
	T878A	10.0	16.2	10.6	12.9	5.6
	F877L	5.0	7.9	5.0	7.9	2.78
20:1	H875Y	95.0	75.6	73.4	82.2	86.9
	T878A	5.0	5.8	5.5	4.0	3.9
	F877L	2.5	3.9	2.9	1.7	0.07
50:1	H875Y	98.0	86.6	64.5	84.4	81.9
	T878A	2.0	1.8	2.3	3.0	2.8
	F877L	1.0	1.24	1.8	1.5	0.83

[a] EAF: Expected Allele Frequency. [b] OAF: Observed Allele Frequency.
The discrepancy between the expected and observed allele frequencies at 20:1 dilution level (1ng) is most likely due to a sudden drop in read depth which occurred during the sequencing of this sample.

Table A.6 AmpliSeq Calibration Experiment - Read Statistics for different variants.

		T878A(chrX:66943552)					F877L (chrX:66943549)					H875Y (chrX:66943543)				
Dilution	DNA (ng)	Read	A	C	G	T	Read	A	C	G	T	Read	A	C	G	T
5:1	10	2917	2236	0	681	0	2769	0	320	0	2449	3271	1	1058	0	2212
	5	3365	2791	0	574	0	3153	0	286	1	2866	3762	2	1006	0	2754
	2.5	2742	2004	1	737	0	2603	0	385	0	2218	3131	2	1168	1	1960
	1	2225	1731	0	494	0	2104	0	238	0	1866	2328	4	598	0	1726
10:1	10	3708	3104	1	602	1	3577	1	283	0	3293	4851	4	1780	0	3067
	5	2508	2243	0	265	0	2348	0	119	0	2229	2953	3	724	0	2226
	2.5	6220	5416	0	803	1	5894	0	465	0	5429	6412	7	1072	0	5333
	1	1979	1868	0	111	0	1898	0	53	0	1845	2558	1	720	1	1836
20:1	10	5247	4943	0	304	0	4378	0	173	2	4203	6221	7	1506	1	4707
	5	4969	4697	0	272	0	4068	0	120	0	3948	6192	5	1641	1	4545
	2.5	7183	6893	0	290	0	6003	0	102	0	5901	8015	19	1408	0	6588
	1	3614	3473	0	141	0	2807	0	2	1	2804	3866	9	499	0	3358
50:1	10	7130	7003	0	127	0	5866	0	73	1	5792	7754	11	1023	3	6717
	5	2995	2926	0	69	0	2490	0	44	0	2446	4350	3	1540	0	2807
	2.5	4646	4504	0	142	0	3852	0	58	1	3793	5084	6	785	2	4291
	1	4006	3893	0	113	0	3241	2	27	0	3212	4585	11	816	0	3758

Reads supporting small indels have not been included in the total read depth.

Table A.7 Illumina calibration data - expected and observed allele frequencies

			DNA Amount			
			50ng	25ng	10ng	5ng
Dilution	Mutation	EAF[a] (%)	OAF[b](%)	OAF(%)	OAF(%)	OAF(%)
10:1	H875Y	90.0	53.5	46.4	72.8	40.4
	T878A	10.0	13.8	25.9	8.8	0.5
	F877L	5.0	9.0	11.9	2.0	1.0
20:1	H875Y	95.0	64.2	60.2	47.6	38.1
	T878A	5.0	6.9	7.4	27.8	1.1
	F877L	2.5	4.4	4.0	0.8	0.00
50:1	H875Y	98.0	61.8	54.1	49.5	65.0
	T878A	2.0	4.0	2.0	0.2	0.3
	F877L	1.0	1.3	1.6	0.9	1.8

[a] EAF: Expected Allele Frequency. [b] OAF: Observed Allele Frequency.

Table A.8 Illumina Calibration Experiment - Read Statistics for different variants.

		T878A(chrX:66943552)					F877L (chrX:66943549)					H875Y (chrX:66943543)				
Dilution	DNA (ng)	Read	A	C	G	T	Read	A	C	G	T	Read	A	C	G	T
10:1	50	921	796	0	125	0	941	2	85	2	852	1535	4	710	0	821
	25	908	673	0	235	0	926	0	110	2	814	1487	0	794	0	690
	10	991	904	0	87	0	1015	0	20	0	995	1270	0	345	1	924
	5	387	383	0	2	2	402	0	4	0	398	993	2	590	0	401
20:1	50	1496	1392	0	100	4	1519	0	67	0	1452	2222	5	786	0	1427
	25	1070	989	0	79	2	1084	2	43	0	1039	1665	4	652	0	1003
	10	521	376	0	145	0	532	2	4	0	526	824	1	431	0	392
	5	363	359	0	0	4	375	0	0	0	375	992	1	613	0	378
50:1	50	1542	1472	4	62	4	1561	8	20	0	1533	2432	7	922	0	1503
	25	1330	1300	2	26	2	1368	1	22	0	1345	2484	2	1139	0	1343
	10	442	441	0	0	1	462	4	0	0	458	925	0	465	2	458
	5	645	643	0	2	0	654	2	12	0	640	994	2	346	0	646

Reads supporting small indels have not been included in the total read depth.

Table A.9 Validated[a] somatic mutations detected by SiNVICT in cfDNA samples (AmpliSeq sequencing data) at progression on enzalutamide

Position	Ref	Alt	Function	Gene	Sample	OAF(%)	Coverage
chr3:41266097	G	A	exonic	CTNNB1	enza-proton-VC-022-progression	20.0	17592
chr3:41266101	C	A	exonic	CTNNB1	enza-proton-VC-063-progression	24.5	14986
chr3:41266113	C	A	exonic	CTNNB1	enza-proton-VC-017-progression	15.1	20153
chr3:41266137	C	T	exonic	CTNNB1	enza-proton-VC-017-progression	19.3	15932
chr3:178936091	G	A	exonic	PIK3CA	enza-proton-VC-019-progression	14.1	14966
chr10:89711902	T	G	exonic	PTEN	enza-proton-VC-057-progression	7.7	15418
chr14:38061317	G	T	exonic	FOXA1	enza-proton-VC-056-progression	24.1	6131
chr17:7576572	A	C	exonic	TP53	enza-proton-VC-019-progression	33.4	10250
chr17:7577544	A	T	exonic	TP53	enza-proton-VC-093-progression	35.9	13234
chr17:7578521	G	-	exonic	TP53	enza-proton-VC-085-progression	44.1	7919
chr20:31024102	C	T	exonic	ASXL1	enza-proton-VC-085-progression	18.2	9191
chrX:66931463	T	A	exonic	AR	enza-proton-VC-008-progression	52.4	24859
chrX:66931463	T	A	exonic	AR	enza-proton-VC-017-progression	44.4	8754
chrX:66931463	T	A	exonic	AR	enza-proton-VC-022-progression	21.8	19972
chrX:66931463	T	A	exonic	AR	enza-proton-VC-063-progression	28.3	17005
chrX:66943543	C	T	exonic	AR	enza-proton-VC-081-progression	6.5	4310

[a] By Vancouver Prostate Centre (VPC).

The purpose of this experiment was not to test the limits of mutation detection on cfDNA like the previous cell-line experiments, but rather prove that cfDNA sequencing is a promising technology for clinical use. Our aim with presenting these results is to demonstrate that SiNVICT does indeed work on actual cfDNA as well and not only on simulated samples and cell-lines. Note: Our coverage and OAF statistics might slightly differ from the VPC validation results due to different bioinformatics software used for mapping, trimming, etc. before running SiNVICT on the data.

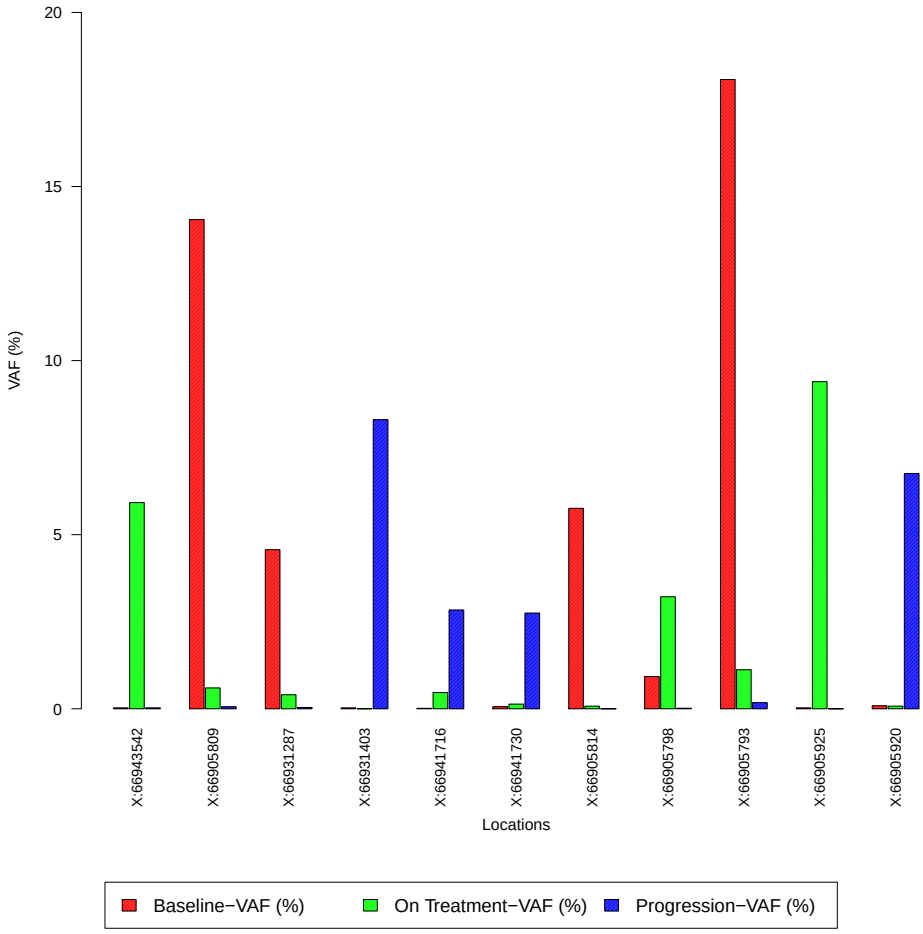

Figure A.1 The y-axis shows the variant allele frequencies for the locations given on the x-axis for patient VC-007 from the dataset described in Section 3.3 at the three time points shown in the legend. We observed a trend in which one time point shows an increase in the VAF despite the other two time points showing little evidence of a variant being present, 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.

Table A.10 Time-Series Analysis for VC-007

		Baseline				On Treatment				Progression			
Position	Ref. Base	Depth	VAF[a] (%)	p-orig[b]	p-new[c]	Depth	VAF (%)	p-orig	p-new	Depth	VAF (%)	p-orig	p-new
X:66943542	G	8004	0.024	0.99	0.01489	2736	5.921	10^{-9}	0.00018	8006	0.024	0.99	0.01489
X:66905809	A	8007	14.050	10^{-9}	0.00046	1329	0.598	0.95553	0.00124	6795	0.058	0.99	0.01044
X:66931287	A	8011	4.568	10^{-9}	0.00018	1236	0.402	0.99434	0.01039	8003	0.037	0.99	0.03130
X:66931403	A	8005	0.024	0.99	0.01636	1280	0.0	0.99	1.0	8009	8.303	10^{-9}	$9.78 * 10^{-5}$
X:66941716	G	8007	0.012	0.99	0.05068	1499	0.466	0.99231	0.00456	8007	2.835	10^{-9}	0.00072
X:66941730	T	8007	0.062	0.99	0.00558	1499	0.133	0.99	0.00697	8007	2.747	10^{-9}	0.00039
X:66905814	A	8007	5.757	10^{-9}	0.00258	1336	0.074	0.99	0.01390	6795	0.0	0.99	1.0
X:66905798	G	8004	0.924	0.76497	0.02081	1337	3.216	$9.81 * 10^{-11}$	0.00090	6795	0.014	0.99	0.10425
X:66905793	A	8006	18.073	10^{-9}	0.00012	1321	1.121	0.36311	0.00386	6791	0.176	0.99	0.00516
X:66905925	T	8013	0.024	0.99	0.07371	1341	9.395	10^{-9}	0.00011	7812	0.0	0.99	1.0
X:66905920	C	8013	0.087	0.99	0.01869	1341	0.074	0.99	0.12334	7812	6.758	10^{-9}	0.00021

[a] VAF: Variant Allele Frequency. [b] p-orig: Original p-value assigned to the location by SiNVICT (see equation 1) prior to time-series analysis.
[c] p-new: New p-value assigned to the location after doing the time-series analysis as described in Section 2.4.
These locations were selected from patient VC-007 sequenced at all three time points of interest - baseline, on-treatment (12-weeks), and progression - from the dataset described in Section 3.3. We then plotted the variant allele frequencies for these locations for this 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 The recalculated p-values were significantly different than the original p-values 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.

Table A.11 SiNVICT - Effect of filters on the number of calls.

	Experiment Dataset			
	AmpliSeq Calibration	Illumina Calibration	Simulation (No Het.)	Simulation (With Het.)
	Number of Locations (bp)			
Total Size of the Panel (Before any calls)	103036	92857	8938	31485
After Poisson CDF	101193	1907	2223	10084
After Min. Read Depth Filter	98420	1874	2223	10084
After Strand Bias Filter	7168	959	21	23
After SNR Filter	90	635	21	3
After Homopolymer Regions Filter	61	506	16	3

We have performed the initial SNV and indel calling with SiNVICT using Poisson CDF with more relaxed settings (confidence score threshold Q = 20) to maximize sensitivity, which results in a large number of locations called. This may be avoided by using a much higher confidence score threshold (like Q = 95), but this is not recommended since our layered filtering approach reduces this number in the later stages and the initially large number of calls can always be returned to later for safety checks. To reduce the number of initial calls, SiNVICT then used our filters to discard potential false positives. It can be seen that after applying each of our filters, the size of the set of locations called as possible mutations has been reduced to a feasible number, which can then be manually curated by a specialist. Note-1: Homopolymer Region Filter is recommended to be only used with AmpliSeq, not Illumina or Illumina based simulations (our simulations were based on Illumina reads, generated by the program wgsim). The HRF results for all experiments were kept as is, but they were not necessary except for the AmpliSeq experiment. Note-2: The strand-bias filtering can usually be more conservative for AmpliSeq (Ion Torrent) technology as mentioned in the Methods section of the text. However, due to the level of noise in our calibration experiment, SiNVICT filters a larger number of locations than normally expected.