SOUND
Statistical Multi-Omics Understanding

Project Deliverable

Project acronym: **SOUND**	GA number: **633974**
Project title: **Statistical Multi-Omics Understanding of Patient Data**	
Funding Scheme: Collaborative Project (H2020-PHC-2014-2015/H2020-PHC-2014-two-stage) Health, novel medical developments	
Project start date: **01 September 2015**	Duration: **36 months**
Project's coordinator: Dr Wolfgang Huber (European Molecular Biology Laboratory, Heidelberg)	

D2.2 Technical report on existing and novel methods

Due date of deliverable: Month 36 - 30.08.2018
Actual submission date: 23.07.2018

Organization name of lead contractor for this deliverable: University of Cambridge (UCAM)
Organization name of other involved partners: UCAM
Personnel involved: Meltem Gürel, Daniele Biasci

Project co-funded by the European Commission within the H2020 Program (2015-2018)		
Dissemination Level		
PU	Public	X
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Technical report: assessment of different methods to identify somatic mutations in ctDNA

Meltem Gürel[1] and Daniele Biasci[1]

[1]Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge CB2 0RE, UK

July 28, 2018

1 Introduction

Among the three main disease groups (circulatory diseases, respiratory diseases, and cancer), cancer was the most common in 2013 for both women and men in the UK: 352,197 new cases were reported in the UK in 2013 [1]. As a leading cause of death globally, cancer urgently requires pertinent diagnosis and treatment methods. Information about patients' genomic data are already used to tailor therapy and to determine which patients are most likely to respond to certain drugs [2]. However, genotyping tumors is accompanied by certain challenges. For instance, during therapy most tumors acquire resistance in time due to intratumoral heterogeneity and molecular evolution. Thus primary tumors and metastases need to be monitored throughout treatment to determine drug efficacy. Removing a piece of tissue, or a sample of cells, from the tumor site is a well-established method of tumor monitoring. However relying on tissue biopsies comes with certain limitations. Performing these "solid biopsies", given the complexities of tumor heterogeneity, both within a tumor and between a primary tumor and metastases, may not give an accurate representation of the tumor's molecular profile [3]. Moreover, during progression of the disease, tumors evolve and sub-clones develop, causing differences between the primary tumor and metastases or recurrences. Due to these genetic alterations, biopsies must be taken at multiple time points from multiple sites. These procedures are invasive and both risky and painful for the patient. For these reasons, minimally invasive methods that can monitor cancer in real-time are direly needed. Liquid biopsies that detect disease biomarkers in bodily fluids have the potential to eliminate the need for tumor tissue samples.

Current methods focus on two main approaches: analysis of circulating tumor cells (CTCs) [4], and analysis of circulating tumor DNA (ctDNA) [5]. The ratio of ctDNA to cell free DNA is greater than that of CTCs to circulating normal cells [6]. Also ctDNA is more accessible and easier to process [7]. Faster and cheaper high-throughput sequencing of cancer genomes has produced a large database of somatic mutations occurring in various tumor types. It is now necessary to monitor the presence of these genetic alterations in cancer patients through accurate and minimally invasive methods. ctDNA analysis fulfills these criteria, but it has yet to be used routinely in clinical settings. Robust open-source statistical tools for the interpretation of ctDNA data are lacking, causing a major bottleneck in the applicability of the technology. Statistics play a huge role in the improvement of evidence-based medicine and is reciprocally advanced through challenges brought forward by progresses in the medical field. Current techniques that use ctDNA data to investigate the longitudinal changes in the genomic profile of different cancer types rely heavily on elaborate lab protocols [3, 8]. Methods that can shift the burden to the dry lab, hence enable the clinical application of these techniques, are required to make liquid biopsies more routinely applicable.

In this report we provide a uniform technical assessment of different computational methods aimed to identify somatic mutations in ctDNA data.

2 Methods

For this report, we used *VarBench* [9], an open-source software for the standardized application of liquid biopsy data analysis we previously developed, to assess the performance of different methods in identifying and quantifying the frequency of somatic mutations in ctDNA. In particular, we leveraged on *VarBench::compare*, a software pipeline designed to assess performance of different ctDNA analysis methods [10].

3 Results

Methods included in this report

We assessed the performance of the following methods against the same set of somatic mutations introduced in real ctDNA data at defined allele frequencies: VarScan [11], SomaticSniper [12], VarDict [13], Strelka [14] and deepSNV [15].

	precision	**recall**	F_1**-score**
Strelka	0.79600	0.252	0.3790
deepSNV	0.20400	0.507	0.2900
SomaticSniper	0.04180	0.361	0.0748
VarScan	0.01710	0.178	0.0432
VarDict	0.00732	0.659	0.0145

Table 1 – Performance metrics of assessed variant callers ordered by F_1-score, descending.

Precision, recall, and F_1-score

Looking at Table 1, which shows the average *precision*, *recall*, and F_1-*score* for each variant caller, we see that both Strelka and deepSNV could be preferable variant callers for the ctDNA data at hand. While Strelka has higher precision, deepSNV has higher sensitivity. The other variant callers, especially VarDict, seem to call too many false positives, thus decreasing their precision performance greatly. However, VarDict has the highest recall meaning that, out of all the mutations present in our ctDNA data, it called the most. It is not uncommon to use multiple variant callers [16]; we could call mutations with Strelka first, then with deepSNV or VarDict, and intersect the called variants to get a more accurate picture. Figures 1, 2, and 3 report the distributions of each performance metric.

Predicted and actual allele frequencies

Allele frequencies (AF) of somatic mutations occurring in ctDNA can potentially be used as a surrogate measure of tumor burden [17]. For this reason, we assessed the ability of different variant callers to correctly estimate the AF of the detected somatic mutations. We looked at correlation between actual and predicted AF to determine which variant callers provided the best AF estimates. Results are reported in Table 2.

	R^2
Strelka	0.943
VarDict	0.752
VarScan	0.707
deepSNV	0.514
SomaticSniper	0.514

Table 2 – R^2 of predicted versus actual AFs for the variant callers assessed in this report.

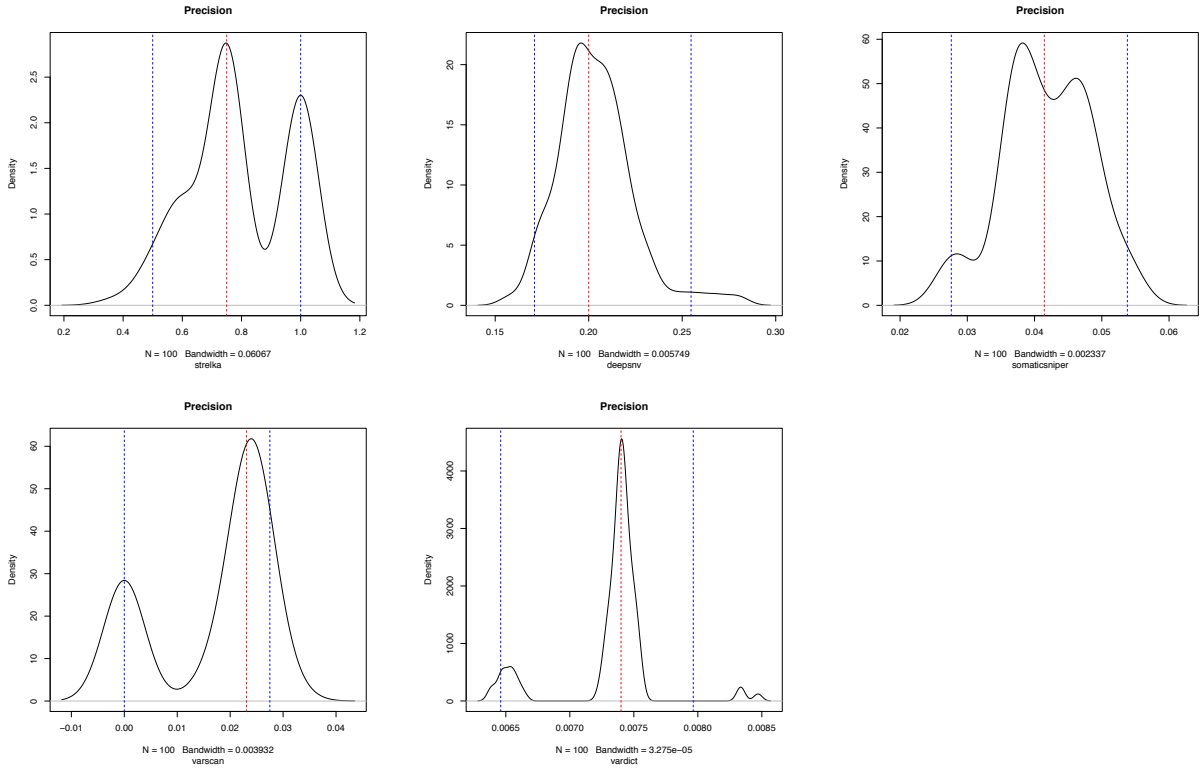

Figure 1 – Precision is the ratio of correctly predicted variants to the total predicted variants. Strelka has high precision which relates to a low false positive rate.

Table 2 shows that Strelka estimates are well correlated with the actual AF. However, this results should be considered together with the observation that Strelka achieves high detection precision at the expenses of recall (see Table 1). Although VarDict and VarScan showed worst performances in detecting somatic mutations, these callers perform better than deepSNV and SomaticSniper in estimating the correct AF of the detected mutations (see Table 2). Figure 4 shows the data points used to calculate the R^2 values reported in Table 2.

4 Conclusion

One of the mainstream approaches taken to manage cancer is early detection; accordingly researchers are trying to devise strategies to detect cancer from the very early stages [17]. Effective cancer tests need to be non-invasive for the patient and yield results for healthcare professionals in a reasonable amount of time. Liquid biopsies, which require a simple blood draw, could allow clinicians to screen patients for cancer in a minimally invasive way. Data obtained from ctDNA will

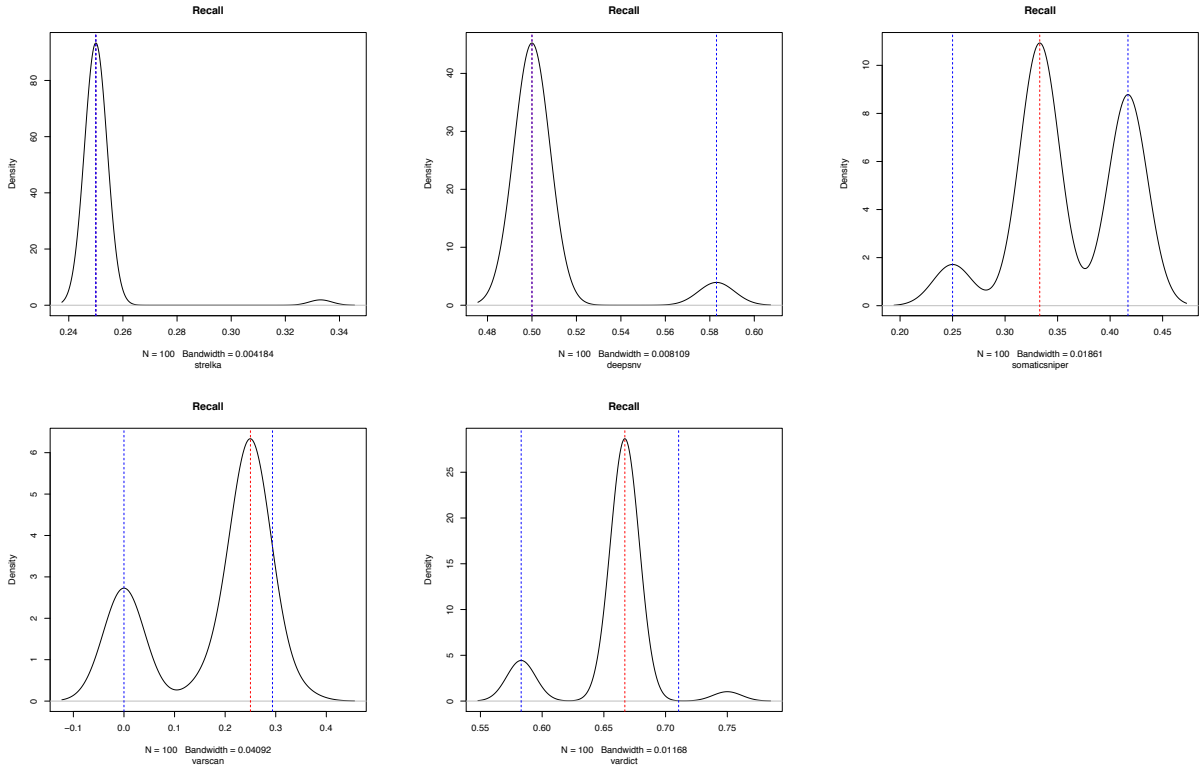

Figure 2 – Recall (or sensitivity) answers the question "Of all the present mutations, how many did the variant caller detect?". It is calculated with the below formula (1) where TP denotes true positives, and FN false negatives.

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

deepSNV and VarDict call mutations with good sensitivity.

yield faster and more accurate results as computational methods for ctDNA somatic variant calling increase in precision, recall and provide better AF estimates. Such methods will also improve tumor monitoring throughout treatment, as minimally invasive tests will reduce patients exposure to the risks of repeated tissue biopsies. To improve the likelihood of calling variants correctly, many ctDNA detection methods rely on targeted deep sequencing. Indeed, increasing the coverage is expected to reduce false positives. Other pre-data-generation techniques to reduce errors could be increasing sample size, and improving library preparation protocols. However, any pre-data-generation technique is intrinsically dependent on the particular platform or technology at hand. In order to maximize the utility of our results, we took a platform agnostic approach and we focused on post-data-generation strategies instead. With *VarBench*, we provided a streamlined way to benchmark different variant callers to select the best performing one for the ctDNA

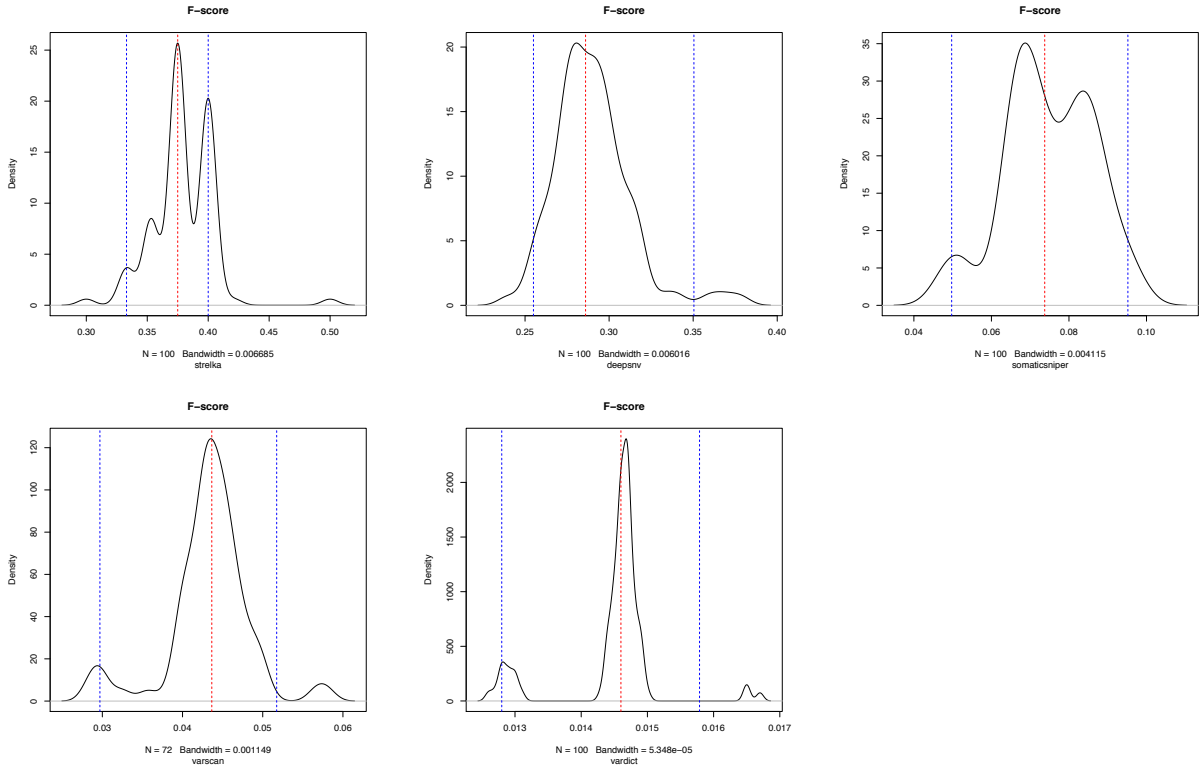

Figure 3 – F_1-score is the weighted average of precision and recall; if the cost of false positives and false negatives are not similar, it is better to look at both precision and recall, otherwise F_1-score can be a good overall performance measure.

data at hand [9, 10]. We also pointed out that reporting called variants and point estimates of the allele frequencies may not be sufficient, as clinical decisions could need to be taken on the basis of allele frequency changes over time (i.e., ctDNA data across multiple time points). Instead, for each variant, interval estimates of allele frequencies should be reported. In this report, we provided an example of such an approach when we calculated the R^2 between called and actual allele frequencies (see Figure 4). Ranking the variant callers using performance metrics precision, recall, and F_1-score, can allow the users to select the best variant caller for their purposes, and also possibly lead them to run multiple variant callers in parallel. With interval estimates of called-actual allele frequency R^2, users can decide the level of confidence in the call, thus providing a stronger rationale for downstream clinical decisions.

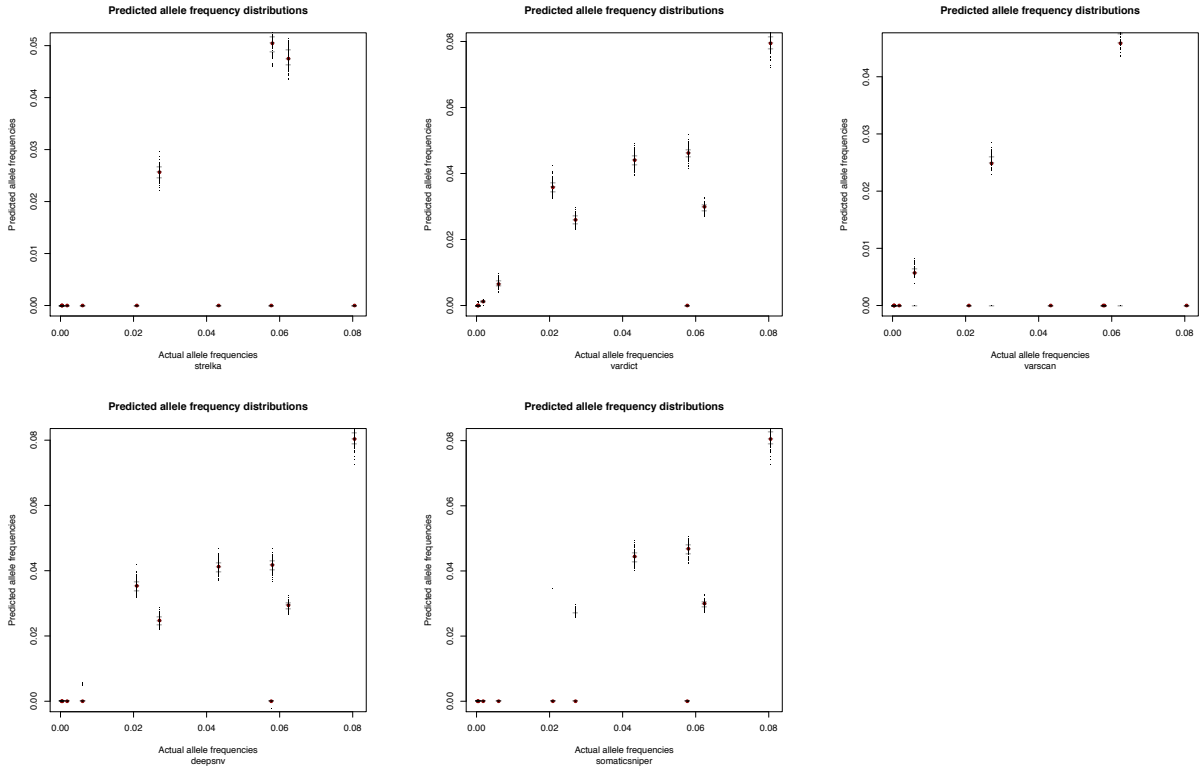

Figure 4 – Each plot reports the actual allele frequencies for selected somatic mutations (x axis) versus the allele frequencies estimated by each variant caller for the same mutations (y axis). Each point represents the results of running the selected variant caller on a bootstrap resample of the original dataset using the pipeline *VarBench::compare* [9, 10].

References

[1] Cancer Research UK. *Cancer Statistics for the UK*, 2016 (accessed May 14, 2016).

[2] Iain A McNeish, Amit M Oza, Robert L Coleman, Clare L Scott, Gottfried E Konecny, Anna Tinker, David M O'Malley, James Brenton, Rebecca Sophie Kristeleit, Katherine Bell-McGuinn, et al. Results of ariel2: A phase 2 trial to prospectively identify ovarian cancer patients likely to respond to rucaparib using tumor genetic analysis., 2015.

[3] Tim Forshew, Muhammed Murtaza, Christine Parkinson, Davina Gale, Dana WY Tsui, Fiona Kaper, Sarah-Jane Dawson, Anna M Piskorz, Mercedes Jimenez-Linan, David Bentley, et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma dna. *Science translational medicine*, 4(136):136ra68–136ra68, 2012.

[4] Catherine Alix-Panabières and Klaus Pantel. Circulating tumor cells: liquid biopsy of cancer. *Clinical chemistry*, pages clinchem–2012, 2012.

[5] SA Leon, B Shapiro, DM Sklaroff, and MJ Yaros. Free dna in the serum of cancer patients and the effect of therapy. *Cancer research*, 37(3):646–650, 1977.

[6] Chetan Bettegowda, Mark Sausen, Rebecca J Leary, Isaac Kinde, Yuxuan Wang, Nishant Agrawal, Bjarne R Bartlett, Hao Wang, Brandon Luber, Rhoda M Alani, et al. Detection of circulating tumor dna in early-and late-stage human malignancies. *Science translational medicine*, 6(224):224ra24–224ra24, 2014.

[7] Elizabeth A Punnoose, Siminder Atwal, Weiqun Liu, Rajiv Raja, Bernard M Fine, Brett GM Hughes, Rodney J Hicks, Garret M Hampton, Lukas C Amler, Andrea Pirzkall, et al. Evaluation of circulating tumor cells and circulating tumor dna in non–small cell lung cancer: Association with clinical endpoints in a phase ii clinical trial of pertuzumab and erlotinib. *Clinical Cancer Research*, 2012.

[8] Aaron M Newman, Scott V Bratman, Jacqueline To, Jacob F Wynne, Neville CW Eclov, Leslie A Modlin, Chih Long Liu, Joel W Neal, Heather A Wakelee, Robert E Merritt, et al. An ultrasensitive method for quantitating circulating tumor dna with broad patient coverage. *Nature medicine*, 20(5):548, 2014.

[9] Meltem Gürel. *Varbench*, 2018 (accessed July 21, 2018).

[10] Daniele Biasci Meltem Gürel. *Open-source software for the standardized application of existing methods in liquid biopsy data analysis*, 2018 (accessed July 21, 2018).

[11] Daniel C Koboldt, Ken Chen, Todd Wylie, David E Larson, Michael D McLellan, Elaine R Mardis, George M Weinstock, Richard K Wilson, and Li Ding. Varscan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285, 2009.

[12] David E Larson, Christopher C Harris, Ken Chen, Daniel C Koboldt, Travis E Abbott, David J Dooling, Timothy J Ley, Elaine R Mardis, Richard K Wilson, and Li Ding. Somaticsniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, 2011.

[13] Zhongwu Lai, Aleksandra Markovets, Miika Ahdesmaki, Brad Chapman, Oliver Hofmann, Robert McEwen, Justin Johnson, Brian Dougherty, J Carl Barrett, and Jonathan R Dry. Vardict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic acids research*, 44(11):e108–e108, 2016.

[14] Christopher T Saunders, Wendy SW Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 2012.

[15] Moritz Gerstung, Christian Beisel, Markus Rechsteiner, Peter Wild, Peter Schraml, Holger Moch, and Niko Beerenwinkel. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature communications*, 3:811, 2012.

[16] Maurizio Callari, Stephen-John Sammut, Leticia De Mattos-Arruda, Alejandra Bruna, Oscar M Rueda, Suet-Feung Chin, and Carlos Caldas. Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. *Genome medicine*, 9(1):35, 2017.

[17] Jonathan CM Wan, Charles Massie, Javier Garcia-Corbacho, Florent Mouliere, James D Brenton, Carlos Caldas, Simon Pacey, Richard Baird, and Nitzan Rosenfeld. Liquid biopsies come of age: towards implementation of circulating tumour dna. *Nature Reviews Cancer*, 17(4):223, 2017.