Comparative analysis of somatic variant calling on matched FF and FFPE WGS from a metastatic prostate sample

Louise de Schaetzen van Brienen
Universiteit Gent https://orcid.org/0000-0002-5882-9458
Maarten Larmuseau
Universiteit Gent - imec
Kim Van der Eecken
Universitair Ziekenhuis Gent
Jan Fostier
Universiteit Gent - imec
Piet Ost
Universitair Ziekenhuis Gent
Kathleen Marchal (kathleen.marchal@ugent.be)
https://orcid.org/0000-0002-2169-4588

Research article

Keywords:

Posted Date: October 10th, 2019

DOI: https://doi.org/10.21203/rs.2.15860/v1

Version of Record: A version of this preprint was published on July 6th, 2020. See the published version at https://doi.org/10.1186/s12920-020-00746-5.

Abstract

Background. Research grade Fresh Frozen (FF) DNA material is not yet routinely collected in clinical practice. Many hospitals, however, do collect and store Formalin Fixed Paraffin Embedded (FFPE) tumor samples. Consequently, the sample size of whole genome cancer cohort studies could be increased tremendously by including FFPE samples, although the presence of artifacts might obfuscate the variant calling. To assess whether FFPE material can be used for cohort studies, we performed an in-depth comparison of somatic SNVs called on matching FF and FFPE Whole Genome Sequence (WGS) samples extracted from the same prostate metastatic tumor.

Results. We first compared the calls between FF and FFPE, showing that on average 50% of the calls in FF are recovered in FFPE, with notable differences between variant callers. Remarkably, this overlap was better than the overlap between different variant callers on the same sample. Inspecting the Variant Allele Frequency (VAF), we observed that many of the calls common to FF and FFPE belonged to the same clonal subpopulation but were detected at a lower VAF in FFPE. We also demonstrated that these calls receive higher significance scores and are often identified by more than one variant caller. Based on this observation, we propose a simple heuristic to perform reliable variant calling in FFPE samples. Our heuristic identified 3684 common calls at a F1-score of 0.83.

Conclusion. This study illustrates that when using the correct variant calling strategy, the overlap between the FF and FFPE sample in somatic SNVs increases to such an extent that a large fraction of the calls detected in the FFPE sample are contained in the FF sample and the number of variants unique to each sample remains restricted. These results suggest that somatic variants derived from WGS of FFPE material can be used in cohort studies.

Introduction

Cohort analysis in which comprehensive genomic data of large patients cohorts are being coupled with clinical information offers a vast potential for precision oncology. So far, most large cohort studies relied on whole exome sequencing (WES) or WGS of FF tumor material. The preservation of and access to FF tissues can be limited. Indeed, in routine clinical practice, FF samples are rarely available due to logistic reasons: they are difficult to collect, prepare and are expensive to store. Optimally exploiting available patients cohorts would therefore require collecting sequence information from FFPE samples which are collected in routine standard of care for histopathological diagnosis. This poses a problem as DNA extracted from FFPE specimens presents degradation such as nucleic acid fragmentation, DNA crosslinks, abasic sites leading to localized DNA denaturation, strand breaks, and deamination leading to C>T mutations (1–3).

Several studies have established that sufficiently high-quality DNA can be derived from FFPE material. Although the processing and storage affect the quality of the DNA and subsequent next generation sequencing (NGS) data (4–7), for most samples enough qualitative DNA can be collected to perform NGS

assays in order to identify copy number variations (CNVs) and single nucleotide variations (SNVs) (8–10). Most studies that compared somatic variants, obtained from sequencing matched FF and FFPE samples, are based on WES. Depending on the study (11–13) an overlap between 54% and 90% was found in somatic variants obtained from matched FFPE and FF samples. Differences in results can be attributed to differences in studied cancer types (which might differ in intra-tumor heterogeneity) and the fact that different variant callers and quality thresholds were used.

In this work, we performed an in-depth comparison of the degree to which somatic variants can be called using WGS of an FF and FFPE sample extracted from the same prostate metastatic tumor. In line with what was shown by Robbe *et al.* (14), we found that FFPE material can be a good proxy for an FF sample and that a recent FFPE sample does not generate too many artifacts, at least not at the SNV level. However, in contrast to previous studies, we show that carefully tuning and combining the results of different variant callers can increase the overlap in variants detected between the FF and the FFPE sample. This to such extent that almost all variants that were detected reliably in the FF sample could also be recovered from the FFPE sample (high sensitivity) and that in addition the majority of variants detected in the FFPE sample were also contained in the ones detected by the FF sample (high precision) and this despite the lower effective coverage in the FFPE sample. Acquiring this combination of a high precision but also high sensitivity shows that an FPPE sample is a useful proxy for an FF sample to comprehensively call somatic variants.

Results And Discussion

We aimed at testing to what extent WGS of FF- and FFPE-derived material results in the identification of the same somatic variants (SNVs). Hereto, we used one FF and one FFPE sample of the same metastatic prostate tumor. A matching blood sample, taken from the same patient, was used as a reference to identify germline mutations. We opted for a metastatic sample as this is in general more homogenous than primary samples, reducing differences in variant calls due to intra-tumor and sampling variation.

Quality checking showed that in the matching FF and FFPE samples a comparable number of variants was called (*Suppl. Table 1).*. No particular bias towards specific mutational patterns was observed, indicating the absence of prominent artifacts in the FFPE sample (*Suppl. Table 1).*. Both FF and FFPE samples showed a similar average coverage (*Suppl. Table 2).*. To ensure that comparing the somatic variants called from the FFPE and matching FF samples would be independent of the specificities of the variant caller, we used four somatic variant callers, i.e. Strelka2, Mutect2, VarScan2 and Shimmer.

Comparison between FF and FFPE sample for each caller

All variant callers except VarScan2 were run under default settings (*see Materials and Methods).*. In general, more calls were reported in the FFPE than in the FF sample, except for VarScan2 (*Suppl. Fig. 1).*. *Table 1* shows the overlap between somatic calls in the matching FF and FFPE samples in terms of sensitivity and precision using the calls from the same variant caller on the FF sample as gold standard.

Overall, variant callers have an average sensitivity and precision of respectively 50.97% and 52.41%, meaning that 50.97% of the variants from the FF sample are also detected in the FFPE sample and 52.41% of the FFPE somatic variants are detected in the FF sample. We also report the F1-score which is the harmonic mean of sensitivity and precision. Among the four variant callers considered in this study, Strelka2 achieved both the highest sensitivity and precision.

To get an idea of how good the sensitivity and precision from *Table 1* are, we compared the calls from different variant callers on the same sample. In general, the difference between variant callers on the same sample is larger than the difference between the FF and the FFPE sample (*Suppl. Fig. 1* and *2)*.. This implies that the choice of variant caller is at least as important as whether or not the sample was FF or FFPE. Overall, the overlap between variant callers is low, especially for the FFPE sample (*Suppl. Fig. 2)*.. Shimmer agrees the least with the other variant callers, whereas Strelka2 and Mutect2 tend to mutually agree in both the FF and FFPE samples (*Suppl. Table 3* and *4)*..

While the choice of variant caller turns out to play a pivotal role in the calls that are obtained, the basic analysis above does not account for the properties of the called variants. Indeed, we will demonstrate that the variants considered in *Table 1* represent both true variants and artifacts of the variant caller, consistently made in the FF and the FFPE sample. To assess the relevance of the calls made by each of the callers in the overlap between the FF and the FFPE sample, we first assessed to what extent each of the callers tends to reconstruct the same clonal subpopulations in both the FF and the FFPE sample and secondly whether high confidence calls are consistent between both sample types.

Comparison of the clonal subpopulation

Given that the FF and the FFPE sample should have a similar subclonal structure, as they derive from the same metastatic tumor, we judged the relevance of the variants called by either method by assessing whether they corresponded in the FF and the FFPE to the same clonal subpopulations. To visualize the different subpopulations in each sample, we plotted for all detected variants thecoverage as a function of the VAF and the corresponding histogram representing the distribution of the VAFs (15). *Fig. 1* shows the results for Strelka2 in both the FF and the FFPE sample. Similar figures for the three other variant callers can be found in the *Supplementary Materials.* Note that the VAF estimate of the major distributions slightly differs between the different tools because of small discrepancies in their definition of the VAF (*Suppl. Fig. 3)*..

According to Strelka2 (*Fig. 1)*,, two distinct clusters of variants can be observed in the FF sample, denoted by the blue bars in the histogram: a small cluster at a VAF of 0.05 and a larger cluster around a VAF of 0.25. In the FFPE sample, only one cluster can be observed around a VAF of 0.15. The calls that are common to the FF and the FFPE sample are shown in orange, demonstrating that the peak at a VAF of 0.25 in FF does indeed shift to 0.15 in FFPE. Almost all variants from the largest cluster in the FF sample, i.e. at a VAF of 0.25, are identified in the FFPE sample but most of the FF calls belonging to the smaller peak at 0.05 are not detectable in the FFPE sample.

We observed similar results for Mutect2 with the peak representing the major cluster at VAF 0.25 in the FF sample being shifted to a lower VAF in the FFPE sample (*Suppl. Fig. 4).*. For VarScan2, only a small proportion of the variants detected at an average VAF of 0.25 in the FF is recovered in the FFPE sample (*Suppl. Fig 5).*. Shimmer could barely detect the cluster of variants at VAF 0.25 in the FF and completely misses the major cluster in the FFPE sample (*Suppl. Fig. 6).*. A deeper analysis shows that Shimmer reports many somatic variants in the tumor sample that have a VAF above zero in the normal sample (putative germline calls). This behavior is not expected and also not observed for any of the other variant callers, explaining the poor overlap between Shimmer and the other variant callers (*Suppl. Fig 6, Suppl. Table 3* and *4).*. Only keeping for Shimmer the variants with zero VAF in the normal sample filtered those putative germline calls and allows to better recover a cluster of variants at VAF 0.25 in the FF, but still not in the FFPE sample (*Suppl. Fig. 7).*.

In addition to the cluster at VAF 0.25, VarScan2 also reports a peak of variants at an average VAF of 0.5 in the FF sample. These are likely germline variants as more than 65% of variants with VAF above 0.5 are present in dbSNP database (*Suppl. Table 5).*. Except for Mutect2, which has a similar number of germline calls above and below a VAF of 0.5, other callers reported up to 8 times more calls from dbSNP at a VAF above 0.5. This suggests that calls reported with VAF above 0.5 are more likely to be false positives. Indeed, most of the calls with VAF above 0.5 detected consistently in both the FF and the FFPE sample tend to be residual germline calls.

Hence, the cluster of variants detected at VAF 0.25 likely *represents the major subpopulation in the metastatic sample.* All variant callers except Shimmer could at least partially recover this subpopulation in the FFPE sample albeit at lower VAF. This shows that the lower effective VAF in the FFPE sample is independent of the variant caller and thus an intrinsic property of the FFPE sample. As a result, calls common to the FF and the FFPE sample were typically reported with lower VAF in the FFPE sample (*Suppl. Fig. 8).*. Imposing a zero VAF criterion in the normal sample helped Shimmer to recover the major cluster of variants that was also recovered by the other callers, but only in the FF sample. In addition, only a minority of calls (813) are retained of which most (604) were also reported by other callers (*Suppl. Fig. 12).*. This indicates that many of the calls uniquely made by Shimmer without imposing the zero VAF criteria and reported in *Table 1* are likely spurious calls, despite being consistently detected in both the FF and the FFPE sample. For the remainder of the analysis, variants reported by Shimmer with a positive VAF in the normal sample were filtered.

Assessing the significance levels in FFPE

To assess the relevance of the calls in the overlap between the FF and the FFPE sample, we also compared for each caller the distributions of the significance scores of the somatic variant calls detected in the FFPE sample that were also called in the FF sample, hereby assuming that the FF sample is less prone to artifacts and constitutes the reference as to what should be detected. *Table 2* shows that calls

reported in both samples typically have lower significance scores in the FFPE than in the FF sample, which can be explained by the discrepancies in VAF observed in *Fig. 1.*

In addition, the boxplots in *Fig. 2* show that the FFPE calls that were also made in the FF sample, received a relatively higher significance score than FFPE calls not made in the FF sample. This shows that the most reliable variants in the FFPE sample generally correspond to those detected in the FF sample.

For each caller, we calculated the correlation between the significance levels for common calls and compared it in FF versus FFPE. *Suppl. Table 6* shows that the significance levels of Strelka2 and Mutect2 are significantly correlated in both FF and FFPE samples. In addition, the significance levels of between the callers themselves seems to be consistent between the FF and the FFPE sample. For Shimmer and VarScan2, this consistency between both samples cannot be observed. Furthermore, the ranking of these callers is consistent in the FF, but not in the FFPE sample, again indicating that these variant callers are not performing well in the FFPE sample. A possible explanation for this could be that the underlying hypergeometric testing procedure cannot cope with the lower VAFs present in this sample.

Relation between the subclonal structure and the significance level

To investigate the relation between the major subpopulation and the significance of the calls, we map for each variant caller the 25% highest confidence calls on the Coverage versus VAF plots. The upper panel of *Fig. 3* shows how for Strelka2 these most significant calls are located around a VAF of 0.25 in the FF and 0.15 in the FFPE sample, and hence make up the aforementioned major subpopulation that was detected in both the FF and the FFPE sample.

For Mutect2, the majority of significant calls also belonged to this cluster of variants (*Suppl. Fig. 9).*. For VarScan2 and Shimmer (*Suppl. Fig. 10* to *12),*, a similar effect could be observed, although not as pronounced as in Strelka2 and Mutect2. Subsequently, many of the highly significant calls belong to the major subpopulation at a VAF of 0.25 in FF and 0.15 in FFPE, such that the analysis of the significance levels and the clonal subpopulations points at the existence of a highly confident subset of variants, present in both the FF and the FFPE sample.

In addition, the lower panel in *Fig. 3* shows that when comparing the variants obtained from different callers, many variants that are called by other callers are also located in the variant cluster (see also the lower panels of *Suppl. Fig. 9–12).*. This is in line with the observation that for each variant caller (except Shimmer), calls made by any of the other three variant callers in general received a higher significance especially in the FF sample (*Suppl. Fig. 13* and *14).*. Using two criteria, based on clonal subpopulation (*Fig. 1)* and significance score of the variants (*Fig. 2),*, we could see that calls common to the FF and the FFPE sample tend to belong to the major subpopulation and are highly significant. Importantly, the lower panel of *Fig. 3* and *Suppl. Fig. 13* and *14* show that variants that are called by more than one caller tend to satisfy these two criteria as well. In the next section, we investigated how well these calls, reported by more than one variant caller, overlap with the calls common to the FF and the FFPE sample.

Sensitive and precise variant calling on FF, FFPE threshold optimization

Previous analysis also showed that at least some of the calls that were made consistently between the FF and the FFPE sample by the same caller tend to be spurious or at least non-somatic. To assure that only biologically relevant calls are considered, we first identify a highly reliable subset of calls in the FF sample. This subset, referred to as the ground truth, can then be used to more qualitatively assess how well each caller can recover these highly reliable calls in the FFPE sample. From our analysis above, an ideal ground truth would consist of all highly reliable calls made in the reference FF sample that also represent the major subpopulation in the metastatic sample. We have shown that calls made by at least two callers tend to satisfy these criteria. Therefore, the ground truth was defined as the union of all calls that were detected by at least two callers in the FF sample.

Our previous analysis also shows that the subpopulation represented by the ground truth is also present in the FFPE sample albeit at lower effective coverage. Hence, fully recovering the subpopulation of highly significant variants from the FFPE sample will require the identification of a significance threshold in the FFPE sample. Because calls in the FF sample were reported with a higher significance score, the threshold in the FFPE will typically be lower than the threshold that would be necessary in the FF sample to capture the ground truth (*Suppl. Fig. 13* and *14)*.. In addition, previous analysis shows that it is feasible to set a threshold in the FFPE sample as calls common to the FF and the FFPE sample tend to have a higher significance level in the FFPE sample (see *Fig. 2)*.. However, the lower effective coverage in the FFPE sample complicates identifying a threshold that distinguishes the true calls from the noise (the significance distribution of noisy and true calls starts overlapping (see *Suppl. Fig. 14)*.. We determined for each variant caller a threshold in the FFPE sample that optimized the F1-score with the FF-derived ground truth (i.e. that optimizes the tradeoff between precision and sensitivity in recovering the ground truth).

Table 3 shows for the different variant callers their sensitivity and precision in recovering this ground truth. This table shows that Strelka2 and Mutect2 are performing the best in recovering from the FFPE sample the calls that belong to the ground truth. *Table 3* now quantitatively illustrates how Shimmer underperforms on the ground truth despite calling consistently the same mutations in the FF and the FFPE sample, which was shown in *Table 1.* This indicates, as was shown above, that most of the calls made by Shimmer in the overlap between the FF and the FFPE sample are likely spurious. The same is to some extent true for VarScan2 because of the high number of residual germline calls.

Table 3 gives an estimate of the expected overlap between the FF and the FFPE sample after optimizing the thresholds in the FFPE sample using the ground truth based on the variants detected in the FF sample. However, often only FFPE samples are available, such that the stringency thresholds cannot be optimized based on observations in the FF sample. Therefore, rather than optimizing the threshold for each caller separately, we assessed to what extent combining the output of different variant callers in the FPPE sample allows recovering the ground truth from the FF sample.

Table 4 shows how taking the intersection of the four variant callers maximizes the precision but comes at the expense of losing almost all sensitivity. As discussed above, Shimmer can barely retrieve the cluster representing the major subpopulation in the FFPE sample (*Suppl. Fig. 7)* and most of the calls retrieved by Shimmer in the FFPE sample are unique (even after correcting for the so-called somatic calls with high VAF in the normal sample). Because the intersection seems too strict and limits the sensitivity, we considered calls reported by at least three callers. It results in a high precision but still relatively low sensitivity in recovering the ground truth (*Table 4)* because VarScan2 reports less calls in the FFPE than in the FF sample and loses many true positive calls. Using the calls returned by at least two of the four callers in the FFPE sample drastically increases sensitivity while only slightly decreasing precision. The performance here is even better than the performance of the best caller (Strelka2) that was obtained after optimizing its stringency thresholds based on the ground truth. This indicates that there is definitely some complementarity in the calls made by different callers. Given the lower performances of VarScan2 and Shimmer in recovering the ground truth, one can wonder what their added value is when taking the union of all variants that were detected by at least two callers. To assess the added value of Shimmer, we compared the performances obtained when retaining the 'union of all variants obtained by at least two callers' where Shimmer was respectively included and discarded from the ensemble of used callers (*Table 5)*.. Omitting Shimmer from the ensemble resulted in a slightly higher F1-score while only 4 variants belonging to the ground truth could no longer be detected.

Along the same lines, we assessed the added value of VarScan2 by comparing performances of two ensemble consisting of Strelka2 and Mutect2 respectively with and without VarScan2 using again the criterion being called by at least two callers (*Table 5)*.. Compared to the situation without Shimmer, omitting VarScan2 resulted in a slight increase in precision but came at the expense of a slight decrease in sensitivity, 70 variants could no longer be detected by excluding VarScan2 from the ensemble strategy. Hence, including Varscan2 allows detecting more variants common to FF and FFPE samples, but comes with a few additional false positives.

These results show that the FFPE sample can recapitulate approximatively 80% of the highly reliable calls detected in the FF sample, while maintaining a precision of 0.87. Taking calls reported by at least two callers allows effectively distinguishing the true from the spurious calls in the FFPE sample, removing the residual germline variants obtained by VarScan2 and the spurious calls made by Shimmer. It gives an overlap with the ground truth that outcompetes the best variant caller in our hand (Strelka2, with threshold optimized on the ground truth) in terms of both precision and recall.

The relatively low fraction of variants that are unique to the FFPE sample (539/4219) shows that in our sample FFPE artifacts do not heavily bias the calls. The fraction of calls unique to the FF sample corresponds to the less significant calls in that sample (low VAF) and hence might, because of the lower effective coverage in the FFPE sample, no longer detectable (false negatives). Small discrepancies between the FF and the FFPE sample can of course also be ascribed to sampling and tumor heterogeneity (14).

Material And Methods

Patient and samples

For this study, a patient with isolated pulmonary recurrence of prostate cancer after initial definitive local therapy was selected for who we had both FFPE and FF samples from the solitary pulmonary metastasis. The prostatic origin of the lung metastatic adenocarcinoma was confirmed by pathological review (J. V. D., S. V. and K. V. D. E.). Microscopically, the pulmonary metastasis was composed of eosinophilic tumor cells with very large pleomorphic hyperchromatic nuclei and prominent nucleoli, which exhibited a cribriform pattern (*Suppl. Fig. 15 A)* with negative staining for CK7 and TTF–1 and positive PSA staining (*Suppl. Fig. 15 B, C* and *D);*; these findings were compatible with metastatic prostate cancer. Two samples from this pulmonary metastasis had been obtained, one had been stored as FFPE and one as FF. Whole blood was collected and informed consent was obtained at time of clinical follow-up.

Pathologic quality control (QC)

For both FF and FFPE samples, 5 µm-tick haematoxylin and eosin-stained slides were prepared and independently evaluated by two genitourinary pathologists (J. V. D. and S. V.) to determine the tumor cellularity. For FFPE tissue, eleven adjacent 5 µm-tick sections were prepared. The first ten sections were used for DNA extraction, whilst the last section served as reference to indicate a tumor-rich area suitable for macro-dissection (more than 70% tumor cellularity). Manual macro-dissection was performed using sterile scalpel blades. Information about input materials is displayed in *Table 6.*

Preparation steps and sequencing

The genomic DNA (gDNA) was extracted from the FFPE tissue using the proprietary method of Wuxi (NextCODE SeqPlus extraction protocol) and from the FF tissue with QIAamp DNA Mini Kit (Qiagen) according to the manufacturer's instructions. gDNA was extracted from a 200 µL EDTA-whole blood sample using the QIAamp® Blood Mini Kit (Qiagen) with QIAcube according to the manufacturer's instructions. The DNA samples were quantified with a Qubit 3.0 fluorescence spectrometer (Life Technologies, Waltham, MA USA) using a Qubit dsDNA BR assay kit. Covaris has been used for DNA shearing. TruSeq® Nano DNA Library Prep (Illumina) has been used for library construction. The Illumina sequencing platform HiSeqX PE150 has been used for WGS. The mean coverage was of 30X for the blood sample and 100X for the tumor samples.

Variant calling

Quality checking obtained by running GATK Picard tools (https://broadinstitute.github.io/picard/). For somatic variant calling, we used Strelka2 (16), Mutect2 (17), and Shimmer (18) with default settings.

VarScan2 (19) was run without imposing a minimal VAF threshold. To select the most reliable somatic calls, *FilterMutectCalls* with default parameters was applied on Mutect2, *somaticFilter* to the VarScan2 output without imposing the default threshold of minimal VAF and no additional somatic filters were applied to Strelka2 and Shimmer output. *Table 7* provides a summary of the main parameters of the four variant callers used.

The default settings used by VarScan2 impose a threshold for the minimum VAF at 0.2 (see *somaticFilter* in *VarScan2's Online Manual)*.. This prevents VarScan2 from detecting the low frequent variants in the FFPE sample. To recover also these variants and hence maximize the overlap with the other callers, we ran VarScan2 without the constraint on the minimal VAF threshold. Using this non-default setting resulted in VarScan2detecting more variants with VAF lower than 0.2 and on overall increased the overlap in somatic variants detected by Strelka2, Mutect2 and Shimmer on the same sample while also decreasing significantly the number of variants uniquely called by VarScan2 (*data not shown)*.. Although we expected intuitively that most calls obtained by VarScan2 without the VAF constraints would also be present in calls from VarScan2 with default parameters, this appeared not to be the case (and we could not find any reasonable explanation for this).

Identifying the stringency of the calls | For Strelka2 the stringency of the call was determined by the *Somatic EVS,* for Mutect2 by the *TLOD* scores, for VarScan2 by the *somatic p-values* and for Shimmer by *q-values.* The higher the scores were for Strelka2 and Mutect2, the more significant were the variants. For VarScan2 and Shimmer, the smaller values were the most significant.

Measure to evaluate the concordance between FF and FFPE sample |The overlap between reported calls in matching FF and FFPE samples is reported in terms of sensitivity and precision using the variants obtained on the FF sample as gold standard. A somatic variant was considered present in both samples if in both samples the variant was located at an identical chromosomal position, and if reference and alternative alleles were identical.

Identifying the threshold maximizing the overlap between FF and FFPE sample |For each variant caller, we searched for an optimal significance threshold in FFPE sample to obtain the largest concordance between samples (maximal F1-score). The screening space for the optimization of the significance threshold is displayed in *Table 8.*

Conclusions

In this work, we have investigated whether a metastatic FFPE sample, embedded with recent protocols and subjected to DNA extraction using specialized procedures, can be used as a proxy for an FF sample to call somatic variants for cohort analysis. In contrast to previous studies, which focused on comparing the extent to which a small fraction of the most reliable variants compares between an FF and an FFPE sample (the fraction enriched in drivers or actionable mutations) (14), cohort analysis requires that as many true somatic variants as possible are called (high sensitivity) so that subsequent statistical analysis over a cohort can identify driver variants. Because of the subsequent statistical analysis, cohort

analysis can tolerate some false positives and thus allows for a less stringent precision. Using four different variant callers (Strelka2, Mutect2, VarScan2 and Shimmer), we compared the somatic calls on the FFPE sample to its FF counterpart. At first sight, each variant caller recovered about 50% of the FF calls in the FFPE sample. Interestingly, we observed a larger discrepancy between variant callers on the same sample than between samples using the same variant caller. This implies that the choice of variant calling tool is at least as important as whether FF or FFPE material is being used.

Using Coverage vs. VAF plots on the FF sample, a clear subpopulation of calls was distinguishable and was enriched in highly significant calls, these were the calls we aimed to recover in the FFPE sample. However, while many of the calls detected in the FF were effectively present in the FFPE, they suffered from a lower effective VAF in the FFPE sample. This effect reduces the resolution of variant callers for the identification of low-VAF variants in the FFPE sample, reducing the overlap between the FF and FFPE samples. The effect was especially prominent in variant callers that rely on hypergeometric testing, i.e. VarScan2 and Shimmer. By choosing for each variant caller a threshold on the significance level of the identified variants, the overlap between FF and FFPE can be optimized in terms of sensitivity and precision. However, in many real-life situations, there is no matching FF sample available, and there is a need for a good strategy to perform a precise yet sensitive variant calling. Simply taking the intersection between the different callers, turned out to be too simplistic, as the low resolution of certain variant callers in the FFPE sample (in this case VarScan2 and Shimmer) obfuscated the final intersection. Indeed, for these two variants callers, the calls common to the FFPE sample and the FF gold standard are not assigned a more significant score. Nevertheless, when considering only calls reported by at least two variant callers (in our hands Strelka2, Mutect2, VarScan2 and Shimmer), we obtain almost 3700 calls present in both FF and FFPE, with an F1-score higher than 80%. Using the correct variant calling strategy, the overlap between the FF and FFPE sample in somatic SNVs increases to such an extent that a large fraction of the calls detected in the FFPE sample are contained in the FF sample and the number of variants unique to each sample remains restricted. Our results indicate that somatic SNVs derived from FFPE WGS samples can be used for cohort analysis, provided a careful variant calling strategy was used.

List Of Abbreviations

AD – allelic depth

CK7 – cytokeratin 7

CNV – copy number variation

DNA – deoxyribonucleic acid

DP – depth

FF – fresh frozen

FFPE – formaldehyde fixed-paraffin embedded

FWO – Fonds Wetenschappelijk Onderzoek-Vlaanderen

gDNA – genomic DNA

GHU – Ghent University Hospital

H&E – haematoxylin and eosin

IWT – Innovatie door Wetenschap en Technologie

PSA – prostate specific antigen

NGS – next generation sequencing

SNV – single nucleotide variation

TTF–1 – thyroid transcription factor 1

VAF – variant allele frequency

WES – whole exome sequencing

WGS – whole genome sequencing

Declarations

Ethic approval and consent to participate

The Ghent University Hospital (GHU) Ethics Committee (EC 2015/0260) approved this work.

Consent for publication

Not applicable

Availability of data and materials

Data used in this study is available on request.

Competing interests

The authors declare that they have no competing interests.

Funding

Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) [3G046318, G.0371.06]; Agentschap voor Innovatie door Wetenschap en Technologie (IWT) [NEMOA]; Katholieke Universiteit Leuven [PF/10/010] (NATAR). Funding for open access charge: Fonds Wetenschappelijk Onderzoek [G.0371.06].

Author's contribution

K. V. D. E. extracted and prepared samples for the analysis. L. D. S. V. B. performed variant calling analysis, interpreted the somatic calls and wrote the original manuscript. K. M. and M. L. were major contributor in writing the manuscript. JF helped to revise the manuscript. K. M. and P. O. supervised the project. All the authors read and approved the final manuscript.

References

1. Do H, Dobrovic A. Sequence artifacts in DNA from formalin-fixed tissues: Causes and strategies for minimization. Clin Chem. 2015;61(1):64–71.
2. Wong SQ, Li J, Tan AYC, Vedururu R, Pang JMB, Do H, et al. Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. BMC Med Genomics. 2014;7(1):1–10.
3. Haile S, Corbett RD, Bilobram S, Bye MH, Kirk H, Pandoh P, et al. Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples. Nucleic Acids Res. 2019;47(2):e12.
4. Beltran H, Tagawa ST, Nanus DM, Yelensky R, Frampton GM, Downing SR, et al. Targeted next-generation sequencing of advanced prostate cancer identifies potential therapeutic targets and disease heterogeneity. Eur Urol. 2013;63(5):920–6.
5. Hedegaard J, Thorsen K, Lund MK, Hein AMK, Hamilton-Dutoit SJ, Vang S, et al. Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. PLoS One. 2014;9(5).
6. Spencer DH, Sehn JK, Abel HJ, Watson MA, Pfeifer JD, Duncavage EJ. Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. J Mol Diagnostics [Internet]. 2013;15(5):623–33. Available from: http://dx.doi.org/10.1016/j.jmoldx.2013.05.004
7. Carrick DM, Mehaffey MG, Sachs MC, Altekruse S, Camalier C, Chuaqui R, et al. Robustness of next generation sequencing on older formalin-fixed paraffin-embedded tissue. PLoS One. 2015;10(7):3–10.
8. Schweiger MR, Kerick M, Timmermann B, Albrecht MW, Borodina T, Parkhomchuck D, et al. Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues

for copy-number-and mutation-analysis. PLoS One. 2009;4(5):3–9.

9. Wood HM, Belvedere O, Conway C, Daly C, Chalkley R, Bickerdike M, et al. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucleic Acids Res. 2010;38(14).
10. Kerick M, Isau M, Timmermann B, Sültmann H, Herwig R, Krobitsch S, et al. Targeted high throughput sequencing in clinical cancer Settings: Formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. BMC Med Genomics [Internet]. 2011;4(1):68. Available from: http://www.biomedcentral.com/1755–8794/4/68
11. Allen EM Van, Wagle N, Stojanov P, Perrin DL, Marlow S, Jane-valbuena J, et al. Whole-exome sequencing and clinical interpretation of FFPE tumor samples to guide precision cancer medicine. Nat Genet. 2014;20(6):682–8.
12. Oh E, Choi Y La, Kwon MJ, Kim RN, Kim YJ, Song JY, et al. Comparison of accuracy of whole-exome sequencing with formalin-fixed paraffin-embedded and fresh frozen tissue samples. PLoS One. 2015;10(12):1–13.
13. De Paoli-Iseppi R, Johansson PA, Menzies AM, Dias KR, Pupo GM, Kakavand H, et al. Comparison of whole-exome sequencing of matched fresh and formalin fixed paraffin embedded melanoma tumours: Implications for clinical decision making. Pathology. 2016;48(3):261–6.
14. Robbe P, Popitsch N, Knight SJL, Becq J, He M, Kanapin A. Europe PMC Funders Group Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens : pilot study for the 100, 000 Genomes Project. 2019;20(10):1196–205.
15. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The life history of 21 breast cancers. Cell. 2012;149(5):994–1007.
16. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods [Internet]. 2018;15(8):591–4. Available from: http://dx.doi.org/10.1038/s41592–018–0051-x
17. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol [Internet]. 2013;31(3):213–9. Available from: http://dx.doi.org/10.1038/nbt.2514
18. Hansen NF, Gartner JJ, Mei L, Samuels Y, Mullikin JC. Shimmer: Detection of genetic alterations in tumors using next-generation sequence data. Bioinformatics. 2013;29(12):1498–503.
19. Wilson RK, Mardis ER, McLellan MD, Koboldt DC, Shen D, Zhang Q, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568–76.

Tables

Table 1. Performance measures of calls considering the FF sample as gold standard for each variant caller.

	FF	FFPE	Overlap	Sensitivity	Precision	F1-score
Strelka2	6292	6761	4225	0.6715	0.6249	0.6474
Mutect2	10460	11815	5755	0.5502	0.4871	0.5167
VarScan2	4067	1760	883	0.2171	0.5017	0.3031
Shimmer	8109	10080	4865	0.6000	0.4826	0.5349

Table 2. Average significance score for somatic variants reported in both samples for each variant caller.

	Strelka2	Mutect2	VarScan2	Shimmer
FF	17.33	51.88	0.0024	0.0160
FFPE	14.34	26.00	0.0038	0.0166

Table 3. Optimized F1-scores of calls made by each variant caller considering FF sample as gold standard.

Caller - threshold	FF (gold std.)	FFPE	Overlap	Sensitivity	Precision	F1-score
Strelka2 – 9.5	4656	4559	3316	0.7122	0.7274	**0.7197**
Mutect2 – 13	4656	5658	3418	0.7341	0.6041	**0.6628**
VarScan2 – 0.00995	4656	1755	425	0.1045	0.2422	**0.1460**
Shimmer – 0.0495	4656	262	16	0.0197	0.0611	**0.0298**

Table 4. Strategies to retrieve the ground truth of calls from FF in the FFPE sample.

Reported by... (in FFPE)	FF (gold std.)	FFPE	Overlap	Sensitivity	Precision	F1-score
at least 1 caller	4656	16020	4155	0.8924	0.2594	**0.4019**
at least 2 callers	4656	4232	3684	0.7912	0.8705	**0.8290**
at least 3 callers	4656	340	325	0.0698	0.9559	**0.1301**
all 4 callers	4656	8	8	0.0017	1	**0.0034**

Table 5. Value assessment of Shimmer and VarScan2 in recovering the ground truth from FF in FFPE.

Reported by...	FF (gold std.)	FFPE	Overlap	Sensitivity	Precision	F1-score
At least 2 callers (with Shimmer)	4656	4232	3684	0.7912	0.8705	**0.8290**
At least 2 callers (without Shimmer)	4656	4219	3680	0.7904	0.8722	**0.8293**
Only Strelka2 and Mutect2	4656	4065	3610	0.7753	0.8881	**0.8279**

Table 6. Input material (in ng) per sample analyzed.

WGS ID	Sample type	Input (ng)
R18044561	FFPE	500
R18044562	Blood	300
R18044563	FF	300

Table 7. Summary of the main parameters used for Strelka2, Mutect2, VarScan2 and Shimmer.

Strelka2	Shimmer
Min Somatic EVS = 7	Max q-value acceptable FDR = 0.05
Mutect2	**VarScan2**
Min base quality score = 10 Min Phred-scaled confidence threshold = 10 Min TLOD = 5.3 Min NLOD = 2.3 Sample ploidy = 2 *FilterMutectCalls:* Min MedianBaseQuality = 20 Min MedianMappingQuality = 30	Min coverage in normal, in tumor = 8, 6 Min variant allele frequency = 0.01 Max somatic p-value = 0.05 *somaticFilter:* Min variant allele frequency = 0 Min read depth = 10 Min average quality = 20 Max somatic p-value = 0.01

Table 8. Screening space for the threshold optimization for each variant caller.

Strelka2	Shimmer
Somatic EVS from 5 to 20 (steps of 0.25)	Q-value from 0.0005 to 0.05 (steps of 0.0005)
Mutect2	**VarScan2**
TLOD from 0 to 200 (steps of 1)	Somatic p-value from 0.00005 to 0.01 (steps of 0.00005)

Figures

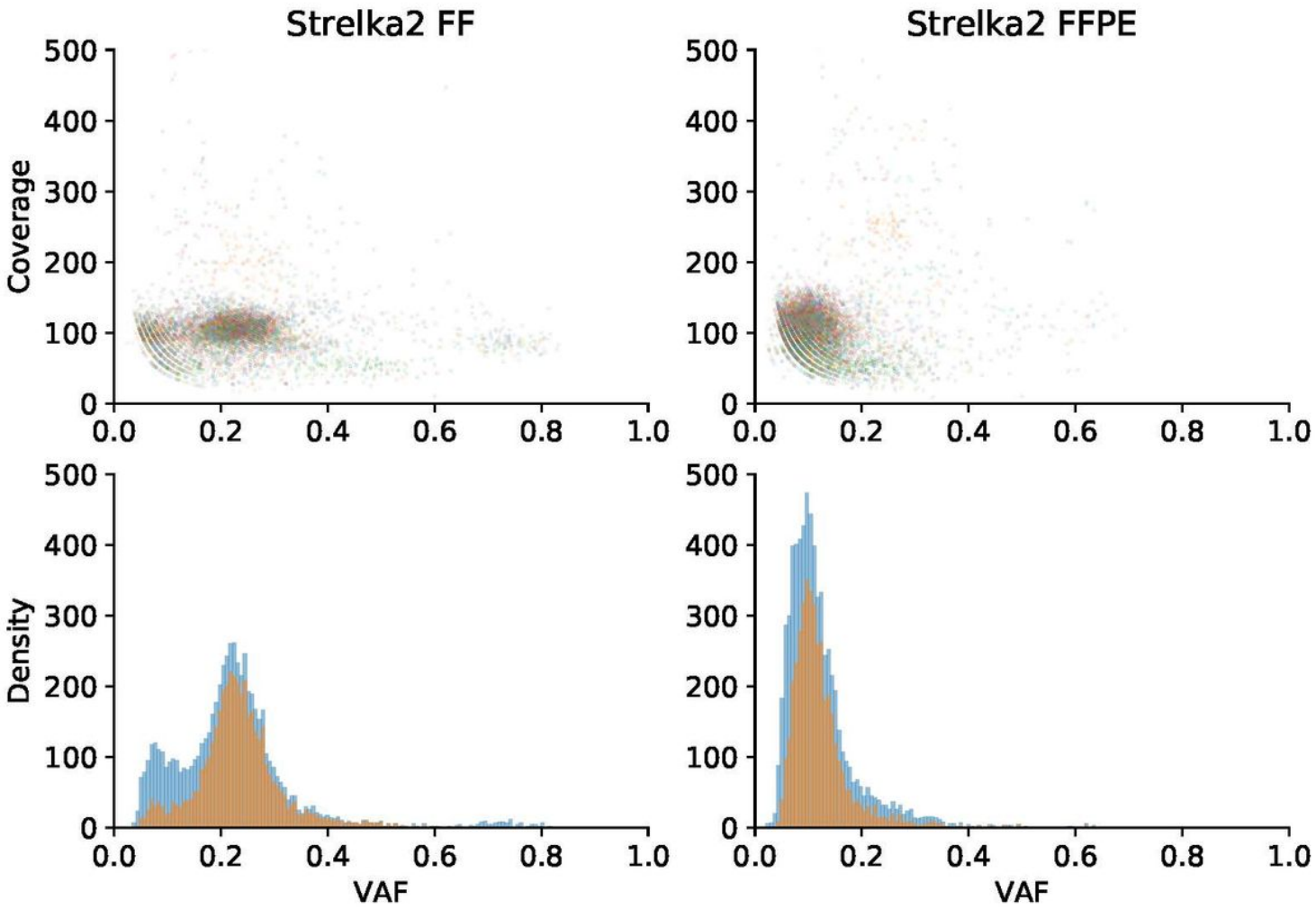

Figure 1

Clonal population detection in the FF (left) and the FFPE (right) sample using Strelka2. The upper panel shows coverage as a function of the VAF (15), where a higher variance in the coverage can be observed for FFPE. The lower panel shows the distribution of the VAFs. The blue distribution denotes all calls made in that sample, while the orange distribution shows only the calls common to FF and FFPE.

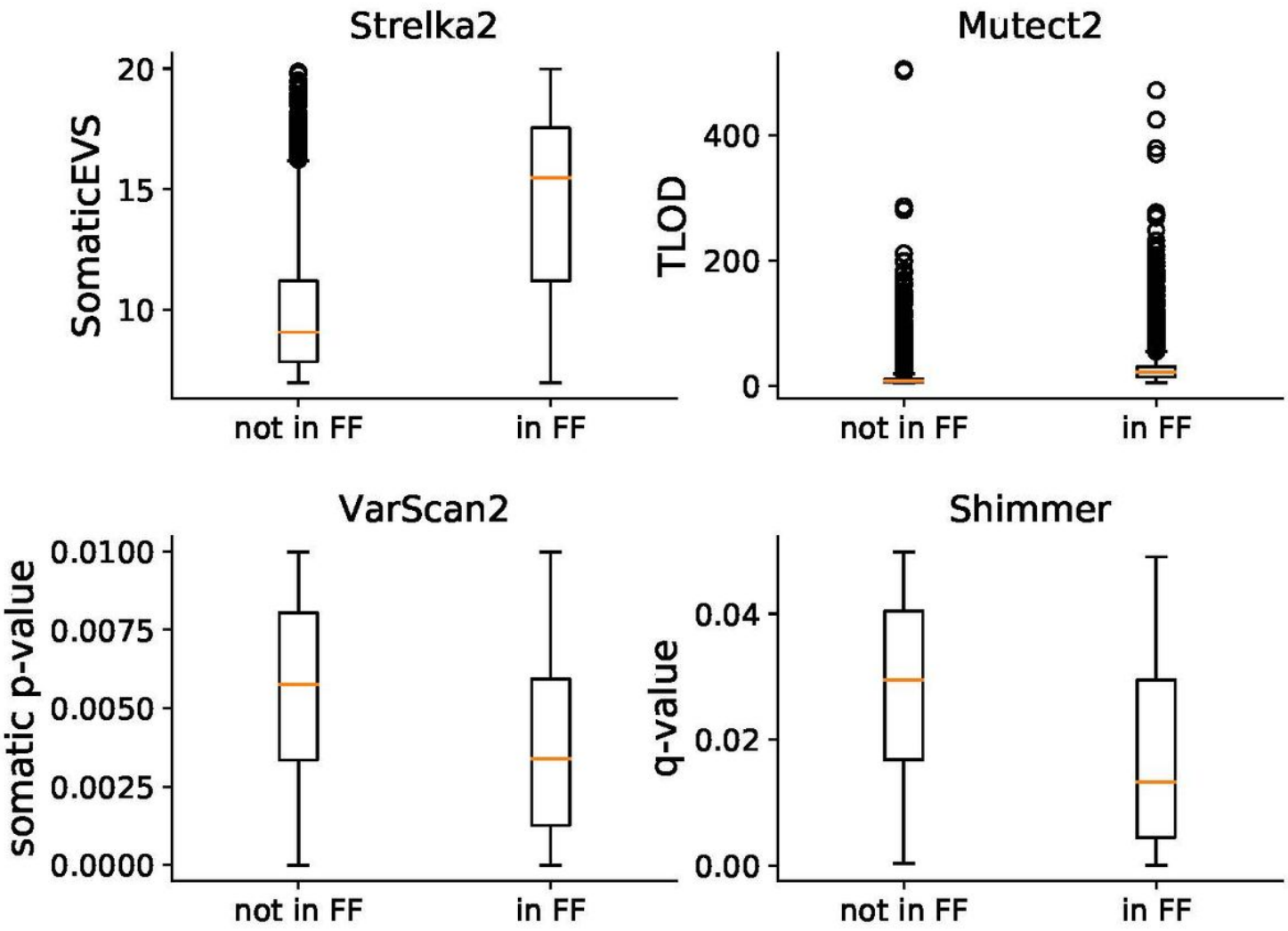

Figure 2

Boxplots comparing the significance level of calls from FFPE reported or not in FF sample. For Strelka2 and MuTect2 a higher Somatic EVS and TLOD, means a higher confidence in the calls, while for Varscan2 and Shimmer a lower value implies a higher confidence.

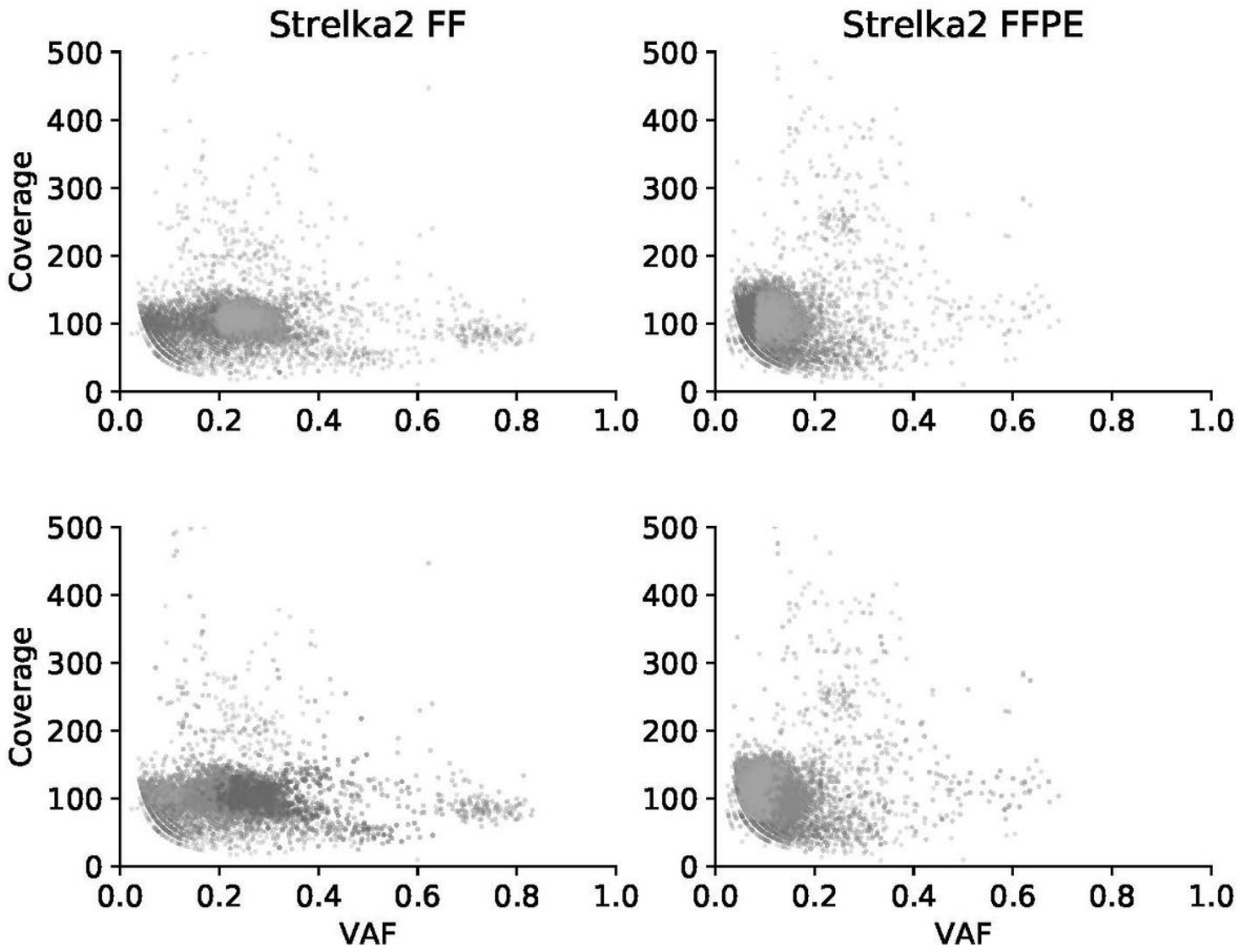

Figure 3

Coverage vs. VAF for somatic variants reported by Strelka2, comparing FF (left) against FFPE (right). This plot is identical to the upper panel of Fig. 1 but with a color used to indicate the most significant calls. The upper panel shows the 25% highest confidence calls in orange and the lower confidence in blue. The lower panel shows which calls are also found by other callers where blue= unique calls, orange = calls reported by 2 callers, green = calls reported by 3 callers, red = calls reported by 4 callers.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfigure13boxplotff.pdf
- Supplementaryfigure11shimmer.pdf
- Supplementaryfigure6shimmer.pdf

- Supplementaryfigure9mutect2.pdf
- Supplementaryfigure7shimmervaf0.pdf
- Supplementaryfigure8VAFffpeff.pdf
- Supplementaryfigure10varscan2.pdf
- SupplementaryMaterials.docx
- Supplementaryfigure3realvaf.pdf
- Supplementaryfigure2vennall.png
- Supplementaryfigure5varscan2.pdf
- Supplementaryfigure1simplevenn.pdf
- Supplementaryfigure14boxplotffpe.pdf
- Supplementaryfigure12shimmervaf0.pdf
- Supplementaryfigure15histopathology.png
- Supplementaryfigure4mutect2.pdf