The consequences of variant calling decisions in secondary analyses of cancer sequencing data

Carlos García-Prieto[1], Alfonso Valencia[1,2*] and Eduard Porta-Pardo[1,3*]

[1] Barcelona Supercomputing Center (BSC)

[2] Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[3] Josep Carreras Leukaemia Research Institute, Badalona, Spain

*To whom correspondence should be addressed.

Abstract

The analysis of tumor genomes is a key step to understand many aspects of cancer biology, ranging from its aetiology to its treatment or the oncogenic processes driving cell transformation. The first crucial step in the analysis of any tumor genome is the identification of somatic genetic variants that cancer cells have acquired during their evolution. For that purpose, a wide range of somatic variant callers have been created in recent years. However, it is still unclear which variant caller, or combination of them, is best suited to analyze tumor sequencing data.

Here we present a study to elucidate if different variant callers (MuSE, MuTect2, SomaticSniper, VarScan2) and strategies to combine them (Consensus and Union) lead to different downstream results in these three important aspects of cancer genomics: driver genes, mutational signatures and clinically actionable targets identification. To this end, we assess their performance in five different projects from The Cancer Genome Atlas (TCGA). Our results show that variant calling decisions have a significant impact on these downstream analyses, rendering important differences in driver genes prediction and mutational status among variant call sets, as well as in the identification of clinically actionable targets. More importantly, it seems that there is not a one-size-fits-all variant calling strategy, as the optimal decision seems to depend on both, the cancer type and the goal of the analysis.

Contact: eduard.porta@bsc.es; alfonso.valencia@bsc.es

1 Introduction

The widespread access to genomic data of tumor samples and cancer patients is transforming all aspects of this disease, from basic research to its clinical care (1). For example, thanks to sequencing data, we are beginning to understand the aetiology of the mutational mechanisms that affect cancer cells (2). Furthermore, since cancer progression relies on the expansion of clonal cell populations, we are now able to track and reconstruct the phylogenetic tree of tumor evolution (3). Similarly, the large cohorts of cancer patients that have been sequenced so far, have helped us identify germline and somatic mutations that predispose or drive carcinogenesis respectively (4,5), laying the foundation of personalized cancer care.

The first crucial step in analyzing cancer genomic data is the identification of genetic variants, particularly those of somatic origin. In that sense, the research community has made great efforts to assess the performance of the many different somatic variant callers available (6–11). However, so far, there has been no agreement on which variant caller nor which strategy to combine them is the most suitable. For instance, The Cancer Genome Atlas (TCGA) implemented different variant callers on multiple papers throughout its history (12–15). This led to the Multi-Center Mutation Calling in Multiple Cancers (MC3) project (16) to address standardization and reproducibility issues in the final part of the project. During MC3, many groups worked together to define a clear and unique strategy to combine the output of multiple variant calling tools. Another alternative to combine the output of different tools is to use machine learning approaches (17,18). However, despite all these efforts, it is still unclear which tool, or combination of tools, should be used to analyze cancer genomics data.

One likely reason for this lack of consensus is that it is difficult to define what is actually best in cancer genomics. At the end of the day somatic variant calling is a means to an end. Sequencing data from tumors can be used for many different secondary analyses, from finding cancer driver genes and mutations to determining the presence of actionable mutations or quantifying the effects of mutational signatures. Since none of the somatic variant callers nor strategies are perfect, it is possible that the answer to all these secondary analyses differs depending on which somatic variant calling tool or strategy is used.

While there have been benchmarking studies comparing how mutation callers find somatic mutations, to the best of our knowledge there has been no systematic study of the impact on variant calling tools in secondary analyses. Here we aim to elucidate if different variant calling strategies lead to different downstream results in three important aspects of cancer genomics: driver genes, mutational signatures and clinically actionable targets identification.

2 Methods

Variant calling datasets

To compare the effects of different mutation calling approaches in secondary analyses, we selected five different projects from The Cancer Genome Atlas (TCGA) (19): bladder urothelial carcinoma (BLCA) (20), prostate adenocarcinoma (PRAD) (21), adrenocortical carcinoma (ACC) (22), breast invasive carcinoma (BRCA) (13,23) and uterine corpus endometrial carcinoma (UCEC) (24). We chose these five projects to explore the impact of the somatic variant calling strategy in downstream analyses cohorts with different sizes, mutational signatures and mutational burdens. The Genomic Data Commons (GDC) portal (https://portal.gdc.cancer.gov) gives access to all the processed whole-exome sequencing (WXS) data for all the TCGA projects. In particular, the GDC created the DNA-Seq pipeline to process all TCGA samples in a uniform way (25). This pipeline includes sample preprocessing, alignment to the human reference genome GRCh38.d1.vd1 followed by BAM cleaning and somatic variant calling with variant annotation and aggregation. Somatic variants were identified in WXS data by comparing allele frequencies in matched tumor-normal samples. The GDC used four different variant calling tools to identify somatic mutations: MuSE (26), MuTect2 (27), SomaticSniper (28) and VarScan2 (29). After analyzing the whole-exome sequencing data for each individual sample, the GDC pipeline has an aggregation step that combines variants from all cases of a cancer cohort into a single TCGA project mutation annotation format (MAF) file.

For each of the five different cancer types, we downloaded the four different aggregated MAF files with all the somatic mutations for each variant caller. Additionally, we computed two extra mutation call sets per TCGA project. First, we created a Consensus file with those variants that were called by at least

3 out of the 4 aforementioned variant callers. Finally, we also created a Union file with every somatic variant called by any variant caller.

Identifying cancer driver genes

To identify cancer driver genes we used the IntoGen pipeline (30). Specifically, we analyzed every somatic variant file (MuSE, MuTect2, SomaticSniper, VarScan2, Consensus and Union) of each of the five TCGA projects analysed separately. IntoGen integrates the result of multiple computational tools to find cancer driver genes: OncodriveFML (31), OncodriveCLUSTL (32), dNdScv (33) , CBaSE (34), HotMAPS (35), smRegions (36). Most of the tools to find cancer driver genes in IntoGen rely on a series of randomizations to create a background mutational rate model and compute the likelihood of a given gene to present more mutations than expected by chance. These tools also identify genes with a significant linear or 3D clustering of mutations and genes with bias towards high functional mutations. The results of these tools are then combined to produce a consensus ranking of genes using a TIER classification while accounting for each tool credibility to detect driver genes on a given cohort and by combining the p-values that the different methods compute for the candidate genes.

For the purpose of our analysis, we only consider as true driver genes those within TIER 1 and TIER 2 (q-value <0.05). We discarded genes in TIER 3 (those that require manual curation) and TIER 4 genes (those deemed as likely false positives).

Table 1 Benchmarking metrics

Metric	Description
$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$	Proportion of true driver genes among the total number of driver genes.
$Precision = \frac{TP}{TP + FP}$	Also known as Positive Predictive Value (PPV). It is the ratio of correct driver genes among all driver genes detected with a given somatic variant calling dataset.
$Recall = \frac{TP}{TP + FN}$	Also known as Sensitivity. It is the ratio of correct driver genes among all driver genes detected with any other somatic variant calling dataset but the given, that are within the truth set.
$FDR = \frac{FP}{TP + FP}$	It is the ratio of incorrect driver genes among all driver genes detected with a given somatic variant calling dataset.
$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$	Harmonic average of precision and recall. The best value is 1 and the worst is 0.
$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$	A measure of the quality of binary classifications. It is the correlation coefficient between the observed and predicted binary classifications. It ranges from -1 (wrong binary classification) to +1 (perfect prediction).

MCC: Matthews Correlation Coefficient

Benchmarking variant calling strategies with driver genes

We created a truth set of 831 known driver genes to benchmark how the different mutation call sets can be used to detect cancer driver genes. To this end we combined the list of 299 driver genes published by the PanCancerAtlas project (5) with the 723 driver genes (TIER 1 and 2) present in the Cancer Gene

Census (CGC) (37) (https://cancer.sanger.ac.uk/census, 4/12/2019) of Catalogue Of Somatic Mutations In Cancer (COSMIC)(38).

We used multiple metrics (**Table 1**) to assess the performance of the different variant calling strategies for downstream IntoGen driver genes analysis. We defined our True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN) as follows (**Figure S1**):

- TP: driver genes identified with a given variant call set that fall within the truth set.
- FP: driver genes identified with a given variant call set that fall outside the truth set.
- FN: driver genes identified by any other variant call set but the given that fall within the truth set.
- TN: rest of the genes not identified with the given variant call set nor with any other strategy or identified with the latter but outside the truth set. We considered our universe as 20,000 genes.

Mutational signature analysis

We used *deconstructSigs* 1.8.0 R package (39) to quantify the presence of different mutational signatures in the different mutation call sets. In brief, deconstructSigs accounts for the trinucleotide context of each mutation to classify the six different base substitutions (C>A, C>G, C>T, T>A, T>C, T>G) into 96 possible mutation types (2). The signature matrix with the number of times a mutation is found within each trinucleotide context is compared against COSMIC Single Base Substitution (SBS) signatures (40).

Then, it uses an iterative approach, to assign different weights to each signature and estimate their contribution to the mutational profile of the tumor sample. We filtered out those samples with less than 50 mutations. Since we have analyzed WXS samples, the signature matrix was normalized to reflect the absolute frequency of each trinucleotide context as it would have taken place in the whole genome.

To assess the different contribution of mutational signatures to a sample mutational profile we used Kruskal-Wallis test comparisons across all variant call sets for a given cancer. Additionally, we also used pairwise comparisons (Wilcoxon test) between one strategy versus all the others for each signature and cancer.

Clinically actionable variants analysis

We used the Cancer Genome Interpreter (CGI, (41)) to detect alterations that might be therapeutically actionable. To detect alterations as biomarkers of drug response, CGI uses an in-house database of genomic events that influence tumor drug response (sensitivity, toxicity and resistance) classified with different levels of clinical evidence (FDA guidelines, case report, early trials, pre-clinical). Since CGI only supports hg19 mapped alterations, we transformed our MAF files to a BED-like format (keeping only “chr”, “start”, “end”, “ref allele”, “alt allele” and “sample ID” columns) and performed a liftover using *CrossMap* (42) version 0.3.4 (99.99% of variants were successfully remapped).

Purity dataset

We used purity/ploidy ABSOLUTE annotations for all TCGA samples available at https://gdc.cancer.gov/about-data/publications/pancanatlas (43).

Clinical dataset

We retrieved tumor stage information from the TCGA-Clinical Data Resource (44) file in https://gdc.cancer.gov/about-data/publications/pancanatlas.

3 Results

3.1 Effects of variant calling in the detection of cancer driver genes

One of the most widespread uses of somatic mutation data from cohorts of cancer patients is the identification of cancer driver genes. The tools to detect cancer driver genes can be sensitive to which somatic mutations are included in the final analysis, as they can bias some aspects of the randomization in which most cancer driver detection tools rely to make their analysis (31–36).

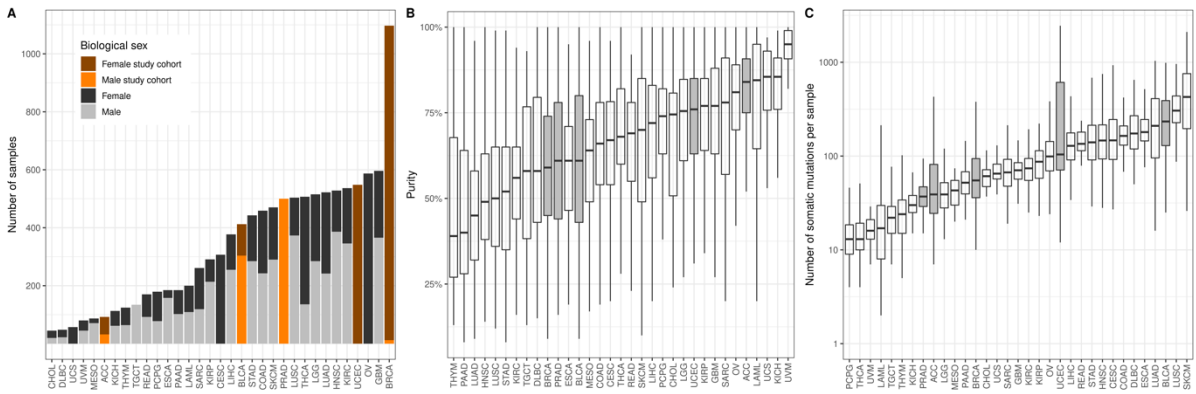

Fig. 1. TCGA projects characteristics. A)Biological sex distribution. Coloured bars represent the five cohorts we have analysed. **B)Purity distribution. C)Mutational load distribution**. Gray bars represent the five cohorts we have analysed.

To assess to what extent decisions about calling somatic variants can affect the detection of cancer driver genes, we used IntoGen to find cancer driver genes in thirty different mutation call sets for five different cancer types from TCGA. The six mutation call sets of each cancer type are distributed as follows: one mutation set with all the calls from one of the variant calling tools (MuSe (26), MuTect2 (27), SomaticSniper (28) and VarScan2 (29)), another mutation set -Consensus- with the consensus of the four variant callers (i.e. all those mutations found by, at least, three of the four callers) and a final mutation set with all the mutations found by any mutation caller -Union-.

Regarding variant calling strategies, we must point out that only MuTect2 and VarsCan2 were able to call small indels, thus being single nucleotide variants (SNVs) the main focus of our study. Therefore, Consensus call set was entirely comprised of SNVs. Furthermore, indels represented less than 10% of variants in MuTect2 and VarScan2 call sets. Consequently we decided not to split up our analysis and consider all variants together.

We chose the five cancer types (ACC (22), BLCA (20), BRCA (13,23), PRAD (21) and UCEC (24)) so that they spanned a variety of mutational processes, ranges of purity, mutation rates and cohort sizes within TCGA. For example, ACC is one of the smallest cohorts within TCGA (n = 92), as well amongst those with the highest tumor purity (average purity 80%) (45). On the other hand, BRCA is the largest cohort in TCGA (n = 986). Another factor that can alter the efficiency of tools to detect cancer driver genes is the mutation rate of the cohort, hence why we included UCEC, which is amongst the cancer types with highest mutation rates (5). Finally, BLCA and PRAD are amongst the cohorts that are closest to the TCGA average in all these aspects, making them good representatives of the average tumor sample (**Figure 1**).

In terms of absolute number of somatic mutations, the mutation call sets from SomaticSniper and MuTect2 are, respectively, the smallest and largest (not accounting for the Union) in all cancers but ACC (**Figure S2**). One possible factor affecting the number of somatic mutations identified by each caller is the sample purity. To assess the impact of this variable for each patient sample we compared its purity with the ratio of somatic mutations called by the Union that were recalled by each variant call set. We found a positive correlation between purity and Union recall for every call set except for MuTect2, where the correlation was negative in all cancers (**Figure S3**). This reflects that most of the calls in low purity samples were made by MuTect2, thus being a variant caller with high sensitivity when dealing with low purity tumor samples (10,46).

We next correlated sample purity and tumor stage to observe if the lower the purity the later stage, since these tend to be more heterogeneous and have more subclonal populations, but we only find this trend in BLCA (**Figure S4**). To further assess stage implication in variant calling, we analysed the range of variant allele frequencies (VAFs) spanned by each call set (**Figure S5**). This information, additionally, may help guide patient treatment since is used to classify mutations as clonal (0.5 VAF) or subclonal. We observe that SomaticSniper call set spans a median VAF range of 0.4-0.5 already adjusted by purity and ploidy (43), whereas MuTect2 and Union call sets span lower VAF ranges calling more subclonal mutations. This may suggest SomaticSniper as a reliable caller for early stage tumors, where we do not

expect to find many subclonal populations. On the other hand, MuTect2 and Union may be more suitable approaches for late stage tumors.

Having assessed the influence of various tumor properties in the number of mutations called by each tool, we next wondered what the effect would be when trying to identify cancer driver genes. The number of consensus genes in each cancer type positively correlated with the median mutation burden (spearman R = 0.53, p.value = 0.0029) (**Figure 2A**), as already described by the MC3 (5). However, if we split by tumor, a significant negative correlation is found for PRAD (spearman p.value = 0.019) (**Figure 2B**).

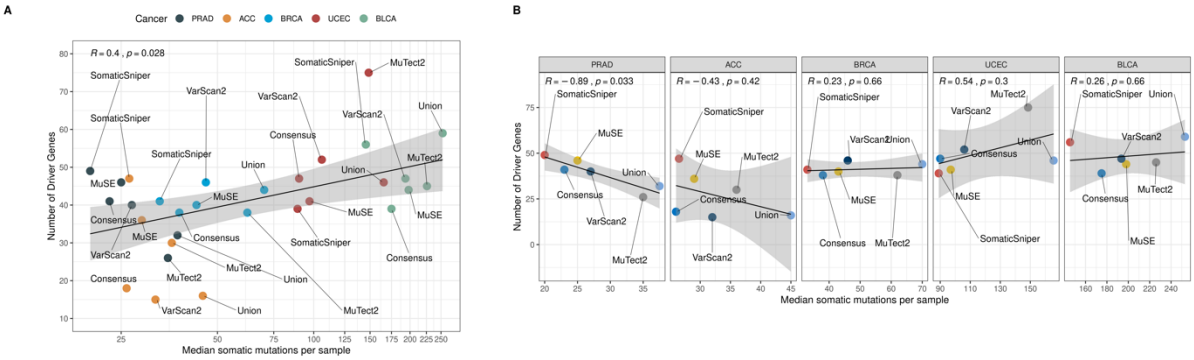

Fig. 2. Mutational load vs number of driver genes predicted by IntoGen. Shaded area indicates 95% bootstrapped confidence interval. A)Results per callset and B) per TCGA project.

Strikingly, despite having the lowest number of mutations, the analysis of the variant call set from SomaticSniper identifies the largest number of driver genes in ACC and PRAD and the second largest in BLCA (**Figure 2**). In addition, there were between 7 and 8 driver genes identified only with this call set in each of the aforementioned cancers (**Figure S6**, **Tables S1-5**). For instance, we could only identify BRCA1 as a driver gene in BLCA in the analysis of the SomaticSniper variant call set. This is a good example of how variant call sets impact downstream analysis. Another interesting result was found in UCEC, where using MuTect2 call set IntoGen identified 75 driver genes, 29 of those being only identified with this call set.

To compare the performance of the IntoGen results with the different variant call sets, we did a benchmarking analysis by creating a truth set of well known driver genes. To this end, we merged those genes present in the PanCancerAtlas MC3 project and those in the CGC of COSMIC. Different benchmarking metrics results were found depending on the tumor being analysed, suggesting there is not a best variant call set overall (**Figure 3**). For instance, MuTect2 call set performance in UCEC was the worst with the highest false discovery rate (FDR). However, MuTect2 outperformed the rest of variant call sets in PRAD. The call set from VarScan2 was the most regular overall, showing consistency among the different tumors. SomaticSniper and MuSE had very similar benchmarking results, also being consistent among all cohorts, though not as much as VarScan2.

As explained above, we also created two additional variant call sets that result from the combination of the results of the four variant calling tools. One of them, the Consensus, is very important, as it is a very common strategy in the cancer genomics community. The Consensus mutation call set had the lowest recall in BLCA, but it seems to be the best in UCEC and ACC. Therefore, is important to bear in mind that this common approach is not always the best one and, depending on the purpose and interest of our study, one should consider alternative strategies.

The other additional variant call set is the Union. This results from the combination of all the mutations identified by any variant caller. Given the likely high number of false positive mutations in this set, we would expect it to score a high FDR but this was only the case in BRCA. In all the other cancer types, the Union proved to be a pretty robust strategy overall scoring best F1-Score and Matthews correlation coefficient (MCC) in BLCA and second best in ACC, PRAD and UCEC. This shows how starting from a call set with supposedly lots of false positive calls, one can still identify the majority of cancer driver genes in a cohort.

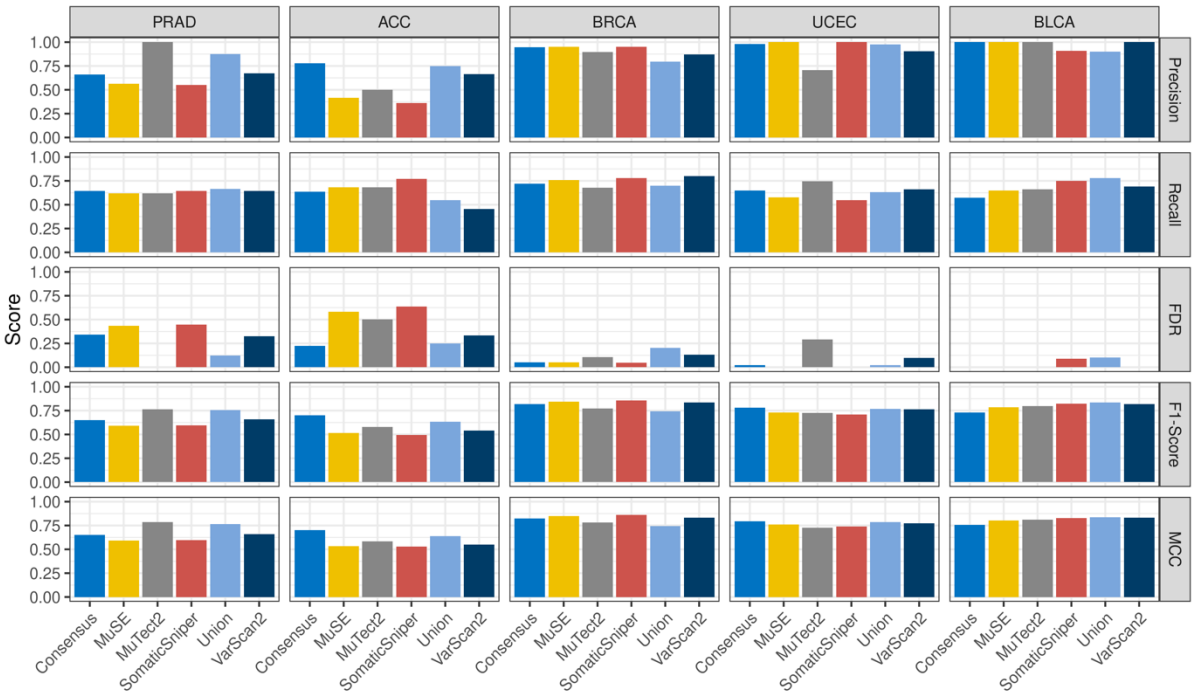

Fig. 3. Performance metrics of the different variant calling strategies when identifying cancer driver genes with IntOGen

3.2 Mutational signatures

Mutational signatures analysis is relevant for the study of biological mechanisms underlying somatic mutations such as defective DNA repair, mutagenic exposures, DNA replication infidelity or enzymatic DNA modifications. These mutational processes have implications in the understanding of cancer aetiology and may inform patient treatment.

The mutational signatures contributing the most to individual tumor genomes were SBS2 and SBS13 in BLCA and BRCA, both related to APOBEC family of cytidine deaminases activity. Significant differences are found when comparing mutational signatures contribution to individual tumor genomes among variant call sets (**Figure 4**). Interestingly, we observe the most significant difference in BRCA for SBS13 (Kruskal-Wallis p value = 3.7e-06). Other significant differences are found in BRCA (SBS1, SBS2, SBS5, SBS22 and SBS39), BLCA (SBS17a) and UCEC (SBS37) (**Figure S7**). Furthermore, no significant differences are found in ACC neither in PRAD. Finally, no signature is found to be significantly different in more than one tumor type. Therefore, we conclude that mutational signatures analysis is pretty robust regardless of the variant call set selected, although significant differences are found in specific signatures and cancers.

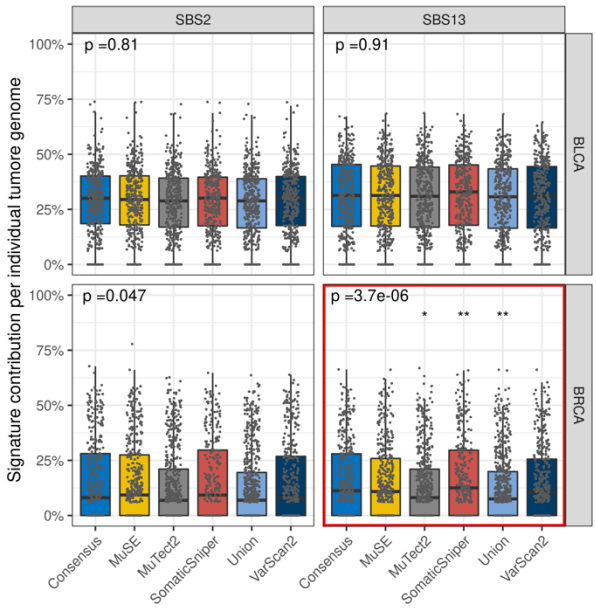

Fig. 4. Mutational signatures contributing the most to individual tumor genomes. Displayed only BLCA and BRCA because are the ones with larger signatures contribution. Squared in red the one with the most significant differences among variant call sets. Global p.value was computed with Kruskal-Wallis test comapring all variant calling strategies. Pairwise comparison were performed with Wilcoxon test comparing each caller against all others.

3.3 Somatic mutations in cancer driver genes

Even if one can identify a gene as a driver in a cohort using a certain variant call set, it is possible that some individual mutations in the gene in certain samples are missing. This could have important implications for patients, as the presence or absence of mutations in cancer driver genes can determine whether patients will receive certain treatments or not.

To evaluate the impact of decisions about somatic variant calling in finding mutations in cancer driver genes, we calculated the ratio of patients with differing mutation status depending on the mutation set used. To this end, we first computed the total number of samples in the Union call set with a mutation in a given driver gene. Secondly, for the given driver gene, we computed the percentage of those Union samples that were missing in each variant call set (**Figure 5**).

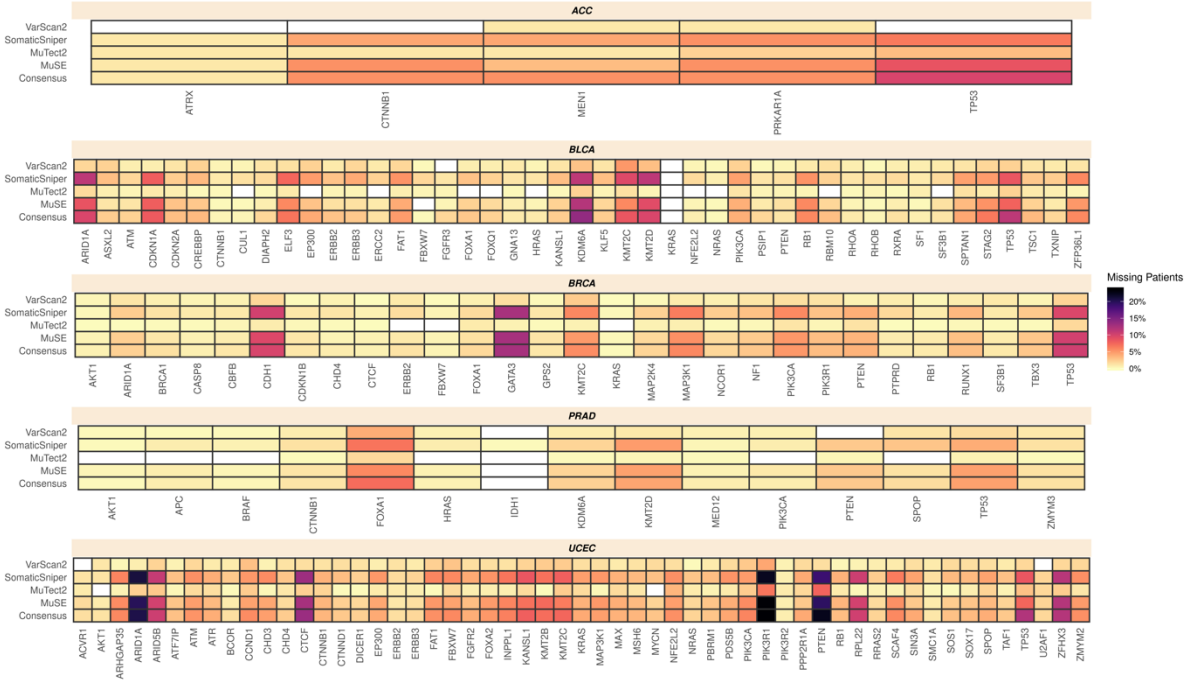

Fig. 5. Patients with driver genes differing mutation status according to the variant call set used. Only the MC3 driver genes of each cancer are displayed. Union call set is used as reference to consider all possible variants.

Overall, as expected, there is a correlation between the total number of mutations called by each method and the number of mutations detected in cancer driver genes. MuTect2 and specially VarsCan2 detected more mutations in cancer driver genes than SomaticSniper, MuSe and the Consensus. Interestingly, driver genes detected in multiple cancer types, such as TP53, ARID1A or PTEN, also showed more differences across call sets. In particular, TP53 is the gene with more differences in the number of patients with detected somatic mutations, with 5-15% patients that vary their TP53 mutational status across cancers depending on the variant call set selected. Other cancer driver genes with significant number of patients with differing somatic mutation status, between 10% and 15%, are ARID1A, KDM6A, KMT2C and KMT2D in BLCA or CDH1 and GATA3 in BRCA. In terms of cancer types, UCEC shows the largest variation across call sets with more than 20% of patients with diverging mutation status in ARID1A, PIK3R1 and PTEN depending on which mutation call set is used.

3.4 Drug biomarkers

Another important goal of the analysis of somatic cancer genomes is the identification of clinically actionable variants. These are some somatic variants that help oncologists and physicians decide whether they should give a treatment to a cancer patient, as they are associated with sensitivity or resistance to some cancer treatments. Therefore, properly assessing the presence of such variants in the genome of cancer cells is of ultimate clinical importance. Thus, we used the Cancer Genome Interpreter to detect and classify variants as biomarkers of drug response according to different levels of clinical evidence.

Importantly, we found many differences in the number of clinically actionable variants depending on which variant caller is used (**Figure 6**). As expected from their total number of mutations, the mutation sets from MuTect2, Union and VarScan2 have between 5-15% more patients with at least one clinically actionable mutation compared to Consensus, MuSE and SomaticSniper. These results reflect the fact that the larger the variant call set is, the more likely it is to find a clinically actionable target.

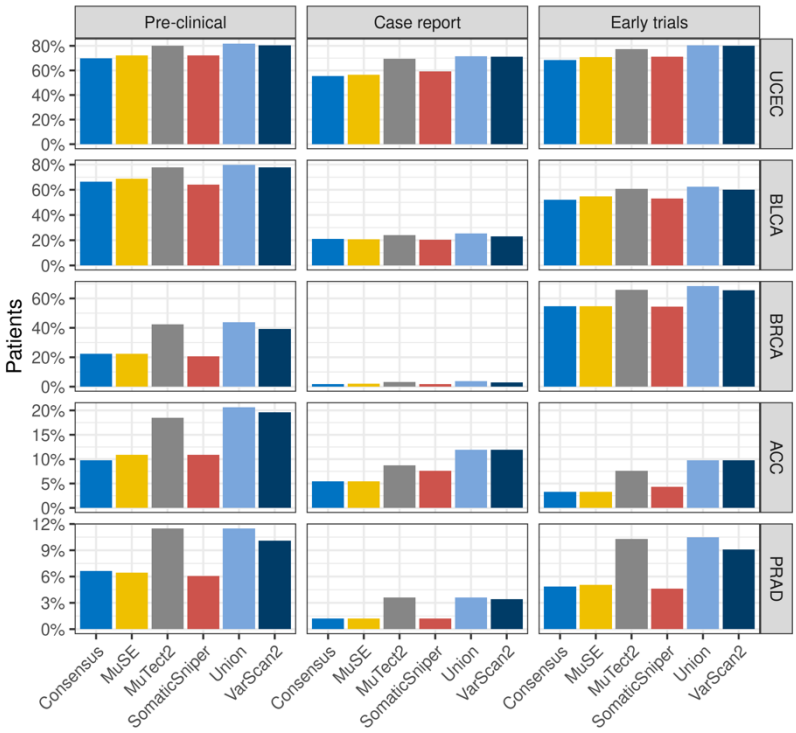

Fig. 6. Patients with at least one clinically actionable target identified by Cancer Genome Interpreter (CGI) classified per clinical evidence level.

Furthermore, we observed larger differences among variant call sets in BLCA, BRCA and UCEC. Additionally, we found the most important differences in BRCA with MuTect2, Union and VarScan2 call sets identifying 10% and 20% more patients bearing at least one clinically actionable mutation within the early trials and preclinical evidence levels respectively.

The actionable alterations contributing to the most number of samples are TP53 mutations, found in 30% and 50% of BRCA and BLCA patients respectively (**Figure S8**). PIK3CA mutations affect more than 30% and 40% of BRCA and UCEC patients respectively. Moreover, we found great differences in UCEC patients bearing PTEN clinically actionable mutations: between 40% to 60% of patients bear this actionable mutation depending on the variant calling strategy. This 20% difference is important since these mutations may predict sensitivity to PI3K pathway inhibitors, mTOR and PARP inhibitors.

4 Discussion

The comprehensive characterization of tumor somatic events may inform about cancer aetiology and help guide treatment. The analysis of large cohorts of cancer patients genomic data has revealed new driver genes and mutational mechanisms underlying oncogenesis. Additionally, novel actionable targets have been discovered. However, to the best of our knowledge no systematic study has been conducted to assess the impact of variant calling approaches in all these analyses.

We have shown that variant calling decisions can have a significant impact on downstream analyses yielding different outcomes in driver genes, mutational signatures and clinically actionable targets identification. More importantly, it seems that there is not a one-size-fits-all variant calling strategy, as the optimal decision seems to depend on both, the cancer type and the goal of the analysis.

If we aim to identify new driver genes in a cohort we should consider using SomaticSniper. This tool has shown to be a pretty stringent variant caller retrieving the smallest mutation call set in most cancer types. We believe that this could reduce the noise of false positive calls, which seems to be especially useful for driver gene detection. Moreover, different studies have shown that SomaticSniper is the variant caller with the smallest number of false positive calls and lowest FDR (16,18). However, as shown by our benchmarking analysis, we must account for the cancer being analysed.

We have discovered very important differences among strategies in the detection of patients with variants in driver genes. In some cases, more than 20% of patients with a variant in a specific driver gene could differ in their mutation status depending on the variant call set selected.

Given our results in cancer driver detection and identifying somatic mutations in cancer driver genes, one possible strategy could be to use the mutation call set from SomaticSniper to detect the cancer driver genes and then run a less stringent variant caller, such as MuTect2, to look for mutations in driver genes on a second pass.

We would like to point out that the Consensus, a very widely used strategy by the research community, showed noticeable limitations in some aspects. Reducing the number of suspected false positive mutation calls at the expense of true positive ones did not necessarily translate into a better performance in many experiments, as this variant call set led to many cancer driver genes not being detected and driver and actionable mutations missing. On the other hand, the Union has proved to be a legit strategy for downstream clinical analysis with a robust performance overall.

Another important aspect to consider while conducting variant calling is sample purity. We have shown that variant callers spann different VAF ranges. This reflects their different capabilities to call mutations in samples with different purities, which can also correlate with tumor stage where heterogeneity increase shapes VAF distribution. Since most of the variants called in low purity samples were made by MuTect2, we consider it to be a more sensitive variant caller for late stage tumors composed of mixtures of subclonal and normal cell populations. On the other hand, SomaticSniper was able to better call clonal mutations, thus being more sensitive for early stage tumor samples calling.

The study of somatic genomes mutational signatures helps to better characterize the mutational processes harboring somatic mutations. We have shown that the analysis of the mutational signatures contributing the most to individual tumor genomes is mostly robust regardless of the variant call set selected. However, we observed significant differences in some signatures and cancers.

If the goal of the analysis of the somatic genome is to find clinically actionable mutations, we need to be aware that there are considerable differences depending on the somatic mutation calling used. On average, 10% of patients with at least one clinically actionable target can either be detected or not depending on the variant calling strategy.

We acknowledge several limitations in our study. For example, we are not considering results for copy number and structural variants in the mutation call sets. Additionally, the limited availability of experimentally validated variants with orthogonal sequencing prevents us from extracting final conclusions regarding which variant caller performance is more robust.

Finally, we hope this study will help researchers consider the whole picture while performing variant calling. It is important to account for the clinical implications that variant calling decisions have on different downstream analyses, especially in such important aspects of cancer genomics like driver genes, mutational signatures and actionable targets identification. Moreover, we aim this study will help guide variant calling design while considering the needs and goals of the different research projects.

Acknowledgements

The authors would like to thank the patients that donated the samples for The Cancer Genome Atlas, without them this work would not be possible.

Funding

This work has been supported by the La Caixa Junior Leader Fellowship

Conflict of Interest: none declared.

References

1. Hyman DM, Taylor BS, Baselga J. Implementing Genome-Driven Oncology. Cell. 2017 Feb 9;168(4):584–99.
2. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013 Aug;500(7463):415–21.
3. 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 May 25;149(5):994–1007.
4. Huang K, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, et al. Pathogenic Germline Variants in 10,389 Adult Cancers. Cell. 2018 Apr 5;173(2):355-370.e14.
5. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018 Apr 5;173(2):371-385.e18.

6. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J. 2018 Feb 6;16:15–24.
7. Cai L, Yuan W, Zhang Z, He L, Chou K-C. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. Scientific Reports. 2016 Nov 22;6(1):1–9.
8. Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nature Communications. 2015 Dec 9;6:10001.
9. Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH, et al. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. Sci Rep. 2017 24;7:43169.
10. Consortium TSMWG of the S-I. Achieving reproducibility and accuracy in cancer mutation detection with whole-genome and whole-exome sequencing. bioRxiv. 2019 Jun 2;626440.
11. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, et al. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Medicine. 2013 Oct 11;5(10):91.
12. Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, et al. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. Cell. 2017 Oct 19;171(3):540-556.e25.
13. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell. 2015 Oct 8;163(2):506–19.
14. Abeshouse A, Adebamowo C, Adebamowo SN, Akbani R, Akeredolu T, Ally A, et al. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. Cell. 2017 Nov 2;171(4):950-965.e28.
15. Robertson AG, Shih J, Yau C, Gibb EA, Oba J, Mungall KL, et al. Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma. Cancer Cell. 2017 Aug 14;32(2):204-220.e15.
16. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. cels. 2018 Mar 28;6(3):271-281.e7.
17. Wood DE, White JR, Georgiadis A, Emburgh BV, Parpart-Li S, Mitchell J, et al. A machine learning approach for somatic mutation discovery. Science Translational Medicine. 2018 Sep 5;10(457):eaar7939.
18. Anzar I, Sverchkova A, Stratford R, Clancy T. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. BMC Medical Genomics. 2019 May 16;12(1):63.
19. The Cancer Genome Atlas Program [Internet]. National Cancer Institute. 2018 [cited 2019 Oct 30]. Available from: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
20. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014 Mar;507(7492):315–22.
21. Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, et al. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015 Nov 5;163(4):1011–25.
22. Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, et al. Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. Cancer Cell. 2016 May 9;29(5):723–36.
23. Comprehensive molecular portraits of human breast tumours. Nature. 2012 Oct;490(7418):61–70.
24. Levine DA, The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. Nature. 2013 May;497(7447):67–73.
25. Grossman RL, Heath AP, Ferretti V, Varmus H, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. The New England journal of medicine. 2016;375(12):1109–12.
26. Fan Y, Liu X, T HDS, Zhang J, Zhang J, Futreal PA, et al. Accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling for sequencing data. bioRxiv. 2016 May 25;055467.
27. 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. Nature Biotechnology. 2013 Mar;31(3):213–9.
28. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012 Feb 1;28(3):311–7.
29. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012 Mar 1;22(3):568–76.
30. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nature Methods. 2013 Nov;10(11):1081–2.
31. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biology. 2016 Jun 16;17(1):128.
32. Arnedo-Pac C, Mularoni L, Muiños F, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. Bioinformatics. 2019 Nov 1;35(22):4788–90.
33. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Loo PV, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017 Nov 16;171(5):1029-1041.e21.
34. Weghorn D, Sunyaev S. Bayesian inference of negative and positive selection in human cancers. Nature Genetics. 2017 Dec;49(12):1785–8.

35. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer Res. 2016 Jul 1;76(13):3719–31.
36. Martínez-Jiménez F, Muiños F, Lopez-Arribillaga E, Lopez-Bigas N, Gonzalez-Perez A. Disruption of ubiquitin mediated proteolysis is a widespread mechanism of tumorigenesis. bioRxiv. 2018 Dec 28;507764.
37. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nature Reviews Cancer. 2018 Nov;18(11):696.
38. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017 Jan 4;45(D1):D777–83.
39. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biology. 2016 Feb 22;17(1):31.
40. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Ng AW, Boot A, et al. The Repertoire of Mutational Signatures in Human Cancer. bioRxiv. 2018 May 15;322859.
41. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Medicine. 2018 Mar 28;10(1):25.
42. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014 Apr 1;30(7):1006–7.
43. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018 Apr 5;173(2):291-304.e6.
44. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018 Apr 5;173(2):400-416.e11.
45. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nature Communications. 2015 Dec 4;6:8971.
46. Bohnert R, Vivas S, Jansen G. Comprehensive benchmarking of SNV callers for highly admixed tumor data. PLOS ONE. 2017 Oct 11;12(10):e0186175.