


SOFTWARE

Open Access



pyCancerSig: subclassifying human cancer with comprehensive single nucleotide, structural and microsatellite mutational signature deconstruction from whole genome sequencing

Jessada Thutkawkorapin¹, Jesper Eisfeldt^{1,2}, Emma Tham^{1,2} and Daniel Nilsson^{1,2*} 

Abstract

Background: DNA damage accumulates over the course of cancer development. The often-substantial amount of somatic mutations in cancer poses a challenge to traditional methods to characterize tumors based on driver mutations. However, advances in machine learning technology can take advantage of this substantial amount of data.

Results: We developed a command line interface python package, pyCancerSig, to perform sample profiling by integrating single nucleotide variation (SNV), structural variation (SV) and microsatellite instability (MSI) profiles into a unified profile. It also provides a command to decipher underlying cancer processes, employing an unsupervised learning technique, Non-negative Matrix Factorization, and a command to visualize the results. The package accepts common standard file formats (vcf, bam). The program was evaluated using a cohort of breast- and colorectal cancer from The Cancer Genome Atlas project (TCGA). The result showed that by integrating multiple mutations modes, the tool can correctly identify cases with known clear mutational signatures and can strengthen signatures in cases with unclear signal from an SNV-only profile. The software package is available at <https://github.com/jessada/pyCancerSig>.

Conclusions: pyCancerSig has demonstrated its capability in identifying known and unknown cancer processes, and at the same time, illuminates the association within and between the mutation modes.

Keywords: Human cancer, Mutational signatures, Unsupervised learning, Cancer processes

Introduction

Cancer is a genomic disorder, involving different kinds of DNA damage. DNA damage and imperfect repair occurs frequently in human cells. Changes accumulate over time, starting with our first cell, the fertilized egg, and progressively over the course of cell division [1]. Cellular proteins pertaining to replication, damage sensing and repair are

important in limiting the damage. Most induced DNA variations will have no effect on the cell (so called passenger variants). However, over time there is a risk that accumulating DNA damage can result in tumorigenesis and later, metastasis, regardless of whether the initial driver mutation stems from inheritance, induced damage or imperfections in the replication of the hereditary material. The continued accumulation of DNA changes is caused by a combination of exposure to sources of DNA damage, both endogenous and exogenous and a broken DNA damage response mechanism. Each of these have their own

* Correspondence: daniel.nilsson@ki.se

¹Department of Molecular Medicine and Surgery, Karolinska Institutet, SE-171 76 Stockholm, Sweden

²Department of Clinical Genetics, Karolinska University Hospital, SE-171 76 Stockholm, Sweden



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

respective profile with regard to the type of damage inflicted and bias in the repair mechanisms including double-strand DNA breaks [2], single-strand DNA breaks [3], and microsatellite instability [4].

Conventional methods to characterize tumor behavior are based on mutations in functional domains of cancer driver genes. Identifying the driver mutation is difficult, since the majority of the somatic events are passenger mutations [5]. An alternative is to classify tumors based on tumor behavior directly, by observing the pattern of somatic mutations. With an exploding amount of whole exome and whole genome sequencing data, together with advances in machine learning technology, a computational approach to characterize tumors based on their base-substitution profile was initiated, capturing mutation signatures [6]. These pioneering results were extended, showing e.g. how the mutational signatures associated with the presence of particular somatic driver mutations [7] and how a mutational signature enables classification of germline missense mutation [8]. However, base substitution is only one result from DNA damage. The others include copy number variation, structural rearrangement and microsatellite instability. Sensing both single nucleotide and structural variation (SV) has been shown efficient in predicting *BRCA1/BRCA2* null mutation [9]. Tumor microsatellite instability (MSI) profiling has been proven useful clinically to inform testing for mutation in mismatch repair genes (MMR) [10] and more recently, to identify tumors eligible for immunotherapy [11, 12]. Tools to classify MSI status from tumor genome short read sequencing data [13] have been successfully developed.

Integrating several types of genetic aberrations from different types of DNA damage may not only reveal patterns and connections between these aberrations, but may also reveal novel combinations of cancer processes and possibly allow novel clinically actionable distinctions between tumors and inform interpretation of larger classes of hereditary variants of uncertain significance.

Most tools available for mutational signatures are implemented in R or as web interfaces, and require custom file formats [14]. Tools otherwise used in large scale processing of short read sequencing data in production environments are nearly exclusively command line based, and operate on a limited set of standardized file formats.

In this article, we describe pyCancerSig: an open source, command line interface python package for deciphering cancer signatures; integrating SNV, SV and MSI profiles in signatures decomposed using non-negative matrix factorization; and producing pdf reports.

Materials and methods

Workflow processes

The workflow consists of 4 steps, data pre-processing, profiling, deciphering and visualizing (Figs. 1, 2). Example input files and example output files, together with corresponding

command line examples, for each of the following steps are available at <https://github.com/jessada/pyCancerSig>.

Data pre-processing

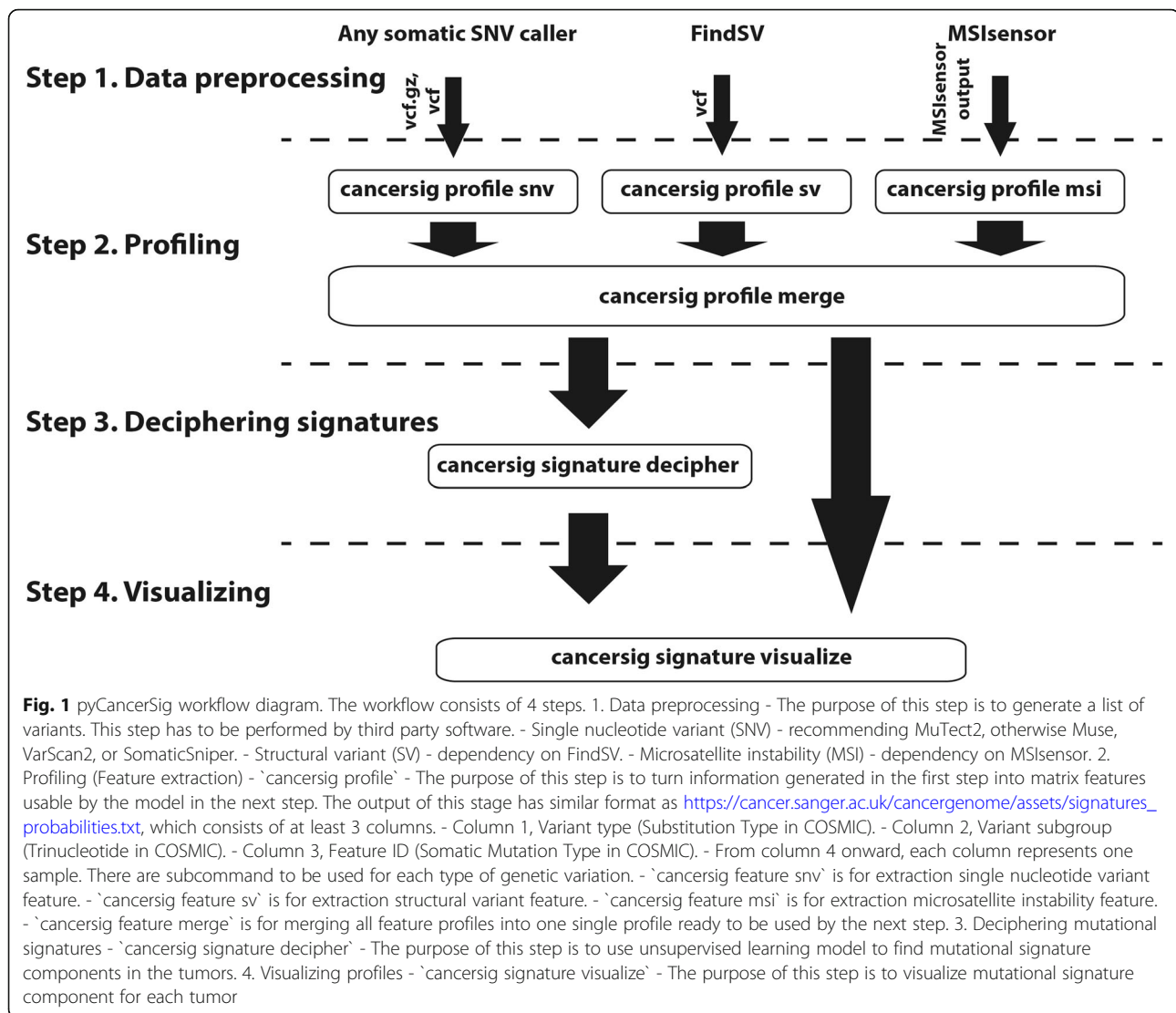
The purpose of this step is to generate a list of variants. This step has to be performed by third party software. For SNVs, any somatic callers that produce variant calls in VCF format can be used. The same is true for SV calls, but even though the VCF standard has support for SVs, callers may not always be fully interchangeable. Specifically, the “END” tag added by many callers and a “CHR2” tag are parsed out from the INFO field. Other information not evident from the VCF definition could be parsed by replacement or modification of a custom parseVCFLine function, as was done for FindSV. We have tested with, and recommend FindSV [15]. This includes support for TIDBIT [16], CNVnator [17] and Manta [18]. For MSI, the MSI profiling in the stage currently has a dependency on MSIsensor [13].

Profiling

The purpose of this step is to quantify information generated in the first step into matrix features usable by the deciphering process. There are three types of profiles, SNV, SV and MSI. The feature weights for SNV:SV:MSI were selected as 7:2:1 for this study. The weights can be adjusted in a feature selection file to the program package as needed.

SNV profiling scans the VCF file and quantifies six types of base substitution C:G > A:T, C:G > G:C, C:G > T:A, T:A > A:T, T:A > C:G, and T:A > G:C. Then, the changes are further subclassified by the sequence context, 5' and 3', of the mutation. In total, there are 96 mutation types (6 types of substitution × 4 types of 5' base × 4 types of 3' base). For example, if a sample has a variant T > C at position 47,672,720 (hg19/GRCh37) on chromosome 2, which has a neighbouring 5' G and also a 3' G, this will be counted as one mutation for the mutation type G [T > C] G in the feature output file.

SV profiling scans the VCF record for the INFO field SVTYPE to classify the mutation into structural duplication, structural deletion, inversion, and translocation. Then it checks another INFO field, END, to calculate the length of the event, in order to subclassify the event based on its approximated size in log10 scale (size between 100-1Kbps, 1K-10K, 10K-100K, 100K-1M, 1M-10M, 10M-100M, 100M-1000M, and whole chromosome). In total, there were 32 mutation types (4 types of variation × 8 length groups). For example, if a VCF record has an INFO field SVTYPE set to “DUP” and a length of 15,000 bp, this would be counted as one mutation for structural duplication of length between 10K-100K.



MSI profiling quantifies all possible repeat patterns of a repeat unit with size between 1 and 3 nucleotides (A, C, AC, AG, AT, CG, AAC, AAG, AAT, ACC, ACG, ACT, AGC, AGG, ATC, CCG). For repeat unit sizes of 4 and 5 base pairs, it performs quantification without checking repeat patterns. The profile merger merges all sample profiles into one ready to be used by the deciphering- and visualization processes.

Deciphering

The purpose of this step is to use an unsupervised learning model to find mutational signature components in the tumors. The process was developed based on a well-established framework [6] and uses non-negative matrix factorization (NMF) to decipher signature probabilities, matrix P , from given input mutation profiles, matrix M , where $M \approx P \times E$. Matrix M represents fraction of each mutation type in each sample, each column for one

sample and each row for one mutation type. Matrix P represents fraction of mutation type in each cancer process, each column for one cancer signature process and each row for one mutation type. Matrix E represents exposure, fraction of cancer signature process in each sample, each column for one sample and each row for one cancer signature process.

Visualizing

The purpose of this step is to decompose the tumor mutational signature components and visualize them together with mutation profiles of the tumor. The visualized profiles are generated in pdf format files consisting of 4 pages (Fig. 2), adapted from [19] but reimplemented in Python. The first page shows a pie chart representing mutational signature components in the profile. The second page displays the actual tumor profile. The third page displays the profile reconstructed from the predicted mutational

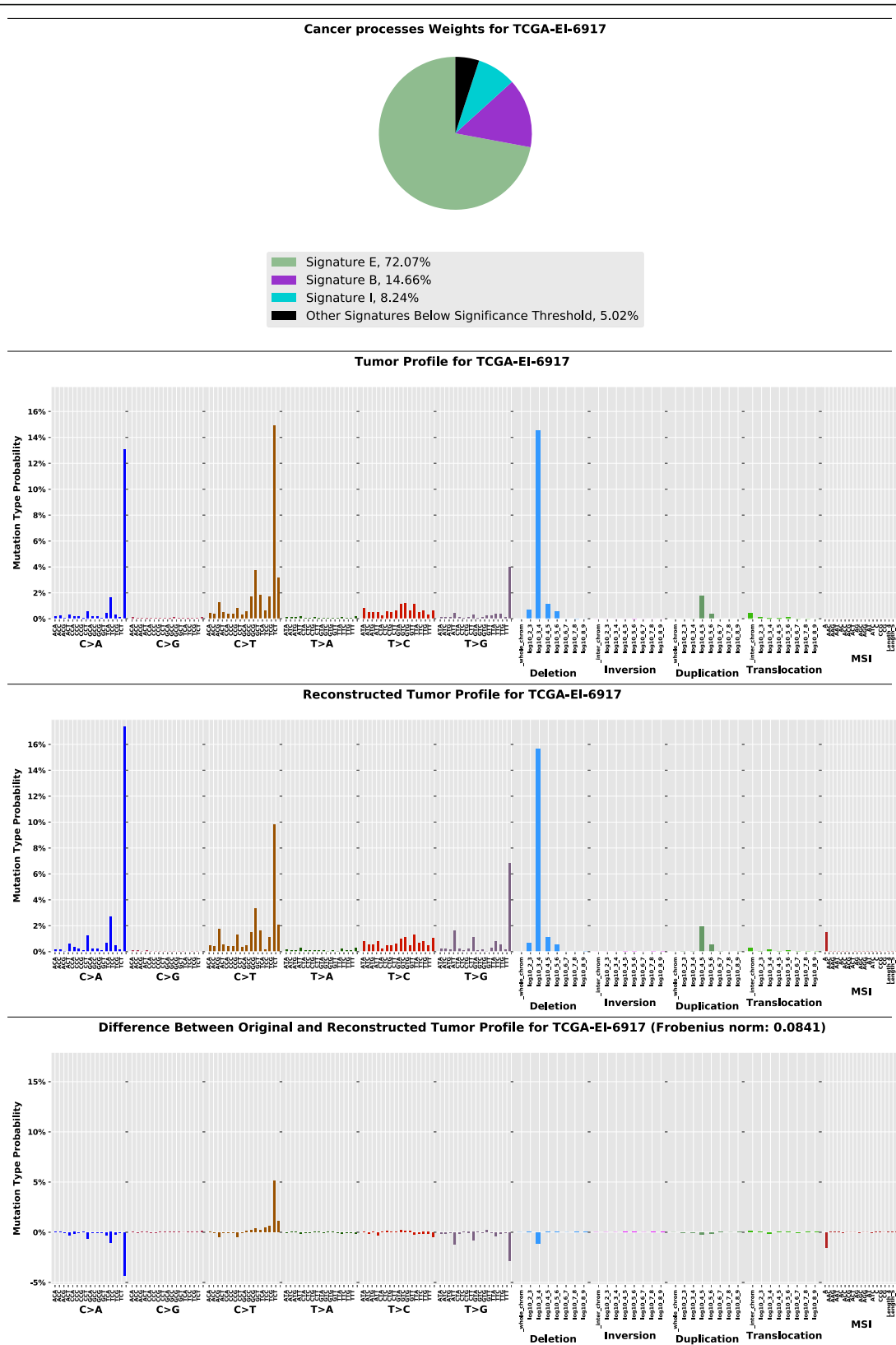


Fig. 2 Example of a visualized tumor profile

signatures component in the first page. The last page displays the prediction error, which is the difference between the actual profile and the reconstructed profile.

Evaluation

Benchmarking data

A cohort of 92 breast cancer cases, TCGA-BRCA, and 38 colorectal cancer cases, TCGA-COAD and TCGA-READ, from The Cancer Genome Atlas project (TCGA) was used to evaluate the tool. These 130 cases were all of TCGA breast cancer and colorectal cancer cases where we could obtain both whole genome and whole exome sequencing data from dbGAP. The data consists of tumor-normal pair whole genome sequence in BAM format and SNV exome sequencing data called by MuTect2 [20] in VCF format (See Additional file 1: Benchmarking samples). The whole genome sequencing data was used for benchmarking the SV- and the combined profiles. The whole exome sequencing data was used for replicating the previous result as well as benchmarking the combined profiles. The purpose of the colorectal cancer cases was for replicating signature SBS10 described in COSMIC. The cases were also used for evaluating the deciphering part of the package in the combined profile to see if the cases with signature SBS10 would have a combined signature with the SNV part similar to signature SBS10. The purpose of the breast cancer cases was for evaluating the SV-only profile and to see if there are any associations in patterns of SNVs and SVs in the combined profile. SBS3 has been associated with failure of DNA double-strand break-repair by homologous recombination, and the latter has been found in breast cancers.

Preprocessing of the benchmarking data

For SNVs, the downloaded somatic VCF files called by MuTect2 were in a format ready to be used for SNV profiling in the next stage. For SVs, a singularity container implementation of FindSV was used with downloaded aligned bam files. Briefly, balanced structural rearrangements were called by TIDDIT [16]. Copy number variations were called by TIDDIT [16] and CNVnator [17]. Somatic aberrations were called by subtracting normal tissue events from tumor tissue events using SVDB [16]. For MSI, the instability loci were called by msisensor [13] from aligned bam files.

Sample labeling

Clinical information on the cases was provided by TCGA. Annotation of somatic SNVs in genes of interest (See Additional file 1: Pathway and related genes) was done using VEP [21]. ClinVar is a clinical database used for variant interpretation [22].

Profile types and evaluation processes

We evaluated the tool using three profile types: SNV-only profile, SNV + SV + MSI profile and SV-only profile. The purpose of the SNV-only profile was to verify the SNV profiling- and the visualization processes by comparing the results with [23], and to setup a baseline for the other two comparisons. The combined, SNV + SV + MSI, profile was used to evaluate the efficiency of the profile when integrating several groups of mutations. The SV-only profile was used to evaluate the use of structural variation as the sole input compared to the combined profile.

Signature probabilities and signature IDs

There are three groups of mutational signatures described in this article, one for each profile type. The signature probabilities used in the evaluation of SNV-only profile were downloaded from [24]. Signature IDs are represented by COSMIC nomenclature, "signature SBS1", "signature SBS2", "signature SBS3", etc., the same as the ones in the downloaded signature probabilities file and the ones in [23]. For the combined profile, the signature probabilities were generated from the deciphering process. The signature IDs are represented using italic uppercase English alphabets, "signature A", "signature B", "signature C", etc. For the SV-only profile, the signature probabilities were also generated from the deciphering process. The signature IDs are represented using italic lower case English alphabets, "signature a", "signature b", "signature c", etc.

Sample-signature association

If a sample shows a decomposed fraction of more than 30% of a particular signature cancer process, it is defined as having an association between the signature and the sample.

Tumor mutation burden (TMB)

TMB was quantified as the total number of base substitutions in the SNV profile of an exome divided by 30 to give a roughly scaled estimate of SNVs per Mb exonic sequence.

Structural variation burden (SVB)

SVB was quantified as the total number of somatic structural variations in the sample, as determined from the SV profile.

Results

We used pyCancerSig to profile the benchmark data on three profile types, SNV-only, SNV + SV + MSI (combined), and SV-only. The SNV-only signatures used in the package verification were from COSMIC Mutational Signatures (v2 - March 2015) [25]. pyCancerSig was

used to decipher the combined- and the SV-only profile, resulting in nine (See Additional file 2: Figure 1) and twelve signatures (See Additional file 2: Figure 2) respectively. Tumor profiles were visualized for each profile type, together with their mutational signature components (Additional files 3, 4 and 5).

Verification using the SNV-only profile

POLE gene

Using SNV profiling by pyCancerSig, five cases were found to have signature SBS10, which has altered activity of the mutated *POLE* as proposed etiology. All of these cases were found to have somatic mutation in *POLE* and *POLD1* (Additional file 6). Moreover, these cases were found to have hypermutation with a median of 881 SNVs/Mbp (range 667–1686/Mbp), compared to the median 36/Mbp in the total data set and 22/Mbp in the tumors without *POLE/D1* variants. (Fig. 3), another property of tumors with bi-allelic loss-of-function variants in *POLE* [26].

DNA mismatch repair (MMR) genes

Signatures SBS6, SBS15, SBS20 and SBS26 were described to be associated with defective DNA mismatch repair. None of the cases in our cohort were associated with signature SBS15, SBS20 or SBS26. Using SNV profiling by pyCancerSig, two cases: TCGA-A6-6781 and TCGA-AA-A01R, were found to be associated with signature SBS6. Both of them were identified to be MSI-positive by MSIsensor. Both cases harbored pathogenic

germline (and in one case also somatic) truncating variants in *MSH2* or *MSH6* (Additional file 6).

Double strand break pathway

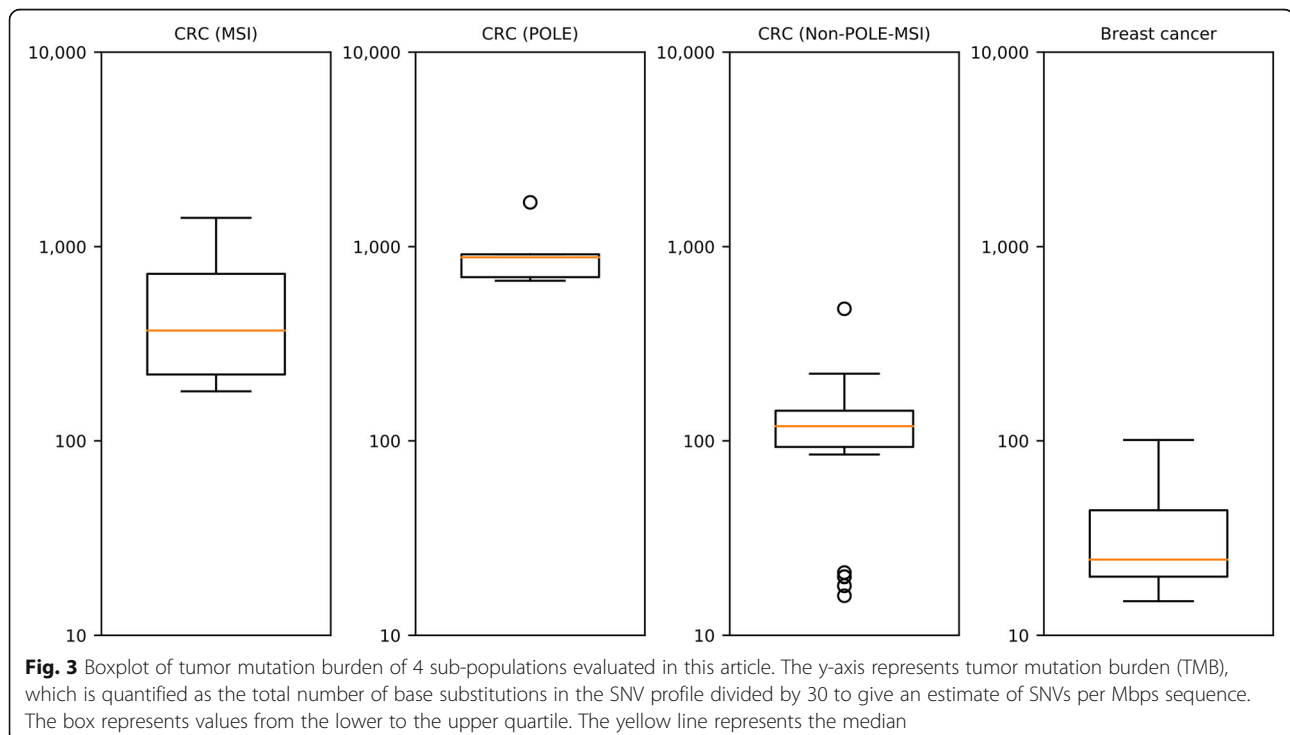
Signature SBS3 was described to be associated with failure of DNA double-strand break (DSB) repair by homologous recombination. Using SNV profiling by pyCancerSig, 78 breast cancer- and 10 colorectal cancer cases were found to be associated with signature SBS3. Among these, eleven breast cancer cases and one colorectal case had pathogenic truncating variants in DSB-pathway genes. Four cases have truncating variants in *BRCA1*, p.Val1734Ter, p.Glu23ValfsTer17 (both germline) and p.Glu720Ter, c.4739-1G>A (both somatic). One case has a somatic truncating variant in *BRCA2*, c.7805+1G>A. One breast cancer case had bi-allelic truncating variants in *BARD1*. The rest had truncating variants in *BLM*, *ATM* and *ATRX*. In addition, 30 other cases (27 breast cancer and 3 CRC) had in-frame or missense variants in 22 DSB genes (Additional file 6).

Comparison with deconstructSigs

Visual comparison with a leading tool specialized on profiling, deconstruction and visualisation (deconstructSigs) showed effectively identical results in the five cases with signature SBS10.

Evaluation using the combined profile

The optimal number of mutational signatures for the combined profile suggested by pyCancersig was nine



(See Additional file 2: Figure 1). Signatures *E* and *I* were only associated with colorectal cancer cases, while signatures *C*, *D*, *G* and *H* were only associated with breast cancer cases and signatures. *A*, *B* and *F* were found to be associated with both cancer types (Additional file 6).

POLE and POLD1 genes

Signature *E* exhibited a pattern of TCT > TAT, TCG > TTG and TTT > TGT, similar to those of signature SBS10, together with a pattern of structural deletion with size between 1 K–10 K. Signature *E* was associated with five cases and four of them were associated with signature SBS10. Three of these had known mutations in *POLE* and the fourth had a somatic truncating variant in *POLE* and a high TMB of 1686/Mbps. The fifth case with signature *E* had 26% signature SBS10, just below the 30% threshold applied in this paper and had a known mutation in *POLE* (p.Ser297Phe) as well as a high TMB (881/Mbps).

MMR genes

Signature *I* was the only signature with a pattern of MSI identified by msisensor [13]. It exhibited specific dominant characteristics in all mutation modes. SNVs were dominated by C > T substitutions, particularly on CG > TG backgrounds. The majority of SV events were structural deletion of 1 K–10 K. Signature *I* was associated with four colorectal cancer cases, two of which were the two cases associated with signature SBS6.

However, there were 2 other cases, TCGA-AD-6964 and TCGA-AZ-6601 with signature *I*. TCGA-AZ-6601 had both a germline truncating variant in *MSH6* (p.Lys537Ilefs*33) and a somatic in frame change in *MSH6* suggesting a mismatch repair defect, while no pathogenic variants were found in TCGA-AD-6964. (Additional file 6).

Double strand break pathway

From the aforementioned eleven breast cancer cases and one colorectal cancer case with signature SBS3 and pathogenic or likely-pathogenic variants in the DSB pathway, eleven were associated with seven different combined signatures (*A*, *B*, *C*, *D*, *F*, *G* and *H*) and one was not associated with any signatures. The median SV burden in these eleven cases (1004 events) did not differ significantly from the median of the entire cohort (Additional file 6).

Structural variation burden

The 10 cancer cases with highest SVB have a median of 10,016 events compared to the median 885 events in the total cohort. All ten were breast cancer samples and were divided between four signatures (*A*, *C*, *D* and *F*). (Additional file 6). To compare with the SNV-only profile,

all of these cases, except TCGA-B6-A0IJ, were associated with signature SBS3. TCGA-B6-A0IJ was not associated with any signature using the SNV-only profile.

Evaluation using the SV-only profile

The optimal number of mutational signatures for the SV-only profile suggested by pyCancerSig were 12 (See Additional file 2: Figure 2). Signatures *c*, *e*, *f*, *g* and *l* were associated with only breast cancer, while signatures *a*, *b* and *d* were associated with both cancer types.

POLE, POLD1 and MMR genes

Signature *a* exhibited a very unique pattern. 96% of the mutation in this signature was structural deletion with size between 1 K–10 K. This pattern was found in signature *E* and *I*. Eleven breast cancer- and fifteen colorectal cancer cases were identified to have this signature. Among them, five cases were the ones with combined signature *E* and four cases were the ones with combined signature *I*.

Double strand break

Out of the aforementioned eleven breast cancer cases and one colorectal cancer case with signature SBS3 and pathogenic or likely-pathogenic variants in the DSB pathway, four were associated with SV-only signatures: *c*, *d* and *e*. The colorectal cancer case with an *ATM* mutation was associated with signature *a*. One thing in common between these four signatures was that they each represented one specific structural event (more than 95%), signature *a* for structural deletion 1 K–10 K, signature *c* for structural inversion of size 1 K–10 K, signature *d* for structural deletion of size 10 K–100 K and signature *e* for structural duplication of size 10 K–100 K. The rest were not associated with any signatures (Additional file 6).

Structural variation burden (SVB)

The 10 cancer cases with the highest SVB were all breast cancers and were mainly associated with two SV-only signatures: *c* and *d*. Both of these had one main SV event at high frequency. (Additional file 6).

Runtime estimation

The amount of time needed for processing variants may depend on the size of data and configuration of the machine. The following performance was based on execution on the Uppsala Multidisciplinary Center for Advanced Computational Science computational cluster “bianca”, on a single Intel Xeon E5–2630 v3 core with 8 Gb RAM allocated. SNV profiler can process 7046 variants in 2 s (3523 variants/second). SV profiler can process 1755 variants in 0.1 s (17,550 variants/second). MSI profiler can process 65,535 loci in 0.3 s (218,450 loci/second). Deciphering a combined profile for 130

samples took 33 min. Visualizer took 9 s to generate the pdf file for one sample.

Discussion

In the present study, a command line tool was developed and tested on two different cancer types. We report a working tool, and study the use of a combined SV, SNV and MSI profile, contrasting to SNV or SV only profiles.

Comparison between signatures generated by the SNV-only profile and signatures generated by the combined profile

This study has shown that combining SNV, SV and MSI into one single profile can correctly identify mutational signature processes caused by altered activity of the DNA polymerases *POLE* and *POLD1*: signature *E*, and loss-of-function in MMR genes: signature *I*. Moreover, beside backward compatibility to SNV-only signature SBS6, signature *I* generated by the combined profile has identified two additional cases, TCGA-AD-6964 and TCGA-AZ-6601, with limited association to signature SBS6 (23 and 18%) which were also MSI positive according to msisensor. In principle, it could be possible to reduce the required level of association to signature SBS6 in order to increase the sensitivity of the SNV-only profile. However, this also reduces the specificity. In total, 10 colorectal cancer and 20 breast cancer cases displayed a contribution from signature SBS6, ranging from weight 6 to 42% of the total mutational profile. Four of them had weights in the range 20–30%, including one MSI and three MSS cases. Integrating SNV, SV and MSI into one single profile significantly increased the strength of the signal (signature *I*) due to a specific SV pattern (structural deletion with size between 1 K–10 K) and MSI.

COSMIC signature SBS3 is generally described as associated with deficiency in homologous recombination repair of double stranded breaks [27]. SNV-only profiles categorized 78 of 92 (85%) breast cancer cases as having contributions from signature SBS3. Thus, this signature very likely appears in cancers with heterogeneous etiologies. The eleven breast cancer samples with signature SBS3 and a defect DSB pathway due to truncating mutations in DSB genes did not cluster together in any specific combined or SV-only signature, thus adding SV information did not identify a DSB-specific pattern. In fact, the four *BRCA1/2*-mutated samples displayed a nonspecific SV pattern and did not contain a high number of SV events. This is in contrast to previous work where *BRCA1/2* deficient structural variation-only profiles were detected as clustered events [9, 28, 29]. The present tool does not differentiate between clustered and non-clustered variants and indeed the SV callers used

are not designed to detect such clusters, being essentially overlapping structural events.

Comparing SV-only signatures to combined signatures

Whether to use a combined or SV-only signature depends on the problem at hand. Combining an informative profile with a non-informative one may weaken the signal, depending on the amount of training data available. On the other hand, combining two informative profiles of two different mutation modes will increase accuracy, as in the cases with MMR gene mutation.

When analyzing SVB, a clear difference between cases with high SVB and low SVB (See Additional file 7: Figure 1) can be seen. The high-SVB cases each have at least one dominant structural variant type, accounting for more than 30% (Additional file 7: Figure 2), which is never present in any of the low SVB cases. The ten high SVB samples clustered in two different SV-only profiles, each showing one single dominant structural variant type. However, when SNV data was added, the SV signal was weakened with two samples clustering in signatures that lacked SV specific patterns. In this scenario, interpretation is simplified by the SV-only model.

On the contrary, it would be disadvantageous to use our SV-only profile alone to identify tumors with a *POLD1/E*-defect or a loss of function in MMR genes. The analyzed colorectal cancer cases with pathogenic mutation in *POLE*, *POLD1* and *MSI* were all identified to have signature *a*, which was characterized by structural deletion with size between 1 K–10 K. However, signature *a* was not specific to only these cases. There were eleven breast cancer and six colorectal cancer cases with signature *a* which did not have any coding *POLD1/E* or MMR variants. This suggested a possible advantage of combining two informative mutational modes in increasing accuracy.

Signatures representing unknown biological mechanisms

Mathematically, NMF results represent sets of variables that capture most of the variation in the data. Hence, the signatures represent the underlying biological processes of cancer only when the captured variations are the result of the oncogenic mechanism in the analyzed tumor. The types of variation seen in dysfunction of one mechanism do not necessarily aid explanation of a dysfunction in another mechanism. Increasing sample numbers and number of cancer types under study does aid discovering more underlying signatures. As the study progressed to include also CRC samples, additional signatures were indeed discovered. Application to a larger and more diverse data set was, however, outside the scope of the current study.

It is worth noting that at least half of the COSMIC signatures have yet to be ascribed an underlying process.

Adding sensors to detect variations caused by different cancer mechanisms will make it possible for the signatures to closer represent the underlying biological processes. Many different sensors of DNA change could be incorporated, and have indeed been used in different studies: multi nucleotide variants, small indels, clustered/non-clustered SVs, breakpoint characterization sensors, including microhomology and mutations clustered around the breakpoints. Many different structural variant callers have been developed, and our results only reflect only one possible, though proven, combination of such. As an example, previous work has advised that TMB be used for tumor classification, genetic testing, and clinical trial design [30]. Including TMB as part of a molecular profile may strengthen accuracy. While we use TMB to characterise benchmark samples, different bioinformatic workflows can result in different TMB values [31]. Second, TMB analysis compares the number of mutations in one tumor to the others. But signature analysis compares relative incidences of mutation types within a tumor, resulting in a proportion of mutation types. Research on how TMB can be included in a form of a proportion is needed to ensure that TMB of hypermutated tumors does not overpower other mutation types.

Having more quantitative information of omics data may also strengthen the power of the model. Other omics data that can be included in the model include methylation data, RNA expression data and protein data. Depending on the amount of correlation between different sensor variables, signatures may be useful biomarkers even if the variant type repertoire is far from complete.

Implementation

While it is being increasingly recognized that mutational signatures are useful for the analysis of tumors, they are not yet in use in many sequencing centers. We believe this is partly due to the lack of a modern, command line implementation of any such tools. Most available tools operate in the R environment, well suited for prototyping of mathematical methods, or by web interface. While conversion programs or wrappers can be written to supply such tools with data from standard files, this is perhaps not given priority. Many tools expect data in domain custom or rare formats. Sequencing facilities prefer to operate with command line interface tools that can be readily included into pipelines, and a small number of standardized file formats. One notable exception is EMu [32] which also presents a command line interface. The EMu tool accepts copy number variation data, but relies on custom file formats for data input. It does not handle balanced structural variation or MSI. We opted for an implementation in python, with a command line interface accepting standard genomics file formats (vcf, bam) to be suitable for

integration into standard pipelines. pyCancerSig also produces pdf reports. pyCancerSig requires whole genome sequencing input of both tumor and germline DNA in bam and VCF format. pyCancerSig optionally accepts COSMIC/WTSI format signatures. No particular constraint is made on SNV or SV VCF input, but SNV calls compatible with data used for Alexandrov et al. [6] are accommodated, and the structural variant calling pipeline (FindSV) used in the present study is also documented for ease of use and reproduction.

Choice of feature extraction model

Developing a feature extraction method that would reasonably represent the global mutation profile was challenging. The mathematical model, NMF, used in this study took a matrix M as an input and returned matrix P , where $M \approx P \times E$. The model treated every number in the matrix equally, regardless of the type of variation represented.

Relative differences in the number of mutation events in each mutation group played an important role in choosing the extraction method. It is known that mutation burden varies between different cancer types [33]. The number of somatic indels are significantly different between MSI and non-MSI cases [13]. In this study, we decided to limit the total percentage of features from each mutation group. For instance, the total fraction of base-substitution will always be 70% regardless of the relative difference between the number of base substitution events compared to copy number variation events. This prevented a very large number of mutations from one mutation group, like hypermutation due to defect MMR or the enormous amount of structural inversion events found in cases with signature D to overpower the others.

The numerical encoding of each type of structural event sensor variable also had an impact on the design. On the surface, the events might look equivalent. In detail, the quantitative values encode different molecular events. In structural deletion or structural duplication, the size of an event was the total number of nucleotides lost or gained. But for translocation events, the size of an event was a distance between the original position and the new position. If a piece of 100 bp DNA is translocated from a coordinate at 20 Mb to a coordinate at 25 Mb, in the same chromosome, the size of this event will be 5 Mb, instead of 100 bp. If the event is to a different chromosome, the size will be set to infinite. Moreover, there were also differences behind the quantitative value within each mutation type. For example, larger structural deletion events are not as numerous as their smaller counterparts. It is common to see 100 events of structural deletion with size between 100 and 1000 bp but it is impossible to see the same number of events for whole chromosome deletions. In fact, having only 1 or 2

whole chromosome deletion events is a pathogenic event. At first, we attempted a logarithmic scale for the quantitative values of structural events. However, in order to make SVs comparable with the base substitution profile and in order to keep the profile representation as a percentage (in line with previous publications [6], the present study does not use logarithmic values. This resulted in a risk that certain mutation types can overpower the others within the group, like in cases with signature *D*. Moreover, events that are rare but pathologic, such as the loss of entire chromosomes or chromosome arms, were also underpowered by this choice.

Usage of tumor profile

Cancer signatures can be applied in both supervised-and/or unsupervised learning models. Using cancer signatures in a supervised learning model can result in a high accuracy classifier, which is very efficient in answering questions with a limited set of answers. One such classifier is HRDetect [9], which can identify *BRCA1/BRCA2*-deficient tumors with 98.7% sensitivity.

However, the implementation in this study was in an unsupervised fashion. This means that the cases have not been labelled, including cancer type, cancer-related pathway, cancer-related genes or tumor heterogeneity prior to incorporation in the model. As a result, the model has brought out the patterns of profile that were commonly found in the cases, regardless of whether there were one or more cancer processes in the tumors, whether the processes are in known or unknown pathways, whether the processes are in known or unknown genes, or whether the damage was caused by one or more sources of DNA damage.

The unsupervised method aids interpretation by reducing the dimensionality of the data and presenting it in a way that humans can recognize. Therefore, this model can be used to support molecular classification of tumors; for instance, to compare a tumor to previously analyzed ones in order to delineate possible cancer processes. Given that the model has been trained with enough data, they can be used as supportive evidence in a clinic by associating a new unsolved case with signatures from solved cases, and perhaps in the future aid medical decisions regarding personalized therapy. We envision the use of an unsupervised model with multiple sensor types of particular use in the exploration of disease where no supervised model is yet available, or when investigation of the relation between signatures of multiple modes of mutation types is desired. One way to make use of visualization of this clustering method is however to use it as an intermediate result to infer the mechanism behind the tumor development and to

suggest potential targets for therapy based on the suggested profile. For example, in MSI cases, if we profile a tumor and its profile is similar to sample TCGA-A6-6781 or TCGA-AA-A01R, the profile can be a supporting evidence suggesting a similar causative mechanism, in this case defect MMR genes. Also, since the profiles of the MSI cases suggest similar molecular mechanisms caused by the mutation, this might be used as evidence that similar treatment on cases with similar profiles is likely to have similar effect. Thus using combined mutation profiles may identify groups of cancers that may profit from specific therapies as has been previously suggested [34].

Profile interpretation

This software package includes a visualization tool, “cancersig signature visualize”, that can help in clinical interpretation. The visualization output files are in both graphical and numerical representation.

As graphical format in a pdf file (Fig. 2), the pie chart suggests mutational signature components found in the tumor together with their percentage. For suggesting similarity between two samples, the tumor profile, the second figure in the pdf file, and the pie chart of the two samples can be inspected. If the two look similar, it may suggest similar mechanisms of causation and potentially inform diagnosis and treatment. The last figure in the pdf file can help determine error and uncertainty. The higher the error, the more likely there are unidentified mutational signature components in the tumor.

As numerical representation, the tool generates an output file called “normalized_weights.txt”. This file contains the percentage of mutational signature components from all of the samples from one run (the same percentage numbers of those in the pie charts). It is useful for systematic interpretation for multiple tumors and can be worked on with basic data processing operations, like “sort”, “not zero”, “more than”, etc.

Workflow flexibility

The aim of this python package was to have a complete integrated set of tools written in Python to work with and answer questions related to mutational signatures of multiple variant types. Even though the package was benchmarked with the recommended variant callers, other bioinformaticians may use different variant callers which may produce output in different formats, which may not be compatible with our tools. So, we have designed our workflow (Fig. 1) to be flexible and allow the bioinformaticians to not have to change their callers in order to use our tools. Most tools in the package were designed to be replaceable, particularly the profiling. For example, if a user uses another SV caller with entirely different format, the user can replace the SV profiler

with the parser written by the user. We have also provided examples input/output for each tool in the package. As long as the parser can generate the output in the same format as the example, the parser should work perfectly.

Conclusions

This pattern-based study has shown that applying an unsupervised learning method on joint information from multiple variant types can identify the cases suggested by the SNV-only profile and demonstrated that integrating SV and MSI into the profile can strengthen previously unclear signals. This illuminated the association within and between the mutation modes. The tool presented, pyCancerSig, accepts common file formats, and is easily incorporated into a genome center pipeline.

Availability and requirements

Project name: pyCancerSig

Project home page: <https://github.com/jessada/pyCancerSig>

Operating system(s): Platform independent

Programming language: Python

Other requirements: Python 3.0 or higher

License: GNU v.3.0.

Any restrictions to use by non-academics: None.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12859-020-3451-8>.

Additional file 1.

Additional file 2.

Additional file 3.

Additional file 4.

Additional file 5.

Additional file 6.

Additional file 7.

Abbreviations

DSB: Double-strand break; Mbp: Mega basepair; MMR: Mismatch repair genes; MSI: Microsatellite instability; MSS: Microsatellite Stable; NMF: Non-negative matrix factorization; SNV: Single nucleotide variation; SV: Structural variation; SVB: Structural variation burden; TCGA: The Cancer Genome Atlas project; TMB: Tumor mutation burden

Acknowledgements

We thank Annika Lindblom for help in applying for access to dbGaP TCGA data.

Authors' contributions

JT applied for computational resources, gathered data, developed and evaluated the software, helped design the study and was a major contributor in writing the manuscript. JE was involved in structural variant analysis. ET performed pathogenicity interpretation of cancer-related genes and was involved in writing the manuscript. DN designed the study and was involved in designing software architecture and command line interface, and was involved in writing the manuscript. All authors approved of the final manuscript.

Funding

This study was supported by grants provided by the Stockholm County Council (ALF project) and Swedish cancer society. Open access funding provided by Karolinska Institute.

Availability of data and materials

The datasets used for evaluation of this software are available at <https://portal.gdc.cancer.gov>.

Ethics approval and consent to participate

This study utilized previously obtained TCGA data, made available by dbGap after ethical review of our dbGap project #14084; and used under approval of this request [#55693–2].

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 30 September 2019 Accepted: 10 March 2020

Published online: 03 April 2020

References

1. Tomasetti C, Li L, Vogelstein B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*. 2017;355(6331):1330–4.
2. Chen J, Silver DP, Walpita D, Cantor SB, Gazdar AF, Tomlinson G, et al. Stable interaction between the products of the BRCA1 and BRCA2 tumor suppressor genes in mitotic and meiotic cells. *Mol Cell*. 1998;2(3):317–28.
3. Saffran WA, Greenberg RB, Thaler-Scheer MS, Jones MM. Single strand and double strand DNA damage-induced reciprocal recombination in yeast. Dependence on nucleotide excision repair and RAD1 recombination. *Nucleic Acids Res*. 1994;22(14):2823–9.
4. Papadopoulos N, Nicolaides NC, Wei YF, Ruben SM, Carter KC, Rosen CA, et al. Mutation of a mutL homolog in hereditary colon cancer. *Science*. 1994;263(5153):1625–9.
5. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. *Nature*. 2007;446(7132):153–8.
6. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*. 2013;3(1):246–59.
7. Poulos RC, Wong YT, Ryan R, Pang H, Wong JWH. Analysis of 7,815 cancer exomes reveals associations between mutational processes and somatic driver mutations. *PLoS Genet*. 2018;14(11):e1007779.
8. Polak P, Kim J, Braunstein LZ, Karlic R, Haradhavala NJ, Tiao G, et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat Genet*. 2017;49(10):1476–86.
9. Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med*. 2017;23(4):517–25.
10. Terdiman JP, Gum JR Jr, Conrad PG, Miller GA, Weinberg V, Crawley SC, et al. Efficient detection of hereditary nonpolyposis colorectal cancer gene carriers by screening for tumor microsatellite instability before germline genetic testing. *Gastroenterology*. 2001;120(1):21–30.
11. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N Engl J Med*. 2015;372(26):2509–20.
12. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science*. 2017;357(6349):409–13.
13. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics*. 2014;30(7):1015–6.
14. Omichessan H, Severi G, Perduca V. Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance. *PLoS One*. 2019;14(9):e0221235.
15. FindSV. <https://github.com/J35P312/FindSV>. Accessed 24 Sept 2019.
16. Eisfeldt J, Vezzi F, Olason P, Nilsson D, Lindstrand A. TIDBIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. *F1000Res*. 2017;6:664.

17. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 2011;21(6):974–84.
18. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics.* 2015;32(8):1220–2.
19. 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 Biol.* 2016;17:31.
20. 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.* 2013;31(3):213–9.
21. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl variant effect predictor. *Genome Biol.* 2016;17(1):122.
22. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018;46(D1):D1062–D7.
23. Signatures of Mutational Processes in Human Cancer (v2 - March 2015). https://cancer.sanger.ac.uk/cosmic/signatures_v2. Accessed 24 Sept 2019.
24. Patterns of mutational signatures. https://cancer.sanger.ac.uk/cancergenome/assets/signatures_probabilities.txt. Accessed 24 Sept 2019.
25. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in Cancer. *Nucleic Acids Res.* 2019; 47(D1):D941–D7.
26. Shlien A, Campbell BB, de Borja R, Alexandrov LB, Merico D, Wedge D, et al. Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. *Nat Genet.* 2015; 47(3):257–62.
27. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell.* 2012;149(5):979–93.
28. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature.* 2016;534(7605):47–54.
29. Letouze E, Shinde J, Renault V, Couchy G, Blanc JF, Tubacher E, et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat Commun.* 2017;8(1):1315.
30. Campbell BB, Light N, Fabrizio D, Zatzman M, Fuligni F, de Borja R, et al. Comprehensive Analysis of Hypermutation in Human Cancer. *Cell.* 2017; 171(5):1042.
31. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* 2017;9(1):34.
32. Fischer A, Illingworth CJ, Campbell PJ, Mustonen V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol.* 2013;14(4):R39.
33. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. *Science.* 2013;339(6127):1546–58.
34. Fabrizio DA, George TJ Jr, Dunne RF, Frampton G, Sun J, Gowen K, et al. Beyond microsatellite testing: assessment of tumor mutational burden identifies subsets of colorectal cancer who may respond to immune checkpoint inhibition. *J Gastrointest Oncol.* 2018;9(4):610–7.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

