Published online 1 November 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 4 **1**
https://doi.org/10.1093/nargab/lqab099

MutScape: an analytical toolkit for probing the mutational landscape in cancer genomics

Cheng-Hua Lu[1,†], Chia-Hsin Wu[2,†], Mong-Hsun Tsai[3,4], Liang-Chuan Lai[3,5,*] and Eric Y. Chuang[1,2,3,6,*]

[1]Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan, [2]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 10617, Taiwan, [3]Bioinformatics and Biostatistics Core, Centers of Genomic and Precision Medicine, National Taiwan University, Taipei 10055, Taiwan, [4]Institute of Biotechnology, National Taiwan University, Taipei 10672, Taiwan, [5]Graduate Institute of Physiology, College of Medicine, National Taiwan University, Taipei 10051, Taiwan and [6]Master Program for Biomedical Engineering, China Medical University, Taichung City 40402, Taiwan

Received May 24, 2021; Revised September 28, 2021; Editorial Decision October 04, 2021; Accepted October 28, 2021

ABSTRACT

Cancer genomics has been evolving rapidly, fueled by the emergence of numerous studies and public databases through next-generation sequencing technologies. However, the downstream programs used to preprocess and analyze data on somatic mutations are scattered in different tools, most of which require specific input formats. Here, we developed a user-friendly Python toolkit, MutScape, which provides a comprehensive pipeline of filtering, combination, transformation, analysis and visualization for researchers, to easily explore the cohort-based mutational characterization for studying cancer genomics when obtaining somatic mutation data. MutScape not only can preprocess millions of mutation records in a few minutes, but also offers various analyses simultaneously, including driver gene detection, mutational signature, large-scale alteration identification and actionable biomarker annotation. Furthermore, MutScape supports somatic variant data in both variant call format and mutation annotation format, and leverages caller combination strategies to quickly eliminate false positives. With only two simple commands, robust results and publication-quality images are generated automatically. Herein, we demonstrate the ability of MutScape to correctly reproduce known results using breast cancer samples from The Cancer Genome Atlas. More significantly, discovery of novel results in cancer genomic studies is enabled through the advanced features in MutScape. MutScape is freely available on GitHub, at https://github.com/anitalu724/MutScape.

INTRODUCTION

Next-generation sequencing technologies are developing rapidly, leading to substantial improvements in cancer genomic research (1). Among these, whole-genome sequencing (WGS) and whole-exome sequencing (WES) are two major paradigms that enable researchers to uncover mutational landscapes in the least amount of time. In recent years, The Cancer Genome Atlas (TCGA), a publicly available genomics database, has released a large amount of somatic mutation data for researchers, leading to improvements in cancer diagnosis, treatment and prevention. With the sudden augmentation of studies, a surfeit of large-scale mutation data, comprising single-nucleotide variants and small insertion/deletions, has been generated for advancing our understanding of cancer etiology and biology. Those generated data offer opportunities for several useful analyses, such as significantly mutated gene (SMG) detection (2) and mutational signatures (3). However, these bioinformatics analyses require complicated statistical approaches. Additionally, researchers find it difficult to visualize these diverse and feature-rich data from genomic studies. Even though several tools and software packages for probing genomic data have already been developed, distinct formats of input files still make analysis cumbersome. Tools such as Maftools (4) offer various analyses in a single R package, while VIVA (5) supports data processing. However, they require still other tools to complement their inability to

*To whom correspondence should be addressed. Tel: +886 2 33663660; Fax: +886 2 33224179; Email: chuangey@ntu.edu.tw
Correspondence may also be addressed to Liang-Chuan Lai. Tel: +886 2 23123456 (Ext 88241); Fax: +886 2 33224179; Email: llai@ntu.edu.tw
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Present addresses:
Eric Y. Chuang, Bioinformatics and Biostatistics Core, Centers of Genomic and Precision Medicine, National Taiwan University, Taipei 10055, Taiwan.
Liang-Chuan Lai, Bioinformatics and Biostatistics Core, Centers of Genomic and Precision Medicine, National Taiwan University, Taipei 10055, Taiwan.

filter out false-positive mutations and perform data transformation. Also, these tools accept either mutation annotation format (MAF) or variant call format (VCF) as inputs, but not both. With the cost reduction of sequencing technologies, researchers could design more large-scale studies for hundreds or thousands of cancer patients than past. Therefore, the growth of sequencing data and the emergence of various analysis approaches highlight the need for a comprehensive, integrated toolkit for efficiently studying cancer genomics.

To cope with these problems, we developed a useful Python package, which we named 'MutScape'. MutScape provides two modules for preprocessing input data and performing comprehensive analyses and visualizations. This package is intended for WGS, WES or gene panels. It facilitates cancer genomic studying and enables researchers to efficiently perform downstream analysis for cancer cohorts when obtaining the mutation data from various variant callers, even for newcomers getting a hang of how these analyses are. To be user-friendly, it accepts large amounts of VCF and MAF as input for the data preprocessing module. With only two simple commands and a single customized tab-separated values (TSV) file, which records the paths of the input folders and the criteria used for data preprocessing, MutScape will automatically implement data filtering, combination and transformation. All preprocessed files are ultimately combined into a single MAF summary file for easily performing a multitude of downstream analyses and a wide variety of visualizations. The analysis module is divided into nine main functions that provide SMG detection and mutational signature analysis, as well as several other computational and statistical analyses, such as homologous recombination deficiency (HRD) score (6) and chromosome instability (CIN) score (7). Moreover, MutScape can be used to generate multiple high-quality statistical graphics in order to assist researchers with further exploration. Finally, usage of MutScape is publicly available as an open-source Python package in the GitHub repository (https://github.com/anitalu724/MutScape).

MATERIALS AND METHODS

The functions of MutScape are separated into two main modules (Figure 1): data preprocessing and analysis and visualization. The specifications for both modules and their key functions are described below.

Data preprocessing module

Loading data. MutScape can handle a large number of VCF files or MAF files as input in a single process. For VCF files as input data, users must submit a limited-format TSV file, which lists input paths and simple parameters for data preprocessing. In this package, we can deal with five TCGA-adopted variant callers and one popular, commercial variant caller—MuSE (8), Mutect2 (9), SomaticSniper (10), Strelka2 (11), VarScan 2 (12) and DRAGEN (13)—so that users can choose several callers of VCF files for filtering and combination. For MAF files as input data, users should also enter a one-column TSV file, which includes every path of the MAF files, for downstream analysis and visualization.

VCF filtering. The VCF filtering step, which is optional, provides four types of filters (Supplementary Figure S1). The filter function named *genome interval* (GI) is used to choose specific intervals of chromosomes. Users can enter a list of genomic positions or specific chromosomal regions to select the mutations. The *caller information* (CI) filter is utilized to sift valid data based on the information in the VCF file. Users can enter a self-defined criteria list that is composed of some items from the VCF—the total depth of reads (DP), allelic depth (AD) of the reference and alternative alleles, mutant allele fraction (AF), and tumor and normal log odds (LOD) scores. We process the data based on the variant caller for each VCF file. Specifically, the LOD scores are only calculated and recorded in VCF files of Mutect2. In this step, we exclude the variants that are smaller than the provided criteria. The so-called *PASS* (PA) filter is employed to extract mutations, for which the 'FILTER' column presents 'PASS' in the VCF file. To filter false-positive calls from formalin-fixed paraffin-embedded tissues, the *artifact variant* (AV) filter uses the absolute value, defined as

$$\frac{F1R2_{\text{alt}} - F2R1_{\text{alt}}}{F1R2_{\text{alt}} + F2R1_{\text{alt}}},$$

as the threshold that is only for Mutect2 VCF files (14). Mutations with values below the threshold are discarded. Furthermore, users can enter a VCF or TSV file containing the chromosome, position, reference allele and alternative allele information from each variant as an accept list or reject list of variants. MutScape will forcibly retain variants present in the accept list during data preprocessing, whereas it will exclude variants recorded in the reject list (Supplementary Figure S2). All of these filters are simultaneously applied to the data for the creation of more reliable and robust results.

VCF combination. As a unique feature in MutScape, the purpose of VCF combination is to combine VCF files with distinct variant callers from the same sample. During this step, we merge overlapping mutations that have identical values in the 'CHROM', 'POS', 'REF' and 'ALT' columns of the VCF files. For each mutation recorded in the combined VCF, we add two extra types of information—'CALLS' and 'REJECT'—in the 'INFO' column. Any variant callers that identified the sequence data as having a mutation are recorded in the 'CALLS' information. For instance, if a candidate mutation is identified only by MuSE, the 'CALLS' information will be recorded as 'MuSE', whereas if another candidate mutation is detected by both MuSE and Mutect2, the 'CALLS' information will be recorded as 'MuSE_Mutect2' (top right in Supplementary Figure S3). In contrast, 'REJECT' information is created to record names of variant callers that the mutation failed to pass. In the input TSV file, the 'At Least CALLS' column states the minimum number of variant callers, whereas the 'At Most REJECT' column mentions the maximum number of variant callers recorded in REJECT. Using the input TSV file, users can define the threshold of CALLS and REJECT to include in the combined file (bottom right in Supplementary Figure S3). Moreover, users should utilize unfiltered VCFs as inputs when employing the 'REJECT' information to filter variants that failed

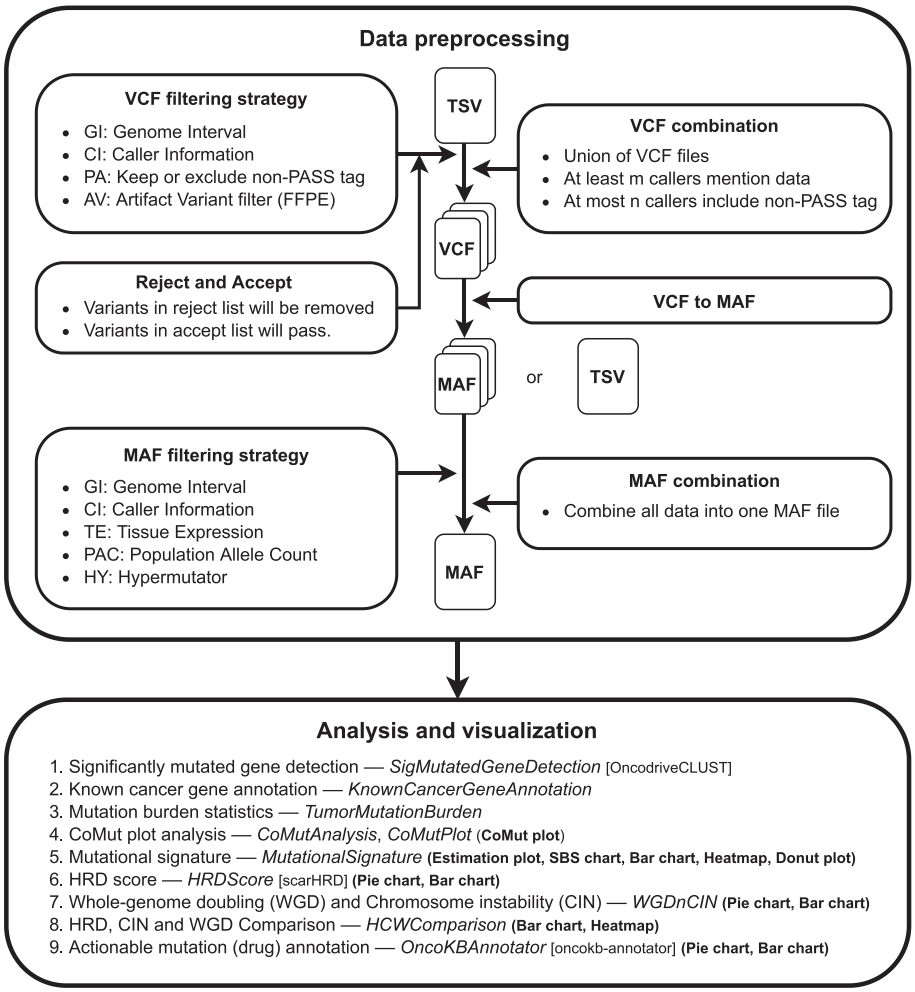

Figure 1. Overview of the MutScape toolkit. Block headers define the two available modules, Data preprocessing and Analysis and visualization. **Data preprocessing module**: The main workflow starts with a TSV file, which records the input paths of either VCF or MAF files and the criteria of the filtering and combination strategies. The VCF inputs are transformed to a combined, filtered MAF file, which is the basis of the analysis and visualization module. **Analysis and visualization module**: This consists of functions to perform common and advanced analyses in cancer genomics and generate publication-ready plots and tables (shown in bold). The square brackets indicate the libraries or packages we implemented in MutScape. A simple specification and the related function (italics) are shown. HRD: homologous recombination deficiency

to pass the threshold. After combination, only one VCF file will remain for each sample.

VCF to MAF. VCF transformation to MAF is implemented by the vcf2maf utility (https://github.com/mskcc/vcf2maf), which processes variant annotation and transcript prioritization.

MAF filtering. In addition to VCF filtering, MutScape also includes functions to filter MAF files (Supplementary Figure S4). Five diverse filters are illustrated as follows: The GI filter is similar to the one in VCF filtering, and the usage and input types are all identical. In the CI filter, users should enter a list with only two values—DP and AD. The *tissue expression* (TE) filter is used to exclude non- or low-expressed gene mutations in the specific tissue. The usage also supports self-defined criteria and gene expression data. The *population frequency* (PF) filter is used to identify data whose 'FILTER' column does not present 'common_variant', which is annotation by population allele frequency from the ExAC database (15). The *hypermutator* (HY) filter is a strategy that avoids statistical bias from hypermutators. If samples have mutation counts larger than the entered threshold, their data will be removed from downstream analysis.

Analysis and visualization module

This module simultaneously provides numerous analysis methods and several plotting functions to produce readable and publication-quality plots and tables. A comutation (CoMut) plot is generated using the comut Python module (16), SMGs are detected using the OncodriveCLUST Python package (17), an estimation plot is produced by Scikit-learn nonnegative matrix factorization (NMF), Nimfa is a Python library for NMF, HRD status is evaluated using the scarHRD R package (18), known cancer genes are annotated using the Cancer Gene Census database (19) and known actionable genes are summarized based on the levels of evidence in the OncoKB database (20). Plots

generated to demonstrate results from the analysis module, such as bar charts, heatmaps and donut plots, among others, are plotted using matplotlib and the seaborn Python library.

Tumor mutation burden

Tumor mutation burden (TMB) is defined as the total number of mutations per megabase (21). The mutational types included in the TMB statistics are missense mutation, silent mutation, splice site mutation, nonstop mutation, nonsense mutation, frameshift mutation, in-frame mutation and translation start site mutation. The number of *synonymous* mutations is the sum of the data that have the 'silent' mutation type, while the number of *nonsynonymous* mutations includes the rest of the data.

Large-scale genomic events

The evaluations of three large-scale genomic events are described below in detail.

Homologous recombination deficiency. A genome aberration-based scoring system (HRD score) derived from the unweighted sum of the HRD loss of heterozygosity (HRD_LOH) score (22), the telomeric allelic imbalance (Telomeric_AI) (23) and the large-scale state transition (LST) score (24) is employed to assess the underlying tumor HRD (6). The HRD_LOH score is defined as the number of LOH events >15 Mb without covering the whole chromosome. The Telomeric_AI score is the number of regions with allelic imbalance, which extend toward the telomeric ends of a chromosome without crossing the centromere. The LST score is the total number of breakpoints between regions of at least 10 Mb, with a distance between them <3 Mb. Based on the allele-specific copy numbers for each region as inputs, the extent of HRD is quantified using scarHRD (18). A predefined HRD threshold of ≥42 is employed to distinguish the HRD phenotype from nondeficient tumors (6).

Whole-genome doubling. Patients were considered to have undergone whole-genome doubling (WGD) if >50% of their autosomal genome had a major copy number ≥2 (25).

Chromosomal instability. CIN is a broad concept that encompasses a wide range of chromosome-level abnormalities. The CIN burden is defined as the proportion of the genome's length that is affected by copy number alterations (CNAs) (7) and is given by

$$\mathrm{CIN} = \frac{\sum_{i=1}^{n} u_i}{L},$$

where L is the total length of the autosome and u_i represents the altered length in CNA i.

Example datasets

The datasets for the CoMut plot analysis were obtained from TCGA breast cancer database (https://www.cancer.gov/tcga). Only 50 samples were randomly chosen for better visualization and they were simulated as 19 patients with metastatic cancer for the need of the demonstration of CoMut plot analysis. VCF files used to assess the results of mutational signatures were from a published study (26), and MutScape accurately reproduced the results in it. The dataset of 30 paired esophageal adenocarcinoma genomes before and after neoadjuvant chemotherapy from the previous study was collected to demonstrate that MutScape can uncover additional noteworthy findings (27). Simulation datasets for displaying other analysis functions originated from rearrangement of encrypted clinical data in Taiwan. All input data are provided on the GitHub website of MutScape.

RESULTS

Mutational landscape discovery with the CoMut plot

Since the development of WGS and WES, a large amount of the mutational landscape has been uncovered and requires a comprehensive visualization to present the analysis results. Dees *et al.* had come up with the concept of SMGs, which describes genes that show a conspicuously higher mutation rate than expected by chance (2). In MutScape, we provide the function named *SigMutatedGeneDetection* to detect SMGs. Additionally, TMB is also a common cancer hallmark in cancer genomic studies (21). Thus, we offer a class called *TumorMutationBurden* to calculate TMB automatically. A general representation of these results is a CoMut plot, which demonstrates mutation status with clinical and genomic characteristics by sample level.

To demonstrate the performance of MutScape, we used the mutation data from a TCGA breast cancer cohort (N = 50) with simulated CNA data (Figure 2). The top bar chart (Figure 2A) showing the mutational classification was based on the results of TMB statistics. In these data, we can see that most of the TMB scores were below 300, while only a single sample had a higher burden and was classified as a hypermutator. Generally, samples with high TMB may be suitable for immunotherapy (21). The mutational signature bar chart (Figure 2B) shows sample-level relative proportions of four different signatures. These signatures represent origins of mutation (see the 'Mutational signature analysis' section for more details). In this subplot, it is obvious to see that signature 2 has the highest proportion, followed by signature 3. The purity heatmap (Figure 2C) indicates the original sample purity, which tells the validity and reliability of data. It is clear that most of the sample purities are high, which promises reliable analyses of these samples. The oncoprint plot (Figure 2D) indicates the mutational type-specific genes by sample. These specific genes are mostly the result of SMG detection or of known cancer gene annotation. In this cohort, we can see that the SMGs include *MAP3K1*, *CDH1*, *TP53* and *PIK3CA*. Notably, the mutated *PIK3CA* genes in the plot are all missense mutations. The CoMut plot can also be used to visualize and summarize CNAs (Figure 2E). This plot clearly presents the alteration type of distinct genes in each patient. Candidate genes for this plot are identified using GISTIC2 (28). In this cohort, *ERBB2* is the most common CNA. Furthermore, WGD, which involves the duplication of a com-

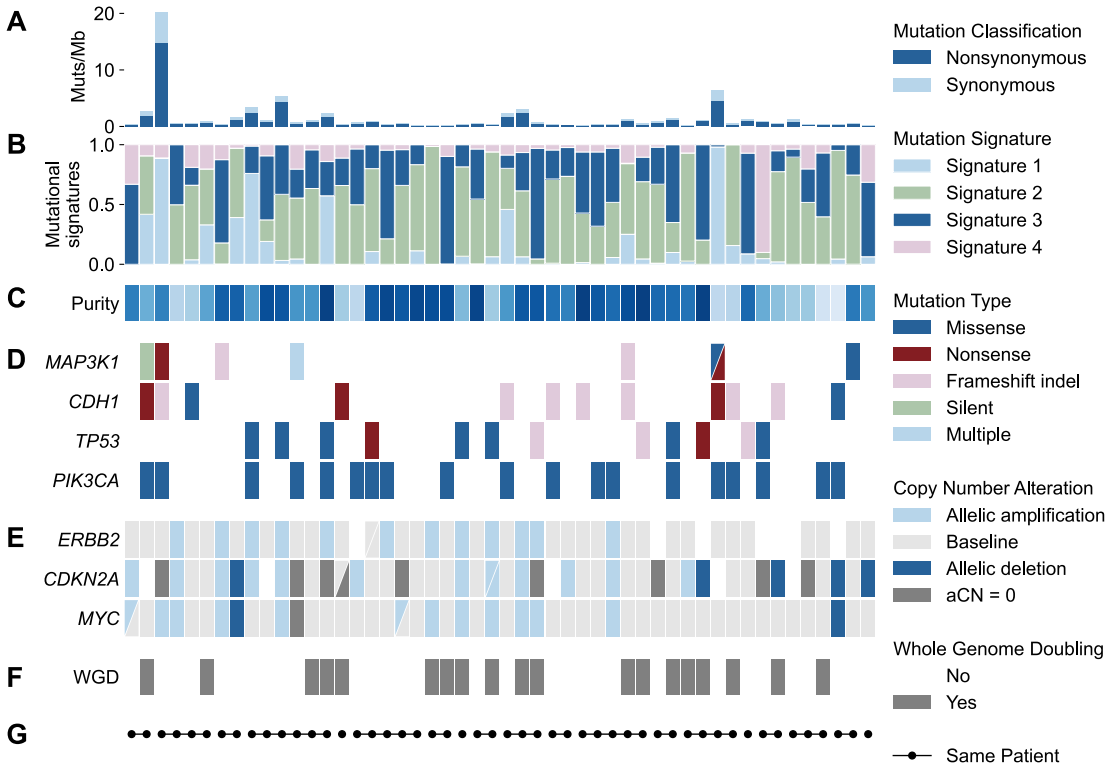

Figure 2. CoMut plot generated by the MutScape analysis and visualization module using a TCGA breast cancer dataset. For better visualization, only 50 samples are shown. The CoMut plot illustrates mutational profiles for each sample. (**A**) The top bar chart displays the mutation burden for each sample. (**B**) The stacked bar chart shows the relative proportion of the mutational signatures in each sample. (**C**) The purity heatmap manifests the degree of purity for every sample. The darker blue blocks indicate higher purity. (**D**) The oncoprint plot demonstrates the mutation type of specific significantly mutated genes for each sample. (**E**) The copy number alteration heatmap illustrates the alteration type for commonly mutated genes. (**F**) The WGD plot highlights samples with a WGD phenotype. (**G**) The dot plot indicates the relationship between the samples; those from the same patient are linked by a black line between two dots.

plete set of chromosomes, is a common feature in cancer genomics. From this plot (Figure 2F), we can observe that ∼40% of samples in these data have WGD. Moreover, our CoMut plot clearly illustrates whether the samples came from the same patient using dot-connection labeling (Figure 2G). This feature is useful for the study of cancer relapse or metastasis cohorts.

Mutational signature analysis

Somatic mutations are ubiquitous in all human cells, which accumulate distinct mutational characteristics. For each cancer, a specific mutational pattern is left behind, namely a mutational signature. Recently, Alexandrov *et al.* reported many more single base substitution (SBS) signatures than they found in 2013 (3). They also mentioned that such mutagenic events can be recognized by dimensional reduction techniques such as NMF. To complete this analysis, we implemented a class named *MutationalSignature*, which contains an estimation method, analysis approaches and multiple plotting functions to construct a smooth pipeline.

First, we utilized a Python library, Nimfa (29), to estimate the factorization rank, which is the optimal number of mutational signatures. We further applied these functions to accurately reproduce the previous results of human adult stem cell data (26). The estimation result is indicated in Supplementary Figure S5. According to the NMF package (30), the most common method to determine the optimal factorization rank is to choose the smallest rank for which the cophenetic correlation coefficient starts decreasing. Thus, based on the cophenetic correlation metric plot, we determined three signatures to be the optimal factorization rank in this dataset.

In order to complete the NMF process, we integrated the scikit-learn decomposition NMF Python package to handle the computation of nonnegative matrices. The three identified substitution signatures perfectly match the results reported by Blokzijl *et al.* (Figure 3A). Furthermore, to compare the three identified signatures with the COSMIC mutational signature dataset, we computed the cosine similarity between each pair of signatures (Figure 3B). Based on the cosine similarity heatmap, we can see that signature 1 was quite similar to COSMIC SBS1, while signature 2 had high similarity with COSMIC SBS5. In the COSMIC dataset, SBS1 and SBS5 are both correlated with the age of the individual. For a more detailed analysis, we also provide the heatmap and bar chart of the relative contribution for each signature so that the mutational activities involved in each sample are clear (Figure 3C and D). The donut plot indicates the total proportion of the three signatures and shows

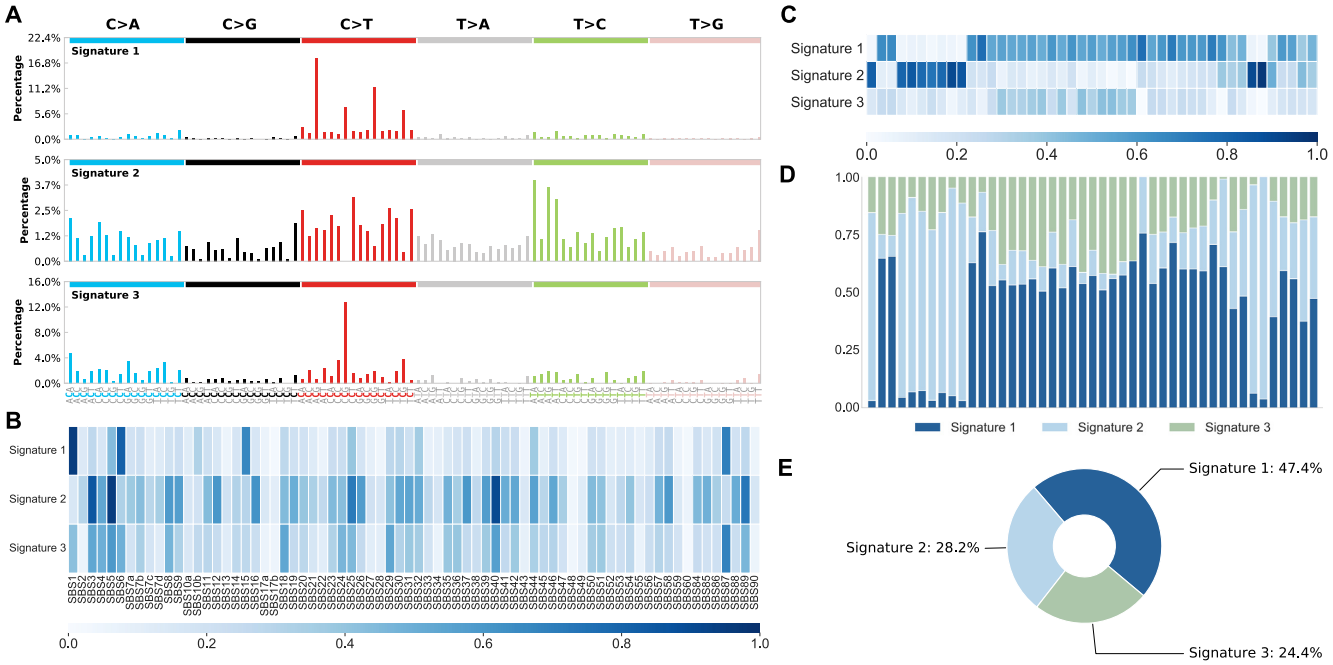

Figure 3. Visualization of mutational signature analysis using data from Blokzijl *et al.* (26). (**A**) SBS signature bar chart. The x-axis indicates the conventional 96 mutation type classification, which is based on the six substitution subtypes: C>A, C>G, C>T, T>A, T>C, T>G. The y-axis indicates the percentage of 96 trinucleotide motifs in the signature. (**B**) Cosine similarity heatmap between identified signatures (y-axis) and SBS signatures of the COSMIC database (x-axis). The blocks with darker blue mean higher similarity between the two signatures. (**C**) The heatmap displays the contribution of the three signatures to mutations in each sample. The blocks with darker blue signify a greater contribution. (**D**) The stacked bar chart shows the relative distribution of the three signatures for each sample. (**E**) The donut plot displays the proportion of the three signatures in the overall cohort.

that signature 1 is predominant in this cohort. To put it another way, the somatic mutations from these testing data are mainly caused by aging. Furthermore, MutScape can also allow users to perform signature refitting to evaluate the contribution of COSMIC signatures for small cohorts or individual samples.

Large-scale genomic event analysis

As a benefit of the development of cancer genomics, more treatments are available to improve the efficacy of therapy. HRD, which means a defect in DNA double-strand break repair, can be quantified by the sum of HRD_LOH, Telomeric_AI and LST. The degree of HRD is an important therapeutic biomarker when deciding to use specific treatments. We provided a class named *HRDScore* to perform HRD score analysis. In previous studies, samples whose HRD score was >42 were eligible to use PARP inhibitors or cisplatin as treatment (6). Using simulated data, it was observed that ∼24% of samples had the HRD phenotype (Figure 4A and B).

CIN score means the level of unstable chromosomes in the sample, expressed as the proportion of abnormal autosomal regions (7), while WGD is a hallmark in cancer genomics (25). MutScape implemented a class called *WGDnCIN* to display the results of CIN and WGD evaluation. In this simulated cohort, most of the samples exhibited high CIN scores (Figure 4C). Furthermore, we can see that 28% of the samples had WGD (Figure 4D) during cancer progression.

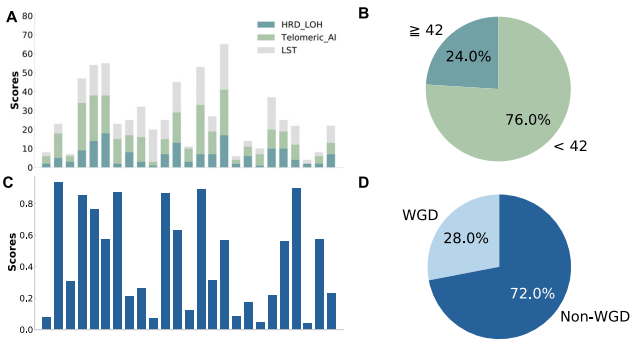

Figure 4. Evaluation of HRD, CIN and WGD. (**A, B**) HRD score analysis. The bar chart displays the HRD score for each sample. HRD score is the sum of HRD_LOH, LST and Telomeric_AI. The pie chart states the proportion of samples, whose HRD scores are ≥42. (**C**) The bar chart indicates the CIN score for each sample. (**D**) The pie chart demonstrates the distribution of WGD status in the cohort.

We further implemented the *HCWComparison* function to demonstrate that MutScape could provide noteworthy findings even from the previous study, which explored changes of driver alterations and clonal evolution following neoadjuvant chemotherapy in esophageal adenocarcinomas (27). The WGD phenotype in pretreatment samples showed an absence after treatment in most good responders (Figure 5A). In case 23, the pretreatment samples with WGD had no major change after treatment, probably due to failure to pass through a genetic bottleneck. Similarly, the HRD

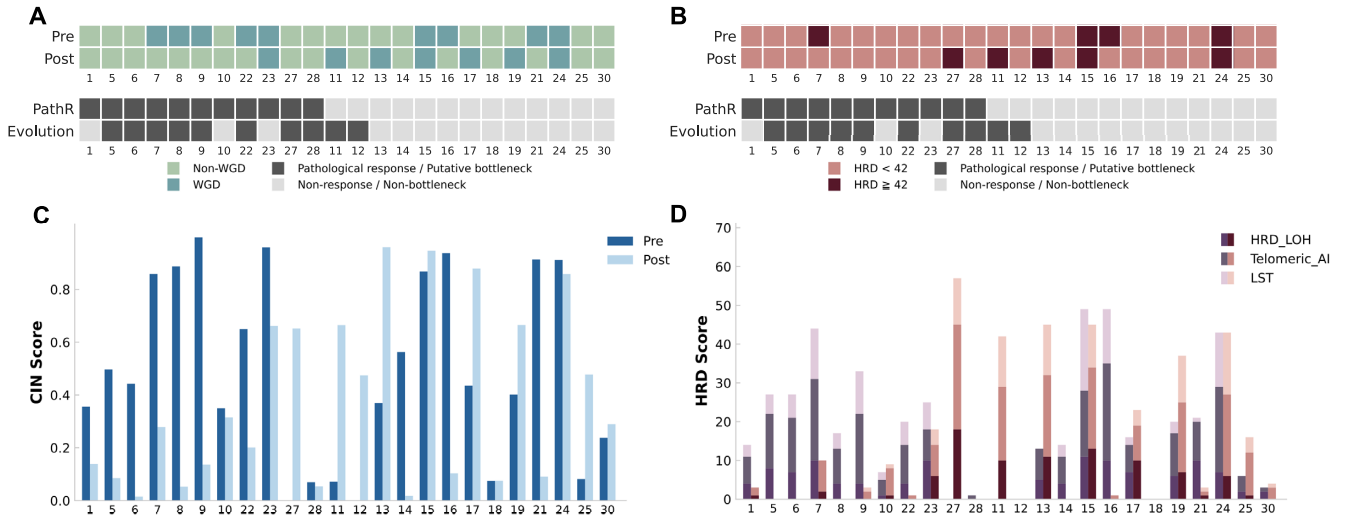

Figure 5. WGD, HRD and CIN in pre- and posttreatment esophageal cancers. Oncoprint showing the (**A**) WGD status and (**B**) HRD phenotype defined as high HRD score (≥42) in pre- and posttreatment tumors. Clinical characteristics and molecular features are indicated with color for each tumor. PathR: pathological response. Bar chart showing the (**C**) CIN and (**D**) HRD score in pre- and posttreatment tumors. HRD score is the sum of HRD_LOH, LST, and Telomeric_AI.

phenotype in pretreatment samples was absent in posttreatment ones under pathological responses (Figure 5B). Notably, in case 27, the posttreatment samples harbored the HRD phenotype in our assessment. The previous study found that this case gained *TP53* and *CNTNAP5* mutations after treatment, with the expansion of a minor pretreatment subclone. Moreover, these findings could also reflect in the CIN and HRD scores that most good responders, having evidence of passing through a genetic bottleneck, exhibited relatively stable genomes after treatment (Figure 5C and D). In contrast, nonresponders showed relatively varied mutational profiles compared to good responders in WGD and HRD assessments.

Actionable biomarker annotation

Since the mainstream of cancer care has gradually become based on the sequencing of tumors, the analysis of actionable mutations has become more imperative than ever before. The precise oncology knowledge base can be a great auxiliary system when medical personnel are making a diagnosis. The database called OncoKB (20) is a precision oncology knowledge base that contains information for actionable cancer gene alterations. MutScape includes the function *OncoKBAnnotator* to evaluate actionability based on OncoKB. Results from this function showed that 44.9% of samples in the TCGA breast cancer cohort have actionable biomarkers in the OncoKB database (Figure 6A). The bar chart shows the distribution of biomarkers among the 44.9% of samples. From this bar chart, we can clearly see that *PIK3CA* is the most common actionable biomarker in this cohort (Figure 6B).

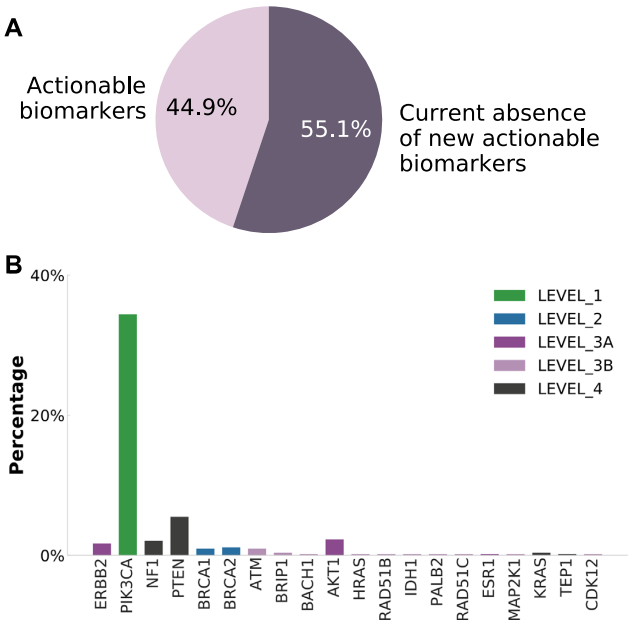

Figure 6. Actionable biomarker annotation analysis using a TCGA breast cancer dataset. (**A**) The pie chart shows the proportion of samples that have actional biomarkers listed in OncoKB. (**B**) The bar chart indicates the proportion of each actionable biomarker in the cohort. The color of the bar signifies OncoKB therapeutic levels of evidence. Level 1 (green), FDA-approved biomarker; level 2 (blue), standard care biomarker; level 3 (purple), clinical evidence-supported biomarker; level 4 (dark gray), biological evidence-supported biomarker.

DISCUSSION

The dramatic growth of cancer genomic sequencing data highlights the need for efficient and potent analysis tools. There are numerous common tools that offer several useful analysis and visualization modules. Maftools (4), for instance, is a widely used R package that provides SMG detection, known cancer gene annotation and mutational signature analysis. However, its application is limited by its input file format requirements, which hinder its usage due to

Table 1. Comparison of features for VCF and MAF analysis and visualization tools

Categories	Features	MutScape	Maftools	VCFtools	MuSiC	VIVA
Technical details	One-step command	✓	✓	✓	✓	✓
	Language	Python	R	C++, Perl	Perl	Julia
Preprocessing	Genomic range filter	✓		✓	✓	✓
	Caller information filter	✓		✓		
	PASS filter	✓		✓		✓
	Sample selection			✓		✓
	Tissue expression filter	✓				
	File format transformation	✓				
	File combination	✓		✓		
Analysis	SMG detection	✓	✓		✓	
	Known cancer gene annotation	✓	✓			
	Mutation burden statistics	✓	✓			
	Mutational signature	✓	✓			
	Homologous recombination deficiency	✓				
	Whole-genome doubling	✓				
	Chromosome instability	✓				
	Actionable mutation annotation	✓				
	Pfam annotation		✓		✓	
Visualization	CoMut plot	✓	✓			
	Cosine similarity heatmap	✓				
	SBS signature bar chart	✓	✓			
	Lollipop plot		✓			

SMG, significantly mutated gene; SBS, single base substitution.

the lack of filtering false-positive mutations. Another tool, VCFtools (31), lacks the function of data transformation, which is a major limitation, since most of the analyses for cancer genomics require MAF files as the input format. VCFtools is also devoid of visualization functions. Other tools such as MuSiC (2) and VIVA (5) can be used for visualization and may also handle some analyses. Nevertheless, these tools provide limited capability that must be complemented by other tools, seriously impeding the efficiency and accuracy of studies in cancer genomics (Table 1). Furthermore, utilizing multiple variant callers can provide more robust results than using single ones. TCGA, for example, organized the Multi-Center Mutation Calling in Multiple Cancers (MC3) project to apply an ensemble of seven mutation-calling algorithms for robust cross-tumor-type analyses. However, none of these tools can process these mutational data from a combination of variant callers, for example by assessing the intersection/union between two or more callers.

Here, we describe a user-friendly Python package, MutScape, which combines well-contrived data preprocessing and a plethora of representative analyses that are commonly utilized in cancer genomics. This package is available for mutation data from WGS, WES or gene panels. MutScape also provides multiple choices for visualization that produce high-quality images for publication. Along with specially designed strategies for data filtering and combination, users can easily produce more reliable and robust results than with the other tools mentioned above using only few lines of commands. Moreover, since MutScape skillfully integrates data preprocessing and analysis, it paves the way for complex computation and statistics approaches for bioinformaticians and will also substantially boost the efficiency of research in cancer genomics.

Moreover, the rationale for choices of tools is described below in detail. Currently, most tools depicting the mutational landscape only plot specific genomic and phenotypic data types, such as CoMutPlotter (32), jsComut (33), Maftools (4) and GenVisR (34). CoMut can visualize various types of genomic data and more complex phenotypic information. OncodriveCLUST identifies SMGs with a bias toward mutation clustering that may complement other methods of detecting SMGs. Furthermore, only the scarHRD package could assess three HRD-related biomarkers to calculate the HRD score based on the previous study. This scoring system was further developed as an FDA-approved HRD test, namely 'myChoice CDx'. Also, the oncokb annotator was directly developed by the OncoKB team. The MC3 project organized by TCGA employed the vcf2maf script for VCF annotation and transformation (35).

However, there are three limitations that exist in this package. The quantification of the extent of HRD and the assessment of WGD status was based on allele-specific copy numbers for genome regions, relying on the WGS and WES data. Users need to provide such information as inputs through tools developed for performing an allele-specific copy number analysis. Moreover, this package does not include germline mutation analysis. Thus, the HRD assessment of MutScape did not consider the BRCA deficiency. Notably, a previous study showed that the HRD scoring measurement has the ability beyond BRCA deficiency by identifying more patients with high sensitivity to PARP inhibitors or platinum agents (36). Furthermore, record formats in VCFs were different based on the variant callers, such as AD and DP information. As we employed the vcf2maf script for VCF annotation and transformation, we mainly followed its method, which handled VCFs according to variant callers. Thus, MutScape currently supports VCFs from six commonly used variant callers in the data preprocessing module. We may also extend the functionality of MutScape for VCFs from more variant callers in the future.

In summary, MutScape provides a comprehensive and easy-to-use pipeline and also exactly reproduced many known results from published datasets. More significantly, MutScape is capable of producing novel results via advanced analyses. In the future, we will work to diversify the analysis and visualization modules in MutScape by adding the ability to handle Pfam annotation and GISTIC2 (28) results, resulting in a comprehensive suite of functions for omics data analysis.

DATA AVAILABILITY

The source code for the updated MutScape is freely available at https://github.com/anitalu724/MutScape.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank Melissa Stauffer for editorial assistance. We also benefited from insightful discussion from NTU Core Consortiums.

Authors contributions: EYC and LCL conceived and supervised the study; CHL and CHW designed the algorithm; CHL implemented the algorithm; CHW and MHT collected all testing data; CHL and CHW designed the programming structure of MutScape; CHL and CHW tested each function of MutScape; CHL and CHW wrote the paper; and CHL and CHW wrote README on GitHub. All authors read and approved the manuscript.

FUNDING

Industrial Technology Research Institute [110HT654005]; YongLin Healthcare Foundation [FB002-7]; National Taiwan University [GTZ300; 109L104704-2].

Conflict of interest statement. None declared.

REFERENCES

1. Malone,E.R., Oliva,M., Sabatini,P.J., Stockley,T.L. and Siu,L.L. (2020) Molecular profiling for precision cancer therapies. *Genome Med.*, **12**, 8.
2. Dees,N.D., Zhang,Q., Kandoth,C., Wendl,M.C., Schierding,W., Koboldt,D.C., Mooney,T.B., Callaway,M.B., Dooling,D. and Mardis,E.R. (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res.*, **22**, 1589–1598.
3. Alexandrov,L.B., Kim,J., Haradhvala,N.J., Huang,M.N., Ng,A.W.T., Wu,Y., Boot,A., Covington,K.R., Gordenin,D.A. and Bergstrom,E.N. (2020) The repertoire of mutational signatures in human cancer. *Nature*, **578**, 94–101.
4. Mayakonda,A., Lin,D.-C., Assenov,Y., Plass,C. and Koeffler,H.P. (2018) Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.*, **28**, 1747–1756.
5. Tollefson,G.A., Schuster,J., Gelin,F., Agudelo,A., Ragavendran,A., Restrepo,I., Stey,P., Padbury,J. and Uzun,A. (2019) VIVA (VIsualization of VAriants): a VCF file visualization tool. *Sci. Rep.*, **9**, 12648.
6. Telli,M.L., Timms,K.M., Reid,J., Hennessy,B., Mills,G.B., Jensen,K.C., Szallasi,Z., Barry,W.T., Winer,E.P. and Tung,N.M. (2016) Homologous recombination deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. *Clin. Cancer Res.*, **22**, 3764–3773.
7. Pereira,B., Chin,S.-F., Rueda,O.M., Vollan,H.-K.M., Provenzano,E., Bardwell,H.A., Pugh,M., Jones,L., Russell,R. and Sammut,S.-J. (2016) The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat. Commun.*, **7**, 11479.
8. Fan,Y., Xi,L., Hughes,D.S., Zhang,J., Zhang,J., Futreal,P.A., Wheeler,D.A. and Wang,W. (2016) MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.*, **17**, 178.
9. Benjamin,D., Sato,T., Cibulskis,K., Getz,G., Stewart,C. and Lichtenstein,L. (2019) Calling somatic SNVs and indels with Mutect2. bioRxiv doi: https://doi.org/10.1101/861054, 02 December 2019, preprint: not peer reviewed.
10. Larson,D.E., Harris,C.C., Chen,K., Koboldt,D.C., Abbott,T.E., Dooling,D.J., Ley,T.J., Mardis,E.R., Wilson,R.K. and Ding,L. (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, **28**, 311–317.
11. Kim,S., Scheffler,K., Halpern,A.L., Bekritsky,M.A., Noh,E., Källberg,M., Chen,X., Kim,Y., Beyter,D. and Krusche,P. (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, **15**, 591–594.
12. Koboldt,D.C., Zhang,Q., Larson,D.E., Shen,D., McLellan,M.D., Lin,L., Miller,C.A., Mardis,E.R., Ding,L. and Wilson,R.K. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
13. Miller,N.A., Farrow,E.G., Gibson,M., Willig,L.K., Twist,G., Yoo,B., Marrs,T., Corder,S., Krivohlavek,L., Walter,A. *et al.* (2015) A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Med.*, **7**, 100.
14. Diossy,M., Reiniger,L., Sztupinszki,Z., Krzystanek,M., Timms,K.M., Neff,C., Solimeno,C., Pruss,D., Eklund,A.C. and Toth,E. (2018) Breast cancer brain metastases show increased levels of genomic aberration-based homologous recombination deficiency scores relative to their corresponding primary tumors. *Ann. Oncol.*, **29**, 1948–1954.
15. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J. and Cummings,B.B. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
16. Crowdis,J., He,M.X., Reardon,B. and Van Allen,E.M. (2020) CoMut: visualizing integrated molecular information with comutation plots. *Bioinformatics*, **36**, 4348–4349.
17. Tamborero,D., Gonzalez-Perez,A. and Lopez-Bigas,N. (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, **29**, 2238–2244.
18. Sztupinszki,Z., Diossy,M., Krzystanek,M., Reiniger,L., Csabai,I., Favero,F., Birkbak,N.J., Eklund,A.C., Syed,A. and Szallasi,Z. (2018) Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. *NPJ Breast Cancer*, **4**, 16.
19. Sondka,Z., Bamford,S., Cole,C.G., Ward,S.A., Dunham,I. and Forbes,S.A. (2018) The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*, **18**, 696–705.
20. Chakravarty,D., Gao,J., Phillips,S., Kundra,R., Zhang,H., Wang,J., Rudolph,J.E., Yaeger,R., Soumerai,T. and Nissan,M.H. (2017) OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.*, **2017**, PO.17.00011.
21. Chan,T.A., Yarchoan,M., Jaffee,E., Swanton,C., Quezada,S.A., Stenzinger,A. and Peters,S. (2019) Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann. Oncol.*, **30**, 44–56.
22. Abkevich,V., Timms,K., Hennessy,B., Potter,J., Carey,M., Meyer,L., Smith-McCune,K., Broaddus,R., Lu,K. and Chen,J. (2012) Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer*, **107**, 1776–1782.
23. Birkbak,N.J., Wang,Z.C., Kim,J.-Y., Eklund,A.C., Li,Q., Tian,R., Bowman-Colin,C., Li,Y., Greene-Colozzi,A. and Iglehart,J.D. (2012) Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov.*, **2**, 366–375.
24. Popova,T., Manié,E., Rieunier,G., Caux-Moncoutier,V., Tirapo,C., Dubois,T., Delattre,O., Sigal-Zafrani,B., Bollet,M. and Longy,M.

(2012) Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.*, **72**, 5454–5462.
25. Bielski,C.M., Zehir,A., Penson,A.V., Donoghue,M.T., Chatila,W., Armenia,J., Chang,M.T., Schram,A.M., Jonsson,P. and Bandlamudi,C. (2018) Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.*, **50**, 1189–1195.
26. Blokzijl,F., De Ligt,J., Jager,M., Sasselli,V., Roerink,S., Sasaki,N., Huch,M., Boymans,S., Kuijk,E. and Prins,P. (2016) Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, **538**, 260–264.
27. Findlay,J.M., Castro-Giner,F., Makino,S., Rayner,E., Kartsonaki,C., Cross,W., Kovac,M., Ulahannan,D., Palles,C., Gillies,R.S. *et al.* (2016) Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy. *Nat. Commun.*, **7**, 11111.
28. Mermel,C.H., Schumacher,S.E., Hill,B., Meyerson,M.L., Beroukhim,R. and Getz,G. (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.*, **12**, R41.
29. Zitnik,M. and Zupan,B. (2012) Nimfa: a Python library for nonnegative matrix factorization. *J. Mach. Learn. Res.*, **13**, 849–853.
30. Gaujoux,R. and Seoighe,C. (2010) A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, **11**, 367.
31. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T. and Sherry,S.T. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
32. Huang,P.J., Lin,H.H., Lee,C.C., Chiu,L.Y., Wu,S.M., Yeh,Y.M., Tang,P., Chiu,C.H., Lyu,P.C. and Tsai,P.C. (2019) CoMutPlotter: a web tool for visual summary of mutations in cancer cohorts. *BMC Med. Genomics*, **12**, 99.
33. Pearce,T.M., Nikiforova,M.N. and Roy,S. (2019) Interactive browser-based genomics data visualization tools for translational and clinical laboratory applications. *J. Mol. Diagn.*, **21**, 985–993.
34. Skidmore,Z.L., Wagner,A.H., Lesurf,R., Campbell,K.M., Kunisaki,J., Griffith,O.L. and Griffith,M. (2016) GenVisR: genomic visualizations in R. *Bioinformatics*, **32**, 3012–3014.
35. Ellrott,K., Bailey,M.H., Saksena,G., Covington,K.R., Kandoth,C., Stewart,C., Hess,J., Ma,S., Chiotti,K.E., McLellan,M. *et al.* (2018) Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.*, **6**, 271–281.
36. von Wahlde,M.K., Timms,K.M., Chagpar,A., Wali,V.B., Jiang,T., Bossuyt,V., Saglam,O., Reid,J., Gutin,A., Neff,C. *et al.* (2017) Intratumor heterogeneity of homologous recombination deficiency in primary breast cancer. *Clin. Cancer Res.*, **23**, 1193–1199.