Examining clustered somatic mutations with SigProfilerClusters

Erik N. Bergstrom[1-3], Mousumy Kundu[1-3], Noura Tbeileh[1-3], and Ludmil B. Alexandrov[1-3*]

[1]Department of Cellular and Molecular Medicine, UC San Diego, La Jolla, CA, 92093, USA
[2]Department of Bioengineering, UC San Diego, La Jolla, CA, 92093, USA
[3]Moores Cancer Center, UC San Diego, La Jolla, CA, 92037, USA

[*]Correspondence should be addressed to L2alexandrov@health.ucsd.edu.

ABSTRACT

Summary: Clustered mutations are found in the human germline as well as in the genomes of cancer and normal somatic cells. Clustered events can be imprinted by a multitude of mutational processes, and they have been implicated in both cancer evolution and development disorders. Existing tools for identifying clustered mutations have been optimized for a particular subtype of clustered event and, in most cases, relied on a predefined inter-mutational distance (IMD) cutoff combined with a piecewise linear regression analysis. Here we present SigProfilerClusters, an automated tool for detecting all types of clustered mutations by calculating a sample-dependent IMD threshold using a simulated background model that takes into account extended sequence context, transcriptional strand asymmetries, and regional mutation densities. SigProfilerClusters disentangles all types of clustered events from non-clustered mutations and annotates each clustered event into an established subclass, including the widely used classes of doublet-base substitutions, multi-base substitutions, *omikli*, and *kataegis*. SigProfilerClusters outputs non-clustered mutations and clustered events using standard data formats as well as provides multiple visualizations for exploring the distributions and patterns of clustered mutations across the genome.

Availability: SigProfilerClusters is freely available at https://github.com/AlexandrovLab/SigProfilerClusters with support across most operating systems and extensive documentation at https://osf.io/qpmzw/wiki/home/.

Contact: ebergstr@eng.ucsd.edu or L2alexandrov@health.ucsd.edu

INTRODUCTION

Mutations are found on the genomes of all cells in the human body (Martincorena and Campbell, 2015; Stratton, et al., 2009). Most single-base substitutions and small insertions and deletions (indels) accumulate independently across the genome, but a subset of mutations cluster in a non-random manner (Lawrence, et al., 2013; Supek and Lehner, 2017). Previous studies have revealed that clustered mutations are imprinted by a plethora of endogenous and exogenous mutational processes (Alexandrov, et al., 2020; Boichard, et al., 2017; Brash, 2015; Buisson, et al., 2019; Chan, et al., 2015; Chen, et al., 2013; Consortium, 2020; Mas-Ponte and Supek, 2020; Matsuda, et al., 1998; Nik-Zainal, et al., 2012; Nik-Zainal, et al., 2019; Pfeifer, et al., 2005; Roberts, et al., 2013; Roberts, et al., 2012; Supek and Lehner, 2017; Taylor, et al., 2013; Wang, et al., 2020). Some clustered mutations have been implicated in cancer evolution (Bergstrom, et al., 2022; Chen, et al., 2013; Consortium, 2020; Mas-Ponte and Supek, 2020; Supek and Lehner, 2017; Taylor, et al., 2013), while *de novo* clustered mutations have been identified in the human germline and shown to contribute to developmental disorders (Kaplanis, et al., 2019; Veltman and Brunner, 2012). In recent years, sets of simultaneously occurring clustered substitutions have been further subclassified into independent events (Bergstrom, et al., 2022; Mas-Ponte and Supek, 2020), including: *(i)* doublet-base substitutions (DBSs); *(ii)* multi-base substitutions (MBSs); *(iii)* diffuse hypermutation termed *omikli*; *(iv)* longer strand-coordinated events termed *kataegis*; and *(v)* recurrent hypermutation of extra-chromosomal DNA (ecDNA) termed *kyklonas*.

Traditional methods separate clustered mutations based upon a predefined inter-mutational distance (IMD) threshold typically between 1 and 2 kilobases (Alexandrov, et al., 2020;

55 Alexandrov, et al., 2013; Chan, et al., 2015; D'Antonio, et al., 2016; Maciejowski, et al., 2020;
56 Nik-Zainal, et al., 2019; Taylor, et al., 2013). Many of these approaches utilize a piece-wise
57 linear regression to segment each chromosome, which, in most cases, is optimized for calling
58 larger strand-coordinated *kataegic* events (**Supplementary Fig. 1**) (Alexandrov, et al., 2013;
59 Lin, et al., 2021; Yin, et al., 2020). Most existing methods have also ignored confounding effects
60 attributed to localized differences in mutation rates, copy number alterations, or the mutational
61 burden across each chromosome within a given sample leading to an accumulation of false
62 positive clustered events (**Supplementary Fig. 1**). Further, the majority of existing tools focus
63 on detecting only a specific class of clustered events including doublet-base substitutions and
64 multi-nucleotide variants (Chen, et al., 2013; Matsuda, et al., 1998; Wang, et al., 2020), *kataegis*
65 (D'Antonio, et al., 2016; Lin, et al., 2021; Taylor, et al., 2013), or APOBEC3-associated events
66 (Chan, et al., 2015; Nik-Zainal, et al., 2012) while ignoring the larger landscape of clustered
67 mutations. For example, a recent study (Mas-Ponte and Supek, 2020) developed an algorithm
68 focused on the detection of APOBEC3-associated *omikli* and *kataegis* events in cancer genomes
69 by incorporating simulations of somatic mutations and estimates of cancer cell fractions.
70
71 Separation and classification of clustered events is required to fully elucidate the mutational
72 processes operating in cancer and normal somatic cells (Bergstrom, et al., 2022; Supek and
73 Lehner, 2017). Here we present SigProfilerClusters, a tool to comprehensively characterize and
74 subclassify clustered mutations from the complete catalog of mutations within the genome of a
75 single sample (**Fig. 1*a***). SigProfilerClusters classifies all types of clustered mutations, including
76 *(i)* doublet-base substitutions; *(ii)* multi-base substitutions; *(iii) omikli*; *(iv) kataegis*; and *(v)*
77 clustered small insertions and deletions (indels). The tool calculates a sample-dependent IMD

threshold that considers regional differences in mutation rates, variant allele fractions and cancer cell fractions of adjacent mutations to reduce the false positive rate and provides visualizations for downstream analyses (**Fig. 1*b&c***; **Supplementary Fig. 1**). Further, SigProfilerClusters integrates within the larger suite of SigProfiler tools (Bergstrom, et al., 2020; Bergstrom, et al., 2019; Islam, et al., 2020) to facilitate downstream mutational signature analysis of both non-clustered and clustered single-base substitutions and indels, thus, allowing the accurate detection of mutational processes giving rise to even low levels of clustered events (**Fig. 1*d***) (Bergstrom, et al., 2019; Bergstrom, et al., 2022; Islam, et al., 2020).

METHODS

SigProfilerClusters derives an IMD cutoff that is unlikely to occur purely by chance given the observed mutational burden and the mutational patterns within the genome of a given sample. To calculate the genome-dependent IMD, the tool leverages SigProfilerSimulator (Bergstrom, et al., 2020) to generate background models by randomizing the distribution of mutations across the genome. By default, the genome of each sample is simulated 100 times in order to derive 95% confidence intervals for the expected genomic mutational landscape, with every simulation maintaining the penta-nucleotide sequence context for each substitution, the ratio of all mutations in genic and inter-genic regions, the transcriptional strand asymmetries of all mutations in genic regions, and the mutational burden on each chromosome (Bergstrom, et al., 2020; Bergstrom, et al., 2019). Importantly, this randomization procedure is highly customizable (Bergstrom, et al., 2020) and can be altered based upon the needs of a given study design, thus, allowing the incorporation of other factors that affect the accumulation of mutations such as nucleosome occupancy, presence of histone modifications, and many others. A binary search

101 algorithm is implemented to efficiently derive the global IMD threshold for each genome. The
102 final global IMD threshold is selected by ensuring that 90% of mutations below the chosen
103 cutoff are unlikely to appear by chance given the simulated distribution of mutations (q-
104 value<0.01; **Supplementary Fig. 1**) with a maximum global IMD cutoff of 10 kilobases. The
105 algorithm also considers regional heterogeneities of mutation rates, generally associated with
106 replication timing (Stamatoyannopoulos, et al., 2009) or differential gene expression (Buisson, et
107 al., 2019; Hess, et al., 2019; Lawrence, et al., 2013; Pleasance, et al., 2010; Polak, et al., 2015),
108 by correcting for variance in clonality as well as variance in both mutation-dense and mutation-
109 poor regions using a sliding genomic window (default size of 1 megabase). Specifically, an
110 additional regional IMD cutoff is corrected within each genomic window based upon the fold
111 difference between the number of real and the number of simulated mutations, while maintaining
112 the original criteria of less than 10% of mutations below the IMD cutoff appearing by chance (q-
113 value<0.01). Lastly, when data are available, SigProfilerClusters ensures that adjacent mutations
114 are in the same cells by introducing a maximum difference in variant allele frequencies (VAF) or
115 cancer cell fraction (CCF), which incorporates copy-number changes, below a certain threshold
116 (default cutoff value of 0.10 and 0.25; respectively).
117
118 After identifying the set of clustered mutations, SigProfilerClusters subclassifies each clustered
119 substitution into a single category of previously established clustered events (Bergstrom, et al.,
120 2022; Mas-Ponte and Supek, 2020). Briefly, all clustered substitutions with consistent VAFs or
121 consistent CCFs are classified into one of four categories. Two mutations with an IMD of 1 are
122 classified as *doublet-base substitutions*, while clusters of three or more adjacent mutations each
123 with an IMD of 1 are classified as *multi-base substitutions*. Clusters of two or three mutations

with IMDs less than the sample-dependent cutoff and with at least a single IMD greater than 1 are classified as *omikli* (Bergstrom, et al., 2022), while clusters of four or more mutations with IMDs less than the sample-dependent cutoff and with at least a single IMD greater than 1 are classified as *kataegis* (Bergstrom, et al., 2022). All remaining clustered mutations with inconsistent VAFs or CCFs are classified as *other*. Clustered indels are not subclassified into different categories due to a lack of previously defined subtypes.

USAGE

SigProfilerClusters is freely available as a Python package, distributed under the permissive BSD-2 clause license, and can be used on most operating systems including Windows, MacOS, and Linux-based machines. The tool is compatible with large-scale deployments on high-performance computing clusters as well as on cloud infrastructures such as Amazon Web Services (AWS). Input data can be provided in the form of common mutation formats including the Variant Call Format (VCF), the Mutation Annotation Format (MAF), or in the form of a simple text file. The output of SigProfilerClusters results in the partitioning of all mutations into a clustered or non-clustered directory. All clustered mutations are then classified into distinct subcategories of events and provided individually in VCF files for downstream visualization and analyses. The output for each subclass of clustered event can be directly utilized by additional SigProfiler tools including SigProfilerExtractor for mutational signature analysis (Islam, et al., 2020) and SigProfilerPlotting for examining patterns of somatic mutations (Bergstrom, et al., 2019). The results for each sample are also summarized using two individual visualizations that include: *(i)* a rainfall plot depicting the minimum global IMD between all adjacent mutations, where each individual set of adjacent mutations is colored based upon its clustered classification;

147 and *(ii)* a multi-panel figure that displays the mutational patterns across all mutations, clustered
148 mutations, and non-clustered mutations, separately along with the distribution of IMDs across
149 the real and simulated data for each sample (**Fig. 1*a***).
150

151 **CONCLUSION**

152 Elucidating the compendium of clustered somatic mutations in the genome of a sample allows
153 further understanding of the mutational process that give rise to these events and can provide
154 novel insights into disease etiology (Bergstrom, et al., 2022; Mas-Ponte and Supek, 2020; Supek
155 and Lehner, 2017). Previous studies have traditionally interrogated the complete mutational
156 catalogs of cancer genomes, which can lead to the inability to detect processes active at low
157 levels or those which have been transiently activated. Our prior analysis of clustered mutations
158 (Bergstrom, et al., 2022) have revealed an enrichment of clustered mutations within known
159 cancer driver events, hypermutation of extrachromosomal DNA fueling the evolution of cancers,
160 and ultimately, resulting in a differential patient outcome. Here we provide SigProfilerClusters,
161 an automated and freely available Python based tool that comprehensively identifies and
162 classifies clustered mutations enabling users to interrogate the mutational processes giving rise to
163 such events.
164

165 **FIGURE LEGENDS**

166 **Figure 1. Detection and characterization of clustered mutations with SigProfilerClusters. *a***)
167 An example workflow used to detect clustered mutations in a single cancer genome. As an input,
168 SigProfilerClusters accepts common formats for mutations, such as ones in the variant calling
169 format (VCF), and the tool separates all clustered mutations from the complete mutational

170 catalog of the provided sample. Final partitions of mutations in the sample are outputted as VCF
171 files and visualized using the mutational spectra of all mutations, only clustered mutations, and
172 only non-clustered mutations along with a rainfall plot commonly used to show the distribution
173 of inter-mutational distances across a cancer genome (Alexandrov, et al., 2013; Bergstrom, et al.,
174 2022; Nik-Zainal, et al., 2012). ***b)*** Schematic demonstrating the process of calculating a sample
175 dependent IMD threshold to separate clustered from non-clustered mutations across each
176 genome. A binary search algorithm is used to efficiently detect the optimal global IMD threshold
177 for each sample. Detection of the global IMD threshold is illustrated using grey arrows. Regional
178 corrections are performed to identify local IMD thresholds based on variance of mutation rates
179 across the genome. ***c)*** Every clustered mutation is classified into a single subcategory of
180 clustered event. ***d)*** Rainfall plot illustrating the distribution of IMDs across a single glioblastoma
181 sample (*left*). The mutational spectra for *omikli* and *kataegic* events reveal a different mutational
182 pattern compared to the pattern of all non-clustered somatic mutations (*right*).
183

184 **SUPPLEMENTARY FIGURE LEGENDS**

185 **Supplementary Figure 1. Benchmarking of existing tools for detecting clustered mutations.**
186 ***a***) Assessing the degree of overlap between two tools that detect clustered mutations and
187 SigProfilerCusters. P-MACD locates regions of clustered mutations by implementing a negative
188 binomial distribution to model the probability of observing a given cluster within a given
189 window of the genome. When calculating p-values, this model assumes that all mutations occur
190 randomly across the genome limiting clustered events to those that occur at most 10 kilobases
191 (kb) from one another. Kataegis implement a piece-wise linear model with set thresholds
192 requiring all clustered events to be composed of 6 or more mutations with an average inter-

193 mutational distance of 1 kb between adjacent mutations. All tools were applied to 2,703 whole-
194 genome sequenced samples from PCAWG and were split into low (TMB<1), intermediate
195 (1<=TMB<10), and high (TMB>=10) TMB. ***b)*** The percentage of mutations overlapping each
196 subclass of clustered events called by SigProfilerClusters (*left*), the percentage of mutations that
197 were missed by each of the two tools that were called by SigProfilerClusters (*middle*), and the
198 percentage of total mutations called by each tool that were missed by SigProfilerClusters
199 separated by low, intermediate, and high TMB (*right*). ***c)*** Schematic workflow to assess the false
200 positive rate of each tool using simulations of 211 breast cancer genomes from PCAWG. ***d)*** The
201 average number of false positive mutations per simulated sample detected across the simulated
202 breast cancer dataset from PCAWG. Kataegis did not detect any *kataegic* clustered events within
203 the simulated data using the tools' default parameters. Similarly, SigProfilerClusters did not
204 detect any *kataegic* clustered events within these simulated data.
205
206 **ACKNOWLEDGEMENTS**
207 **Funding:** This was work was funded by Cancer Research UK Grand Challenge Award
208 C98/A24032 as well as US National Institute of Health grants R01ES030993-01A1 and
209 R01ES032547. LBA is also supported by a Packard Fellowship for Science and Engineering.
210 The funders had no roles in study design, data collection and analysis, decision to publish, or
211 preparation of the manuscript.
212
213 **Conflicts of Interest:** LBA is a compensated consultant and has equity interest in io9, LLC. His
214 spouse is an employee of Biotheranostics, Inc. LBA is also an inventor of a US Patent
215 10,776,718 for source identification by non-negative matrix factorization. ENB and LBA declare

provisional patent applications for “Clustered mutations for the treatment of cancer” (U.S. provisional application serial number 63/289,601) and “Artificial intelligence architecture for predicting cancer biomarker” (serial number 63/269,033). All other authors declare no competing interests.

Contributions: ENB developed the Python code and wrote the manuscript. MK performed all benchmarking. ENB, MK, and NT tested and documented the code. LBA supervised the overall development of the code, benchmarking, and writing of the manuscript. All authors read and approved the final manuscript.

REFERENCES

Alexandrov, L.B., *et al.* The repertoire of mutational signatures in human cancer. *Nature* 2020;578(7793):94-101.

Alexandrov, L.B., *et al.* Signatures of mutational processes in human cancer. *Nature* 2013;500(7463):415-421.

Bergstrom, E.N., *et al.* Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator. *BMC Bioinformatics* 2020;21(1):438.

Bergstrom, E.N., *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* 2019;20(1):685.

Bergstrom, E.N., *et al.* Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA. *Nature* 2022.

Boichard, A., Tsigelny, I.F. and Kurzrock, R. High expression of PD-1 ligands is associated with kataegis mutational signature and APOBEC3 alterations. *Oncoimmunology* 2017;6(3):e1284719.

Brash, D.E. UV signature mutations. *Photochem Photobiol* 2015;91(1):15-26.

Buisson, R., *et al.* Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* 2019;364(6447).

Chan, K., *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet* 2015;47(9):1067-1072.

Chen, J.M., Ferec, C. and Cooper, D.N. Patterns and mutational signatures of tandem base substitutions causing human inherited disease. *Hum Mutat* 2013;34(8):1119-1130.

Consortium, I.T.P.-C.A.o.W.G. Pan-cancer analysis of whole genomes. *Nature* 2020;578(7793):82-93.

D'Antonio, M., *et al.* Kataegis Expression Signature in Breast Cancer Is Associated with Late Onset, Better Prognosis, and Higher HER2 Levels. *Cell Rep* 2016;16(3):672-683.

Hess, J.M., *et al.* Passenger Hotspot Mutations in Cancer. *Cancer Cell* 2019;36(3):288-301 e214.

Islam, S.M.A., *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *bioRxiv* 2020:2020.2012.2013.422570.

Kaplanis, J., *et al.* Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations. *Genome Res* 2019;29(7):1047-1056.

Lawrence, M.S., *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;499(7457):214-218.

Lin, X., *et al.* kataegis: an R package for identification and visualization of the genomic localized hypermutation regions using high-throughput sequencing. *BMC Genomics* 2021;22(1):440.

Maciejowski, J., *et al.* APOBEC3-dependent kataegis and TREX1-driven chromothripsis during telomere crisis. *Nat Genet* 2020;52(9):884-890.

Martincorena, I. and Campbell, P.J. Somatic mutation in cancer and normal cells. *Science* 2015;349(6255):1483-1489.

Mas-Ponte, D. and Supek, F. DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers. *Nat Genet* 2020;52(9):958-968.

Matsuda, T., *et al.* Specific tandem GG to TT base substitutions induced by acetaldehyde are due to intra-strand crosslinks between adjacent guanine bases. *Nucleic Acids Res* 1998;26(7):1769-1774.

Nik-Zainal, S., *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* 2012;149(5):979-993.

Nik-Zainal, S., *et al.* Author Correction: Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2019;566(7742):E1.

272 Pfeifer, G.P., You, Y.H. and Besaratinia, A. Mutations induced by ultraviolet light. *Mutat Res*
273 2005;571(1-2):19-31.
274 Pleasance, E.D.*, et al.* A comprehensive catalogue of somatic mutations from a human cancer
275 genome. *Nature* 2010;463(7278):191-196.
276 Polak, P.*, et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer.
277 *Nature* 2015;518(7539):360-364.
278 Roberts, S.A.*, et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in
279 human cancers. *Nat Genet* 2013;45(9):970-976.
280 Roberts, S.A.*, et al.* Clustered mutations in yeast and in human cancers can arise from damaged
281 long single-strand DNA regions. *Mol Cell* 2012;46(4):424-435.
282 Stamatoyannopoulos, J.A.*, et al.* Human mutation rate associated with DNA replication timing.
283 *Nat Genet* 2009;41(4):393-395.
284 Stratton, M.R., Campbell, P.J. and Futreal, P.A. The cancer genome. *Nature*
285 2009;458(7239):719-724.
286 Supek, F. and Lehner, B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair
287 Targets Mutations to Active Genes. *Cell* 2017;170(3):534-547 e523.
288 Taylor, B.J.*, et al.* DNA deaminases induce break-associated mutation showers with implication
289 of APOBEC3B and 3A in breast cancer kataegis. *Elife* 2013;2:e00534.
290 Veltman, J.A. and Brunner, H.G. De novo mutations in human genetic disease. *Nat Rev Genet*
291 2012;13(8):565-575.
292 Wang, Q.*, et al.* Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708
293 genomes. *Nat Commun* 2020;11(1):2539.
294 Yin, X.*, et al.* Multiregion whole-genome sequencing depicts intratumour heterogeneity and
295 punctuated evolution in ovarian clear cell carcinoma. *J Med Genet* 2020;57(9):605-609.
296
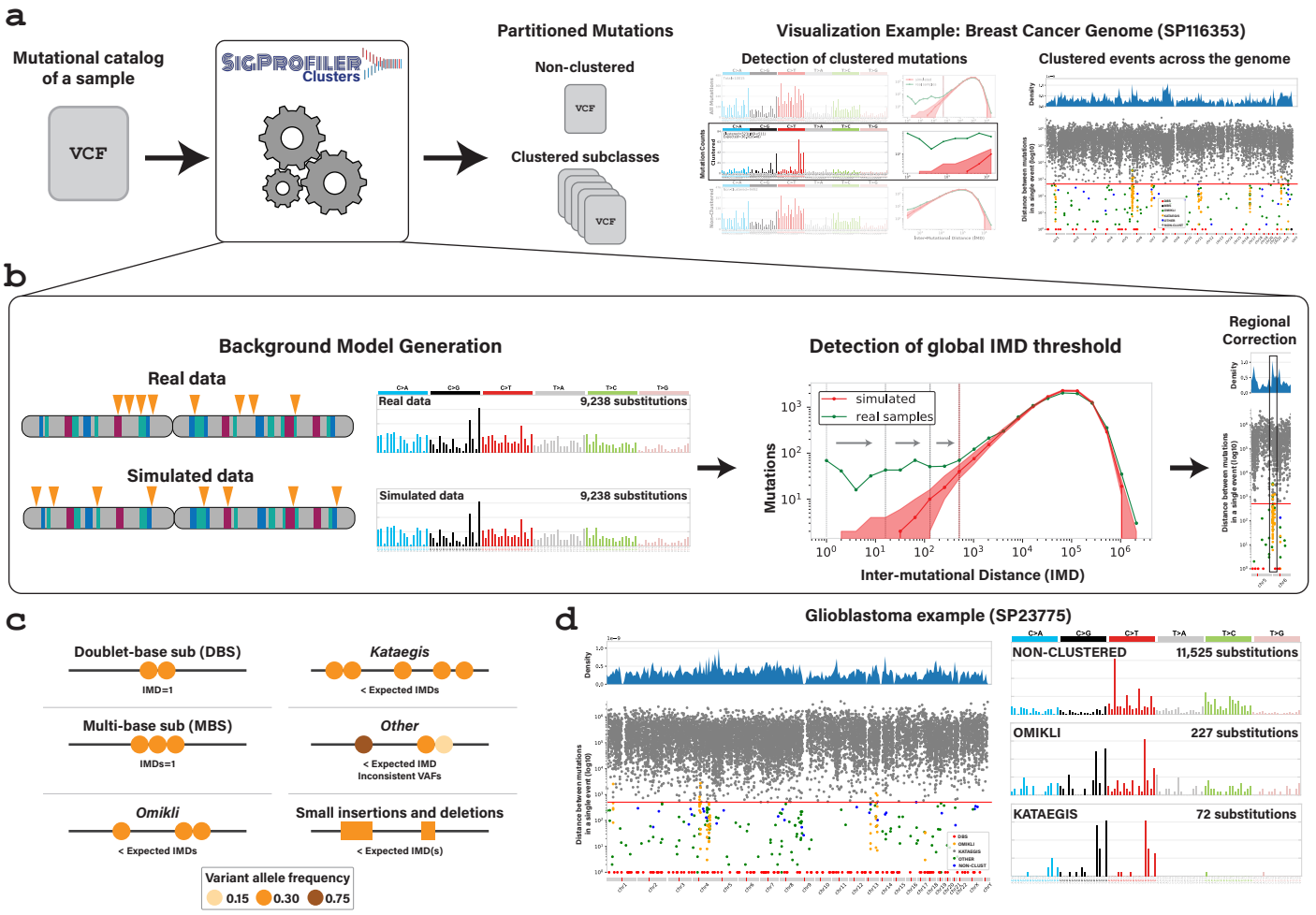
Figure 1. Detection and characterization of clustered mutations with SigProfilerClusters