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. Prior 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 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., 2012; Supek and Lehner, 2017; Taylor, et al., 2013). 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). 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; Alexandrov, et al., 2013; Chan, et al., 2015; D'Antonio, et al., 2016; Maciejowski, et al., 2020;

55 Nik-Zainal, et al., 2019; Taylor, et al., 2013). Many of these approaches utilize a piece-wise
56 linear regression to segment each chromosome, which, in most cases, is optimized for calling
57 larger strand-coordinated *kataegic* events (Alexandrov, et al., 2013; Lin, et al., 2021; Mas-Ponte
58 and Supek, 2020). Most prior methods have also ignored confounding effects attributed to
59 localized differences in mutation rates, copy number alterations, or the mutational burden across
60 each chromosome within a given sample. Further, the majority of existing tools focus on
61 detecting only a specific class of clustered events including doublet-base substitutions (Chen, et
62 al., 2013; Matsuda, et al., 1998), *kataegis* (D'Antonio, et al., 2016; Lin, et al., 2021; Taylor, et
63 al., 2013), or APOBEC3-associated events (Chan, et al., 2015; Nik-Zainal, et al., 2012) while
64 ignoring the larger landscape of clustered mutations. For example, a recent study (Mas-Ponte and
65 Supek, 2020) developed an algorithm focused on the detection of APOBEC3-associated *omikli*
66 and *kataegis* events in cancer genomes by incorporating simulations of somatic mutations and
67 estimates of cancer cell fractions.
68
69 Separation and classification of clustered events is required to fully elucidate the mutational
70 processes operating in cancer and normal somatic cells (Bergstrom, et al., 2022; Supek and
71 Lehner, 2017). Here we present SigProfilerClusters, a tool to comprehensively characterize and
72 subclassify clustered mutations from the complete catalog of mutations within the genome of a
73 single sample (**Fig. 1*a***). SigProfilerClusters classifies all types of clustered mutations, including
74 *(i)* doublet-base substitutions; *(ii)* multi-base substitutions; *(iii) omikli*; *(iv) kataegis*; and *(v)*
75 clustered small insertions and deletions (indels). The tool calculates a sample dependent IMD
76 threshold that considers regional differences in mutation rates, variant allele fractions of adjacent
77 mutations, and provides visualizations for downstream analyses (**Fig. 1*b&c***). 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 give 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 algorithm is implemented to efficiently derive the global IMD threshold for each genome. The final global IMD threshold is selected by ensuring that 90% of mutations below the chosen cutoff are unlikely to appear by chance given the simulated distribution of mutations (q-

101 value<0.01) with a maximum global IMD cutoff of 10 kilobases. The algorithm also considers
102 regional heterogeneities of mutation rates, generally associated with replication timing
103 (Stamatoyannopoulos, et al., 2009) or differential gene expression (Buisson, et al., 2019; Hess, et
104 al., 2019; Lawrence, et al., 2013; Pleasance, et al., 2010; Polak, et al., 2015), by correcting for
105 variance in clonality as well as variance in both mutation-dense and mutation-poor regions using
106 a sliding genomic widow (default size of 1 megabase). Specifically, an additional regional IMD
107 cutoff is corrected within each genomic window based upon the fold difference between the
108 number of real and the number of simulated mutations, while maintaining the original criteria of
109 less than 10% of mutations below the IMD cutoff appearing by chance (q-value<0.01). Lastly,
110 when data are available, SigProfilerClusters ensures that adjacent mutations are in the same cells
111 by introducing a maximum difference in variant allele frequencies (VAF) below a certain
112 threshold (default cutoff value of 0.10).
113
114 After identifying the set of clustered mutations, SigProfilerClusters subclassifies each clustered
115 substitution into a single category of previously established clustered events (Bergstrom, et al.,
116 2022). Briefly, all clustered substitutions with consistent VAFs are classified into one of four
117 categories. Two mutations with an IMD of 1 are classified as *doublet-base substitutions*, while
118 clusters of three or more adjacent mutations each with an IMD of 1 are classified as *multi-base*
119 *substitutions*. Clusters of two or three mutations with IMDs less than the sample-dependent
120 cutoff and with at least a single IMD greater than 1 are classified as *omikli*, while clusters of four
121 or more mutations with IMDs less than the sample-dependent cutoff and with at least a single
122 IMD greater than 1 are classified as *kataegis*. All remaining clustered mutations with

123 inconsistent VAFs are classified as *other*. Clustered indels are not subclassified into different
124 categories due to a lack of previously defined subtypes.
125
126 **USAGE**
127 SigProfilerClusters is freely available as a Python package, distributed under the permissive
128 BSD-2 clause license, and can be used on most operating systems including Windows, MacOS,
129 and Linux-based machines. The tool is compatible with large-scale deployments on high-
130 performance computing clusters as well as on cloud infrastructures such as Amazon Web
131 Services (AWS). Input data can be provided in the form of common mutation formats including
132 the Variant Call Format (VCF), the Mutation Annotation Format (MAF), or in the form of a
133 simple text file. The output of SigProfilerClusters results in the partitioning of all mutations into
134 a clustered or non-clustered directory. All clustered mutations are then classified into distinct
135 subcategories of events and provided individually in VCF files for downstream visualization and
136 analyses. The output for each subclass of clustered event can be directly utilized by additional
137 SigProfiler tools including SigProfilerExtractor for mutational signature analysis (Islam, et al.,
138 2020) and SigProfilerPlotting for examining patterns of somatic mutations (Bergstrom, et al.,
139 2019). The results for each sample are also summarized using two individual visualizations that
140 include: *(i)* a rainfall plot depicting the minimum global IMD between all adjacent mutations,
141 where each individual set of adjacent mutations is colored based upon its clustered classification;
142 and *(ii)* a multi-panel figure that displays the mutational patterns across all mutations, clustered
143 mutations, and non-clustered mutations, separately along with the distribution of IMDs across
144 the real and simulated data for each sample (**Fig. 1*a***).
145

146 **CONCLUSION**

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

159

160 **FIGURE LEGENDS**

161 **Figure 1. Detection and characterization of clustered mutations with SigProfilerClusters.** ***a***)
162 An example workflow used to detect clustered mutations in a single cancer genome. As an input,
163 SigProfilerClusters accepts common formats for mutations, such as ones in the variant calling
164 format (VCF), and the tool separates all clustered mutations from the complete mutational
165 catalog of the provided sample. Final partitions of mutations in the sample are outputted as VCF
166 files and visualized using the mutational spectra of all mutations, only clustered mutations, and
167 only non-clustered mutations along with a rainfall plot commonly used to show the distribution
168 of inter-mutational distances across a cancer genome (Alexandrov, et al., 2013; Bergstrom, et al.,

169 2022; Nik-Zainal, et al., 2012). ***b)*** Schematic demonstrating the process of calculating a sample
170 dependent IMD threshold to separate clustered from non-clustered mutations across each
171 genome. A binary search algorithm is used to efficiently detect the optimal global IMD threshold
172 for each sample. Detection of the global IMD threshold is illustrated using grey arrows. Regional
173 corrections are performed to identify local IMD thresholds based on variance of mutation rates
174 across the genome. ***c)*** Every clustered mutation is classified into a single subcategory of
175 clustered event. ***d)*** Rainfall plot illustrating the distribution of IMDs across a single glioblastoma
176 sample (*left*). The mutational spectra for *omikli* and *kataegic* events reveal a different mutational
177 pattern compared to the pattern of all non-clustered somatic mutations (*right*).
178

179 **ACKNOWLEDGEMENTS**

180 **Funding:** This was work was funded by Cancer Research UK Grand Challenge Award
181 C98/A24032 as well as US National Institute of Health grants R01ES030993-01A1 and
182 R01ES032547. LBA is also supported by a Packard Fellowship for Science and Engineering.
183 The funders had no roles in study design, data collection and analysis, decision to publish, or
184 preparation of the manuscript.
185

186 **Conflicts of Interest:** LBA is a compensated consultant and has equity interest in io9, LLC. His
187 spouse is an employee of Biotheranostics, Inc. LBA is also an inventor of a US Patent
188 10,776,718 for source identification by non-negative matrix factorization. ENB and LBA declare
189 a provisional patent application for utilizing clustered mutations as clinical prognostic
190 biomarkers in cancer. All other authors declare no competing interests.
191

Contributions: ENB developed the Python code and wrote the manuscript. ENB, MK, and NT tested and documented the code. LBA supervised the overall development of the code 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.

242 Pfeifer, G.P., You, Y.H. and Besaratinia, A. Mutations induced by ultraviolet light. *Mutat Res*
243 2005;571(1-2):19-31.
244 Pleasance, E.D.*, et al.* A comprehensive catalogue of somatic mutations from a human cancer
245 genome. *Nature* 2010;463(7278):191-196.
246 Polak, P.*, et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer.
247 *Nature* 2015;518(7539):360-364.
248 Roberts, S.A.*, et al.* Clustered mutations in yeast and in human cancers can arise from damaged
249 long single-strand DNA regions. *Mol Cell* 2012;46(4):424-435.
250 Stamatoyannopoulos, J.A.*, et al.* Human mutation rate associated with DNA replication timing.
251 *Nat Genet* 2009;41(4):393-395.
252 Stratton, M.R., Campbell, P.J. and Futreal, P.A. The cancer genome. *Nature*
253 2009;458(7239):719-724.
254 Supek, F. and Lehner, B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair
255 Targets Mutations to Active Genes. *Cell* 2017;170(3):534-547 e523.
256 Taylor, B.J.*, et al.* DNA deaminases induce break-associated mutation showers with implication
257 of APOBEC3B and 3A in breast cancer kataegis. *Elife* 2013;2:e00534.
258
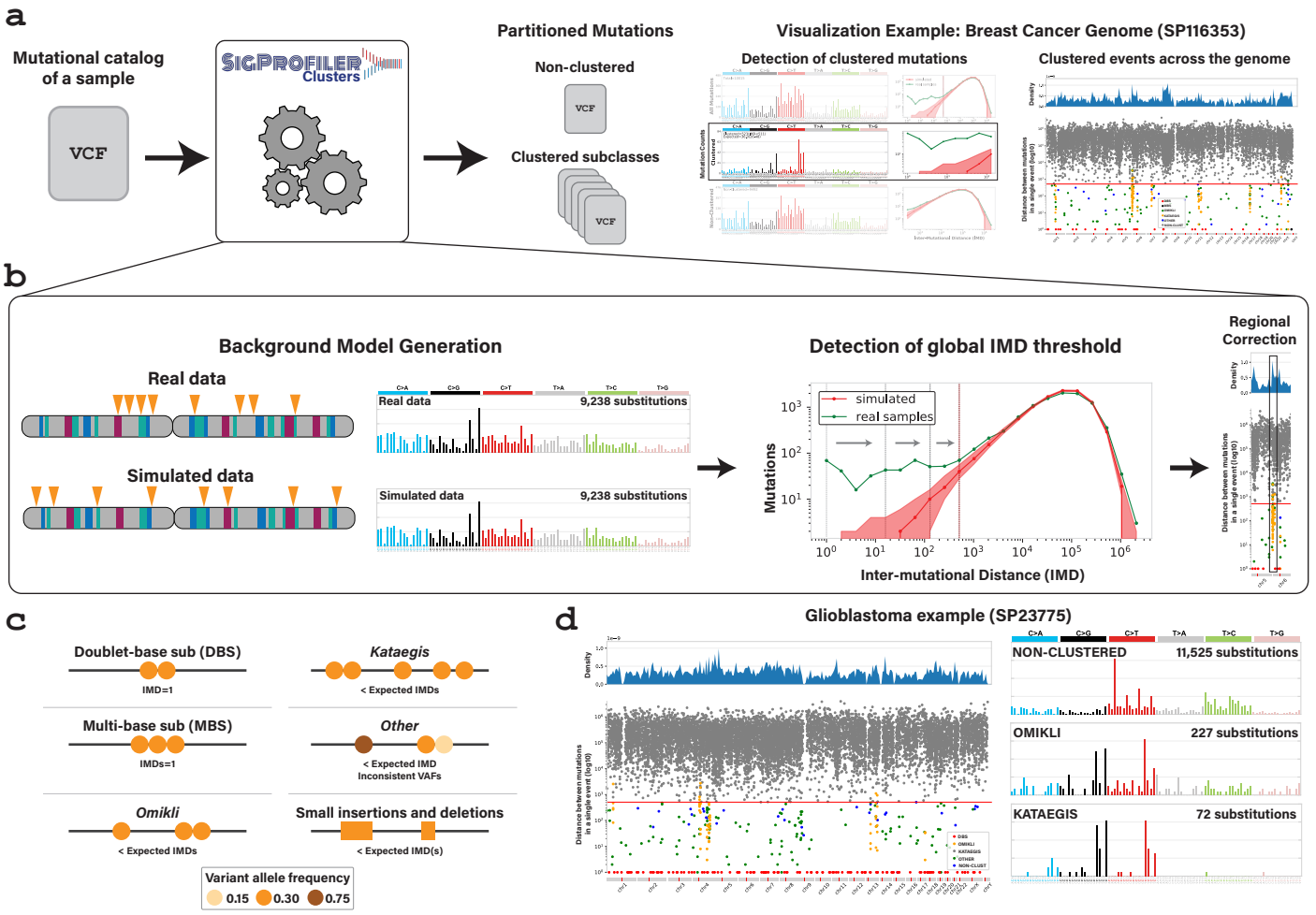
Figure 1. Detection and characterization of clustered mutations with SigProfilerClusters