CMSC499A: Interactive Mutation Signature Explorer

Mark Keller *

Spring 2018

Abstract

Identification and analysis of patterns in data can be difficult without visualization tools. Datasets of somatic mutations in cancer are no exception. Recent whole-genome sequencing projects have produced large amounts of mutation data to be explored. Of great importance is the classification of different mutational processes and the mutational signatures[2] they leave behind. As new mutational signatures continue to be discovered, observation of the levels of signature activity, called signature exposure, helps show how the underlying mutational processes differ across cancer types, time, and environmental variables. Visualizations of mutational signatures and mutation datasets in the context of these variables can help researchers to quickly observe trends and relationships in this data, verifying current knowledge or prompting further analysis. This paper describes a mutation visualization tool developed this semester that allows simple somatic mutation datasets and mutation signatures to be explored interactively and provides users with publication-ready graphics.

1 Introduction

Mutation signatures are distinctive patterns of somatic mutations generated by different mutational processes over the course of a lifetime[2]. Deamination, a biochemical reaction that removes an amine group from a molecule, is one process that has been linked to several mutational signatures[15]. These patterns have been extracted from mutation datasets using computational and mathematical techniques[3]. The concept of mutational signatures has been validated by in vitro experiments that perform knockouts of DNA repair genes and then observe that mutational signatures result[26,20].

Observation of mutational signatures and knowledge of the mutational processes that cause them can improve cancer diagnosis and treatment. Mutational signatures have been used to predict deficiency of BRCA1 and BRCA2, genes that produce tumor suppressor proteins[10]. Knowledge of such deficiencies in breast cancers can allow therapeutic response to be predicted as early as the first biopsy[8]. In colorectal cancers, mutational signatures have enabled identification of base excision repair (BER) defects[22]. BER is a repair process that enables altered DNA bases to be recognized and removed, with defects in this process causing multiple mutation patterns to arise[15]. Being able to attribute genomic instability to BER can prompt genetic counselling and immunotherapy measures that take into consideration a particular mutation load[22].

There are many open questions about mutational signatures. Currently, mutational signatures are represented as probability distributions over 96 categories of single-nucleotide variant (SNV)

*advised by Professor Max Leiserson

mutations. These 96 categories include the reference base, the variant base, and one flanking base pair in each direction. It is not known how significant these flanking base pairs are, how many to consider, or to what extent these depend on one another[23]. The mutational processes that cause some of the currently published mutational signatures remain unknown[2]. Methods to consider other types of mutations (indels and rearrangements) are only being formulated now[26].

Many interactive cancer visualization tools and browsers exist, but few consider or provide support for mutational signatures. For example, the cBio Cancer Genomics Portal provides interactive visualizations of mutation data with a focus on copy number alterations, mRNA expression changes, DNA methylation values, and protein and phosphoprotein levels[6,11]. The cBio Portal also contains features for computing mutual exclusivity and co-occurrence between pairs of genes[6,11]. The UCSC Xena Browser supports gene, exon, miRNA and protein expression, copy number, DNA methylation and somatic mutation data along with phenotypes, subtype classifications and genomic biomarkers[13,14]. FireBrowse is a cancer data browser that does include plots of mutation signature exposure that can be compared to clinical data, but only includes 5 signatures for analysis. It is unclear what these 5 signatures represent or if any of these are related to any of the widely-used signatures published by Catalogue Of Somatic Mutations In Cancer (COSMIC).

1.1 Contributions

A browser that can easily visualize mutation datasets and contributions of the over thirty currently known mutational signatures is needed, as it could allow trends and relationships with clinical data to be spotted easily. This could further understanding of the underlying mutational processes that cause certain signatures to be expressed.

Interactive Mutation Signature Explorer (iMuSE) is a web-based (`https://imuse.lrgr.io`) mutation data browser that enables exploration of mutational signatures, SNV mutations, and mutation density. This tool allows users to choose sequencing project datasets by cancer type, as well as mutational signature combinations (of which can be selected based on cancer-type-specific presets) before plotting this data. Web-based interactive visualizations allow for exploration across datasets and features in a way that is accessible and fast. Data exploration through these visualizations can be used to make initial inferences that lead to further statistical analysis, or, alternatively, can be used to verify and communicate relationships found through prior analysis.

Each type of visualization produced by iMuSE aims to enable further examination of a trend or relationship which has been hypothesized or presented in existing literature. The following are types of visualizations that can be generated and their motivations:

1. It is well known that tobacco and alcohol usage increases risk of many types of cancers[9,21]. A question of interest is how mutational signatures relate to smoking status and other clinical variables. Alexandrov *et al.* 2016 finds that smokers exhibit increases in mutations attributed to COSMIC signatures 2, 4, 5, 13, and 16[1]. Kim *et al.* 2016 also notes an association between signature 5 and smoking[17].

   To illustrate estimated contributions of a selected combination of mutational signatures, a stacked bar plot can be generated, showing estimated signature exposures for a sample, along with tobacco and alcohol usage indicators. Exposure values can be normalized to sum to one for each sample, or kept relative to the total number of mutations in each sample. This plot type was inspired by Figure 4 in Kim *et al.* 2016 that displays estimated contributions of 4 different signatures and tobacco usage over cohorts of urothelial cancer tumor samples[17].

2. Localized hypermutation, known as kataegis, was first discussed by Taylor *et al.* 2013[24], in which whole-genome sequencing of breast cancers showed regions with mutation "rainfalls" - greater than 5 mutations with significantly short intermutational distances. Mutations in regions of kataegis have been shown to occur frequently at C base pairs preceded by a 5-prime T base pair, due to APOBEC activity during double-strand break repair[24,2]. Since this introduction in 2013, kataegis has been shown to occur in additional tumor types[2]. COSMIC signatures 2 and 13 have been associated with APOBEC activity, and by extension, kataegis[2]. There is recent interest in rigorous statistical identification and verification of kataegis[25].

   To examine mutation clusters, specifically those occurring in regions of kataegis, a rainfall plot can be generated for each sample. Mutations are plotted horizontally based on their genome location, and vertically by the distance (in bp) to the previous mutation. Mutations are colored according to one of 96 mutation categories (5' flanking base pair, single-nucleotide variant, 3' flanking base pair). Rainfall plots are commonly used to examine regions of hypermutation, as they allow these regions to be easily identified by tight vertical clusters of mutations. Similar rainfall plots can be seen in Figure 4 of Nik-Zainal *et al.* 2012[19] and in Figure 6 of Alexandrov *et al.* 2013[2].

   To identify samples containing instances of localized hypermutation, kataegic events can be highlighted along the genome in a second type of plot. Along the vertical axis are samples, and along the horizontal axis is the genome. Users can zoom and pan along each chromosome, and easily pinpoint kataegis events by the dark bars located on mutations in kataegis regions. Samples are grouped by sequencing project and cancer type. This plot acts as a rainfall plot selector, as each sample bar can be clicked to generate a corresponding rainfall plot. To our knowledge, this style of plot has not been used before to visualize kataegis .

3. A question that prompted the development of this visualization tool is whether certain mutational signatures are associated with mutations in specific genomic locations or regions. One way to examine this is to look at mutation signature activity across the genome with a "Manhattan plot". The Manhattan plot gets its name from the resemblance of a skyline along the genome, and is typically used to visualize the output of statistical significance tests used in genome-wide association studies[12]. To generate this plot, signatures are assigned to individual mutations (see Methods section for details) and then grouped into bins by chromosome location and signature. The horizontal axis represents the genome location, and the vertical axis represents the number of mutations within a bin. To our knowledge, this is a new method for exploring mutational signatures along the genome.

The rest of this paper is organized into the following sections: Methods, Case Studies, and Conclusions. Included in the Methods section are discussions of open-source libraries and frameworks used to create the browser application, the steps taken to process raw mutation and clinical datasets and the sources of these datasets, and the algorithms used to compute signature exposures. The Case Studies section presents two use cases for this tool: exploration of the relationship between alcohol usage and mutational signatures, and verification of the mutation categories typically present in regions of hypermutation. Future directions for development of iMuSE are outlined in the Conclusions section.

2 Methods

2.1 Data

Data processing for plots occurs in two stages. Simple somatic mutation datasets and donor clinical datasets from sequencing projects are downloaded and processed into a uniform format. This conversion must be specified for each sequencing project, as each provides datasets in a different format. Currently, processing has been completed for simple somatic mutation and clinical datasets from the Pan-Cancer Analysis of Whole Genomes (PCAWG) study. Once into a uniform format, these datasets are stored in a bucket in the UMIACS object store, and can be processed dynamically by a web server as specific requests are made for visualizations. Usage of the object store allows future developers to quickly get the application up and running without needing to perform this initial time-consuming data processing step. Dynamic processing is performed by a web application written in Python with the Flask framework. The Pandas and NumPy packages are used for data manipulation in both of these processing stages.

2.2 Exposures

Signature exposures are estimated using a quadratic programming approach detailed by Huang *et al.* 2017[16]. To generate the aforementioned Manhattan Plot showing signatures across the genome, signatures are assigned to mutation categories for each sample by first computing estimated signature exposures, then taking the maximum of the product of each signature's probability for the mutation category and the sample's exposure to the same signature. Signatures present in each cancer type, or signature combination "presets", are based on those specified in publications of mutational signatures. For example, the presences of COSMIC signatures can be found in Figure 3 of Alexandrov *et al.* 2013[2].

2.3 Implementation and Availability

Using the Vue JavaScript framework, the application is made up of reusable components that encapsulate templates, functions, and variables. This encapsulation promotes modularity, and therefore ease of maintainability. Using the data-driven documents JavaScript library (D3.js)[5], each type of plot is tailored to suit the data types and data sets presented. Showing donor clinical variables, such as smoking and alcohol usage, and their relationship to mutation signatures, requires this fine control that D3 provides. In addition, D3 contains APIs for easy implementation of custom interactive features, such as highlighting, panning, and zooming. Interactivity extends beyond single plots, linking plots together based on variables such as chromosome region and donor.

Code for iMuSE is open-source and available through the following repositories: `https://github.com/lrgr/mutation-signature-explorer`, `https://github.com/lrgr/mutation-signature-exp` While this software can be run locally, accessing a hosted version may be easier for some users. Currently, an instance of the Flask-powered web server is deployed to Heroku. An instance of the front end (Vue/D3) component is deployed through GitHub pages, and uses Travis CI for continuous integration and deployment. This makes access and usage as simple as connecting via a web browser.

3 Case Studies

3.1 Alcohol Usage and Mutational Signatures

One use case for this tool is the analysis of relationships between signature exposures and clinical variables. Interactive sorting and dynamic calculations of signature exposures allow trends to be observed quickly. Although the relationship between liver cancer and alcohol consumption has long been documented[4], the relationship between mutational signatures and alcohol consumption has not been extensively examined. Recently, Li *et al.* 2018 identified a mutational signature associated with alcohol usage in esophageal squamous cell carcinoma that resembles COSMIC signatures 1, 2, 13 and 16[18]. To look for a relationship in liver cancer with iMuSE, we can select the PCAWG Liver Cancer - NCC, JP cohort, along with the COSMIC LICA signatures preset for liver cancer. Then, to examine the alcohol usage clinical variable, we can select the Signature Exposures with Clinical Data plot. Using the sorting functionality, we first sort by the alcohol usage clinical variable. With no trends immediately apparent, we can now sort by signature values. Sorting by COSMIC signature 12 and normalizing the signature exposures, we can see that only two alcohol-free donors are in the top 14 of 28 donors (Figure 1). Hovering over the samples and

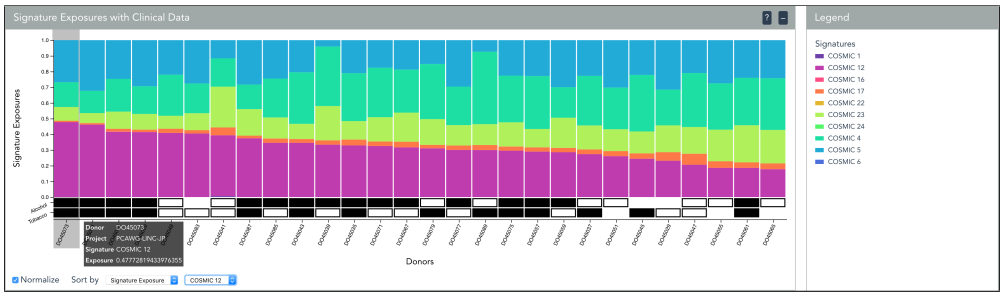


Figure 1: Screenshot of Signature Exposures with Clinical Data plot with PCAWG-LINC-JP data, normalized and sorted by COSMIC signature 12

looking at the y-axis, we can see that the contribution of signature 12 ranges from over 10% to just under 50% for all samples. Using this information, one could perform further statistical analysis on this data or seek out additional liver cancer mutation data to analyze.

Interestingly, when sorting by COSMIC signatures 1 and 16 - signatures both present in liver cancer and that resemble the Li *et al.* 2018 signature, a similar trend in alcohol usage arises (only two alcohol-free donors in the top half of donors). However, the contributions from these two signatures are too low to be visible for any sample in the plot.

3.2 Hypermutation and Mutation Category

Another use case is the analysis of the relationship between hypermutation and mutation category. For SNVs, the mutation category is the 5' flanking base pair, the mutation reference base pair, the variant base pair, and the 3' flanking base pair.

Mutations are assigned colors based on mutation categories on the rainfall plots to allow this type of relationship (or lack thereof) to be quickly spotted. The prevalence of mutations with reference base C and 5' T in regions of kataegis is well documented[24].

To verify this relationship with iMuSE, we can choose the PCAWG UK Breast Triple Negative/Lobular Cancer cohort and add a Kataegis Plot. Next, Rainfall Plots can be generated by choosing donors with kataegis events on chromosome 1. By hovering over mutations in kataegis regions on these plots and looking at the categories on the legend, we see an abundance of orange-colored C[C>T]X mutations and some blue T[C>T]X mutations in these regions. Figure 2 shows these rainfall plots and the mutation category legend. These plots could be used to discover further trends occurring in regions of hypermutation, such as significant rates of mutations that fall into categories with patterns other than T[C>T]X and C[C>T]X in specific cancer types. Such discoveries can be communicated to others easily, as iMuSE URLs include information about current plot states.

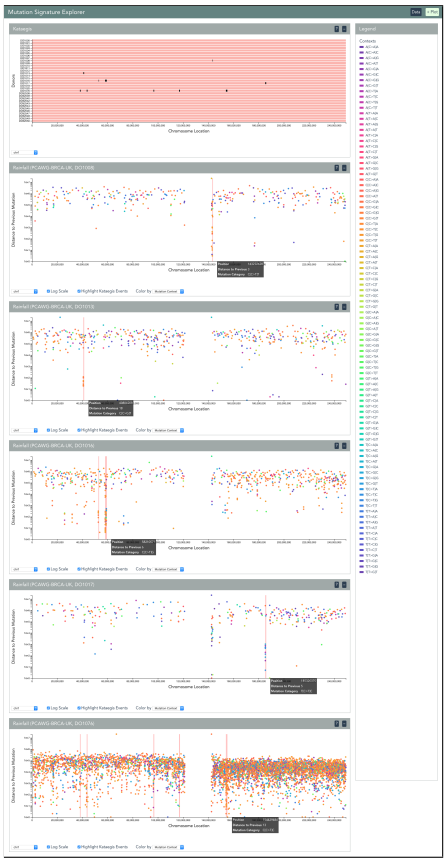


Figure 2: Screenshot of Rainfall Plots with PCAWG-BRCA-UK data for Chromosome 1

3.3 Discovery of Hypermutation Regions

As previously mentioned, there have been recent efforts to better detect instances of kataegis[25]. Alexandrov *et al.* 2013 describes an early definition of kataegis as a set of 6 or more mutations with an average intermutational distance of less than 1,000 base pairs[2]. To understand why a more rigorous statistical definition of kataegis is needed, we can look at regions of hypermutation with iMuSE rainfall plots. First, we can generate the kataegis plot for the PCAWG UK Breast Triple Negative/Lobular Cancer cohort and look at chromosome 1. Selecting donor with ID DO1017, the rainfall plot shows one kataegis region, highlighted near location 180,000,000. But to the left of this, we see another kataegic event that is not highlighted because it was not detected by this method (Figure 3).

4 Conclusions

In this paper, the motivations, features, design, and use cases of iMuSE, a web-based interactive visualization of mutational signature contributions, are presented. The ability to dynamically compute and visualize selected combinations of signatures, clinical variables, and cancer types makes this a powerful tool for data exploration.

Future development of iMuSE should focus on the addition of features that enable better analysis of single tumors. This could be implemented by allowing a user to trigger the creation of donor-specific visualizations in the same way that the rainfall plots are currently generated. These features would help to better understand the processes acting in a single tumor.

Functionality can also be expanded through implementation of more plot types. Specifically, visualization of mutational signatures themselves - as distributions over mutation categories - would be a simple but informative feature to add. This would give users that are unfamiliar with

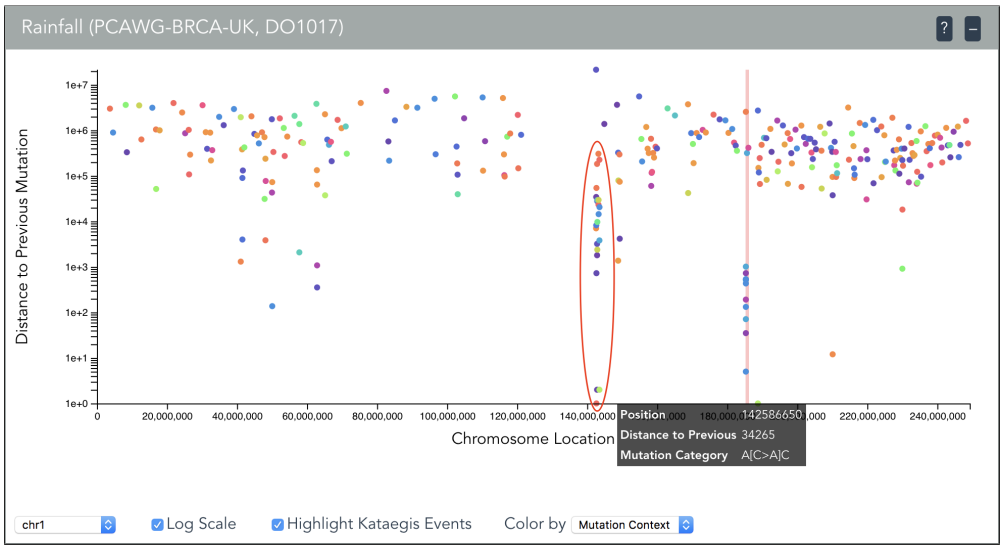


Figure 3: Screenshot of Rainfall Plot with overlay identifying undetected kataegis region

mutational signatures a better sense of how the signatures are used in the other plot types. These could look like interactive versions of the plots presented in Figure 2 of Alexandrov *et al.* 2013[2].

Visualization of mutation rates, in the form of interactive mutational prevalence plots, should be used to show relationships between mutation rates and variables such as cancer type, clinical data, mutational signatures, and kataegis rates. A mutation prevalence plot by cancer type is presented in Figure 1 of Alexandrov *et al.* 2013[2].

Another Manhattan plot should be implemented, showing mutation rates along the genome by mutation category.

A plot showing the location of genes should be linked to existing plots that present data along the genome. This type of plot showing gene locations can be found in Figure 1 of Chelaru *et al.* 2014[7].

Existing plots should be modified to allow analysis of mutation types other than SNVs. For example, since kataegis has been hypothesized to be associated with rearrangements, the rainfall and kataegis plots should be modified to show structural somatic mutations. The kataegis plot should also be modified to show survival status, which has also been hypothesized to be associated with kataegis.

Mutation datasets from additional sources should be processed to allow for stronger conclusions to be made using this tool. When additional mutational signatures are discovered, they should be added to the database to ensure that users have access to the most relevant and accurate visualizations possible. Local dataset visualization should be further enabled, to extend functionality beyond the public datasets we have processed.

References

[1] Ludmil B Alexandrov, Young Seok Ju, Kerstin Haase, Peter Van Loo, Iñigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, Akihiro Fujimoto, Hidewaki Nakagawa, Tatsuhiro Shibata, et al. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622, 2016.

[2] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415, 2013.

[3] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Peter J Campbell, and Michael R Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell reports*, 3(1):246–259, 2013.

[4] F Xavier Bosch, Josepa Ribes, Mireia Diaz, and Ramon Cleries. Primary liver cancer: worldwide incidence and trends. *Gastroenterology*, 127(5):S5–S16, 2004.

[5] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.

[6] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, 2012.

[7] Florin Chelaru, Llewellyn Smith, Naomi Goldstein, and Héctor Corrada Bravo. Epiviz: interactive visual analytics for functional genomics data. *Nature methods*, 11(9):938, 2014.

[8] Helen Davies, Dominik Glodzik, Sandro Morganella, Lucy R Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, Sancha Martin, Sandrine Boyault, Anieta M Sieuwerts, et al. Hrdetect is a predictor of brca1 and brca2 deficiency based on mutational signatures. *Nature medicine*, 23(4):517, 2017.

[9] Richard Doll and A Bradford Hill. Lung cancer and other causes of death in relation to smoking. *British medical journal*, 2(5001):1071, 1956.

[10] Susan M Domchek, Tara M Friebel, Christian F Singer, D Gareth Evans, Henry T Lynch, Claudine Isaacs, Judy E Garber, Susan L Neuhausen, Ellen Matloff, Rosalind Eeles, et al. Association of risk-reducing surgery in brca1 or brca2 mutation carriers with cancer risk and mortality. *Jama*, 304(9):967–975, 2010.

[11] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci. Signal.*, 6(269):pl1–pl1, 2013.

[12] Greg Gibson. Hints of hidden heritability in GWAS. *Nature genetics*, 42(7):558, 2010.

[13] Mary Goldman, Brian Craft, Teresa Swatloski, Melissa Cline, Olena Morozova, Mark Diekhans, David Haussler, and Jingchun Zhu. The ucsc cancer genomics browser: update 2015. *Nucleic acids research*, 43(D1):D812–D817, 2014.

[14] Mary Goldman, Brian Craft, Jingchun Zhu, Teresa Swatloski, Melissa Cline, and David Haussler. The ucsc xena system for integrating and visualizing functional genomics, 2016.

[15] Thomas Helleday, Saeed Eshtad, and Serena Nik-Zainal. Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics*, 15(9):585, 2014.

[16] Xiaoqing Huang, Damian Wojtowicz, and Teresa M Przytycka. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics*, 34(2):330–337, 2017.

[17] Jaegil Kim, Kent W Mouw, Paz Polak, Lior Z Braunstein, Atanas Kamburov, Grace Tiao, David J Kwiatkowski, Jonathan E Rosenberg, Eliezer M Van Allen, Alan D D'Andrea, et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature genetics*, 48(6):600, 2016.

[18] XC Li, MY Wang, M Yang, HJ Dai, BF Zhang, W Wang, XL Chu, X Wang, H Zheng, RF Niu, et al. A mutational signature associated with alcohol consumption and prognostically significantly mutated driver genes in esophageal squamous cell carcinoma. *Annals of Oncology*, 2018.

[19] Serena Nik-Zainal, Ludmil B Alexandrov, David C Wedge, Peter Van Loo, Christopher D Greenman, Keiran Raine, David Jones, Jonathan Hinton, John Marshall, Lucy A Stebbings, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, 2012.

[20] Serena Nik-Zainal, Jill E Kucab, Sandro Morganella, Dominik Glodzik, Ludmil B Alexandrov, Volker M Arlt, Annette Weninger, Monica Hollstein, Michael R Stratton, and David H Phillips. The genome as a record of environmental exposure. *Mutagenesis*, 30(6):763–770, 2015.

[21] Collaborative Group on Hormonal Factors in Breast Cancer et al. Alcohol, tobacco and breast cancer–collaborative reanalysis of individual data from 53 epidemiological studies, including 58 515 women with breast cancer and 95 067 women without the disease. *British journal of cancer*, 87(11):1234, 2002.

[22] Camilla Pilati, Jayendra Shinde, Ludmil B Alexandrov, Guillaume Assié, Thierry André, Zofia Hélias-Rodzewicz, Romain Ducoudray, Delphine Le Corre, Jessica Zucman-Rossi, Jean-François Emile, et al. Mutational signature analysis identifies mutyh deficiency in colorectal cancers and adrenocortical carcinomas. *The Journal of pathology*, 242(1):10–15, 2017.

[23] Yuichi Shiraishi, Georg Tremmel, Satoru Miyano, and Matthew Stephens. A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS genetics*, 11(12):e1005657, 2015.

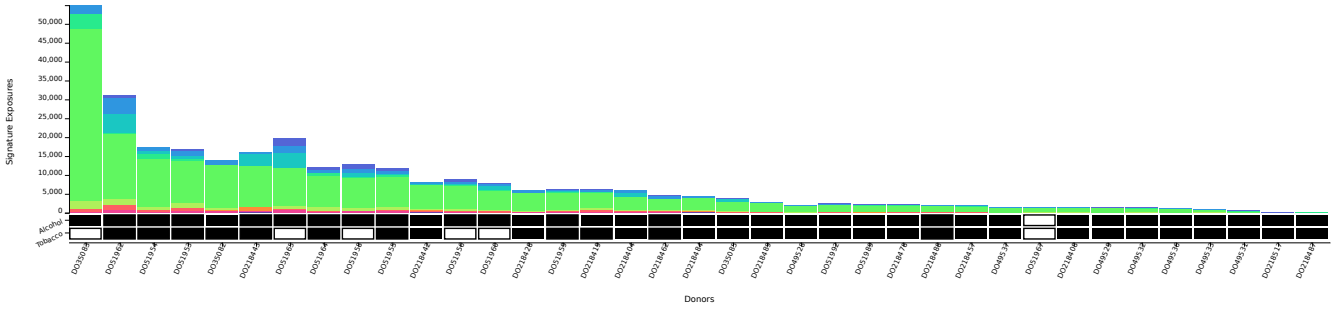
[24] Benjamin JM Taylor, Serena Nik-Zainal, Yee Ling Wu, Lucy A Stebbings, Keiran Raine, Peter J Campbell, Cristina Rada, Michael R Stratton, and Michael S Neuberger. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife*, 2, 2013.

[25] Fouad Yousif, Stephenie Prokopec, Ren X Sun, Fan Fan, Christopher M Lalansingh, David H Park, Lesia Szyca, PCAWG Network, and Paul C Boutros. The origins and consequences of localized and global somatic hypermutation. *bioRxiv*, page 287839, 2018.

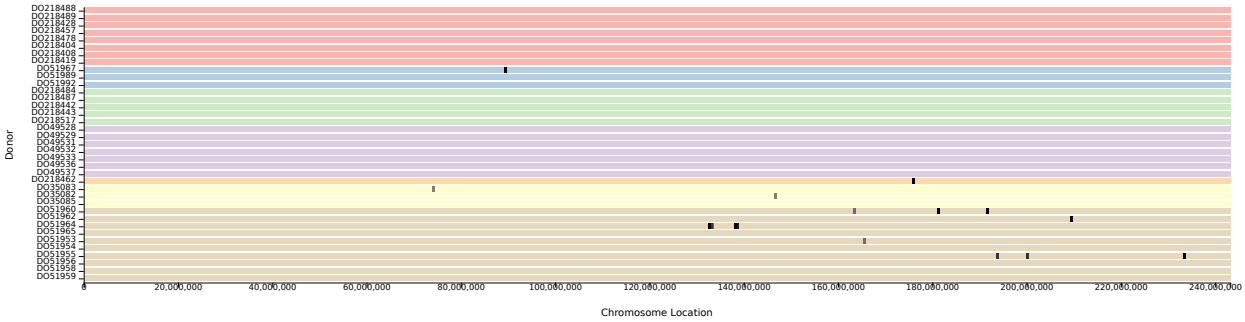
[26] Xueqing Zou, Michel Owusu, Rebecca Harris, Stephen P Jackson, Joanna I Loizou, and Serena Nik-Zainal. Validating the concept of mutational signatures with isogenic cell models. *Nature communications*, 9(1):1744, 2018.

Figure 4: Screenshot displaying all plot types

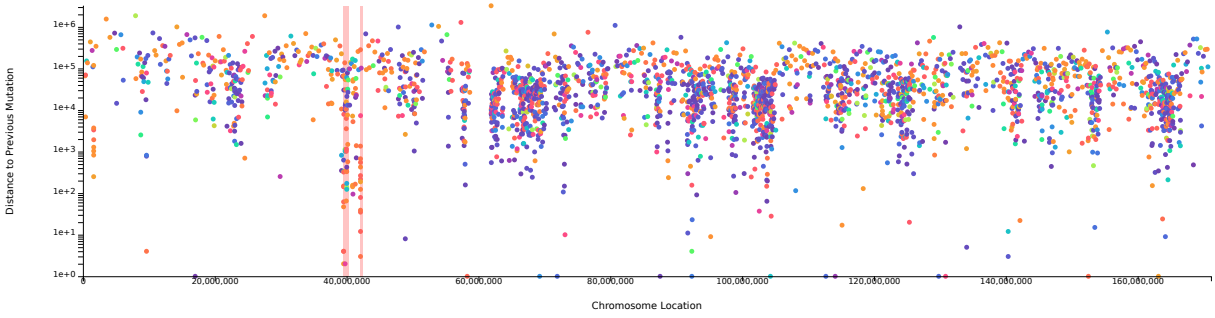
A Examples of Publication-Ready Figures



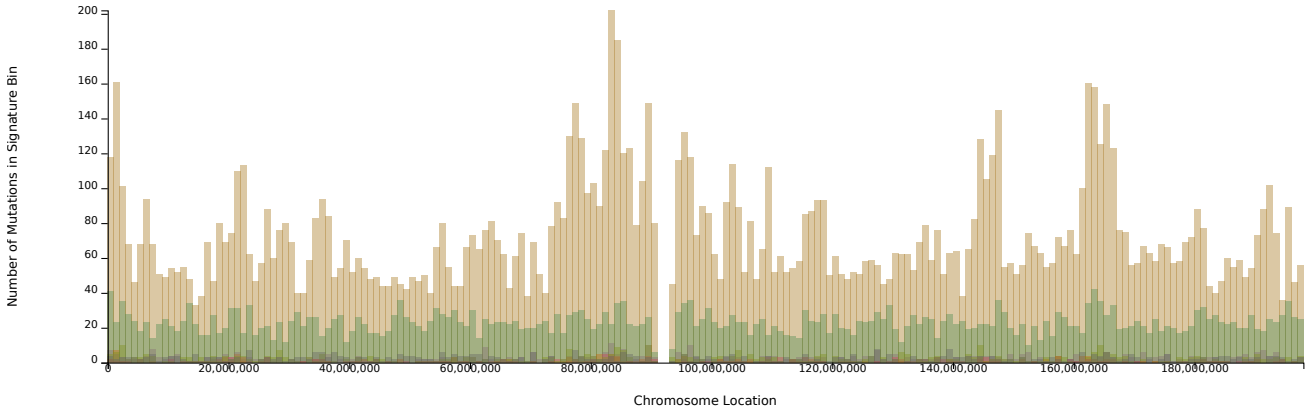
Example Figure 1: Signature Exposures with Clinical Variables



Example Figure 2: Kataegis Plot (Chromosome 2)



Example Figure 3: Rainfall Plot (Chromosome 6)



Example Figure 4: Manhattan Plot (Chromosome 2)