CMSC499A: 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. 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 Signature Explorer is a web-based (`https://mutation-signature-explorer.lrgr.io`) mutation data browser that enables exploration of mutational signatures, single-nucleotide variant (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 the Mutation Signature Explorer 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 [6,12]. 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 [9].

   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

---

[*]advised by Professor Max Leiserson

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[9].

2. Localized hypermutation, known as kataegis, was first discussed by Taylor *et al.* 2013[13], 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[13]. 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[14].

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[11] 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 genome-wide association studies[7]. To generate this plot, signatures are assigned to individual mutations and then grouped into bins by chromosome location and signature. A signature is assigned to each mutation category (5' flanking base pair, single-nucleotide variant, 3' flanking base pair) for each sample by first computing estimated signature exposures for a sample, 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. 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.

2 Case Study

2.1 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[3], 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[10]. To look for a relationship in liver cancer with the Mutation Signature Explorer, 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 looking at the y-axis, we can see that the contribution of signature
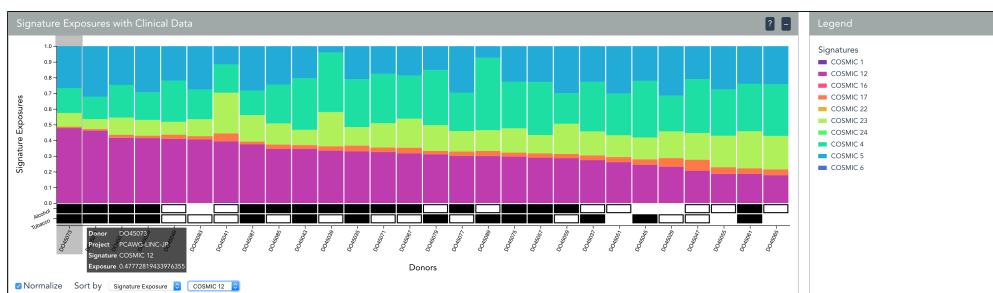

Figure 1: Screenshot of Signature Exposures with Clinical Data plot with PCAWG-LINC-JP data, normalized and sorted by COSMIC signature 12

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.

2.2 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[13].

To verify this relationship with the Mutation Signature Explorer, 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.
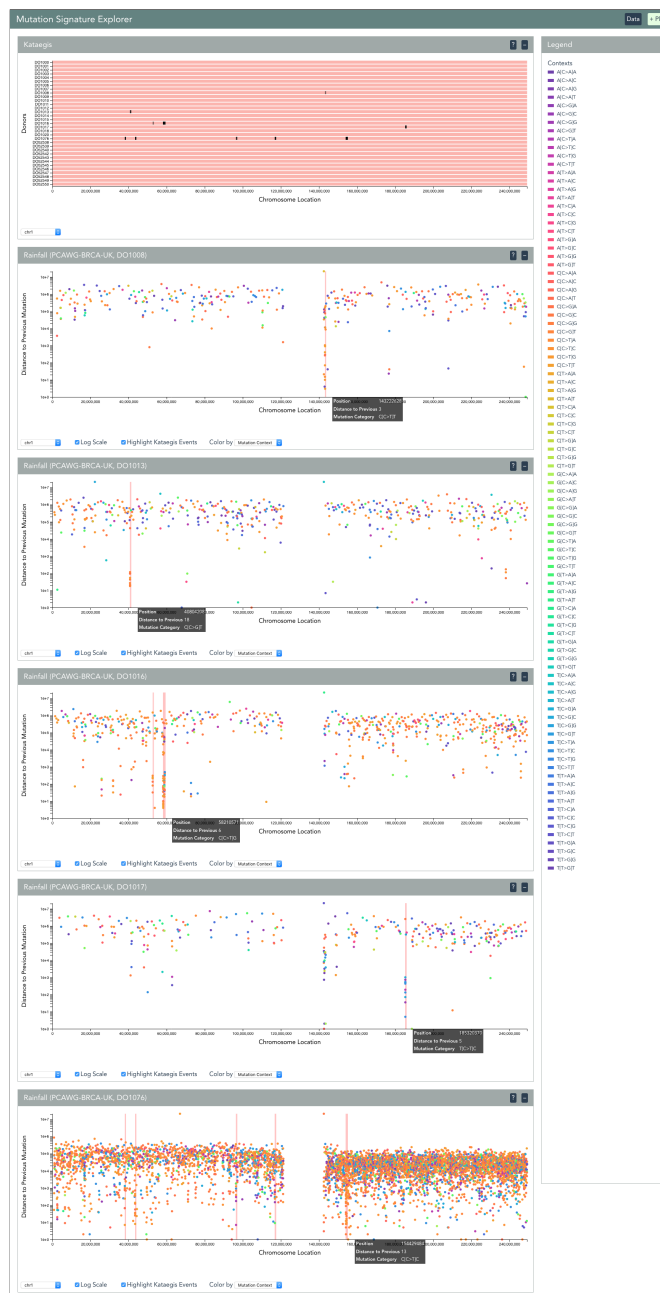

Figure 2: Screenshot of Rainfall Plots with PCAWG-BRCA-UK data for Chromosome 1

3   Methods

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)[4], 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.

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. 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.

Signature exposures are estimated using a quadratic programming approach detailed by Huang *et al.* 2017[8]. 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].

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.

4   Future Directions

Future development of the Mutation Signature Explorer should include the expansion of functionality 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 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.*

Figure 3: Screenshot
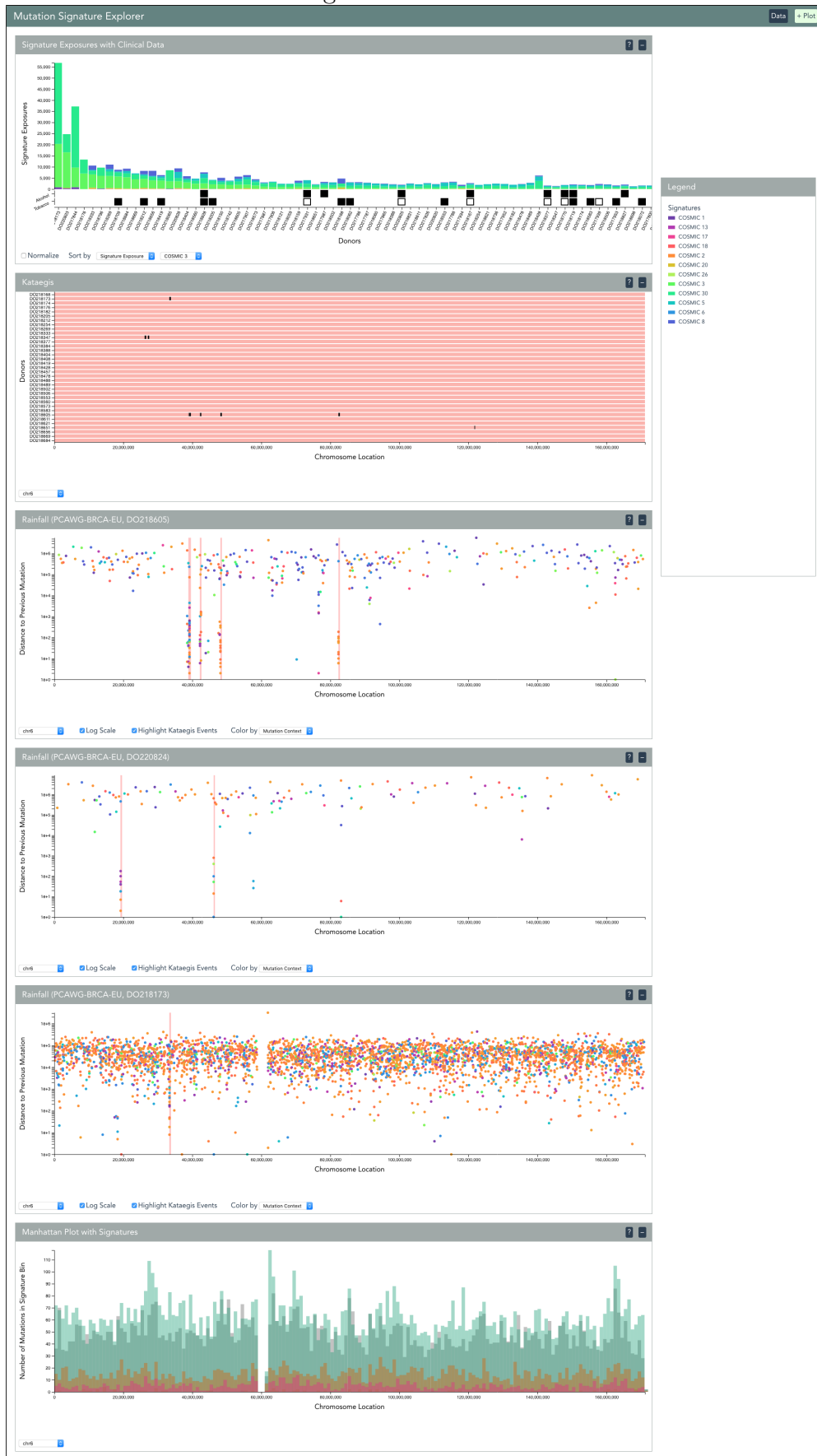
2014[5].

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

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

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

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

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

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

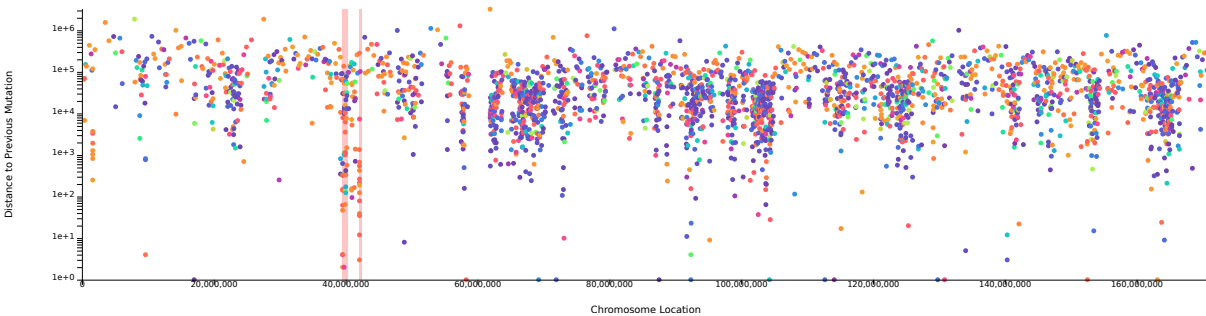# A   Examples of Publication-Ready Figures

Example Figure 1: Signature Exposures with Clinical Variables
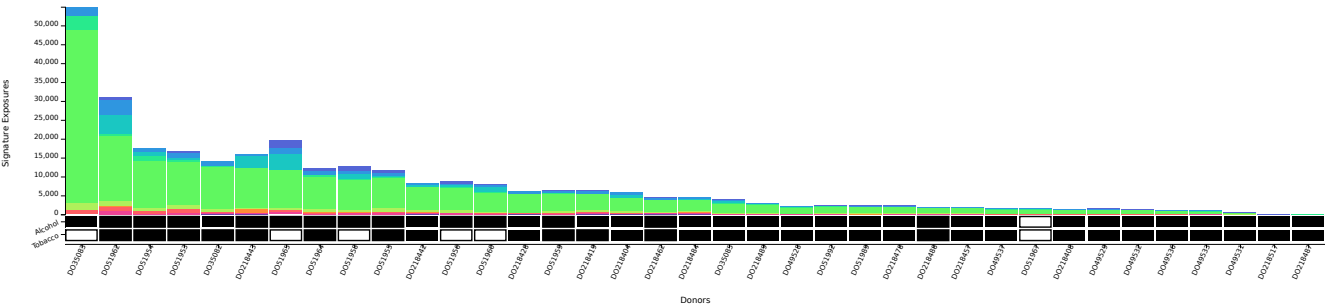
Example Figure 2: Kataegis Plot (Chromosome 2)
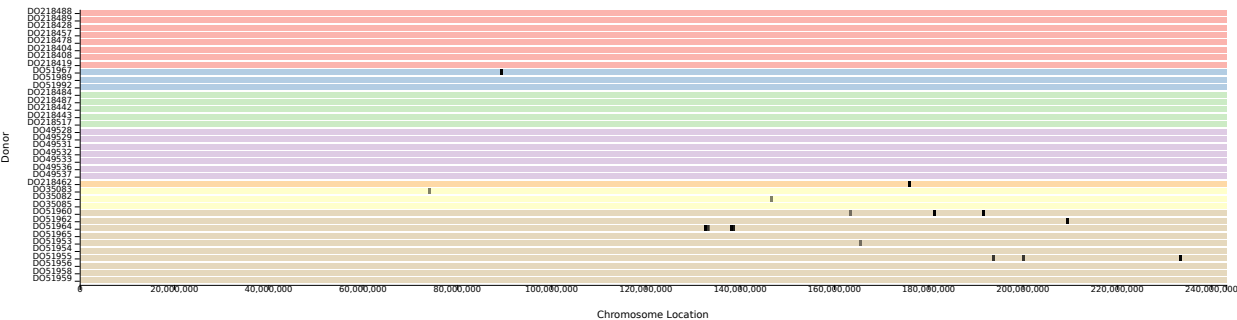
Example Figure 3: Rainfall Plot (Chromosome 6)

Example Figure 4: Manhattan Plot (Chromosome 2)