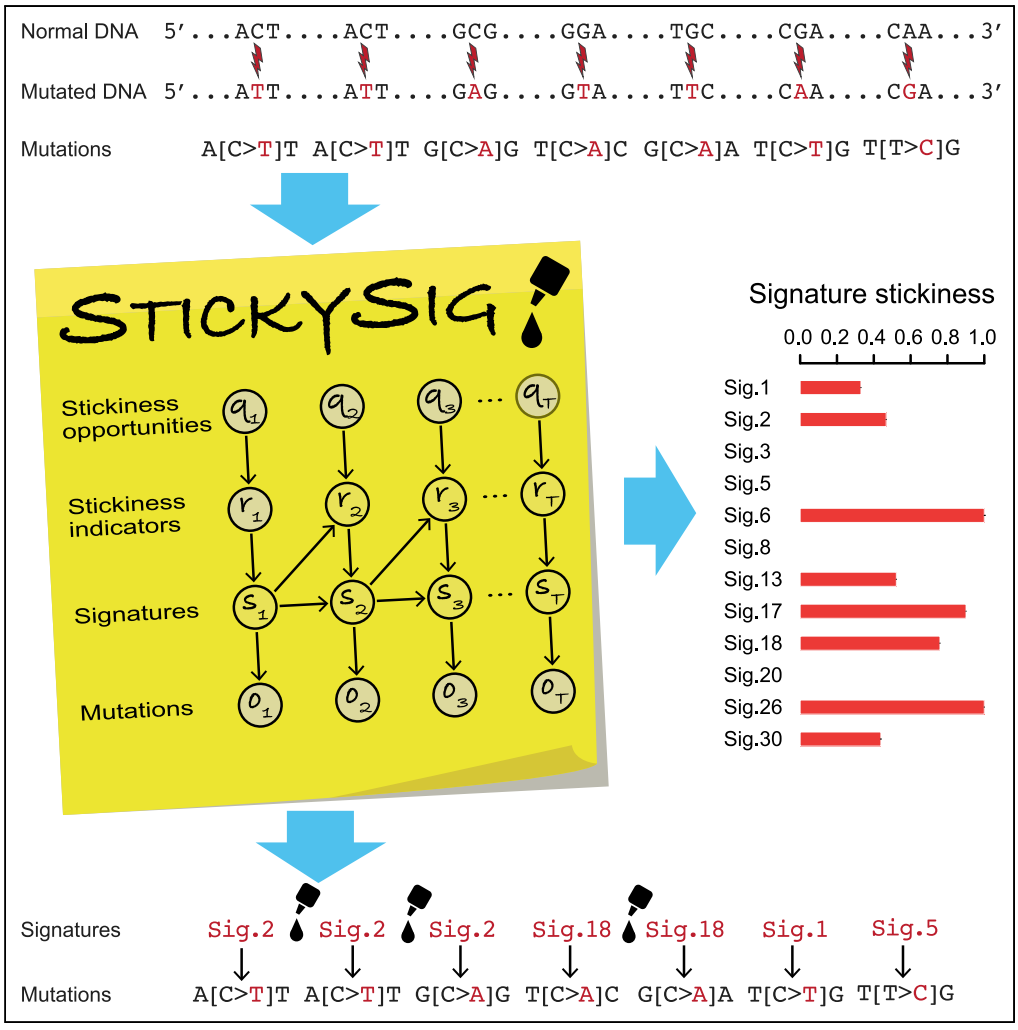
iScience

CellPress

Article

A Sticky Multinomial Mixture Model of Strand-Coordinated Mutational Processes in Cancer



Itay Sason, Damian Wojtowicz, Welles Robinson, Mark D.M. Leiserson, Teresa M. Przytycka, Roded Sharan

roded@tauex.tau.ac.il

HIGHLIGHTS

StickySig is a method for learning mutational signatures and their exposures

StickySig models the stickiness and strand coordination of mutational signatures

StickySig outperforms existing independent mutation models

StickySig reveals the role of Signature 18 in generating processive groups

DATA AND CODE AVAILABILITY

https://cancer.sanger.ac.uk/cosmic/signatures

Sason et al., iScience 23, 100900
March 27, 2020
https://doi.org/10.1016/j.isci.2020.100900

iScience

CellPress

Article

A Sticky Multinomial Mixture Model of Strand-Coordinated Mutational Processes in Cancer

Itay Sason,[1] Damian Wojtowicz,[2] Welles Robinson,[3] Mark D.M. Leiserson,[3] Teresa M. Przytycka,[2] and Roded Sharan[1,4,*]

SUMMARY

The characterization of mutational processes in terms of their signatures of activity relies mostly on the assumption that mutations in a given cancer genome are independent of one another. Recently, it was discovered that certain segments of mutations, termed processive groups, occur on the same DNA strand and are generated by a single process or signature. Here we provide a first probabilistic model of mutational signatures that accounts for their observed stickiness and strand coordination. The model conditions on the observed strand for each mutation and allows the same signature to generate a run of mutations. It can both use known signatures or learn new ones. We show that this model provides a more accurate description of the properties of mutagenic processes than independent-mutation achieving substantially higher likelihood on held-out data. We apply this model to characterize the processivity of mutagenic processes across multiple types of cancer.

INTRODUCTION

Mutational processes are key factors in shaping cancer genomes (Alexandrov et al., 2013a, 2013b; Helleday et al., 2014; Tubbs and Nussenzweig, 2017), and their characterization has important implications for understanding the disease and choosing targeted therapies (Davies et al., 2017a, 2017b; Polak et al., 2017; Gulhan et al., 2019). Multiple algebraic and statistical approaches have been suggested for the detection of mutational processes from somatic mutation data (Alexandrov et al., 2013a, 2013b; Fischer et al., 2013; Shiraishi et al., 2015; Kim et al., 2016; Rosales et al., 2016). These methods, which focus on single-base substitutions (consult Figure 1), are based on learning the pattern of mutations of each potential process as well as its activity (aka *exposure*) in any given tumor in a way that will best explain the observed mutation data. State-of-the-art approaches for learning mutational signatures include non-negative matrix factorization (NMF) methods (Alexandrov et al., 2013a, 2013b; Fischer et al., 2013; Kim et al., 2016; Rosales et al., 2016) that aim to explain the mutation counts as a sum over all signatures of the probability of a specific mutation to be generated by the respective signature times its exposure. Other approaches that borrow from the world of topic modeling (discovery of abstract topics in text documents) aim to provide a probabilistic model of the data so as to maximize the model's likelihood (Shiraishi et al., 2015; Funnell et al., 2018). However, most of these methods assume that mutations are independent of one another and cannot capture processes that create dependencies among them.

Recently, it was observed that some signatures operate in a strand-coordinated manner where consecutive mutations tend to mutate from the same reference allele and occur on the same strand (Nik-Zainal et al., 2014). Morganella et al. generalized these observations and found segments of such mutations (i.e., same reference allele and same strand) that they termed processive groups (Morganella et al., 2016). The length of a processive group, that is, the number of such consecutive mutations attributed to the same signature, is signature dependent. The significance and abundance of these processive groups suggested that certain mutational processes display stickiness and strand-coordination properties. In a previous work we have suggested a hidden Markov-based model for capturing sequential dependencies between close-by mutations (Wojtowicz et al., 2019). Here we follow a similar path and suggest novel probabilistic models for consecutive, albeit not necessarily close-by, mutations that occur on the same strand.

The biological reasons for this strand coordination are related, at least in part, to the asymmetric role that the two strands play in many cellular processes that operate on DNA. For example, the APOBEC C-to-U

[1]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, USA

[3]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA

[4]Lead Contact

*Correspondence: roded@tauex.tau.ac.il

https://doi.org/10.1016/j.isci.2020.100900

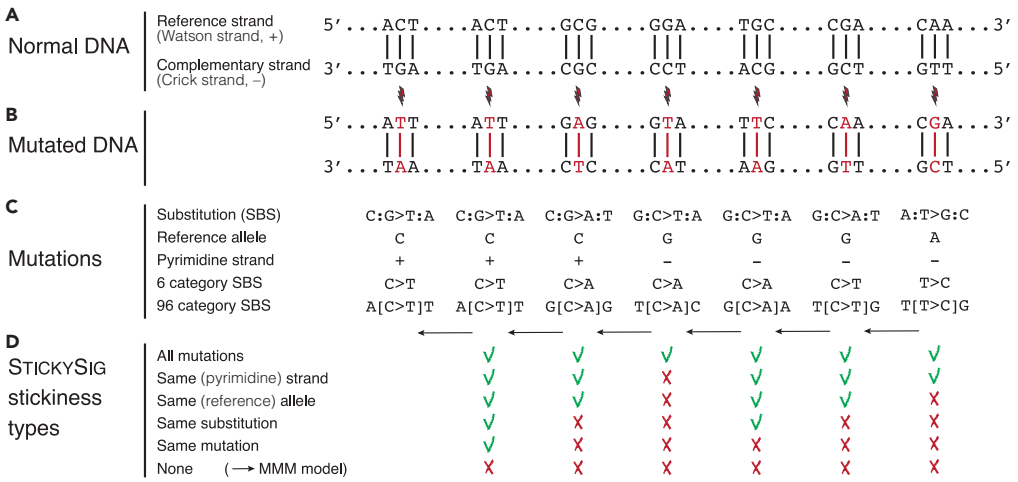


Figure 1. Definitions and Conventions

The figure shows normal DNA, mutated DNA, representation and characteristics of DNA mutations, and different types of stickiness used in StickySig model variants.

(A) The genome consists of the reference strand (the strand whose 5′-end is on the short arm of the chromosome), also known as the Watson strand or the plus strand, and the complementary strand, also known as the Crick strand or the minus strand.

(B) In the mutated DNA, changes in DNA base pair sequence are shown in red.

(C) Single base-pair substitution (SBS) can be represented as a transition in which one base pair in normal DNA is replaced by another in mutated DNA. The reference allele refers to the nucleotide that is found in the reference strand of normal genomic DNA. The pyrimidine strand is the strand (+or –) containing the pyrimidine base (C or T) in normal DNA. It is usually not known which base in a pair was the source of a mutation; thus, the convention is to annotate mutations from the pyrimidine base of the mutated base pair, leading to 6 substitution types (when context is not considered) or to 96 possible combinations of substitution types and neighboring bases.

(D) The StickySig model can use several types of stickiness opportunities: *all mutations* can be sticky, *same strand* stickiness, mutations having a pyrimidine base in the normal DNA on the same strand as the previous mutation; *same allele* stickiness, mutations having the same reference allele as the previous mutation; *same substitution* stickiness, mutations having exactly the same base-pair substitution as the previous mutation; *same mutation* stickiness, mutations having the same mutation features (96 mutation category and pyrimidine strand) as the previous mutation; and *none*, no stickiness allowed that leads to MMM model. Other types of stickiness can be also considered.

editing enzymes are a major source of mutations in many cancer types and are known to act on single-stranded DNA (Refsland and Harris, 2013). Many cellular processes, including replication and transcription, require strand separation leaving one or both strands exposed. Importantly, if the strands are separated, one of the strands is often more exposed than the other, leading to asymmetric strand coordination of APOBEC mutations. In particular, during DNA replication, the two DNA strands are processed differently. In this process, one of the strands (the lagging strand) is more exposed than the other strand (the leading strand). Owing to these differences, APOBEC asymmetry between these two strands is particularly strong (Haradhvala et al., 2016; Morganella et al., 2016; Tomkova et al., 2018; Seplyarskiy et al., 2016). In addition, leading and lagging strands are, among other differences, also processed by different polymerases, which might introduce different types of error in each strand leading to replication-related strand coordination. Transcription-coupled repair is another source of strand-specific mutagenesis. Another process leading to coordinated mutations and strand asymmetry is the formation DNA/RNA duplexes—the so-called R-loops. R-loops are thought to form co-transcriptionally when nascent messenger RNA hybridizes with the DNA template and thus can protect this strand from APOBEC activity and other types of mutations that act on single-stranded DNA. Indeed, multiple signatures have been found to have mutation strand bias in template versus non-template strands (Alexandrov et al., 2013a, 2013b; Haradhvala et al., 2016; Morganella et al., 2016; Tomkova et al., 2018).

Our suggested probabilistic model, StickySig, accounts for the stickiness and strand coordination of mutational signatures. The model captures independent mutations as well as processive groups in one probabilistic framework. In cross-validation tests on multiple datasets, StickySig outperforms independent-mutation models or sticky models that do not account for the strand information. We apply our model to gain

Cancer Type	#Samples	#Mutations	COSMIC Signatures
BRCA	560	3,479,652	1, 2, 3, 5, 6, 8, 13, 17, 18, 20, 26, 30
MALY	100	1,220,526	1, 2, 5, 9, 13, 17
CLLE	100	270,870	1, 2, 5, 9, 13

Table 1. Datasets Analyzed in This Study: Breast Cancer (BRCA), Malignant Lymphoma (MALY), and Chronic Lymphocytic Leukemia (CLLE)

new insights about the stickiness and exposures of known signatures, as well as in a *de novo* setting to learn new signatures.

RESULTS AND DISCUSSION

Mutation Data

For each dataset, we followed the standard approach introduced by Alexandrov et al. (2013a, 2013b) and classified mutations into $M = 96$ categories based on the 5′ flanking base, substitution type, and 3′ flanking base, following the convention that simple base-pair substitutions can be classified into six subtypes (Figures 1A–1C). (It is usually not known which base in a pair was the source of a mutation, thus the convention is to annotate substitutions from the pyrimidine base, i.e., G:C > A:T is written as C > T rather than G > A.) These mutations are assumed to be the result of the activity of K mutational processes, each of which is associated with a signature $S_i=(e_i(1), \ldots, e_i(M))$ of probabilities to emit each of the mutation categories. Henceforth, we denote the mutation categories observed in a given tumor by $o_1, \ldots, o_T$. We assume that o_i was emitted by signature s_i (whose identity is hidden from us).

In addition to mutation categories, the mutation data include information on the reference allele (the nucleotide that is on the Watson strand) and the pyrimidine strand (the Watson or Crick strand containing the pyrimidine base in normal DNA). By convention, the Watson strand is the reference genome strand (the strand whose 5′-end is on the short arm of the chromosome) and the Crick strand is the complementary strand. See Figures 1A–1C.

Data Description

We analyzed breast cancer (BRCA), chronic lymphocytic leukemia (CLLE), and malignant lymphoma (MALY) mutation datasets from whole-genome sequences from the International Cancer Genome Consortium (ICGC) (more information is available in the Data and Code Availability supplement section). We chose to study BRCA, CLLE, and MALY because they are known to have active signatures (Signatures 2 and 13) that were previously shown to display strand coordination (Morganella et al., 2016). In addition, each of the corresponding datasets contained at least 100 samples.

A Comparative Evaluation

We evaluated our suggested models and compared them with previous approaches using the datasets outlined above (Table 1). Here MMM serves as a stand-in for state-of-the-art non-probabilistic mutation signature methods such as non-negative matrix factorization, as MMM is a probabilistic method that encodes the standard assumption that each mutation in a tumor is independent of all others. In the Supplemental Informationwe show that MMM is in fact equivalent to a statistical variant of NMF, which is widely used for mutational signature analysis (Fischer et al., 2013; Kim et al., 2016).

We performed this comparison in two modes; the first was testing the models in a *refitting* mode when the signatures are known, and the second was testing the models in a *de novo* mode when the signatures are unknown. In the following comparisons, owing to running time considerations, we used a maximum of 100 iterations.

First, we focused on the refitting scenario, using the known COSMIC signatures. To this end, we use the leave-one-chromosome-out (LOCO) method. Specifically, we split samples by chromosomes and learned for each sample i the exposure vector π^i and the stickiness for the cosmic signatures α using all the chromosomes but one. We then report the log likelihood of the model on the left-out chromosome (summed

Model	LOCO Log likelihood		
	BRCA	MALY	CLLE
MMM	−13743198	−5235042	−1178024
StickySig	−13739451	−5232119	−1177527
StickySig-same-strand	−13736711	−5233842	−1177905
StickySig-same-allele	−13696283	**−5205381**	−1173271
StickySig-same-substitution	**−13549757**	−5206208	**−1173221**
StickySig-same-mutation	−13683356	−5227289	−1176916

Table 2. Performance Evaluation of MMM and StickySig Variants in a Refitting Setting Using the Leave-One-Chromosome-Out (LOCO) Method

In bold are the best values for each dataset.

across all samples and chromosomes). The results are summarized in Table 2 and clearly show the superiority of StickySig across the three cancer types analyzed. In each cancer type, StickySig has higher held-out likelihood than the independent mutation MMM, demonstrating that mutation signatures have stickiness that is shared across samples and that modeling this stickiness provides greater predictive power for held-out data. Furthermore, the difference between the variants of StickySig and MMM becomes much larger when StickySig is restricted to allow stickiness only between mutations with the same reference allele or the same base-pair substitution.

Next, we study the de novo scenario, where signatures are learned as part of the model training. For comparison purpose, we set the number of signatures to be the same as the number of active COSMIC signatures used above. To evaluate the models with respect to signature learning we used 10-fold sample cross-validation, where we learned e and α across all samples of the train data; then, for each sample X in the test set we fitted π and computed $\Pr[X|\pi,e,\alpha]$ and summed across all samples. This tests the model's capability to produce meaningful signatures that can explain well a new given sample. The results are summarized in Table 3 and show again that stickiness adds power to the model. Here again the leading models are StickySig-same-allele and StickySig-same-substitution.

Finally, we wished to assess the signatures learned by the algorithm. We trained the two best performing variants of StickySig (StickySig-same-allele and StickySig-same-substitution) on each of the three datasets. For evaluation purpose, we matched each signature learned by each model to its most similar COSMIC signature known to be active in the corresponding cancer type (measured via cosine similarity). The results are summarized in Figure 2. Evidently, StickySig-same-allele yields signatures that are more similar to the COSMIC ones. Note that the definition of signatures in our model is different from the standard one since they are coupled with a stickiness value, which in this application was always one. This may partially explain the deviation (particularly in the same-substitution case) from the COSMIC signatures.

Strand-Coordinated StickySig Defines Processive Groups in Breast Cancers

Morganella et al. defined processive groups as sets of adjacent substitutions of the same mutational signature sharing the same reference allele (Morganella et al., 2016). Our model, StickySig-same-allele, allows us to compute maximum likelihood estimates that sequences of mutations are generated in processive groups; hence, we could apply it to characterize the processivity of the different signatures in breast cancer. In order to compare and contrast our findings with those of Morganella et al. (2016), we used the same statistical test for the significance of a processive group of a given length. We confirmed the association of processive groups with Signatures 2, 6, 13, 17, and 26 when using the same length threshold of more than 10 (Morganella et al., 2016). In addition, our strand-coordinated model revealed that processivity is also a feature of Signature 18 (Figure 3A). The number of processive groups of length more than 10 was particularly high for Signatures 2 and 13 (Figure 3B), which is consistent with previous studies showing that APOBEC-related signatures demonstrated strand-coordinated mutagenesis (Nik-Zainal et al., 2014; Morganella et al., 2016).

Model	10-Fold CV Log likelihood		
	BRCA	MALY	CLLE
MMM	−13713230	−5200582	−1167727
StickySig	−13708734	−5187470	−1167238
StickySig-same-strand	−13702999	−5202680	−1167628
StickySig-same-allele	−13231739	**−4987710**	**−1121929**
StickySig-same-substitution	**−13135156**	−5020921	−1128093
StickySig-same-mutation	−13597846	−5187859	−1165856

Table 3. Performance Evaluation of MMM and StickySig Variants in a *De Novo* Setting Using 10-Fold Sample Cross-Validation (CV)

In bold are the best values for each dataset.

Next, we tested whether using the strand-coordinated StickySig rather than the mutation-independent MMM or the regular StickySig was important for the accurate discovery of processive groups. On average, the StickySig-same-allele model uncovered 133.9 groups of length greater than 10 in 41.9 patients, whereas MMM and StickySig models captured only 38.5 in 11 patients and 38.3 in 11.9 patients, respectively (Figure 3B). This is consistent with the large number of sticky mutations found by StickySig-same-allele model, even though StickySig has more sticky opportunities (Figures 3C and 3D). All these differences underscore the higher sensitivity of the strand-coordinated model for detecting processive groups, which may in part explain the observed differences in likelihood between MMM and StickySig models on held-out data (Table 2).

Processive groups, as summarized in Figures 3A and 3B, capture statistically significant patterns in cancer genomes and are considered features of specific signatures. An alternative characterization signature could be provided by the model parameter α—the "stickiness" of a signature—which is learned by the strand-coordinated model. Thus, we analyzed how these two views of strand-coordinated mutagenesis relate to each other. We considered only signatures for which there is a sufficient number of sticky mutations to properly learn this parameter (Figure 3E). For comparison purposes, we also included stickiness values computed with the strand-oblivious StickySig. We found that, in the strand-coordinated variant, the most sticky signatures were 1, 2, 6, 13, 17, 18, 26, and 30 (Figure 3F). Signatures 2, 6, 13, 17, 18, and 26 are exactly the same signatures that were found to be associated with processive groups. Although Signature 30 did not make the length 10 cutoff for processive groups, its processive segments are also relatively long. Interestingly Signatures 3 and 5 were not found to be sticky despite the fact that their processive groups were also quite long. In contrast, there is some stickiness to Signature 1 while its processive groups

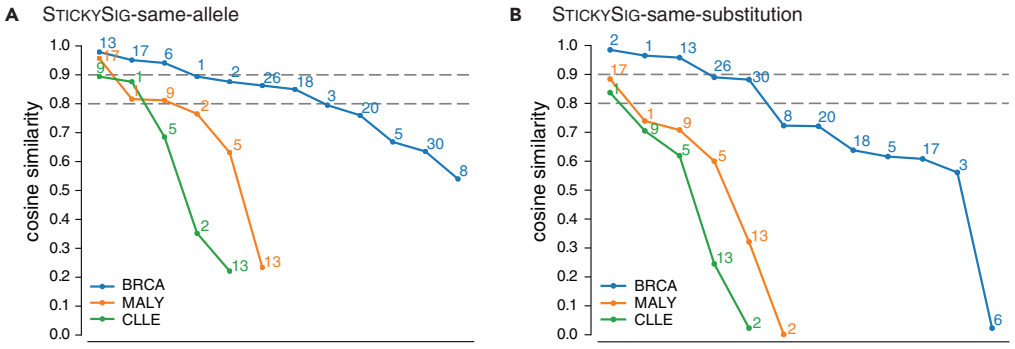


Figure 2. Signatures Learning

Performance of signature learning by StickySig-same-allele (A) and StickySig-same-substitution (B) on the three cancer datasets. For each case, depicted are the cosine similarities of the learned signatures to known COSMIC signatures, sorted from highest to lowest and computed by a maximum matching algorithm to prevent repetitions.

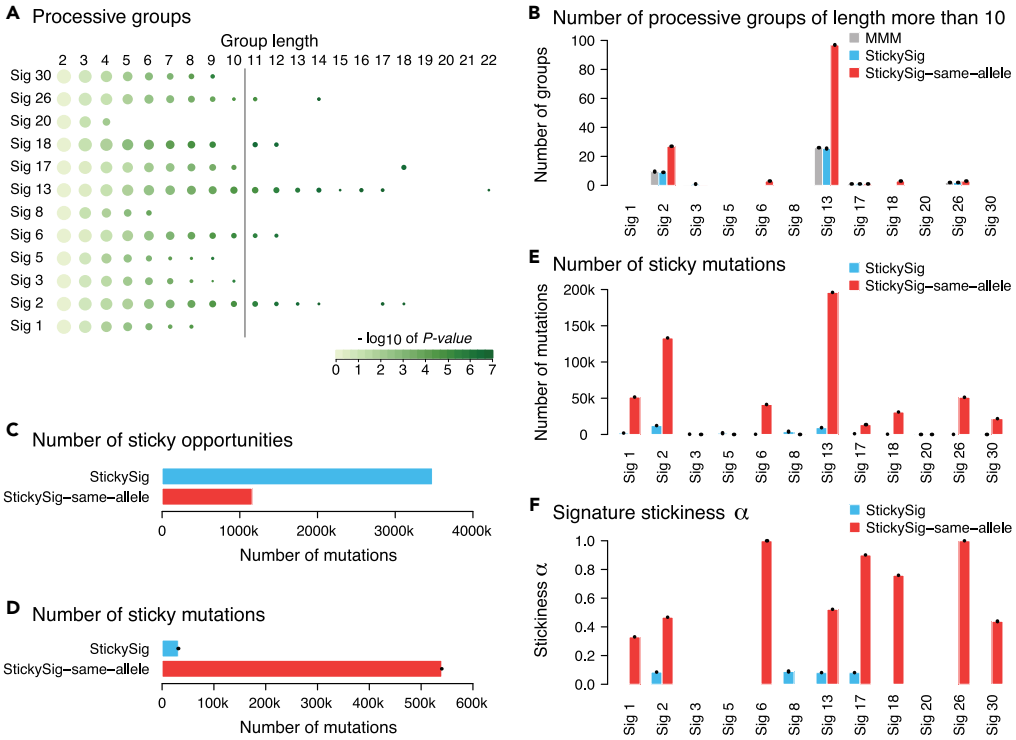


Figure 3. Stickiness in BRCA

(A) Relationship between processive group lengths (columns) and mutational signatures (rows) modeled by StickySig-same-allele. The size of each circle represents the number of groups (log10) observed for the specified group length and for each signature. The color of each circle corresponds to the p value of detecting a processive group of a given length in randomized data (-log10).

(B) The number of processive groups of length more than 10 for all signatures modeled by MMM (gray), StickySig (blue), and StickySig-same-allele (red).

(C) The total number of mutations can be sticky in StickySig (blue) and StickySig-same-allele (red).

(D) The total number of sticky mutations as modeled by StickySig (blue) and StickySig-same-allele (red).

(E) The number of sticky mutations for each signature as modeled by StickySig (blue) and StickySig-same-allele (red).

(F) Signature stickiness α as learned by StickySig (blue) and StickySig-same-allele (red). All bar plots show mean values with standard error of the mean (small black bars) from 10 random initializations of StickySig models.

are shorter than for 3, 5, and 30. The signature stickiness in the strand-oblivious StickySig model is minimal. The meaning of these intriguing findings is a subject for further investigation.

Limitations of the Study

Although in this work we focused on Watson/Crick strands, there is compelling evidence that other strand definitions may affect signatures and their activities. As we reviewed above, such categorization may be based on replication-based characteristics such as leading and lagging or transcription-based characteristics such as template and non-template. In that vein, a promising next step may be modeling multiple strand characteristics simultaneously, rather than considering them individually. For example, there is evidence in humans and other species that transcription and replication are co-oriented (Huvet et al., 2007; Srivatsan et al., 2010). Testing these different variants may reveal the role of strand characteristics in mutagenesis.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND CODE AVAILABILITY

For BRCA, we used the data from (Nik-Zainal et al., 2016) (ICGC release 22). For CLL and MALY we used ICGC release 27, analyzed the sample with the most mutations per patient, and restricted to those samples

annotated as "study = PCAWG" (Campbell et al., 2017). We used version 2 COSMIC signatures (https://cancer.sanger.ac.uk/cosmic/signatures) (Forbes et al., 2017) known to be active in the corresponding cancer type (enumerated in Table 1). StickySig and data download and processing is implemented in Python 3 and is available at https://github.com/itaysason/StickySig.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.100900.

ACKNOWLEDGMENTS

This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. D.W. and T.M.P. are supported by the Intramural Research Programs of the National Library of Medicine (NLM), National Institutes of Health, USA. R.S. was supported by Len Blavatnik and the Blavatnik Family Foundation as well as by a research grant from the Israel Science Foundation (grant no. 715/18). We thank Mark Keller for his help in processing mutation datasets. This manuscript is a development of a manuscript published in the proceedings of RECOMB 2019 (Sason et al., 2019).

AUTHOR CONTRIBUTIONS

I.S. and R.S. conceived the study and methodology. I.S., D.W., and W.R. performed the computational experiments and analyzed the results. M.D.M.L. and T.M.P. provided expertise and feedback. I.S. implemented the software. All authors discussed the results and contributed to the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 18, 2019
Revised: January 23, 2020
Accepted: February 5, 2020
Published: March 27, 2020

REFERENCES

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al. (2013a). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013b). Deciphering signatures of mutational processes operative in human cancer. Cell Reports *3*, 246–259.

Campbell, P.J., Getz, G., Stuart, J.M., Korbel, J.O., and Stein, L.D. (2017). Pan-cancer analysis of whole genomes. bioRxiv, 162784, https://doi.org/10.1101/162784.

Davies, H., Glodzik, D., Morganella, S., Yates, L.R., Staaf, J., Zou, X., Ramakrishna, M., Martin, S., Boyault, S., Sieuwerts, A.M., and Simpson, P.T. (2017a). HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat. Med. *23*, 517–525.

Davies, H., Morganella, S., Purdie, C.A., Jang, S.J., Borgen, E., Russnes, H., Glodzik, D., Zou, X., Viari, A., Richardson, A.L., and Børresen-Dale, A.L. (2017b). Whole-genome sequencing reveals breast cancers with mismatch repair deficiency. Cancer Res. *77*, 4755–4762.

Fischer, A., Illingworth, C.J., Campbell, P.J., and Mustonen, V. (2013). EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biol. *14*, 1–10.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., et al. (2017). COSMIC: somatic cancer genetics at high- resolution. Nucleic Acids Res. *45*, D777–D783.

Funnell, T., Zhang, A., Shiah, Y.-J., Grewal, D., Lesurf, R., McKinney, S., Bashashati, A., Wang, Y.K., Boutros, P.C., and Shah, S.P. (2018). Integrated single-nucleotide and structural variation signatures of DNA-repair deficient human cancers. bioRxiv, 267500, https://doi.org/10.1101/267500.

Gulhan, D.C., Lee, J.J., Melloni, G.E., Cortés-Ciriano, I., and Park, P.J. (2019). Detecting the mutational signature of homologous recombination deficiency in clinical samples. Nat. Genet. *51*, 912–919.

Haradhvala, N.J., Polak, P., Stojanov, P., Covington, K.R., Shinbrot, E., Hess, J.M., Rheinbay, E., Kim, J., Maruvka, Y.E., Braunstein, L.Z., et al. (2016). Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. Cell *164*, 538–549.

Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. Nat. Rev. Genet. *15*, 585–598.

Huvet, M., Nicolay, S., Touchon, M., Audit, B., d'Aubenton-Carafa, Y., Arneodo, A., and Thermes, C. (2007). Human gene organization driven by the coordination of replication and transcription. Genome Res. *17*, 12781285.

Kim, J., Mouw, K.W., Polak, P., Braunstein, L.Z., Kamburov, A., Tiao, G., Kwiatkowski, D.J., Rosenberg, J.E., Van Allen, E.M., D D'Andrea, A., and Getz, G. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat. Genet. *48*, 600–606.

Morganella, S., Alexandrov, L.B., Glodzik, D., Zou, X., Davies, H., Staaf, J., Sieuwerts, A.M., Brinkman, A.B., Martin, S., Ramakrishna, M., et al. (2016). The topography of mutational processes in breast cancer genomes. Nat. Commun. *7*, 11383.

Nik-Zainal, S., Wedge, D.C., Alexandrov, L.B., Petljak, M., Butler, A.P., Bolli, N., Davies, H.R., Knappskog, S., Martin, S., Papaemmanuil, E., and Ramakrishna, M. (2014). Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. Nat. Genet. *46*, 487–491.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560

breast cancer whole-genome sequences. Nature *534*, 4754.

Polak, P., Kim, J., Braunstein, L.Z., Karlic, R., Haradhavala, N.J., Tiao, G., Rosebrock, D., Livitz, D., Kübler, K., Mouw, K.W., and Kamburov, A.A. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nat. Genet. *49*, https://doi.org/10.1038/ng.3934.

Refsland, E., and Harris, R. (2013). The APOBEC3 family of retroelement restriction factors. Curr. Top. Microbiol. Immunol. *371*, 1–27.

Rosales, R.A., Drummond, R.D., Valieris, R., Dias-Neto, E., and da Silva, I.T. (2016). signeR: an empirical Bayesian approach to mutational signature discovery. Bioinformatics *33*, 8–16.

Sason, I., Wojtowicz, D., Robinson, W., Leiserson, M.D., Przytycka, T.M., and Sharan, R. (2019). A sticky multinomial mixture model of strand-coordinated mutational processes in cancer. In Research in Computational Molecular Biology, L.J. Cowen, ed. (Springer), pp. 243–255.

Seplyarskiy, V., Soldatov, R.A., Popadin, K.Y., Antonarakis, S.E., Bazykin, G.A., and Nikolaev, S.I. (2016). APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. Genome Res. *26*, 174–182.

Shiraishi, Y., Tremmel, G., Miyano, S., and Stephens, M. (2015). A simple model-based approach to inferring and visualizing cancer mutation signatures. PLoS Genet. *11*, e1005657.

Srivatsan, A., Tehranchi, A., MacAlpine, D.M., and Wang, J.D. (2010). Co-orientation of replication and transcription preserves genome integrity. PLoS Genet. *6*, e1000810.

Tomkova, M., Tomek, J., Kriaucionis, S., and Schuster-Böckler, B. (2018). Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol. *19*, 129.

Tubbs, A., and Nussenzweig, A. (2017). Endogenous DNA damage as a source of genomic instability in cancer. Cell *168*, 644–656.

Wojtowicz, D., Sason, I., Huang, X., Kim, Y.A., Leiserson, M.D., Przytycka, T.M., and Sharan, R. (2019). Hidden Markov models lead to higher resolution maps of mutation signature activity in cancer. Genome Med. *11*, 49.

iScience, Volume *23*

Supplemental Information

A Sticky Multinomial Mixture Model of Strand-Coordinated Mutational Processes in Cancer

Itay Sason, Damian Wojtowicz, Welles Robinson, Mark D.M. Leiserson, Teresa M. Przytycka, and Roded Sharan

Transparent Methods

Model specification and training

A *multinomial mixture model (MMM)* assumes the following generative process for the mutation data. For each mutation, independently of all others, a signature $s \in S$ is drawn from a multinomial distribution $\pi = (\pi_1, \ldots, \pi_K)$; subsequently, the mutation category is drawn from S_i according to the signature's emission probabilities. The model parameters can be learned using the Expectation-Maximization (EM) algorithm (Wojtowicz et al. 2019).

Here we propose a sticky MMM model, StickySig, that allows signatures to emit more than one mutation at a time. In the basic model, any two consecutive mutations can stick. In a refined version of the model, two consecutive mutation can stick if they share some predefined property, such as being on the same strand. Examples of stickiness opportunities are shown in Figure 1D and the StickySig model is sketched in Figure 1S.

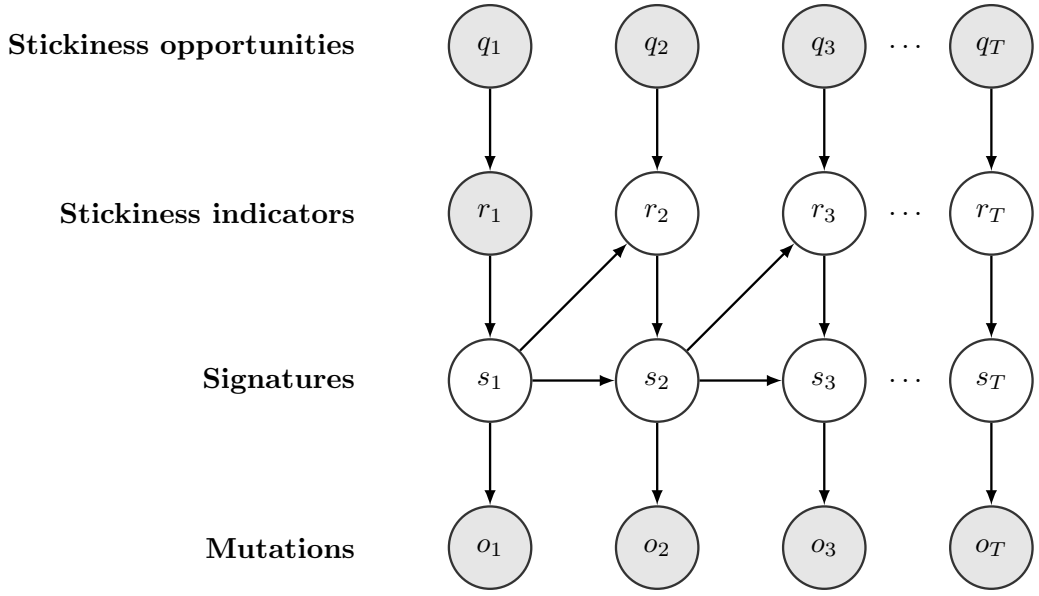


Figure 1S: **A sketch of StickySig. Related to Figure 1.**

Formally, StickySig is parameterized by a $K \times M$ matrix e of signature emission probabilities, signature start probabilities $\pi = (\pi_1, \ldots, \pi_K)$ that are assumed to be sample-specific, and signature stickiness values $\alpha =$

$(\alpha_1, \ldots, \alpha_K)$ that are shared across samples. The observed data for this model are the mutations sequence O and the stickiness opportunity Q which is 1 if there is opportunity and 0 otherwise. The hidden variables for this model are the signatures S and the stickiness indicator R which indicates whether or not the current mutation came from the same signature as the previous one. For simplicity, we omit sample indices below and focus the description on a single sample. The model can be described by the following conditional probability distributions:

- $\Pr[o_t = m | s_t = S_i] = e_i(m)$
- $\Pr[r_{t+1} = 1 | s_t = S_i, q_{t+1} = q] \begin{cases} 0 & q = 0 \\ \alpha_i & q = 1 \end{cases}$
- $\Pr[s_{t+1} = S_j | r_{t+1} = 1, s_t = S_i] = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$
- $\Pr[s_{t+1} = S_i | r_{t+1} = 0] = \pi_i$

Here, t is a mutation index that ranges from 1 to T (T – number of mutations) and m is a mutation category. Note that if we choose q_t to be always 0, the model reduces to MMM.

We developed an Expectation Maximization (EM) algorithm to learn the parameters of this model from data. The EM algorithm is described in (Sason et al. 2019) and runs in $O(Tn)$ time per iteration for T mutations and n signatures. The EM model training is controlled by two parameters: Maximum number of iterations and Tolerance, which is used to decide on convergence when the relative improvement in log-likelihood falls below it, and is set to $1e-10$ throughout.

Model variants

We derived several model variants implementing different types of stickiness (Figure 1D):

- MMM - no stickiness allowed, i.e $q_t = 0$.
- StickySig - this is the regular StickySig with no strand information, thus any pair of consecutive mutations can stick, i.e $q_t = 1$.

- StickySig-same-strand - same-strand stickiness, i.e $q_t = \begin{cases} 0, & \text{ps}_{t-1} \neq \text{ps}_t \\ 1, & \text{ps}_{t-1} = \text{ps}_t \end{cases}$ where ps_t is the pyrimidine strand of mutation t.
- StickySig-same-allele - same reference allele stickiness, i.e $q_t = \begin{cases} 0, & \text{ra}_{t-1} \neq \text{ra}_t \\ 1, & \text{ra}_{t-1} = \text{ra}_t \end{cases}$ where ra_t is the reference allele of mutation t.
- StickySig-same-substitution - same base-pair substitution stickiness, i.e $q_t = \begin{cases} 0, & \text{bs}_{t-1} \neq \text{bs}_t \\ 1, & \text{bs}_{t-1} = \text{bs}_t \end{cases}$ where bs_t is the base-pair substitution of mutation t.
- StickySig-same-mutation - same mutation stickiness, i.e $q_t = \begin{cases} 0, & \text{mf}_{t-1} \neq \text{mf}_t \\ 1, & \text{mf}_{t-1} = \text{mf}_t \end{cases}$ where mf_t are the mutation features (mutation category – o_t, and pyrimidine strand – ps_t) of the mutation in position t.

For each model variant, we performed training in two modes: *refit*, where signatures are fixed in advance, and *de-novo*, where the signatures are learned as part of the training.

Equivalence of MMM and statistical-NMF

In statistical-NMF (Lee and Seung 2000; Regli and Silva 2018), which is widely used for analyzing mutational signatures (Fischer et al. 2013; Kim et al. 2016), we optimize following optimization problem:

$$\begin{aligned} \min_{W,H} \quad & \sum_{i,j}(\sum_{k=1}^{K} w_{ik}h_{kj}) - v_{ij}\log(\sum_{k=1}^{K} w_{ik}h_{kj}) \\ \text{subject to} \quad & w_{ik}, h_{kj} \geq 0 \end{aligned} \tag{1}$$

where V is the mutation counts matrix of size $N \times M$ and W, H are of size $N \times K, K \times M$ respectively. In the MMM model we optimize the following:

$$\begin{aligned} \min_{W,H} \quad & -\sum_{i,j} v_{ij}\log(\sum_{k=1}^{K} w_{ik}h_{kj}) \\ \text{subject to} \quad & (w_{i1}, ..., w_{iK}) \in \Delta^K \\ & (h_{k1}, ..., h_{kM}) \in \Delta^M \end{aligned} \tag{2}$$

Where $\Delta^n = \{(x_1, ...x_n) | x_i \geq 0 \ \ \& \ \sum_{i=1}^{n} x_i = 1\}$ the simplex.

We begin by observing that for equation (1), each solution is a actually a class of solution by moving the weights of H's rows to W's columns. We can lose this ambiguity by requiring that H's rows will sum to 1:

$$\begin{aligned} \min_{W,H} \quad & \sum_{i,j} (\sum_{k=1}^{K} w_{ik} h_{kj}) - v_{ij} \log(\sum_{k=1}^{K} w_{ik} h_{kj}) \\ \text{subject to} \quad & w_{ik} \geq 0 \\ & (h_{k1}, ..., h_{kM}) \in \Delta^M \end{aligned} \tag{3}$$

Now denote by r_i the weight of row i in W, and $\tilde{W}$ to be W with rows summed to 1, i.e $r_i = \sum_{k=1}^{K} w_{ik}$, $\tilde{w}_{ik} = w_{ik}/r_i$. We change the optimization problem again:

$$\begin{aligned} \min_{R,\tilde{W},H} \quad & \sum_{i,j} r_i (\sum_{k=1}^{K} \tilde{w}_{ik} h_{kj}) - v_{ij} \log(\sum_{k=1}^{K} \tilde{w}_{ik} h_{kj}) - v_{ij} \log r_i \\ \text{subject to} \quad & r_i \geq 0 \\ & (\tilde{w}_{i1}, ..., \tilde{w}_{iK}) \in \Delta^K \\ & (h_{k1}, ..., h_{kM}) \in \Delta^M \end{aligned}$$

Because of the conditions we see that $\sum_{j,k} \tilde{w}_{ik} h_{kj} = 1$, so we can re-write this as:

$$\begin{aligned} \min_{R,\tilde{W},H} \quad & \sum_{i} \left(r_i - \log(r_i) \sum_{j=1}^{M} v_{ij} - \sum_{j=1}^{M} v_{ij} \log(\sum_{k=1}^{K} \tilde{w}_{ik} h_{kj}) \right) \\ \text{subject to} \quad & r_i \geq 0 \\ & (\tilde{w}_{i1}, ..., \tilde{w}_{iK}) \in \Delta^K \\ & (h_{k1}, ..., h_{kM}) \in \Delta^M \end{aligned}$$

We can derive this function with respect to r_i and get:

$$\frac{\partial f}{\partial r_i} = 1 - \frac{\sum_{j=1}^{M} v_{ij}}{r_i} = 0$$

And from here we can see that $\sum_{j=1}^{M} v_{ij}$ is the value for r_i to optimize this problem. After assigning the optimal value for R and dropping constants we are left with the optimization problem (2).
We can now use this process to go from a solution of the MMM problem to a solution of the statistical-NMF problem: Given (W, H) that solves (2), denote $R = diag(\sum_{j=1}^{M} v_{1j}, ..., \sum_{j=1}^{M} v_{Nj})$, then (RW, H) is a solution for (1). Similarly we can convert a solution for (1) to a solution for (2).

References

Wojtowicz, D. et al. (2019). "Hidden Markov models lead to higher resolution maps of mutation signature activity in cancer". In: *Genome Medicine* 11, p. 49. DOI: 10.1186/s13073-019-0659-1.

Sason, I. et al. (2019). "A Sticky Multinomial Mixture Model of Strand-Coordinated Mutational Processes in Cancer". In: *Research in Computational Molecular Biology*. Ed. by L. J. Cowen. Springer, pp. 243–255. DOI: 10.1007/978-3-030-17083-7_15.

Lee, D. D. and H. S. Seung (2000). "Algorithms for Non-negative Matrix Factorization". In: *Proceedings of the 13th International Conference on Neural Information Processing Systems*. NIPS'00. Denver, CO: MIT Press, pp. 535–541.

Regli, J.-B. and R. Silva (2018). *Alpha-Beta Divergence For Variational Inference*. arXiv: 1805.01045 [stat.ML].

Fischer, A. et al. (2013). "EMu: probabilistic inference of mutational processes and their localization in the cancer genome". In: *Genome Biology* 14.4, pp. 1–10. DOI: 10.1186/gb-2013-14-4-r39.

Kim, J. et al. (2016). "Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors". In: *Nature Genetics* 48.6, pp. 600–606. DOI: 10.1038/ng.3557.