A sequence context-based germline filter for structural variant calling from tumor samples without paired normal

Wolu Chukwu[1,2,*], Siyun Lee[1,2,*], Alexander Crane[1,2,*], Shu Zhang[1,2,*], Ipsa Mittra[1], Marcin Imielinski[3], Rameen Beroukhim[1,2], Frank Dubois[1,2,4,,#], and Simona Dalin[1,2,#]

1. Cancer Program, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.
2. Departments of Cancer Biology and Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA.
3. Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA; New York Genome Center, New York, NY, USA; Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA; Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA; Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA; Department of Pathology and Perlmutter Cancer Center, NYU Grossman School of Medicine, New York, NY, USA.
4. Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin; Humboldt-Universität zu Berlin, Institute of Pathology; Berlin Institute of Health at Charité-Universitätsmedizin Berlin, Berlin, Germany.

* Equal contributors
# Jointly supervised

Corresponding Authors: Simona Dalin, sdalin@broadinstitute.org and Frank Dubois frank.dubois@charite.de

Abstract

There is currently no method to distinguish between germline and somatic structural variants (SVs) in tumor samples that lack a matched normal sample. In this study, we analyzed several features of germline and somatic SVs from a cohort of 974 patients from The Cancer Genome Atlas (TCGA). We identified a total of 21 features that differed significantly between germline and somatic SVs. Several of the germline SV features were associated with each other, as were several of the somatic SV features. We also found that these associations differed between the germline and somatic classes, for example, we found that somatic inversions were more likely to be longer events than their germline counterparts. Using these features we trained a support vector machine (SVM) classifier on 555,849 TCGA SVs to computationally distinguish germline from somatic SVs in the absence of a matched normal. This classifier had an ROC curve AUC of 0.984 when tested on an independent test set of 277,925 TCGA SVs. In this dataset, we achieved a positive predictive value (PPV) of 0.81 for an SV called somatic by the classifier being truly somatic. We further tested the classifier on a separate set of 7,623 SVs from pediatric high-grade gliomas (pHGG). In this non-TCGA cohort, our classifier achieved a PPV of 0.828, showing robust performance across datasets.

Introduction

Large-scale whole-genome sequencing (WGS) efforts have revealed a prominent role of recurring structural variants (SVs) as cancer drivers[1] and as biomarkers of disruption of the DNA damage response and other processes[2,3]. These studies relied on paired normal tissue to call germline SVs and remove them from the somatic calls. However, while paired tumor-normal WGS was the standard in these well-funded initiatives, in many cases normal tissue is not available. This is frequently the case in clinical settings[4] or in the study of cell line models. Furthermore, the study of the signatures resulting from the treatment effects of drugs on cell lines as a biomarker for various processes is easily overshadowed by the large number of preexisting germline events in cell lines[5,6].

Due to their wide span, SV can affect multiple genes and up to a third of the average cancer genome[7]. SV can also alter the activity of distant cancer genes in multiple complex ways[8,9]. Thus, SVs can affect the cancer genome at least as profoundly as SNVs. However, we currently lack a tool that allows the reliable detection of SVs from tumor-only sequencing comparable to our capacity to detect SNVs.

In contrast to the identification of somatic SNVs (single nucleotide variants) from tumor-only sequencing, a simple filter against a catalog of known germline variants has limitations in the case of SVs. A database of germline SVs has been developed by the gnomAD group, yet it is magnitudes smaller than that of SNVs[10]. This is in part due to the rarity of germline SVs compared to germline SNVs. Additionally, all SVs are intrinsically harder to detect than SNVs[11]. Despite recent improvements in SV-detection algorithms, different callers frequently result in slightly different breakpoint positions and show performance differences depending on the span of the SV[10,12]. A simple reference matching by positions is therefore not possible to filter germline SVs.

There are several features of the sequence context of an individual SV that could be predictive of it being a germline event. For example, the presence of repeat elements, the span of the SV, the length of homology at its breakpoints, and proximity to protein-coding sequences all have been shown to be predictive of the probability of somatic SV formation in cancer[2]. Thanks to the recent large WGS efforts[13] we now have sufficient WGS data to train a machine-learning model to differentiate germline from somatic SVs based on these characteristics. SVMs have been recently demonstrated to be suitable for classifying cancer subtypes[14–17]. We therefore chose an analogous approach to establish an SVM-based SV germline filter that is applicable to the SvABA SV caller[11]. SVMs also have intrinsic advantages. They avoid overfitting by not seeking to approach zero error when learning on training data. Additionally, an SVM approach is not disadvantaged by having a large feature set as the model will determine which features contribute to the support vector[18]. Therefore, an SVM approach allows for liberal addition of features while preserving the utility of a generalizable classifier.

Here we present an easily implementable tool that uses a support vector machine architecture to reliably distinguish germline SVs from somatic SVs.

Results

To explore the differences between somatic and germline SVs we used the TCGA dataset of paired tumor-normal WGS encompassing 974 tumors from 24 cancer types. Across the 974 tumors, germline SVs outnumbered somatic SVs by 12:1 (**Fig. 1A**, median of 2,009 germline SVs and 57 somatic SVs per tumor). The number of germline events in a given individual was constant irrespective of their age (**Fig. 1B**).

Extending SNV classification approaches to SVs

We first attempted to filter germline SVs using the gnomAD v2.1 population dataset[10] analogous to the commonly used filtering approach for removing germline SNVs[19,20]. The distance from gnomAD-listed germline SVs was significantly different for germline and somatic SVs (KS-test, p-value < 2.2e-16, **Fig 1C**). However only a small fraction of germline SVs matched within a 3bp of a gnomAD SV (~910K SVs out of 2.1M germline SVs). About a third of germline SVs lay more than 1000bp from a gnomAD SV. To map SVs to the gnomAD reference SV we determined the average base pair distance of each pair of breakpoints from its closest SV in gnomAD. This “fuzzy matching” approach had an overall AUC of 0.87. The optimal cutoff-point on the ROC curve of ~1400bp average distance from the nearest gnomAD SV resulted in a high degree of sensitivity in correctly classifying somatic SVs, (TPR ≃ 95% of true somatic events). However, this cutoff resulted in a large fraction of germline contamination (67%) in the set of SVs called somatic (**Fig 1D)**. Given that a single SV can have profound driver impacts on the cancer genome[8], these data show that a simple filter based on the closest germline SV in gnomAD cannot sufficiently differentiate between germline and somatic SVs in tumor-only SV calls.

Our group and others have previously established that several characteristics of a DNA sequence increase the probability of a somatic SV occurring at a particular locus[1,2]. Additionally, SVMs using a combination of functional genomic and context-specific features to each SV have been developed for structural variant pathogenicity classification[21]. Therefore, to enable germline vs. somatic structural variant classification without a paired normal sample, we set out to build an SVM classifier considering a variety of features describing the genomic context of each structural variant.

Differences in Genomic Features Between Germline and Somatic SVs

We first assessed features of each SV (**Supplementary Table 1**) that we hypothesized could differentiate germline from somatic SVs. These included SV type (Deletion, Duplication,

Inversion, Translocation), distance between the SV breakpoints (span), distance to the closest reference gnomAD germline SV, distance to the nearest SV, number of SVs within a 5Mbp window of each SV breakpoint, microhomology length at the SV breakpoints, number of bases inserted between breakpoints and the replication timing of the SV breakpoints. Each SV was also annotated to indicate if it impacted a protein-coding region or caused a copy-number change. Finally, we annotated the distance to the closest repeat elements: short interspersed nuclear elements (SINE) and long interspersed nuclear elements (LINE), and the *TP53* mutation status of the sample.

Several SV features showed significant differences between somatic and germline SVs. Somatic SV had significantly larger spans (on average by a factor of 4x) than those of germline SVs (KS test, p-value < 2.2e-16; **Fig 2A**). Peaks in the germline SV span distribution corresponded to the typical spans of SINE and LINE repeat elements[22],[23]. These span distributions varied by SV type, with the greatest differences between germline deletions, which tend to be short, and somatic deletions, whose span distribution is more uniform (**Fig 2B-D**). As span increased, the enrichment for somatic SVs became more pronounced. Somatic SVs were about twice as likely to have spans greater than 1,000 bp compared to germline SVs (54% of all somatic SVs vs. 31% of all germline SVs), a number that swells to 40 times more likely at 1 Mb (17% of all somatic SVs vs. 0.4% of all germline SVs). Therefore, larger SV span was associated with somatic SVs.

The distance to the nearest LINE or SINE repeat elements also differed significantly between germline and somatic SVs (KS-test, p-value < 2.2e-16, **Fig. 2E**); both types of repeat elements were similarly distant from SVs (**Fig. 2F-G**). However, of all germline SVs, 57% were within 100bp of a repeat element in contrast to 33% of somatic SVs.

As a proxy for inferring complexity, we looked at the distance of any given SV to its nearest neighbor as well as the number of SVs within a 5Mbp window of a breakpoint. Somatic SVs tend to cluster closer than germline SVs suggesting that they may be part of more complex rearrangements (KS-test, p-value < 2.2e-16, **Fig. 2H-I**). For example, 46% of somatic SVs were within 10kb of their nearest SV within a sample compared to 19% of germline SVs. Furthermore, only 11% of germline SVs had at least 20 SVs in a 5Mbp neighborhood, whereas 64% of somatic SVs crossed this threshold.

SV types also differed between germline and somatic SVs. Somatic SVs were six times more likely to be translocations (Chi-Squared test, p-value < 2.2e-16). Deletion events made up the majority (about 72%) of germline events (Chi-Squared test, p-value < 2.2e-16). Inversion events, on the other hand, comprised the majority (50%) of somatic events (Chi-Squared test, p-value < 2.2e-16, **Fig 2J)**.

SVs that disrupted coding sequences or spanned a whole gene were more frequently somatic (Fisher's Exact Test, both p-values < 2.2e-16, **Fig 2K**). We observed 23.2% of

somatic SVs spanning over the whole length of at least one gene as opposed to 0.008% of germline events, a nearly 2900x enrichment. Somatic SVs were also nearly nine times more likely to disrupt a coding sequence. Despite this enrichment, only 34.5% of somatic SVs directly affected the exome.

Single features are insufficient to distinguish germline and somatic SVs

To test the predictive value of individual SV features, we constructed single-feature logistic regression models with a training set of 200,000 SVs. We tested each model on 100,000 SVs and obtained an average single-feature AUC of 0.663 (SD 0.006, **Fig 2L)**. The distance from either breakpoint of an SV to its nearest gnomAD SV performed best with similar AUCs of ~0.939 (**Supplementary Table 2**). Even with this performance, neither feature achieved a positive predictive value of greater than 0.23 resulting in a large amount of germline SVs in the set of predicted-somatic events. Replication timing, insertion GC%, insertion length, and distance to SINE elements were poor individual differentiators between germline and somatic SVs. We conclude no single SV characteristic is sufficiently predictive to perform the germline or somatic classification.

Associations between SV characteristics differ between germline and somatic SVs

We evaluated whether SV features (**Supplementary Table 1**) were associated with each other, performing the analysis separately among somatic and germline SVs. For SV features measurable in continuous values, we investigated the frequency of association with a Spearman correlation of $|\rho|>0.5$. We found that most continuous variables were not correlated within either germline or somatic SVs. For example, longer distances to repeat elements did not predict longer homology or insertion lengths (**Fig. 3A**). We also tested if continuous SV features differed across different levels of the categorical variables within germline and somatic SVs. This showed significant differences in all pairings of continuous and categorical variables within SV classes (Mann-Whitney U q-value < 0.05, **Fig. 3B**). For example, SVs that impacted a gene had longer spans across both germline and somatic SVs. Finally, we performed chi-square tests to evaluate if different categories of SVs were more or less likely to impact genes or exons in germline and somatic genomes (q-value < 0.05). This showed that both exon and gene impact show significant associations with all SV types (**Fig. 3C**). These data support that most continuous SV features are independent of each other, but categorical features were associated with both continuous and other categorical features, among both germline and somatic SVs.

Almost all SV feature associations (180/184) differed between germline and somatic SVs indicating that combinations of feature pairs have varying strengths and directionalities of associations. For example, long somatic SVs tend to be farther from a reference germline event, whereas no such correlation is observed among germline SVs. Germline SVs display a weak correlation between the length of their insertion sequence and their distance to SINE elements, which is mostly absent in somatic SVs. Most notably, homology length exhibited a slight anti-correlation to many contextual features among somatic SVs including: distance to LINE and SINE elements, distance to gnomAD, and SPAN. Homology length is indicative of the type of DNA damage repair process utilized[2]. Rearrangements caused during tumorigenesis frequently have deficiencies in DNA damage repair pathways decreasing the amount of homology used during the repair process[2]. We observed similar differences between germline and somatic associations of continuous/categorical SV features and categorical/categorical SV features. For instance, we found that somatic inversions were more likely to be longer events than their germline counterparts. Additionally, somatic deletions more frequently impacted genes than germline deletions (**Fig. 3D-F**). These data suggest that combinations of SV features differ between somatic and germline SV and may provide more information to distinguish them.

Classifier Development and Performance

We therefore developed an SVM classifier to distinguish between germline and somatic SVs based on the combination of SV features. We performed a grid-search approach to hyperparameter tuning, resulting in the selection of an RBF kernel, a cost parameter of 10, and a gamma parameter of 0.1. The classifier was trained on 555,849 SVs, two-thirds of all SVs available in TCGA passing our filter requirements. Once trained, the classifier was tested on 277,925 SVs, one-third of the SVs in the sample set.

In addition to a binary classification, our SVM generates a probability of classification for each variant. These probability values range from 0 to 1. Values close to 1 represent a higher likelihood of a variant being somatic and vice versa. These probabilities allow for flexibility in choosing a classification cutoff by letting the user prioritize certain performance metrics, including AUC, PPV, etc. For instance, having a cutoff farther from 0 would reduce the number of germline SVs falsely called somatic by the classifier (increasing the PPV), while increasing the number of true somatic SVs falsely called germline (decreasing the TPR). We aimed to reduce germline contamination in our set of SVs called somatic for future analysis of somatic variants, so we decided on a cutoff of 0.232 for the rest of this paper. This cutoff maximized the sum of the PPV and TPR in our test set.

Overall we achieved an AUC of 0.984, with a sensitivity (TPR) of 0.913 and specificity (1 - FPR) of 0.963 (**Fig. 4A**). The positive predictive value of our classifier was 0.808.

SVM performance varied across tumor types due largely to differences in the prevalence of somatic SVs in each tumor type. A higher fraction of SVs that were called somatic actually were somatic (PPV) in tumors with many somatic SVs. In our test set, sarcomas had the highest somatic SV burdens (~155 SVs per tumor) and the highest proportion of correctly classified somatic events (PPV = 0.937), whereas acute myeloid leukemia with the fewest SVs per tumor (~1.2 SVs) had the lowest proportion of correctly classified somatic events (0.13) (**Fig. 4B**).

Performance in an Independent Dataset

In order to test how well our tool classifies SVs in tumors from other datasets, we gathered a test set consisting of 7,623 SVs from six pHGG patients. The tumor samples used in this dataset were collected with blood normals, which provided a truth set of somatic and germline SVs. The features observed in the pHGG dataset had similar distributions to those of the tumors in the TCGA dataset (**Fig. S1**). Our SVM achieved a sensitivity (TPR) of 0.945 and specificity (1-FPR) of 0.887. Of the SVs that were called somatic in this set, 83% were truly somatic (PPV: 0.828). We conclude that the performance of our SVM is robust across different datasets.

Discussion

Here we present a tool to distinguish between somatic and germline SVs for situations where a paired normal is not present. We found that SVs of germline vs. somatic origin have distinguishing characteristics; although each is insufficient to distinguish between germline and somatic SVs on its own, their combination (including their interaction terms) effectively distinguishes germline vs. somatic SVs. By having the ability to extract somatic variants, our tool has applications in clinical genotyping, research in tumors where a normal sample was never collected, and research in cancer cell lines.

The classifier's success in the separation of somatic and germline rearrangements suggests that there are inherent differences between the two classes. One clear example of this was seen with the SV types. The majority of germline rearrangements were short nonclustered deletions. These potentially originate in small errors during normal DNA replication as previously described[24]. The somatic SV types were much more varied, likely as a consequence of a higher rate of genome instability and tolerance to gene losses in cancer[25,26]. The homogeneity of the germline SV types may derive largely because anything other than short deletions has a detrimental effect during development. For example, small

repeats resulting from duplication have been found to decrease cell fitness[27]. This also explains the infrequency of germline SVs in the coding genome.

In constructing our classifier, we chose a probability cutoff that optimized for the sum of TPR and PPV, in order to minimize germline contamination from rearrangements called somatic. However, the probability cutoff chosen for our classifier can be modified for other use cases, including those that demand correct germline calls or a high overall accuracy, in which case a cutoff of the maximum sum of the TNR and NPV or the maximum AUC in the same TCGA validation set can be used.

One major limitation of our classifier was the ancestry representation of our training data. Although our tool had good performance on the TCGA test set, the TCGA cohort we used was composed of at least 77% patients of European ancestry (750/974). On the other hand, there were only 79 patients of known African descent (8%) in our dataset. Considering that there is more genetic diversity amongst those of African descent than amongst those from the rest of the world, it is likely that the training data has insufficient diversity for adequate performance on rearrangements from African patients[28].

Another potential limitation of our classifier was the dependence on SvABA outputs. Although the features input into the classifier are not unique to SvABA, our classifier was trained on the rearrangements, microhomology, and non-mapped insertion sequences called by SvABA. If rearrangements from another SV caller were input into the SVM, the performance may suffer due to potential differences in convention or SV calling method. We suggest retraining the classifier based on the outputs of these callers instead.

Coverage is also an important consideration for our classifier inputs. The TCGA cohort used to train our data had an average genomic coverage of 50-60x[13]. By nature of most SV calling methods, an increase in sequencing coverage means that more SVs are called for a given sample. This also means that a difference in coverage is likely to impact multiple features, particularly those relating to SV complexity—number of SVs in a sample and distance to the nearest SV. As such, the performance of our classifier may suffer for lower sequencing coverages.

Although our classifier performed well overall, it tended to misclassify germline-like somatic events and vice versa. For example, the SVM incorrectly labeled many somatic deletions shorter than 100kb as a germline rearrangement. This suggests that it could be relying on the SPAN and SV type features to make the classification. One potential improvement could be to examine the commonly misclassified types of SVs and see if there are any other differences between the somatic and germline classes in these events that we can add as features.

There could also be inherent characteristics of SVM classification that may lead to incorrect classification for certain SVs. SVMs rely on the construction of a geometric hyperplane which attempts to not only separate points belonging to different classes but also increase the distance between the points and the hyperplane[29]. As such, SVMs tend to have decreased performance on overlapping or noisy data[30]. A consensus ensemble classifier might overcome such classifier-specific issues, with this SVM being one of the models used in the ensemble classification[31,32].

In conclusion, our classifier provides the first publicly available computational method that efficiently classifies SVs into somatic and germline variants. We expect it to be practical for analyses of WGS samples without paired normals including unpaired tumors collected prior to the genomics era such as many of the most highly-used cell lines.

Methods and Materials

Data Acquisition

Whole genome sequencing tumor samples with matched normals were obtained from the TCGA patient cohort (**Supplementary Table 3**). We used SvABA to call SVs on all samples as previously described, which were used for our training and test sets[11]. The pHGG test set consisted of six samples from our recent study[33]. We also ran SvABA on the pHGG samples, which output the SVs used in our external test set.

Quality Control and Filtering

We obtained the filtered “sv.vcf” SvABA outputs for germline and somatic breakpoints from all samples of our cohort. We then grouped pairs of breakpoints according to their MATEID. Any breakpoint without an identified MATEID pair was removed. We will subsequently refer to these MATEID pairs as rearrangements (or SVs). Once combined, we selected SVs that had the max MAPQ value (60) for both breakpoints, were not detected solely by discordant reads, had a span of 1000bp or greater or were translocations, and had at least two SV-supporting split reads.

Once the TCGA rearrangements were filtered, two-thirds (555,849 SVs) were randomly selected to be part of the training set, and the remaining one-third (277,925 SVs) were labeled as the test set.

Feature Annotation

Once the filtering step was complete, we annotated each SV with the following features: distance to the closest gnomAD reference SV, DNA replication timing of each breakpoint, GC content and length of any novel sequence insertion, GC content and length of homology associated with a breakpoint, TP53 gene mutation status, number of other SVs within a 5Mbp window, total number of SVs in the sample, distance to a long interspersed nuclear element (LINE), distance to a short interspersed nuclear element (SINE), distance to the nearest SV, span of the SV, type of the SV (categorized as deletion, duplication, inversion, or translocation), impact on a gene, and impact on an exon.

The distance to the closest gnomAD reference SV for a given rearrangement was encoded as two features, each representing the distance of a breakpoint to its "corresponding" breakpoint of the nearest reference gnomAD SV (i.e. comparing the left breakpoint of the rearrangement to the left breakpoint of the nearest gnomAD reference SV and vice versa). The nearest reference SV was determined by iterating through the reference SVs in the gnomAD database, calculating the distance from each breakpoint, and selecting the reference SV that had the lowest sum of distances.

The DNA replication timing was likewise encoded as two features, corresponding to the replication timing of each breakpoint. This was determined by looking up each breakpoint location in the DNA replication timing table for the hg19 reference genome (from http://mskilab.com/fishHook/hg19/RT_NHEK_Keratinocytes_Int92817591_hg19.rds).

The GC content and length of insertion and microhomology sequences were features derived from the SvABA output. When assembling the contigs, SvABA frequently finds short gaps or overlaps between regions that map to the reference genome. These gaps or overlaps were output as "INSERTION" or "HOMSEQ" respectively. The gaps (i.e. the short region of the sample genome that did not map to the reference genome) and overlaps were analyzed for GC content and length and each of these were added as features.

The TP53 mutation status was determined using the consensus SNV, MNV, and indel calls[13] from TCGA samples available on the ICGC Data Portal (PCAWG 2020). These consensus calls were filtered for the TP53 gene and functional mutations (i.e. variants classified as "3'UTR", "5'UTR", "lincRNA", "Intron", "Silent", "IGR", "5'Flank", "RNA" were removed). XX of our samples were not analyzed in the PCAWG consensus calls. For these, we assigned the TP53 status as indeterminate. Each sample was then assigned a value of 1, 0, or -1, which corresponded to TP53 mutant, indeterminate, or TP53 wild-type.

The number of SVs in a 5Mbp window and the total number of SVs in the sample were calculated by counting the number of rearrangements that were within 5Mbp of the SV and within the specific sample respectively. The distances to the nearest SINE and LINE events were calculated by taking the distance from a given rearrangement to the closest SINE and LINE event (as defined in https://www.repeatmasker.org/species/hg.html). The distance to the

nearest SV was calculated by iterating through other rearrangements in the same sample, determining the distance between the given rearrangement and the other SV, and selecting the lowest value.

The span of the SV was taken from the SvABA output. The SV type feature was separated into four binary features, each corresponding to insertion, duplication, inversion, or translocation events based on the position and read orientation of the supporting reads

The impact on a gene or exon region were binary factors representing whether a given rearrangement overlapped a gene or exon region, which was downloaded from Gencode and Ensembl genome browser BioMart, respectively. If the SV breakpoint was located within one of these regions, it was labeled with a 1. If no overlap was found, the feature was labeled with a 0.

Moreover, we log-transformed any feature that related to genomic distances or counts of SVs to reduce noise. These features included the SV span, homology length, insertion length, distance to the closest gnomAD reference SV, distance to the nearest SINE and LINE element, the number of SVs in a given sample, the distance to the nearest SV, and the number of SVs within 5Mbp.

Once all the features were added, each feature was then scaled. The scaling was done by creating a scaling matrix containing values representing the mean and standard deviation of each feature (**Supplementary Table 4**). These values were calculated by taking 10 random samples of 50,000 SVs from the training dataset and evaluating the mean and standard deviation of the features for all of these rearrangements. Each input rearrangement feature—including those in both training and test sets—was scaled by subtracting the mean and dividing by the standard deviation in this matrix.

SVM

We used an SVM with an RBF kernel from the e1071 package in R and set cost and gamma parameters as 10 and 0.1 respectively. These hyperparameters were determined by a grid-search, of all combinations of the following values: cost of 0.0001, 0.001, 0.01, 0.05, 0.1, 1, gamma of 1, 10, 100, 500, 1000 for both RBF and linear kernels. For each combination of hyperparameters, we performed a 5-fold cross-validation, splitting the training set into five groups. For each of the five iterations of the cross-validation, one group was selected as a validation set, and the other four groups made up the sub-training set. During each iteration, the model was trained on the sub-training set and evaluated on the validation set. The resulting AUC and PPV for all iterations were averaged for each combination of hyperparameters, and the model that performed well with respect to both metrics was chosen.

Although SVMs typically do not have an inherent probability metric, there is a probability feature built into e1071 implementation of the SVM—based on Platt Scaling. We used this to not only generate the ROC curve for the AUC above but also to select a probability cutoff that optimizes for our specific needs, instead of just accuracy. Because we aimed to reduce germline contamination for analysis of somatic variants, we decided on a cutoff that resulted in a high PPV.

In order to achieve this, we treated the probability cutoff as another hyperparameter and performed another 5-fold cross-validation for each probability value between 0 and 1 in 0.001 increments. This time, we optimized for the maximum TPR + PPV value, since we found that PPV itself was monotonically increasing with respect to the probability cutoff. We did not change the cost, gamma, and kernel type parameters in this probability tuning step, and were kept as 10, 0.1, and RBF kernel respectively.

Signature Analysis

The signature analysis was performed primarily using the signature.tools.lib package published by the Nik-Zainal group (Degasperi et al. 2020).

Code availability

All annotation functions, reference files, and trained SVM will be available at https://github.com/shu-z/germline_svm

References

1. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
2. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
3. Degasperi, A. *et al.* A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. *Nat Cancer* **1**, 249–263 (2020).
4. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **7**, 818–831 (2017).
5. Levatić, J., Salvadores, M., Fuster-Tormo, F. & Supek, F. Mutational signatures are markers of drug sensitivity of cancer cells. *bioRxiv* 2021.05.19.444811 (2021) doi:10.1101/2021.05.19.444811.
6. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
7. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
8. Bandopadhayay, P. *et al.* MYB-QKI rearrangements in angiocentric glioma drive tumorigenicity through a tripartite mechanism. *Nat. Genet.* **48**, 273–282 (2016).
9. Dubois, F. *et al.* Structural variants shape driver combinations and outcomes in pediatric high-grade glioma. (2021) doi:10.21203/rs.3.rs-389596/v1.
10. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
11. Wala, J. A. *et al.* SvABA: genome-wide detection of structural variants and indels by local

assembly. *Genome Res.* **28**, 581–591 (2018).

12. Gong, T., Hayes, V. M. & Chan, E. K. F. Detection of somatic structural variants from short-read next-generation sequencing data. *Brief. Bioinform.* **22**, (2021).
13. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
14. Vural, S., Wang, X. & Guda, C. Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. *BMC Syst. Biol.* **10 Suppl 3**, 62 (2016).
15. Rapaport, F., Barillot, E. & Vert, J.-P. Classification of arrayCGH data using fused SVM. *Bioinformatics* **24**, i375–82 (2008).
16. Segal, N. H. *et al.* Classification and subtype prediction of adult soft tissue sarcoma by functional genomics. *Am. J. Pathol.* **163**, 691–700 (2003).
17. Kim, S., Jhong, J.-H., Lee, J. & Koo, J.-Y. Meta-analytic support vector machine for integrating multiple omics data. *BioData Min.* **10**, 2 (2017).
18. Awad, M. & Khanna, R. Support Vector Machines for Classification. in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers* (eds. Awad, M. & Khanna, R.) 39–66 (Apress, 2015).
19. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* 861054 (2019) doi:10.1101/861054.
20. Adelson, R. P. *et al.* Empirical design of a variant quality control pipeline for whole genome sequencing data using replicate discordance. *Sci. Rep.* **9**, 16156 (2019).
21. Kumar, S., Harmanci, A., Vytheeswaran, J. & Gerstein, M. B. SVFX: a machine learning framework to quantify the pathogenicity of structural variants. *Genome Biol.* **21**, 274 (2020).
22. Vassetzky, N. S. & Kramerov, D. A. SINEBase: a database and tool for SINE analysis.

Nucleic Acids Res. **41**, D83–9 (2013).

23. Alisch, R. S., Garcia-Perez, J. L., Muotri, A. R., Gage, F. H. & Moran, J. V. Unconventional translation of mammalian LINE-1 retrotransposons. *Genes Dev.* **20**, 210–224 (2006).
24. Bzymek, M., Saveson, C. J., Feschenko, V. V. & Lovett, S. T. Slipped misalignment mechanisms of deletion formation: in vivo susceptibility to nucleases. *J. Bacteriol.* **181**, 477–482 (1999).
25. Duijf, P. H. G., Schultz, N. & Benezra, R. Cancer cells preferentially lose small chromosomes. *Int. J. Cancer* **132**, 2316–2326 (2013).
26. Abbas, T., Keaton, M. A. & Dutta, A. Genomic instability in cancer. *Cold Spring Harb. Perspect. Biol.* **5**, a012914 (2013).
27. Adler, M., Anjum, M., Berg, O. G., Andersson, D. I. & Sandegren, L. High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. *Mol. Biol. Evol.* **31**, 1526–1535 (2014).
28. Tucci, S. & Akey, J. M. The long walk to African genomics. *Genome biology* vol. 20 130 (2019).
29. Boser, B. E., Guyon, I. M. & Vapnik, V. N. A training algorithm for optimal margin classifiers. in *Proceedings of the fifth annual workshop on Computational learning theory* 144–152 (Association for Computing Machinery, 1992).
30. Li, H.-X., Yang, J.-L., Zhang, G. & Fan, B. Probabilistic support vector machines for classification of noise affected data. *Inf. Sci.* **221**, 60–71 (2013).
31. Mohammed, A. & Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences* **35**, 757–774 (2023).
32. Huang, M.-W., Chen, C.-W., Lin, W.-C., Ke, S.-W. & Tsai, C.-F. SVM and SVM

Ensembles in Breast Cancer Prediction. *PLoS One* **12**, e0161501 (2017).

33. Dubois, F. P. B. *et al.* Structural variants shape driver combinations and outcomes in pediatric high-grade glioma. *Nat Cancer* (2022) doi:10.1038/s43018-022-00403-z.

Figures and Figure Legends

Figure 1

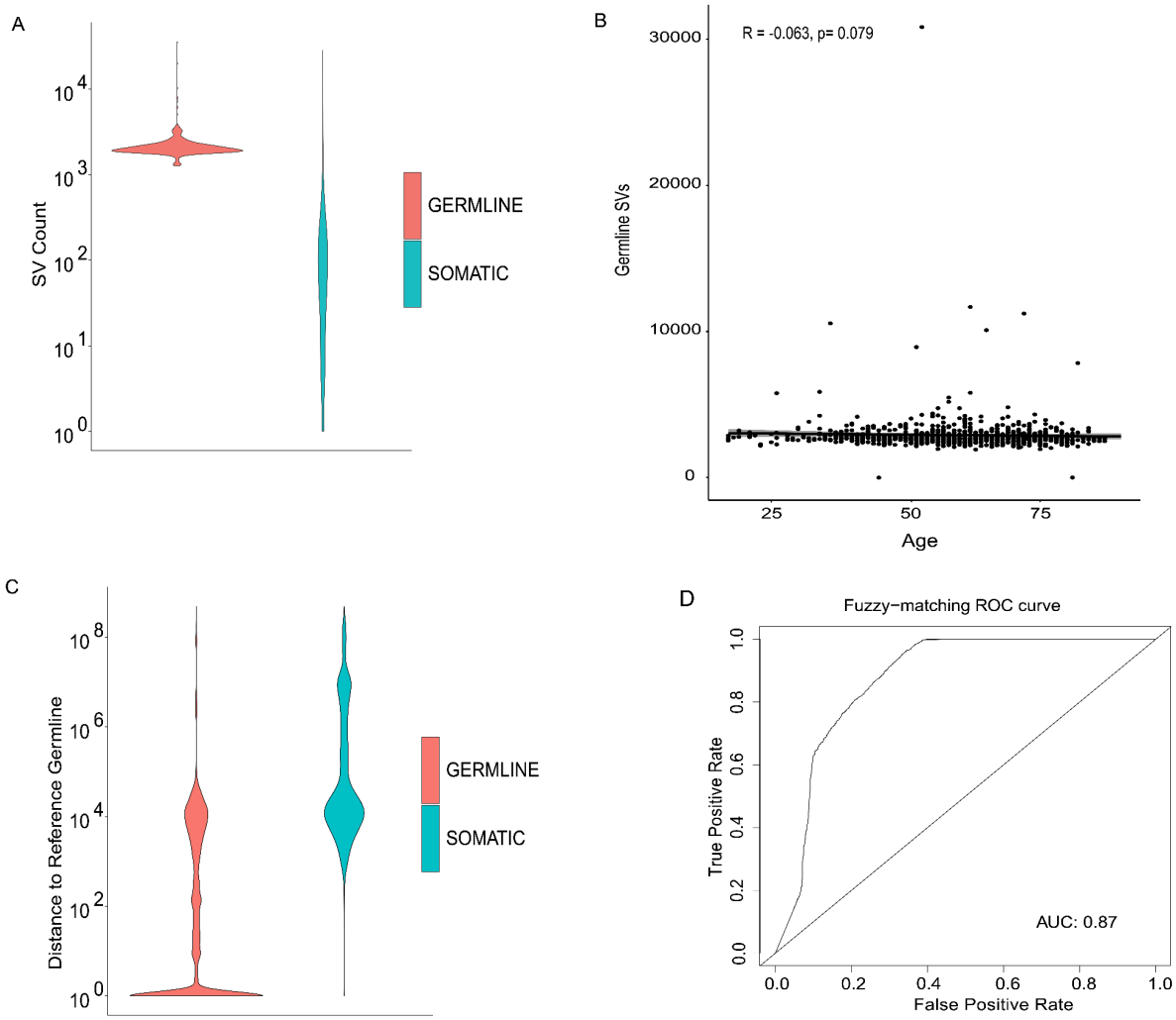


Figure 1. Distance to nearest reference germline SV does not perform well as a somatic/germline classifier. **A**, Distribution of germline and somatic SV frequencies in PCAWG tumors. **B**, Pearson correlation analysis between germline SV frequency and patients' age. **C**, Distribution of total distance of germline and somatic SVs to the closest SV in gnomAD (reference) SV database. **D**, ROC for classifying SVs based on distance to gnomAD reference (fuzzy-matching).

Figure 2

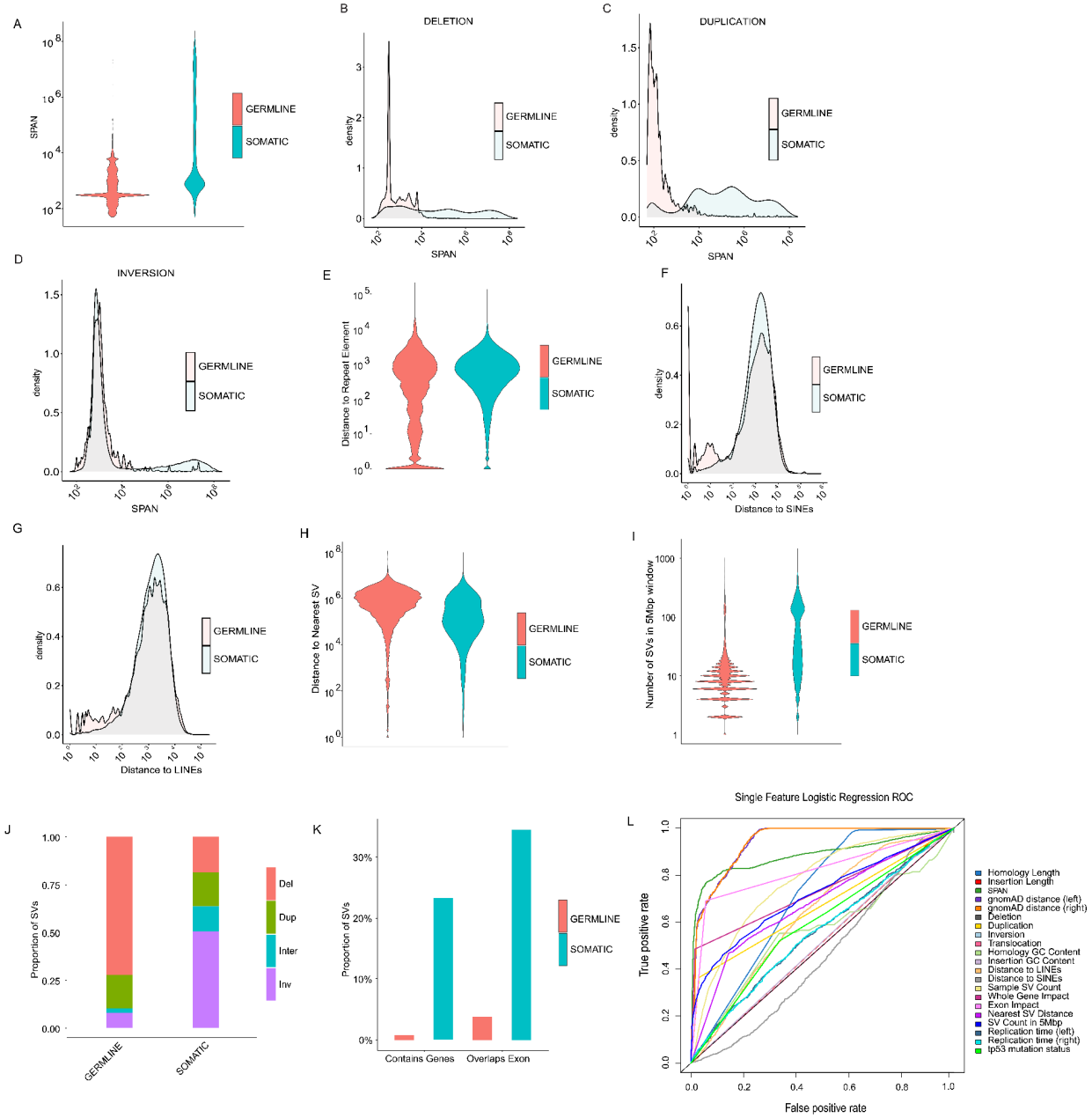


Figure 2. Genomic features differ significantly between SV classes in TCGA and are individually insufficient to distinguish germline and somatic SVs. **A**, Overall span distribution of germline and somatic SVs. **B-D**, Span distribution of deletions (**B**), duplications (**C**), and Inversions (**D**). **E**, Distance of germline and somatic SVs to nearest repeat element (LINE or SINE). **F**, Distribution of distance to SINE elements. **G**, Distribution of distance to LINE elements. **H**, Distance to nearest SV within a sample. **I**, Number of SVs within a 5Mbp window of each SV breakpoint. **J**, Proportion of each SV type - Del(deletion), Dup(duplication), Inter(interchromosomal/translocation), Inv (inversion) present in germline and somatic SVs. **K**, Proportion of germline and somatic SVs that impact a gene or overlap an exon. **L**, ROCs for single-feature logistic regression.

Figure 3

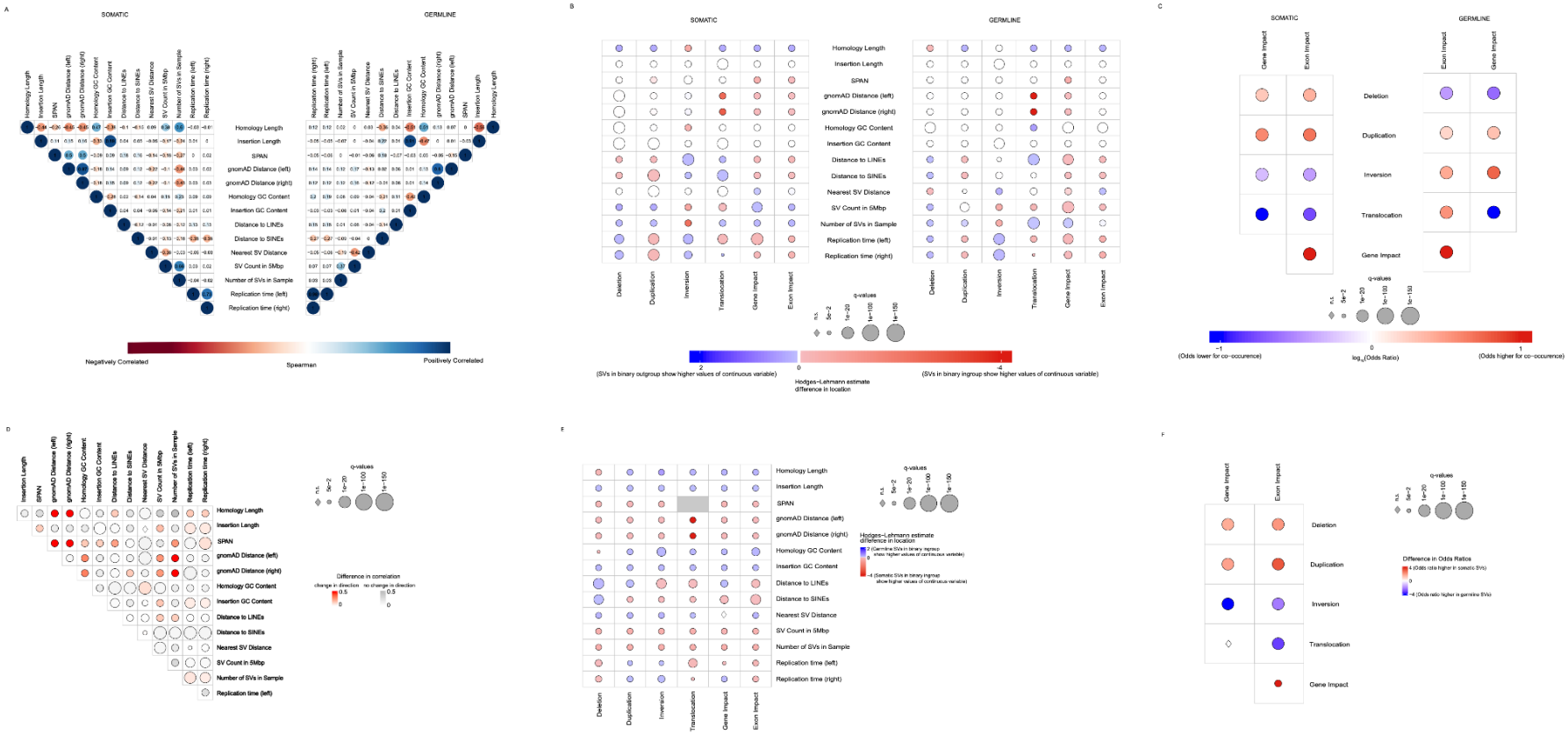


Figure 3. Comparisons between SV features show associations within somatic and germline SVs that differ across SV classes. **A**, Correlations between continuous features within somatic and germline SVs. **B**, Mann-Whitney U test estimates for associations between continuous and categorical SV features within germline and somatic SVs. **C**, Odds ratios comparing categorical variables of gene and exon impact to SV type in germline and somatic SVs. **D**, Signed difference in spearman correlation values between continuous features across germline and somatic SVs. **E**, Association between continuous and categorical variables across germline and somatic SV. **F**, Difference in odds ratios of categorical variable comparisons across germline and somatic SVs.

Figure 4

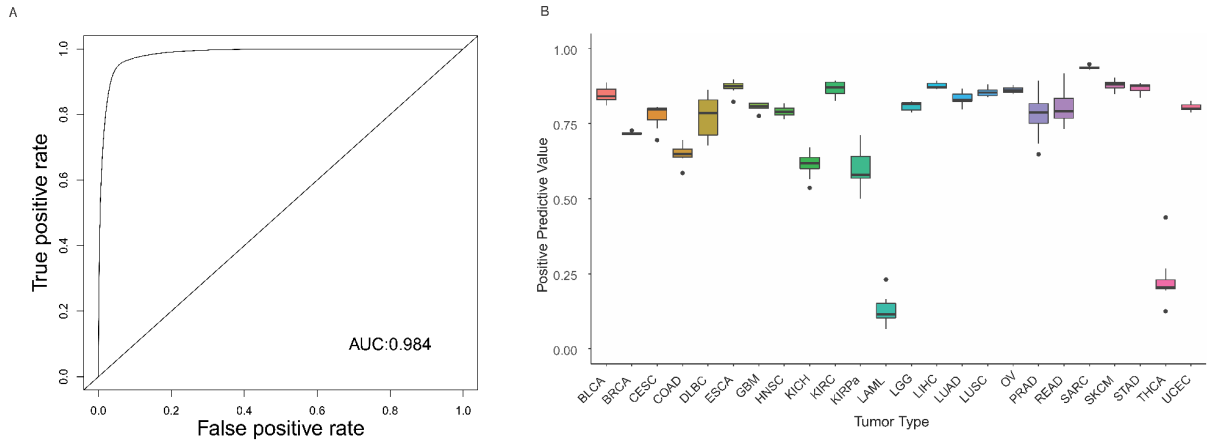
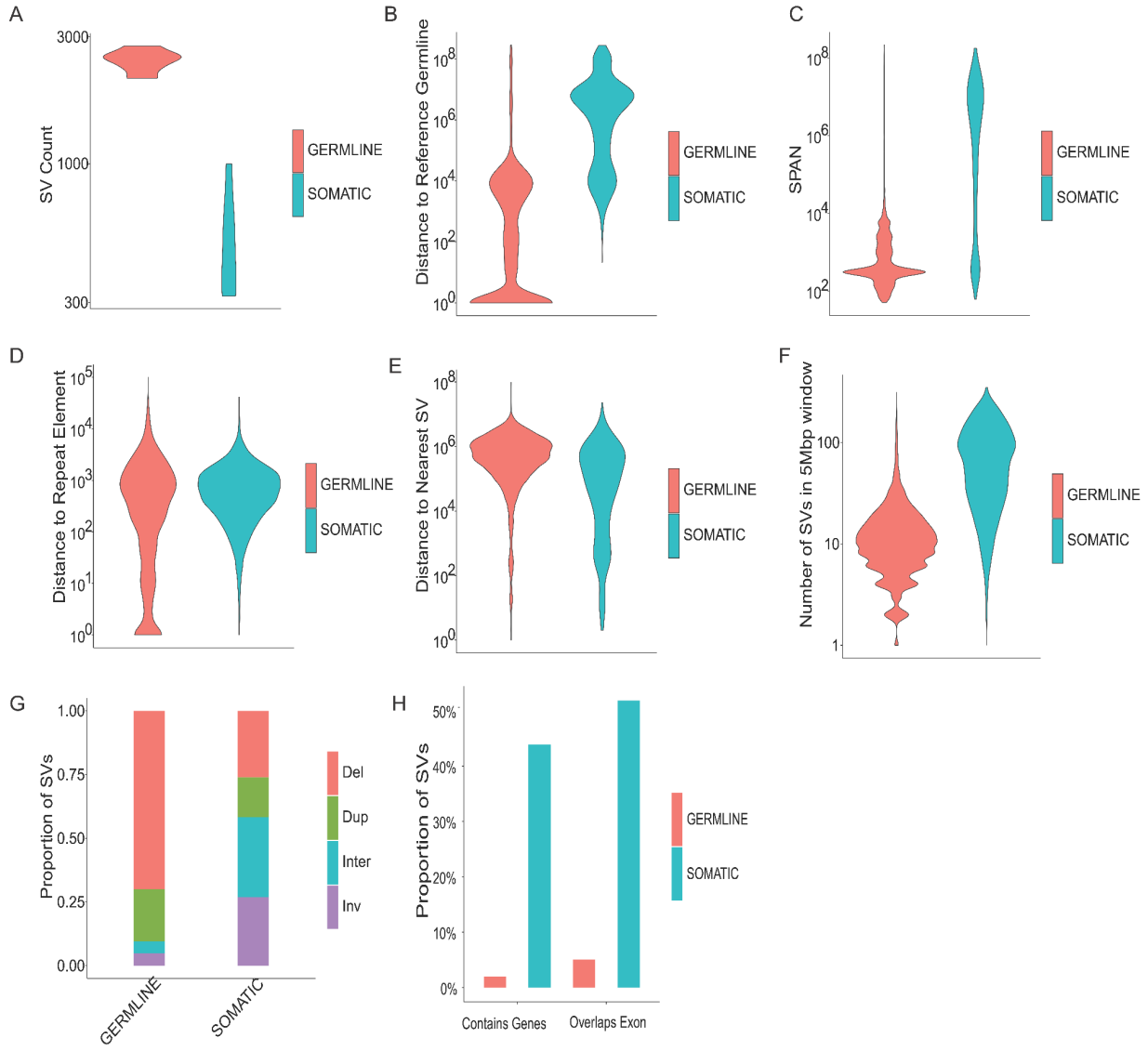


Figure 4. SVM performance in TCGA. **A**, SVM ROC on TCGA test set of 277,925 SVs. **B**, Positive Predictive Value (PPV) of SVM by tumor type showing variance in performance across tissue types.

Supplementary Figure 1



Supplementary Figure 1. SV Feature Distribution in the pHGG test set is similar to TCGA. **A**, Germline and somatic SV frequencies in test samples. **B**, Total distance of somatic and germline SVs to closest SV in gnomAD reference. **C**, Span distribution of somatic and germline SVs. **D**, Distance of somatic and germline SVs to nearest repeat element (LINE or SINE). **E**, Distance of germline and somatic SVs to the nearest SV within a sample. **F**, Number of SVs within a 5Mbp window of each SV breakpoint. G, Proportion of each SV type - Del(deletion), Dup(duplication), Inter(interchromosomal/translocation), Inv (inversion) present in germline and somatic SVs. **H**, Proportion of germline and somatic SVs that impact a gene or overlap an exon.