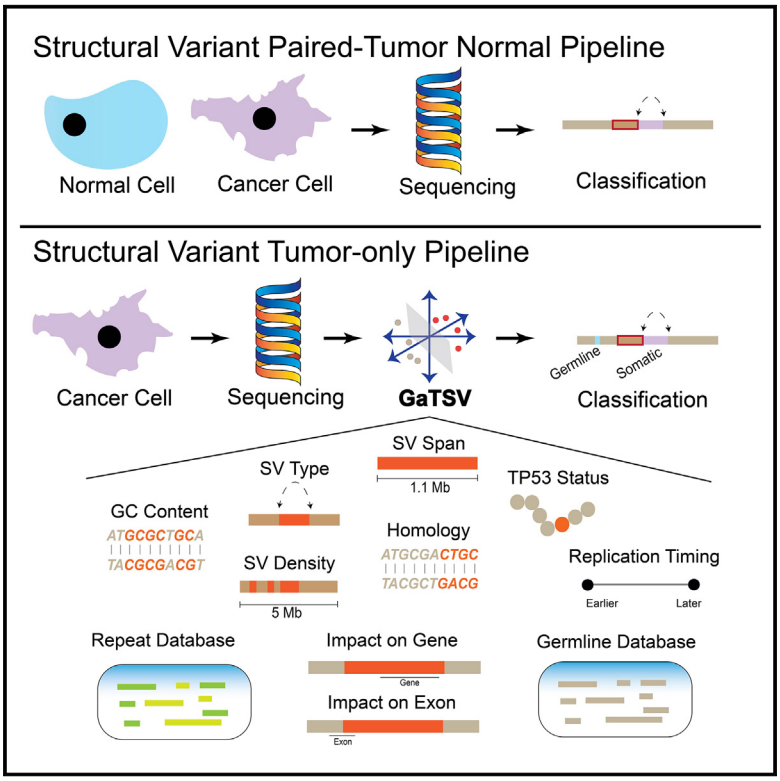
Cell Reports
Methods

Article

A sequence context-based approach for classifying tumor structural variants without paired normal samples

Graphical abstract



Highlights

- Genomic sequence contexts differ between germline and somatic SVs
- GaTSV can separate somatic from germline SVs without a paired normal sample
- GaTSV-classified somatic SVs recapitulate known somatic SV signatures

Authors

Wolu Chukwu, Siyun Lee, Alexander Crane, ..., Rameen Beroukhim, Frank Dubois, Simona Dalin

Correspondence

frank.dubois@charite.de (F.D.), sdalin@broadinstitute.org (S.D.)

In brief

Genomic structural variants (SVs) are key to understanding cancer and can inform treatment response. Distinguishing cancer from normal SVs has the limitation of requiring paired normal samples. Chukwu et al. introduce GaTSV, a classifier that accurately identifies cancer-specific SVs without matched normal tissue, based on genomic sequence context differences.

Chukwu et al., 2025, Cell Reports Methods *5*, 100991
March 24, 2025
https://doi.org/10.1016/j.crmeth.2025.100991

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

A sequence context-based approach for classifying tumor structural variants without paired normal samples

Wolu Chukwu,[1,2,10] Siyun Lee,[1,2,10] Alexander Crane,[1,2,10] Shu Zhang,[1,2,10] Sophie Webster,[1,2] Oumayma Dakhama,[1] Ipsa Mittra,[1] Carlos Rauert,[3] Marcin Imielinski,[4,5,6,7,8,9] Rameen Beroukhim,[1,2] Frank Dubois,[1,2,3,11,12,*] and Simona Dalin[1,2,11,*]

[1]Cancer Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA
[3]Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Institute of Pathology, Berlin, Germany
[4]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA
[5]New York Genome Center, New York, NY, USA
[6]Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA
[7]Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA
[8]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA
[9]Department of Pathology and Perlmutter Cancer Center, NYU Grossman School of Medicine, New York, NY, USA
[10]These authors contributed equally
[11]Senior author
[12]Lead contact
*Correspondence: frank.dubois@charite.de (F.D.), sdalin@broadinstitute.org (S.D.)
https://doi.org/10.1016/j.crmeth.2025.100991

MOTIVATION While structural variants (SVs) harbor rich insight into the pathogenicity of diseases such as cancer, the models used to study these SVs often lack matched normal samples. This makes it difficult to separate germline from cancer-related variants. Our ability to confidently associate disease phenotypes with their underlying genomic events depends on differentiating germline from cancer SVs. Given that germline and somatic SVs are generated by different processes, we hypothesized that they will be distinguishable based on their genomic contexts. Here, we introduce a machine-learning-based classifier to distinguish germline from cancer-related SVs in tumors without matched normal samples.

SUMMARY

Although several recent studies have characterized structural variants (SVs) in germline and cancer genomes independently, the genomic contexts of these SVs have not been comprehensively compared. We examined similarities and differences between 2 million germline and 115 thousand tumor SVs from a cohort of 963 patients from The Cancer Genome Atlas. We found significant differences in features related to their genomic sequences and localization that suggest differences between SV-generating processes and selective pressures. For example, our results show that features linked to transposon-mediated processes are associated with germline SVs, while somatic SVs more frequently show features characteristic of chromoanagenesis. These genomic differences enabled us to develop a classifier—the Germline and Tumor Structural Variant or "the great GaTSV" —that accurately distinguishes between germline and cancer SVs in tumor samples that lack a matched normal sample.

INTRODUCTION

Structural variants (SVs) are rearrangements of genomic material that result from incorrect double-strand break (DSB) repair. Large-scale whole-genome sequencing (WGS) efforts have revealed a prominent role of recurrent somatic SVs as cancer drivers[1] and as biomarkers of disruption of the DNA damage response and other processes.[2,3] Similar studies on normal tissue have shown that human genetic diversity results in large part from germline SVs.[4] The SVs in these different contexts may result from different biological constraints.

Please cite this article in press as: Chukwu et al., A sequence context-based approach for classifying tumor structural variants without paired normal samples, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100991

All SVs including somatic variants as well as benign and deleterious germline variants arise from DSBs that are repaired to new loci in the genome. The resulting rearrangements can delete, duplicate, invert, or translocate segments of DNA. The responsible mechanisms of DSB repair include non-homologous end joining (NHEJ), which pastes broken ends of DNA together irrespective of adjacent sequence homology; microhomology-mediated end joining, which utilizes 3–10 bp of breakpoint-adjacent microhomology to form repair intermediates; and homologous recombination, which uses more bases of homology.

Germline and somatic SVs are likely to arise from different DSB repair mechanisms. Previous studies report that germline SVs primarily result from non-allelic homologous recombination, where substantial amounts of sequence homology at distinct loci are employed in double-stranded break repair, often resulting in the deletion of the intervening sequence.[5,6] Conversely, somatic SVs, which present with more varied spans and clustering of breakpoints, indicate a tendency toward NHEJ and replication-based mechanisms of repair such as microhomology-mediated break-induced replication, which are more error prone.[5–7] However, a direct and comprehensive comparison between the features of germline and somatic SVs is likely to provide a more detailed view of the differences in their generation, with implications for the activity of different DNA damage and repair processes in human vs. somatic cell evolution.

One benefit of recognizing the different features of germline and somatic SVs is that it provides an opportunity to distinguish germline and somatic SVs when only data from somatic tissue have been collected. Currently, confidently designating an SV as "somatic" requires comparing sequencing data from highly clonal somatic tissue such as cancers or single cells with multiclonal "normal" tissues that represent the germline. However, there are many situations in which normal tissue is unavailable, including clinical settings[8] and the study of long-term cell line models.[9] There is one existing method to deplete germline SVs from samples lacking matched normals. This method, which we term "fuzzy matching," involves matching SV breakpoints to germline reference databases, allowing for a certain amount of "slop," or distance from the observed SV's breakpoints to the closest germline reference SV's breakpoints.[10–12] However, this method has limitations, including its inability to detect rare germline variants, the lack of a standardized definition of slop parameter, and variability in breakpoint calls across different SV callers, which arise from methodological differences and conventions for assigning ambiguous breakpoint locations. Considering these issues, as well as the fact that about half of every individual's germline SVs are extremely rare or completely unique,[13] the fuzzy matching method is not able to accurately remove germline SVs from samples without matched normals. The ability to distinguish germline from somatic SVs in the absence of normal tissue is therefore valuable both clinically and in cancer research.

Here, we comprehensively evaluate similarities and differences between germline and somatic SVs, confirming previous findings with similar comparisons[10] and extending upon these results. We find that they are strikingly different in molecular context, enabling us to develop a machine-learning-based classifier (the Germline and Tumor Structural Variant classifier, also known as the great GaTSV) that can distinguish germline from somatic SVs called by SvABA in the absence of a matched normal sample with extremely high accuracy.

RESULTS

To explore the differences between somatic and germline SVs, we used a TCGA (The Cancer Genome Atlas) dataset of paired tumor-normal WGS encompassing 963 tumors from 24 cancer types. We used the SvABA SV caller to ascertain the breakpoints and types of the SVs in this dataset.[14] Across the 963 tumors, germline SVs outnumbered somatic SVs 17:1 (Figure 1A, median of 2,007 germline SVs and 53 somatic SVs per tumor). The number of germline events in a given individual was constant irrespective of age, whereas the number of somatic SVs in a sample showed a slight positive correlation with age (Figures 1B and 1C) as shown previously.[10]

Germline and somatic SVs have different SV features

To evaluate the differing impacts of SV generation and selection processes between germline and somatic contexts, we compared features of germline vs. somatic SVs (also see Table S1 and Figures S1A–S1D). The most striking difference was SV span, the distance between intrachromosomal breakpoints. Somatic SVs had spans 60 times larger than those of germline SVs (KS test, $p < 2.2e\text{-}16$; Figure 1D) and were about twice as likely to have spans greater than 1,000 bp (67% of all somatic SVs vs. 31% of all germline SVs), a number that swells to 60 times more likely at 1 Mb (27% of all somatic SVs vs. 0.4% of all germline SVs). These results align with the intuition that SVs of larger spans are likely to result in changes to the genome that are not tolerated during normal development. Peaks in the germline SV span distribution corresponded to the typical spans of SINE and LINE transposable elements.[15,16] Germline and somatic 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 (Figure S1E). These results are consistent with previous studies of germline and somatic SVs.[2,10]

The second most striking difference between germline and somatic SVs was the much higher levels of breakpoint homology attributed to germline SVs (KS test, $p < 2.2e\text{-}16$; Figure 1E), suggesting a transposon-mediated origin. Closer examination of the distribution of homology lengths revealed a peak between 13 and 17 bp in germline SVs but not somatic SVs. This peak was specifically present in germline deletions of ~300 bp span (Figures S1F and S1G), corresponding to the spans of Alu elements—a type of SINE element and the most abundant transposons in the human genome.[17] Previous studies have shown that Alu elements comprise a significant proportion of transposable element-mediated rearrangements in the genome and use ~15 bp of homology.[18] Germline SVs were closer to SINE and LINE elements than somatic SVs, regardless of SV type (KS test, $p < 2.2e\text{-}16$; Figures 1F–1H, S1H, S2A, and S2B), and—of all repeat elements—these showed the greatest difference in range of distances between germline SVs and somatic SVs (Figures S2C–S2Q). Interestingly, somatic SVs were closer to all classes of RNA pseudogenes compared with germline SVs.

Please cite this article in press as: Chukwu et al., A sequence context-based approach for classifying tumor structural variants without paired normal samples, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100991

CellPress
OPEN ACCESS

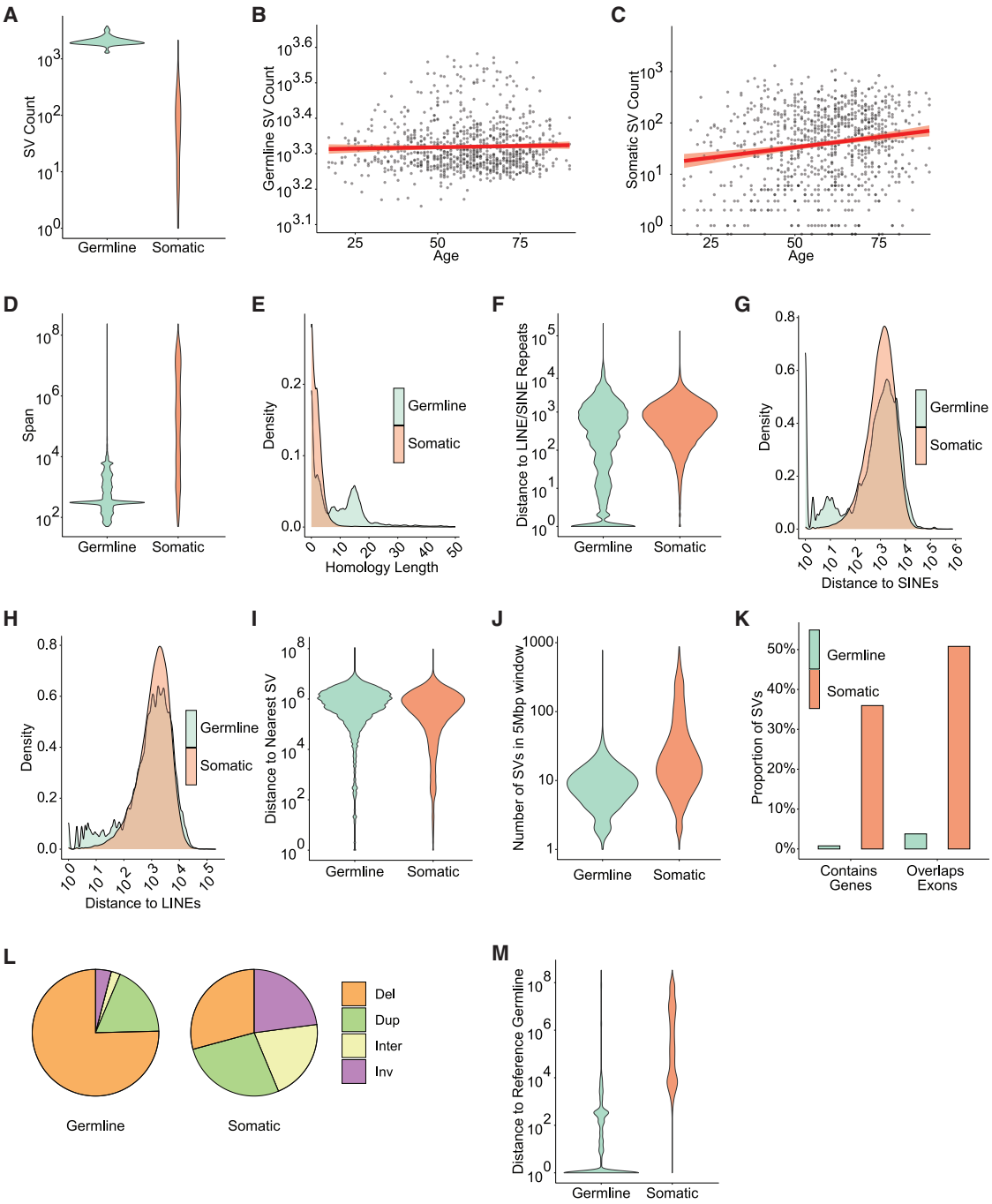


Figure 1. SV features differ significantly between germline and somatic genomes

(A) Distribution of germline and somatic SV frequencies in TCGA genomes.

(B and C) Pearson correlation analysis between germline SV frequency and patients' age in germline (B) and somatic (C) genomes.

(D) Overall span distribution of germline and somatic SVs.

(E) Homology length distribution of germline and somatic SVs.

(F) Distance of germline and somatic SVs to nearest repeat element (LINE or SINE).

(G and H) Distribution of distance to SINE (G) and LINE (H) elements.

(I) Distance to nearest SV within a sample.

(J) Number of SVs within a 5 Mbp window of each SV breakpoint within a sample.

(K) Proportion of germline and somatic SVs that impact a gene or overlap an exon.

(L) Proportion of each SV type—Del, deletion; Dup, duplication. Inter(interchromosomal/translocation), Inv (inversion) present in germline and somatic SVs.

(M) Distribution of total distance of germline and somatic SVs to the closest SV in gnomAD (reference) SV database.

Please cite this article in press as: Chukwu et al., A sequence context-based approach for classifying tumor structural variants without paired normal samples, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100991

Overall, these data suggest that germline SV generation is linked to the activity of transposable elements.

In contrast, somatic SVs are more likely to be formed by chromoanagenesis[19] and affect gene structure. Somatic SVs were more likely to be found in proximity to each other (KS test, $p < 2.2e-16$, Figures 1I and 1J) and were more likely to disrupt coding sequences or span entire genes (Fisher's exact test, both $p < 2.2e-16$; Figure 1K). Strikingly, 51% of somatic SVs directly affected the exome, in contrast to only 3.8% of germline SVs. Deletion events made up about 75% of germline events but only 29% of somatic events (chi-squared test, $p < 2.2e-16$; Figure 1L), whereas somatic SVs were nine times more likely to be translocations (chi-squared test, $p < 2.2e-16$). Finally, germline SVs were much closer to reference SVs in the gnomAD database[13] (median of 12 bp) than somatic SVs (median of 51k bp) ($p < 2.2e-16$; Figure 1M). Overall, our analysis of single features of SVs showed pronounced differences in the characteristics of germline compared with somatic SVs.

If there are specific variant-generating or selection processes that result in germline and somatic SVs, we expect the features of SVs (Table S1) that are linked to such processes to vary together. Indeed, characteristic combinations of SV features between germline and somatic SVs reflected the pronounced differences in the relationships between SV features (Figures 2A–2C). Somatic SVs with shorter spans tended to have longer homology (Figure 2A), while no such associations were observed for germline SVs. Germline SVs had longer homology if they were closer to known SINE elements or were deletions (Figures 2A and 2B), while no such association existed for somatic SVs (Figures 2A and 2D). Longer germline SVs also showed higher homology GC content while we observed the opposite in somatic SVs (Figure 2D). Interestingly, germline and somatic translocations were closer to LINE and SINE repeat elements than other SV types (Figures 2B and S1H). Altogether, these data further suggest that transposon-mediated processes dominate germline SV generation while only a subset of somatic SVs originate from this pathway. Recently published long-read WGS also confirmed the presence of highly prevalent germline transposons.[20] In contrast, somatic SVs showed an anti-correlation between homology length and GC content with the number of SVs within 5 Mbp as well as SV span with its distance to the nearest SV (Figures 2A and 2D). These data indicate a subset of (longer) somatic SVs generated in clusters by NHEJ in a more complex process like chromothripsis.

The differences in feature associations between germline and somatic SVs also hinted at the strong differences in the selection pressures they face. Somatic deletions were significantly more likely to span gene/exons than germline deletions. (Figures 2C and 2F). Also, while duplications were also more likely to span genes than other SV types in both germline and somatic SVs (chi-squared test, $q < 2.2e-16$; Figure 2C), this association was significantly stronger in somatic SVs (Figure 2F). Germline inversions, which could be more neutral in their effects on gene expression, were more likely to affect genes/exons than other germline SV types (Figure 2C) and also more likely to span genes/exons than somatic inversions (Figure 2F). Our analysis of the association between replication timing and SV type also found results consistent with strong selection pressures conserving the coding genome in the germline. We observed a positive correlation between homology length and replication timing among germline SVs that was absent or insignificant among somatic SVs (Figure 2A). This shows an enrichment for high-fidelity repair in gene-dense regions in the germline context that is not observed in cancer-associated SVs. Also, while there was an overall bias for inversions and deletions toward later replicating genomic loci and duplications for earlier replicating loci (Figure 2B), striking differences arise when comparing across germline and somatic SVs. Notably, somatic duplications and inversions showed significantly higher replication timing values, corresponding to earlier replicating regions, than germline duplications and inversions (Figure 2E). Early-replicating regions tend to be gene-dense and highly expressed.[21] Together, these data indicate strong selection pressures to conserve the full coding genome in the germline attenuated in cancer cells.

Extending SNV classification approaches to SVs proves insufficient

Next, we considered how we might use this information to distinguish between germline and somatic SVs when paired germline DNA sequencing data are unavailable. As a control, we first attempted to filter germline SVs using the gnomAD v.4.0 population dataset,[13] analogous to the commonly used filtering approach for removing germline SNVs.[22,23] However, only a fraction of germline SVs matched within 3 bp of a gnomAD SV (~1.0 million SVs out of 2.0 million germline SVs with only ~31 thousand SVs matching exactly to a reference gnomAD SV). The paucity of exact matches may be due to differences in SV callers between gnomAD and our dataset as well as the fact that most germline SVs are not recurrent.[13] In fact, when we assessed the likelihood of finding a particular SV from a patient within gnomAD using a 200 bp slop, we found that only 50% of SVs are found in the gnomAD reference database. An earlier study by Chen et al. reported that 85% of germline SVs in their cohort were recurrent using the same slop. However, their analysis considered the total set of germline SVs across the cohort including multiple entries of the same variant if it appeared in multiple individuals. Since we were interested in determining the fraction of germline SVs in any single individual that would be present in gnomAD, we only considered a deduplicated set of SVs, where each SV had a single entry even if it was detected in many people.

Due to the fact that about half of germline SVs are recurrent, the distance from gnomAD-listed germline SVs was significantly different for germline and somatic SVs ($p < 2.2e-16$; Figure 1M), and about a third of germline SVs lay more than 1,000 bp from a gnomAD SV. Therefore, to match TCGA SVs to the gnomAD reference SVs, we determined the average base pair distance of each pair of breakpoints from each TCGA SV to its closest SV in gnomAD. This fuzzy matching approach had an overall AUC of 0.90. The optimal cutoff point on the ROC curve of ~1,400 bp average distance from the nearest gnomAD SV resulted in a high degree of sensitivity in correctly classifying somatic SVs (TPR = 96% of true somatic events). However, this cutoff still resulted in a large fraction of germline contamination (60%) in the set of SVs called somatic (positive predictive value

Please cite this article in press as: Chukwu et al., A sequence context-based approach for classifying tumor structural variants without paired normal samples, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100991

CellPress
OPEN ACCESS

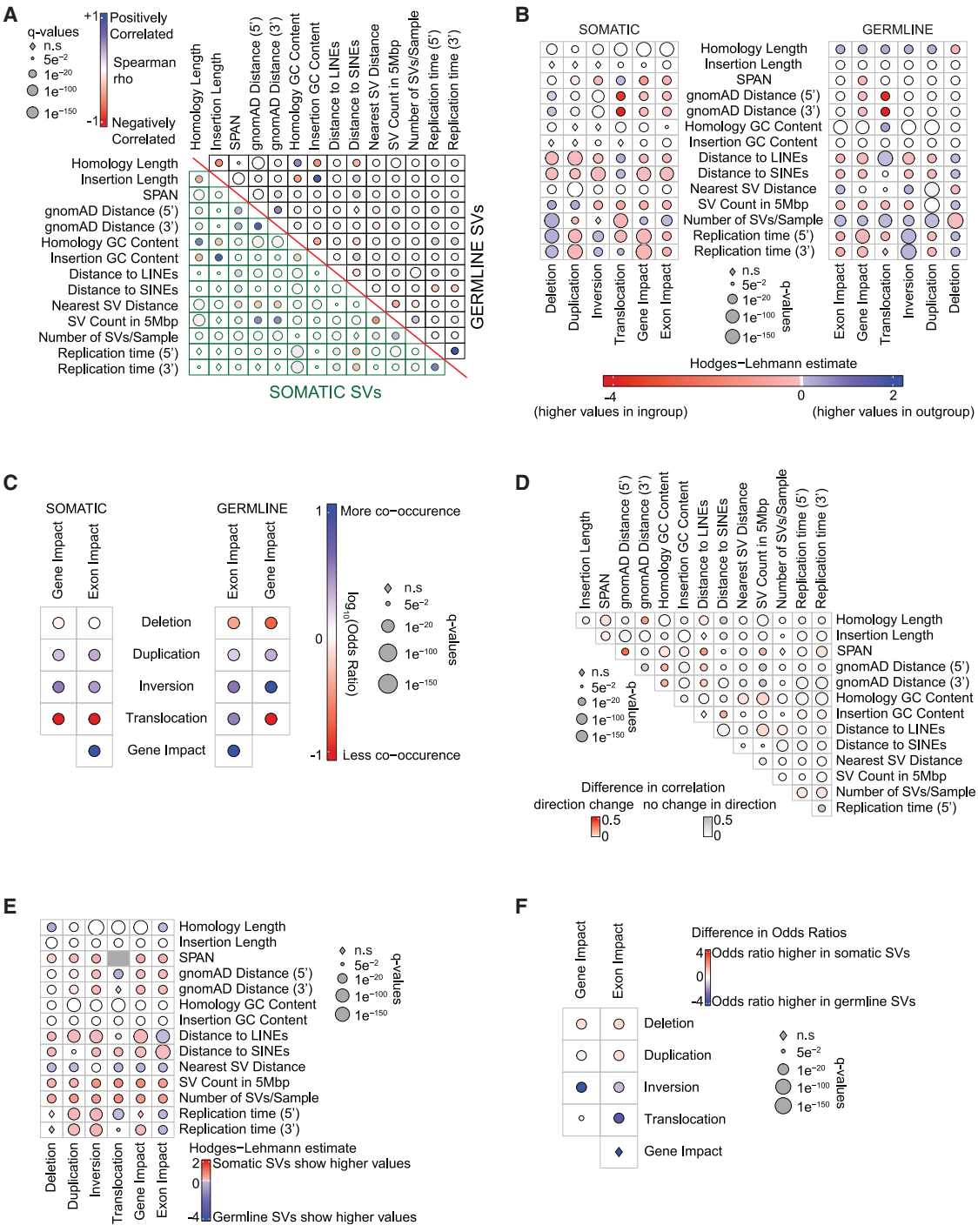


Figure 2. SV feature associations differ between somatic and germline SVs

(A) Spearman correlations between continuous features within somatic and germline SVs (positive correlations in blue, anticorrelation in red).

(B) Differences in the values of continuous SV features between categorical SV features within germline and somatic SVs.

(C) Odds ratios of the associations between the categorical variables gene or exon impact and SV type in germline and somatic SVs.

(D) Difference in Spearman correlation values between continuous features across germline and somatic SVs.

(E) Differences in the values of continuous SV features between germline and somatic SVs within categorical SV features.

(F) Difference in odds ratios of categorical variable comparisons across germline and somatic SVs. For all panels, the circle size represents the significance of the statistic represented.

Please cite this article in press as: Chukwu et al., A sequence context-based approach for classifying tumor structural variants without paired normal samples, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100991

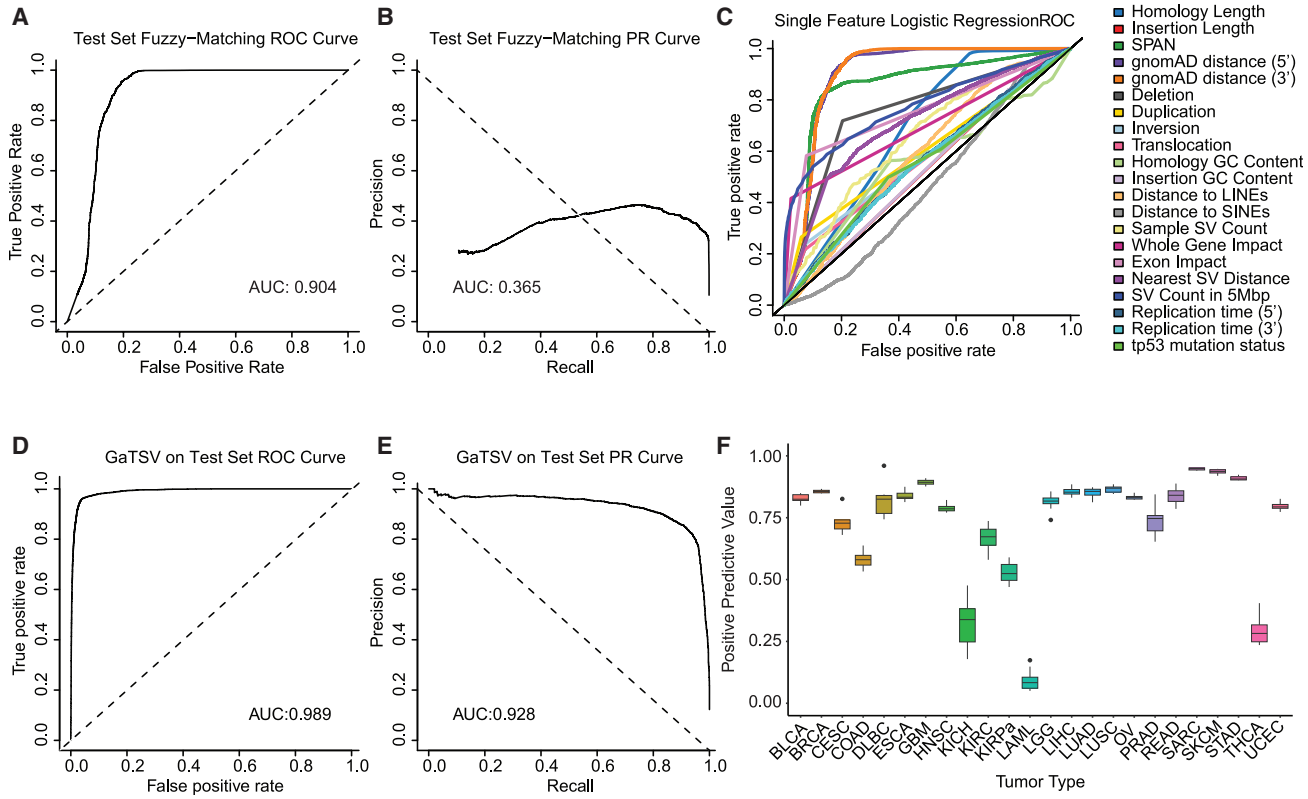


Figure 3. Classifier performances on the TCGA test set

(A) ROC for classifying SVs based on a cutoff distance to gnomAD reference (fuzzy matching).

(B) PR curve for classifying SVs using fuzzy matching.

(C) ROC for separate single-feature logistic regression classifiers.

(D) ROC for the GaTSV classifier.

(E) PR curve for the GaTSV classifier.

(F) Positive predictive value (PPV) of the GaTSV classifier by tumor type showing variance in performance across tissue types. Boxes represent the interquartile range (IQR) of PPVs; whiskers extending from the boxes represent the range 1.5× beyond the IQR.

[PPV]) (Figures 3A and 3B). 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.

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 0.656 (SD 0.127, Figure 3C). The distance from either breakpoint of an SV to its nearest gnomAD SV performed best, with similar AUCs of ∼0.90 (Table S2). Even with this performance, neither feature achieved a positive predictive value of greater than 0.50 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 concluded no single SV characteristic was sufficiently predictive to perform the germline or somatic classification.

GaTSV distinguishes SVs based on somatic and germline identity

As individual features cannot sufficiently distinguish germline and somatic SVs, we used the combination of SV features to develop a support vector machine (SVM)-based classifier—the great GaTSV classifier—to distinguish between germline and somatic SVs. We trained GaTSV on 509,433 SVs from 634 samples, two-thirds of all samples available in our TCGA cohort. Once trained, we tested the classifier on 262,118 SVs from 329 samples, one-third of the samples in our TCGA cohort.

In addition to a binary classification, GaTSV generates a probability of classification for each variant, with values closer to 1 representing a higher likelihood of a variant being somatic. These probabilities allow for the selection of a classification cutoff to 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). To balance these considerations, GaTSV uses a cutoff of 0.268, which maximizes the sum of the PPV and TPR in our validation set.

Please cite this article in press as: Chukwu et al., A sequence context-based approach for classifying tumor structural variants without paired normal samples, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100991

CellPress
OPEN ACCESS

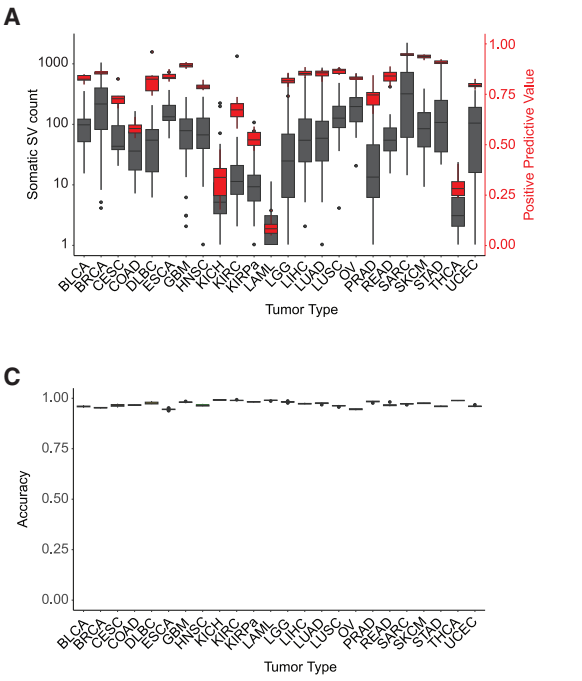


Figure 4. GaTSV classifier performance is related to somatic SV count

(A) Somatic SV count and PPV for each tumor type in the TCGA dataset. Boxes indicate the IQR and whiskers the range of 1.5× IQR.

(B) PPV vs. somatic SV count for all tumors in the TCGA dataset.

(C) Accuracy of the GaTSV classifier across each tumor type. Boxes indicate the IQR and whiskers the range 1.5× IQR.

Overall, GaTSV achieved an AUC of 0.989, with a sensitivity (TPR) of 0.915, specificity (1 – FPR) of 0.977 (Figures 3D and 3E), and PPV of 0.849.

GaTSV's performance varied across tumor types, largely due to differences in the prevalence of somatic SVs in each tumor type. The fraction of GaTSV-classified somatic SVs that were truly somatic was higher in tumors with many somatic SVs. In other words, the PPV was higher in tumors with more somatic SVs. In our test set, sarcomas had the highest somatic SV burdens (~391 SVs per tumor) and the highest proportion of called-somatic SVs that were truly somatic (PPV = 0.947). In contrast, acute myeloid leukemia, with the fewest SVs per tumor (~2 SVs), had the lowest proportion of called-somatic SVs that were truly somatic (0.09) (Figure 3F). This finding was corroborated at the sample level. Samples with a lower SV burden in the test set tended to have lower PPV than samples with a higher SV burden (Figures 4A and 4B). Specifically, 86% of samples with fewer than 10 somatic SVs had PPV below 0.5. In contrast, only 0.5% of samples with 10 or more SVs had a PPV below 0.5 (Fisher's exact test, $p < 2.2e\text{-}16$). These data show that GaTSV performs well overall but is subject to falsely calling true germline SVs somatic SVs in samples with very low somatic SV counts. GaTSV's performance also differs between germline SVs that recur in the TCGA test set and those that do not, with GaTSV classifying recurring SVs at a higher accuracy. The same is true for SVs that are found within the gnomAD database. Although GaTSV is less accurate for non-recurrent SVs (SVs more than 10 bp from other TCGA SVs) and SVs not found in gnomAD, GaTSV performs much better on these SVs than the fuzzy matching approach, likely due to the other SV features it analyzes (Table S3).

GaTSV achieved an accuracy of 97% on average across all tumor types (Figure 4C). To uncover potential weaknesses of our classifier, we examined the features of the misclassified SVs. The feature distribution of germline SVs mislabeled as somatic largely mimics that of true somatic SVs in the test set (Figure S3). This suggests that GaTSV poorly differentiated these somatic-like germline events, leading to germline contamination in the somatic calls. For example, 98% of misclassified germline SVs were more than 1 kb away from a gnomAD reference (Figure S3B). Nonetheless, GaTSV correctly classified 91% of germline SVs with a distance of 1 kb or more from a gnomAD SV; therefore, GaTSV performs well on most SVs.

GaTSV performs reliably in an independent dataset

To test how well GaTSV classifies SVs in tumors from other datasets, we gathered a test set consisting of 7,623 SVs from 6 pediatric high-grade glioma (pHGG) patients with comparable somatic SV burden to TCGA samples. 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 (Figure S4). GaTSV achieved a sensitivity (TPR) of 0.975 and specificity (1 – FPR) of 0.892. Of the SVs that were called somatic in this set, 84% were truly somatic (PPV = 0.839). We conclude that GaTSV performs robustly across different datasets.

GaTSV shows reduced performance in underrepresented ancestries

We hypothesized that imbalance in representation of individuals from certain ancestries in our training set would impair GaTSV's performance on individuals of those ancestries—specifically, over 77% of individuals in the TCGA dataset are of European descent. Indeed, GaTSV performed better on SVs from European individuals than on SVs from East Asian and African individuals on all metrics considered, including the AUC of the ROC and PR curves as well as the PPV (Figures 5A–5F). Among the ancestries considered, GaTSV performed the worst on African SVs across all metrics. African genomes are considerably more diverse than any other ancestry group, which means that increased representation in the training set is necessary compared with other ancestries.

Please cite this article in press as: Chukwu et al., A sequence context-based approach for classifying tumor structural variants without paired normal samples, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100991

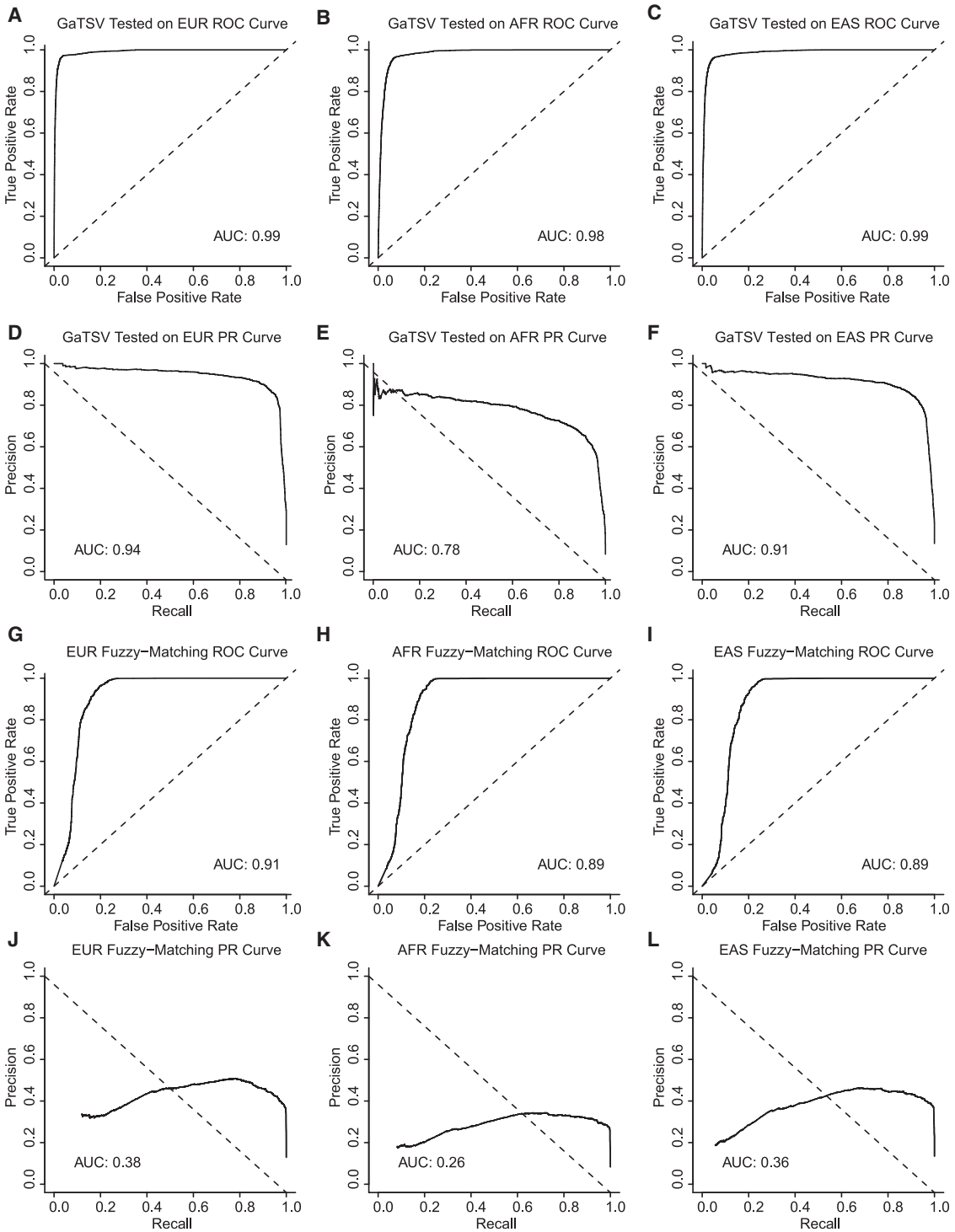


Figure 5. Classifier performances on ancestry-specific subsets of the TCGA test set

(A–C) ROC for the GaTSV classifier on individuals descending from European (EUR) (A), African (AFR) (B), and East Asian (EAS) (C) ancestry.

(D–F) PR curve for the GaTSV classifier on individuals descending from European (D), African (E), and East Asian (F) ancestry.

(G–I) ROC for the fuzzy-matching method to the gnomAD database on individuals descending from European (G), African (H), and East Asian (I) ancestry.

(J–L) PR curve for the fuzzy-matching method on individuals descending from European (J), African (K), and East Asian (L) ancestry.

Please cite this article in press as: Chukwu et al., A sequence context-based approach for classifying tumor structural variants without paired normal samples, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100991

CellPress
OPEN ACCESS

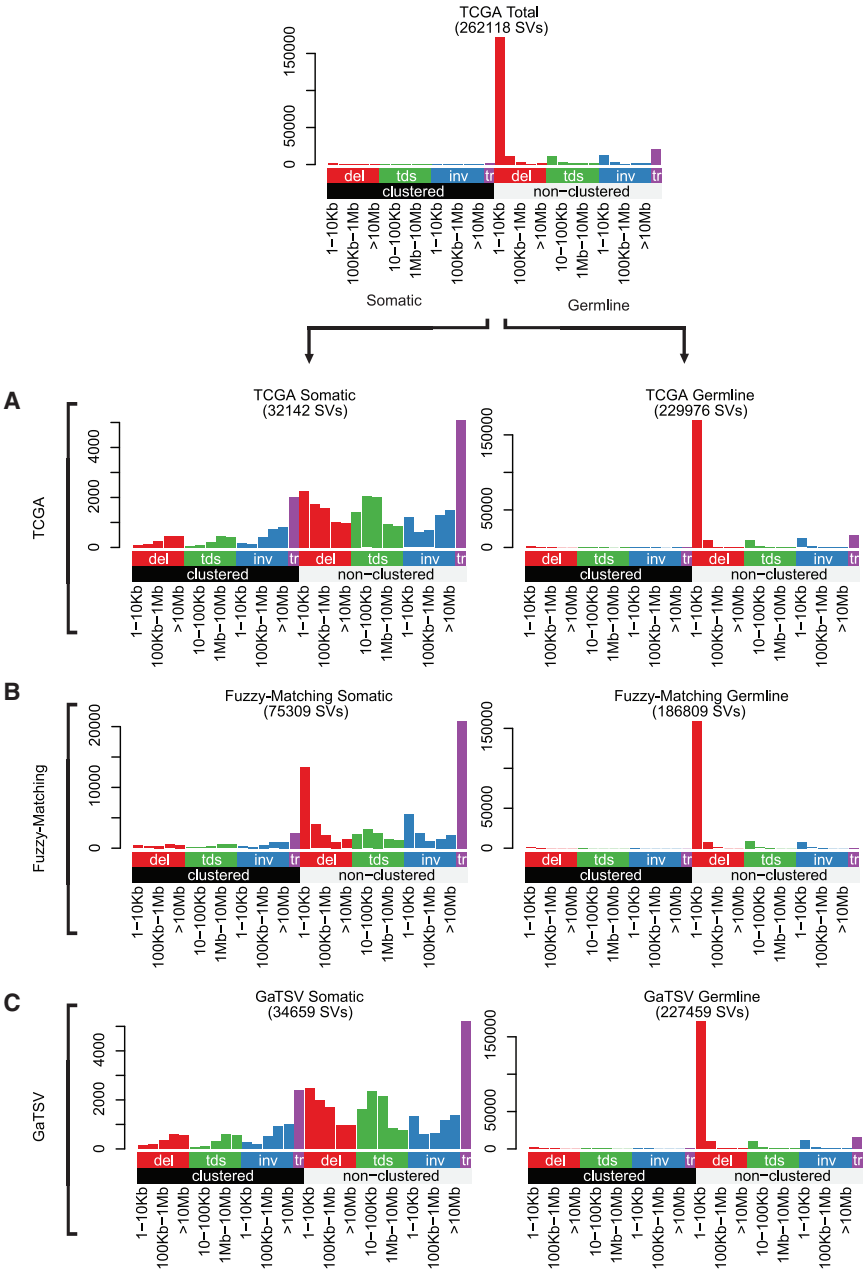


Figure 6. Distribution of SV categories (SV, signature input catalogs) for true and predicted germline and somatic SVs

(A–C) SV catalogs for the GaTSV-classified SVs (C) match those of the true TCGA SVs (A) more closely than the fuzzy matching to the gnomAD database (B).

GaTSV enables the extraction of somatic SV signatures in the absence of paired normal SV calls

The effects of somatic SV-generating processes were recently characterized across TCGA using SV signature analysis.[3] Somatic SV signature analysis has only been described after the removal of all germline SVs from the call set through joint analysis with paired normal tissue. We wanted to test if GaTSV allowed the accurate extraction of these biologically relevant patterns of somatic SVs without a paired normal. Organizing the set of all SVs in the TCGA dataset into the SV signature input catalog showed an abundance of short non-clustered deletions (Figure 6A), the majority of which were germline SVs. Both the gnomAD fuzzy matching-classified germline SVs and the GaTSV-classified germline SVs recaptured this true germline SV deletion peak (Figures 6B and 6C). However, gnomAD fuzzy matching misclassified many of these short non-clustered deletions as somatic, resulting in a falsely inflated total number of somatic SVs to more than double the true count of somatic SVs. In contrast, GaTSV correctly matched both the distribution and counts of SVs seen in the true somatic and germline SV signature input catalogs.

We extracted previously published SV signatures[3] based on these catalogs and primarily detected reference SV signature R9 in the germline profile and only minor contributions of the other signatures (Figures 7A, S5A, and S5B). In contrast, somatic SVs were composed of a variety of signatures (Figures 7A, S5A, and S5B). This difference in the reference signature proportions across germline and somatic SVs suggests that the mechanisms behind somatic SV generation are more varied than those behind germline SV generation. It is also possible that there are additional germline SV signatures that we did not detect because the reference signatures we used were extracted from somatic SVs.

In the comparison of reference SV signature proportions, GaTSV also outperformed gnomAD fuzzy matching. The medians of the GaTSV-predicted "somatic" and "germline" signature

To evaluate whether this drop-off in performance was unique to GaTSV, we also tested the gnomAD fuzzy matching method on different ancestries. Again, we saw the best performance on European SVs and the worst on African SVs, across all metrics (Figures 5G–5L). European genomes are the most represented ancestry group in the gnomAD database at over 77%.[24] These results highlight the issue of ancestry biases in public databases. GaTSV's performance on the worst-performing ancestry group—African—with a PPV of 0.66 was still notably better than the gnomAD fuzzy matching performance with a PPV of 0.42 on the best-performing ancestry group—European.

Please cite this article in press as: Chukwu et al., A sequence context-based approach for classifying tumor structural variants without paired normal samples, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100991

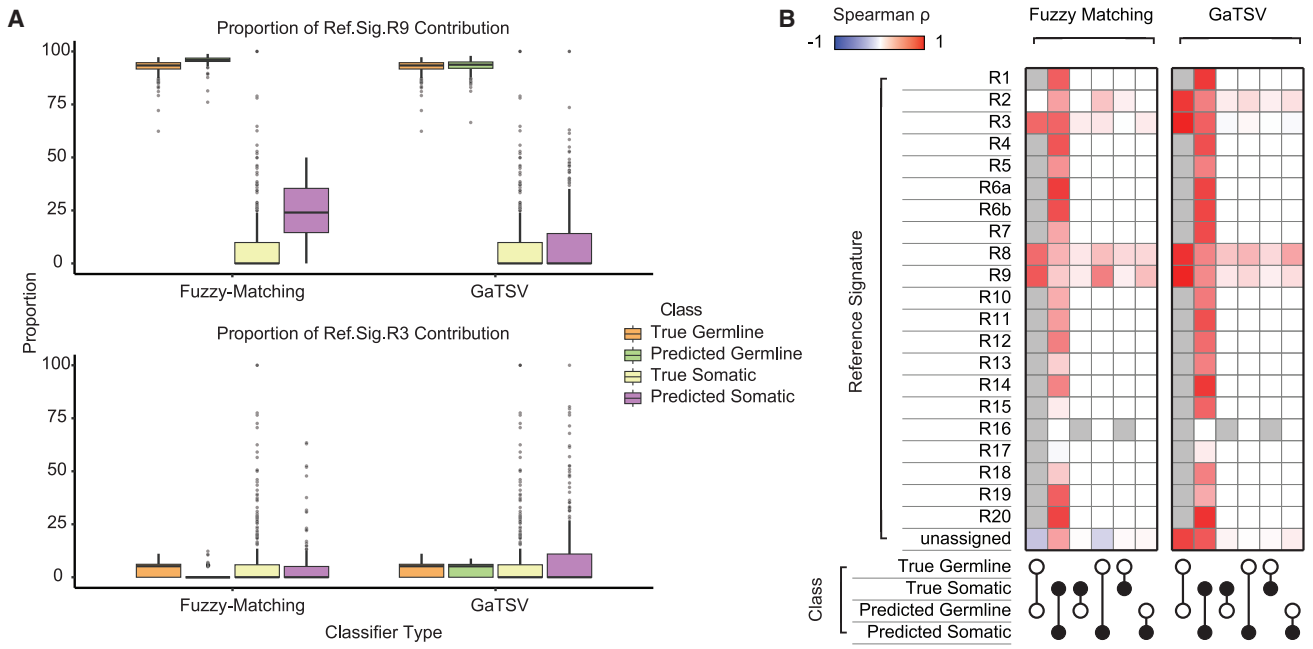


Figure 7. Analysis of classifier performances using rearrangement signatures

(A) R3 and R9 reference signature activity is more accurately recapitulated using GaTSV-classified SVs than fuzzy matching-classified SVs. The proportion of contributions of the R9 reference signature (top) and R3 reference signature (bottom) in each patient. Boxes represent the IQR with whiskers extending to 1.5× IQR and outliers beyond that as individual points. Box colors represent whether the germline or somatic SVs calls originated from SvABA (true), GaTSV, or fuzzy matching (predicted).

(B) Signature contributions based on GaTSV-classified SVs more closely match the true distributions compared with the fuzzy matching-classified SVs. Similarity of signature contributions based on true SVs and SVs called by each classifier were calculated by Spearman correlation for all reference signatures between all combinations of predicted vs. true germline vs. somatic SVs. The gray squares refer to undefined correlations between two zero vectors, or when no samples in either class have any contribution for a given reference signature.

contributions matched their respective true class median for reference signatures R3 (non-clustered translocations and short duplications) and R9 (short deletions); however, this was not the case for the gnomAD fuzzy matching-derived signatures (Figure 7A). GaTSV also showed high levels of correlation for called somatic to true somatic and called germline to true germline for almost all reference signatures (Figure 7B). These results further support our finding that the GaTSV overcomes the germline contamination issue in somatic SV calls without a paired normal.

GaTSV has reliable performance on SVs called by Manta

To test how well GaTSV classifies SVs from SV callers other than SvABA, we ran Manta[25] on the same set of pHGG tumor data used above. This resulted in 6,241 SVs in the test set. Although the specific distributions of features in SvABA- and Manta-called SVs had slight differences (Figures S4 and S6), the overall trends remained consistent. Moreover, in the Manta test set, GaTSV achieved a sensitivity (TPR) of 0.995, specificity (1 – FPR) of 0.747, and a PPV of 0.668. This demonstrates that, although GaTSV is optimized for SVs called by SvABA, it performs reliably on SVs from other SV callers.

DISCUSSION

Our analyses show that germline SVs are shorter, less likely to impact genes, and have more bases of homology adjacent to their breakpoints than somatic SVs. In contrast, somatic SVs are more likely to cluster together and are farther from transposable elements than germline SVs. These results strengthen previous findings that homology and transposon-based repair contribute more to the formation of germline SVs than somatic SVs. They also strengthen the intuition that germline genome structure is under much more stringent fitness constraints than somatic genome structure. We also establish that these differences can be used to computationally classify germline from somatic SVs in the absence of a matched normal, with high sensitivity and specificity.

Differences between DSB repair processes in somatic cells vs. germ cells and their progenitors are reflected in differences between SVs in these different contexts. Germline SVs tend to have more bases of homology than somatic SVs, revealing a preference for more accurate repair processes to maintain genome integrity. Somatic SVs are closer to each other than germline SVs, indicative of a higher likelihood of chromoanagenesis events in cancer cells.

Differences in fitness constraints in germline vs. somatic contexts are also reflected in their SVs. Most common germline SVs are short non-clustered deletions that do not impact coding sequences. Previous work showed that even small repeats resulting from duplication have been found to decrease cell fitness.[26] Therefore, most germline SVs lack gene dosage effects through copy-number alterations of the genomic

Please cite this article in press as: Chukwu et al., A sequence context-based approach for classifying tumor structural variants without paired normal samples, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100991

sequence or changes to protein structure. Similarly, most germline SVs tend to accumulate close to previously described germline loci[13] and are less likely to form in clusters. We postulate that these observations are the result of a selective process that eliminates *de novo* germline rearrangements that impact gene dose or protein functionality. The particular underlying process is worthy of further investigation and is plausibly attributed to embryonic lethality leading to variant extinction. As cells harboring somatic SVs do not need to develop from a single cell to an entire organism, they are likely subject to fewer fitness constraints than germline SVs. As a result, they are more tolerant of impacts on coding sequences and transcriptional dysregulation. We surmise that these SVs are under selection pressure that is divergent from germline SVs.

Previous studies have shown that the tumor context is capable of altering the replication timing landscape. Such transformation from early-to-late replication and vice versa are demonstrated in chromatin remodeling and methylation frequency.[21] The types of SVs were also linked to this altered chromatin compaction density between germline and somatic cells. While this consideration is not accounted for in our analysis, future studies can seek to uncover this potential bidirectional relationship between the cancer context and replication timing.

We took advantage of the differences between germline and somatic SVs to create an SVM classifier to distinguish them in cases where a matched normal sample is unavailable. Our SVM performed well (AUC = 0.99, PPV = 0.85), so we termed it the great GaTSV classifier. One significant strength of GaTSV is its ability to classify SVs that are not present in germline reference databases—this is critical as most SVs are rare and, therefore, not present in reference databases.

Although our classifier performed well overall, it tended to misclassify certain variants, such as its propensity to call somatic deletions shorter than 100 kb as germline rearrangements. These incorrect classifications may reflect true limitations in the great GaTSV and create a false identity for SVs. However, our analysis showed that features of germline SVs misidentified as somatic closely mimicked features of true somatic SVs and vice versa. While SV-generating processes differ substantially between germline and somatic contexts, some processes such as the integration of transposable elements and repeats into the genome are common to both.[27] This overlap could lead to similarities in the SV features, thus making them indistinguishable to the classifier.

The great GaTSV classifier will facilitate several avenues of new research to understand the formation and impact of germline and somatic SVs. Using GaTSV, SVs in clinical samples lacking a matched normal can be interrogated. The power of deeply characterized groups of cell lines such as those in the Cancer Cell Line Encyclopedia[9,28,29] can now be brought to bear on SVs, including discovering relationships between drug sensitivities, CRISPR dependencies, and gene expression on specific SVs, signatures of SVs, and SV abundance. The ability to accurately distinguish germline from somatic SVs in the absence of a matched normal will enable functional assessment of factors guiding SV formation and consequences for therapy development. Population-level databases such as gnomAD may contain somatic variants resulting from aging-related processes; this tool may provide for a more accurate catalog of variants.

Limitations of the study

Our tool has several limitations. First, it did not perform as well on patients of African ancestry compared with patients of European ancestry, likely due to the underrepresentation of individuals of African descent in our TCGA training dataset. Second, our tool is optimized for SV calls from SvABA,[14] as it was both trained and its hyperparameters validated using breakpoints called by SvABA, which may follow different conventions from other SV callers. Although GaTSV does have reliable performance on Manta SV calls, optimal performance on these SVs would require retraining or reselecting hyperparameters specific to each SV caller.

Another potential limitation of our study is the presence of artifactual rearrangements in SV calls. SvABA uses short-read sequencing data, potentially resulting in reads that align to multiple repetitive loci. These multi-mapping reads can result in artifactual SV calls present in many germline and somatic SV call sets. Since these SVs are present in both germline and somatic calls, they will be classified as germline by traditional methods that compare somatic SVs with a matched germline sample, such as our training set. Therefore, the GaTSV classifier will classify these artifactual SVs as germline. Hence, while we expect that the GaTSV classifier can identify somatic variants from the pool of germline and artifactual SVs, it is more difficult to draw biological conclusions from the GaTSV germline calls, as they are contaminated with artifactual SV calls. In future work, these issues can be addressed with long-read sequencing data, diversifying the ancestries represented in the training set, and extending GaTSV to use SV calls from other software such as Manta and GRIDSS2.[25,30]

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Frank Dubois (frank.dubois@charite.de).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- The reference datasets used to support the conclusions of this article are available online and listed in the key resources table.[3,13,31–34]
- TCGA sequence data can be accessed with dbGaP authorization under accession number dbGaP: phs000854.v3.p8.
- pHGG sequence data can be accessed with dbGaP authorization under accession number dbGaP: phs002380.v1.p1.
- All annotation functions, the scripts used to create our figures, and the trained classifier are available on the GaTSV Github (https://github.com/beroukhim-lab/GaTSV). An archival DOI is provided in the STAR Methods key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank and acknowledge the following funding sources: Fund for Innovation in Cancer Informatics (to S.D. and R.B.), the Gray Matters Brain Cancer

Foundation (to R.B.), the Pediatric Brain Tumor Foundation (to R.B.), and Break Through Cancer (to R.B.). F.D. is supported by the Max Eder program of the German Cancer Aid and a participant in the BIH Charité Junior Clinician Scientist program. S.D. was supported by an NIH NRSA award (F32), 1F32CA261024, and is currently supported by an NIH NIGMS career development award (K99/R00), 1K99GM155595.

AUTHOR CONTRIBUTIONS

F.D. and S.D. conceived and designed the work. W.C., S.L., A.C., S.Z., S.W., O.D., I.M., C.R., and M.I. acquired and analyzed data. W.C., S.L., A.C., S.Z., S.W., I.M., R.B., F.D., and S.D. interpreted the data. W.C., S.L., A.C., and S.Z. wrote the code for the GaTSV classifier algorithm. W.C., S.L., A.C., S.Z., S.W., R.B., F.D., and S.D. wrote the manuscript. All authors reviewed the manuscript.

DECLARATION OF INTERESTS

M.I. is on the scientific advisory board of ImmPACT Bio. R.B. consults for and owns equity in Scorpion Therapeutics, owns equity in Karyoverse Therapeutics, and receives research support from Novartis. We, the authors, have a patent application for the GaTSV classifier.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Data acquisition
  - Quality control and filtering
  - Feature annotation
  - gnomAD filtering
  - gnomAD fuzzy matching
  - Logistic regression
  - SVM
  - Ancestry analysis
  - Signature analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2025.100991.

Received: August 2, 2024
Revised: December 13, 2024
Accepted: February 12, 2025
Published: March 12, 2025

REFERENCES

1. Rheinbay, E., Nielsen, M.M., Abascal, F., Wala, J.A., Shapira, O., Tiao, G., Hornshøj, H., Hess, J.M., Juul, R.I., Lin, Z., et al. (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature *578*, 102–111.
2. Li, Y., Roberts, N.D., Wala, J.A., Shapira, O., Schumacher, S.E., Kumar, K., Khurana, E., Waszak, S., Korbel, J.O., Haber, J.E., et al. (2020). Patterns of somatic structural variation in human cancer genomes. Nature *578*, 112–121.
3. Degasperi, A., Amarante, T.D., Czarnecki, J., Shooter, S., Zou, X., Glodzik, D., Morganella, S., Nanda, A.S., Badja, C., Koh, G., et al. (2020). A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. Nat. Cancer *1*, 249–263.
4. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell *185*, 3426–3440.e19.
5. Pócza, T., Grolmusz, V.K., Papp, J., Butz, H., Patócs, A., and Bozsik, A. (2021). Germline Structural Variations in Cancer Predisposition Genes. Front. Genet. *12*, 634217.
6. Carvalho, C.M.B., and Lupski, J.R. (2016). Mechanisms underlying structural variant formation in genomic disorders. Nat. Rev. Genet. *17*, 224–238.
7. Dubois, F., Sidiropoulos, N., Weischenfeldt, J., and Beroukhim, R. (2022). Structural variations in cancer and the 3D genome. Nat. Rev. Cancer *22*, 533–546. https://doi.org/10.1038/s41568-022-00488-9.
8. AACR Project GENIE Consortium (2017). AACR Project GENIE: Powering Precision Medicine through an International Consortium. Cancer Discov. *7*, 818–831.
9. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.
10. Chen, F., Zhang, Y., Sedlazeck, F.J., and Creighton, C.J. (2024). Germline structural variation globally impacts the cancer transcriptome including disease-relevant genes. Cell Rep. Med. *5*, 101446.
11. Kronenberg, Z.N., Osborne, E.J., Cone, K.R., Kennedy, B.J., Domyan, E.T., Shapiro, M.D., Elde, N.C., and Yandell, M. (2015). Wham: Identifying structural variants of biological consequence. PLoS Comput. Biol. *11*, e1004572.
12. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. *15*, R84.
13. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451.
14. Wala, J.A., Bandopadhayay, P., Greenwald, N.F., O'Rourke, R., Sharpe, T., Stewart, C., Schumacher, S., Li, Y., Weischenfeldt, J., Yao, X., et al. (2018). SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. *28*, 581–591.
15. Vassetzky, N.S., and Kramerov, D.A. (2013). SINEBase: a database and tool for SINE analysis. Nucleic Acids Res. *41*, D83–D89.
16. Alisch, R.S., Garcia-Perez, J.L., Muotri, A.R., Gage, F.H., and Moran, J.V. (2006). Unconventional translation of mammalian LINE-1 retrotransposons. Genes Dev. *20*, 210–224.
17. Kim, E.Z., Wespiser, A.R., and Caffrey, D.R. (2016). The domain structure and distribution of Alu elements in long noncoding RNAs and mRNAs. RNA *22*, 254–264.
18. Balachandran, P., Walawalkar, I.A., Flores, J.I., Dayton, J.N., Audano, P.A., and Beck, C.R. (2022). Transposable element-mediated rearrangements are prevalent in human genomes. Nat. Commun. *13*, 7115.
19. Holland, A.J., and Cleveland, D.W. (2012). Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. Nat. Med. *18*, 1630–1638.
20. Kraft, K., Murphy, S.E., Jones, M.G., Shi, Q., Bhargava-Shah, A., Luong, C., Hung, K.L., He, B.J., Li, R., Park, S.K., et al. (2024). Enhancer activation from transposable elements in extrachromosomal DNA. Preprint at bioRxiv. https://doi.org/10.1101/2024.09.04.611262.
21. Du, Q., Bert, S.A., Armstrong, N.J., Caldon, C.E., Song, J.Z., Nair, S.S., Gould, C.M., Luu, P.-L., Peters, T., Khoury, A., et al. (2019). Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. Nat. Commun. *10*, 416.
22. Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., and Lichtenstein, L. (2019). Calling Somatic SNVs and Indels with Mutect2. Preprint at bioRxiv. https://doi.org/10.1101/861054.

23. Adelson, R.P., Renton, A.E., Li, W., Barzilai, N., Atzmon, G., Goate, A.M., Davies, P., and Freudenberg-Hua, Y. (2019). Empirical design of a variant quality control pipeline for whole genome sequencing data using replicate discordance. Sci. Rep. *9*, 16156.

24. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2021). Author Correction: The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *590*, E53.

25. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222.

26. Adler, M., Anjum, M., Berg, O.G., Andersson, D.I., and Sandegren, L. (2014). 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.

27. Zamudio, N., and Bourc'his, D. (2010). Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? Heredity *105*, 92–104.

28. Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., 3rd, Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature *569*, 503–508.

29. Corsello, S.M., Nagari, R.T., Spangler, R.D., Rossen, J., Kocak, M., Bryan, J.G., Humeidi, R., Peck, D., Wu, X., Tang, A.A., et al. (2020). Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. Nat. Cancer *1*, 235–248.

30. Cameron, D.L., Baber, J., Shale, C., Valle-Inclan, J.E., Besselink, N., van Hoeck, A., Janssen, R., Cuppen, E., Priestley, P., and Papenfuss, A.T. (2021). GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. Genome Biol. *22*, 202.

31. Smit, A.F.A., Hubley, R., and Green, P.. RepeatMasker Open-4.0 2013-2015 Repeat Elements Data. https://www.repeatmasker.org/species/hg.html.

32. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nat. Genet. *46*, 1160–1165.

33. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93.

34. Carrot-Zhang, J., Chambwe, N., Damrauer, J.S., Knijnenburg, T.A., Robertson, A.G., Yau, C., Zhou, W., Berger, A.C., Huang, K.-L., Newberg, J.Y., et al. (2020). Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. Cancer Cell *37*, 639–654.e6.

35. Dubois, F.P.B., Shapira, O., Greenwald, N.F., Zack, T., Wala, J., Tsai, J.W., Crane, A., Baguette, A., Hadjadj, D., Harutyunyan, A.S., et al. (2022). Structural variants shape driver combinations and outcomes in pediatric high-grade glioma. Nat. Cancer *3*, 994–1011. https://doi.org/10.1038/s43018-022-00403-z.

36. Howell, D.C. Statistical Methods for Psychology (Eighth Edition).Cengage Learning.

37. Fisher, R.A., and Others (1921). 014: On the" Probable Error" of a Coefficient of Correlation Deduced from a Small Sample.

38. Breslwo, N.E., and Day, N.E. (1980). Statistical Methods in Cancer Research Volume I: The Analysis of Case-Control Studies (IARC Scientific Publications).

Please cite this article in press as: Chukwu et al., A sequence context-based approach for classifying tumor structural variants without paired normal samples, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100991

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA sequence data	Weinhood et al.[32]	dbGaP: phs000854.v3.p8
pHGG sequence data	Dubois et al.[35]	dbGaP: phs002380.v1.p1
gnomAD v4.0	Collins et al.[13]	https://gnomad.broadinstitute.org/downloads#v4-structural-variants
Replication timing	Weddington et al.[33]	http://mskilab.com/fishHook/hg19/RT_NHEK_Keratinocytes_Int92817591_hg19.rds
Repeat elements database	Repeatmasker via UCSC Genome Browser	https://genome.ucsc.edu/cgi-bin/hgTrackUi?g=rmsk
Software and algorithms		
GaTSV	This paper	https://github.com/beroukhim-lab/GaTSV; https://doi.org/10.5281/zenodo.14756714
SvABA	Wala et al.[14]	https://github.com/walaj/svaba
signature.tools.lib	Degasperi et al.[3]	https://github.com/Nik-Zainal-Group/signature.tools.lib
Manta	Chen et al.[25]	https://github.com/Illumina/manta
R version 4.2.2	R Foundation	https://www.r-project.org/

METHOD DETAILS

Data acquisition

Whole genome sequencing tumor samples with matched normals were obtained from the TCGA patient cohort[32] (Table S4). We used SvABA to call SVs on all samples as previously described, which were used for our training and test sets.[14] The pHGG test set consisted of six samples from our recent study.[35] We also ran SvABA and Manta on the pHGG samples, which formed the SVs used in our external test set.

Quality control and filtering

From an initial set of 974 samples, we filtered out samples with more than 5000 total SVs, with greater than 50% being inversions. We believed that these samples, which were also excluded from the PCAWG comprehensive analysis, were primarily composed of artifactual SVs. This filter excluded 11 samples and we proceeded with 963 samples for the rest of our analysis.

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 at least one breakpoint, were not detected solely by discordant reads, had a span of 50bp (1000bp for inputs to GaTSV) 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.

To analyze the samples run through Manta, the "somaticSV.vcf.gz" and "diploidSV.vcf.gz" files were unzipped, and the SVs from each file were labeled as somatic and germline, respectively. Similar to SvABA outputs, the Manta outputs labeled as "MantaBND" in the ID field of the vcfs were grouped according to their breakend pair ID—the part of the ID field following "MantaBND" and the last colon (":"). All other events were analyzed as they appeared in the vcf file. In order to filter out potentially artifactual or low-evidence reads, only SVs with at least 5 split and discordant reads were analyzed. Moreover, to account for the purity of the normal samples compared to the tumor samples, only SVs with a 4:1 alternate to reference read ratio were taken for germline SVs. Finally, we only tested GaTSV on Manta SVs with a SPAN of greater than or equal to 1000bp and translocations, similar to how GaTSV was trained and tested for SvABA-called SVs.

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 5' breakpoint of the rearrangement to the 5' breakpoint of a gnomAD reference SV and vice versa; also see Figure S7A). These distances were averaged, and the gnomAD SV with the lowest average was labeled as the closest SV. If a matching gnomAD SV was not found, as is the case for many translocations, we input an artificial distance of 1e9 for each breakpoint.

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 from TCGA samples available on the ICGC Data Portal.[33] 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). 187 of our samples were not analyzed in the PCAWG consensus calls. For these, we assigned the *TP53* status as indeterminate (Table S4). 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 events.[31] 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 SVs, 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 (also see Figure S7B).

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 (Table S5). 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 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.

gnomAD filtering

We obtained the gnomAD v4.0 release from the open-source gnomAD browser.[13] We filtered for variants that passed all gnomAD filters and had resolved breakpoints. In addition, we excluded all complex (CPX) SVs due to unclear breakpoint origins, as well as insertion (INS) SVs that lacked a source sequence. For insertion SVs that reported a source sequence, we resolved the breakends to reflect a corresponding intrachromosomal or interchromosomal translocation.

gnomAD fuzzy matching

For each rearrangement in our test set, we first determined the distance to the closest gnomAD reference SV as described in the feature annotation section (also see Figure S7A).

Each of the discrete values of average distance to gnomAD was treated as a cutoff, and we constructed an ROC curve by classifying all rearrangements with an average distance less than that cutoff as germline. For each of the cutoffs, the predicted and actual classifications were used to calculate the specificity and sensitivity, which were used to generate the ROC curve.

Logistic regression

We trained 21 logistic regression models - one for each feature - using the glm method from the R stats package. Our training set included a random sample of 200k SVs from the full training set described in the quality control and filtering section that were scaled as described in the feature annotation section. We used each single-feature logistic regression model to compute prediction

Please cite this article in press as: Chukwu et al., A sequence context-based approach for classifying tumor structural variants without paired normal samples, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100991

probabilities on 100k scaled SVs sampled from the full test set. Prediction probabilities were derived using the R stats predict method. Finally, we calculated the AUC, TPR, and FPR values to evaluate the performance of each model (Figure 3C).

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 1, 10, 100, 500, 1000 for both RBF and linear kernels and gamma of 0.0001, 0.001, 0.01, 0.05 0.1, 1 for the RBF kernel. For each combination of hyperparameters, we performed a 5-fold cross-validation, splitting the training set into five groups. One group was selected as a validation set for each of the five cross-validation iterations, 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 for 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 generate the GaTSV ROC curve and 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.

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; these were kept as 10, 0.1, and RBF kernel respectively.

In constructing our classifier, we chose a probability cutoff that optimized for the sum of TPR and PPV, 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 such cases, a cutoff of the maximum sum of the TNR and NPV or the maximum AUC in the same TCGA validation set can be used.

Ancestry analysis

Using consensus ancestry calls of TCGA patients from a previous study, we conducted a lookup of patients in our cohort to those in the consensus calls.[34] After excluding the 25 patients who were not in the consensus calls, the test set of 277,925 TCGA SVs—each with an associated patient—were labeled with their corresponding ancestry call. For our analysis, we did not consider any of the 50 admixed patients. We also did not include the "amr" and "sas", corresponding to American and South Asian, ancestries, as there were only seven and two patients within those categories, respectively. For the remaining "afr", eas", and "eur", corresponding to African, East Asian, and European, ancestries, we took subsets of the test set according to their ancestry labels, and analyzed the performance of the classifier on each subset (Figure 5).

Signature analysis

The signature analysis was performed primarily using the signature.tools.lib package published by the Nik-Zainal group.[3] From bedpe files containing SVs for each patient in our cohort, we first created true germline and somatic signature catalogs for each patient using the bedpeToRearrCatalogue function in the signature.tools.lib package. These signature catalogs consisted of counts of the number of rearrangements organized across three categories: SV span, type, and clustering of all SVs of span above 1kb. This process was repeated for predicted germline and somatic rearrangements from both the fuzzy matching approach and the SVM classifier.

These catalogs were then fit to the reference signature profile for rearrangements from the same study using the Fit and plotFit functions.[3] This resulted in the true and predicted germline and somatic reference signature exposures for each patient, which shows how much contribution each reference signature has in the collection of SVs in each patient. For Figures 7A and S5, we determined the ratio of a given reference signature exposure to the sum of all exposures within a patient for all reference signatures. In the case that there were no SVs in a patient—for a given class—the proportion was simply set as 0, instead of an undefined value.

In Figure 7B, we first conducted a modified Spearman correlation between each predicted and true class. Each vector used for the Spearman correlation had a length of 963 and was composed of the proportion values for a given reference signature and a given class. Each predicted vector was paired with another true vector with the same reference signature. If both vectors were nonzero, then a normal Spearman correlation was conducted. If this was not the case, we calculated the correlation as follows: If both vectors were zero, a Spearman rho of 1 was assigned. If one vector was zero and the other was nonzero, a Spearman rho of 0 was assigned.

QUANTIFICATION AND STATISTICAL ANALYSIS

We performed Kolmogorov-Smirnov tests to determine if the distributions of continuous features between germline and somatic SVs significantly differed. Our p-value significance threshold was set to 0.05. For features quantified in discrete variables, we used Chi-squared tests with a significance threshold of 0.05.

To assess the significance of associations between variables both within SV classes and across classes, we performed a variety of statistical tests (Figure 2). For continuous/continuous feature comparisons within germline and somatic SVs, we computed

Spearman ρ and defined moderate to high correlation as $|\rho| \geq 0.5$. We then obtained p-values for the Spearman ρ values using two-sided t-tests as described previously.[36] Next, we performed false discovery rate (FDR) multiple hypothesis correction with a q-value threshold of 0.05. To check whether these correlations significantly differed across germline and somatic SVs, we derived z-scores from the ρ values using Fisher's z-transformation[37] and subsequently obtained p-values. We performed FDR correction on these p-values and defined significance at a q-value threshold of 0.05.

For continuous/discrete variables, we calculated Wilcoxon ranked sum tests. Within both somatic and germline SVs, we compared observations in the binary ingroup to those in the binary outgroup. We reported the Hodges Lehmann estimator of location shift and the FDR-corrected p-values with a significance threshold of 0.05. For across-group comparisons, we also computed Wilcoxon ranked sum tests. Here, we compared continuous variable observations in the binary ingroup of germline SVs against the binary ingroup of somatic SVs.

For binary/binary variables, we performed Chi-squared tests on 2x2 contingency tables within germline and somatic SVs. We only compared binary variables that were non-mutually exclusive to obtain a relevant statistic. We reported the odds ratio and FDR-corrected p-values with a significance threshold of 0.05. For across-group binary/binary comparisons, we implemented the Breslow-Day test for homogeneity of associations on the 2x2x2 contingency tables.[38] This test checks if the odds-ratio across different strata - in this case, germline and somatic - are significantly different for any pair of binary variables. We reported the difference in odds ratios across the different strata and FDR-corrected p-values with a 0.05 significance threshold.