Comparison of germline and somatic structural variants in cancers reveal systematic differences in variant generating and selection processes

Wolu Chukwu[1,2,*], Siyun Lee[1,2,*], Alexander Crane[1,2,*], Shu Zhang[1,2,*], Sophie Webster[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;

* Equal contributors

\# Jointly supervised

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

Author email addresses:

W.C.: wchukwu@broadinstitute.org

S.L.: siyun@broadinstitute.org

A.C.: acrane@broadinstitute.org

S.Z.: shu@broadinstitute.org

S.W.: swebster@broadinstitute.org

I.M.: ipsam001@gmail.com

M.I.: mai9037@med.cornell.edu

R.B.: rameen@broadinstitute.org

F.D.: frank.dubois@charite.de

S.D.: sdalin@broadinstitute.org

Abstract

Although several recent studies have characterized structural variants (SVs) in germline and cancer genomes, the features of SVs in these different contexts have not been directly 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 (TCGA). 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, we found that transposon-mediated processes shape germline much more than somatic SVs, while somatic SVs more frequently show features characteristic of chromoanagenesis. These differences were extensive enough to enable us to develop a classifier - “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 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, yet germline and cancer SVs have not been directly compared.

56 SVs arise from DSBs that are repaired to incorrect loci in the genome. The resulting
57 rearrangements can delete, duplicate, invert, or translocate segments of DNA. The responsible
58 mechanisms of DSB repair include non-homologous end joining (NHEJ), which pastes broken
59 ends of DNA together irrespective of adjacent sequence homology; microhomology-mediated
60 end joining (MMEJ), which utilizes 3-10bp of breakpoint-adjacent microhomology to form repair
61 intermediates; and homologous recombination (HR), which uses more bases of homology.
62
63 Germline and somatic SVs are likely to arise from different DSB repair mechanisms. Previous
64 studies report that germline SVs primarily result from non-allelic homologous recombination
65 (NAHR), where substantial amounts of sequence homology at distinct loci are employed in
66 double-stranded break repair, often resulting in the deletion of the intervening sequence [5,6].
67 Conversely, somatic SVs, which present with more varied spans and clustering of breakpoints,
68 indicate a tendency towards NHEJ and replication-based mechanisms of repair such as
69 microhomology-mediated break-induced replication (MMBIR) which are more error-prone[5–7].
70 However, a direct and comprehensive comparison between the features of germline and
71 somatic SVs is likely to provide a more detailed view of the differences in their generation, with
72 implications for the activity of different DNA damage and repair processes in human vs. somatic
73 cell evolution.
74
75 One benefit of recognizing the different features of germline and somatic SVs is that it provides
76 an opportunity to distinguish germline and somatic SVs when only data from somatic tissue
77 have been collected. Currently, confidently designating an SV as “somatic” requires comparing
78 sequencing data from highly clonal somatic tissue such as cancers or single cells with
79 multiclonal “normal” tissues that represent the germline. However, there are many situations in

80 which normal tissue is unavailable, including clinical settings[8] and the study of long-term cell
81 line models[9]. The ability to distinguish germline from somatic SVs in the absence of normal
82 tissue is therefore valuable both clinically and in cancer research.
83
84 Here, we comprehensively evaluate similarities and differences between germline and somatic
85 SVs. We find that they are strikingly different, enabling us to develop a machine learning-based
86 classifier (the Germline and Tumor Structural Variant classifier, aka great GaTSV) that can
87 distinguish germline from somatic SVs in the absence of a matched normal sample with
88 extremely high accuracy.

89 Results

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

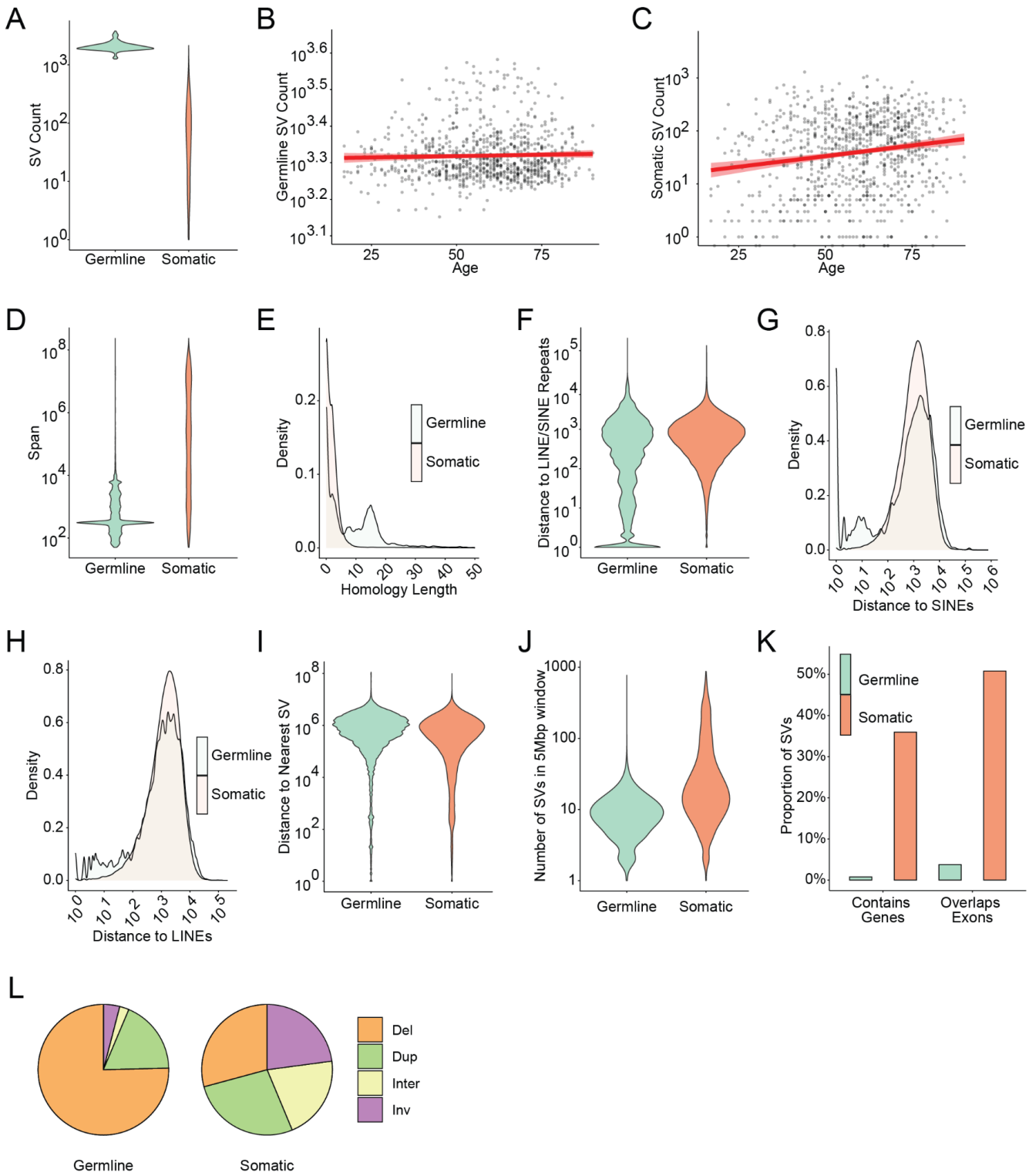


Figure 1. *SV features differ significantly between germline and somatic genomes* ***A****, Distribution of germline and somatic SV frequencies in TCGA genomes.* ***B-C****, Pearson*

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-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 5Mbp 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.*

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 **Supplementary Table 1 and Supp. Fig S1A-D**). 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-value < 2.2e-16; **Fig 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[11,12]. 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 (**Supp Fig. S1E**). These results are consistent with previous studies of germline and somatic SVs [2].

124

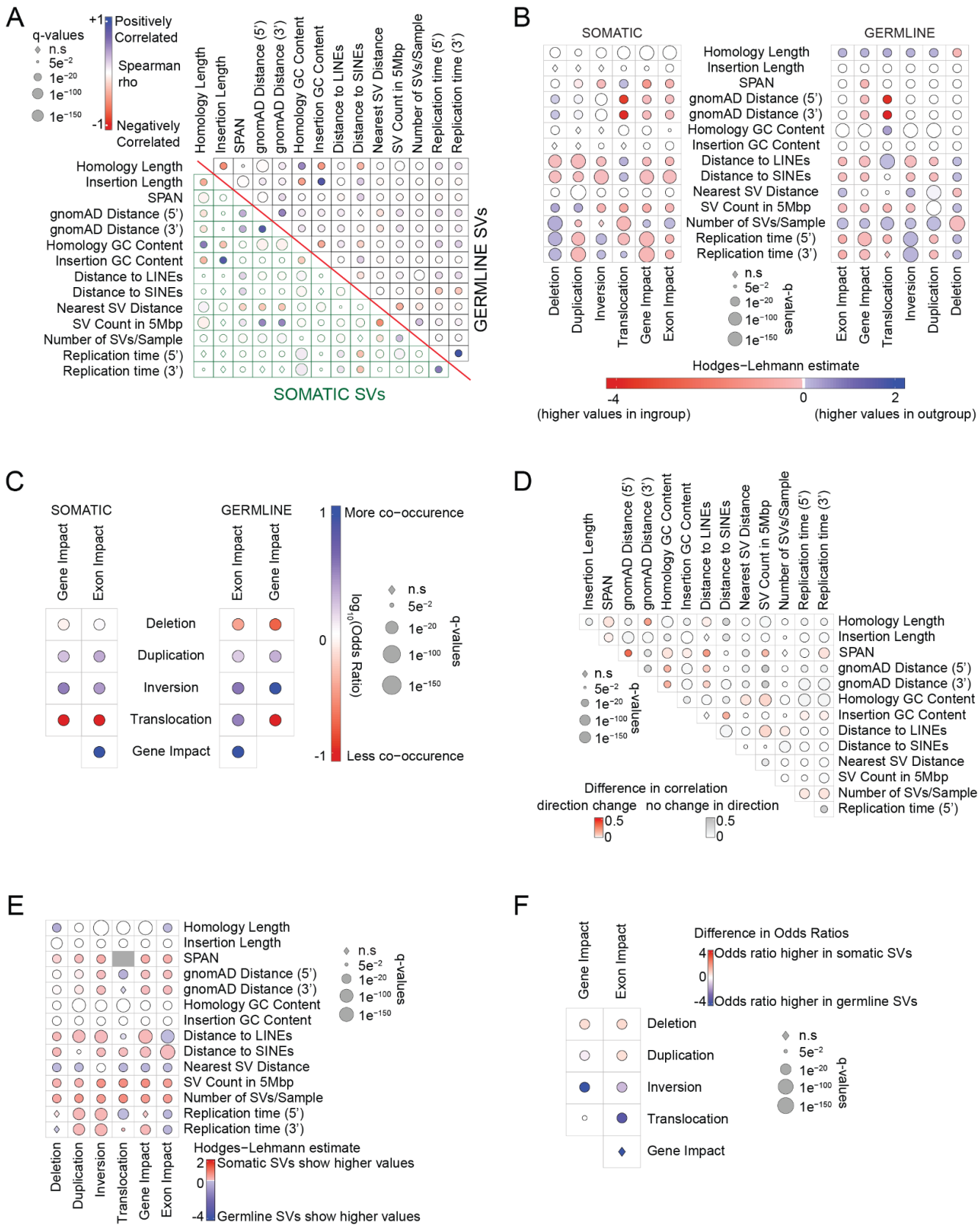
125 The second most striking difference between germline and somatic SVs was the much higher
126 levels of breakpoint homology attributed to germline SVs (KS test, p-value < 2.2e-16; **Fig 1E**),
127 suggesting a transposon-mediated origin. Closer examination of the distribution of homology
128 lengths revealed a peak between 13-17bp in germline SVs but not somatic SVs. This peak was
129 specifically present in germline deletions of ~300bp span (**Supp Fig. S1F-G**), corresponding
130 to the spans of Alu elements - a type of SINE element and the most abundant transposons in
131 the human genome[13]. Previous studies have shown that Alu elements comprise a significant
132 proportion of transposable element-mediated rearrangements in the genome and use ~15 bp
133 of homology[14]. Germline SVs were closer to SINE and LINE elements than somatic SVs (KS-
134 test, p-value < 2.2e-16, **Fig. 1F-H**, **Supp Fig. S2A-B**), and—of all repeat elements—these
135 showed the greatest difference in range of distances between germline SVs and somatic SVs
136 (**Supp Fig. S2C-Q**). Interestingly, somatic SVs were closer to all classes of RNA pseudogenes
137 compared to germline SVs.

138

139 In contrast, somatic SVs are more likely to be formed by chromoanagenesis[15] and affect gene
140 structure. Somatic SVs were more likely to be found in proximity to each other (**Fig. 1I-J,** p-
141 value < 2.2e-16, KS-test) and were more likely to disrupt coding sequences or span entire
142 genes (Fisher's Exact Test, both p-values < 2.2e-16, **Fig 1K**). Strikingly, 51% of somatic SVs
143 directly affected the exome, in contrast to only 3.8% of germline SVs. Finally, deletion events
144 made up about 75% of germline events but only 29% of somatic events (**Fig 1L**, Chi-Squared
145 test, p-value < 2.2e-16), whereas somatic SVs were nine times more likely to be translocations
146 (Chi-Squared test, p-value < 2.2e-16). Overall, our analysis of single features of SVs showed
147 pronounced differences in the characteristics of germline compared to somatic SVs.

148

149 If there are specific variant-generating or selection processes that result in germline and
150 somatic SVs, we expect the features of SVs (**Supplementary Table 1**) that are linked to such
151 processes to vary together. Indeed, characteristic combinations of SV features between
152 germline and somatic SVs reflected the pronounced differences in the relationships between
153 SV features (**Fig. 2A-C**). Somatic SVs with shorter spans tended to have longer homology (**Fig.**
154 **2A**), while no such associations were observed for germline SVs. Germline SVs had longer
155 homology if they were closer to known SINE elements or were deletions (**Fig. 2A, 2B**), while
156 no such association existed for somatic SVs (**Fig. 2A, 2D**). Longer germline SVs also showed
157 higher homology GC content while we observed the opposite in somatic SVs (**Fig. 2D**). These
158 data suggest that transposon-mediated processes dominate germline SV generation while only
159 a subset of somatic SVs originate from this pathway. In contrast, somatic SVs showed an anti-
160 correlation between homology length and GC content with the number of SVs within 5Mbp as
161 well as SV span with its distance to the nearest SV (**Fig 2A, 2D**). These data indicate a subset
162 of (longer) somatic SVs generated in clusters by NHEJ in a more complex process like
163 chromothripsis.


A
q-values
n.s
5e−2
1e−20
1e−100
1e−150
+1 Positively Correlated
Spearman rho
-1 Negatively Correlated
Homology Length
Insertion Length
SPAN
gnomAD Distance (5')
gnomAD Distance (3')
Homology GC Content
Insertion GC Content
Distance to LINEs
Distance to SINEs
Nearest SV Distance
SV Count in 5Mbp
Number of SVs/Sample
Replication time (5')
Replication time (3')
GERMLINE SVs
SOMATIC SVs
B
SOMATIC
GERMLINE
Deletion
Duplication
Inversion
Translocation
Gene Impact
Exon Impact
q-values
Hodges−Lehmann estimate
-4
0
2
(higher values in ingroup)
(higher values in outgroup)
C
SOMATIC
GERMLINE
Gene Impact
Exon Impact
Deletion
Duplication
Inversion
Translocation
Gene Impact
1 More co−occurence
0
-1 Less co−occurence
log10(Odds Ratio)
q-values
D
Difference in correlation
direction change
no change in direction
0.5
0
E
Hodges−Lehmann estimate
2 Somatic SVs show higher values
0
-4 Germline SVs show higher values
F
Difference in Odds Ratios
4 Odds ratio higher in somatic SVs
0
-4 Odds ratio higher in germline SVs

164

165

166 ***Figure 2.*** *SV feature associations differ between somatic and germline SVs.* ***A****, Spearman*
167 *correlations between continuous features within somatic and germline SVs (positive*
168 *correlations in blue, anticorrelation in red).* ***B****, Differences in the values of continuous SV*
169 *features between categorical SV features within germline and somatic SVs.* ***C****, Odds ratios of*
170 *the associations between the categorical variables gene or exon impact and SV type in*
171 *germline and somatic SVs.* ***D****, Difference in Spearman correlation values between continuous*
172 *features across germline and somatic SVs.* ***E****, Differences in the values of continuous SV*
173 *features between germline and somatic SVs within categorical SV features.* ***F****, Difference in*
174 *odds ratios of categorical variable comparisons across germline and somatic SVs. For all*
175 *panels, the circle size represents the significance of the statistic represented.*
176
177 The differences in feature associations between germline and somatic SVs also hinted at the
178 strong differences in the selection pressures they face. Somatic deletions were significantly
179 more likely to span gene/exons than germline deletions. (**Fig 2C, F**). Also, while duplications
180 were also more likely to span genes than other SV types in both germline and somatic SVs
181 (**Fig. 2C**, q-value < 2.2e-16, Chi-Squared test), this association was significantly stronger in
182 somatic SVs (**Fig. 2F**). Germline inversions, which could be more neutral in their effects on
183 gene expression, were more likely to affect genes/exons than other germline SV types (**Fig.**
184 **2C**) and also more likely to span genes/exons than somatic inversions (**Fig. 2F)**. Our analysis
185 of the association between replication timing and SV type also found results consistent with
186 strong selection pressures conserving the coding genome in the germline. We observed a
187 positive correlation between homology length and replication timing among germline SVs that
188 was absent or insignificant among somatic SVs (**Fig. 2A**). This shows an enrichment for high-
189 fidelity repair in gene-dense regions in the germline context that is not observed in cancer-

190 associated SVs. Also, while there was an overall bias for inversions and deletions towards later
191 replicating genomic loci and duplications for earlier replicating loci (**Fig. 2B**), striking differences
192 arise when comparing across germline and somatic SVs. Notably, somatic duplications and
193 inversions showed significantly higher replication timing values, corresponding to earlier
194 replicating regions, than germline duplications and inversions (**Fig. 2E**). Early-replicating
195 regions tend to be gene-dense and highly expressed.[16] Together, these data indicate strong
196 selection pressures to conserve the full coding genome in the germline that are attenuated in
197 cancer cells.

198 Extending SNV classification approaches to SVs

199 Next, we considered how we might use this information to distinguish between germline and
200 somatic SVs when paired germline DNA sequencing data are unavailable. As a control, we first
201 attempted to filter germline SVs using the gnomAD v4.0 population dataset[17], analogous to the
202 commonly-used filtering approach for removing germline SNVs[18,19]. However, only a fraction
203 of germline SVs matched within 3bp of a gnomAD SV (~1.0M SVs out of 2.0M germline SVs
204 with only ~31k SVs matching exactly to a reference gnomAD SV). The paucity of exact matches
205 may be due to differences in SV callers between gnomAD and our dataset as well as the fact
206 that the vast majority of germline SVs are not recurrent[17]. While the distance from gnomAD-
207 listed germline SVs was significantly different for germline and somatic SVs (**Supp. Fig S3A**,
208 p-value < 2.2e-16, **Supp. Fig S3A**), about a third of germline SVs lay more than 1000bp from
209 a gnomAD SV. Therefore, to match TCGA SVs to the gnomAD reference SVs, we determined
210 the average base pair distance of each pair of breakpoints from each TCGA SV to its closest
211 SV in gnomAD. This “fuzzy matching” approach had an overall AUC of 0.90. The optimal cutoff-
212 point on the ROC curve of ~1400bp average distance from the nearest gnomAD SV resulted

213 in a high degree of sensitivity in correctly classifying somatic SVs (TPR = 96% of true somatic

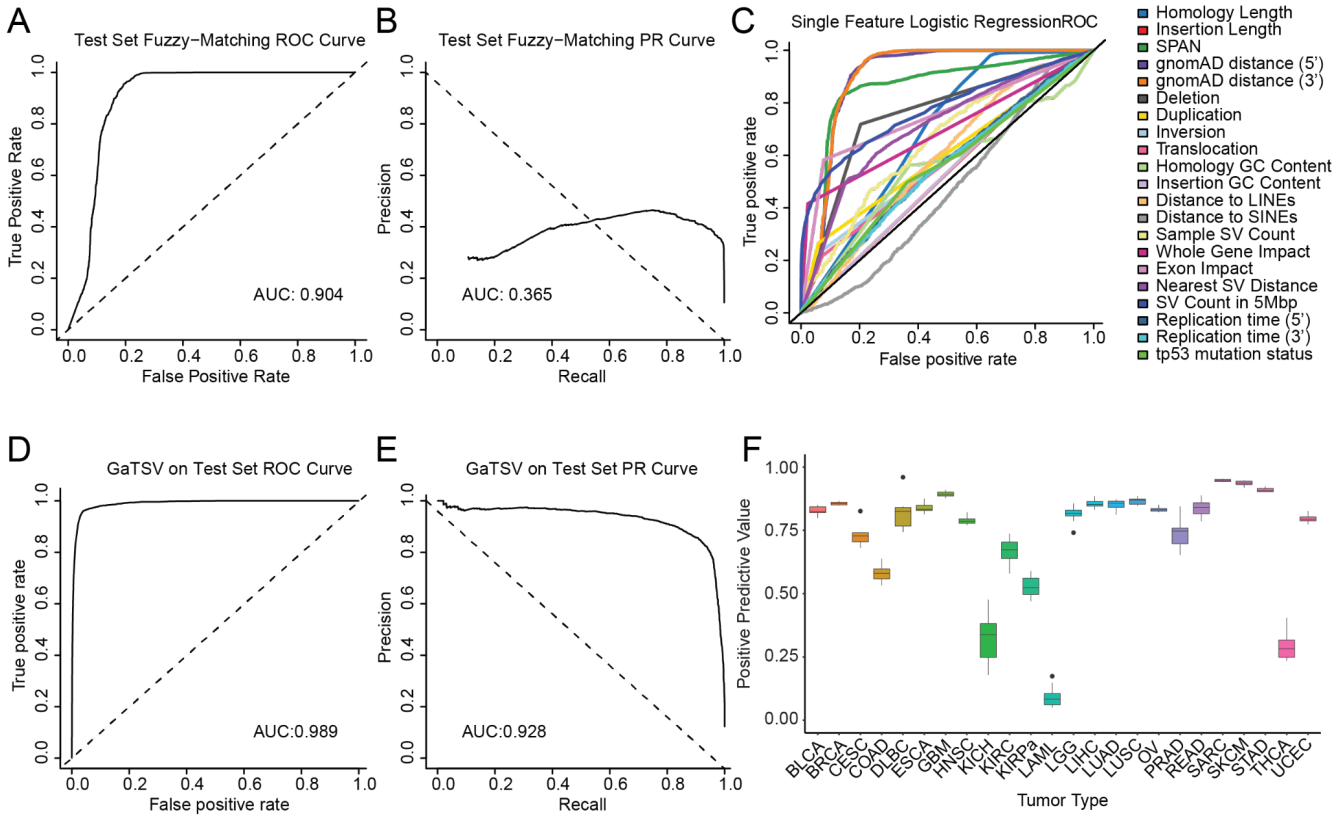
214 events). However, this cutoff still resulted in a large fraction of germline contamination (60%)

215 in the set of SVs called somatic (positive predictive value, PPV) (**Fig 3A)**. These data show

216 that a simple filter based on the closest germline SV in gnomAD cannot sufficiently differentiate

217 between germline and somatic SVs in tumor-only SV calls.

218



219

220 ***Figure 3.*** *Classifier performances on TCGA test set.* ***A****, ROC for classifying SVs based on a*

221 *cutoff distance to gnomAD reference (fuzzy-matching).* ***B****, PR curve for classifying SVs using*

222 *fuzzy-matching.* ***C****, ROC for separate single-feature logistic regression classifiers.* ***D****, ROC for*

223 *the GaTSV classifier.* ***E****, PR curve for the GaTSV classifier.* ***F****, Positive Predictive Value (PPV)*

224 *of the GaTSV classifier by tumor type showing variance in performance across tissue types.*

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, **Fig 3C)**. The distance from either breakpoint of an SV to its nearest gnomAD SV performed best, with similar AUCs of ~0.90 (**Supplementary Table 2**). 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.

Classifier Development and Performance

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 Germline and Tumor Structural Variant (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.

247 These probabilities allow for the selection of a classification cutoff to prioritize certain
248 performance metrics, including AUC, PPV, etc. For instance, having a cutoff farther from 0
249 would reduce the number of germline SVs falsely called somatic by the classifier (increasing
250 the PPV), while increasing the number of true somatic SVs falsely called germline (decreasing
251 the TPR). To balance these considerations, GaTSV uses a cutoff of 0.268 which maximizes
252 the sum of the PPV and TPR in our test set.
253
254 Overall, GaTSV achieved an AUC of 0.989, with a sensitivity (TPR) of 0.915, specificity (1 -
255 FPR) of 0.977 (**Fig. 3D-E**), and PPV of 0.849.
256
257 GaTSV's performance varied across tumor types, largely due to differences in the prevalence
258 of somatic SVs in each tumor type. The fraction of GaTSV-classified somatic SVs that were
259 truly somatic was higher in tumors with many somatic SVs. In other words, the PPV was higher
260 in tumors with more somatic SVs. In our test set, sarcomas had the highest somatic SV burdens
261 (~391 SVs per tumor) and the highest proportion of called-somatic SVs that were truly somatic
262 (PPV = 0.947). In contrast, acute myeloid leukemia, with the fewest SVs per tumor (~2 SVs),
263 had the lowest proportion of called-somatic SVs that were truly somatic (0.09) (**Fig. 3**). This
264 finding was corroborated at the sample level. Samples with a lower SV burden in the test set
265 tended to have lower PPV than samples with higher SV burden (**Supp Fig. S3B-C**).
266 Specifically, 86% of samples with fewer than 10 somatic SVs had PPV below 0.5. In contrast,
267 only 0.5% of samples with 10 or more SVs had a PPV below 0.5 (p-value < 2.2e-16, Fisher's
268 Exact Test). These data show that GaTSV performs well overall but is subject to falsely calling
269 true germline SVs as somatic SVs in samples with very low somatic SV counts.
270

271 GaTSV achieved an accuracy of 97% on average across all tumor types (**Supp Fig. S3D)**. To
272 uncover potential weaknesses of our classifier, we examined the features of the misclassified
273 SVs. The feature distribution of germline SVs mislabeled as somatic largely mimics that of true
274 somatic SVs in the test set. The patterns of correlations between the SV features also resemble
275 true somatic SVs more than true germline SVs (**Supp. Fig. S4**). This suggests that GaTSV
276 poorly differentiated these somatic-like germline events, leading to germline contamination in
277 the somatic calls. For example, 98% of misclassified germline SVs were more than 1kb away
278 from a gnomAD reference (**Supp. Fig. S4B**). Nonetheless, GaTSV correctly classified 91% of
279 germline SVs with a distance of 1kb or more from a gnomAD SV; therefore, GaTSV performs
280 well on most SVs.

281 GaTSV's performance in an Independent Dataset

282 To test how well GaTSV classifies SVs in tumors from other datasets, we gathered a test set
283 consisting of 7,623 SVs from six pediatric high-grade glioma (pHGG) patients with comparable
284 somatic SV burden to TCGA samples. The tumor samples used in this dataset were collected
285 with blood normals, which provided a truth set of somatic and germline SVs. The features
286 observed in the pHGG dataset had similar distributions to those of the tumors in the TCGA
287 dataset (**Supp**. **Fig. S5**). GaTSV achieved a sensitivity (TPR) of 0.975 and specificity (1-FPR)
288 of 0.892. Of the SVs that were called somatic in this set, 84% were truly somatic (PPV: 0.839).
289 We conclude that GaTSV performs robustly across different datasets.

290 GaTSV's performance across Different Ancestries

291 We hypothesized that imbalance in representation of individuals from certain ancestries in our
292 training set would impair GaTSV's performance on individuals of those ancestries - specifically,

293 over 77% of individuals in the TCGA dataset are of European descent. Indeed, GaTSV
294 performed better on SVs from European individuals than on SVs from East Asian and African
295 individuals on all metrics considered, including the AUC of the ROC and PR curves as well as
296 the PPV (**Supp Fig. S6A-F**). Among the ancestries considered, GaTSV performed the worst
297 on African SVs across all metrics. African genomes are considerably more diverse than any
298 other ancestry group, which means that increased representation in the training set is
299 necessary compared to other ancestries.
300
301 To evaluate whether this drop-off in performance was unique to GaTSV, we also tested the
302 gnomAD fuzzy matching method on different ancestries. Again, we saw the best performance
303 on European SVs and the worst on African SVs, across all metrics (**Supp Fig. S6G-L**).
304 European genomes are the most represented ancestry group in the gnomAD database at over
305 77%[20]. These results highlight the importance of ancestry biases in public databases. GaTSV's
306 performance on the worst-performing ancestry group—African—with a PPV of 0.66 was still
307 notably better than the gnomAD fuzzy matching performance with a PPV of 0.42 on the best-
308 performing ancestry group—European.

309 GaTSV uncovers associations between somatic SV signatures and 310 genetic variants in the absence of paired normal SV calls.

311 The effects of somatic SV-generating processes were recently characterized across TCGA
312 using SV signature analysis[3]. Somatic SV signature analysis has only been described after the
313 removal of all germline SVs from the call set through joint analysis with paired normal tissue.
314 We wanted to test if GaTSV allowed the accurate extraction of these biologically relevant
315 patterns of somatic SVs without a paired normal. Organizing the set of all SVs in the TCGA

316 dataset into the SV signature input catalog showed an abundance of short non-clustered
317 deletions (**Supp Fig. S7A**), the majority of which were germline SVs. Both the GaTSV-
318 classified germline SVs and the gnomAD fuzzy matching-classified germline SVs recaptured
319 this true germline SV deletion peak. However, gnomAD fuzzy matching misclassified many of
320 these short non-clustered deletions as somatic, resulting in a falsely inflated total number of
321 somatic SVs to more than double the true count of somatic SVs. In contrast, GaTSV correctly
322 matched both the distribution and counts of SVs seen in the true somatic and germline SV
323 signature input catalogs.
324
325 We extracted the published SV signatures[3] based on these catalogs and primarily detected
326 reference SV signature R3 in the germline profile and only minor contributions of the other
327 signatures. In contrast, somatic SVs were composed of a variety of signatures (**Fig. 4A**, **Supp**
328 **Fig. S8A-B**). This difference in the reference signature proportions across germline and
329 somatic SVs suggests that the mechanisms behind somatic SV generation are more varied
330 than those behind germline SV generation. It is also possible that there are additional germline
331 SV signatures that we did not detect because the reference signatures we used were extracted
332 from somatic SVs.
333

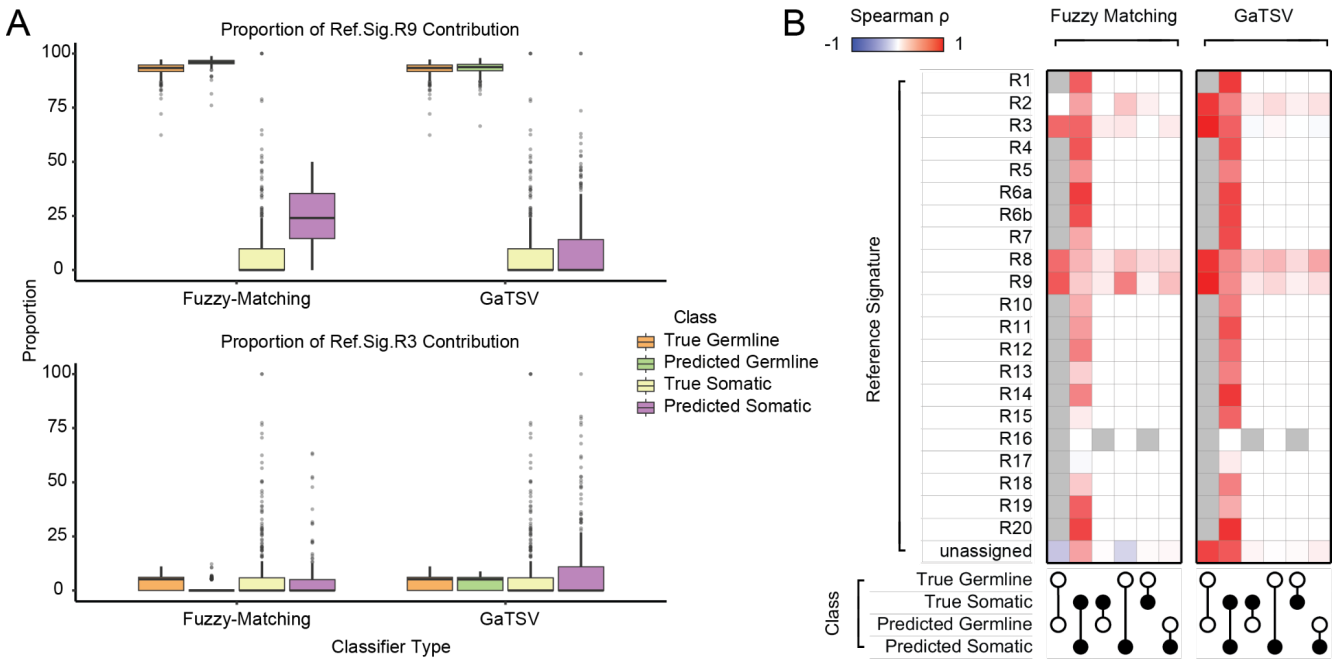


Figure 4. *Analysis of classifier performances using rearrangement signatures.* ***A****, R3 and R9 reference signature activity is more accurately recapitulated using GaTSV classified SVs than fuzzy-matching classified SVs. The proportion of contributions of the R9 reference signature (top) and R3 reference signature (bottom) in each patient. 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 to 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.*

In the comparison of reference SV signature proportions, the GaTSV also outperformed gnomAD fuzzy matching. The medians of gnomAD fuzzy matching derived “somatic” signature

350 contributions were closer to the true germline than the true somatic contributions for reference
351 signatures R3 (nonclustered translocations) and R9 (short inversions). In contrast, the GaTSV-
352 predicted "somatic" signature contribution medians matched the respective true class medians
353 (**Fig. 4A**). GaTSV also showed high levels of correlation for called somatic to true somatic and
354 called germline to true germline for almost all reference signatures (**Fig. 4B**). These results
355 further support our finding that the GaTSV overcomes the germline contamination issue in
356 somatic SV calls without a paired normal.
357

358 Discussion

359 Our analyses show that germline SVs are shorter, less likely to impact genes, and have more
360 bases of homology adjacent to their breakpoints than somatic SVs. In contrast, somatic SVs
361 are more likely to cluster together and are farther from transposable elements than germline
362 SVs. These results strengthen previous findings that homology and transposon-based repair
363 contribute more to the formation of germline SVs than somatic SVs. They also strengthen the
364 intuition that germline genome structure is under much more stringent fitness constraints than
365 somatic genome structure. We also establish that these differences can be used to
366 computationally classify germline from somatic SVs in the absence of a matched normal, with
367 high sensitivity and specificity.
368

369 Differences between DSB repair processes in somatic cells vs. germ cells and their progenitors
370 are reflected in differences between SVs in these different contexts. Germline SVs tend to have
371 more bases of homology than somatic SVs, revealing a preference for more accurate repair

372 processes to maintain genome integrity. Somatic SVs are closer to each other than germline
373 SVs, indicative of a higher likelihood of chromoanagenesis events in cancer cells.
374
375 Differences in fitness constraints in germline vs. somatic contexts are also reflected in their
376 SVs. Most common germline SVs are short unclustered deletions without impact on a coding
377 sequence. Previous work showed that even small repeats resulting from duplication have been
378 found to decrease cell fitness[21]. Therefore, most germline SVs lack gene dosage effects or
379 changes in protein composition. Ostensibly, such a rearrangement typically does not impact
380 transcriptional regulation. Similarly, most germline SVs tend to accumulate close to previously
381 described germline loci[17] and are less likely to form in clusters. We postulate that these
382 observations are the result of a selective process that eliminates de novo germline
383 rearrangements that impact gene dose or protein functionality. The particular underlying
384 process is worthy of further investigation and is plausibly attributed to embryonic lethality
385 leading to variant extinction. As cells harboring somatic SVs do not need to develop from a
386 single cell to an entire organism, they are likely subject to fewer fitness constraints than
387 germline SVs. As a result, they are more tolerable to impacts on coding sequences and
388 transcriptional dysregulation. We surmise that these SVs are under selection pressure that is
389 divergent from germline SVs.
390
391 Previous studies have shown that the tumor context is capable of altering the replication timing
392 landscape. Such transformation from early-to-late replication and vice versa are evidenced in
393 chromatin remodeling and methylation frequency[16]. The types of SVs were also linked to this
394 altered chromatin compaction density between germline and somatic cells. While this

395 consideration is not accounted for in our analysis, future studies can seek to uncover this
396 potential bidirectional relationship between the cancer context and replication timing.
397
398 We took advantage of the differences between germline and somatic SVs to create an SVM
399 classifier to distinguish them in cases where a matched normal sample is unavailable. Our SVM
400 performed well (AUC = 0.99, PPV = 0.85), so we termed it the great Germline and Tumor
401 Structural Variant classifier - the great GaTSV classifier. One significant strength of GaTSV is
402 its ability to classify SVs that are not present in germline reference databases - this is critical
403 as most SVs are rare and therefore not present in reference databases. Our tool has several
404 limitations. First, it did not perform as well on patients of African ancestry compared to patients
405 of European ancestry, likely due to the underrepresentation of individuals of African descent in
406 the TCGA training dataset we used. Second, our tool can only be used with SV calls from
407 SvABA[10] because it was trained on breakpoints called by SvABA, which may follow different
408 conventions from other SV callers.
409
410 Another potential limitation of our study is the presence of artifactual rearrangements in SV
411 calls. SvABA uses short-read sequencing data, potentially resulting in reads that each align to
412 multiple repetitive loci. These multi-mapping reads can result in artifactual SV calls present in
413 many germline and somatic samples. Since these SVs are present in both germline and
414 somatic samples, they will be called germline by traditional methods that compare somatic SVs
415 with a matched germline sample, such as our training set. Therefore, the GaTSV classifier will
416 classify these artifactual SVs as germline. Hence, while we expect that the GaTSV classifier
417 can identify somatic variants from the pool of germline and artifactual SVs, it is more difficult to
418 draw biological conclusions from the GaTSV germline calls, as they are contaminated with

419 artifactual SV calls. In future work, these issues can be addressed with long-read sequencing
420 data, diversifying the ancestries represented in the training set, and extending GaTSV to use
421 SV calls from other software such as MANTA and GRIDSS2[22,23].
422
423 Although our classifier performed well overall, it tended to misclassify certain variants, such as
424 its propensity to call somatic deletions shorter than 100kb as germline rearrangements. These
425 incorrect classifications may reflect true limitations in the great GaTSV and create a false
426 identity for SVs. However, our analysis showed that features of germline SVs misidentified as
427 somatic closely mimicked features of true somatic SVs and vice versa. While SV-generating
428 processes differ substantially between germline and somatic contexts, some processes such
429 as the integration of transposable elements and repeats into the genome are common to both[24].
430 This overlap could lead to similarities in the SV features, thus making them indistinguishable
431 to the classifier.
432
433 The great GaTSV classifier will facilitate several avenues of new research to understand the
434 formation and impact of germline and somatic SVs. Using GaTSV, SVs in clinical samples
435 lacking a matched normal can be interrogated. The power of deeply characterized groups of
436 cell lines such as those in the Cancer Cell Line Encyclopedia[9,25,26] can now be brought to bear
437 on SVs, including discovering relationships between drug sensitivities, CRISPR dependencies,
438 and gene expression on specific SVs, signatures of SVs, and SV abundance. The ability to
439 accurately distinguish germline from somatic SVs in the absence of a matched normal will
440 enable functional assessment of factors guiding SV formation and consequences for therapy
441 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.

Methods and Materials

Data Acquisition

Whole genome sequencing tumor samples with matched normals were obtained from the TCGA patient cohort (**Supplementary Table 3**)[27]. We used SvABA to call SVs on all samples as previously described, which were used for our training and test sets[10]. The pHGG test set consisted of six samples from our recent study[28]. We also ran SvABA on the pHGG samples, which called 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.

465

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

469

470 **gnomAD Filtering**

471

472 We obtained the gnomAD v4.0 release from the open-source gnomAD browser[29]. We filtered
473 for variants that passed all gnomAD filters and had resolved breakpoints. In addition, we
474 excluded all complex (CPX) SVs due to unclear breakpoint origins, as well as insertion (INS)
475 SVs that lacked a source sequence. For insertion SVs that reported a source sequence, we
476 resolved the breakends to reflect a corresponding intrachromosomal or interchromosomal
477 translocation.

478

479 **gnomAD Fuzzy Matching**

480

481 For each rearrangement in our test set, we first determined the distance to the closest gnomAD
482 reference SV. The distances of a given rearrangement breakpoint to its "corresponding"
483 breakpoint of all reference gnomAD SVs were calculated on each side (i.e. comparing the 5'
484 breakpoint of the rearrangement to the 5' breakpoint of a gnomAD reference SV and vice
485 versa). These distances were averaged, and the gnomAD SV with the lowest average was
486 labeled as the closest SV.

487

488 Each of these discrete values of average distances was treated as a cutoff, and we constructed
489 an ROC curve by classifying all rearrangements with an average distance less than that cutoff
490 as germline. For each of the cutoffs, the predicted and actual classifications were used to
491 calculate the specificity and sensitivity, which were used to generate the ROC curve.
492
493 **Feature Annotation**
494
495 Once the filtering step was complete, we annotated each SV with the following features:
496 distance to the closest gnomAD reference SV, DNA replication timing of each breakpoint, GC
497 content and length of any novel sequence insertion, GC content and length of homology
498 associated with a breakpoint, TP53 gene mutation status, number of other SVs within a 5Mbp
499 window, total number of SVs in the sample, distance to a long interspersed nuclear element
500 (LINE), distance to a short interspersed nuclear element (SINE), distance to the nearest SV,
501 span of the SV, type of the SV (categorized as deletion, duplication, inversion, or translocation),
502 impact on a gene, and impact on an exon.
503
504 The distance to the closest gnomAD reference SV for a given rearrangement was encoded as
505 two features, each representing the distance of a breakpoint to its “corresponding” breakpoint
506 of the nearest reference gnomAD SV, similar to the method used for gnomAD fuzzy matching.
507 If a matching gnomAD SV was not found, as is the case for many translocations, we input an
508 artificial distance of 1e9 for each breakpoint.
509
510 The DNA replication timing was likewise encoded as two features, corresponding to the
511 replication timing of each breakpoint. This was determined by looking up each breakpoint

512 location in the DNA replication timing table for the hg19 reference genome (from
513 http://mskilab.com/fishHook/hg19/RT_NHEK_Keratinocytes_Int92817591_hg19.rds).
514
515 The GC content and length of insertion and microhomology sequences were features derived
516 from the SvABA output. When assembling the contigs, SvABA frequently finds short gaps or
517 overlaps between regions that map to the reference genome. These gaps or overlaps were
518 output as “INSERTION” or “HOMSEQ” respectively. The gaps (i.e. the short region of the
519 sample genome that did not map to the reference genome) and overlaps were analyzed for
520 GC content and length and each of these were added as features.
521
522 The TP53 mutation status was determined using the consensus SNV, MNV, and indel calls
523 from TCGA samples available on the ICGC Data Portal[30,31]. These consensus calls were
524 filtered for the TP53 gene and functional mutations (i.e. variants classified as “3'UTR”, “5'UTR”,
525 “lincRNA”, “Intron”, “Silent”, “IGR”, “5’Flank”, “RNA” were removed). XX of our samples were
526 not analyzed in the PCAWG consensus calls. For these, we assigned the TP53 status as
527 indeterminate. Each sample was then assigned a value of 1, 0, or -1, which corresponded to
528 TP53 mutant, indeterminate, or TP53 wild-type.
529
530 The number of SVs in a 5Mbp window and the total number of SVs in the sample were
531 calculated by counting the number of rearrangements that were within 5Mbp of the SV and
532 within the specific sample respectively. The distances to the nearest SINE and LINE events
533 were calculated by taking the distance from a given rearrangement to the closest SINE and
534 LINE events[32]. The distance to the nearest SV was calculated by iterating through other

535 rearrangements in the same sample, determining the distance between the given
536 rearrangement and the other SV, and selecting the lowest value.
537
538 The span of the SV was taken from the SvABA output. The SV-type feature was separated into
539 four binary features, each corresponding to insertion, duplication, inversion, or translocation
540 events based on the position and read orientation of the supporting reads
541
542 The impact on a gene or exon region were binary factors representing whether a given
543 rearrangement overlapped a gene or exon region, which was downloaded from Gencode and
544 Ensembl genome browser BioMart, respectively. If the SV breakpoint was located within one
545 of these regions, it was labeled with a 1. If no overlap was found, the feature was labeled with
546 a 0 (also see **Supp. Fig. S9A**).
547
548 Moreover, we log-transformed any feature that related to genomic distances or counts of SVs
549 to reduce noise. These features included the SV span, homology length, insertion length,
550 distance to the closest gnomAD reference SV, distance to the nearest SINE and LINE element,
551 the number of SVs in a given sample, the distance to the nearest SV, and the number of SVs
552 within 5Mbp.
553
554 Once all the features were added, each feature was then scaled. The scaling was done by
555 creating a scaling matrix containing values representing the mean and standard deviation of
556 each feature (**Supplementary Table 4**). These values were calculated by taking 10 random
557 samples of 50,000 SVs from the training dataset and evaluating the mean and standard
558 deviation of the features for all of these rearrangements. Each input rearrangement feature—

559 including those in both training and test sets—was scaled by subtracting the mean and dividing
560 by the standard deviation in this matrix.
561
562 **Statistical Analyses**
563
564 We performed Kolmogorov-Smirnov tests to determine if the distributions of continuous
565 features between germline and somatic SVs significantly differed. Our p-value significance
566 threshold was set to 0.05. For features quantified in discrete variables, we used Chi-squared
567 tests with a significance threshold of 0.05.
568
569 To assess the significance of associations between variables both within SV classes and
570 across classes, we performed a variety of statistical tests. For continuous/continuous feature
571 comparisons within germline and somatic SVs, we computed Spearman ρ and defined
572 moderate to high correlation as $|\rho| \geq 0.5$. We then obtained p-values for the Spearman ρ values
573 using two-sided t-tests as described previously[33]. Next, we performed false discovery rate
574 (FDR) multiple hypothesis correction with a q-value threshold of 0.05. To check whether these
575 correlations significantly differed across germline and somatic SVs, we derived z-scores from
576 the ρ values using Fisher's z-transformation [34] and subsequently obtained p-values. We
577 performed FDR correction on these p-values and defined significance at a q-value threshold of
578 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[35]. 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.

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

603 compute prediction probabilities on 100k scaled SVs sampled from the full test set. Prediction
604 probabilities were derived using the R stats predict method. Finally, we calculated the AUC,
605 TPR, and FPR values to evaluate the performance of each model.
606
607 **SVM**
608
609 We used an SVM with an RBF kernel from the e1071 package in R and set cost and gamma
610 parameters as 10 and 0.1 respectively. These hyperparameters were determined by a grid-
611 search, of all combinations of the following values: cost of 1, 10, 100, 500, 1000 for both RBF
612 and linear kernels and gamma of 0.0001, 0.001, 0.01, 0.05 0.1, 1 for the RBF kernel. For each
613 combination of hyperparameters, we performed a 5-fold cross-validation, splitting the training
614 set into five groups. One group was selected as a validation set for each of the five cross-
615 validation iterations, and the other four groups made up the sub-training set. During each
616 iteration, the model was trained on the sub-training set and evaluated on the validation set. The
617 resulting AUC and PPV for all iterations were averaged for each combination of
618 hyperparameters, and the model that performed well for both metrics was chosen.
619
620 Although SVMs typically do not have an inherent probability metric, there is a probability feature
621 built into e1071 implementation of the SVM—based on Platt Scaling. We used this to generate
622 the ROC curve for the AUC above and to select a probability cutoff that optimizes for our
623 specific needs, instead of just accuracy. Because we aimed to reduce germline contamination
624 for analysis of somatic variants, we decided on a cutoff that resulted in a high PPV.
625

626 To achieve this, we treated the probability cutoff as another hyperparameter and performed
627 another 5-fold cross-validation for each probability value between 0 and 1 in 0.001 increments.
628 This time, we optimized for the maximum TPR + PPV value, since we found that PPV itself was
629 monotonically increasing with respect to the probability cutoff. We did not change the cost,
630 gamma, and kernel type parameters in this probability tuning step, and were kept as 10, 0.1,
631 and RBF kernel respectively.
632
633 In constructing our classifier, we chose a probability cutoff that optimized for the sum of TPR
634 and PPV, to minimize germline contamination from rearrangements called somatic. However,
635 the probability cutoff chosen for our classifier can be modified for other use cases, including
636 those that demand correct germline calls or a high overall accuracy, in which case a cutoff of
637 the maximum sum of the TNR and NPV or the maximum AUC in the same TCGA validation
638 set can be used.
639
640 **Ancestry Analysis**
641
642 Using consensus ancestry calls of TCGA patients from a previous study, we conducted a
643 lookup of patients in our cohort to those in the consensus calls[36,37]. After excluding the 25
644 patients who were not in the consensus calls, the test set of 277,925 TCGA SVs—each with
645 an associated patient—were labeled with their corresponding ancestry call. For our analysis,
646 we did not consider any of the 50 admixed patients. We also did not include the “amr” and
647 “sas”, corresponding to American and South Asian, ancestries, as there were only seven and
648 two patients within those categories, respectively. For the remaining “afr”, eas”, and “eur”,
649 corresponding to African, East Asian, and European, ancestries, we subsetted the test set

650 according to their ancestry labels, and analyzed the performance of the classifier on each
651 subset.
652
653 **Signature Analysis**
654
655 The signature analysis was performed primarily using the signature.tools.lib package published
656 by the Nik-Zainal group[3]. From bedpe files containing SVs for each patient in our cohort, we
657 first created true germline and somatic signature catalogs for each patient using the
658 `bedpeToRearrCatalogue` function in the signature.tools.lib package. These signature
659 catalogs consisted of counts of the number of rearrangements organized across three
660 categories: SV span, type, and clustering of all SVs of span above 1kb. This process was
661 repeated for predicted germline and somatic rearrangements from both the fuzzy matching
662 approach and the SVM classifier.
663
664 These catalogs were then fit to the reference signature profile for rearrangements from the
665 same study using the `Fit` and `plotFit` functions[3,38]. This resulted in the true and predicted
666 germline and somatic reference signature exposures for each patient, which shows how much
667 contribution each reference signature has in the collection of SVs in each patient. For Fig 5A
668 and Supplemental Fig S8, we determined the ratio of a given reference signature exposure to
669 the sum of all exposures within a patient for all reference signatures. In the case that there
670 were no SVs in a patient—for a given class—the proportion was simply set as 0, instead of an
671 undefined value.
672

In Fig 5B, we first conducted a modified Spearman correlation between each predicted and true class. Each vector used for the Spearman correlation had a length of 974 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 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.

Availability of Data and Materials

All annotation functions, scripts used to create our figures, and trained classifier will be available on Github[39]. The datasets used to support the conclusions of this article are available online[27,29,31,32,37,38]. Any data generated during the current study are available from the corresponding author on reasonable request.

Acknowledgments

We thank and acknowledge the following funding sources: Fund for Innovation in Cancer Informatics (S.D. and R.B.), the Gray Matters Brain Cancer Foundation (R.B.), the Pediatric Brain Tumor Foundation (R.B.), and Break Through Cancer (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. is supported by an NIH NRSA award (F32), 1F32CA261024.

Competing 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. The remaining authors declare no competing interests.

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. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
5. Pócza, T. *et al.* Germline Structural Variations in Cancer Predisposition Genes. *Front. Genet.* **12**, 634217 (2021).
6. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
7. Dubois, F., Sidiropoulos, N., Weischenfeldt, J. & Beroukhim, R. Structural variations in cancer and the 3D genome. *Nat. Rev. Cancer* (2022) doi:10.1038/s41568-022-00488-9.
8. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **7**, 818–831 (2017).
9. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
10. Wala, J. A. *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
11. Vassetzky, N. S. & Kramerov, D. A. SINEBase: a database and tool for SINE analysis.

725 *Nucleic Acids Res.* **41**, D83–9 (2013).
726 12. Alisch, R. S., Garcia-Perez, J. L., Muotri, A. R., Gage, F. H. & Moran, J. V.
727 Unconventional translation of mammalian LINE-1 retrotransposons. *Genes Dev.* **20**,
728 210–224 (2006).
729 13. Kim, E. Z., Wespiser, A. R. & Caffrey, D. R. The domain structure and distribution of Alu
730 elements in long noncoding RNAs and mRNAs. *RNA* **22**, 254–264 (2016).
731 14. Balachandran, P. *et al.* Transposable element-mediated rearrangements are prevalent in
732 human genomes. *Nat. Commun.* **13**, 7115 (2022).
733 15. Holland, A. J. & Cleveland, D. W. Chromoanagenesis and cancer: mechanisms and
734 consequences of localized, complex chromosomal rearrangements. *Nat. Med.* **18**, 1630–
735 1638 (2012).
736 16. Du, Q. *et al.* Replication timing and epigenome remodelling are associated with the
737 nature of chromosomal rearrangements in cancer. *Nat. Commun.* **10**, 416 (2019).
738 17. Collins, R. L. *et al.* A structural variation reference for medical and population genetics.
739 *Nature* **581**, 444–451 (2020).
740 18. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* 861054
741 (2019) doi:10.1101/861054.
742 19. Adelson, R. P. *et al.* Empirical design of a variant quality control pipeline for whole
743 genome sequencing data using replicate discordance. *Sci. Rep.* **9**, 16156 (2019).
744 20. Karczewski, K. J. *et al.* Author Correction: The mutational constraint spectrum quantified
745 from variation in 141,456 humans. *Nature* **590**, E53 (2021).
746 21. Adler, M., Anjum, M., Berg, O. G., Andersson, D. I. & Sandegren, L. High fitness costs
747 and instability of gene duplications reduce rates of evolution of new genes by
748 duplication-divergence mechanisms. *Mol. Biol. Evol.* **31**, 1526–1535 (2014).

749 22. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and
750 cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
751 23. Cameron, D. L. *et al.* GRIDSS2: comprehensive characterisation of somatic structural
752 variation using single breakend variants and structural variant phasing. *Genome Biol.* **22**,
753 202 (2021).
754 24. Zamudio, N. & Bourc'his, D. Transposable elements in the mammalian germline: a
755 comfortable niche or a deadly trap? *Heredity* **105**, 92–104 (2010).
756 25. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia.
757 *Nature* **569**, 503–508 (2019).
758 26. Corsello, S. M. *et al.* Discovering the anti-cancer potential of non-oncology drugs by
759 systematic viability profiling. *Nat Cancer* **1**, 235–248 (2020).
760 27. TCGA Primary Sequence Data. https://gdc.cancer.gov. Accessed 1 November 2021.
761 28. Dubois, F. P. B. *et al.* Structural variants shape driver combinations and outcomes in
762 pediatric high-grade glioma. *Nat Cancer* (2022) doi:10.1038/s43018-022-00403-z.
763 29. gnomAD Structural Variant Data. https://gnomad.broadinstitute.org/downloads.
764 Accessed 28 March 2024.
765 30. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis
766 of whole genomes. *Nature* **578**, 82–93 (2020).
767 31. ICGC Data Portal SNV and Indel Calls.
768 https://dcc.icgc.org/releases/PCAWG/consensus_snv_indel. Accessed 12 September
769 2023.
770 32. RepeatMasker Repeat Elements Data. https://www.repeatmasker.org/species/hg.html.
771 Accessed 28 March 2024.
772 33. Howell, D. C. Statistical methods for psychology (8th, International ed.). *Boston, MA:*

773 *Cengage Learning*.

774 34. Fisher, R. A. & Others. 014: On the‘ Probable Error’ of a Coefficient of Correlation
775 Deduced from a Small Sample. (1921).

776 35. Breslow, N. E. & Day, N. E. *Statistical Methods in Cancer Research. Volume I - The*
777 *Analysis of Case-Control Studies*. (IARC Scientific Publications, 1980).

778 36. Carrot-Zhang, J. *et al.* Comprehensive Analysis of Genetic Ancestry and Its Molecular
779 Correlates in Cancer. *Cancer Cell* **37**, 639–654.e6 (2020).

780 37. TCGA Consensus Ancestry Calls. https://ars.els-cdn.com/content/image/1-s2.0-
781 S1535610820302117-mmc2.xlsx. Accessed 27 October 2023.

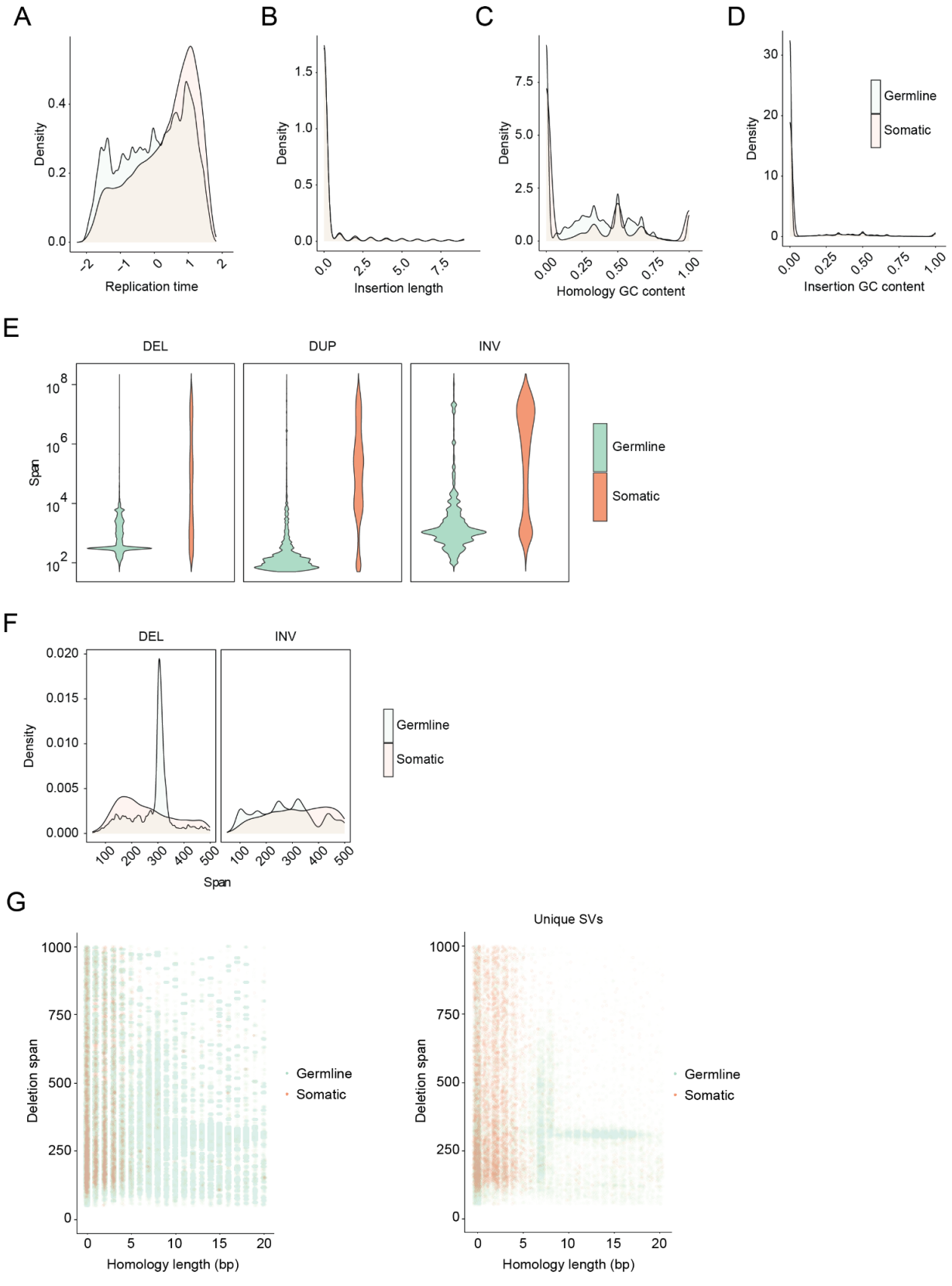
782 38. Degasperi et al. Reference Signatures. https://static-
783 content.springer.com/esm/art%3A10.1038%2Fs43018-020-0027-
784 5/MediaObjects/43018_2020_27_MOESM3_ESM.xlsx. Accessed 22 March 2023.

785 39. GaTSV Github Repository. https://github.com/beroukhim-lab/GaTSV. Accessed 6 June
786 2024.

787

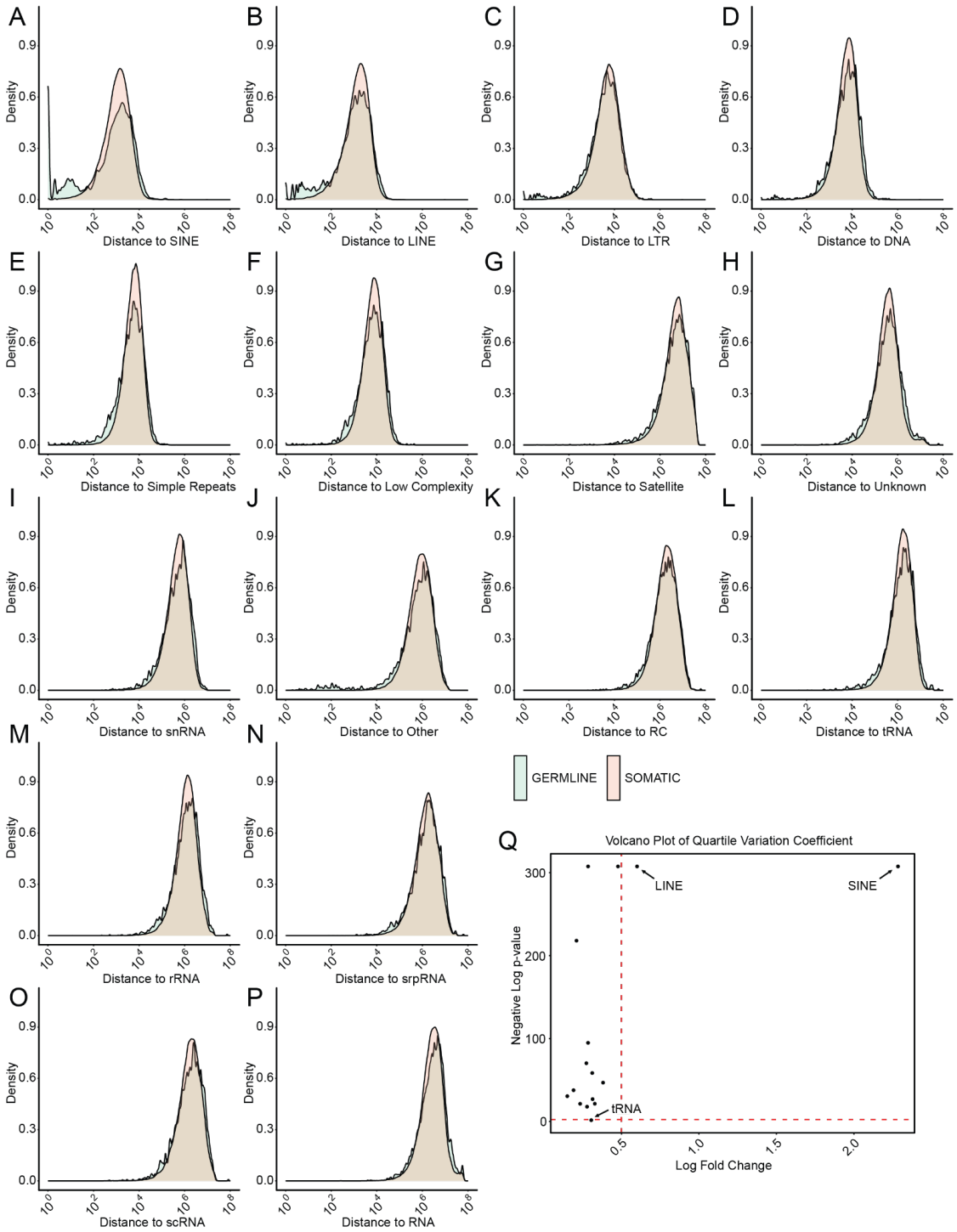
Supplementary Figures and Figure Legends

Supplementary Figure 1


A
Density
Replication time
B
Density
Insertion length
C
Density
Homology GC content
D
Density
Insertion GC content
Germline
Somatic
E
DEL
DUP
INV
Span
Germline
Somatic
F
DEL
INV
Density
Span
Germline
Somatic
G
Unique SVs
Deletion span
Homology length (bp)
Germline
Somatic

792 **Supplementary Figure 1.** Germline and somatic distributions for additional features
793 considered. **A,** Replication timing distributions for SV breakpoints. **B,** Length distribution of SV
794 insertion sequence. **C,** GC content of homology sequence. **D,** GC content of insertion
795 sequence. **E,** Violin plots of span distributions of germline and somatic deletions (DEL),
796 duplications (DUP), and inversions (INV). **F,** Density plots of span distributions of deletions and
797 inversions truncated at 500bp. **G,** Span of deletions vs homology lengths across the entire
798 dataset (left) and downsampled to represent unique SVs (right) showing a peak in span
799 distribution of germline deletions that coincides with homology lengths between 13-17 bp.
800

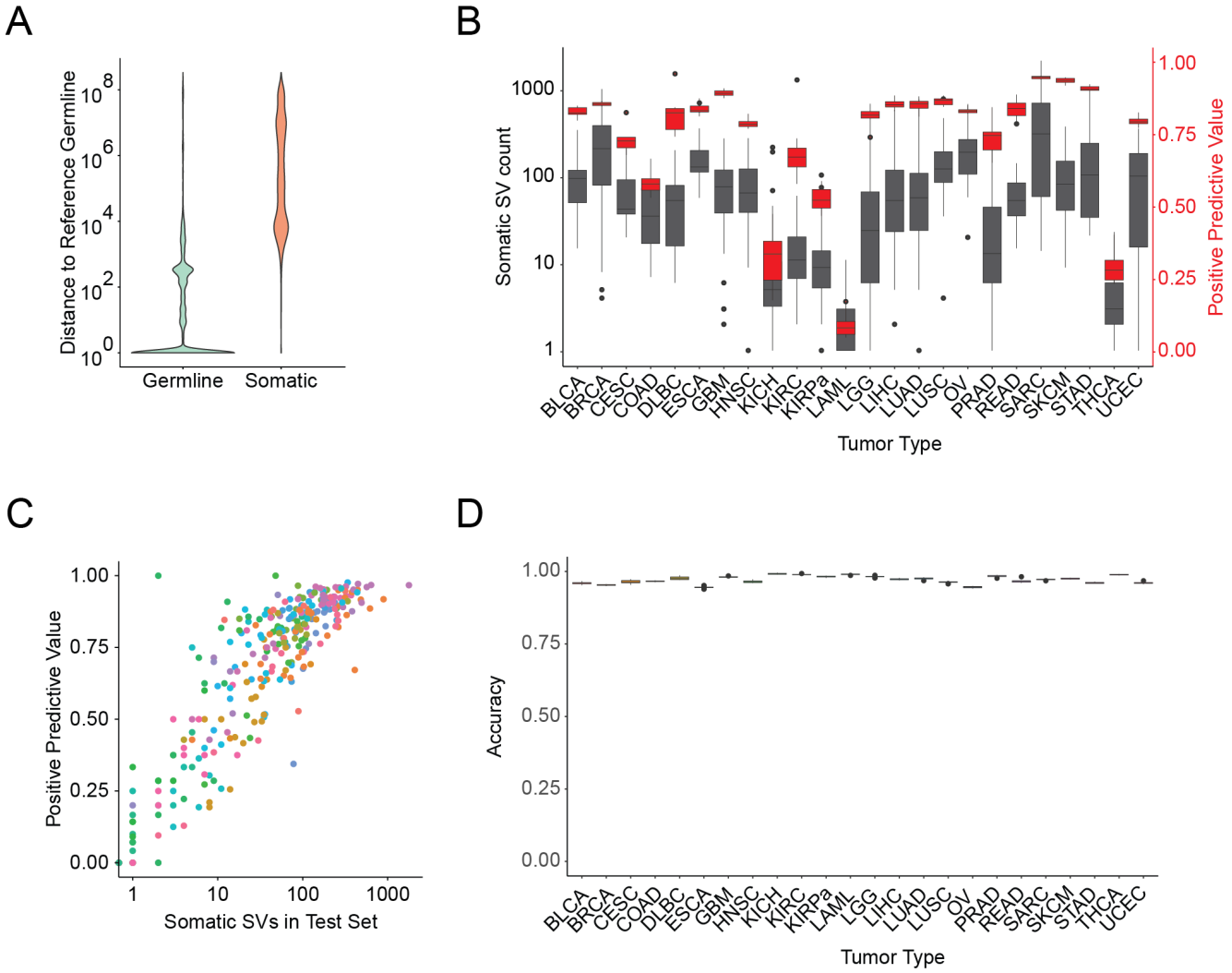
Supplementary Figure 2



806 **Supplementary Figure 2.**

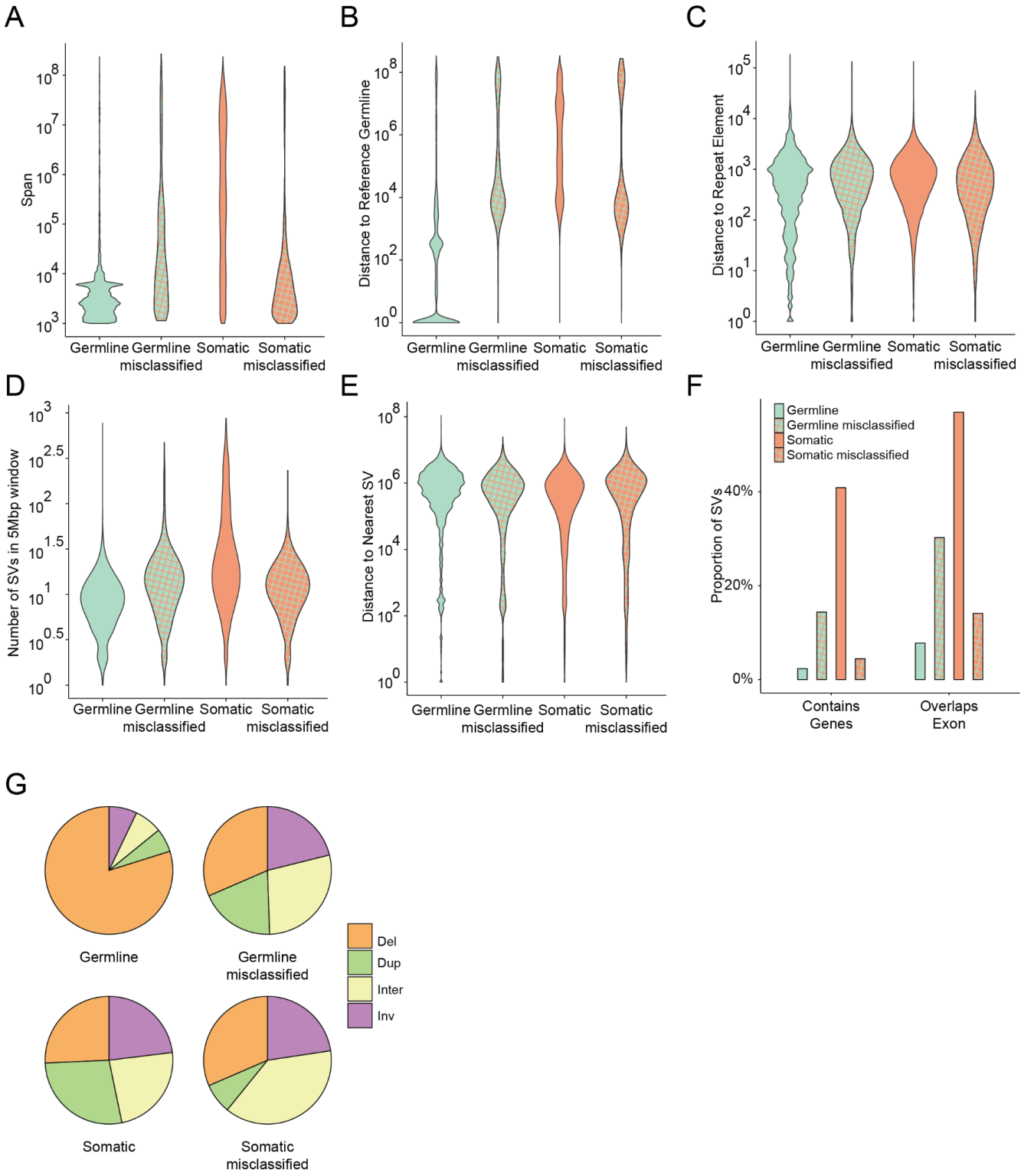
807 Distances to various repeat elements. **A-P,** Distributions of the distances of germline and
808 somatic SVs to various repeat elements ordered by instances of each element in the genome.
809 These include SINE (A), LINE (B), long terminal repeat (C), DNA (D), simple repeat (E), low
810 complexity (F), satellite (G), unknown (H), snRNA (I), other (J), rolling circle (K). Distances to
811 tRNA (L), rRNA (M), srpRNA (N), scRNA (O), RNA (P) pseudogenes were also calculated. **Q,**
812 For each repeat element, a Wilcoxon test was performed, and the BH-corrected p-values were
813 plotted in a volcano plot against the difference in quartile variation coefficient of the
814 distributions. Only SINE and LINE elements passed both cutoffs.

Supplementary Figure 3



Supplementary Figure 3. GaTSV classifier performance is related to somatic SV count. **A,** Distribution of total distance of germline and somatic SVs to the closest SV in gnomAD (reference) SV database. **B**, Somatic SV count, and PPV for each tumor type in the TCGA dataset. **C**, PPV vs. somatic SV count for every tumor in the TCGA dataset. **D,** Accuracy of GaTSV classifier across each tumor type

824 **Supplementary Figure 4**

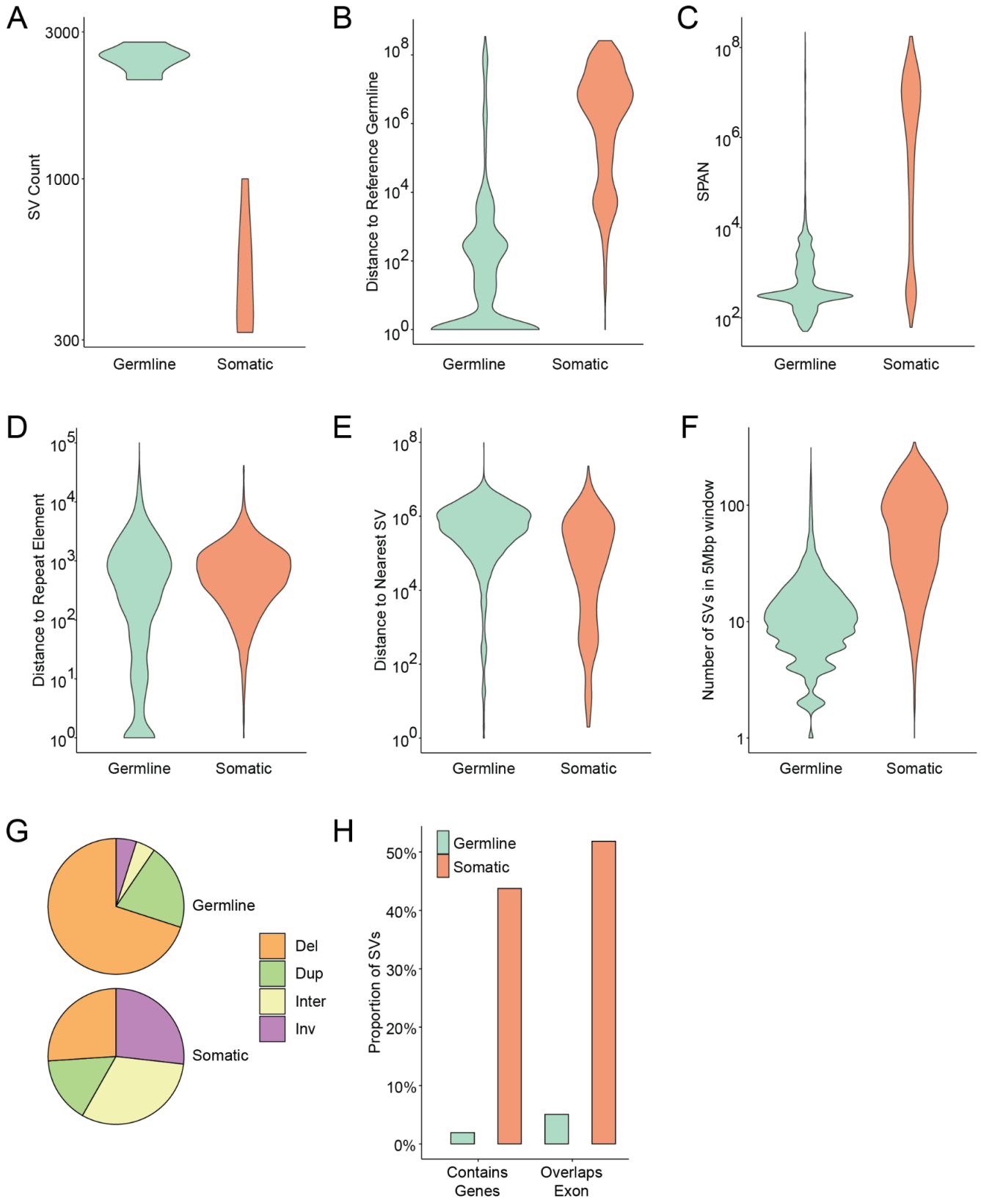


825

826

827 **Supplementary Figure 4.** Misclassified SVs resemble the other class. Span distribution (**A**),
828 distance to the closest SV in gnomAD (reference) database (**B**), distance to nearest repeat
829 element (LINE or SINE) (**C**), number of SVs within a 5Mbp window (**D**), distance to nearest SV
830 within a sample (**E**), proportion of SVs that impact a gene or overlap an exon (**F**), and proportion
831 of each SV type - Del (deletion), Dup (duplication), Inter (interchromosomal/ translocation), Inv
832 (inversion) (**G**) for correctly and incorrectly classified SVs.
833
834

835 **Supplementary Figure 5**



836

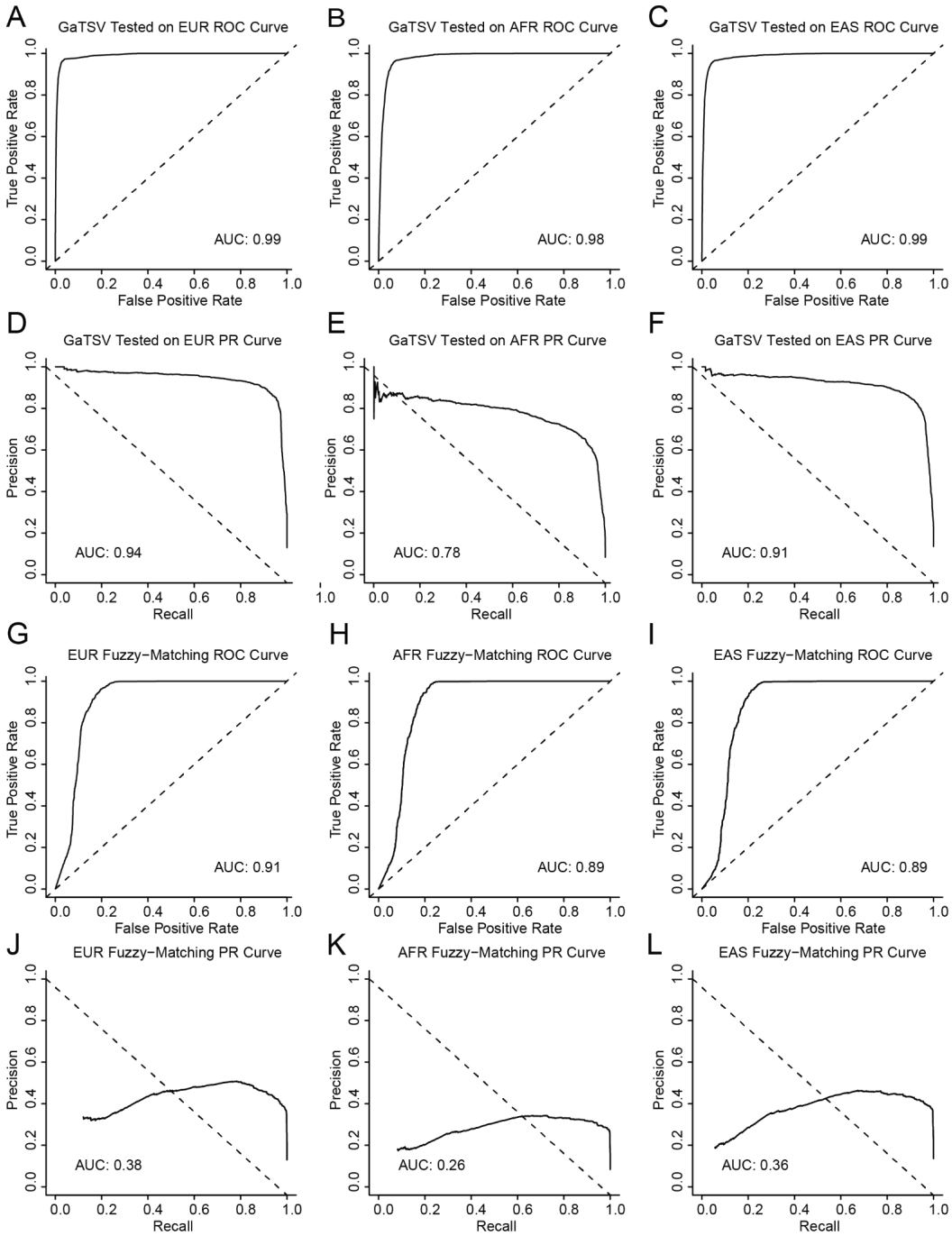
837

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

847

848

849 **Supplementary Figure 6**

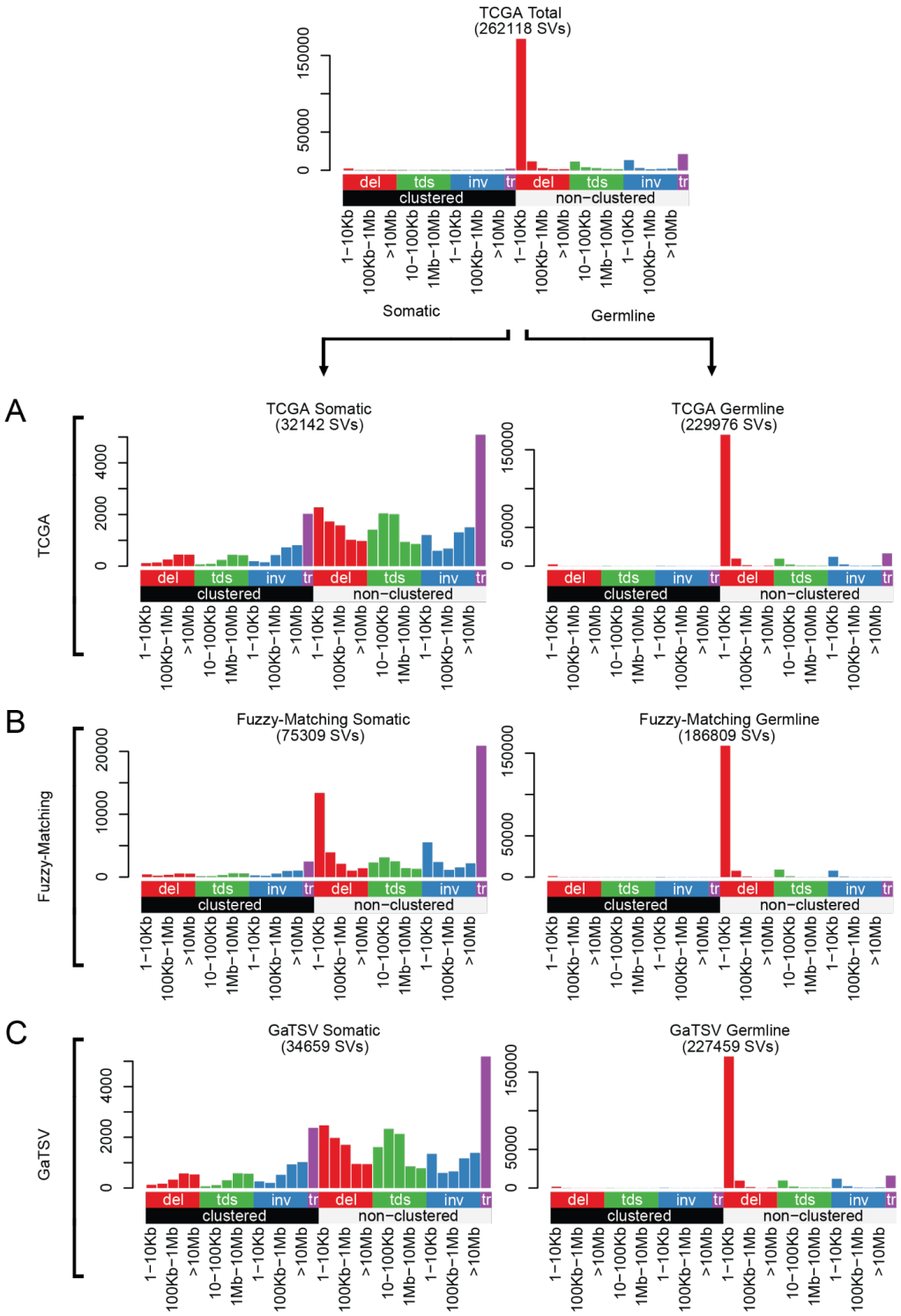


850

851 **Supplementary Figure 6.** Classifier performances on ancestry-specific subsets of the TCGA
852 test set. **A-C**, ROC for the GaTSV classifier on individuals descending from European (EUR)
853 (A), African (AFR) (B), and East Asian (EAS) (C) ancestry. **D-F**, PR curve for the GaTSV
854 classifier on individuals descending from European (D), African (E), and East Asian (F)

855 ancestry. **G-I**, ROC for the fuzzy-matching method to the gnomAD database on individuals
856 descending from European (G), African (H), and East Asian (I) ancestry. **J-L**, PR curve for the
857 fuzzy-matching method on individuals descending from European(J), African(K), and East
858 Asian(L) ancestry.

859 **Supplementary Figure 7**



860

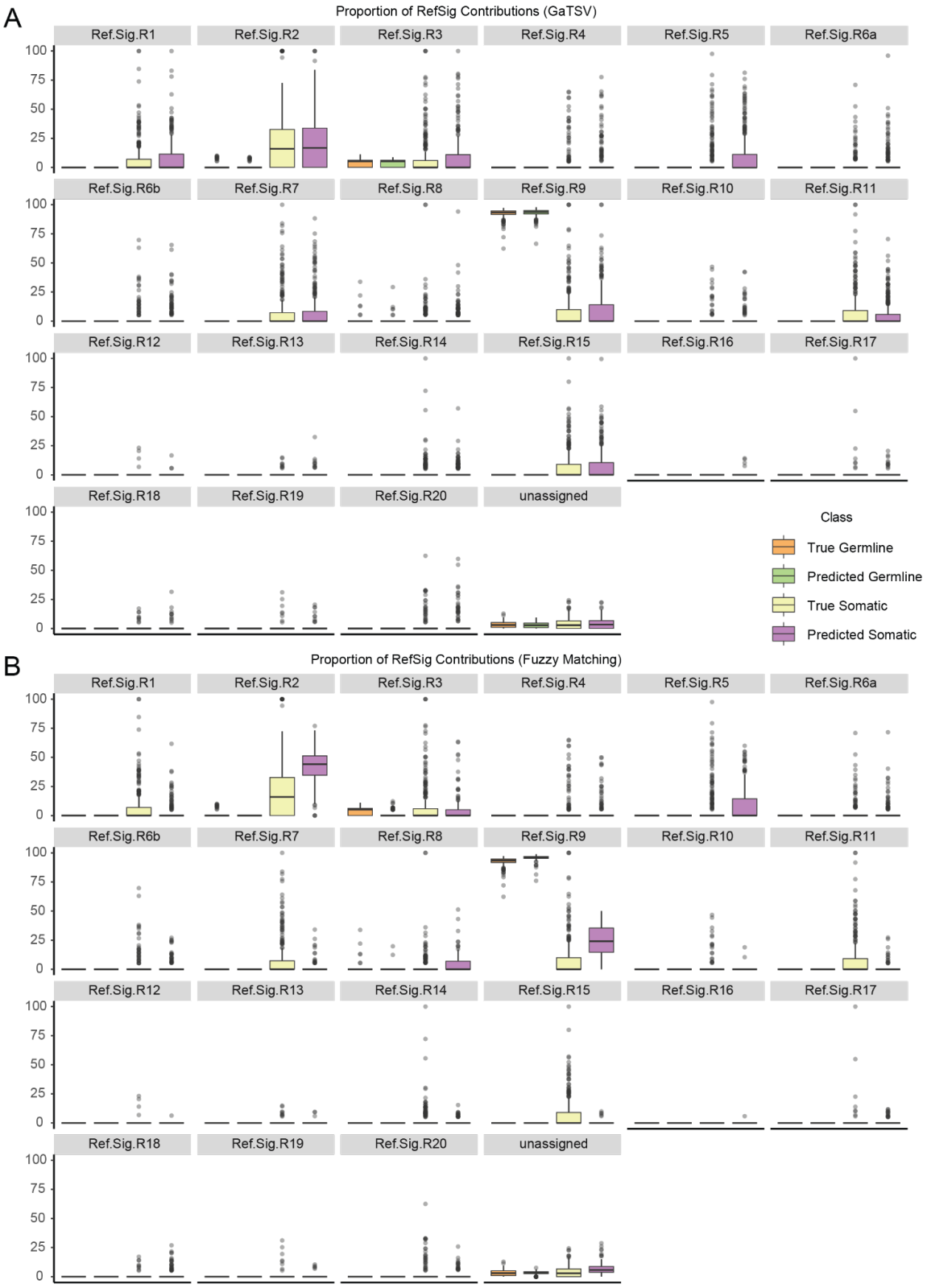
861 **Supplementary Figure 7.** Distribution of SV categories (SV-Signature input catalogs) for true

862 and predicted germline and somatic SVs. **A-C**, SV catalogs for the GaTSV classified SVs (C)

863 match those of the true TCGA SVs (A) more closely than the fuzzy-matching to the gnomAD

864 database (B).

865 **Supplementary Figure 8**

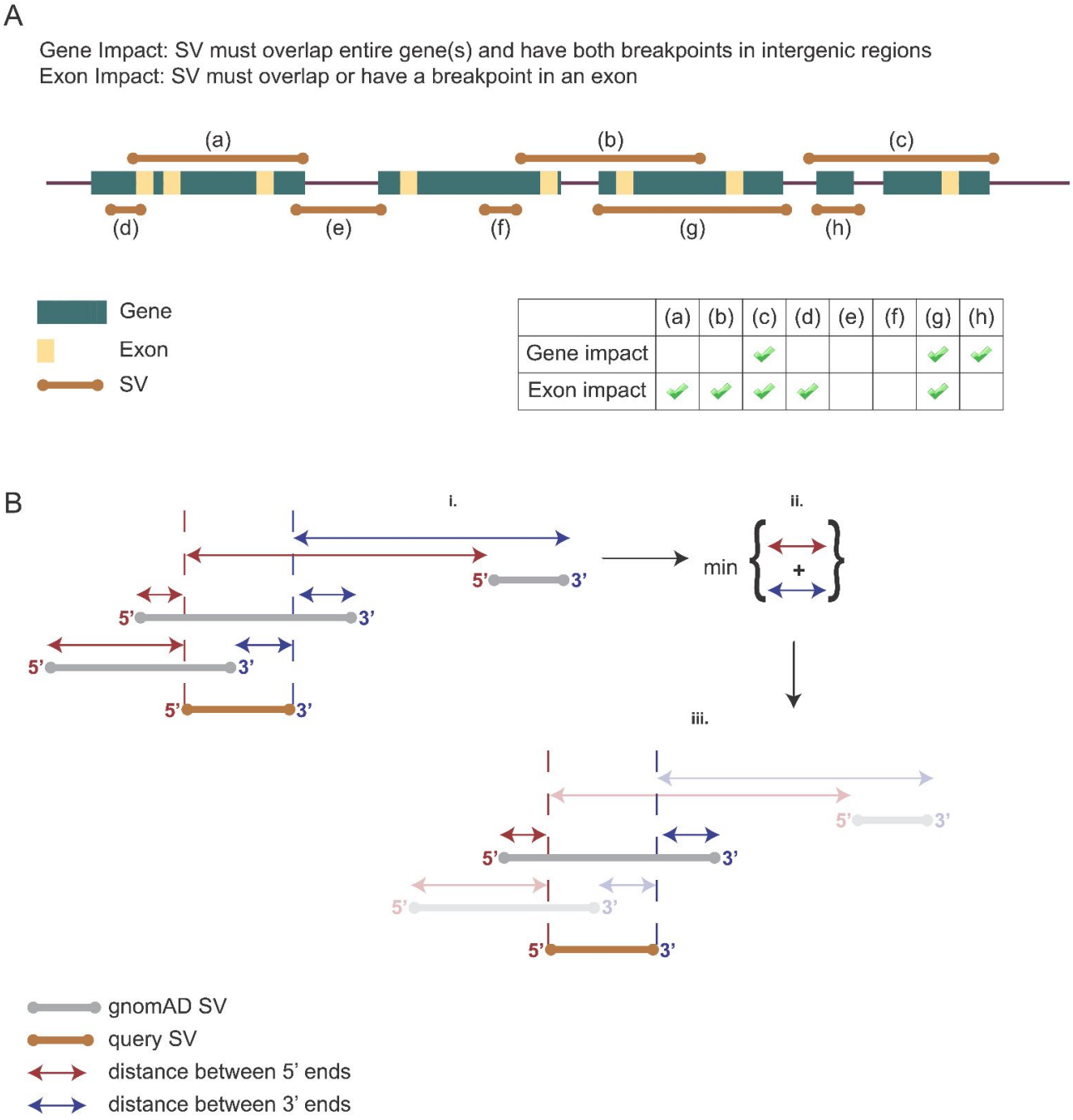


866

867 **Supplementary Figure 8.** The proportion of individual SV signature contributions in each

868 patient. **A**, post GaTSV classification. **B**, post fuzzy-matching to the gnomAD database.

869 **Supplementary Figure 9**



870

871 **Supplementary Figure 9.** Schematics of how gene/exon impact was determined for non-

872 translocation events (**A**) and how distance to nearest gnomAD SV was calculated (**B**).

873

874

875 # Supplementary Table Legends

876

877 **Supplementary Table 1.** Features of the SVM.

878

879

880 **Supplementary Table 2.** Performance of logistic regression models using single SV features.

881

882

883 **Supplementary Table 3.** TCGA cohort patient metadata.

884

885

886 **Supplementary Table 4.** Scaling matrix for SVM features.