



Published in final edited form as:

*Nat Methods*. 2022 April ; 19(4): 445–448. doi:10.1038/s41592-022-01423-4.

## Searching Thousands of Genomes to Classify Somatic and Novel Structural Variants using STIX

Murad Chowdhury<sup>1</sup>, Brent S. Pedersen<sup>2</sup>, Fritz J. Sedlazeck<sup>3</sup>, Aaron R. Quinlan<sup>4,5,6</sup>, Ryan M. Layer<sup>1,7,\*</sup>

<sup>1</sup>BioFrontiers Institute, University of Colorado, Boulder

<sup>2</sup>University Medical Center, Utrecht University, The Netherlands

<sup>3</sup>Human Genome Sequencing Center, Baylor College of Medicine, Houston

<sup>4</sup>Department of Human Genetics, University of Utah, Salt Lake City

<sup>5</sup>Department of Biomedical Informatics, University of Utah, Salt Lake City

<sup>6</sup>Utah Center for Genetic Discovery, University of Utah, Salt Lake City

<sup>7</sup>Department of Computer Science, University of Colorado, Boulder

### Abstract

Structural variants (SVs) are associated with cancers and developmental disorders, but challenges with estimating population frequency remain a barrier to prioritizing mutations over inherited variants. In particular, variability in variant calling heuristics and filtering limits the utility of current SV catalogs. We present STIX, a method that, instead of relying on variant calls, indexes and searches the raw alignments from thousands of samples to enable more comprehensive allele frequency estimation.

### Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

---

Structural variants (SVs), including large deletions, duplications, insertions, inversions, and translocations<sup>1</sup>, are associated with cancer progression and Mendelian disorders<sup>2–5</sup>. Copy number variants and gene fusions have received the most attention, but recent large-scale

---

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: <https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms>

\*Correspondence: [ryan.layer@colorado.edu](mailto:ryan.layer@colorado.edu).

Author Contributions Statement

Author contribution: MC developed software and designed, executed, and analyzed experiments. BSP conceived the study and developed software. ARQ conceived the study. RML conceived the study, developed software, and designed, executed, and analyzed experiments. All authors wrote and edited the manuscript.

Competing Interests Statement

The authors declare no competing interests.

Code Availability

STIX source code can be found at <https://github.com/ryanlayer/stix>. Excord source code can be found at <https://github.com/brentp/excord>. Source code for data analysis can be found at [https://github.com/ryanlayer/stix\\_paper](https://github.com/ryanlayer/stix_paper)

SV studies such as the Pan-Cancer Analysis of Whole Genomes<sup>6</sup> (PCAWG), the 1000 Genome Project<sup>7</sup> (1KG), gnomAD SV<sup>8</sup>, and the Centers for Common Disease Genomics<sup>9</sup> (CCDG) have expanded our understanding of the depth and diversity of somatic SVs in cancer and polymorphic SVs in humans. Despite the importance of SVs, barriers remain to their adoption in disease analysis<sup>1</sup>. In particular, limitations to short-read SV calling, reference biases, and variability in the heuristics and filtering strategies between cohorts lead to an incomplete understanding of SV population frequency that limits our ability to assess a variant's severity and impact<sup>10</sup>.

In cancer studies, SV interpretation requires classifying variants as germline or somatic. The standard is to call variants in the tumor and control tissue from the same individual. SVs found only in the tumor are deemed somatic. This method is susceptible to the sensitivity of the normal sample calls, which are often sequenced at lower coverage. When an inherited SV is missed in the normal tissue, it can be incorrectly classified as somatic. An alternative strategy is to substitute matched normal tissue with a panel of unrelated normal samples (e.g., 1KG, Simons Diversity Panel<sup>11</sup> (SGDP)), but the time and computational costs associated with joint-calling large numbers of samples can be prohibitively high.

SV catalogs from large DNA sequencing projects can filter tumor-only calls as a shortcut to joint calling. Variants found in both the tumor and reference catalog can be classified as inherited since we can reasonably assume that somatic variants, and driver mutations in particular, are likely to be rare and unlikely to share SV breakpoints with polymorphic SVs. While this assumption does not hold in all cases, it is the standard for many diseases studies. The analysis is more complicated for variants found only in the tumor calls. In principle, SVs that are not in the cohort are rare and could be somatic. In practice, several SV-specific factors, including short-read calling limitations<sup>12</sup>, genotyping complexities (see Supplementary Note 1, Supplementary Fig. 1), and aggressive filtering for false positive calls, exclude many real SVs from appearing in these catalogs. For example, among the thousands of cancer-related SVs that are recoverable in 1KG, an order of magnitude fewer are present in the 1KG SV call set<sup>7</sup>. Given these issues, it is impossible to determine whether an SV observed in a patient and not in a reference cohort is absent from the population (true negative) or removed in the filtering step (false negative).

Similarly, in Mendelian disease analyses, causal variants should be either absent or are at very low-frequency in the reference population<sup>13</sup>. Using allele frequencies from gnomAD<sup>14</sup>, a catalog of single nucleotide variants (SNVs) from 141,546 human genomes, can reduce the number of variants under diagnostic consideration by two orders of magnitude<sup>13</sup>. Unfortunately, no equivalent resource exists for SVs since, as with the cancer analysis, static call sets from large populations are inadequate. Pangenomes can help by identifying and genotype SVs<sup>15</sup>, but given the limited number of samples and SVs they can currently represent, they are better suited for common variants and are less useful for somatic and pathogenic variant classification.

To ensure comprehensive and accurate SV detection and allele frequency assignment, we propose searching the raw alignments across thousands of samples using our structural variant index (STIX). For a given deletion, duplication, inversion, or translocation, STIX

reports a per-sample count of every alignment that supports the variant (Fig. 1). Assuming deleterious variants are rare, from these counts, we can conclude that an SV with evidence in many healthy samples is either a common germline variant or the product of systematic noise (e.g., an alignment artifact) and is unlikely pathogenic. By relying on the raw alignments, STIX avoids the previously described false negative issues and removes thousands of variants that could have otherwise been associated with disease.

STIX is built on top of the GIGGLE genome search engine<sup>16</sup>. Sequence alignment files contain mostly normal alignments and typically less than 5% “discordant” alignments (split-reads and paired-end reads with unexpected aligned distance between pairs or strand configuration) due to either the presence of a SV or some noise in the sequencing or alignment process (Fig. 1A). These alignment signals are used for detection by all current methods. STIX extracts and tracks all discordant alignments from each sample’s genome (Fig. 1A), then creates a unified GIGGLE index for all samples. When a user provides the SV type, breakpoint coordinates, and its confidence intervals, STIX returns the count of all alignments that support the variant (Fig. 1B). We have deployed web interfaces for STIX queries of 1KG and SGDP aligned to GRCh37 at <http://stix.colorado.edu> (Fig. 1C). The server also supports direct access for integrating STIX into other programs.

Considering the 1KG SV catalog, STIX shows high accuracy in identifying the samples with deletions (0.998), duplications (0.995), and inversions (0.988) (see Methods, Supplementary Table 1). This result is consistent with a previous report showing STIX outperformed DELLY, SVTyper, and SV2 on simulated and real deletions, and demonstrated the best balance between sensitivity and specificity<sup>17</sup>. The STIX index was also 500X smaller than the original alignments and queries ran 620X faster (see Methods).

Using STIX indexes of 1KG and SGDP, we recovered thousands of the somatic SVs published in Catalogue of Somatic Mutations in Cancer<sup>18</sup> (COSMIC) and PCAWG (Figs. 2A–B, 2D–E). These variants were likely either germline or recurrent mutations and unlikely to be driving tumor evolution. Only a fraction of the SVs found by STIX were in either the 1KG or gnomAD SV lists (Figs. 2C, 2F) (See Supplementary note 2).

STIX’s primary utility is to refine SV calls down to a set that can be assessed manually, especially in the absence of DNA sequences from matched-normal tissue. For example, when applied to 183 prostate cancer samples from PCAWG, the MANTA<sup>19</sup> caller recovered, on average, 3892.8 deletions per sample (Fig. 2G). Using the PCAWG calls as the truth set, removing SVs using the matched-normal tissue resulted in 51.4 false positives, 29.9 true positives, and 3.7 false negatives. Using the STIX 1KG database had 30% fewer false positives (35.8), roughly the same number of true positives (23.1), and some additional false negatives (10.5). The results were similar for inversions (Supplementary Fig. 2A) and duplications (Supplementary Fig. 2B). In addition to being over 50X smaller than the tumor-only call set, the STIX-filtered calls were also enriched for putative consequential variants (Supplementary Fig. 3). Using the 1KG and gnomAD SV calls as germline filters was less effective because the average number of false positives was 88X and 55X higher, respectively. Interestingly, these population filters’ true-positive and false-positive results

were similar to the STIX results, indicating that the PCAWG call set likely retained some common SVs.

STIX enables fast and accurate SV frequency estimates directly from population-scale sequencing data, which wasn't possible in previous SV studies due to inconsistent filtering and calling strategies. It does this by indexing all SV evidence directly from the raw alignments, avoiding detection bias, and compressing large consortia data sets. With STIX, we indexed 2,504 samples from the 1,000 Genomes Project and 279 samples from the Simons Genome Diversity Project. These indexes helped improve somatic SV calls and highlighted the potential for recurrent *de novo* SVs (see Supplementary Note 3). The code is freely available at <https://github.com/ryanlayer/stix>.

A limitation of this approach is that, while population frequency is a powerful metric for isolating rare, potentially functional variants, not all rare variants are pathogenic, and making this classification requires further analysis. Additionally, with STIX, and all alignment-based short-read SV methods, it is difficult to determine whether two discordant alignments support the same SV or similar SVs. Discordant paired-end reads provide indirect evidence of an SV, which leads to breakpoint location ambiguity that can affect STIX's resolution (Supplementary Note 4). STIX also does not track per-sample normal coverage levels (due to high storage cost) and cannot distinguish between no support for an SV and insufficient coverage at a particular locus. When considering a large reference cohort, coverage fluctuations in individual samples minimally impact the results. Quantifying read depth or applying other QC metrics is advisable for smaller cohorts or particularly sensitive experiments involving rare variants.

In the future, we plan to explore how STIX may enable data access with lower consent and privacy issues. Reporting summary statistics reduces the likelihood of re-identifying samples, which would help reconcile different consent rights across patient cohorts. With these improvements, STIX could bring the power of thousands of genomes to the diagnosis and treatment decisions process.

## Methods

### STIX SV evidence collection and classification

When a user submits a query, they specify the SV type (deletion, duplication, inversion, or break end) and breakpoint. Breakpoints are encoded as a pair of left and right coordinates, where each coordinate has a chromosome and start and end positions. The left coordinate is strictly downstream of the right and has a lexicographically equivalent or smaller chromosome. The left and right coordinates are extended to account for the indexed samples' insert size distribution and the SV type. Deletions extend the left coordinate downstream and the right coordinate upstream. Duplications extend the left coordinate upstream and the right coordinate downstream. Inversions extend both coordinates downstream for + strand alignments and upstream for - strand alignments. For break ends, the left and right coordinates are not modified.

STIX searches the index using the left coordinate and only retains alignments that also overlap the right coordinate and have a strand configuration that matches the given SV type (listed below). STIX counts the number matches per sample and reports that total to the user.

SV strand configurations:

- Deletions
  - Paired-end alignments must have a +/- orientation.
  - Split-read alignments must have a +/+ or -/- orientation.
- Duplications:
  - Paired-end alignments must have a -/+ orientation.
  - Split-read alignments must have a +/+ or -/- orientation.
- Inversions:
  - Paired-end alignments must have a ++ or -- orientation.
  - Split-read alignments must have a +/- or -/+ orientation.
- Break ends:
  - The only requirement is that the alignments overlap the left and right query coordinates.

### SV evidence extraction and STIX index creation

SV alignment evidence (discordant reads and split reads) are extracted from BAM and CRAM files using excord (<https://github.com/brentp/excord>). Excord scans each alignment to determine if it contains a split read, has a strand configuration that is not +/-, the two aligned ends are not on the same chromosome, and the distance between the two aligned ends is further away than expected (set by the --discordantdistance command line parameter). The expected distance between two reads depends on the size and variance of fragments and can be measured by finding the mean and standard deviation of normal alignments in the BAM file. We recommend using the mean plus two times the standard deviation for the discordant distance. If any of these conditions is true, then the alignment is recorded as a possible piece of SV evidence. For each piece of evidence, excord stores the position and orientation of the two ends into a sample-specific BED file. For example:

1	10022	10122	1	1	249240455	249240538	1	0
1	10031	10131	1	4	191044177	191044238	1	0
1	10036	10136	1	2	243153001	243153102	-1	0
1	10054	10154	-1	19	59097998	59098033	-1	0
1	10066	10166	-1	1	249239980	249240049	-1	0

Excord was written in the Go programming language. Pre-compiled binaries are available under releases in its GitHub repository.

Each sample BED file is sorted and bgzipped. For example :

```
samtools view -b NA12812.bam \
| excord \
--discordantdistance 500 \
--fasta hs37d5.fa.gz \
/dev/stdin \
| LC_ALL=C sort --buffer-size 2G -k1,1 -k2,2n -k3,3n \
| bgzip -c > alt/NA12812.bed.gz
```

Once all sample BED files have been processed an index is created using giggle. For example:

```
giggle index -i "alt/*.gz" -o alt_idx -s -f
```

The last step is to create a sample database from a cohort pedigree file (PED). At a minimum, this file must contain a file header, and one line per sample where each line must contain the sample name and the name of its associated BED file. The following example has three extra fields:

Sample	Sex	Population	Super_Population	Alt_File
NA12812	1	CEU	EUR	NA12812.bed.gz
HG00672	2	CHS	EAS	HG00672.bed.gz
NA12878	2	CEU	EUR	NA12878.bed.gz
HG00674	1	CHS	EAS	HG00674.bed.gz

Creating the sample database requires the giggle index, input PED file name, output database name, and the column number that contains the name of the sample BED file. For example:

```
stix -i alt_idx -p four.ped -d four.ped.db -c 5
```

Once the BED files have been indexed and the sample database has been created from the PED file, STIX can now query the samples for SV evidence. For each query, the user must specify the index location (-i), sample database (-d), SV type (-t), left (-l) and right (-r) breakpoint coordinates, and window size (-s) to consider around each breakpoint. The window size will depend on the size and variance of the sample fragments. We recommend using the same value used for the discordant distance parameter in the excord extraction. The output of STIX is a per-sample count of alignments that support the existence of the SV in the sample. For example:

```
stix \
  -i alt_idz \
  -d four.ped.db \
  -s 500 \
  -t DEL \
  -l 14:68603030-68603035 \
  -r 14:68603738-68603743
```

Id	Sample Sex	Population	Super_Population	Alt_File	Paired	Split
0	HG00672 2	CHS	EAS	HG00672.13.14.bed.gz	8	0
1	HG00674 1	CHS	EAS	HG00674.13.14.bed.gz	7	0
2	NA12812 1	CEU	EUR	NA12812.13.14.bed.gz	7	0
3	NA12878 2	CEU	EUR	NA12878.13.14.bed.gz	11	0

### 1,000 genomes phase three STIX index

2,504 low-coverage BAMs (GRCh37) and the PED file were downloaded from the 1,000 genomes AWS S3 bucket (s3://1000genomes/phase3/data/). Excord was run on each sample with --discordantdistance set to 500.

### Simons Genome Diversity Panel STIX index

252 30X-coverage FASTQ files and PED file were downloaded from the Simons Foundation (<https://www.simonsfoundation.org/simons-genome-diversity-project/>) and aligned to the human reference genome (GRCh37) using BWA-MEM. Excord was run on each sample with --discordantdistance set to 500.

### STIX speed measurement

To test the speed of STIX versus any other alternative genotyping method that accesses the BAMs directly, we compared the time required for STIX to query a specific SV (DEL, 10:105053143-105054173) across the full 1KG cohort versus how much time was required to read the alignments in the same region of each BAM in the 1KG cohort. The assumption being that any genotyping method would need to at least read the alignments, and the I/O time would be a lower bound for any such method.

```
$ time stix \
  -i 1kg_stix_idx \
  -d 1kg.ped.db \
  -s 500 \
  -t DEL \
  -l 10:105053143-105053143 \
  -r 10:105054173-105054173 -S > /dev/null
real    0m1.531s
```

```
$ time ls 1000G_phaseIII_whole_genome/*.mapped.*.low_coverage.*.bam \
| gargs 'samtools view {} 10:105052643-105053143 > /dev/null'
real    16m45.827s
```

### Source code availability and Snakemake<sup>20</sup> pipeline

To improve readability and reproducibility, the source code for all experiments and analysis in this manuscript are part of a Snakemake pipeline available at [https://github.com/ryanlayer/stix\\_paper/blob/main/Snakefile](https://github.com/ryanlayer/stix_paper/blob/main/Snakefile). In the following sections, the relevant rules within the pipeline are listed.

### Accuracy measurement

To determine STIX's classification performance we considered the 1KG cohort and the phase 3 SVs identified by Sudmant et al.<sup>7</sup>. For each reported deletion, duplication, and inversion, we collected the samples that were identified by 1KG as being non-reference. This analysis only included SVs with the CIEND and CIPOS values specified.

For each of those SVs, we then constructed a similar list of samples where STIX found evidence of the same variant.

Given the list of non-reference samples from the 1KG catalog and the list of samples with supporting evidence from STIX, we computed the following values for deletions, duplications, and inversions separately.

- positives (P): Number of non-reference samples in the 1KG catalog
- negatives (N): Number of reference samples in the 1KG catalog (total samples minus positives)
- true positives (TP): Number of samples with evidence from STIX that were non-reference in the 1KG catalog
- true negatives (TN): Number of samples with no evidence from STIX that were reference in the 1KG catalog
- false positives (FP): Number of samples with evidence from STIX that were reference in the 1KG catalog
- false negatives (FN): Number of samples with no evidence from STIX that were non-reference in the 1KG catalog

From these values we computed:

- accuracy =  $(TP + TN) / (P+N)$
- precision =  $TP / (TP + FP)$
- sensitivity =  $TP / P$
- specificity =  $TN / N$
- F1 =  $2TP / (2TP + FP + FN)$



Relevant Snakemake rules:

- `onekg_classification_stats`
- `onekg_sv_table`

### **COSMIC SV evaluation**

The COSMIC SV catalog was downloaded from the COSMIC website (<https://cancer.sanger.ac.uk/cosmic/download>, Structural Genomic Rearrangements, login required). The chromosomal position of the deletions (intrachromosomal deletion), duplications (intrachromosomal tandem duplication), and inversions (intrachromosomal inversion) were extracted and sorted into a compressed BED file.

Relevant Snakemake rules:

- `cosmic_sv_beds`

To determine the overlap between the COSMIC SVs and the 1KG catalog, we converted the 1KG SV VCF to SV-type specific BED files and intersected these files with the corresponding COSMIC BED files. Intersections required a reciprocal overlap of 90%. From these intersections we compute the 1KG allele frequency.

Relevant Snakemake rules:

- `onekg_gts`
- `Cosmic_1kg_overlap`

Relevant scripts:

- `src/get_1kg_ac.py`

To determine the overlap between the COSMIC SVs and the gnomAD SV catalog, we retrieved the version 2.1 SV BED file from the gnomAD web site (<https://gnomad.broadinstitute.org/downloads/#v2-structural-variants>) and split the BED file into SV-type specific BED files and intersected these files with the corresponding COSMIC BED files. Intersections required a reciprocal overlap of 90%.

Relevant Snakemake rules:

- `cosmic_gnomad_overlap`

To determine the overlap between the COSMIC SVs and the STIX for 1KG and SGDP, we submitted a STIX query for each SV in the COSMIC SV-type BED files using a 500 base pair window. For each SV we compute the number of samples with some supporting evidence.

Relevant Snakemake rules:

- `cosmic_stix_1kg_overlap_stats`

Relevant scripts:

- `src/qdel.sh`

**Pan-Cancer Analysis of Whole Genomes SV evaluation**—The PCAWG SV catalogs were downloaded from the ICGC data portal website ([https://dcc.icgc.org/releases/PCAWG/consensus\\_sv/](https://dcc.icgc.org/releases/PCAWG/consensus_sv/)), and combined SV-type specific call sets.

Relevant scripts:

- `src/get_pcawg_sv.sh`

Similar to the process in *COSMIC SV evaluation*, to determine the overlap between the PCAWG SVs and the 1KG catalog, we converted the 1KG SV VCF to SV-type specific BED files and intersected these files with the corresponding PCAWG BED files. Intersections required a reciprocal overlap of 90%. From these intersections we compute the 1KG allele frequency.

Relevant Snakemake rules:

- `pcawg_1kg_overlap`

To determine the overlap between the PCAWG SVs and the gnomAD SV catalog, we retrieved the version 2.1 SV BED file from the gnomAD web site (<https://gnomad.broadinstitute.org/downloads/#v2-structural-variants>) and split the BED file into SV-type specific BED files and intersected these files with the corresponding PCAWG BED files. Intersections required a reciprocal overlap of 90%.

Relevant Snakemake rules:

- `pcawg_gnomad_overlap`

To determine the overlap between the PCAWG SVs and the STIX for 1KG and SGDP, we submitted a STIX query for each SV in the PCAWG SV-type BEDPE files using a 500 base pair window. For each SV we compute the number of samples with some supporting evidence.

Relevant scripts:

- `src/get_pcawg_stix_1kg_overlap.sh`
- `src/get_pcawg_stix_sgdp_overlap.sh`

**De novo SV evaluation**—The de novo SV catalog was retrieved from the GitHub repository referenced in the publication. Those SVs were reported using the GRCh38 human reference genome. We used the UCSC genome browser tools to lift the SVs to GRCH37, then split the file into SV type specific BED files.

Relevant Snakemake rules:

- `denovo_sv_beds`

Similar to the process in *COSMIC SV evaluation*, to determine the overlap between the de novo SVs and the 1KG catalog, we converted the 1KG SV VCF to SV-type specific BED files and intersected these files with the corresponding PCAWG BED files. Intersections

required a reciprocal overlap of 90%. From these intersections we compute the 1KG allele frequency.

Relevant Snakemake rules:

- `denovo_1kg_overlap`

To determine the overlap between the de novo SVs and the gnomAD SV catalog, we retrieved the version 2.1 SV BED file from the gnomAD web site (<https://gnomad.broadinstitute.org/downloads/#v2-structural-variants>) and split the BED file into SV-type specific BED files and intersected these files with the corresponding PCAWG BED files. Intersections required a reciprocal overlap of 90%.

Relevant Snakemake rules:

- `denovo_gnomad_overlap`

To determine the overlap between the de novo SVs and the STIX for 1KG and SGDP, we submitted a STIX query for each SV in the de novo SV-type BED files using a 500 base pair window. For each SV we compute the number of samples with some supporting evidence.

Relevant Snakemake rules:

- `denovo_stix_1kg_overlap`
- `denovo_stix_sgdp_overlap`

### STIX germline filtering evaluation

Information regarding PCAWG donor IDs, file IDs, and specimen type can be found in Supplemental Table 4. Additionally, we have provided a table mapping PCAWG file ID to BAM sample name for BAMs used for Manta SV calls in Supplemental Table 6. We used Manta version 1.6.0 to call SVs in the PCAWG samples. For each tumor, we called svcs in normal mode as well as matched tumor/normal mode.

To create manta SV calling workflows for the ICGC samples, we used the following commands:

#### single sample (normal) mode

```
$MANTA_INSTALL_PATH/bin/configManta.py \
--bam $BAM_PATH \
--referenceFasta $REF_GENOME_PATH \
--runDir $OUTPUT_DIRECTORY
```

#### paired tumor-normal mode

```
$MANTA_INSTALL_PATH/bin/configManta.py \
--normalBam $NORMAL_BAM \
--tumorBAM $TUMOR_BAM
```

```
--referenceFasta $REF_GENOME_PATH \  
--runDir $OUTPUT_DIRECTORY
```

All bam files are aligned to the hs37d5 reference genome which can be downloaded via the 1KG ftp ([ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2\\_reference\\_assembly\\_sequence/hs37d5.fa.gz](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz)). After running the manta configuration script, a runWorkflow.py script is generated in the designated run directory and can be run as follows:

```
./runWorkflow.py -j $THREADS
```

The germline filtering analysis pipeline and associated scripts are contained within a snakemake pipeline located at [https://github.com/ryanlayer/stix\\_paper/tree/main/germline\\_filtering/stix.smk](https://github.com/ryanlayer/stix_paper/tree/main/germline_filtering/stix.smk). Instructions for how to install dependencies and run the pipeline can be found under germline\_filtering/README.md.

The pipeline performs the 1KG STIX queries using the deletion SVs called from the Manta normal mode callsets. Regions that return evidence from the 1KG STIX query are filtered out. For comparison, we then perform filtering by subtracting sets of deletion regions in GnomAD and 1KG, respectively. For evaluation we intersect the STIX, GnomAD, and 1KG filtered regions along with the Manta tumor/normal SV calls with the PCAWG somatic deletion SVs for each sample. All intersection and subtraction operations were performed with a 90% reciprocal overlap threshold. False positives (FP) were the SVs that passed the filters but were not in the PCAWG calls. True positives (TP) were the SVs that passed the filters and were in the PCAWG calls. False negatives (FN) were SVs that did not pass the filters and were in the PCAWG calls.

### STIX query resolution evaluation

For the 31,762 deletions in the 1KG call set that STIX also found evidence for, we shifted the start and end coordinates up and down stream 500bp at 50bp steps. At each step we submitted the STIX query with the new coordinates and counted the number of samples with supporting evidence, then computed the proportion of the number of samples at each step to the number of samples found by the original query. Finally we plotted the median of proportions at each step.

Relevant Snakemake rules:

- stix\_1kg\_deletion\_resolution\_slide
- Stix\_1kg\_deletion\_resolution\_plot

### Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

RML was funded by NIH/NHGRI grant R00 HG009532. ARQ was funded by NIH/NHGRI grant R01 HG010757.

## Data Availability

For most data availability, the Snakemake pipeline provided by [https://github.com/ryanlayer/stix\\_paper](https://github.com/ryanlayer/stix_paper) downloads data used for analyses. For the somatic SV filtering analysis done using PCAWG alignment files, access to data is restricted. Information regarding PCAWG sample data used for this analysis can be found under Methods subsection “*STIX germline filtering evaluation*”.

## References

1. Mahmoud M et al. Structural variant calling: the long and the short of it. *Genome Biol.* 20, 246 (2019). [PubMed: 31747936]
2. Brady SW et al. Combating subclonal evolution of resistant cancer phenotypes. *Nat. Commun* 8, 1231 (2017). [PubMed: 29093439]
3. Quigley DA et al. Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell* 174, 758–769.e9 (2018). [PubMed: 30033370]
4. Ostrander BEP et al. Whole-genome analysis for effective clinical diagnosis and gene discovery in early infantile epileptic encephalopathy. *NPJ Genom Med* 3, 22 (2018). [PubMed: 30109124]
5. Stefansson H et al. Large recurrent microdeletions associated with schizophrenia. *Nature* 455, 232–236 (2008). [PubMed: 18668039]
6. Li Y et al. Patterns of somatic structural variation in human cancer genomes. *Nature* 578, 112–121 (2020). [PubMed: 32025012]
7. Sudmant PH et al. An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81 (2015). [PubMed: 26432246]
8. Collins RL et al. A structural variation reference for medical and population genetics. *Nature* 581, 444–451 (2020). [PubMed: 32461652]
9. Abel HJ et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* (2020) doi:10.1038/s41586-020-2371-0.
10. Lupski JR & Stankiewicz PT *Genomic Disorders: The Genomic Basis of Disease*. (Springer Science & Business Media, 2007).
11. Mallick S et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538, 201–206 (2016). [PubMed: 27654912]
12. Chaisson MJP et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv* 193144 (2018) doi:10.1101/193144.
13. Lek M et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291 (2016). [PubMed: 27535533]
14. Karczewski KJ et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443 (2020). [PubMed: 32461654]
15. Sirén J et al. Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. *Cold Spring Harbor Laboratory* 2020.12.04.412486 (2020) doi:10.1101/2020.12.04.412486.
16. Layer RM et al. GIGGLE: a search engine for large-scale integrated genome analysis. *Nat. Methods* 15, 123–126 (2018). [PubMed: 29309061]
17. Chander V, Gibbs RA & Sedlazeck FJ Evaluation of computational genotyping of structural variation for clinical diagnoses. *Gigascience* 8, (2019).
18. Forbes SA et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr. Protoc. Hum. Genet* Chapter 10, Unit 10.11 (2008).

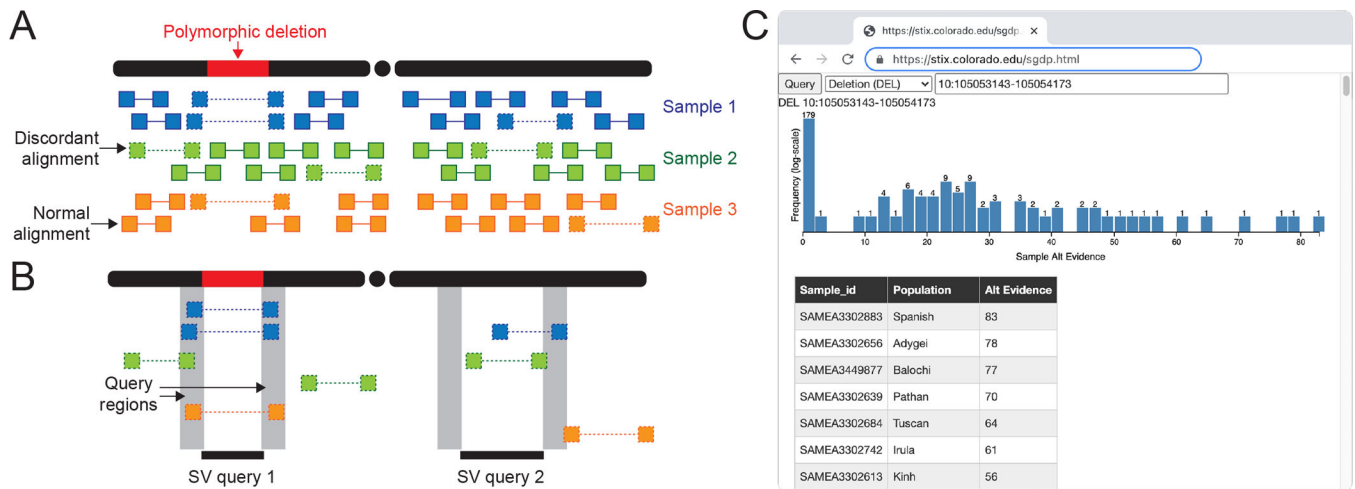
19. Chen X et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* vol. 32 1220–1222 (2016). [PubMed: 26647377]
20. Köster J & Rahmann S Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522 (2012). [PubMed: 22908215]

Author Manuscript

Author Manuscript

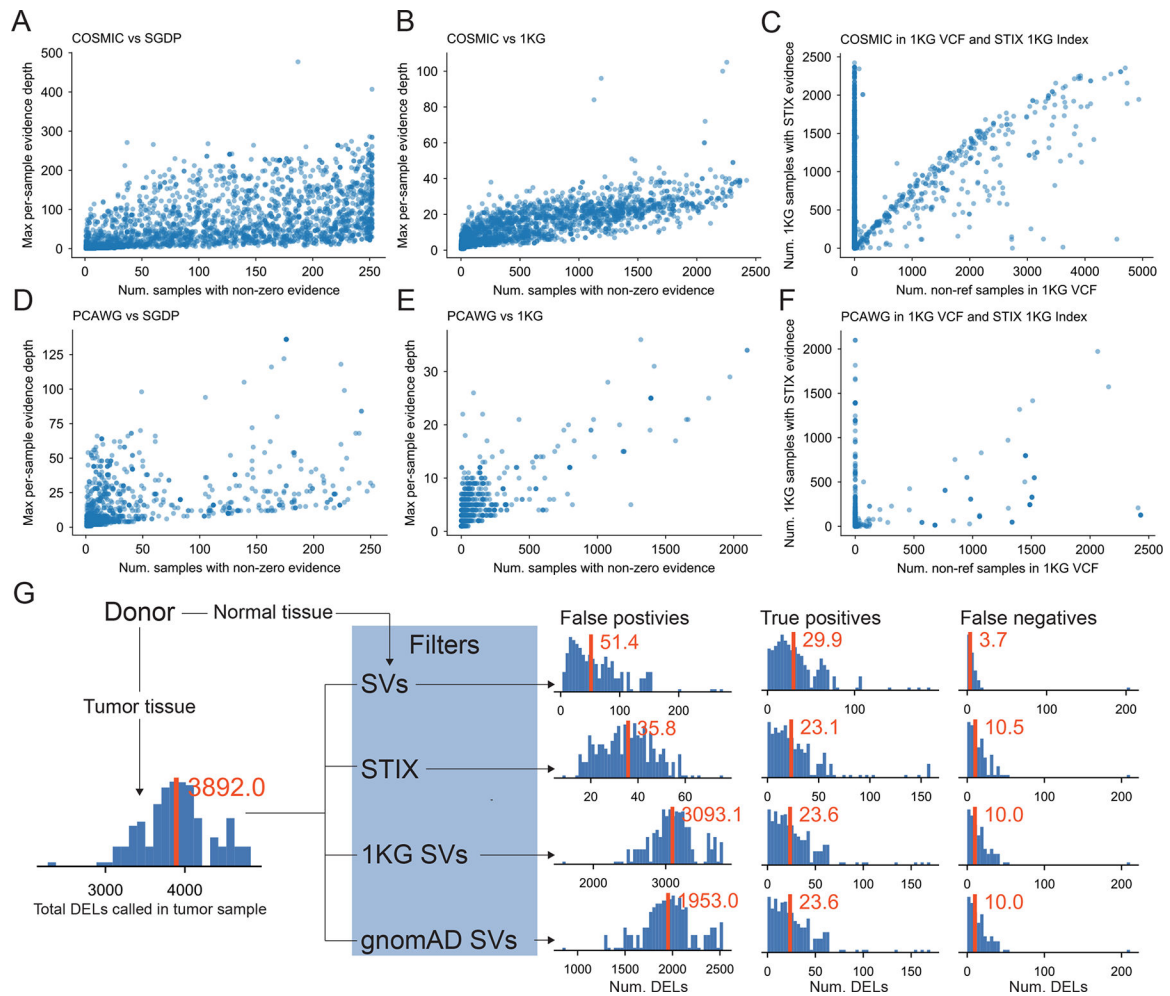
Author Manuscript

Author Manuscript



**Figure 1.**

The STIX structural variant index. **(A,B)** The STIX indexing and query process for three samples and a polymorphic deletion. **(A)** A small number of the alignments that tile the genomes are discordant (designated by a dotted line connected read pairs) because of either an SV or other nonspecific causes (e.g., mapping artifacts). **(B)** Discordant alignments are extracted from all samples and indexed using GIGGLE. Query SVs are mapped to alignments that reside in both regions and are aggregated and returned. The first query returns three alignments in two samples, and the second returns zero alignments. **(C)** The distribution of evidence depths for a deletion searched in the SGDP cohort through the <http://stix.colorado.edu> interface.

**Figure 2.**

(A-F) SVs reported in cancer databases also occur in healthy populations. (A-C) COSMIC contains 46,185 somatic deletions. STIX found evidence for (A) 27.9% of these SVs in SGDP and (B) 27.5% in 1KG. In these two plots (and D and E), we summarize the population-level evidence for each recurring SV (blue dot) by the number of samples with any concordant evidence (x-axis) and the maximum amount of per-sample evidence (y-axis). (C) Only 1% of COSMIC SVs appeared in the 1KG SV call set. The agreement between the STIX and the 1KG call sets is plotted using the population frequency estimates from each method for each SV. (D-F) PCAWG found 84,083 deletions, (D) 3.4% of which were in SGDP and (E) 2% were in 1KG. (F) The 1KG call set contained only 0.2% PCAWG SVs. (G) A comparison of germline filtering strategies for 183 prostate tumor samples that remove tumor deletions found in matched-normal tissue (SV), the STIX index of 1KG, the 1KG SV calls, and the gnomAD SV calls. Histograms show the frequency of sample-level SV counts. Red bars and text give the sample mean. For example, the raw tumor calls had, on average, 3,892.0 SVs, and STIX filtering yielded, on average, 35.8 false positives, 23.1 true positives, and 10.5 false negatives.