SVExpress: identifying gene features altered recurrently in expression with nearby structural variant breakpoints

Instructions for use

Introduction

Somatic structural variations (SVs) are rearrangements of large DNA segments within the cancer genome. SVs may impact the expression of nearby genes in several ways, including forming fusion transcripts or disrupting or repositioning cis-regulatory elements near genes. Recent studies [1-3] have demonstrated an analysis approach to integrate SV with gene expression data, identifying gene-level associations between altered expression and nearby SV breakpoints in proximity to genes. Genes recurrently deregulated with SVs may involve topologically associated domain (TAD) disruption or enhancer hijacking.

The "SVExpress" suite of computational tools allows one to identify SV breakpoint-to-expression associations across a set of cancer samples profiled for both SVs and gene transcription. SVExpress takes as input a table of SV breakpoints and a gene-to-sample expression matrix. SVExpress then constructs a gene-to-sample breakpoint matrix, which the user can then integrate with the expression matrix by linear regression modeling, using the provided R code. Furthermore, using SVExpress, top SV-gene associations identified can be examined in terms of enhancer hijacking (e.g., an enhancer represented by one breakpoint positioned in proximity to a gene nearby the other breakpoint) or in terms of disruption of TADs.

Tool components

SVExpress consists of two main components:

- An Excel macro-enabled workbook file ("SVExpress_Excel_macros.xlm"), with macros that generate the gene-to-sample breakpoint matrix, link enhancers to SV-gene associations, or link TADs to SVs.
- R code ("SVExpress_r-code.R") for running linear modeling to associated altered expression with nearby SV breakpoint pattern, given gene-level data matrices for SV breakpoints, expression, and copy number alterations (CNAs).

The provided Excel file includes three macros:

- "Generate_Gene_to_Sample_SV_Table" macro, which generates a gene-to-sample table of relative SV breakpoint patterns, given an input table of SV breakpoints and a set of gene coordinates.
- Generate_SV_to_Enhancer_Associations" macro, which generates a list of SV-to-gene-to-enhancer associations, given input tables of SV breakpoints, enhancer coordinates, and SV-to-gene associations.
- "Generate_SV_to_TAD_Associations" macro, which annotates SVs in terms of TAD-preserving or TAD-disrupting, given an input table of SV breakpoints and a set of TAD coordinates.

For both the Excel macros and the R code, the downloads provide example data to run using SVExpress tools. Example data include those from the Cancer Cell Line Encyclopedia (CCLE)[4], with these data limited to those involving the set of genes with significant associations between SVs and expression (FDR<10%, using the 1Mb region window).

Input data and formatting

For use with the Excel macros, the input datasets comprise a single Excel workbook with multiple tabs. The Excel file "SVExpress_data_for_Excel_macros.xlsx" includes example data in the proper format. The first row of each data tab has the column headings (which should be in the prescribed order, although the actual headings do not require a specific syntax), and the actual data begin on the second row.

- "SV Breakpoints" tab, with the following columns (in the given order):
    - "SV breakpoint id", the identifier for the SV breakpoint (each SV should consist of two breakpoints). Each breakpoint ID should be unique (including the two breakpoints for a given SV).
    - "sample id," the identifier of the sample with the SV.
    - "chrA," the chromosome of breakpoint A.
    - "posA," the position of breakpoint A.
    - "chrB," the chromosome of breakpoint B.
    - "posB," the position of breakpoint A.
    - "oriA," the orientation of breakpoint A. Denotes whether the upstream (+1) or downstream (−1) sequence is fused relative to the given coordinates.
    - "oriB," the orientation of breakpoint B. Denotes whether the upstream (+1) or downstream (−1) sequence is fused relative to the given coordinates.
- "Genes" tab, with the following columns (in the given order):
    - "gene id," unique gene identifier (e.g., Ensembl ID or Entrez ID).
    - "gene name," gene symbol or description.
    - "chromosome," gene chromosome.
    - "gene start," gene start (in bases).
    - "gene end," gene end (in bases). This should always be greater than gene start.
    - "strand," the gene strand (1 or -1).
- "Enhancers" tab (for "Generate_SV_to_Enhancer_Associations" macro), with the following columns (in the given order):
    - "enhancer id," unique enhancer identifier.
    - "chromosome," enhancer chromosome.
    - "position," enhancer position, e.g., the start position or the position midway between the start and end positions.
- "Genes-SVs" tab (for "Generate_SV_to_Enhancer_Associations" macro), which consists of the gene-to-SV mappings used to generate the gene-to-sample breakpoint matrix. The output Gene-SV-sample associations generated using the "Generate_Gene_to_Sample_SV_Table" macro may be used here, and so the user would not have to assemble this table by hand. The following columns (in the given order) are required:

    - "SV breakpoint id," the identifier for the SV breakpoint.
    - "gene id," the identifier of the gene. For a given gene, the "Generate_Gene_to_Sample_SV_Table" macro assigns the SV breakpoint closest to the gene start.
    - "sample id," the identifier of the sample with the SV breakpoint.
- "TADs" tab (for "Generate_SV_to_TAD_Associations" macro), with the following columns (in the given order):
    - "chromosome," TAD chromosome.
    - "start," TAD start (in bases).
    - "end," TAD end (in bases).

Many SV algorithms output exactly one entry per SV. Where this is the case, the user will need to duplicate the SV entries, but with the "ChrB/posB/oriB" entries under the "ChrA/posA/oriA" fields, and vice versa. For example, only the posA breakpoints from the "SV Breakpoints" tab will be used to generate the gene-to-sample SV breakpoint matrix, so both breakpoints need to be represented separately (with unique SV breakpoint Ids) in the "SV Breakpoints" tab.

Each data entry in the input Excel data should involve chromosomes 1-22 and X/Y (e.g., as "1" or as "chr1"). Entries involving other chromosomes (e.g., "MT") will not be mapped and, in some cases, might lead to the macro stopping abruptly.

Enhancer data in the example file are from Kumar *et al*. [5]. In the example file, TAD coordinates are from Dixon *et al*. [6] (IMR90 cell line).

For the R portion of SVExpress, the input datasets have tab-delimited text file format. The first column of each data file (with no column heading) should be the unique gene identifiers (with the same gene ordering between data files). The column headings should be the sample names. The following data files are needed:

- SV breakpoint matrix, a matrix of gene-to-sample breakpoint associations. For example, if using a specified fixed region window (e.g., 100kb upstream or within the gene), a given entry is "1" if an SV breakpoint is present in relation to the gene for the given sample, and "0" if otherwise. Alternatively, with the "use 1 MB region window surrounding genes" option, the relative distances of the SSV breakpoint closest to the start of each gene are tabulated, and the log2-transformed distance of this closest breakpoint (or log2 of 1Mb if no breakpoint is present) is used to relate genes to samples.
- Gene expression matrix, a matrix of gene expression values. We highly recommend using log-transformed values, as these are more appropriate for linear modeling.
- Copy number alteration matrix, a matrix of copy number alteration values. Log2 tumor/normal ratios are preferred (e.g., gain if greater than zero, loss if less than zero).

A file with the sample cancer type assignments is also required. The first row in the file consists of the sample ids, and the second row lists the cancer types corresponding to the samples.

Running excel macros

In Excel, the user runs macros by first opening the “SVExpress_Excel_macros.xlm” file. Then, with another Excel workbook consisting of the input data tabs being “active” (i.e., in front of any other open files), the user runs one of the Excel macros (e.g., using Alt+F8 keys for Windows or selecting “Macros” from the “View” ribbon). A dialog box with a macros list will appear, and the user selects one of the macros and then clicks the “Run” button.

To use SVExpress, if it does not already, your installed Excel needs to allow the opening and running of macros. When opening the SVExpress_Excel_macros.xlm macro-enabled workbook, select "Enable macros" (not "Disable macros"), if so prompted.

Running R code

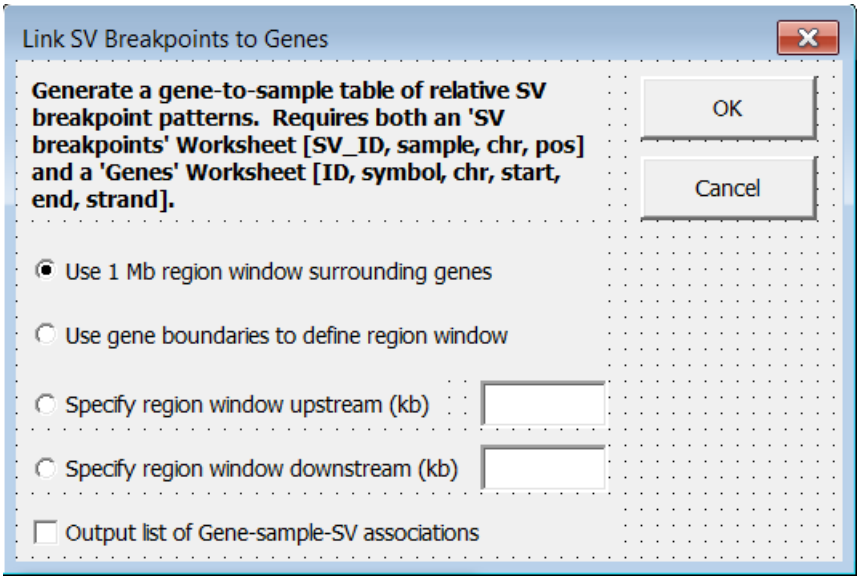
To run the provided R code (“SVExpress_r-code.R” file from the zipped file “SVExpress_r-code.zip”), the user should first open the file using a text editor or using the R program. For the first line of code ("setwd..."), the user edits this to point to the input files on the user's computer. Finally, the user runs the code in R, e.g., copying the code and pasting it into the R console window.

The code provided should work with all versions of R. The code was written using R version 3.1.0, but there are no library dependencies other than the base library.

Generating the gene-to-sample SV breakpoint matrix

The “Generate_Gene_to_Sample_SV_Table” Excel macro generates a gene-to-sample table of relative SV breakpoint patterns. The input workbook requires two worksheet tabs: (1) an “SV breakpoints” worksheet (with columns SV ID, sample, chromosome, position, in this order) and a “Genes” worksheet (with columns ID, symbol, chromosome, start, end, strand, in this order).

On running the macro, the following dialog box will appear:



The fields in the dialog box are as follows:

1) Options to specify the region of interest relative to each gene:

- “Use 1 Mb region window surrounding genes.” With this option, the macro tabulates the relative distances of the SSV breakpoint closest to each gene's start, and the log2-transformed distance of this closest breakpoint (or log2 of 1Mb if no breakpoint is present) relates genes to samples. This option represents the “relative distance metric” method described in ref [2]. With the relative distance metric method, breakpoints that occur close to the gene will have more weight in identifying SV-expression associations, while breakpoints further away but within 1Mb will influence to a lesser extent.
- “Use gene boundaries to define region window.” With this option, for a given gene and sample in the breakpoint matrix, the value is “1” if a breakpoint occurs within the gene and “0” if otherwise. This option is a good one for identifying possible gene fusions or SV-mediated gene disruption.
- “Specify region window upstream (kb).” With this option, for a given gene and sample in the breakpoint matrix, the value is “1” if a breakpoint occurs upstream of the gene (within a user-specified window, e.g., “100” for 100kb upstream) and “0” if otherwise.
- “Specify region window downstream (kb).” With this option, for a given gene and sample in the breakpoint matrix, the value is “1” if a breakpoint occurs downstream of the gene (within a user-specified window, e.g., “100” for 100kb downstream) and “0” if otherwise.

2) Option to output list of gene-to-sample-to-SV associations. If this option is selected, the set of gene-to-SV associations used to construct the SV breakpoint matrix are listed individually. For every gene-to-sample relationship represented by the output matrix, at most one SV is listed, which would be the SV with a breakpoint that is closest to the gene start. If multiple SVs involve a given gene for a given sample, the SV with breakpoint closest to the gene is used for the matrix. This option will generate as output one or more Excel worksheet tabs with the root title “Genes-SVs,” with columns “SV breakpoint id”, “gene id”, and “sample id”. If the above "1Mb" region window option is selected, then the fourth column will give the SV breakpoint's relative distance to the gene. The entire set of gene-to-sample-to-SV associations may span multiple Excel worksheet tabs (as the full number may exceed 1 million rows). The "Genes-SVs" output worksheets from this option can be the input for the "Generate_SV_to_Enhancer_Associations" macro.

After successfully running the “Generate_Gene_to_Sample_SV_Table” macro, the resulting “Output” worksheet tab provides the gene-to-sample SV breakpoint matrix, which may then be run with the R code to generate gene-level associations between altered expression and nearby SV breakpoints.

<u>Associating enhancers with gene-to-SV mappings</u>

The “Generate_SV_to_Enhancer_Associations” Excel macro generates a list of SV-to-gene-to-enhancer associations. The input workbook requires four worksheet tabs: 1) an “SV breakpoints” worksheet (with columns SV breakpoint id, sample id, chrA, posA, chrB, posB, oriA, oriB, in this order), 2) a “Genes” worksheet (with columns id, symbol, chr, start, end,

strand, in this order), 3) an “Enhancers” worksheet (with columns id, chr, position, in this order), and 4) a “Genes-SVs” worksheet (with columns SV breakpoint id, gene id, sample id, in this order).

The user would typically run the “Generate_SV_to_Enhancer_Associations” macro after the “Generate_Gene_to_Sample_SV_Table” macro. The latter macro can output a list of gene-to-sample-to-SV associations, representing the input as part of the “Genes-SVs” worksheet. The entire set of gene-to-sample-to-SV associations may span multiple Excel worksheet tabs (as the full number may exceed 1 million rows). Therefore, the user may need to run the “Generate_SV_to_Enhancer_Associations” macro multiple times with different “Genes-SVs” worksheets.

On running the “Generate_SV_to_Enhancer_Associations” macro, the dialog provides the option to indicate whether the entries in the “SV breakpoints” worksheet are sorted alphanumerically by SV breakpoint id. If the SV breakpoints are sorted in this way, and the user indicates this using the checkbox, the macro should run much faster than if the breakpoints are not sorted. If the “SV breakpoints” worksheet is not sorted by id, the macro will sort the breakpoint entries in an internal data structure, to facilitate the subsequent searches.

For each gene-to-breakpoint association (as defined using breakpoint A), the potential for enhancer translocation represented by the position B breakpoint in proximity to the gene is evaluated (assuming no other disruptions involving the region). The macro examines the region 1 Mb of the SV position B breakpoint for any enhancers. Also, any enhancers located within 1 Mb of the unaltered gene are identified. Only enhancers positioned upstream of the given gene are considered here. Only SVs with breakpoints on the distal side from the gene are part of the analysis. In other words, for genes on the negative strand, the upstream sequence of the breakpoint (denoted as positive orientation) is fused relative to the breakpoint coordinates, and for genes on the positive strand, the downstream sequence of the breakpoint (denoted as negative orientation) is fused relative to the breakpoint coordinates.

On completion, the macro outputs a worksheet entitled “Output Enhancers-SVs” with the following columns:

- “Gene,” gene identifier
- “Gene chromosome,” gene chromosome
- “Gene start,” gene start position
- “Gene end,” gene end position
- “Gene strand,” gene strand
- “SV,” SV breakpoint identifier
- “Chromosome, breakpoint A,” chromosome of breakpoint A
- “Chromosome, breakpoint B,” chromosome of breakpoint B
- “Position, breakpoint A,” position of breakpoint A
- “Position, breakpoint B,” position of breakpoint B
- “Sample,” sample name
- “Enhancer, SV-associated,” id of translocated enhancer associated with breakpoint B

- “Enhancer position, breakpoint B,” position of translocated enhancer associated with breakpoint B
- “Enhancer, unaltered gene,” id of enhancer associated with the unaltered gene (the enhancer closest to the gene start is used)
- “Enhancer position, unaltered gene,” position of enhancer associated with the unaltered gene
- “Relative distance from gene, SV-associated enhancer,” relative distance from gene start of translocated enhancer associated with breakpoint B
- “Relative distance, enhancer for unaltered gene,” relative distance from gene start of enhancer associated with the unaltered gene

Using the above output, one can look for SV-to-gene associations that involve a translocated enhancer positioned closer to the gene start than what would involve the unaltered gene. The user may wish to input all gene-to-sample-to-SV associations found using the “Generate_Gene_to_Sample_SV_Table” macro. The user may then compare the number of enhancer hijacking events found for the subset of gene-to-sample-to-SV associations involving gene over-expression with the number of enhancer hijacking events found for the entire set of gene-to-sample-to-SV associations. We often observe a significant enrichment of putative enhancer hijacking events involved with the set of SVs associated with gene over-expression [1-3].

Associating TADs with SVs

The “Generate_SV_to_TAD_Associations” Excel macro takes a given set of SV breakpoints and note whether each SV is 'TAD preserving' or 'TAD disrupting'. For TAD preserving SVs, both SV breakpoints locate within the same TAD. For TAD disrupting SVs, the SV breakpoints span TAD boundaries. The input workbook requires two worksheet tabs: 1) an 'SV breakpoints' worksheet (with columns SV breakpoint id, sample id, chrA, posA, chrB, posB, in this order) and 2) a 'TADs' worksheet (with columns chr, start position, end position, in this order).

On completion, the macro adds the results to the 'SV breakpoints' worksheet, in columns I-K. The output columns list the TAD associated with SV breakpoint position A ("TAD, bp A"), the TAD associated with SV breakpoint position B ("TAD, bp B"), and whether the SV is preserving or disrupting ("Preserving/Disrupting"). For SVs associated with gene over-expression, we often observe a significant enrichment of TAD-disrupting SVs [2, 3].

Generating gene-to-breakpoint expression correlations

The “SVExpress_r-code.R” file provides R code that does the linear modeling to assess the correlation between expression of each gene and the presence of nearby SV breakpoints. Four input data files are needed: a gene-to-sample breakpoint matrix file, a gene-to-sample expression matrix file, a gene-to-sample copy number alteration matrix file, and a sample cancer type file. Before running the R code, the user needs to modify the first lines of code with the input files' directory names.

The output file ("lm_test_results.txt," in the same directory as the input files) may be opened in Excel, using the Tab character as the delimiter to separate the data into columns. The order of the rows corresponds to the order of genes in the input data files.

The output of the R code gives the results of three linear models for each gene:

- Expression vs. SV breakpoints, no covariates (columns 2-3)
- Expression vs. SV breakpoints, after correcting for cancer type (columns 4-5)
- Expression vs. SV breakpoints, after correcting for both cancer type and gene-level copy number alteration (columns 6-7)

For each model, the output provides two values for each gene. The first value is the t-statistic for the expression vs. breakpoints correlation, and the second value is the p-value for the significance of this correlation. Some significant genes in the first two models may not be significant in the third model, as genomic rearrangements are often associated with widespread copy-number alteration (CNA) patterns [1-3]. On the other hand, any SV-expression associations that would not entirely be explained by the gene-level CNA patterns should be significant in the third model correcting for CNA. If the "Use 1 Mb region window surrounding genes" option was used with the "Generate_Gene_to_Sample_SV_Table" Excel macro to generate the SV breakpoint pattern matrix (i.e., the relative distance metric was used), then a NEGATIVE t-statistic indicates that SHORTER breakpoint distances from the gene associate with INCREASED expression. In other words, a negative t-statistic means that SVs associate with higher gene expression, while a positive t-statistic means that SVs associate with lower gene expression. When using the other "Generate_Gene_to_Sample_SV_Table" options, then a POSITIVE t-statistic indicates that the presence of nearby breakpoints within the given region associate with INCREASED expression. Given the significance p-values for all genes, corrections for multiple testing can be carried out using standard methods such as Storey and Tibshirani [7].

References

1. Zhang Y, Chen F, Fonseca N, He Y, Fujita M, Nakagawa H, Zhang Z, Brazma A, PCAWG_Transcriptome_Working_Group, PCAWG_Structural_Variation_Working_Group, Creighton C: **High-coverage whole-genome analysis of 1220 cancers reveals hundreds of genes deregulated by rearrangement-mediated cis-regulatory alterations.** *Nat Commun* 2020, **E-pub Feb 5**.
2. Zhang Y, Yang L, Kucherlapati M, Hadjipanayis A, Pantazi A, Bristow C, Lee E, Mahadeshwar H, Tang J, Zhang J, et al: **Global impact of somatic structural variation on the DNA methylome of human cancers.** *Genome Biol* 2019, **20:**209.
3. Zhang Y, Yang L, Kucherlapati M, Chen F, Hadjipanayis A, Pantazi A, Bristow C, Lee E, Mahadeshwar H, Tang J, et al: **A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases.** *Cell Reports* 2018, **24:**515-527.
4. Ghandi M, Huang F, Jané-Valbuena J, Kryukov G, Lo C, McDonald Er, Barretina J, Gelfand E, Bielski C, Li H, et al: **Next-generation characterization of the Cancer Cell Line Encyclopedia.** *Nature* 2019, **569:**503-508.

5. Kumar S, Warrell J, Li S, McGillivray PD, Meyerson W, Salichos L, Harmanci A, Martinez-Fundichely A, Chan CWY, Nielsen MM, et al: **Passenger mutations in more than 2500 cancer genomes: Overall molecular functional impact and consequences.** *Cell* 2020, **180:**915-927.
6. Dixon J, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu J, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions.** *Nature* 2012, **485:**376-380.
7. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci USA* 2003, **100:**9440-9445.