SVGAR_sf: A novel computational algorithm to discover driver structural variant genes within cancer patient cohorts

Xavi Loinaz, Chip Stewart, Johnathan Dagan, JB Alberge, Andrew Dunford, Julian Hess, Esther Rheinbay, Rameen Beroukhim, Gad Getz

Getz Lab, Cancer Genome Computational Analysis Group

ABSTRACT

Major international cancer genomics projects are aimed at finding genomic drivers of cancer in order to develop appropriate treatments. Many such genomic drivers pertain to structural variants (SVs) within the cancer genome, where genomic breakpoints in various regions attach aberrantly to other regions. SVGAR_sf (Structural Variant Gene Annotation-Related significance finder; pronounced "sugar S.F.") is a novel computational algorithm that finds enrichment of certain types of SVs based on their functional annotation to certain genes (i.e. loss-of-function or gain-of-function), leading to discovery of putative SV driver genes within cohorts of cancer patients. Applied to the diffuse large B-cell lymphoma (DLBCL) cohort from the National Cancer Institute's Clinical Trial Sequencing Project as well as a 902-patient multiple myeloma cohort, we recapitulate approximately a dozen known drivers for each cancer type as well as novel candidates, outperforming other existing baseline methods.

BACKGROUND

Structural variants (SVs) are classified as deletions, tandem duplications, inversions, and translocations of the genome of significant length (encompassing at least hundreds of base pairs). They comprise a predominant form of somatic aberration for various cancer types, and a subset of these aberrations are also known to drive such cancer types.

Historically in the field of cancer biology, however, there have been a lack of established methods to infer driver genes via structural variation within cohorts of cancer patients. Such methods have been well-established for mutations (such as MutSig2CV[1] and DIG[2]) and copy number alterations (such as GISTIC2.0[3] and BISCUT[4]) but not as much for structural variation even while structural variant cancer drivers are of comparable, if not sometimes greater, aetiological impact. For example, in chronic myeloid leukemia (CML), the BCR-ABL1 fusion is known to be the main driver of the cancer. Martinez-Fundichely et al. recently published CSVDriver[5] in *Nature Communications* in September 2022 for inferring driver genes via structural variation, but here we show that the method we've developed, SVGAR_sf (Structural Variant Gene Annotation-Related significance finder; pronounced "sugar S.F."), significantly outperforms CSVDriver for the two cohorts we've tested both methods on.

SVGAR_sf METHODOLOGY

We use a binomial test to find significant loss-of-function (LoF) enrichment…

$$P(X \geq k) = \sum_{a=k}^{n} \binom{n}{a} p^a (1-p)^{n-a}$$

3 parameters necessary to specify:
n = number of SVs pertaining to a gene → Easy to find
k = number of LoF SVs for gene
p = probability of events pertaining to gene being LoF
→ Need to find these!

Find k: Infer functional annotations using dRanger annotator

Find p: Split into 1D and 2D events and assume uniform breakpoint distribution for 2D events

Our steps:

1. Find p
2. Label which events are LoF and which are not for each gene (find k)
3. Run binomial test
4. Multiple test correction for all genes tested

Step 1 - Find probability of LoF for SVs for each gene:

1. Find expected dRanger annotations within specified window of gene

Example of number of annotations every 10bp within 1Mbp of CIITA:

exon	3'UTR	5'UTR	promoter	before_gene	after_gene
346	898	14	300	100000	100000
intron1	**intron2**	**intron3**	**intron4**	**intron5**	**intron6**
1790	24	291	19	251	47
intron7	**intron8**	**intron9**	**intron10**	**intron11**	**intron12**
47	93	127	219	41	50
intron13	**intron14**	**intron15**	**intron16**	**intron17**	**intron18**
531	71	198	364	15	73

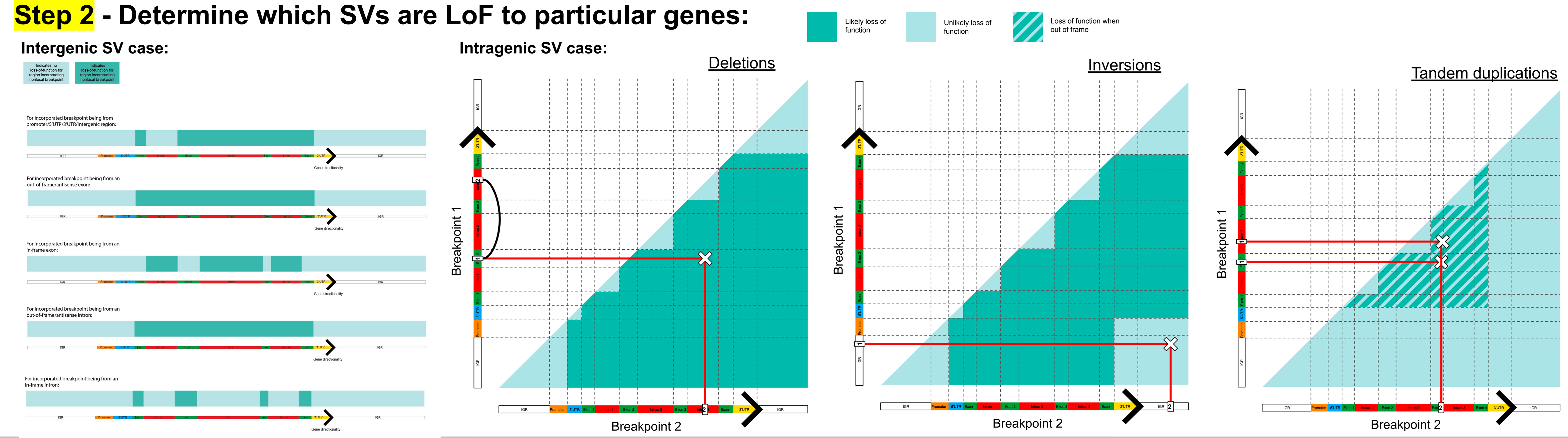
2. Find probabilities of expected dRanger annotations within specified window of gene given uniform distribution of breakpoints (Example of probability matrix for all intragenic events for CIITA)

Placement of breakpoints	Deletion cases	Tandem duplication cases	Inversions cases
before_gene-before_gene	0.105732	0.052510	0.077845
before_gene-promoter	0.000634	0.000315	0.000467
before_gene-5'UTR	0.000030	0.000015	0.000022
before_gene-exon	0.000732	0.000363	0.000539
before_gene-intron	0.008989	0.004464	0.006618
before_gene-3'UTR	0.001899	0.000943	0.001398
before_gene-after_gene	0.211463	0.105020	0.155690
promoter-promoter	0.000001	0.000000	0.000001
promoter-5'UTR	0.000000	0.000000	0.000000
promoter-exon	0.000002	0.000001	0.000002
promoter-intron	0.000027	0.000013	0.000020
promoter-3'UTR	0.000006	0.000003	0.000004
promoter-after_gene	0.000634	0.000315	0.000467
5'UTR-5'UTR	0.000000	0.000000	0.000000
5'UTR-exon	0.000000	0.000000	0.000000
5'UTR-intron	0.000001	0.000001	0.000001
5'UTR-3'UTR	0.000000	0.000000	0.000000
5'UTR-after_gene	0.000030	0.000015	0.000022
exon-exon (in-frame)	0.000000	0.000000	0.000000
exon-exon (out-of-frame)	0.000001	0.000000	0.000001
exon-intron (in-frame)	0.000010	0.000005	0.000008
exon-intron (out-of-frame)	0.000021	0.000010	0.000015
exon-3'UTR	0.000007	0.000003	0.000005
exon-after_gene	0.000732	0.000363	0.000539
intron-intron (in-frame)	0.000050	0.000025	0.000037
intron-intron (out-of-frame)	0.000100	0.000050	0.000074
intron-intron (within same intron)	0.000041	0.000020	0.000030
intron-3'UTR	0.000081	0.000040	0.000059
intron-after_gene	0.008989	0.004464	0.006618
3'UTR-3'UTR	0.000009	0.000004	0.000006
3'UTR-after_gene	0.001899	0.000943	0.001398
after_gene-after_gene	0.105732	0.052510	0.077845

3. Add up all probabilities of all possible events we deem to be LoF

P(LoF for CIITA) = 0.122

Step 2 - Determine which SVs are LoF to particular genes:



Step 3 - Run binomial test

Step 4 - Perform multiple test correction using Benjamini-Hochberg

RESULTS

SVGAR_sf results for CTSP DLBCL:

* indicates known driver

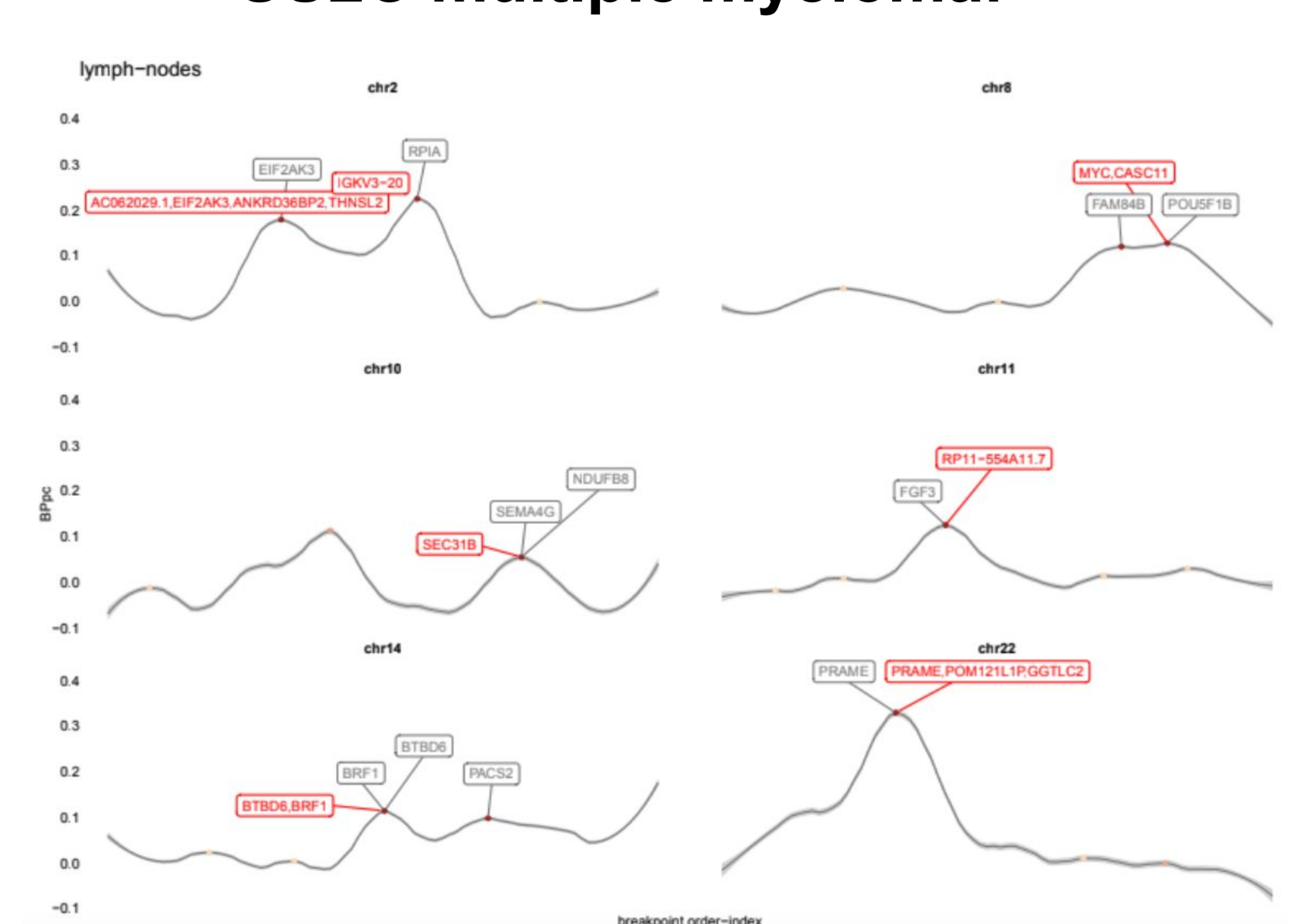
	gene	LOF q-value
*	[redacted]	3.67E-11
	[redacted]	6.30E-09
*	[redacted]	9.20E-08
*	[redacted]	9.20E-08
*	[redacted]	1.14E-04
*	[redacted]	1.96E-03
	[redacted]	9.37E-03
	[redacted]	5.93E-02
*	[redacted]	5.93E-02
*	[redacted]	6.68E-02
	[redacted]	7.78E-02
*	[redacted]	8.89E-02
*	[redacted]	1.05E-01
	[redacted]	1.29E-01
	[redacted]	1.62E-01
*	[redacted]	2.24E-01
	[redacted]	2.24E-01

Note: Had to redact gene names because CTSP DLBCL data is confidential

SVGAR_sf results for SU2C multiple myeloma:

gene	LOF q-value	Expected according to JB	Comment by JB
IGLL5	1.97E-42	TRUE	VDJ recombination, physiologic LOF (IGLL5 is also IGLV1)
TRAF3	1.28E-19	TRUE	14q target common double hit
NSD2	8.52E-09	TRUE	IGH translocation target
SP140	4.77E-06	TRUE	known driver mutated
CDKN2C	1.91E-04	TRUE	known driver del 1p
SP140L	7.7E-04	TRUE	known driver SP140 cluster
PRSS2	7.7E-04		not in hg19? but deletions look OK. region on chr7
WWOX	8.98E-04	TRUE	actually MAF is the target (GOF) and WWOX a long gene neighbor
PRR14L	1.61E-03		
MFF	2.33E-03		
EMSY	2.58E-03		
MTRNR2L6	3.75E-03		strange region on chr7 to be reviewed
RB1	5.51E-03	TRUE	known driver
AMN	5.98E-03		same as TRAF3 / IGH hotspot?
SDCCAG8	6.77E-03		1q
CDKN2A	2.51E-02	TRUE	known driver
GPR180	2.61E-02		TGDS neighbor on chr13
ICE1	5.06E-02		
TBL1XR1	9.27E-02		
DIAPH2	1.04E-01	TRUE	known SV hotspot on chrX
TRAF2	1.13E-01	TRUE	known driver
SP100	1.18E-01	TRUE	known SP140 driver cluster
HERC3	1.44E-01		
NSMCE2	1.44E-01		
TGDS	1.59E-01	TRUE	suspected driver chr13
PLPP5	2.26E-01		
ARHGAP10	2.31E-01		nr3c2 neighbor?

CSVDriver results for SU2C multiple myeloma:



Key points:

- Most CSVDriver hits did not seem relevant or known to multiple myeloma
- Although this analysis has not been finalized for the results of the most recent version of SVGAR_sf, it seems that SVGAR_sf's hits were also depleted in terms of RNA for CTSP DLBCL, acting as an orthogonal signal validating SVGAR_sf's hits
- This analysis has not been finalized for the results of the most recent version of SVGAR_sf, but it seemed that for CTSP DLBCL there were many gene hits from SVGAR_sf that were not picked up by by corresponding peaks on GISTIC

CONCLUSION

Overall, SVGAR_sf shows promising results to predict driver genes via structural variation better than other established methods in the field. Currently work is being done to expand its functionality to discover gain-of-function drivers. More extensive validation of such driver hits can be performed through analyzing RNA-seq data as well as looking at the Cancer Dependency Map. We hope to test SVGAR_sf on other cohorts such as those from PCAWG to further validate it and adjust it as necessary.

REFERENCES

1) Lawrence, M. S., et al. (2014, January 5). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature.*
2) Sherman, M. A., Yaari, A. U., Priebe, O., et al. (2022, June 20). Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. *Nature Biotechnology.*
3) Mermel, C. H., et al. (2011, April 28). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers - genome biology. *Genome Biology.*
4) Shih, J., et al. (2023, June 28). Cancer aneuploidies are shaped primarily by effects on tumour fitness. *Nature.*
5) Martinez-Fundichely, A., Dixon, A., & Khurana, E. (2022, September 26). Modeling tissue-specific breakpoint proximity of structural variations from whole-genomes to identify cancer drivers. *Nature Communications.*