CITATION

For any resultant publications using please cite:
Matthew A. Field, Vicky Cho, T. Daniel Andrews, and Chris C. Goodnow (2015). "Reliably detecting clinically important variants requires both combined variant calls and optimized filtering strategies" PlosOne. doi: 10.1371/journal.pone.0143199

FILE CONTENT/FORMAT

The file contains a summary of all structural variants (SVs) called by either delly or lumpy. Delly calls five SV types (inversions, deletions, duplications, translocations, and insertions) while lumpy call the same SV types except for insertions. SV calls overlapping either an ensemble gene or an exon from a canonical contig are reported. **Note that SVs that overlap >1 gene or exon are grouped together but a line is reported for each gene/exon overlapped.** The lines are tab delimited and designed for loading into excel (set delimiter as tab).

CUTOFFS USED

For this report there are several important cutoffs utilised:

1) **Supporting read pairs**: A minimum of four supporting read pairs is required for a SV to be considered passing. Delly and lumpy will include SVs with 2 or 3 supporting read pairs however these are annotated as failing due to insufficient read pairs
2) **Maximum length**: SVs must be <100kb to be passed within the reports. Delly and lumpy do not filter for length.
3) **Overlap definition**: SVs are combined if both breakpoints are within 250bp of each other. This number has been empirically derived.

SV GROUPINGS

The SVs are divided in several major groups and divided by lines containing '---' in all cells and given priorities of high, medium, low, or lowest:

a) SV type:
SV types of del, dup, ins, inv, and tra are treated separately. **Dels and dups are fundamentally different from the others in that they affect the entire region whereas translocations, inversions, and insertions only affect the breakpoint.**
b) NOVEL_OR_RARE entries:
Novel SVs (i.e. no matching dgv entry) and rare SVs (dgv allele frequency is <2%). The match to dgv uses the overlap definition above of 250bp for each breakpoint.
c) Number of SV callers reporting SV:
Both delly and lumpy are run (no single tool is sufficient) and the calls combined using overlap definition above.
d) Overlap ENSEMBL exon or gene
e) LOWEST_PRIORITY entries
-Remaining variants that don't fall into above.

DEFAULT SORTING

The file is first divided into two main groups; SV calls found by 2 SV callers and SV calls by a single tool. Next SVs are divided into passing (**either novel or rare SVs AND <100kb AND >= 4 supporting read pairs**) followed by all other lower priority variants. The SVs are finally ordered by SVs overlapping exon, gene (but no exon), and no_gene. Example sorting:

SV Type	# SV callers	Exon Overlap	Gene Overlap	Passing SV	Priority
All	2	Yes	Yes	Yes	Highest
del/dup	2	No	Yes	Yes	Medium
All	1	Yes	Yes	Yes	Medium
All	2	Yes	Yes	No	Medium
tra/ins/inv	2	No	Yes	Yes	Low
All	2	No	No	Yes	Low
All	1	No	No	Either	Lowest
All	2	No	Yes	No	Lowest

PRIORITISING SVs

This is a big file with lots of information but there are ways to prioritise for the SVs most likely to be causal.

Cases likely to contain SVs of interest

1) SVs not reported in database of genomic variants → I recommend still examining these candidates using the ucsc genome browser as the DGV coordinates may differ slightly but still represent the same SV
2) SVs where breakpoints truncate exons → More likely to disrupt protein function
3) Translocations that result in candidate gene fusion products → Common mechanism in cancer by either creating novel protein or else placing promotor/enhancer adjacent to a different gene

COLUMN DEFINITIONS

A) **SV Caller Count**: Number of SV callers reporting the SV. Currently using delly and lumpy so will be 1 or 2

B) **Total SV calls:** Total SV calls by both tools. Sometimes >2 as SV calls within each SV caller are grouped together as well. This is typical for inversions and translocations where each side of the breakpoint is reported separately.

C) **Exon overlap:** Whether either SV breakpoint overlaps an ENSEMBL exon from the canonical transcript

D) **Gene overlap:** Whether either SV breakpoint overlaps an ENSEMBL gene (from start of first exon to last exon)

E) **Event ID:** Internal ID used to group SVs together

F) **chr_1:** Coordinate of first breakpoint

G) **coord1:** Genomic coordinate for first breakpoint (currently hs37d5).

H) **chr_2:** Coordinate of second breakpoint

I) **coord2:** Genomic coordinate for second breakpoint (currently hs37d5).

J) **sv_caller:** SV software reporting SV (delly, lumpy, or both)

K) **sv_type:** SV type either del, dup, tra, inv, or ins

L) **sv_id:** Internal SV caller unique ID

M) **quality:** SV quality (delly only) -> max value 60 and closer to 60 the more believable

N) **var_bases:** inserted variant bases (insertions called by delly only)

O) **length:** length of SV for deletions, duplications, and inversions.

P) **split_reads:** Number of read alignments spanning the SV (delly only)

Q) **paired_reads**: Number of read pairs supporting the SV (delly and lumpy)

R) **breakpoint_gene:** List of all genes found at breakpoints

S) **breakpoint_exon:** List of all exons found at breakpoints

T) **gene(hgnc):** HGNC gene name

U) **final_status:** Overall pass/fail status for the SV. Currently fail reasons are:
i) Not enough supporting reads
ii) Length_fail
iii) filter_dgv

V) **dgv_info:** Overlap to database of genomic variants. Either ‘novel’ or lists the event.

e.g. 1:153043412-153066335;gssvL5456;12/2601;0.46%
Lists chr : start_coord – end_coord : dgv ID ; reported_freq count and percent

W) **dgv_freq**: Frequency (0-1) from dgv entry. Inserted to allow easy sorting

X) **filter_dgv:** Final status for DGV, either N/A (for translocations and insertions which aren’t recorded), NOVEL (not reported), RARE_ALLELE (<=2% reported frequency), and FAIL (>2% reported frequency)

Y) **ensemble:** Link to ensembl gene

Z) **ens_canonical_trans:** canonical transcript from ENSEMBL. All variant information is relative to this transcript

AA) **protein_domains:** the source and unique identifier of any protein domains at that coordinate from the variant effect predictor.

AB) **uniprot**: uniprot gene name(s). This is a comma delimited list of all uniprot genes that map to this coordinate.

AC) **ccds**: ccds name(s). This is a comma delimited list of all ccds names that map to this coordinate

AD) **refseq**: refseq gene name(s). This is a comma delimited list of all refseq names that map to this coordinate

AE) **description:** description of gene function from hugo

AF) **omim**: link to omim entry if it exists. Note that these entries are linked by gene name and not coordinates so if there are discrepancies in the gene name this link will not be reported

AG) **pubmed:** link to pubmed ID(s) of publications that cite existing variant

AH) **phenotype:** phenotype information from mouse data from MGI if available

AI) **cosmic**: link to gene in cosmic (catalogue of somatic mutations in cancer)

AJ) **vogelstein**: indicates whether the gene is cited in vogelstein's landmark paper "Cancer genes and the pathways they control". If it is then the column contains information from the original vogelstein paper

AK) **go**: gene GO terms