**Instructions on how to interpret ClinSV results**

To obtain clinical reportable variants from automatically generated ClinSV variant calls, we suggest the following steps 1-4.

1. **Inspect the QC report** (file: *sample*.QC\_report.pdf).   
   Per sample a QC report is generated to assess if the input data and ClinSV results are within expectations obtained from 500 previously analyzed control samples. A detailed description of all metrics can be found at the end of the QC report.
   1. Make sure most metrics are within the expected range (green OK). Focus on individual metrics mostly for troubleshooting.
   2. The coverage by chromosome view (CBCV) should not show any pattern of low or high coverage regions at chromosome tips, which usually indicates an extreme GC bias. This can result in many false positive CNVnator calls and in missed variants.
   3. Look for aneuploidy in the CBCV. Such events may only be visible here.
   4. The CBCV also nicely highlights any larger rare CNVs (>500kb), which may be split up into smaller events in the automated call set.
2. **Short-list reportable variants** (file: *sample*.RARE\_PASS\_GENE.xlsx)  
   A table with rare gene affecting variants is generated containing all variants passing the variant detection threshold (FT=PASS Table 1) and having a population allele frequency ≤1%. Filtering the table by the following columns will enable to short list reportable candidate variants.
   1. Hide variants affecting genes not having a disease annotation (set PHEN to not blank). Shortlist variants matching the patient’s phenotype.
   2. Filter out variants affecting introns (set column GFEAT to != intron)
   3. Hide or sort by variants affecting a candidate gene (column CANDG, if file candGene.ids was provided).
   4. Focus first on variants with copy number change (CNV=1).
   5. If a pedigree file was provided, consider variants only present in affected individuals. (IA>0), and for autosomal dominant diseases filter for variants absent in unaffected individuals (IUA == 0).
   6. Consider the strength of evidence supporting the variant call. High confidence variant calls demonstrate high numbers of discordant pairs (DP) and split reads (SR). Due to repeats at breakpoint, DP/SR evidence can be reduced and some CNVs will have no supporting DP/SR. Unusual GC content, overlapping segmental duplications can be indicators of poor quality calls.
   7. Consider to decrease the threshold for rare variants by applying filters on population variant allele frequency columns PAFSU, PAFDRA, PAFV and PAF1KG

For a description of all column names, see Table 2. A reduced set of annotation columns can be found in *sample*.RARE\_PASS\_GENE.light.xlsx. The full list of variants can be found in files SV-CNV.txt or SV-CNV.vcf

1. **Visualize variants**Similar to SNV, candidate SV and CNV need to be visualized prior to reporting mainly in order to rule out false positive automated variant calls, visually confirm the altering effect on gene function and check if an individual variant is part of a more complex structural rearrangement.  
   First open IGV, then click on the “IGV” link in the excel file to load all tracks from the specified sample, and then click the “GOTO” links to jump to the region containing the variant, or the breakpoints. Clicking on the ClinSV variant will show its properties.  
   Check Table 3 for our recommended manual validation criteria and Figure 1 for a description of the genome browser tracks. Check if there are common overlapping events using the MGRB and DGV track. Note that DGV contains many disease specifc studies, click on a variant to obtain the PubMed ID of the underlying study.  
   To know whether all genome browser tracks are loaded correctly, compare your screen to Figure 1 showing a variant present in most people.

Troubleshooting:

Make sure the paths to the tracks are the same, as when running ClinSV, else tracks cannot be loaded from the IGV session xml file. The URL in the IGV hyperlink contains the path to the IGV session file. Check if you can access it from the terminal.  
Excel on Mac has a known issue: when you click the IGV or GOTO hyperlinks in Excel, it will popup the following message: ‘Cannot locate the Internet server or proxy server’. Whilst this is annoying, IGV does still update its position in response to clicking the link. Alternative spreadsheet programs, like OpenOffice or new Excel versions do not have this issue.

1. **Orthogonal validation**As for SNV it is recommended to confirm reportable SV/CNV through an orthogonal method such as MLPA or Sanger sequencing.

**Table 1** Automatically assignedSV call confidence criteria

|  |  |
| --- | --- |
| **Value** | **Criteria** |
| High | CNVs: length > 100kb or (length >10kb & AMQ>55)  All SVs: DP+SR > 10 & DP > 0 & SR > 0 |
| Pass | CNVs: DOC deviation >20% and (length > 10kb or (DP+SR) >= 2)  Balanced SVs: (DP+SR) >= 6 |
| Low | All other CNV and SV variants |

DOC depth of coverage, DR discordant pairs, SP split reads, AMQ average mapping quality

**Table 2** Column names of ClinSV’s output table containing rare gene affecting variants

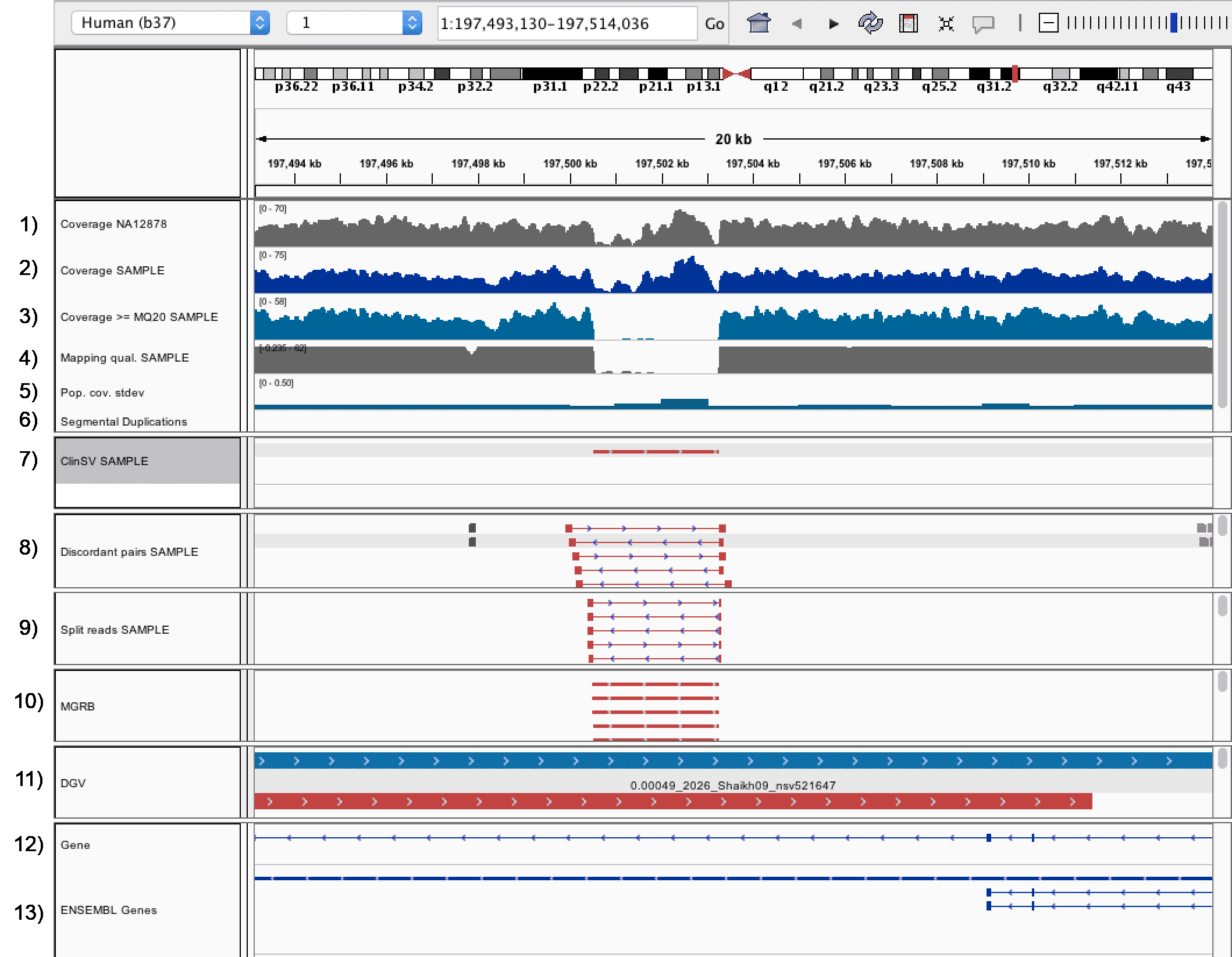
|  |  |
| --- | --- |
| **Column name** | **Description** |
| family 1, 2 | Family ID from ped file (only if ped file was present for analysis) |
| pedInfo 1, 2 | Patient initials |
| affected 1, 2 | 2=affected, 1= unaffected (from ped file) |
| IA1,2 | Number of times a variant was detected **i**n **a**ffected individuals |
| IUA1,2 | Number of times a variant detected **i**n **u**n**a**ffected individuals |
| SAMPLE 2 | Internal sample ID |
| ID 2 | Variant ID |
| FT 2 | Automated call confidence **f**il**t**er column. Values LOW, PASS, HIGH |
| RARE | Is variant **rare**? (1=yes, 0=no) Rare means that PAFV, PAFSU, PAFDRA and PAF1KG are ≤ 1% |
| SU 2 | **Su**m of discordant pairs and split reads supporting the variant |
| PAFSU | **P**opulation variant **a**llele **f**requency estimated from **su**m of discordant pairs (DP) and split reads (SR) in control cohort. Control samples consist of 500 healthy elderly individuals from the Medical Genome Reference Bank (MGRB). Formula  (DP+SR control) / (DP+SR in sample) / (number of control samples). |
| PE | Number of supporting discordant **p**airs (PE field inherited from Lumpy) |
| SR | Number of supporting **s**plit **r**eads |
| DRF 2 | Read **d**epth **r**atio of variant vs **f**lanking regions |
| DRA 2 | Read **d**epth **r**atio of variant vs the **a**verage genome wide coverage  Copy number = DRA x 2 |
| PAFDRA | **P**opulation variant **a**llele **f**requency estimated from normalized **DRA** in control samples (MGRB cohort) |
| PCSD | **P**opulation **c**overage **s**tandard **d**eviation of control cohort |
| GT | **G**eno**t**ype estimation |
| MQBP | Average read **m**apping **q**uality of reads supporting both **b**reak**p**oints |
| CNV 2 | Is the structural variant a **CNV**? 1 = yes, 0 = no. Yes if DRA or DRF <0.8 or >1.2 |
| IGV 2 | Link to load **IGV** session file.  IGV needs to be open for this to work, and only needs to be run once per session. |
| GOTO 2 | Link to **go to** the region containing the variant in IGV |
| LOCATION 2 | Genomic **location** (chr:start-end) |
| SVTYPE 2 | **S**tructural **v**ariant **type**: Deletion, duplication, inversion or break ends (BND). BNDs represent a pair of breakpoints and can represent a translocation.. |
| SVLEN 2 | **Len**gth of **s**tructural **v**ariant, in base pairs |
| TOOL 2 | Varaint detection **tool:** Lumpy and/or CNVnator |
| PAFV 2 | **P**opulation variant **a**llele **f**requency from **v**ariants in control. At the time of publication, this is 500 healthy individuals from the MGRB cohort (https://sgc.garvan.org.au/initiatives/mgrb). |
| PAF1KG | **P**opulation variant **a**llele **f**requency in **1000** **g**enome project |
| GC | **GC** content of the variant |
| CR | Size **r**atio of **c**ompressed vs. uncompressed reference sequence of the variant. Low complexity sequences have smaller compression ratios. |
| MQ | **A**verage read **m**apping **q**uality of the variant |
|  |  |
| SEGD | Overlapping **seg**mental **d**uplications published by Bailey JA et al. 2002. For best match: % variant coverage | % seg-dup coverage | identity | for all matching seg-dup’s: count | merged % variant coverage |
| NUMG 2 | **Num**ber of **g**enes affected by the variant |
| GENES 2 | ENSEMBL genes affected by the variant |
| GFEAT | **G**ene **feat**ure affected by structural variant. If multiple features or genes affected, one feature is reported in order of importance (start\_codon > stop\_codo > CDS > UTR > intron) |
| HPO | **HPO** numbers of affected genes. HPO’s of genes are separated by the “|” symbol and appear in the same order as the gene names in the GENES column. Multiple HPO’s per gene are separated by colon. |
| PHEN 2 | Known **phen**otypes for any genes affected by the variant, obtained from OMIM, DDG2P or Orphanet. If annotation from more than one source is available for a gene, only first source in above order is shown to reduce redundancy of terms. |
| CANDG 2 | Gene names that are also in the **cand**idate **g**ene list (if provided) |

1 If pedigree file was provided

2 minimum set of annotation columns (see light.xlsx)

**Table 3** Manual validation criteria applied when visualizing SVs and their evidence in the genome browser.

|  |  |
| --- | --- |
| **Value** | **Criteria** |
| Pass | * Diverse DP and SR mapping positions: mapping position of supporting DPs and SRs are randomly distributed up and downstream of the SV breakpoint * For CNVs a clear change in DOC at breakpoint: a sharp step shape change of read depth at start and end of the variant * Variant in region with high mapping quality (here MQ>=55): If reads are mapped with low confidence (low mapping quality), they are more likely wrongly placed in the genome resulting in DP/SP or DOC that can be misinterpreted as a variant by automated variant callers. * No competing evidence: e.g. no additional DP/SR linking to other parts of the genome than the predicted SV. |
| Needs further investigation | * Complex variant and competing evidence: Complex variants are composed by more than one adjacent or superposed SV and the effect on a gene can be hard to predict until the complete structure is resolved. Complex events can be harder to distinguish from mapping artifacts. * In close vicinity to common DP/SR: * CNV with avg. mapping quality <40 and no DPs or SRs |
| False call | * Few DP/SR , not being normally distribution around the breakpoint * Presence of same DP/SR in most control samples (high PAFSU) * Contradicting DP/SR evidence solely in region of low mapping quality, representing misaligned reads |



**Figure 1 Default genome browser tracks for manual validation**This region and tracks are display if the IGV session is loaded correctly. It shows a commonly deleted repeat sequence.

1. Depth of coverage of control sample NA12878
2. Depth of coverage of input sample
3. Depth of coverage of sequence read bases with a Phred scaled mapping quality >=20
4. Average mapping quality of aligned reads for current sample
5. Coverage standard deviation in 1kb windows of 500 healthy individuals (MGRB cohort)
6. Regions of segmental duplication as determined by Bailey JA et al. 2002 (Seg-Dup, none in this region)
7. Annotated ClinSV calls from Lumpy and CNVnator
8. Discordant pairs of input sample, followed by discordant pairs of control sample NA12878 (scroll down)
9. Split reads of input sample, followed by split reads of control sample NA12878
10. ClinSV variants from 500 healthy control samples (MGRB cohort) called in batches of 15 samples
11. Structural variants deposited in Database of Genomic Variants (DGV); feature name indicates the population allele frequency, study sample size, author and DGV ID.
12. RefSeq genes
13. ENSEMBL genes (used for annotation)