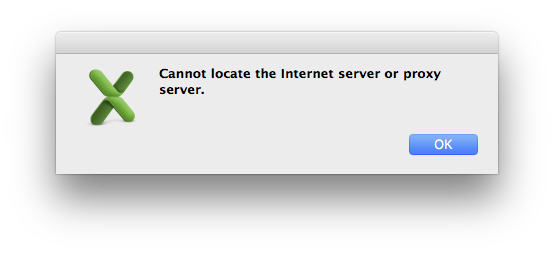
**Instructions on how to interpret ClinSV results**

Go to the results folder located in the project directory

1. Inspect the QC report (*sample*.QC\_report.pdf). A description of the metrics can be found at the end of the report.
2. Open the table containing rare gene affecting variants (i.e., *sample*.RARE\_PASS\_GENE.light.xlsx). These variants are all rare (ie ≤1% population allele frequency), either High quality, or passing all filters.  
   For a description of the column names, see Table 1 below.  
   The full set of annotations can be found in *sample*.RARE\_PASS\_GENE.xlsx
3. Inspect CNV (CNV=1) that have a known phenotype (see PHEN column), or that were included in the candidate gene list (see CANDG column), if this was provided.
4. If a pedigree file was provided, consider variants only present in affected individuals (IA), and not in unaffected individuals (IUA).
5. Consider the strength of evidence supporting the variant call, eg FT=High, GT matching the expected pattern (0/1, or 1/1 in affected), High values for SU, SE, PE demonstrate calls with strong support from split reads. Due to flanking repeats, not all CNV have split or discordant reads though. Unusual GC content, overlapping segmental duplications can be indicators of poor quality calls.
6. Consider how rare the variant is, by assessing PAFSU, PAFDRA, PAFV, PAF1KG
7. For visual validation, you can inspect candidate variants in IGV. First open IGV, Then click 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.   
   Check Table 2 for our recommended manual validation criteria and Figure 1 for a description of the tracks. By comparing to the MGRB and DGV track, check if the candidate variant is indeed rare.  
   To know whether all genome browser tracks are loaded correctly, compare your screen to Figure 1. 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.
8. The full list of variants can be found under in file: SV-CNV.txt or SV-CNV.vcf

**Troubleshooting, using Microsoft Excel on a Mac**

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 do not have this issue.

**Table 1 Description of each ClinSV column.**

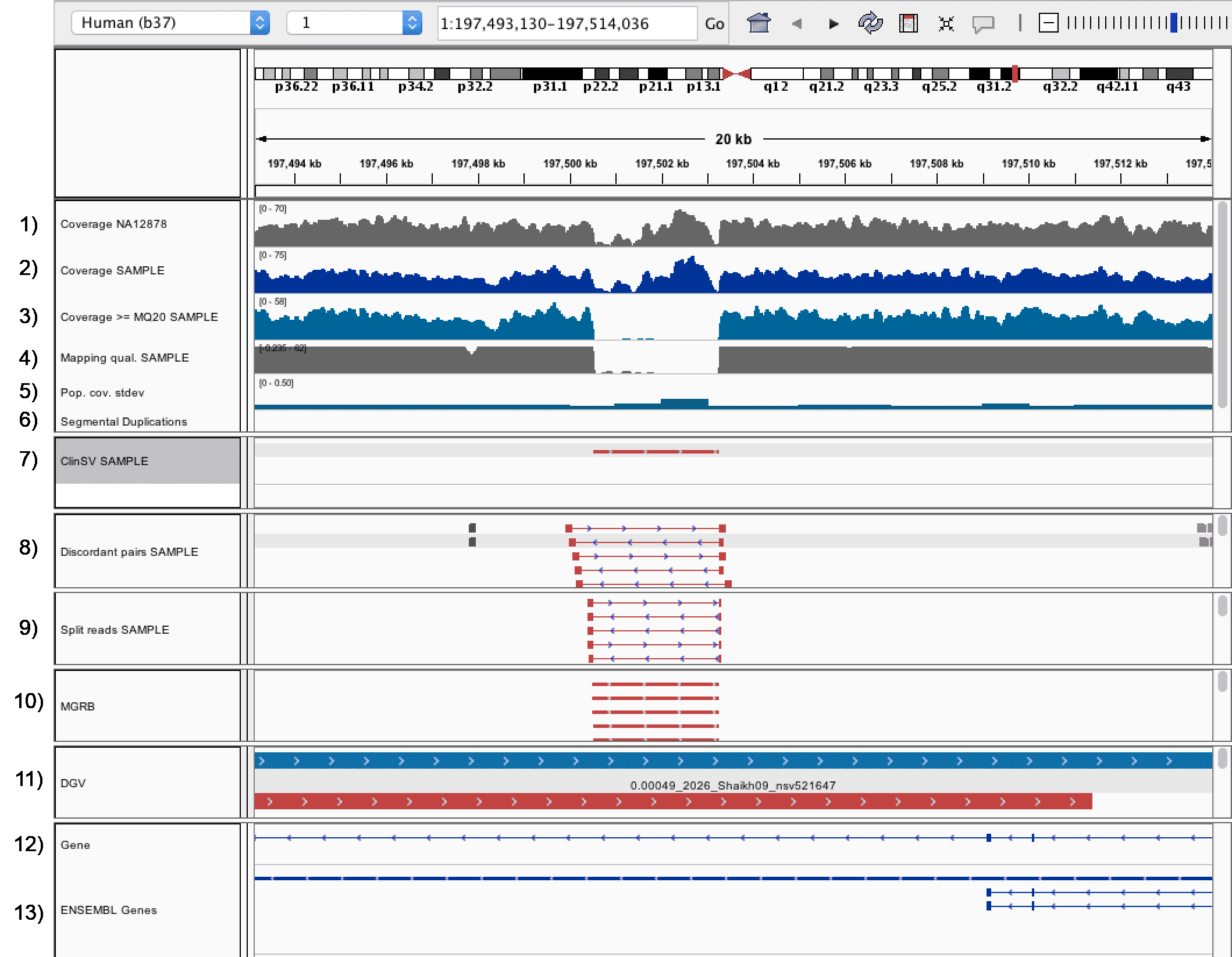
|  |  |
| --- | --- |
| **Column name** | **Description** |
| family 1, 2 | Family ID from ped file (only if ped file was present for analysis) |
| pedInfo 1, 2 | Patient initials |
| affected 1, 2 | 2=affected, 1= unaffected (from ped file) |
| IA1,2 | Number of times a variant was detected **i**n **a**ffected 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 mapping **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 |
| GDESC 2 | Name of the affected gene feature: start\_codon, stop\_codon, CDS, UTR, intron. If multiple genes are affected by a variant, it most severe feature in above order is shown. |
| 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 2 Manual validation criteria**

|  |  |
| --- | --- |
| **Value** | **Criteria** |
| Pass | * Diverse DP and SR mapping positions * For CNVs a clear change in DOC at breakpoint * Variant in region with high MQ * No competing evidence, eg, read orientation not supporting the observed SV type |
| Needs further investigation | * Complex variant / competing evidence * In close vicinity to common PE/SR * CNV with avg. MQ <40 and no DPs or SRs |
| False call | * Few PE/SR with same genomic coordinates * PE/SR evidence in region of low MQ |



**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)