ELIXIR hCNV Community: CNV annotations

Document context: v1
Object : First hCNV Community Implementation Study Implementation Study hCNV - Deliverable D5.1
Producer : Document produced by ELIXIR-HU, (Attila Gyenesei and colleagues, Katalin Monostory)
Review : David Salgado, Christophe Béroud, Michael Baudis
Status : reviewed & approved as deliverable & basis for future discussions (2019-12-31)

Main goal of CNV analysis is to gain functionally, regulatory and clinically relevant information (including clinical assessment, and research evidence tracks) and to provide annotations useful to interpret structural variants (SV) potential pathogenicity and filter out SVs that are potential false positives.

Different types of SV exist including deletions, duplications, insertions, inversions, translocations or more complex rearrangements. They can be either balanced or unbalanced. CNV should be described by coordinates on one chromosome, with the start and end positions of the SV (deletions, insertions, duplications).

CNV annotation can be carried out by finding the overlapping elements in different databases. The following features should be included in the analysis:

- Genes
- Promoters
- CNV prevalence resources (germline and cancer)
- Topologically Associating Domains (TAD)
- Phenotype
- Genotype
- Gene intolerance
- Haploinsufficiency

Genes

The gene annotation aims at providing information for the overlapping known genes with the SV in order to list the genes from a well annotated database. These annotations should include the definition of the genes and corresponding transcripts, the length of the coding sequence (CDS) and of the transcript, the location of the SV in the gene and the coordinates of the intersection between the SV and the transcript.

Promoters

The contribution of SV overlapping with promoters to disease etiology is well established, affecting gene expression, although understanding the consequences of these regulatory variants on the human transcriptome remains a major challenge. The report should contain the list of the genes whose promoters are overlapped by the SV.

CNV prevalence resources (germline and cancer)

Similarly to other variant calling analyses and reports, getting information on the incidence rate of the CNV in the population is essential. Several databases provide this from different populations and in various contexts (constitutional and somatic cases) .

The Database of Genomic Variants (DGV) provides SV defined as DNA elements with a size >50 bp. The content of DGV is only representing SV identified in healthy control samples from large cohorts published and integrated by the DGV team. The annotations give information on whether the identified SV is a rare or a common variant. This is useful in the subsequent filtering step.

The Progenetix resource (progenetix.org) provides genomic CNV frequency profiles for about 500 different cancer diagnoses, based the curation of currently more than 90'000 individual whole-genome cancer CNV profiles from raw data and published, sample specific annotations.

TAD annotation

The spatial organization of the human genome helps to accommodate the DNA in the nucleus of a cell and plays an important role in the control of the gene expression. In this non-random organization, topologically associating domains (TAD) emerge as a fundamental structural unit able to separate domains and define boundaries. Disruption of these structures especially by SV can result in gene misexpression.

Genotype and phenotype

Aim is to identify CNVs that overlap with functional genomic regions, and carry out genome-wide association analysis with gene expression and quantitative phenotypes. For example, OMIM (Online Mendelian Inheritance in Man) focuses on the relationship between phenotype and genotype.

Gene intolerance

Gene intolerance annotations provide the significance deviation from the observed and the expected number of variants for each gene:

- synZ = synonymous Z score
- misZ = missense Z score
- Positive Z scores indicate gene intolerance to variation.

- pLI score (indicates the probability that a gene is intolerant to a loss of function mutation - nonsense, splice acceptor and splice donor variants caused by SNV.)

Haploinsufficiency

Haploinsufficiency, wherein a single functional copy of a gene is insufficient to maintain normal function, is a major cause of dominant disease. As detailed in DECIPHER, over 17,000 protein coding genes have been scored according to their predicted probability of exhibiting haploinsufficiency:

- High ranks (e.g. 0-10%) indicate a gene is more likely to exhibit haploinsufficiency
- Low ranks (e.g. 90-100%) indicate a gene is more likely to NOT exhibit haploinsufficiency.

Most of these annotations provide additional information on each gene overlapped by a SV (independently of the genome build version).

There are already several freely available tools for CNV annotation. Here we would like to highlight only one of them as an example.

- ClinGen - the Clinical Genome Resource is an NIH funded program providing research and clinical genetic testing information from the National Center for Biotechnology Information (NCBI) and the National Library of Medicine (NLM). There are three ClinGen tracks available; Region Dosage Sensitivity (>60 variant regions), Gene Dosage Sensitivity (>1,200 variants), and Overlapping CNVs (>36,000 variants). You can get the following informations:
    - Name and Parent: NCBI's IDs for samples and sample's parent
    - CNV type: Loss or Gain
    - ClinVar Accession and Clinical Interpretation: ClinVar ID and pathogenicity
        - Benign
        - Missing
        - Likely Benign
        - Uncertain
        - Pathogenic
        - Likely Pathogenic
    - Phenotype and ID: Human Phenotype Ontology listed phenotypes and IDs
    - Gender, Var Origin, Zygosity: details of alleles origin
    - Copy Number: copy number associated with loss or gain
    - Inner Start/Stop: positions of each start and stop region for each CNV
    - Remap Score: NCBI's Remap score for multiple assemblies
        - Score around 1: region relatively unchanged between assemblies
        - >1: insertion in target assembly
        - <1: deletion in target assembly
    - Validated: CNV calls validated with additional methods
    - Sample Name and Sample Subset Name: describes any division of samples in the study

The databases mentioned below can provide the necessary information for a detailed CNV annotation. Recommended databases:

- 1kG Phase3 CNVs and Large Variants - https://www.internationalgenome.org/phase-3-structural-variant-dataset/ . Variants are collected from the same 2504 samples used for the variant frequency annotation. It contains descriptions of CNVs, allele frequencies, allele counts, heterozygous and homozygous counts in total and for each population. The frequencies for each subpopulation are also available for CNVs.

- ExAC XHMM CNV - https://gnomad.broadinstitute.org. The Exome Aggregation Consortium utilized XHMM for calling CNVs, and thus Golden Helix has curated the ExAC XHMM CNV Calls. ExAC CNVs provides a PHRED score representing the confidence of the XHMM call and can be used to filter your CNVs. This dataset spans 61,486 unrelated individuals and a CNV event from each individual represented independently. Again, ExAC has two source field groups; Summary of CNVs and Overlapping CNVs.

  The Overlapping CNVs fields:
  - Region and Span (Size): Genomic regions & total number of base pairs for annotated CNV
  - Similarity Coefficient: % overlap of CNV to annotated CNV
  - Type: type of CNV event
  - Population: individual with CNV's associated population
  - Quality Score: Phred Scaled Likelihood Score

- ClinVar CNVs and Large Variants - https://www.ncbi.nlm.nih.gov/clinvar/. ClinVar CNVs and Large Variants (>21,000 variants) contain all variants in the ClinVar database that are over 200 base pairs long. CNV information coming from ClinVar is also linked to NCBI's dbVar database. Within ClinVar, there is access to multiple IDs (allele, gene, HGNC, dbVar nsv/esv, MedGen phenotypeIDs, and others), study origin, and clinical significance. Overall, ClinVar provides data about phenotypes and supporting evidence for variants.

- DGV CNVs - http://dgv.tcag.ca/dgv/app/about?ref=. The Database of Genomic Variants seeks to provide a comprehensive summary of structural variation in the human genome. This database includes not only CNVs (greater than 1000 bp) but also insertions/deletions (InDels) (as low as 100 bp), and inversions/inversion breakpoints. The variants in this database were identified in healthy control samples only. Both tracks supply similar cohort/study information and hyperlinks for fields such as PubMedIDs, method descriptions, and cohort descriptions. The main difference between DGV Variants and Supporting Variants is the depth of information

- DECIPHER GENE ANNOTATIONS - https://decipher.sanger.ac.uk/ddd#overview. The Deciphering Developmental Disorders (DDD) Study (Firth, et al., 2011) has recruited nearly 14,000 children with severe undiagnosed developmental disorders,

and their parents from around the UK and Ireland. The patients have been deeply phenotyped by their referring clinician via DECIPHER using the Human Phenotype Ontology. The DNA from these children have been explored using high resolution exon-arrayCGH and exome sequencing (trio) to investigate the genetic causes of their abnormal development. These annotations give additional information on each gene overlapped by a SV (independently of the genome build version).

- OMIM (Online Mendelian Inheritance in Man) (Hamosh, et al., 2000) focuses on the relationship between phenotype and genotype. These annotations give additional information on each gene overlapped by a SV (independently of the genome build version).