Knockout Analysis

A gene is considered to be knocked out for a sample when there is a set of variants that disable each copy of a certain gene.

This analysis obtains the list of knocked out genes for each input sample.

A variant is considered to disable a gene depending on the biotype of the gene, and its annotated consequence type. In **protein_coding** genes, the consequence type must be any from the list of loss of function sequence ontology terms listed below. In genes with other biotypes, the consequence type is not checked. The variants must also match other filter quality criteria.

Loss of function consequence type:

- `frameshift_variant`
- `inframe_deletion`
- `inframe_insertion`
- `start_lost`
- `stop_gained`
- `stop_lost`
- `splice_acceptor_variant`
- `splice_donor_variant`
- `transcript_ablation`
- `transcript_amplification`
- `initiator_codon_variant`
- `splice_region_variant`
- `incomplete_terminal_codon_variant`

There are multiple scenarios where we can ensure that a set of variants are affecting all copies of the gene, therefore, the gene is knocked out.

- **Homozygous Alternate :** The sample presents the same knock-out variant in both copies of the chromosome.
- **Multi-allelic :** The sample presents two different knock-out variants in the same position.
- **Compound Heterozygous :** The sample presents two different variants at a particular gene, one on each chromosome of a pair, each of them inherited from different parents.
- **Structural Variation overlap :** The sample presents a large structural deletion overlapping with a knock out variant.

Table of Contents:

- Implementation
- Input
  - Parameters
- Output
  - Files
    - knockout.summary.json
    - knockout.gene.{gene}.json
    - knockout.sample.{individual}.{sample}.json

Implementation

Implemented at opencga#1455.

Input

Parameters

Parameters can be grouped in three categories:

- Samples selection
  - **sample** : List of samples to analyse. The analysis will produce a file for each sample.
- Genes selection
  If no parameter is provided to select genes, the analysis will consider all **protein_coding** genes.
  - **gene** : List of genes of interest. In combination with parameter **panel**, all genes will be used.
  - **panel** : List of panels of interest. In combination with parameter **gene**, all genes will be used.
  - **biotype** : List of biotypes. Used to filter by transcripts biotype.
- Variants filter

- **consequenceType** : Consequence type as a list of SequenceOntology terms. This filter is only applied on protein_coding genes. By default filters by loss of function consequence types.
- **filter** : List of accepted terms from the **FILTER** column from the VCF file.
- **qual** : Filter by **QUAL** column from the files.

Output

The analysis will produce one **JSON** file per sample with all knocked-out genes for that sample, one **JSON** file per gene with all samples with that gene knocked-out, and one summary **JSON** file with aggregated information.

Files

knockout.summary.json

See KnockoutByGene.java

....

knockout.gene.{gene}.json

See KnockoutByGene.java

- `gene`
    - `id`
    - `name`
    - `chromosome`
    - `start`
    - `end`
    - `biotypes`
    - `transcripts[]`
        - `id`
        - `chromosome`
        - `start`
        - `end`
        - `....`
- `stats`
    - `numIndividuals`
    - `numSamples`
    - byType : Map<KnockoutType, long>
- `individuals[]`
    - `id`
    - `disorders`
    - `phenotypes`
    - `samples[]`
        - `id`
        - `transcripts[] : KnockoutTranscript`
            - `id`
            - `chromosome`
            - `start`
            - `end`
            - `biotype`

        - `variants[] : KnockoutVariant`
            - `id`
            - `genotype`
            - `filter`
            - `qual`
            - `knockoutType : [HOM_ALT, COMP_HET, MULTI_ALLELIC, DELETION_OVERLAP]`
            - `sequenceOntologyTerms[]`

knockout.sample.{individual}.{sample}.json

See KnockoutBySample.java

- `individual`
    - `id`
    - `....`
    - `sample`
        - `id`
        - `phenotypes`
        - `...`
- `stats`
    - `numGenes`
    - `numTranscripts`
    - byType : Map<KnockoutType, long>
- `genes[] : KnockoutGene`
    - `id`
    - `name`
    - `biotype`
    - `status`
    - `chromosome`
    - `start`
    - `end`
    - `strand`
    - `source`
    - `description`
    - `transcripts[] : KnockoutTranscript`
        - `id`
        - `chromosome`
        - `start`
        - `end`
        - `biotype`
        - `variants[] : KnockoutVariant`
            - `id`
            - `genotype`
            - `filter`
            - `qual`

- knockoutType : [HOM_ALT, COMP_HET, MULTI_ALLELIC, DELETION_OVERLAP]
- sequenceOntologyTerms[]

Example of knockout_genes.{sample}.json


```
{
  "sample" : "NA19600",
  "genesCount" : 5,
  "transcriptsCount" : 15,
  "countByType" : {
    "homAltCount" : 15,
    "multiAllelicCount" : 2,
    "compHetCount" : 0,
    "deletionOverlapCount" : 0
  },
  "genes" : [ {
    "id" : "ENSG00000186470",
    "name" : "BTN3A2",
    "transcripts" : [ {
      "id" : "ENST00000377708",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "6:26370605:T:C" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
    }, {
      "id" : "ENST00000508906",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "6:26370605:T:C" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
    }, {
      "id" : "ENST00000356386",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "6:26370605:T:C" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
    }, {
      "id" : "ENST00000527422",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "6:26370605:T:C" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
    }, {
      "id" : "ENST00000396948",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "6:26370605:T:C" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
    }, {
      "id" : "ENST00000527639",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "6:26370605:T:C" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
    }, {
      "id" : "ENST00000396934",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "6:26370605:T:C" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
```

```
    } ]
  }, {
    "id" : "ENSG00000198919",
    "name" : "DZIP3",
    "transcripts" : [ {
      "id" : "ENST00000361582",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "3:108634973:C:A" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
    }, {
      "id" : "ENST00000463306",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "3:108634973:C:A" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
    }, {
      "id" : "ENST00000479138",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "3:108634973:C:A" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
    } ]
  }, {
    "id" : "ENSG00000215182",
    "name" : "MUC5AC",
    "transcripts" : [ {
      "id" : "ENST00000621226",
      "biotype" : "protein_coding",
      "homAltVariants" : [ ],
      "multiAllelicVariants" : [ "11:1158073:T:C", "11:1158073:T:-" ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
    } ]
  }, {
    "id" : "ENSG00000147874",
    "name" : "HAUS6",
    "transcripts" : [ {
      "id" : "ENST00000380496",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "9:19058483:C:A" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
    }, {
      "id" : "ENST00000380502",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "9:19058483:C:A" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
    } ]
  }, {
    "id" : "ENSG00000099937",
    "name" : "SERPIND1",
    "transcripts" : [ {
      "id" : "ENST00000215727",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "22:20780030:-:C" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
      "deletionOverlapVariants" : [ ]
    }, {
      "id" : "ENST00000406799",
      "biotype" : "protein_coding",
      "homAltVariants" : [ "22:20780030:-:C" ],
      "multiAllelicVariants" : [ ],
      "compHetVariants" : [ ],
```

```
      "deletionOverlapVariants" : [ ]
    } ]
  } ]
}
```