Annotating and prioritizing SNVs

Practical

Pre-requisites

- Check your folder for annovar, we will need the following scripts:
  - convert2annovar.pl
  - annotate_variation.pl
  - table_annovar.pl
- There should be a database folder humandb that should contain already some databases
  - Note: these have been downloaded for you with annovar_commands.sh
- You should have your filtered vcf file from the mutation caller ready

Preparation

- Annovar uses their own format for the input files
- Generate this file from the vcf using:

```
$ software/annovar/convert2annovar.pl \
      -format vcf4old  \
      path_to_your_input.vcf.gz \
      > path_to_your_output.avinput;
```

- Output:

```
C02Q11SYFVH6:CRUK_summerschool perner01$ head tmp/HCC1143_vs_HCC1143_BL.annot.muts.avinput
1       10150   10150   C       T       unknown .       449
1       10180   10180   T       C       unknown .       270
1       10241   10241   T       A       unknown .       172
1       10291   10291   C       T       unknown .       191
1       10315   10315   C       G       unknown .       251
1       10348   10348   A       C       unknown .       395
1       10354   10354   C       A       unknown .       574
1       10357   10357   T       C       unknown .       560
1       10394   10394   T       A       unknown .       624
1       10440   10440   C       A       unknown .       600
```

Gene-based annotation

- Annovar performs gene-based annotation as default
- Will generate at once annotation
  - with respect to genes and
  - with respect to functional effect on coding sequence

```
$ software/annovar/annotate_variation.pl \
      --buildver hg19 \
      path_to_your_annovar_file.avinput \
      path_to_your_db_folder;
```

- Output:

```
C02Q11SYFVH6:CRUK_summerschool perner01$ ls tmp/
HCC1143_vs_HCC1143_BL.annot.muts.avinput
HCC1143_vs_HCC1143_BL.annot.muts.avinput.exonic_variant_function
HCC1143_vs_HCC1143_BL.annot.muts.avinput.log
HCC1143_vs_HCC1143_BL.annot.muts.avinput.variant_function
```

Gene-based annotation

- Output:
  - variant_function file:

```
C02Q11SYFVH6:CRUK_summerschool perner01$ head tmp/HCC1143_vs_HCC1143_BL.annot.muts.avinput.variant_function
intergenic      NONE(dist=NONE),DDX11L1(dist=1724)      1       10150   10150   C       T       unknown .       449
intergenic      NONE(dist=NONE),DDX11L1(dist=1694)      1       10180   10180   T       C       unknown .       270
intergenic      NONE(dist=NONE),DDX11L1(dist=1633)      1       10241   10241   T       A       unknown .       172
intergenic      NONE(dist=NONE),DDX11L1(dist=1583)      1       10291   10291   C       T       unknown .       191
intergenic      NONE(dist=NONE),DDX11L1(dist=1559)      1       10315   10315   C       G       unknown .       251
intergenic      NONE(dist=NONE),DDX11L1(dist=1526)      1       10348   10348   A       C       unknown .       395
intergenic      NONE(dist=NONE),DDX11L1(dist=1520)      1       10354   10354   C       A       unknown .       574
intergenic      NONE(dist=NONE),DDX11L1(dist=1517)      1       10357   10357   T       C       unknown .       560
intergenic      NONE(dist=NONE),DDX11L1(dist=1480)      1       10394   10394   T       A       unknown .       624
intergenic      NONE(dist=NONE),DDX11L1(dist=1434)      1       10440   10440   C       A       unknown .       600
```

Variant regions

Distant to closest gene or overlapping gene

Variant details

Gene-based annotation

- Output:
  - exonic_variant_function file:

```
C02Q11SYFVH6:CRUK_summerschool perner01$ head tmp/HCC1143_vs_HCC1143_BL.annot.muts.avinput.exonic_variant_function
line1074        nonsynonymous SNV       CHD5:NM_015557:exon9:c.A1331C:p.N444T,  1       6208966 6208966 T       G       unknown .
line1103        synonymous SNV  ESPN:NM_031475:exon10:c.C2217T:p.L739L, 1       6512048 6512048 C       T       unknown .       25
line1107        stopgain        TNFRSF25:NM_001039664:exon4:c.G370T:p.E124X,TNFRSF25:NM_003790:exon4:c.G370T:p.E124X,TNFRSF25:NM_1
3:c.G235T:p.E79X,TNFRSF25:NM_148965:exon4:c.G370T:p.E124X,      1       6524705 6524705 C       A       unknown .       229
line1148        nonsynonymous SNV       PER3:NM_001289862:exon19:c.G3046A:p.A1016T,PER3:NM_016831:exon19:c.G3019A:p.A1007T,PER3:NM
53      7890053 G       A       unknown .       92
line1149        synonymous SNV  PER3:NM_001289862:exon19:c.T3048A:p.A1016A,PER3:NM_016831:exon19:c.T3021A:p.A1007A,PER3:NM_0012898
55      T       A       unknown .       95
line1224        nonsynonymous SNV       SLC25A33:NM_032315:exon5:c.G464A:p.R155Q,       1       9633452 9633452 G       A       un
line1335        synonymous SNV  PRAMEF1:NM_023013:exon4:c.C1359T:p.G453G,PRAMEF1:NM_001294139:exon2:c.C624T:p.G208G,    1       12
wn      .       132
line1347        nonsynonymous SNV       PRAMEF2:NM_023014:exon4:c.T1330G:p.F444V,       1       12921539        12921539        T
line1454        nonsynonymous SNV       C1orf64:NM_178840:exon2:c.A325C:p.T109P,        1       16332656        16332656        A
line1531        unknown UNKNOWN 1       16907947        16907947        C       T       unknown .       861
```

Line of input file

Type of exonic variant

Gene and effect details

Variant details

Exercises

- Check how many variants/what percentage of variants fall in intergenic or exonic regions?
- What is the most common exonic variant type?
- Which variants affect your favourite gene (e.g. TP53)?

Region-based annotation

- Uses same script but we need to set two more parameters:

```
$ software/annovar/annotate_variation.pl \
    -regionanno \
    -build hg19 \
    -dbtype region_dbname \
    path_to_your_annovar_file.avinput \
    path_to_your_db_folder;
```

- Options for region databases, are for example:
  - cytoband, wgRna, phastConsElements46way, tfbsConsSites, gwasCatalog, genomicSuperDups
  - See also: http://annovar.openbioinformatics.org/en/latest/user-guide/region/

Exercises

- Is there a transcription factor whose binding sites are often hit by mutations?
- Has any of the variants been found as being associated with cancer?
- How many variant should we treat we caution because they fall into segmental duplications?

Filter-based annotation

- Uses same script but we need to change two parameters:

```
$ software/annovar/annotate_variation.pl \
      -filter \
      -build hg19 \
      -dbtype filter_dbname \
      path_to_your_annovar_file.avinput \
      path_to_your_db_folder;
```

- Options for region databases, are for example:
  - snp138, 1000g2015aug, cosmic70, ljb23_sift, …
- Output:

```
HCC1143_vs_HCC1143_BL.annot.muts.avinput.hg19_snp138_dropped
HCC1143_vs_HCC1143_BL.annot.muts.avinput.hg19_snp138_filtered
```

Exercises

- How many SNVs would you filter based on dbSNP?
- How many based on Cosmic?
- Which variants are probably deleterious according to the SIFT score (score of less than 0.05)?

All at once

```
$ software/annovar/table_annovar.pl \
      -buildver hg19 \
      -out path_to_outfile.annovar \
       -remove \
       -protocol refGene,cytoBand,gwasCatalog,
genomicSuperDups,snp138s,cosmic70,nci60,ljb23_sift \
       -operation g,r,r,r,r,f,f,f,f \
       -nastring NA \
       -csvout \
      path_to_your_annovar_file.avinput \
      path_to_your_db_folder
```

Exercise

- Select variants that…
  - are exonic and
  - are non-synonymous and
  - are deleterious according to SIFT and
  - have been found to be mutated in breast cancer and
  - have not been reported in snpdb