# Variant calling and analysis

## From reads to diagnosis

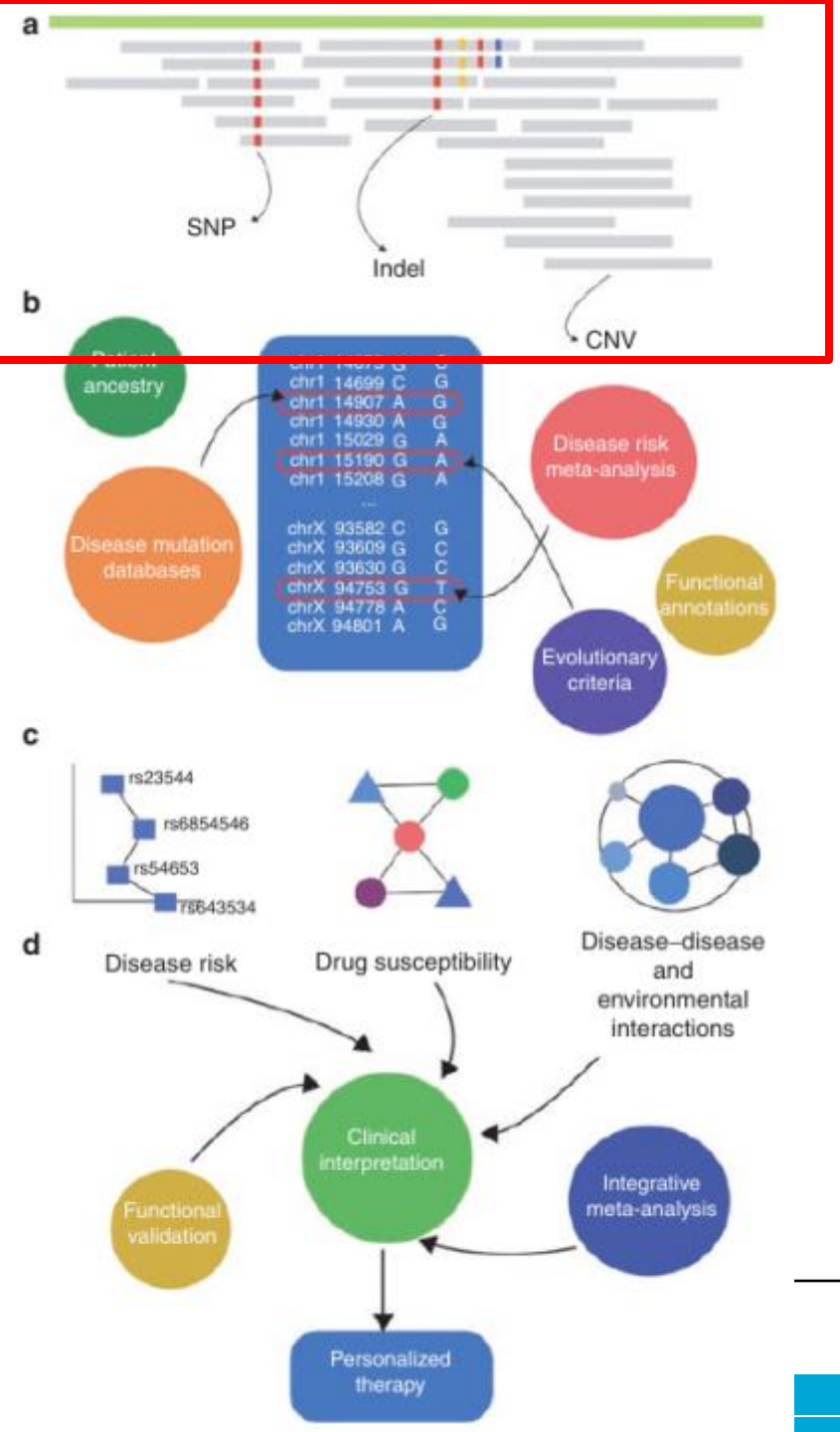
# Overview

Variant calling

Variant filtering

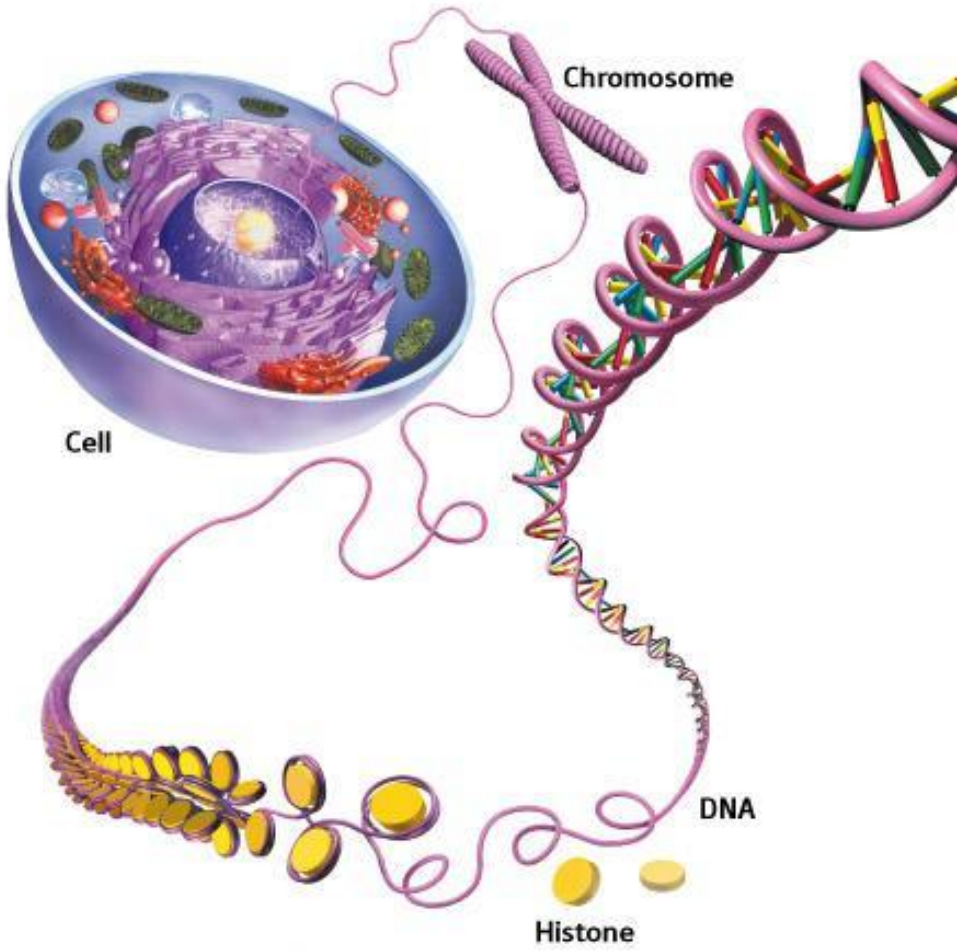
Variant relations

Variant visualization



# Reading the genome

## DNA in living cells



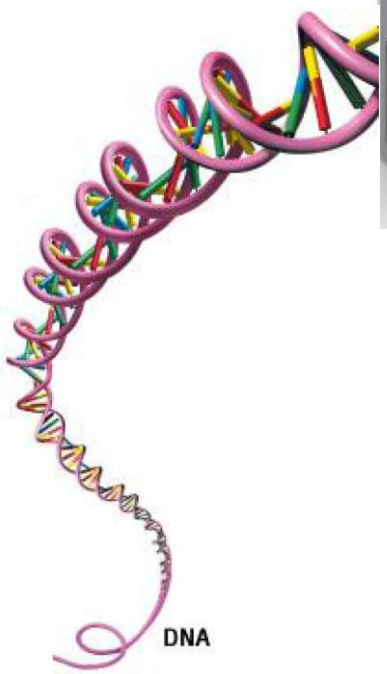
# Sequencing

## As means of reading out the genome


random fragmentation
and readout

GCGACAGCTTCCATCGCAGC
CACCAGCCGGGGATCAAAAC
CGGAGAACCTGCACTTTTTCC
GCGCGTCTCCCGGAGGCGCT
GGCGGGATTTGAGGACGTGC
TCGGCCGGGCAGCAGCGC.....

Reads
±50-1000 letters



# Variant calling

## Requires a reference genome

Read depth: ~60
~180 billion letters!

CCCTGCGCCGCGTGCGCGACAGCTTCCATCGCAGCCTG
CTGTGGATAGGACACCAGCCGGGGATCAAAAC
CCGCCTGACGGCGCGGGAGAACCTGCACTTTTTCCACC
CCGGCGACGGCGCGCGTCTCCCGGAGGCGCTG
GCGCAGGCCGGGCTGGCGGGATTTGAGGACGTGCCGG
TCGCTCAGCTCTCGGCCGGGCAGCAGCGCCGGG.....

23 chromosomes

Reference sequence

# Variant calling

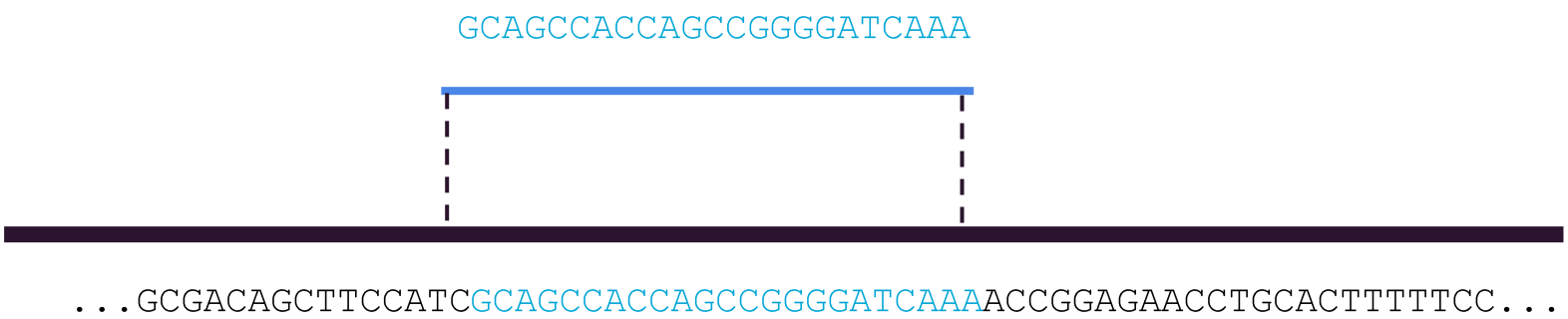
## Requires a reference sequence

- Coordinate system needed
  - for annotations (where are the genes)
  - for reporting variants
- Currently build 38 (GRCh38)
  - 20th version (Hg20)
  - 350 gaps left
  - Based mostly on a few anonymous donors (from Buffallo, NY)
- Many still use Hg19 or Hg18

# Mapping read to a reference (1)

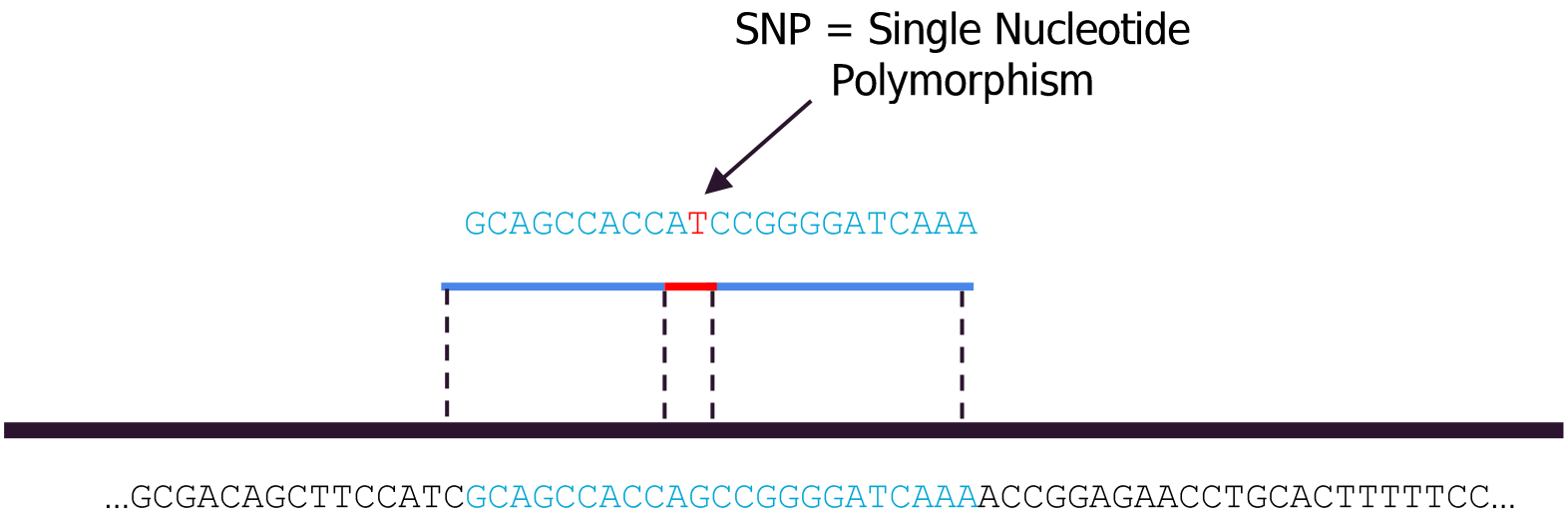
## allows for detecting variation

Perfect match!



# Mapping read to a reference (3)

## allows for detecting variation

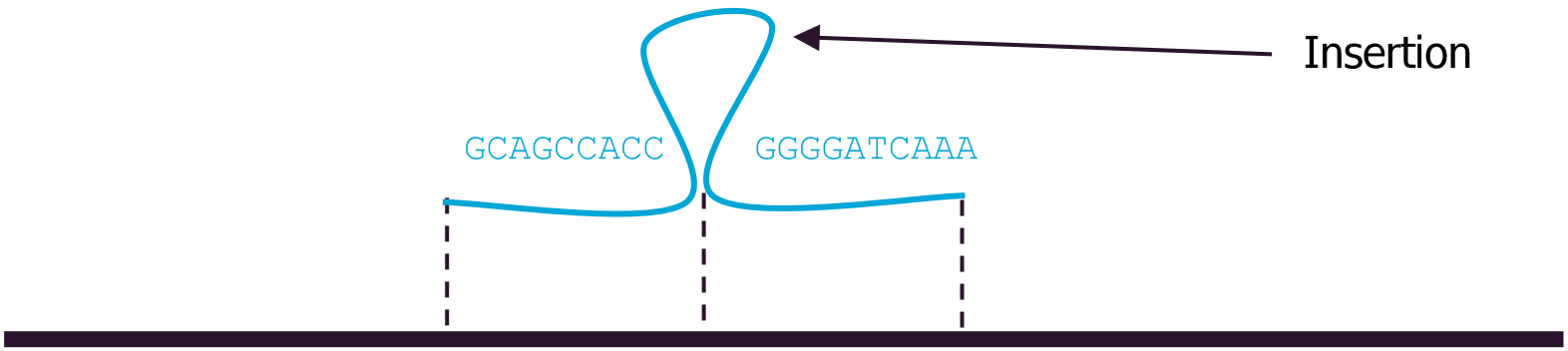


# Mapping read to a reference (4)

## allows for detecting variation

# Mapping read to a reference (5)

## allows for detecting variation




...GCGACAGCTTCCATCGCAGCCACCAGCCGGGGATCAAAACCGGAGAACCTGCACTTTTTCC...

# Types of variants

| | | |
|---|---|---|
| Single nucleotide variant | ATTGGCCTTAACCCCCGATTATCAGGAT<br>ATTGGCCTTAACCTCCGATTATCAGGAT | |
| Insertion–deletion variant | ATTGGCCTTAACCCGATCCGATTATCAGGAT<br>ATTGGCCTTAACCC---CCGATTATCAGGAT | Structural variants |
| Block substitution | ATTGGCCTTAACCCCCGATTATCAGGAT<br>ATTGGCCTTAACAGTGGATTATCAGGAT | |
| Inversion variant | ATTGGCCTTAACCCCCGATTATCAGGAT<br>ATTGGCCTTCGGGGGTTATTATCAGGAT | |
| Copy number variant | ATTGGCCTTAGGCCTTAACCCCCGATTATCAGGAT<br>ATTGGCCTTA-------ACCTCCGATTATCAGGAT | |

# True variant or a sequencing error?

## Statistics helps!

```
Reference: GCGACAGCTTCAATCGCAGCCACCAGCCGGGGATCAAAACCGGAGAACCTGCACTTTTTCC
Sample:    GCGACAGCTTCGATCGCAGCCACCAGCCGGGGATCAAAACCGGAGAACCTGCACTTTTTCC
Read 1        ACAGCTTCGATCGCAGCCACCAG
Read 2            CTTCGATCGCAGCCACCAGCCGGG
Read 3               CGATCGCAGCCACCAGCCGGGGATC
Read 4                           CCAGCCGGGGATCAAAACCGGA
Read 5                                 GGGGATCAAAACCGGAGAACAT
Read 6                                      TCAAAACCGGAGAACCTGCACTTTT
Read 7                                                AGAACCTGCACTTTTTCC
```

# Variant Call Format (VCF) file

- Reports changes w.r.t. to reference genome
- Contains usually millions of variants
- Describes for each variant:
  - Reference sequence (e.g. 'A')
  - Alternate sequence (e.g. 'G')
  - Genotype (homozygote/heterozygote for variant)
  - Quality score
  - and more…

# Variant Call Format (VCF) file

- The first 25 lines of a VCF file of a trio (father, mother, child):

```
##fileformat=VCFv4.0
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/reference/ancestral_alignments/READM
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=HM2,Number=0,Type=Flag,Description="HapMap2 membership">
##INFO=<ID=HM3,Number=0,Type=Flag,Description="HapMap3 membership">
##reference=human_b36_both.fasta
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##INFO=<ID=GP,Number=1,Type=String,Description="GRCh37 position(s)">
##INFO=<ID=BN,Number=1,Type=Integer,Description="First dbSNP build #">
##INFO=<ID=NR,Number=0,Type=Flag,Description="No dbSNP 132 map weight=1 rs number assigned to position">
##INFO=<ID=OR,Number=1,Type=String,Description="Previous rs number">
##INFO=<ID=MP,Number=0,Type=Flag,Description="Maps to multiple positions on GRCh37">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  NA12891 NA12892 NA12878
1       52066   rs28402963      T       C       .       PASS    AA=C;DP=84;GP=1:62203;BN=125    GT:GQ:DP        1/0:44:23       1/0:43:20       1/0:70:36
1       695745  rs72631875      G       A       .       PASS    AA=.;DP=124;GP=1:705882;BN=130  GT:GQ:DP        1|0:100:34      0|0:62:20       1|0:100:56
1       742429  rs3094315       G       A       .       PASS    AA=g;DP=132;HM2;GP=1:752566;BN=103      GT:GQ:DP        1|1:100:38      1|1:59:30       1|1:100:44
1       742584  rs3131972       A       G       .       PASS    AA=a;DP=160;HM3;GP=1:752721;BN=103      GT:GQ:DP        1|1:100:50      1|1:100:33      1|1:100:60
1       744366  rs3115859       G       A       .       PASS    AA=g;DP=127;GP=1:754503;BN=103  GT:GQ:DP        1|1:80:31       1|1:100:34      1|1:100:45
1       746243  rs3131963       T       A       .       PASS    AA=t;DP=105;GP=1:756380;BN=103  GT:GQ:DP        1|1:52:29       1|1:43:24       1|1:100:56
1       746775  rs6699990       A       G       .       PASS    AA=N;DP=120;GP=1:756912;BN=116  GT:GQ:DP        0|1:100:34      0|0:89:30       0|0:100:46
1       747503  rs3115853       G       A       .       PASS    AA=-;DP=113;GP=1:757640;BN=103  GT:GQ:DP        1|1:100:37      1|1:61:25       1|1:100:27
1       747597  rs4951929       C       T       .       PASS    AA=c;DP=129;GP=1:757734;BN=111  GT:GQ:DP        1|1:100:36      1|1:86:28       1|1:100:56
1       747799  rs4951862       C       A       .       PASS    AA=c;DP=97;GP=1:757936;BN=111   GT:GQ:DP        1|1:67:26       1|1:44:13       1|1:100:62
```

↑ Father ↑ Mother ↑ Daughter

# Variant Call Format (VCF) file

```
##fileformat=VCFv4.0
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/reference/ancestral_alignments/READM
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=HM2,Number=0,Type=Flag,Description="HapMap2 membership">
##INFO=<ID=HM3,Number=0,Type=Flag,Description="HapMap3 membership">
##reference=human_b36_both.fasta
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##INFO=<ID=GP,Number=1,Type=String,Description="GRCh37 position(s)">
##INFO=<ID=BN,Number=1,Type=Integer,Description="First dbSNP build #">
##INFO=<ID=NR,Number=0,Type=Flag,Description="No dbSNP 132 map weight=1 rs number assigned to position">
##INFO=<ID=OR,Number=1,Type=String,Description="Previous rs number">
##INFO=<ID=MP,Number=0,Type=Flag,Description="Maps to multiple positions on GRCh37">
#CHROM  POS     ID      REF
1       52066   rs28402963
1       695745  rs72631875
1       742429  rs3094315
1       742584  rs3131972
1       744366  rs3115859
1       746243  rs3131963
1       746775  rs6699990
1       747503  rs3115853
1       747597  rs4951929
1       747799  rs4951862
```

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO |
|---|---|---|---|---|---|---|---|
| 1 | 52066 | rs28402963 | | T | C | . | PASS |
| 1 | 695745 | rs72631875 | | G | A | . | PASS |
| 1 | 742429 | rs3094315 | | G | A | . | PASS |
| 1 | 742584 | rs3131972 | | A | G | . | PASS |
| 1 | 744366 | rs3115859 | | G | A | . | PASS |
| 1 | 746243 | rs3131963 | | T | A | . | PASS |
| 1 | 746775 | rs6699990 | | A | G | . | PASS |
| 1 | 747503 | rs3115853 | | G | A | . | PASS |
| 1 | 747597 | rs4951929 | | C | T | . | PASS |
| 1 | 747799 | rs4951862 | | C | A | . | PASS |

# dbSNP

- Variant database
  - Assign identifiers to variants
    - Remain constant across different builds of reference genome
  - Population frequencies

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO |
|---|---|---|---|---|---|---|---|
| 1 | 52066 | rs28402963 | | T | C | . | PASS |
| 1 | 695745 | rs72631875 | | G | A | . | PASS |
| 1 | 742429 | rs3094315 | | G | A | . | PASS |
| 1 | 742584 | rs3131972 | | A | G | . | PASS |
| 1 | 744366 | rs3115859 | | G | A | . | PASS |
| 1 | 746243 | rs3131963 | | T | A | . | PASS |
| 1 | 746775 | rs6699990 | | A | G | . | PASS |
| 1 | 747503 | rs3115853 | | G | A | . | PASS |
| 1 | 747597 | rs4951929 | | C | T | . | PASS |
| 1 | 747799 | rs4951862 | | C | A | . | PASS |

## Reference SNP (refSNP) Cluster Report: rs3131972

| RefSNP | |
|---|---|
| Organism: | human (Homo sapiens) |
| Molecule Type: | Genomic |
| Created/Updated in build: | 103/138 |
| Map to Genome Build: | 37.5/Weight 1 |
| Validation Status: | |

| Allele | |
|---|---|
| Variation Class: | SNV: single nucleotide variation |
| RefSNP Alleles: | C/T (REV) |
| Allele Origin: | |
| Ancestral Allele: | T |
| Clinical Channel: | unknown |
| Clinical Significance: | NA |
| MAF/MinorAlleleCount: | A=0.321/700 |
| MAF Source: | 1000 Genomes |

| HGVS Names |
|---|
| NC_000001.10:g.752721A>G |
| NT_004350.19:g.231353A>G |
| XR_108280.1:n.-30A>G |

Links

SNP Details are organized in the following sections:

GeneView | Map | Submission | Fasta | Resource | Diversity | Validation

### Integrated Maps (Hint: click on 'Chr Pos' or 'Contig Pos' column value to see variation in NCBI sequence viewer)

| Assembly | Genome Build | Chr | Chr Pos | Contig | Contig Pos | SNP to Chr | Contig allele | Contig to Chr | Neighbor SNP | Map Method |
|---|---|---|---|---|---|---|---|---|---|---|
| GRCh37.p10 | 104.0 | 1 | 752721 | NT_004350.19 | 231353 | Rev | A | Fwd | view | remap |
| NCBI36 | 36.3 | 1 | 742584 | NT_004350.18 | 231353 | Rev | A | Fwd | view | blast |
| HuRef | 36.3 | 1 | 21185 | NW_001838585.1 | 1938 | Rev | G | Fwd | view | blast |
| HuRef | 104.0 | 1 | 21185 | NW_001838585.1 | 1938 | Rev | G | Fwd | view | remap |
| CHM1_1.0 | 104.0 | 1 | 740191 | NW_004077988.1 | 227600 | Rev | G | Fwd | view | remap |

### GeneView

**GeneView via analysis of contig annotation:** FAM87B *family with sequence similarity 87, member B*

*View more variation on this gene (click to hide).*

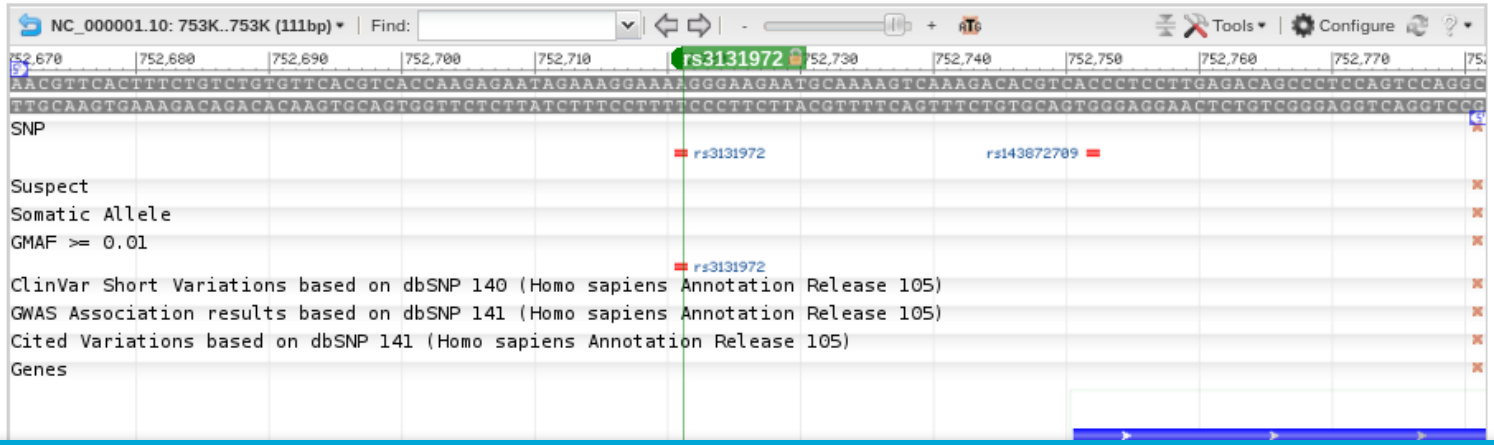
☑ Clinical Source: ○ in gene region ◉ cSNP ○ has frequency ○ double hit Go

Primary Assembly Mapping

| Assembly | SNP to Chr | Chr | Chr position | Contig | Contig position | Allele |
|---|---|---|---|---|---|---|
| GRCh37.p10 | Rev | 1 | 752721 | NT_004350.19 | 231353 | A |

Gene Model(s)

| Function | mRNA | | | | Protein | | |
|---|---|---|---|---|---|---|---|
| | SNP to mRNA | Accession | Position | Allele change | Accession | Position | Residue change |
| nearGene-5 | NA | XR_108280.1 | NA | NA ➡ NA | NA | NA | NA |



## Population Diversity (in rs orientation) Note: rs3131972 allele is reverse to the genome

| | Sample Ascertainment | | | | Genotype Detail | | | | Alleles | |
|---|---|---|---|---|---|---|---|---|---|---|
| ss# | Population | Individual Group | Chrom. Sample Cnt. | Source | C/C | C/T | T/T | HWP | C | T |
| ss118438193 | YRI | | 2 | IG | 1.000 | | | | 1.000 | |
| ss138899069 | ENSEMBL_Venter | | 2 | IG | 1.000 | | | | 1.000 | |
| ss162980826 | YRI | Sub-Saharan African | 2 | IG | 1.000 | | | | 1.000 | |
| ss163702698 | CEU | European | 2 | IG | 1.000 | | | | 1.000 | |
| ss165981005 | PGP | | 2 | IG | | 1.000 | | | 0.500 | 0.500 |
| ss197885385 | BUSHMAN_POP2 | | 2 | IG | | 1.000 | | | 0.500 | 0.500 |
| | BANTU | | 2 | IG | | 1.000 | | | 0.500 | 0.500 |
| ss218190360 | pilot_1_YRI_low_coverage_panel | | 118 | AF | | | | | 0.305 | 0.695 |
| ss230395425 | pilot_1_CEU_low_coverage_panel | | 120 | AF | | | | | 0.858 | 0.142 |
| ss238114952 | pilot_1_CHB+JPT_low_coverage_panel | | 120 | AF | | | | | 0.742 | 0.258 |
| ss78643137 | HapMap-CEU | European | 226 | IG | 0.681 | 0.301 | 0.018 | 0.439 | 0.832 | 0.168 |
| | HapMap-HCB | Asian | 86 | IG | 0.488 | 0.488 | 0.023 | 0.150 | 0.733 | 0.267 |
| | HapMap-JPT | Asian | 172 | IG | 0.547 | 0.407 | 0.047 | 0.439 | 0.750 | 0.250 |
| | HapMap-YRI | Sub-Saharan African | 226 | IG | 0.044 | 0.345 | 0.611 | 1.000 | 0.217 | 0.783 |
| | HAPMAP-ASW | African-American | 98 | IG | 0.143 | 0.531 | 0.327 | 0.527 | 0.408 | 0.592 |
| | HAPMAP-CHB | Asian | 82 | IG | 0.683 | 0.317 | | 0.527 | 0.841 | 0.159 |
| | HAPMAP-CHD | Chinese-Americans | 170 | IG | 0.694 | 0.271 | 0.035 | 0.752 | 0.829 | 0.171 |
| | HAPMAP-GIH | Indian-Americans | 176 | IG | 0.557 | 0.420 | 0.023 | 0.100 | 0.767 | 0.233 |
| | HAPMAP-LWK | Luhya, Kenya | 180 | IG | 0.122 | 0.456 | 0.422 | 1.000 | 0.350 | 0.650 |
| | HAPMAP-MEX | Mexican | 100 | IG | 0.580 | 0.360 | 0.060 | 1.000 | 0.760 | 0.240 |
| | HAPMAP-MKK | Maasai, Kenya | 286 | IG | 0.105 | 0.559 | 0.336 | 0.050 | 0.385 | 0.615 |
| | HAPMAP-TSI | Toscani, Italia | 176 | IG | 0.739 | 0.239 | 0.023 | 1.000 | 0.858 | 0.142 |
| ss97913182 | J. Craig Venter | | 2 | IG | 1.000 | | | | 1.000 | |

| Summary | Average Het.+/- std err: | Individual Count | Founders Count | Individual Overlap | Genotype Conflict |
|---|---|---|---|---|---|
| | 0.436+/-0.167 | 1213 | 1008 | 2 | 0 |

TU Delft

# Genome of the Netherlands

- GoNL: 769 persons genotyped

```
1    752721    rs3131972    A    G    .    PASS    AC=792;AF=0.794;AN=998;DB;set=SNP
```

```
##fileformat=VCFv4.0
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ft
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=HM2,Number=0,Type=Flag,Description="HapMap2 membership">
##INFO=<ID=HM3,Number=0,Type=Flag,Description="HapMap3 membership">
##reference=human_b36_both.fasta
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##INFO=<ID=GP,Number=1,Type=String,Description="GRCh37 position(s)">
##INFO=<ID=BN,Number=1,Type=Integer,Description="First dbSNP build #">
##INFO=<ID=NR,Number=0,Type=Flag,Description="No dbSNP 132 map weight=1 rs number assigned to position">
##INFO=<ID=OR,Number=1,Type=String,Description="Previous rs number">
##INFO=<ID=MP,Number=0,Type=Flag,Description="Maps to multiple positions on GRCh37">
```

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 695745 | rs72631875 | G | A | . | PASS | AA=.;DP=124;GP=1:705882;BN=130 | GT:GQ:DP | 1\|0:100:34 | 0\|0:62:20 | 1\|0:100:56 |
| 1 | 742429 | rs3094315 | G | A | . | PASS | AA=g;DP=132;HM2;GP=1:752566;BN=103 | GT:GQ:DP | 1\|1:100:38 | 1\|1:59:30 | 1\|1:100:44 |
| 1 | 742584 | rs3131972 | A | G | . | PASS | AA=a;DP=160;HM3;GP=1:752721;BN=103 | GT:GQ:DP | 1\|1:100:50 | 1\|1:100:33 | 1\|1:100:60 |
| 1 | 744366 | rs3115859 | G | A | . | PASS | AA=g;DP=127;GP=1:754503;BN=103 | GT:GQ:DP | 1\|1:80:31 | 1\|1:100:34 | 1\|1:100:45 |
| 1 | 746243 | rs3131963 | T | A | . | PASS | AA=t;DP=105;GP=1:756380;BN=103 | GT:GQ:DP | 1\|1:52:29 | 1\|1:43:24 | 1\|1:100:56 |
| 1 | 746775 | rs6699990 | A | G | . | PASS | AA=N;DP=120;GP=1:756912;BN=116 | GT:GQ:DP | 0\|1:100:34 | 0\|0:89:30 | 0\|0:100:46 |
| 1 | 7 | | | | | | | | | | |
| 1 | 7 | | | | | | | | | | |
| 1 | 7 | | | | | | | | | | |

| FORMAT | NA12891 | NA12892 | NA12878 |
|---|---|---|---|
| AA=C;DP=84;GP=1:62203;BN=125 | GT:GQ:DP | 1/0:44:23 | 1/0:43:20 | 1/0:70:36 |
| AA=.;DP=124;GP=1:705882;BN=130 | GT:GQ:DP | 1\|0:100:34 | 0\|0:62:20 | 1\|0:100:56 |
| AA=g;DP=132;HM2;GP=1:752566;BN=103 | GT:GQ:DP | 1\|1:100:38 | 1\|1:59:30 | 1\|1:100:44 |
| AA=a;DP=160;HM3;GP=1:752721;BN=103 | GT:GQ:DP | 1\|1:100:50 | 1\|1:100:33 | 1\|1:100:60 |
| AA=g;DP=127;GP=1:754503;BN=103 | GT:GQ:DP | 1\|1:80:31 | 1\|1:100:34 | 1\|1:100:45 |
| AA=t;DP=105;GP=1:756380;BN=103 | GT:GQ:DP | 1\|1:52:29 | 1\|1:43:24 | 1\|1:100:56 |
| AA=N;DP=120;GP=1:756912;BN=116 | GT:GQ:DP | 0\|1:100:34 | 0\|0:89:30 | 0\|0:100:46 |
| AA=-;DP=113;GP=1:757640;BN=103 | GT:GQ:DP | 1\|1:100:37 | 1\|1:61:25 | 1\|1:100:27 |
| AA=c;DP=129;GP=1:757734;BN=111 | GT:GQ:DP | 1\|1:100:36 | 1\|1:86:28 | 1\|1:100:56 |
| AA=c;DP=97;GP=1:757936;BN=111 | GT:GQ:DP | 1\|1:67:26 | 1\|1:44:13 | 1\|1:100:62 |

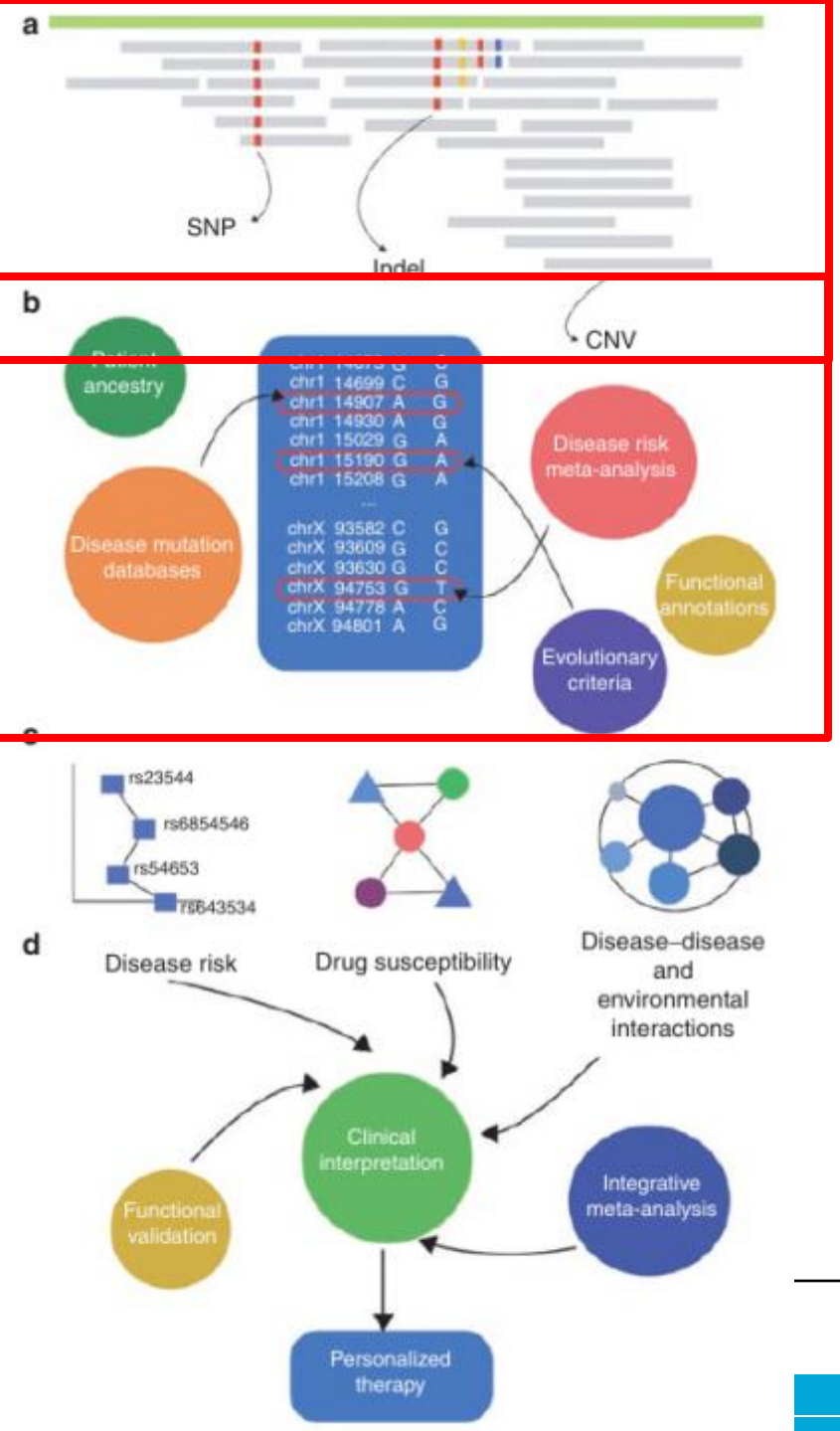
# Overview

Variant calling

Variant filtering

Variant relations

Variant visualization



# Millions of SNPs….

- Europeans have about 3 million SNPs w.r.t. reference genome.
- Which variant is the cause for a disease?

# Method 1: Look for known 'disease mutations'

- Databases: OMIM, HGMD, **ClinVar**
- Clinvar classifies mutations as 'untested',(likely) 'benign', (likely) 'pathogenic'

**Reference SNP (refSNP) Cluster Report: rs80187739** ** With pathogenic,probable-pathogenic,untested allele **

| RefSNP | |
|---|---|
| **Organism:** | human (Homo sapiens) |
| **Molecule Type:** | Genomic |
| **Created/Updated in build:** | 131/138 |
| **Map to Genome Build:** | 37.5/Weight 1 |
| **Validation Status:** | |

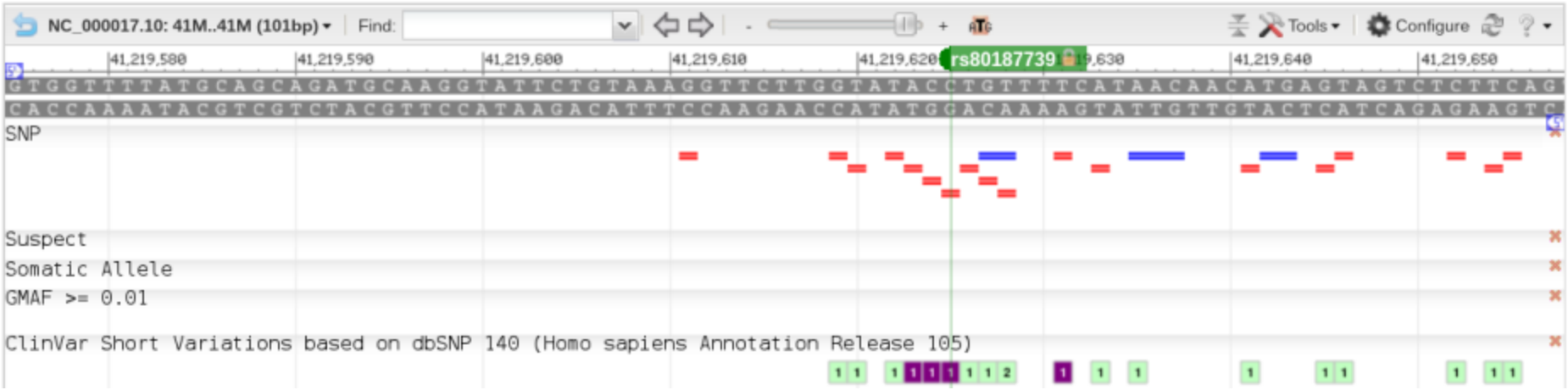
| Allele | |
|---|---|
| **Variation Class:** | SNV: single nucleotide variation |
| **RefSNP Alleles:** | A/C/G/T (REV) |
| **Allele Origin:** | |
| **Ancestral Allele:** | G |
| **Clinical Channel:** | VarView |
| **Clinical Significance:** | With pathogenic,probable-pathogenic,untested allele [ClinVar] |
| **MAF/MinorAlleleCount:** | NA |
| **MAF Source:** | |

NC_
NC_
NC_
NG_
NG_
NG_
NM_
NM_
NM_
NM_

# Method 1: Look for known 'disease mutations'

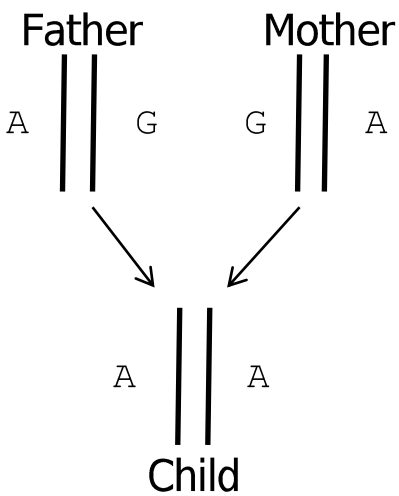
Gene Model(s)

| Function | mRNA | | | | Protein | | |
|---|---|---|---|---|---|---|---|
| | SNP to mRNA | Accession | Position | Allele change | Accession | Position | Residue change |
| ncRNA | Rev | NR_027676.1 | 5210 | NA ⇒ A | NA | NA | NA |
| missense | Rev | NM_007300.3 | 5369 | GAT ⇒ AAT | NP_009231.2 | 1713 | D [Asp] ⇒ N [Asn] |
| missense | Rev | NM_007299.3 | 1956 | GAT ⇒ AAT | NP_009230.2 | 588 | D [Asp] ⇒ N [Asn] |
| missense | Rev | NM_007298.3 | 1781 | GAT ⇒ AAT | NP_009229.2 | 588 | D [Asp] ⇒ N [Asn] |
| missense | Rev | NM_007297.3 | 5214 | GAT ⇒ AAT | NP_009228.2 | 1645 | D [Asp] ⇒ N [Asn] |
| missense | Rev | NM_007294.3 | 5306 | GAT ⇒ AAT | NP_009225.1 | 1692 | D [Asp] ⇒ N [Asn] |

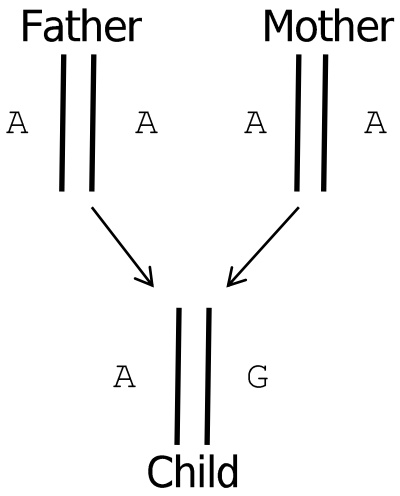


# Method 2: use of family

- Trio: Father, Mother, Child
- If child has disease, and not the father or mother, then:




2 * hetero zygote -> homozygote

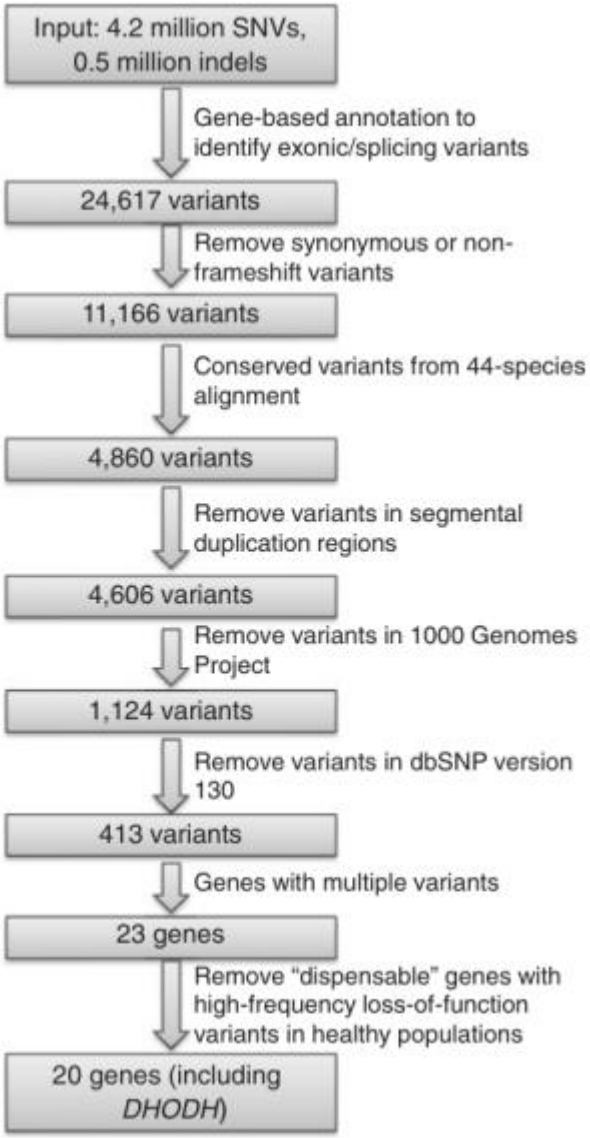



de novo mutation

- Possible to go to extended family

# Method 3: Filtering

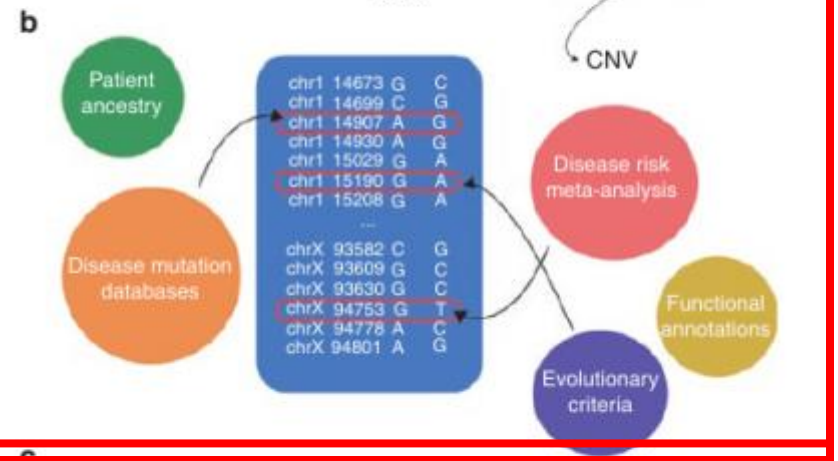
- Is variant:
  - In a gene?
  - Does it cause potential damage?
  - Is sequence conserved in other species?
  - Is it a common, well known variant?
- ANNOVAR pipeline



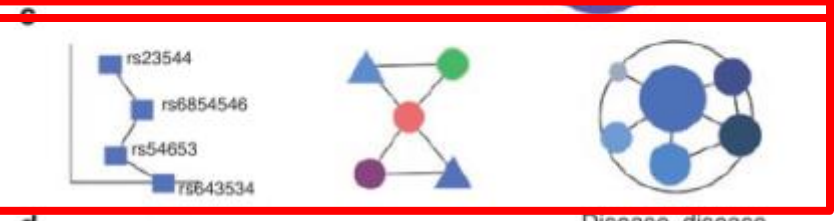
# Overview

Variant calling

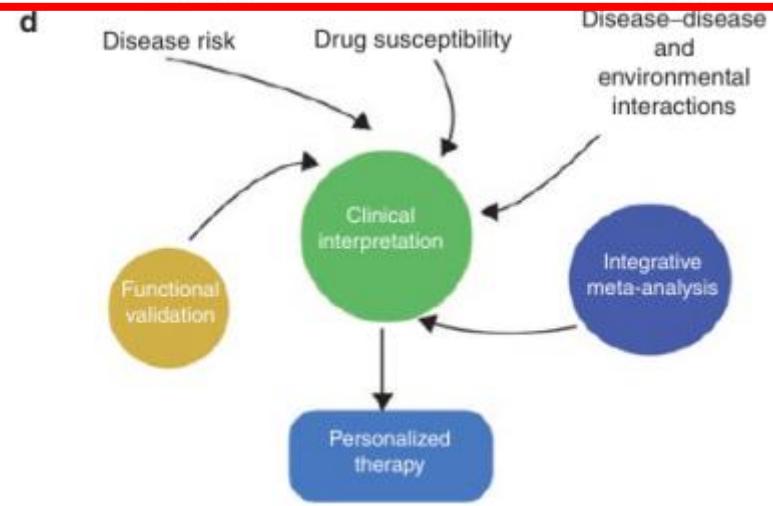

Variant filtering




Variant relations




Variant visualization



# Common versus rare variants

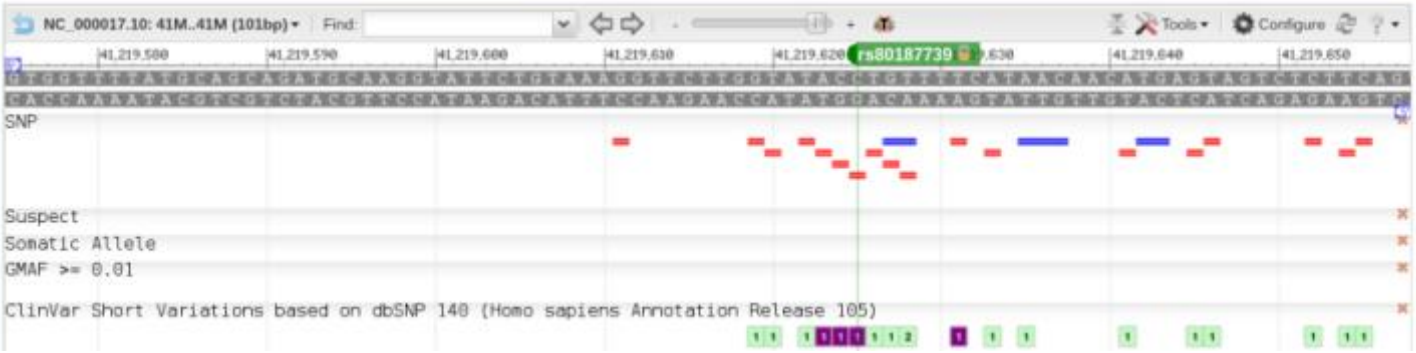
- Early days of GWAS: common variants found for e.g. diabetes
  - but could explain only part of observed heritability
- Hypothesis: most damage due to rare variants
- Problem: rare variants are rare --> not much known about them.

# Method 4: Nearby SNPs

- Are there many other disease-associated mutations in the same gene?

Gene Model(s)

| Function | mRNA | | | | Protein | | |
|---|---|---|---|---|---|---|---|
| | SNP to mRNA | Accession | Position | Allele change | Accession | Position | Residue change |
| ncRNA | Rev | NR_027676.1 | 5210 | NA ⇒ A | NA | NA | NA |
| missense | Rev | NM_007300.3 | 5369 | GAT ⇒ AAT | NP_009231.2 | 1713 | D [Asp] ⇒ N [Asn] |
| missense | Rev | NM_007299.3 | 1956 | GAT ⇒ AAT | NP_009230.2 | 588 | D [Asp] ⇒ N [Asn] |
| missense | Rev | NM_007298.3 | 1781 | GAT ⇒ AAT | NP_009229.2 | 588 | D [Asp] ⇒ N [Asn] |
| missense | Rev | NM_007297.3 | 5214 | GAT ⇒ AAT | NP_009228.2 | 1645 | D [Asp] ⇒ N [Asn] |
| missense | Rev | NM_007294.3 | 5306 | GAT ⇒ AAT | NP_009225.1 | 1692 | D [Asp] ⇒ N [Asn] |

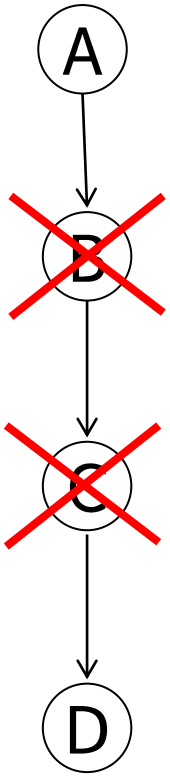



- Burden test: have diseased patients more rare damaging variants in a certain gene than healthy people?

# Method 5: Nearby genes

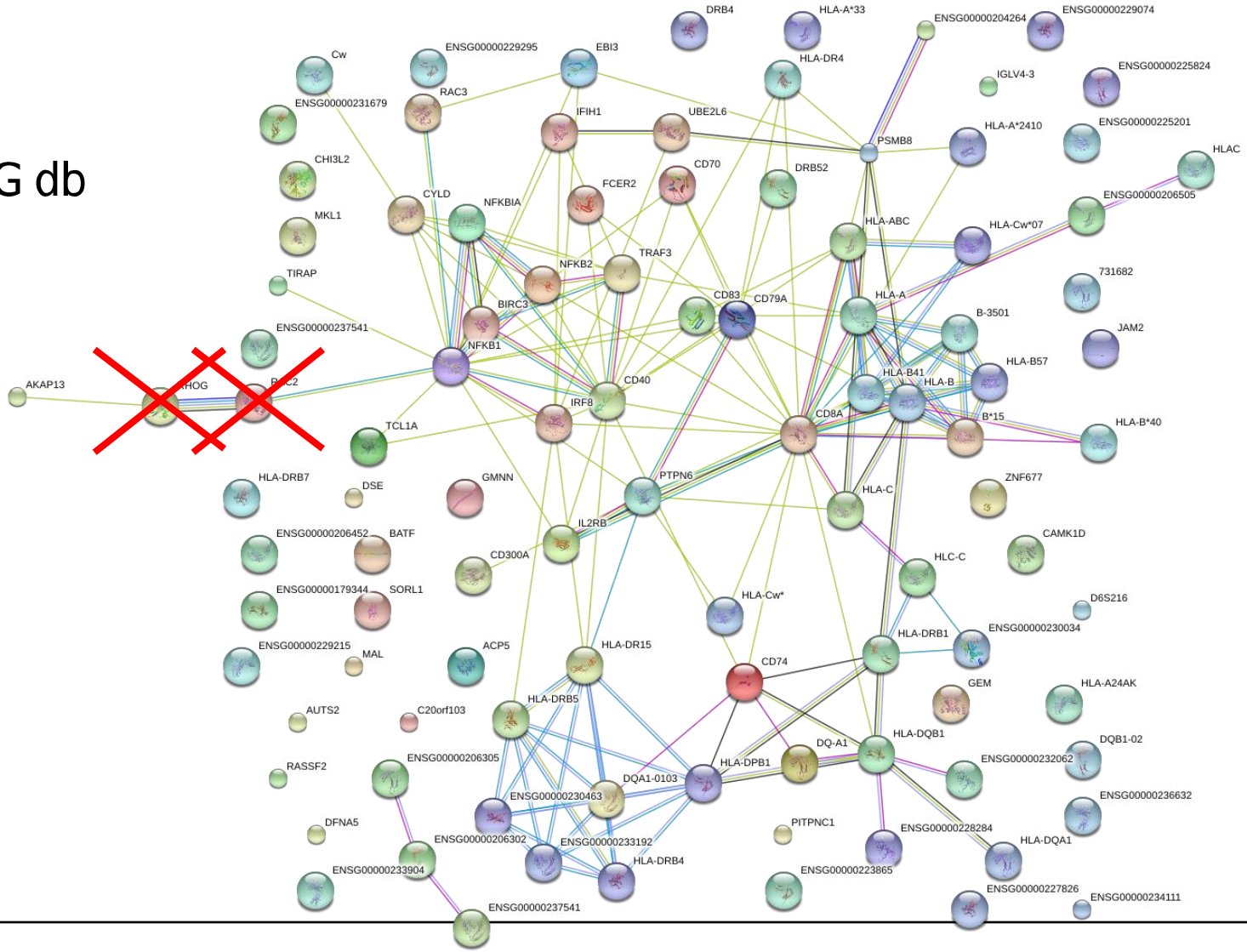
- Gene relations based on common function
  - Similar activity
  - Similar functional annotations
  - Similar sequence
  - Proteins have physical interaction
  - Etc.

- Hypothesis: damage in genes with related functions can cause similar diseases.



# Method 5: Nearby genes

- STRING db



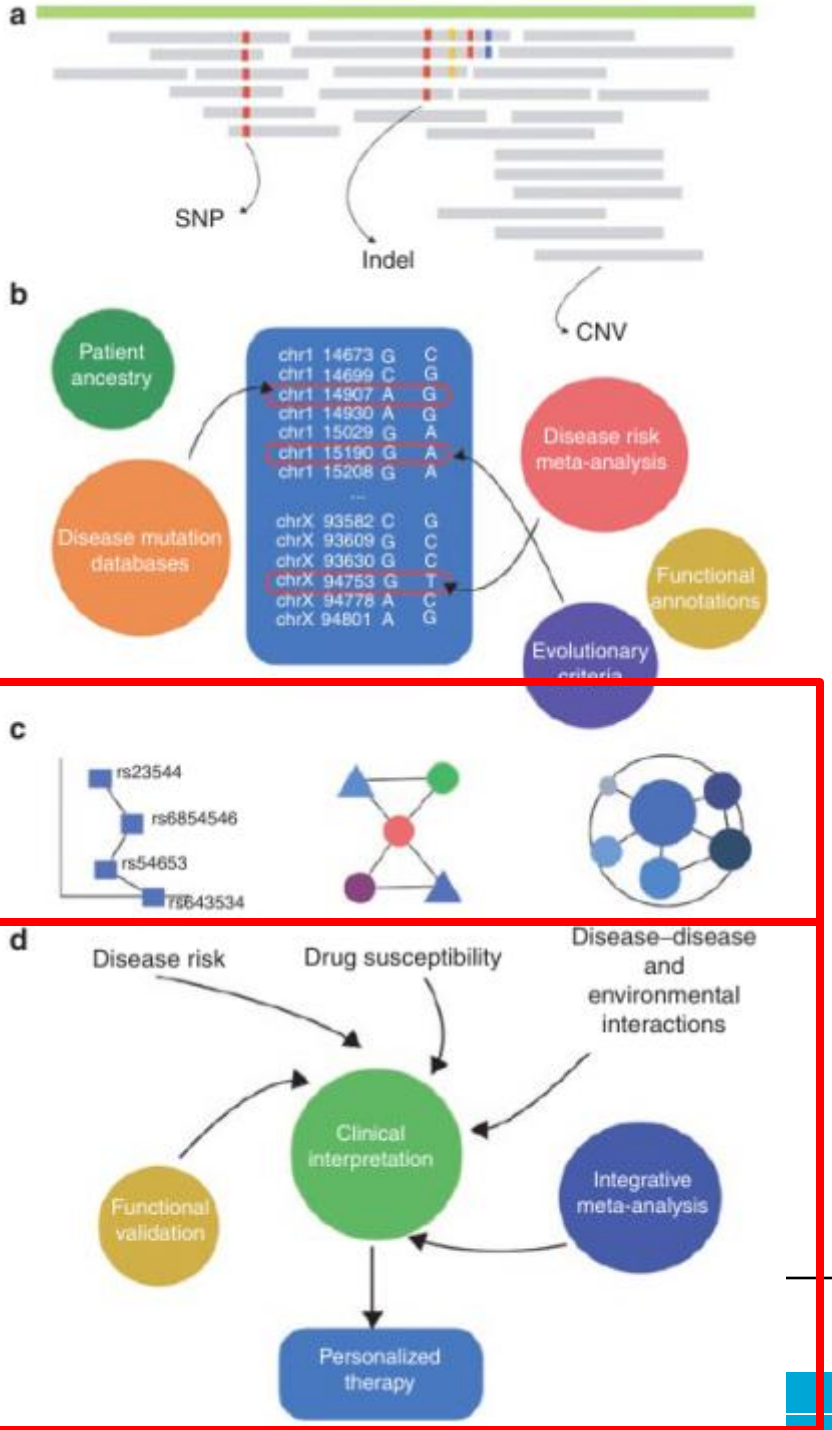
# Overview

Variant calling

Variant filtering

Variant relations

Variant visualization



# Visualizations

- To assist interpretation of variant data

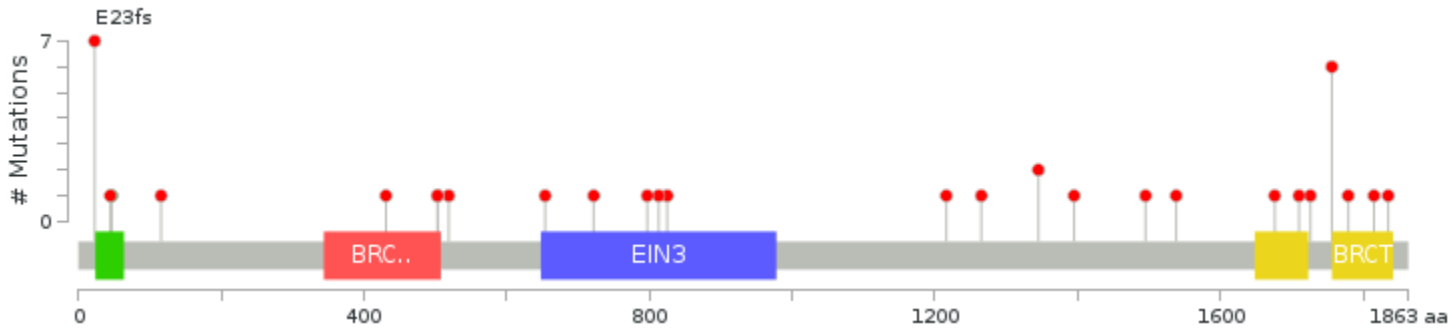
# Visualizations

| Chromosome | Position | Ref/variant | Gene | Amino acid change |
|---|---|---|---|---|
| chr10 | 126691631 | A/T | CTBP2 | Thr626Ser |
| chr11 | 1016928 | G/C | MUC6 | Ser1958Thr |
| chr1 | 156202173 | G/A | PMF1 | Gln75Arg |
| chr12 | 58220841 | T/C | CTDSP2 | Asp98Asn |
| chr12 | 58220844 | T/C | CTDSP2 | Glu97Lys |
| chr14 | 20692643 | T/C | OR11H6 | Cys259Arg |
| chr17 | 45214648 | C/G | CDC27 | Gln595Glu |
| chr17 | 45214651 | G/T | CDC27 | Ile594Leu |
| chr17 | 45214654 | T/C | CDC27 | Ala593Thr |
| chr2 | 97877292 | A/C | ANKRD36 | Pro709Gln |
| chr3 | 75787876 | T/G | ZNF717 | Leu300Ile |
| chr4 | 190876242 | A/G | FRG1 | Gly123Glu |
| chr4 | 47901476 | A/G | NFXL1 | Pro246Leu |
| chr9 | 68455161 | T/C | LOC100287354 | Arg94Trp |

Table 1: Non-synonymous variations found homozygously in the offspring and heterozygously in both parents

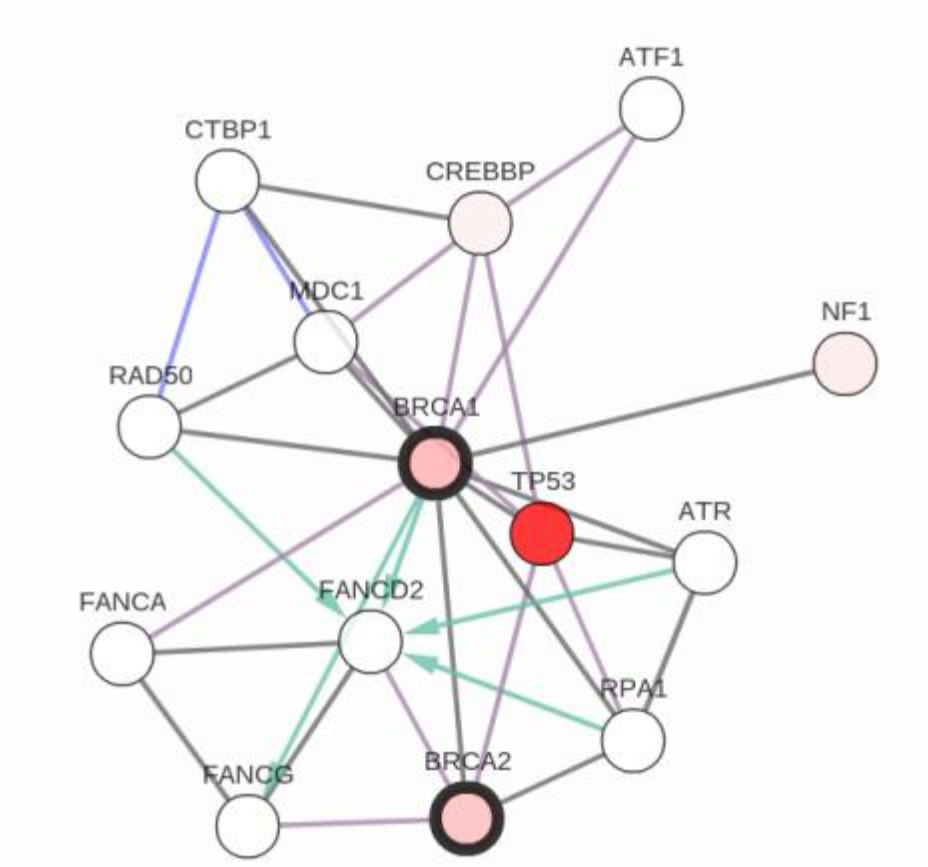
# Visualizations

- From cBioPortal:

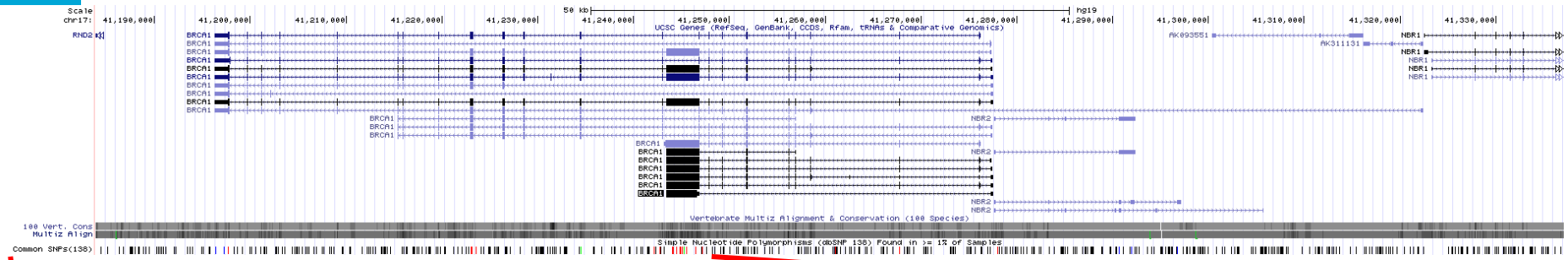


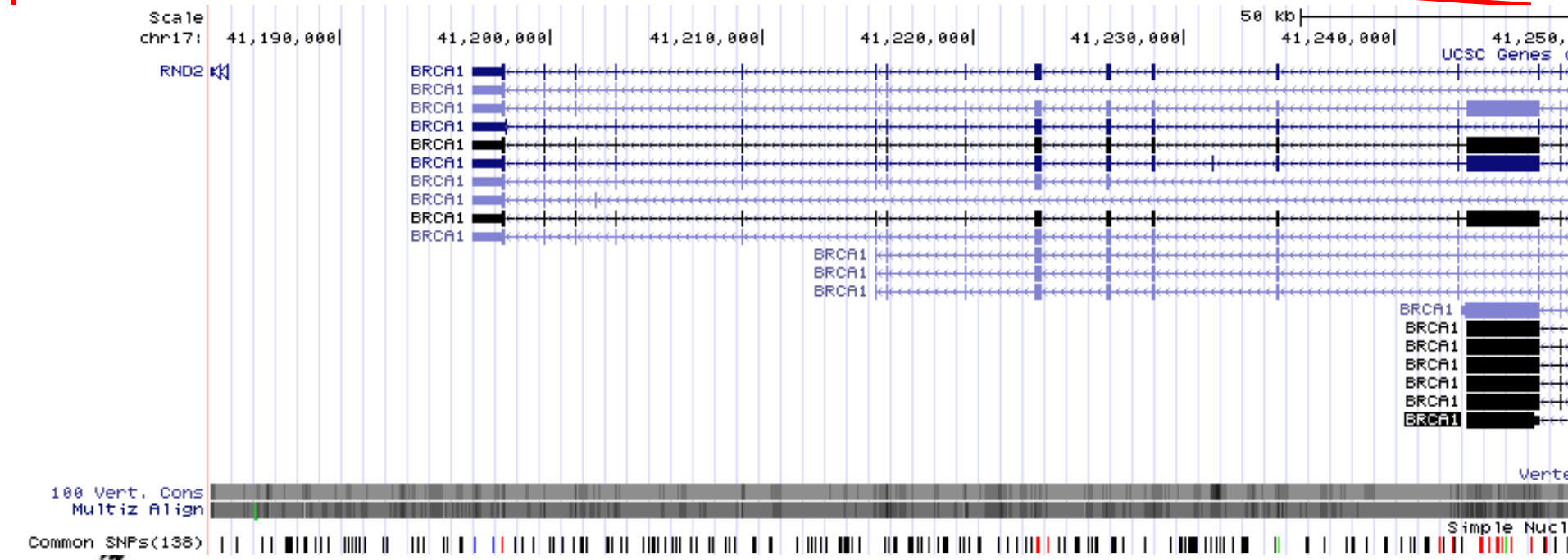
# Visualizations

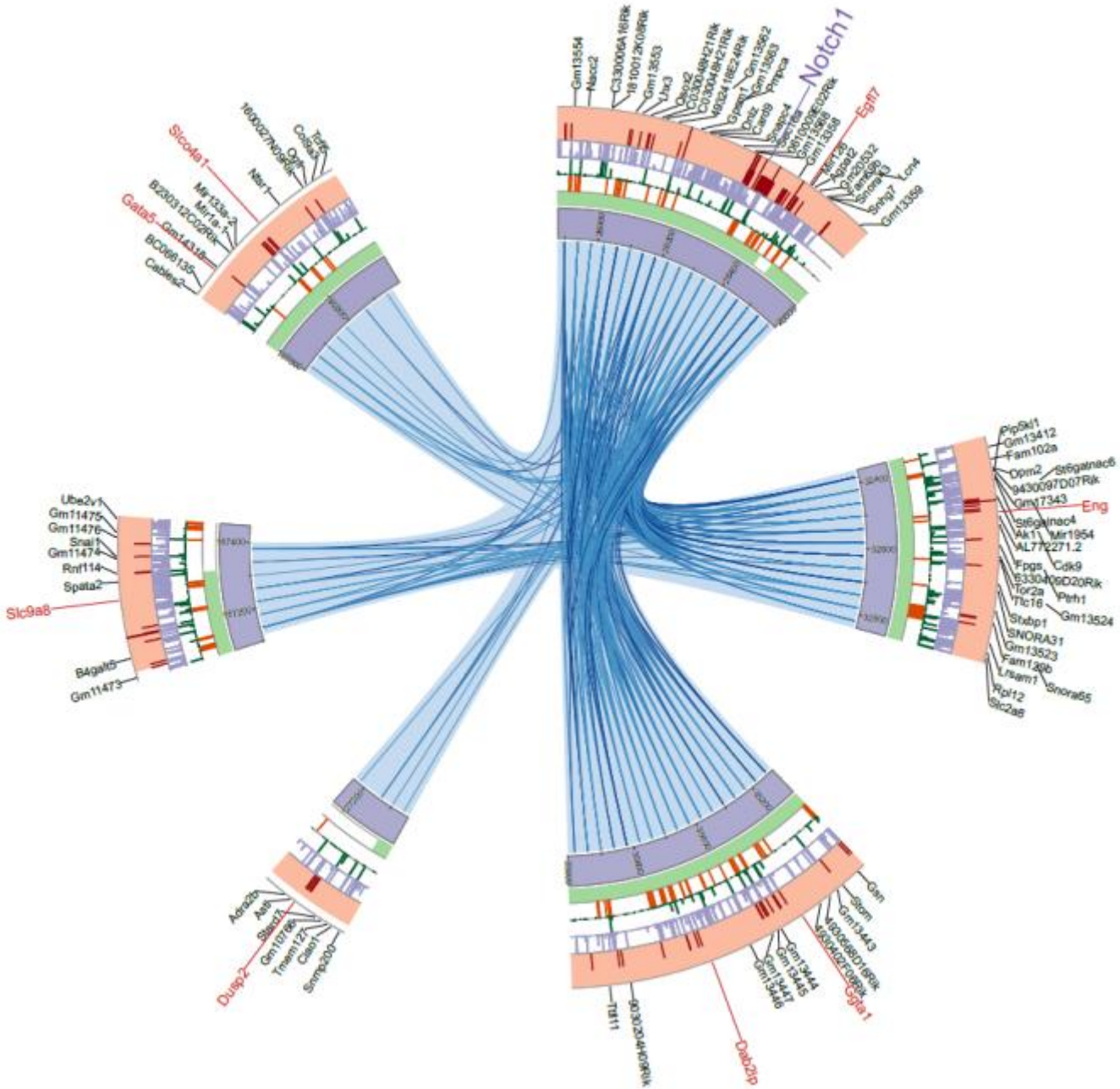
- From cBioPortal:



# Visualizations




Notch1
Egfl7
Slco4a1
Gata5
Eng
Slc9a8
Dusp2
Gja1
Dab2ip

# The project

- Design an application which visualizes and assist in the interpretation of genomic variants
- We are specifically interested in an application that is able to visualize:
  - Trio data
  - Gene interactions

# The project

- Method 1 (known variants):
  - We will give access to dbSNP SQL database
- Method 2 (family relations):
  - We will supply trio variant calls in VCF format
- Method 3 (ANNOVAR):
  - We will supply files containing information useful for filtering
- Method 4 (SNP relations):
  - We will supply a file containing gene locations (GFF3 format)
- Method 5 (gene relations):
  - We will give access to STRING SQL database

# Goal

- Create a useful visualization application which
  - helps clinical geneticists
  - in the interpretation of variants
  - by making use of the aforementioned data sources

# Voor volgende keer (6 mei)

- Lees materiaal voor volgende keer (synthetische biologie)
- Maak samenvatting in en lever deze in voor het begin van het college.