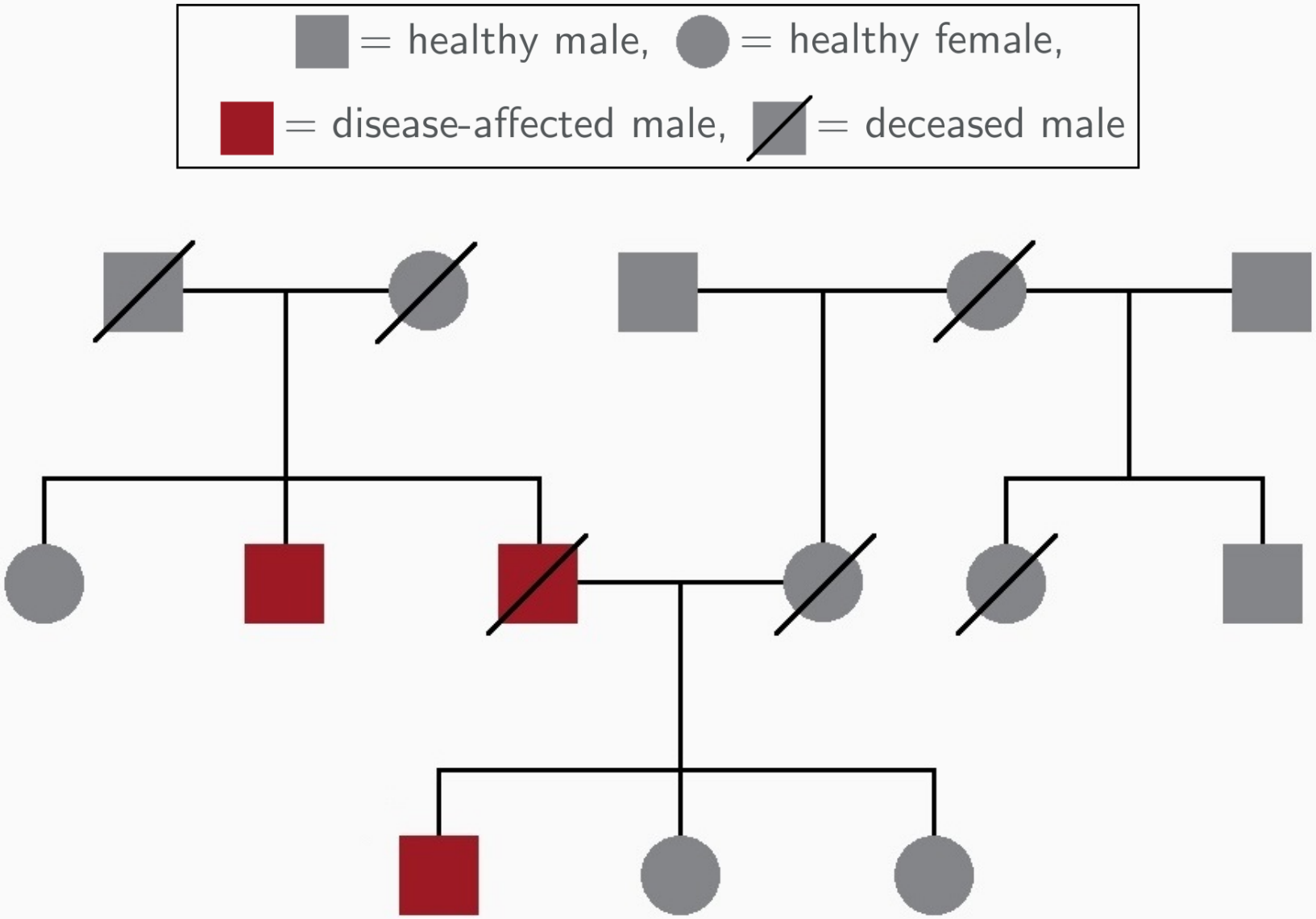
SEEDING THE SIMRVSEQUENCES R PACKAGE

Christina Nieuwoudt, Wen Tian Wang
Supervisor: Jinko Graham

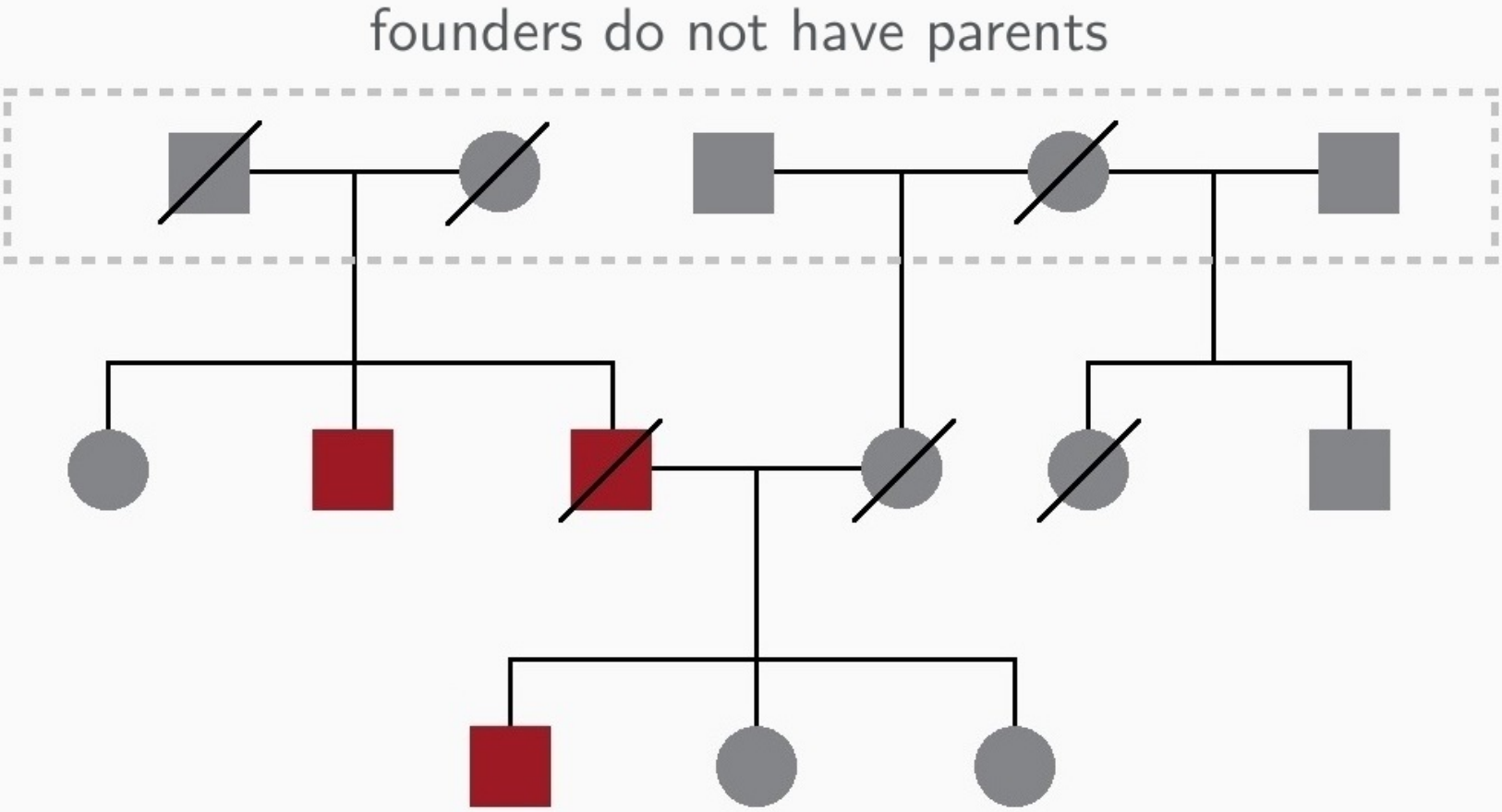
August 14, 2019

Simon Fraser University
Department of Statistics

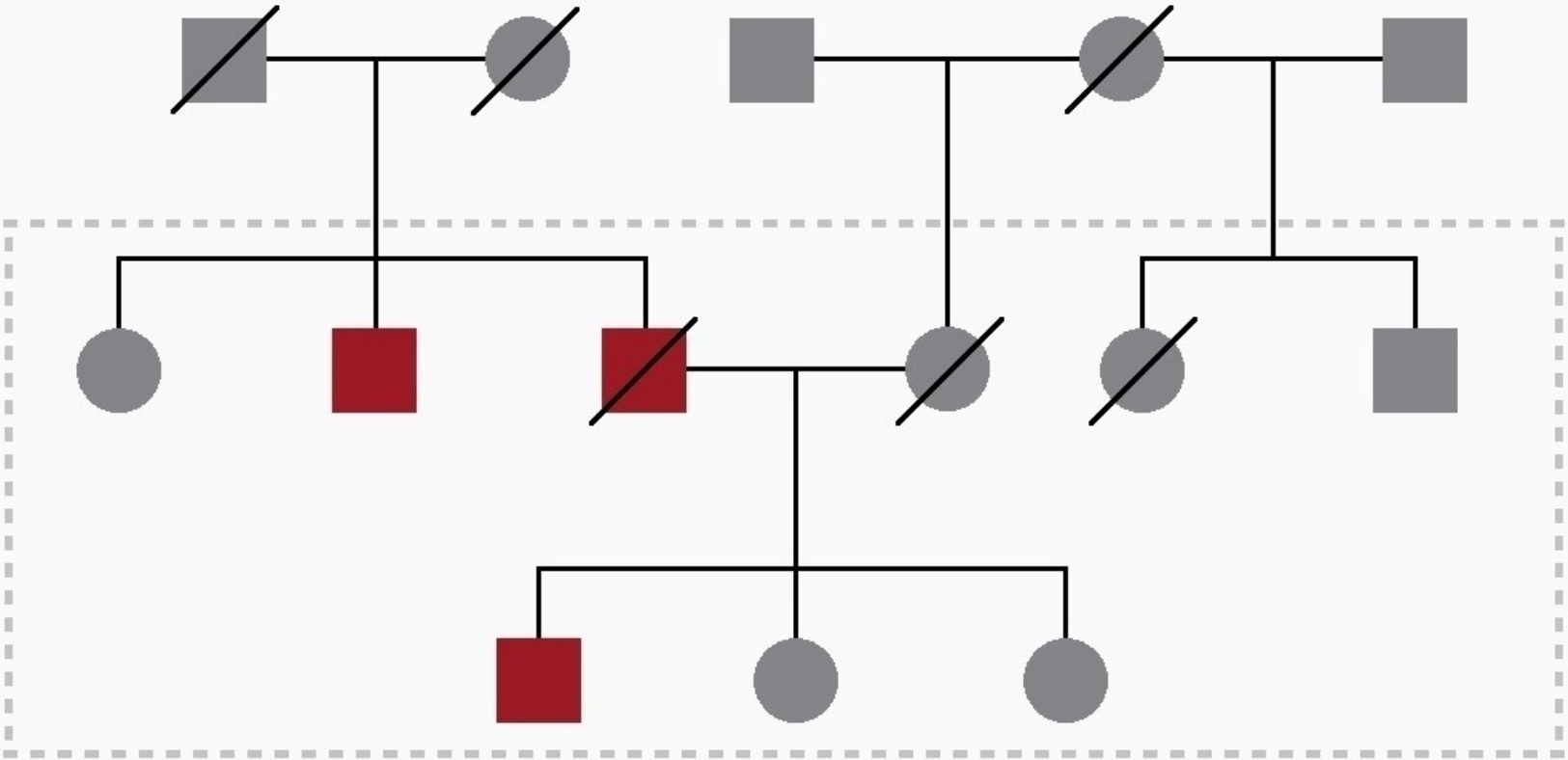
WHAT IS A PEDIGREE?



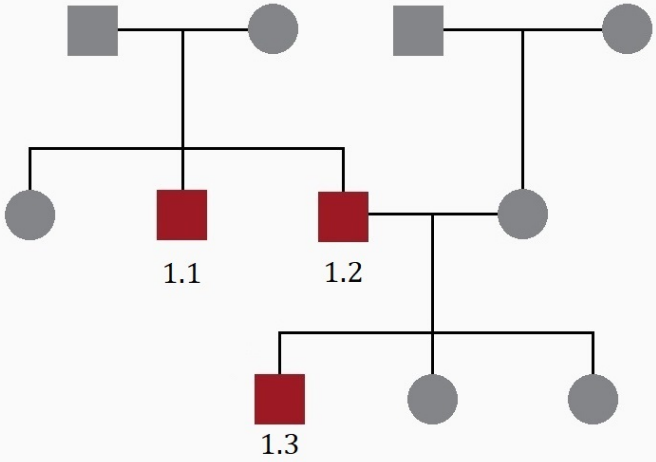
FOUNDERS VS NON-FOUNDERS



FOUNDERS VS NON-FOUNDERS



non-founders have both a mother and a father

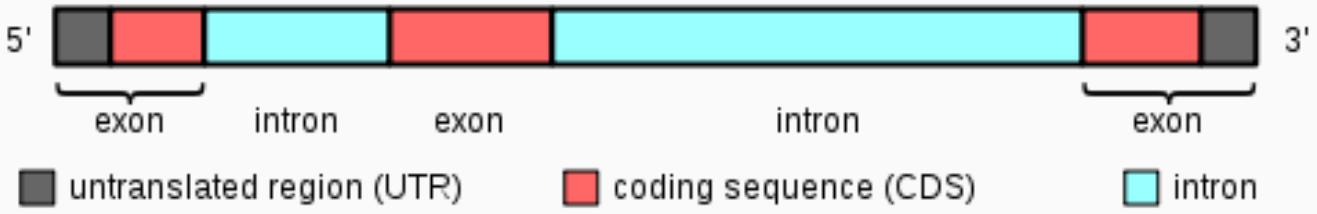


ID	SNV_1	SNV_2	SNV_3	$\cdots$	SNV_p
1.1	1	0	0	$\cdots$	0
1.1	0	1	0	$\cdots$	1
1.2	1	0	1	$\cdots$	0
1.2	0	0	1	$\cdots$	0
1.3	1	0	0	$\cdots$	0
1.3	0	0	1	$\cdots$	1

By convention, reference alleles are 0 and alternate alleles are 1.

To simulate sequence data for a pedigree we require:

1. the pedigree structure, and
2. single-nucleotide variant (SNV) data from a sample of unrelated individuals, representing the population of pedigree founders.



- We assume founders are unrelated and represent a random sample from a global population of individuals.
- Exon-only sequence data may be obtained from:
  - a coalescent simulator (msprime or fastsimcoal),
  - a forward-in-time evolutionary simulator (SLiM), or
  - publicly available sequence data (1000 Genomes Project)

OUTLINE

Datasets

1. 1000 Genome Project Data
2. Exon Positions Data

Software Command for Creating Exon Data

1. Relabel Chromosome ID for Exon Map
2. Remove Duplicated Exon Intervals
3. Extract SNVs in Exons from Chromosome 22

1KG DATA: CHROMOSOME 22

Retrieved from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/

Data file contains both biallelic SNVs and INDEL variants for the 2548 samples

File format:VCF file

Two required files for chromosome 22:

Index file:
ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.vcf.gz.tbi

VCF file:
ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.vcf.gz

1kg data: chromosome 22

```
##fileformat=VCFv4.3
##FILTER=<ID=PASS,Description="All filters passed">
##fileDate=27022019_15h52m43s
##source=IGSRpipeline
##reference=ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa
##FORMAT=<ID=GT,Number=1,Type=String,Description="Phased Genotype">
##contig=<ID=22>
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1)">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=EAS_AF,Number=A,Type=Float,Description="Allele frequency in the EAS populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=EUR_AF,Number=A,Type=Float,Description="Allele frequency in the EUR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=AFR_AF,Number=A,Type=Float,Description="Allele frequency in the AFR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=AMR_AF,Number=A,Type=Float,Description="Allele frequency in the AMR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=SAS_AF,Number=A,Type=Float,Description="Allele frequency in the SAS populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=VT,Number=.,Type=String,Description="indicates what type of variant the line represents">
##INFO=<ID=EX_TARGET,Number=0,Type=Flag,Description="indicates whether a variant is within the exon pull down target boundaries">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##bcftools_viewVersion=1.9-162-g33ecfe8+htslib-1.9-150-gc76b3b2
##bcftools_viewCommand=view ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.vcf.gz; Date=Tue Aug  6 12:01:54 2019
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00096
22      10516173        .       A       G       .       PASS    AC=121;AN=5096;DP=8203;AF=0.02;EAS_AF=0;EUR_AF=0.02;AFR_AF=0.06;AMR_AF=0.02;SAS_AF=0;VT=SNP;NS=2548     GT      0|0
22      10522217        .       G       A       .       PASS    AC=89;AN=5096;DP=9085;AF=0.02;EAS_AF=0;EUR_AF=0;AFR_AF=0.07;AMR_AF=0;SAS_AF=0;VT=SNP;NS=2548     GT      0|0
```

1kg data: chromosome 22

```
##fileformat=VCFv4.3
##FILTER=<ID=PASS,Description="All filters passed">
##fileDate=27022019_15h52m43s
##source=IGSRpipeline
##reference=ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa
##FORMAT=<ID=GT,Number=1,Type=String,Description="Phased Genotype">
##contig=<ID=22>
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1)">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=EAS_AF,Number=A,Type=Float,Description="Allele frequency in the EAS populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=EUR_AF,Number=A,Type=Float,Description="Allele frequency in the EUR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=AFR_AF,Number=A,Type=Float,Description="Allele frequency in the AFR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=AMR_AF,Number=A,Type=Float,Description="Allele frequency in the AMR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=SAS_AF,Number=A,Type=Float,Description="Allele frequency in the SAS populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=VT,Number=.,Type=String,Description="indicates what type of variant the line represents">
##INFO=<ID=EX_TARGET,Number=0,Type=Flag,Description="indicates whether a variant is within the exon pull down target boundaries">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##bcftools_viewVersion=1.9-162-g33ecfe8+htslib-1.9-150-gc76b3b2
##bcftools_viewCommand=view ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.vcf.gz; Date=Tue Aug 6 12:01:54 2019
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00096
22      10516173        .       A       G       .       PASS    AC=121;AN=5096;DP=8203;AF=0.02;EAS_AF=0;EUR_AF=0.02;AFR_AF=0.06;AMR_AF=0.02;SAS_AF=0;VT=SNP;NS=2548     GT      0|0
22      10522217        .       G       A       .       PASS    AC=89;AN=5096;DP=9085;AF=0.02;EAS_AF=0;EUR_AF=0;AFR_AF=0.07;AMR_AF=0;SAS_AF=0;VT=SNP;NS=2548     GT      0|0
```

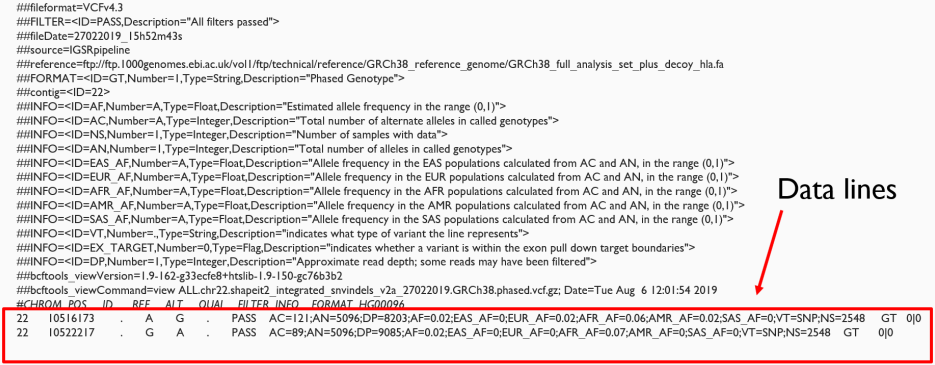
Meta information

1kg data: chromosome 22

```
##fileformat=VCFv4.3
##FILTER=<ID=PASS,Description="All filters passed">
##fileDate=27022019_15h52m43s
##source=IGSRpipeline
##reference=ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa
##FORMAT=<ID=GT,Number=1,Type=String,Description="Phased Genotype">
##contig=<ID=22>
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1)">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=EAS_AF,Number=A,Type=Float,Description="Allele frequency in the EAS populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=EUR_AF,Number=A,Type=Float,Description="Allele frequency in the EUR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=AFR_AF,Number=A,Type=Float,Description="Allele frequency in the AFR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=AMR_AF,Number=A,Type=Float,Description="Allele frequency in the AMR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=SAS_AF,Number=A,Type=Float,Description="Allele frequency in the SAS populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=VT,Number=.,Type=String,Description="indicates what type of variant the line represents">
##INFO=<ID=EX_TARGET,Number=0,Type=Flag,Description="indicates whether a variant is within the exon pull down target boundaries">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##bcftools_viewVersion=1.9-162-g33ecfe8+htslib-1.9-150-gc76b3b2
##bcftools_viewCommand=view ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.vcf.gz; Date=Tue Aug  6 12:01:54 2019
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00096
22      10516173        .       A       G       .       PASS    AC=121;AN=5096;DP=8203;AF=0.02;EAS_AF=0;EUR_AF=0.02;AFR_AF=0.06;AMR_AF=0.02;SAS_AF=0;VT=SNP;NS=2548     GT      0|0
22      10522217        .       G       A       .       PASS    AC=89;AN=5096;DP=9085;AF=0.02;EAS_AF=0;EUR_AF=0;AFR_AF=0.07;AMR_AF=0;SAS_AF=0;VT=SNP;NS=2548     GT      0|0
```

Header line

1KG DATA: CHROMOSOME 22



```
##fileformat=VCFv4.3
##FILTER=<ID=PASS,Description="All filters passed">
##fileDate=27022019_15h52m43s
##source=IGSRpipeline
##reference=ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa
##FORMAT=<ID=GT,Number=1,Type=String,Description="Phased Genotype">
##contig=<ID=22>
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1)">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=EAS_AF,Number=A,Type=Float,Description="Allele frequency in the EAS populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=EUR_AF,Number=A,Type=Float,Description="Allele frequency in the EUR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=AFR_AF,Number=A,Type=Float,Description="Allele frequency in the AFR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=AMR_AF,Number=A,Type=Float,Description="Allele frequency in the AMR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=SAS_AF,Number=A,Type=Float,Description="Allele frequency in the SAS populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=VT,Number=.,Type=String,Description="indicates what type of variant the line represents">
##INFO=<ID=EX_TARGET,Number=0,Type=Flag,Description="indicates whether a variant is within the exon pull down target boundaries">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##bcftools_viewVersion=1.9-162-g33ecfe8+htslib-1.9-150-gc76b3b2
##bcftools_viewCommand=view ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.vcf.gz; Date=Tue Aug  6 12:01:54 2019
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00096
22      10516173        .       A       G       .       PASS    AC=121;AN=5096;DP=8203;AF=0.02;EAS_AF=0;EUR_AF=0.02;AFR_AF=0.06;AMR_AF=0.02;SAS_AF=0;VT=SNP;NS=2548     GT      0|0
22      10522217        .       G       A       .       PASS    AC=89;AN=5096;DP=9085;AF=0.02;EAS_AF=0;EUR_AF=0;AFR_AF=0.07;AMR_AF=0;SAS_AF=0;VT=SNP;NS=2548    GT      0|0
```

Data lines

EXON POSITIONS DATA

RETRIEVED FROM
HTTP://FTP.1000GENOMES.EBI.AC.UK/VOL1/FTP/DATA_COLLECTIONS/1000_GENOMES_PROJECT/WORKING/20190125_COORDS_EXON_TARGET/OUTPUT_1000G_EXOME.V1.BED

FILE FORMAT: BED FILE

EXON POSITIONS DATA

Chromosme ID:
Number from
chr1..22, X,Y

chr1	14642	14882	+	target.0
chr1	14943	15063	+	target.1
chr1	15751	15990	+	target.2
chr1	16599	16719	+	target.3
chr1	16834	17074	+	target.4
chr1	17211	17331	+	target.5
chr1	30275	30431	+	target.6
chr1	69069	70029	+	target.7
chr1	129133	129253	+	target.8
chr1	258482	258603	+	target.9

EXON POSITIONS DATA

Start and end positions of exons

chr1	14642	14882	+	target.0
chr1	14943	15063	+	target.1
chr1	15751	15990	+	target.2
chr1	16599	16719	+	target.3
chr1	16834	17074	+	target.4
chr1	17211	17331	+	target.5
chr1	30275	30431	+	target.6
chr1	69069	70029	+	target.7
chr1	129133	129253	+	target.8
chr1	258482	258603	+	target.9

EXON POSITIONS DATA

Strand orientation

chr1	14642	14882	+	target.0
chr1	14943	15063	+	target.1
chr1	15751	15990	+	target.2
chr1	16599	16719	+	target.3
chr1	16834	17074	+	target.4
chr1	17211	17331	+	target.5
chr1	30275	30431	+	target.6
chr1	69069	70029	+	target.7
chr1	129133	129253	+	target.8
chr1	258482	258603	+	target.9

EXON POSITIONS DATA

Names of exon regions

chr1	14642	14882	+	target.0
chr1	14943	15063	+	target.1
chr1	15751	15990	+	target.2
chr1	16599	16719	+	target.3
chr1	16834	17074	+	target.4
chr1	17211	17331	+	target.5
chr1	30275	30431	+	target.6
chr1	69069	70029	+	target.7
chr1	129133	129253	+	target.8
chr1	258482	258603	+	target.9

EXON POSITIONS DATA

Chromosme ID:
Number from
chr1..22, X,Y

chr1	14642	14882	+	target.0
chr1	14943	15063	+	target.1
chr1	15751	15990	+	target.2
chr1	16599	16719	+	target.3
chr1	16834	17074	+	target.4
chr1	17211	17331	+	target.5
chr1	30275	30431	+	target.6
chr1	69069	70029	+	target.7
chr1	129133	129253	+	target.8
chr1	258482	258603	+	target.9

CHROMOSOME 22

```
##fileformat=VCFv4.3
##FILTER=<ID=PASS,Description="All filters passed">
##fileDate=27022019_15h52m43s
##source=IGSRpipeline
##reference=ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa
##FORMAT=<ID=GT,Number=1,Type=String,Description="Phased Genotype">
##contig=<ID=22>
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1)">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=EAS_AF,Number=A,Type=Float,Description="Allele frequency in the EAS populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=EUR_AF,Number=A,Type=Float,Description="Allele frequency in the EUR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=AFR_AF,Number=A,Type=Float,Description="Allele frequency in the AFR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=AMR_AF,Number=A,Type=Float,Description="Allele frequency in the AMR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=SAS_AF,Number=A,Type=Float,Description="Allele frequency in the SAS populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=VT,Number=.,Type=String,Description="indicates what type of variant the line represents">
##INFO=<ID=EX_TARGET,Number=0,Type=Flag,Description="indicates whether a variant is within the exon pull down target boundaries">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##bcftools_viewVersion=1.9-162-g33ecfe8+htslib-1.9-150-gc76b3b2
##bcftools_viewCommand=view ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.vcf.gz; Date=Tue Aug  6 12:01:54 2019
#CHROM  POS      ID   REF  ALT  QUAL  FILTER  INFO     FORMAT  HG00096
22      10516173  .    A    G    .     PASS    AC=121;AN=5096;DP=8203;AF=0.02;EAS_AF=0;EUR_AF=0.02;AFR_AF=0.06;AMR_AF=0.02;SAS_AF=0;VT=SNP;NS=2548    GT   0|0
22      10522217  .    G    A    .     PASS    AC=89;AN=5096;DP=9085;AF=0.02;EAS_AF=0;EUR_AF=0;AFR_AF=0.07;AMR_AF=0;SAS_AF=0;VT=SNP;NS=2548    GT     0|0
```

SOFTWARE COMMANDS FOR CREATING EXON DATA

Relabel	Chromosome ID for Exon Map
Remove	Duplicated Exon Intervals
Extract	SNVs in Exons from Chromosome 22

```
cut -c4- output_1000G_Exome.v1.bed  > 1KG.exons.bed

sort -V -k1,1 -k2,2 1KG.exons.bed | bedtools merge > 1KG.exons.bash_merged.bed

bcftools view
  --regions-file 1KG.exons.bash_merged.bed
  --types snps --min-alleles 2 --max-alleles 2
  --include 'FILTER="PASS"'
  --output-type z
  --output-file exons_chr22.vcf.gz
ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.vcf.gz;
```

RELABEL CHROMOSOME ID FOR EXON MAP

```
cut -c4- output_1000G_Exome.v1.bed  > 1KG.exons.bed
```

RELABEL CHROMOSOME ID FOR EXON MAP

```
cut -c4- output_1000G_Exome.v1.bed  > 1KG.exons.bed
```

removes the regions
we specified from
each row

RELABEL CHROMOSOME ID FOR EXON MAP

```
cut -c4- output_1000G_Exome.v1.bed  > 1KG.exons.bed
```

selects characters from the fourth index to the last index in each row; in another word, remove the first three characters, "chr".

RELABEL CHROMOSOME ID FOR EXON MAP

```
cut -c4- output_1000G_Exome.v1.bed > 1KG.exons.bed
```

Input

RELABEL CHROMOSOME ID FOR EXON MAP

```
cut -c4- output_1000G_Exome.v1.bed > 1KG.exons.bed
```

creates a new output file and redirects the result to this output file

RELABEL CHROMOSOME ID FOR EXON MAP

```
cut -c4- output_1000G_Exome.v1.bed  > 1KG.exons.bed
```

Output

output of 1kg.exons.bed

```
1       14642   14882   +       target.0
1       14943   15063   +       target.1
1       15751   15990   +       target.2
1       16599   16719   +       target.3
1       16834   17074   +       target.4
1       17211   17331   +       target.5
1       30275   30431   +       target.6
1       69069   70029   +       target.7
1       129133  129253  +       target.8
1       258482  258603  +       target.9
```

REMOVE DUPLICATED EXON INTERVALS

```
sort -V -k1,1 -k2,2 1KG.exons.bed | bedtools merge >
1KG.exons.bash_merged.bed
```

REMOVE DUPLICATED EXON INTERVALS

```
sort -V -k1,1 -k2,2 1KG.exons.bed | bedtools merge >
1KG.exons.bash_merged.bed
```

sorts lines of input

REMOVE DUPLICATED EXON INTERVALS

```
sort -V -k1,1 -k2,2 1KG.exons.bed | bedtools merge >
1KG.exons.bash_merged.bed
```

sorts numbers in natural version

REMOVE DUPLICATED EXON INTERVALS

```
sort -V -k1,1 -k2,2 1KG.exons.bed | bedtools merge >
1KG.exons.bash_merged.bed
```

sort by the first field
(chromosome ID)

REMOVE DUPLICATED EXON INTERVALS

```
sort -V -k1,1 -k2,2 1KG.exons.bed | bedtools merge >
1KG.exons.bash_merged.bed
```

sort by the second
field (start position)

REMOVE DUPLICATED EXON INTERVALS

```
sort -V -k1,1 -k2,2 1KG.exons.bed | bedtools merge >
1KG.exons.bash_merged.bed
```

Input

```
sort -V -k1,1 -k2,2 1KG.exons.bed | bedtools merge
> 1KG.exons.bash_merged.bed
```

Pipe operation, which passes the output from the "sort" command to the "bedtools merge" command. This action is similar as "%>%" in r-studio

REMOVE DUPLICATED EXON INTERVALS

```
sort -V -k1, 1 -k2,2 1KG.exons.bed | bedtools merge >
1KG.exons.bash_merged.bed
```

merge overlapped exons intervals

REMOVE DUPLICATED EXON INTERVALS

```
sort -V -k1,1 -k2,2 1KG.exons.bed | bedtools merge >
1KG.exons.bash_merged.bed
```

Output

output of 1kg.exons.bash_merged.bed

```
1       14642   14882
1       14943   15063
1       15751   15990
1       16599   16719
1       16834   17074
1       17211   17331
1       30275   30431
1       69069   70029
1       129133  129253
1       258482  258603
```

EXTRACT SNVS IN EXONS FROM CHROMOSOME 22

```
bcftools view
  --regions-file 1KG.exons.bash_merged.bed
  --types snps --min-alleles 2 --max-alleles 2
  --include 'FILTER="PASS"'
  --output-type z
  --output-file exons_chr22.vcf.gz
ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.
vcf.gz;
```

EXTRACT SNVS IN EXONS FROM CHROMOSOME 22

```
bcftools view
  --regions-file 1KG.exons.bash_merged.bed
  --types snps --min-alleles 2 --max-alleles 2
  --include 'FILTER="PASS"'
  --output-type z
  --output-file exons_chr22.vcf.gz
ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.
vcf.gz;
```

Use view command for filtering

EXTRACT SNVS IN EXONS FROM CHROMOSOME 22

```
bcftools view
  --regions-file 1KG.exons.bash_merged.bed
  --types snps --min-alleles 2 --max-alleles 2
  --include 'FILTER="PASS"'
  --output-type z
  --output-file exons_chr22.vcf.gz
ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.
vcf.gz;
```

filters to exons specified in bed file

EXTRACT SNVS IN EXONS FROM CHROMOSOME 22

```
bcftools view
  --regions-file 1KG.exons.bash_merged.bed
  --types snps --min-alleles 2 --max-alleles 2
  --include 'FILTER="PASS"'
  --output-type z
  --output-file exons_chr22.vcf.gz
ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.
vcf.gz;
```

includes diallelic SNVs only

EXTRACT SNVS IN EXONS FROM CHROMOSOME 22

```
bcftools view
  --regions-file 1KG.exons.bash_merged.bed
  --types snps --min-alleles 2 --max-alleles 2
  --include 'FILTER="PASS"'
  --output-type z
  --output-file exons_chr22.vcf.gz
ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.
vcf.gz;
```

includes variants that passed all quality filters

EXTRACT SNVS IN EXONS FROM CHROMOSOME 22

```
bcftools view
  --regions-file 1KG.exons.bash_merged.bed
  --types snps --min-alleles 2 --max-alleles 2
  --include 'FILTER="PASS"'
  --output-type z
  --output-file exons_chr22.vcf.gz
ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.
vcf.gz;
```

compresses the output to gzip

EXTRACT SNVS IN EXONS FROM CHROMOSOME 22

```
bcftools view
  --regions-file 1KG.exons.bash_merged.bed
  --types snps --min-alleles 2 --max-alleles 2
  --include 'FILTER="PASS"'
  --output-type z
  --output-file exons_chr22.vcf.gz
ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.
vcf.gz;
```

Output file

EXTRACT SNVS IN EXONS FROM CHROMOSOME 22

```
bcftools view
  --regions-file 1KG.exons.bash_merged.bed
  --types snps --min-alleles 2 --max-alleles 2
  --include 'FILTER="PASS"'
  --output-type z
  --output-file exons_chr22.vcf.gz
ALL.chr22.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.
vcf.gz;
```

Input file

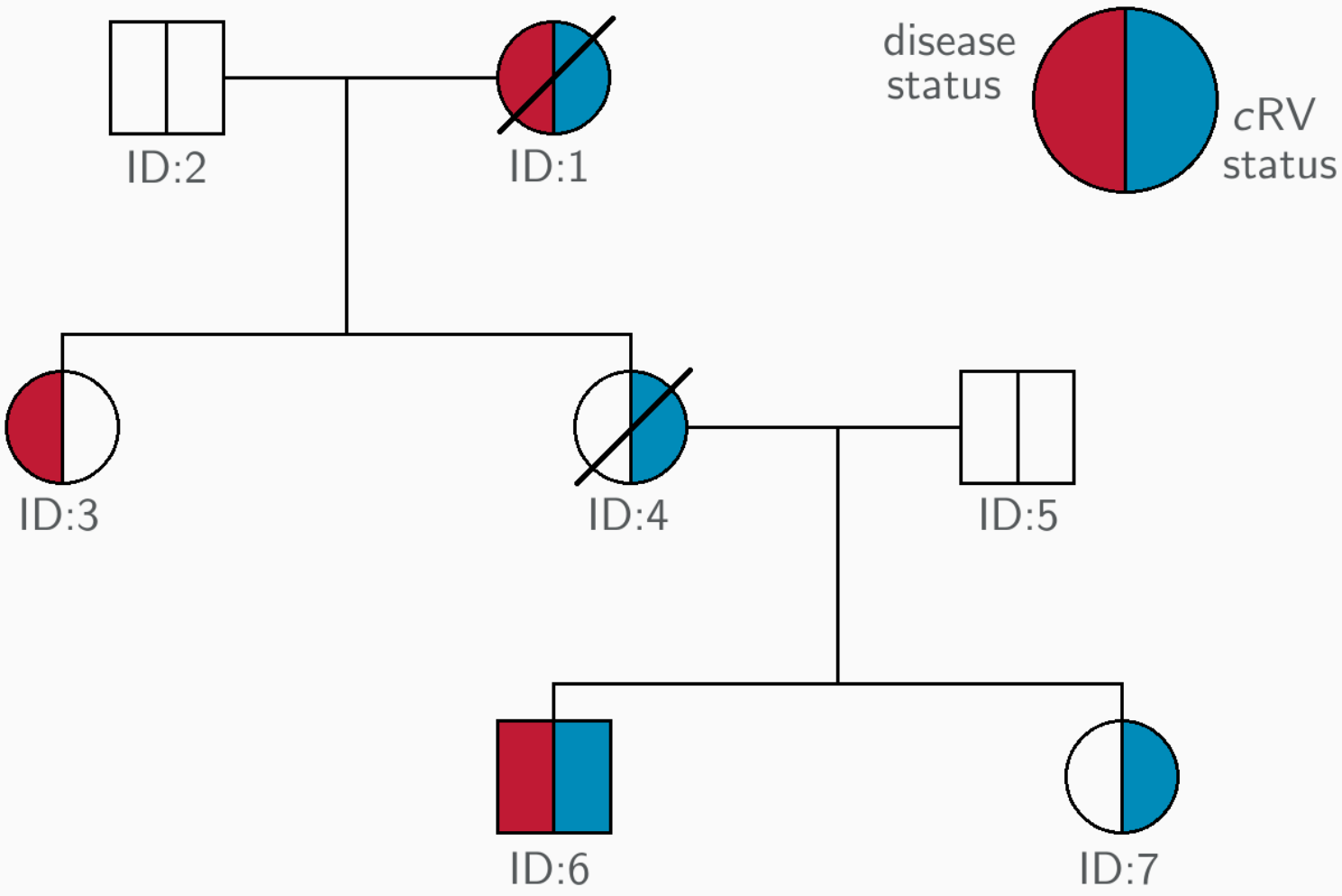
EXAMPLE OUTPUT DATA

```
22	16390584	.	C	T	.	PASS	AC=23;AN=5096;DP=82786;AF=0;EAS_AF=0;EUR_AF=0;AFR_AF=0.01;AMR_AF=0;SAS_AF=0;EX_TARGET;VT=SNP;NS=2548	GT	0|0
22	16390592	.	A	G	.	PASS	AC=1386;AN=5096;DP=86545;AF=0.27;EAS_AF=0.37;EUR_AF=0.34;AFR_AF=0.08;AMR_AF=0.33;SAS_AF=0.32;EX_TARGET;VT=SNP;NS=2548	GT	0|1
22	16390594	.	C	G	.	PASS	AC=1;AN=5096;DP=87990;AF=0;EAS_AF=0;EUR_AF=0;AFR_AF=0;AMR_AF=0;SAS_AF=0;EX_TARGET;VT=SNP;NS=2548	GT	0|0
22	16390595	.	A	G	.	PASS	AC=1;AN=5096;DP=88574;AF=0;EAS_AF=0;EUR_AF=0;AFR_AF=0;AMR_AF=0;SAS_AF=0;EX_TARGET;VT=SNP;NS=2548	GT	0|0
22	16390599	.	G	T	.	PASS	AC=1;AN=5096;DP=90114;AF=0;EAS_AF=0;EUR_AF=0;AFR_AF=0;AMR_AF=0;SAS_AF=0;EX_TARGET;VT=SNP;NS=2548	GT	0|0
```

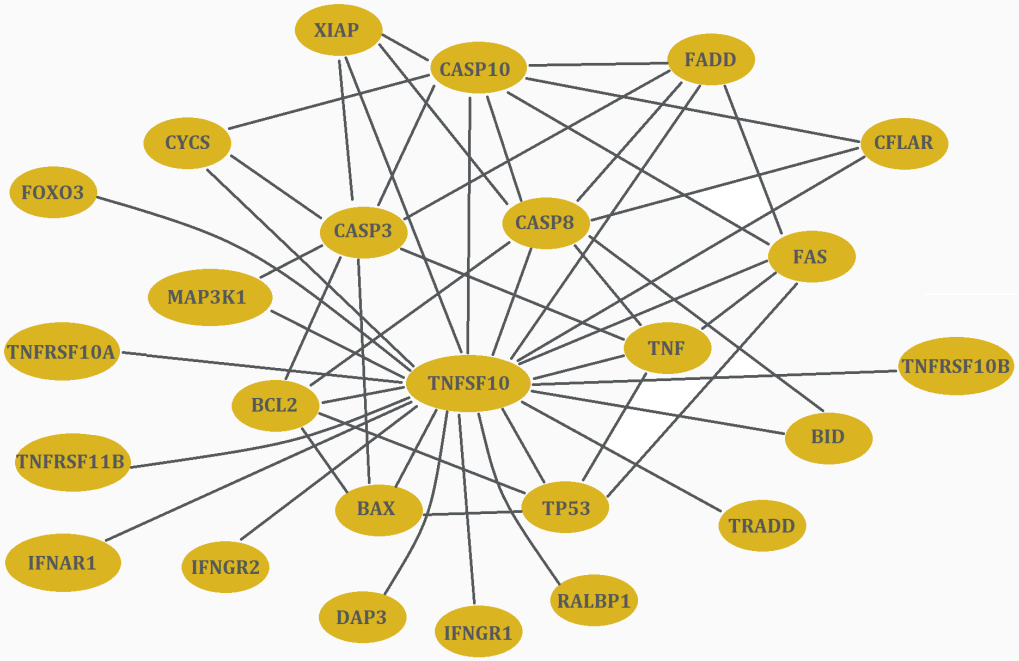
EXAMPLE OUTPUT DATA

22	16390584	.	C	T	.	PASS	AC=23;AN=5096;DP=82786;AF=0;EAS_AF=0;EUR_AF=0;AFR_AF=0.01;AMR_AF=0;SAS_AF=0;EX_TARGET;VT=SNP;NS=2548	GT	0|0
22	16390592	.	A	G	.	PASS	AC=1386;AN=5096;DP=86545;AF=0.27;EAS_AF=0.37;EUR_AF=0.34;AFR_AF=0.08;AMR_AF=0.33;SAS_AF=0.32;EX_TARGET;VT=SNP;NS=2548	GT	0|1
22	16390594	.	C	G	.	PASS	AC=1;AN=5096;DP=87990;AF=0;EAS_AF=0;EUR_AF=0;AFR_AF=0;AMR_AF=0;SAS_AF=0;EX_TARGET;VT=SNP;NS=2548	GT	0|0
22	16390595	.	A	G	.	PASS	AC=1;AN=5096;DP=88574;AF=0;EAS_AF=0;EUR_AF=0;AFR_AF=0;AMR_AF=0;SAS_AF=0;EX_TARGET;VT=SNP;NS=2548	GT	0|0
22	16390599	.	G	T	.	PASS	AC=1;AN=5096;DP=90114;AF=0;EAS_AF=0;EUR_AF=0;AFR_AF=0;AMR_AF=0;SAS_AF=0;EX_TARGET;VT=SNP;NS=2548	GT	0|0

SIMULATED PEDIGREE



Different families may segregate different *c*RVs residing in a set of interacting genes or a pathway.



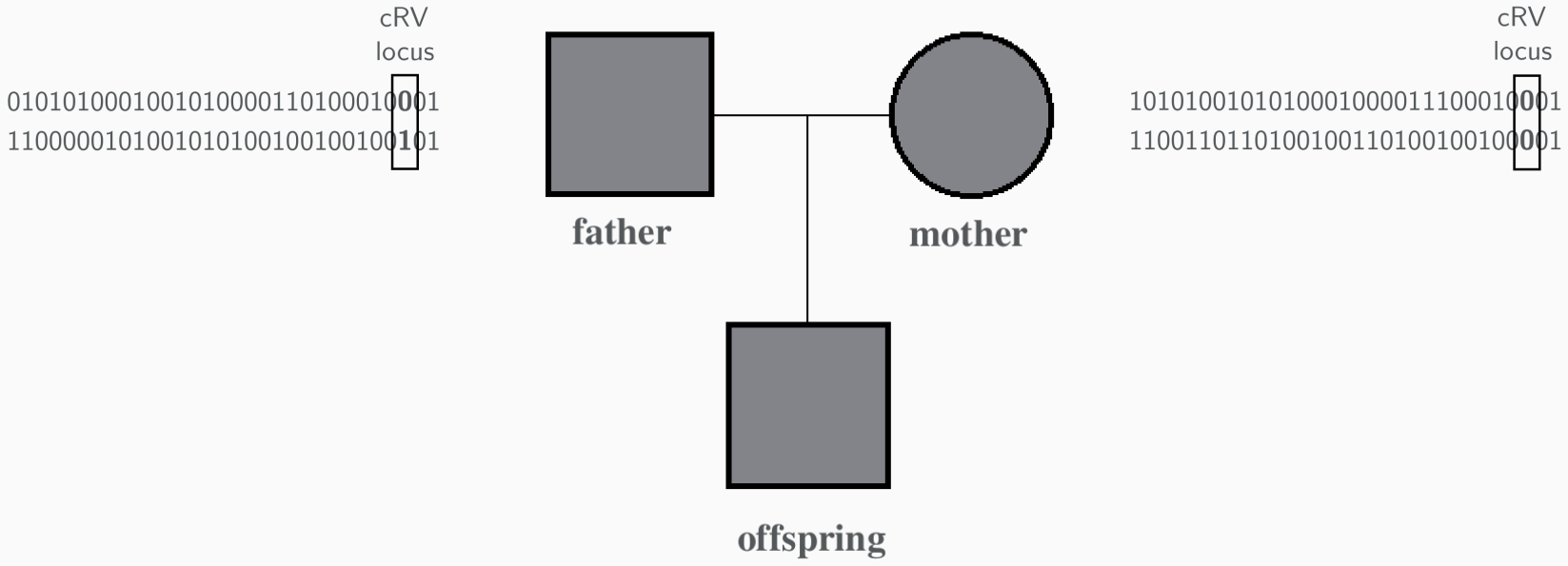
- We specify a pool of *c*RVs from which to sample familial *c*RVs, so that different families can segregate different *c*RVs.

*c*RV
locus

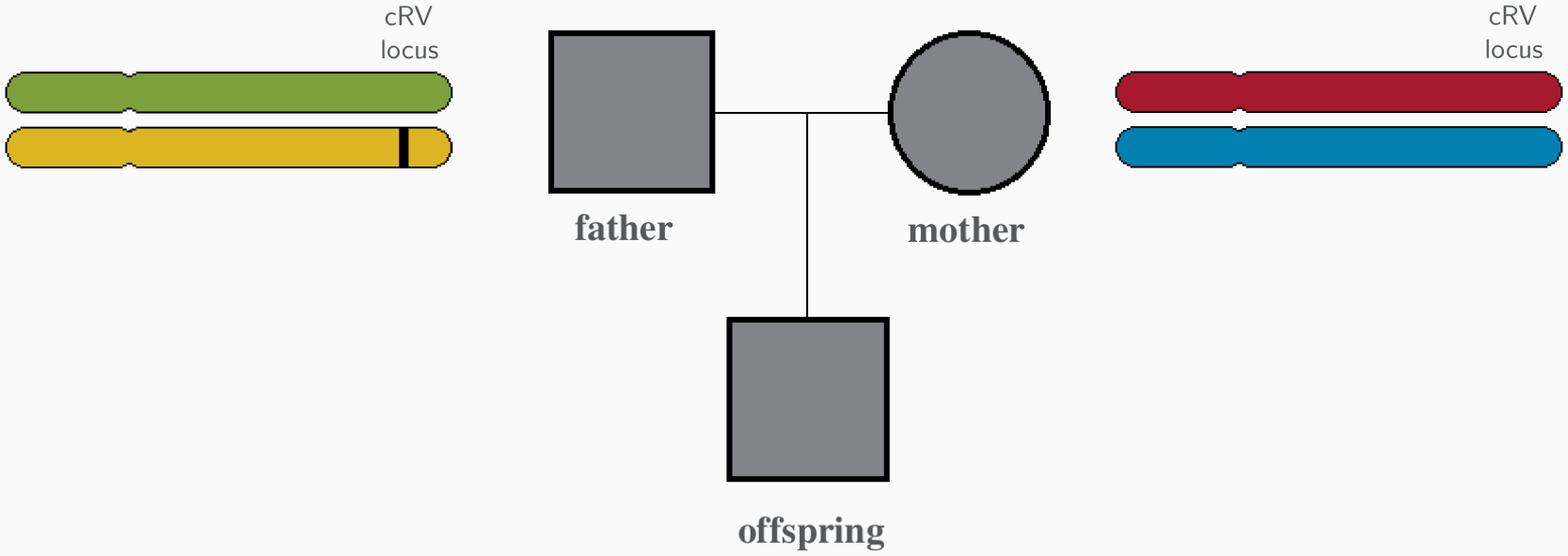
```
0101011011100000010101001 00
0001010000100110001101001 11
0100011010001010011100011 10
0001010100110100100101110 00
0101011011100011010101000 1
0101011011100000010101000 1
```

Founder haplotypes are sampled from the population distribution of haplotypes conditioned on the founder's *c*RV status at the familial disease locus.

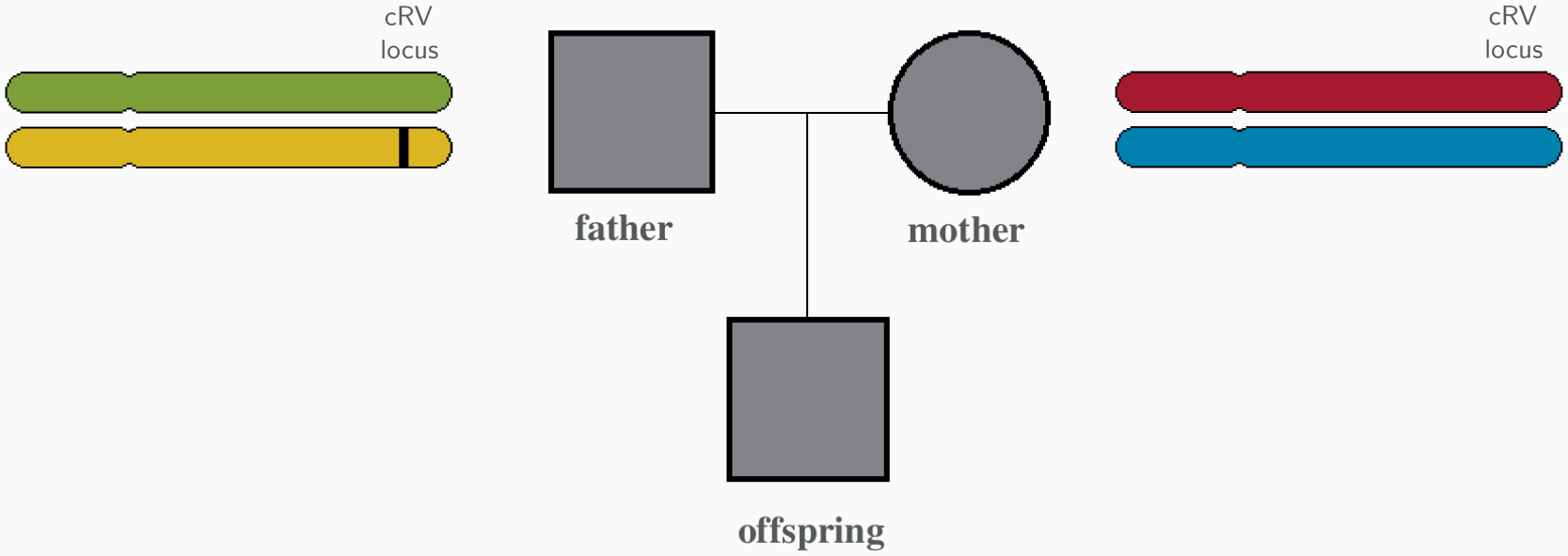
SEQUENCE DATA FOR OFFSPRING



SEQUENCE DATA FOR OFFSPRING



SEQUENCE DATA FOR OFFSPRING

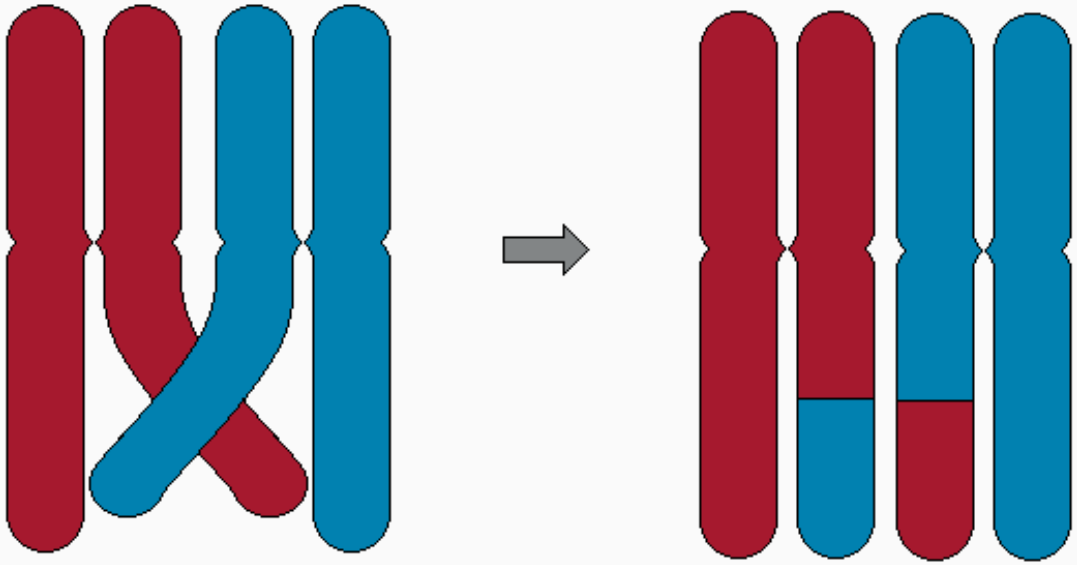


Given the *c*RV status of each pedigree member we perform a conditional gene drop to simulate inheritance.

- Simulate genetic recombination among parental haplotypes.
- Sample the inherited gamete conditionally on *c*RV status.

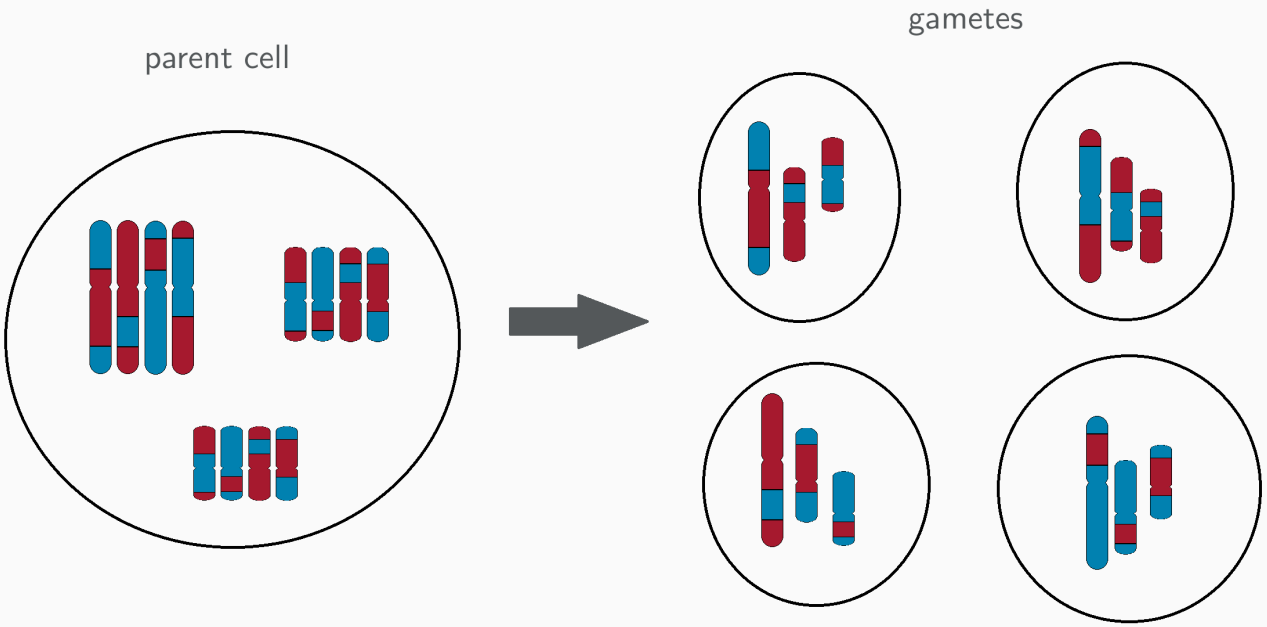
CROSSOVER EVENTS

Each parent's haplotypes participate in recombination, or crossover, events whereby genetic material is exchanged between them.

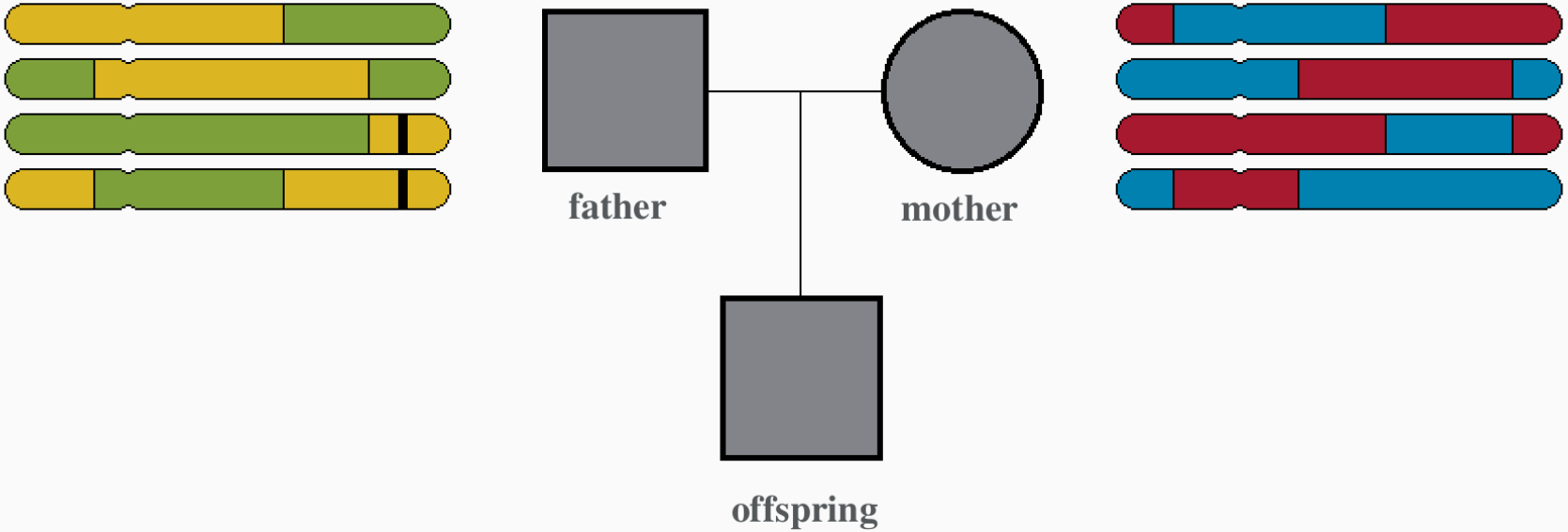


We model the locations of crossover events as stochastic point process with a gamma renewal density [12].

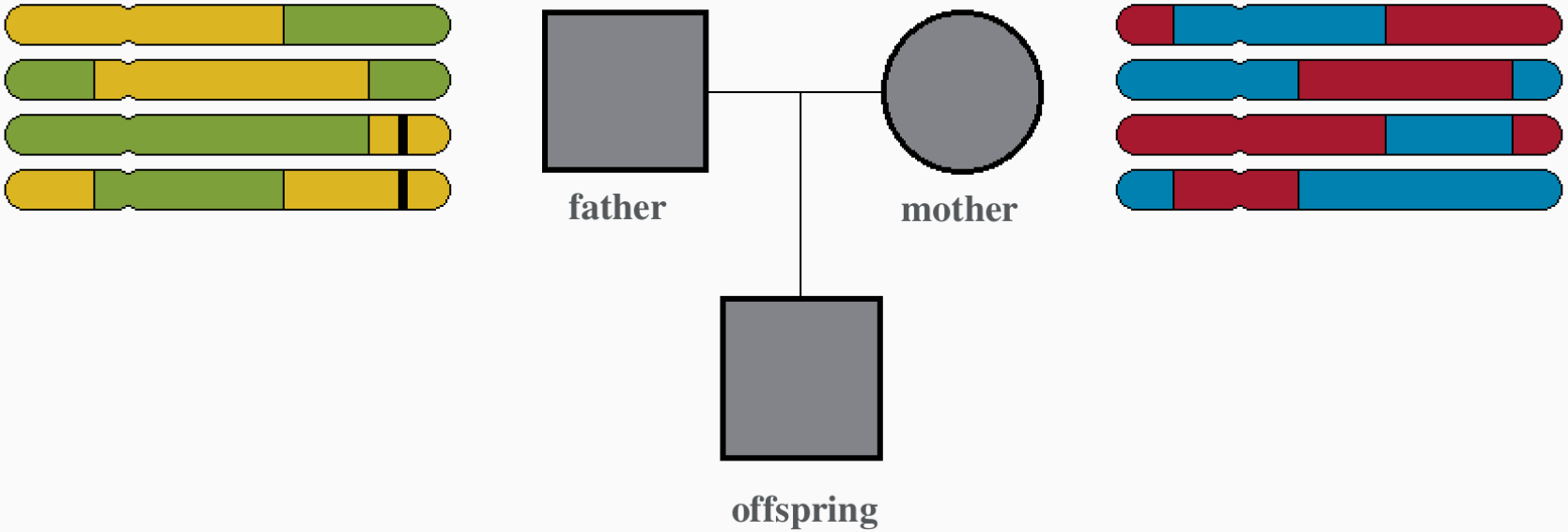
FORMATION OF GAMETES



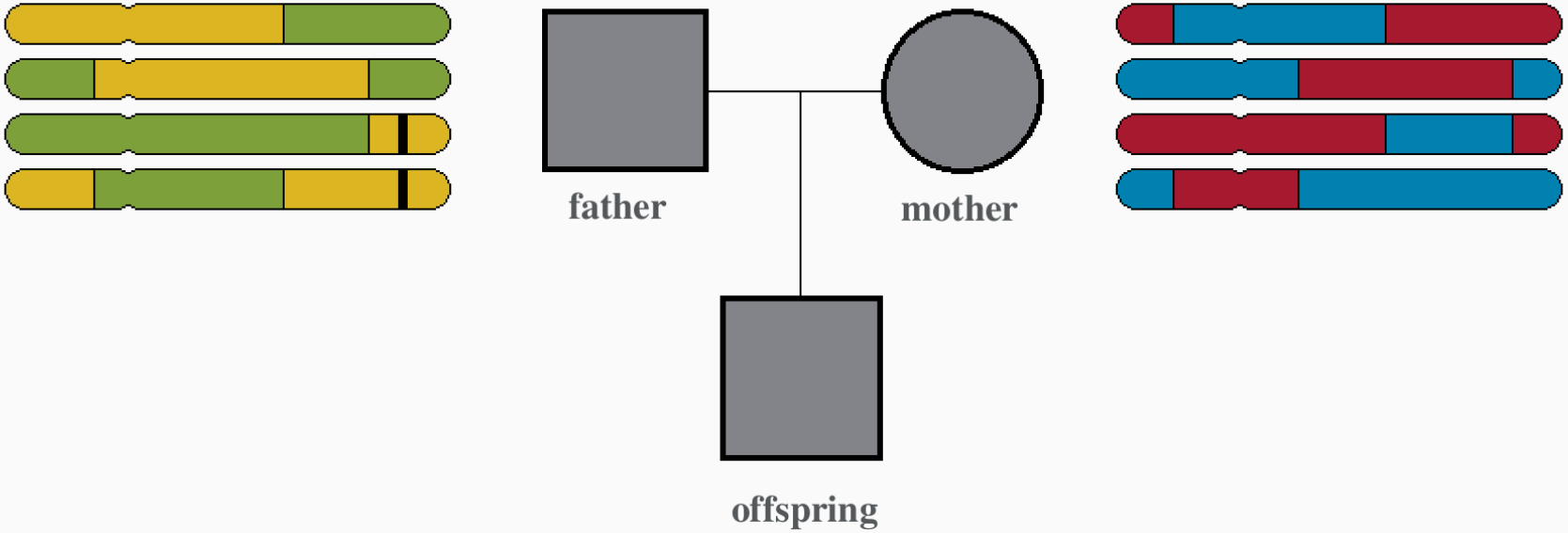
- To simulate the formation of gametes we assume that homologous chromatids are assigned to one of four gamete cells with equal probability.
- This assignment occurs independently for non-homologous chromosomes.



Case 1: If **both the parent and offspring carry the cRV** we sample the inherited gamete from those that carry the *c*RV.



Case 2: If **the parent carries the cRV but the offspring does not**, we sample the inherited gamete from those that do not carry the *c*RV.



Case 3: If a **parent is not a carrier of the cRV**, we sample the inherited gamete from the four parental gametes.

Size Comparison of Genetic Data Objects

Chromosome	R package	Object Name	Size
1	vcfR	vcfR	2.6 Gb
1	sim1000G	vcf	NA (too large)
1	SimRVSequences	SNVdata	157.8 Mb
21	vcfR	vcfR	306.5 Mb
21	sim1000G	vcf	25.8 Mb
21	SimRVSequences	SNVdata	19.3 Mb

Featured genetic data includes SNVs from the exonic regions of chromosomes 1 and 21.

R Packages:

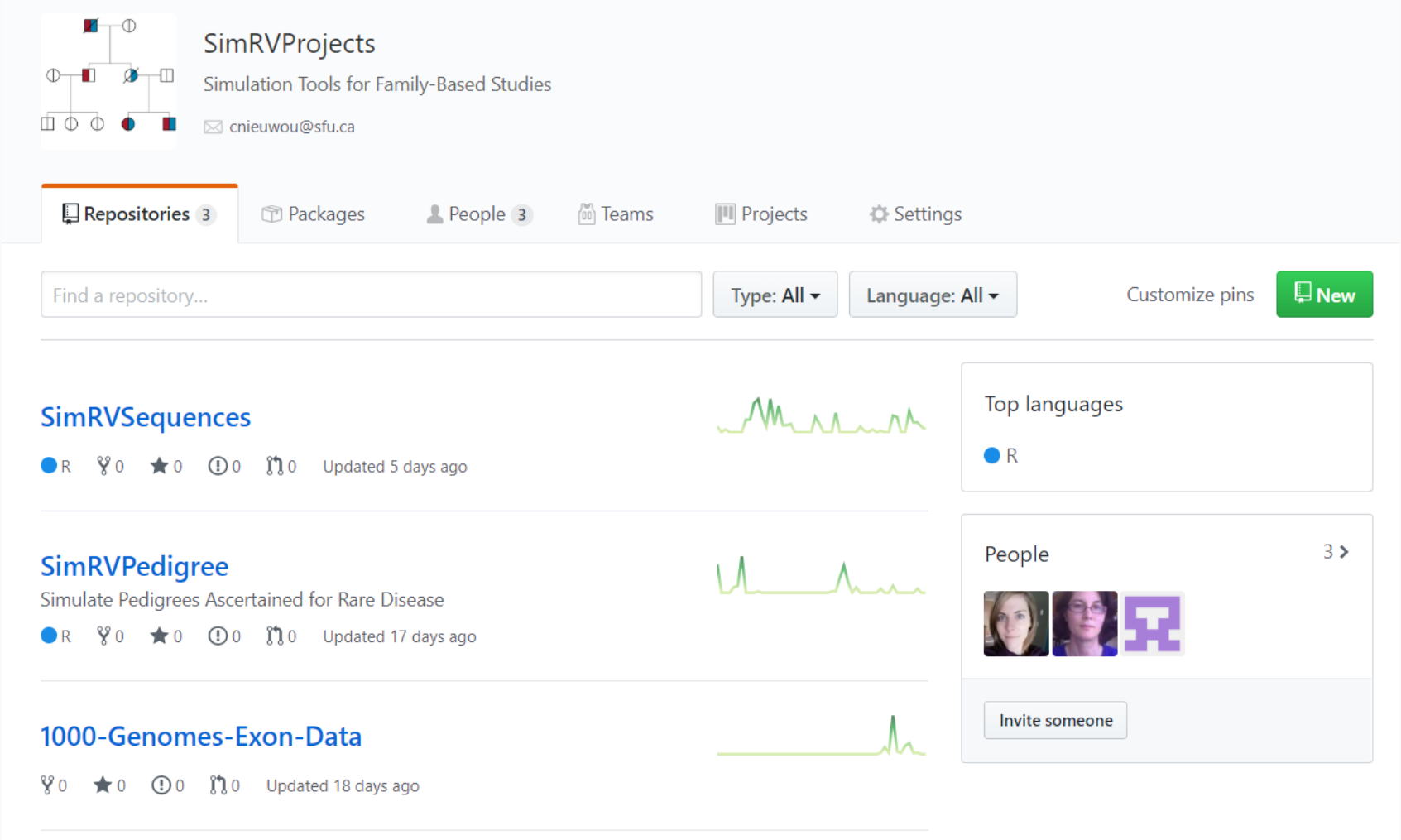
- Nieuwoudt, C., Graham, J., (2017) *SimRVPedigree: Simulate Pedigrees Ascertained for a Rare Disease*. R package version 0.4.0. https://CRAN.R-project.org/package=SimRVPedigree.
- Nieuwoudt, C., Graham, J., (2019) *SimRVSequences: Simulate Genetic Sequence Data for Pedigrees*. R package version 0.1.3. https://CRAN.R-project.org/package=SimRVSequences.

Manuscripts:

- Nieuwoudt, C., Jones, S.J., Brooks-Wilson, A., Graham, J. (2018). Simulating Pedigrees Ascertained for Multiple Disease-Affected Relatives. *Source Code for Biology and Medicine* 13:2.
- *In Submission:* Nieuwoudt, C., Brooks-Wilson, A., Graham, J. (2019). SimRVSequences: An R Package to Simulate Genetic Sequence Data for Pedigrees. *Bioinformatics*.

FIND US ON GITHUB

Source code and data are available on GitHub:
https://github.com/simrvprojects



Supervisor: Jinko Graham

Lymphoid Cancer Families Study
(PI Angela Brooks-Wilson)