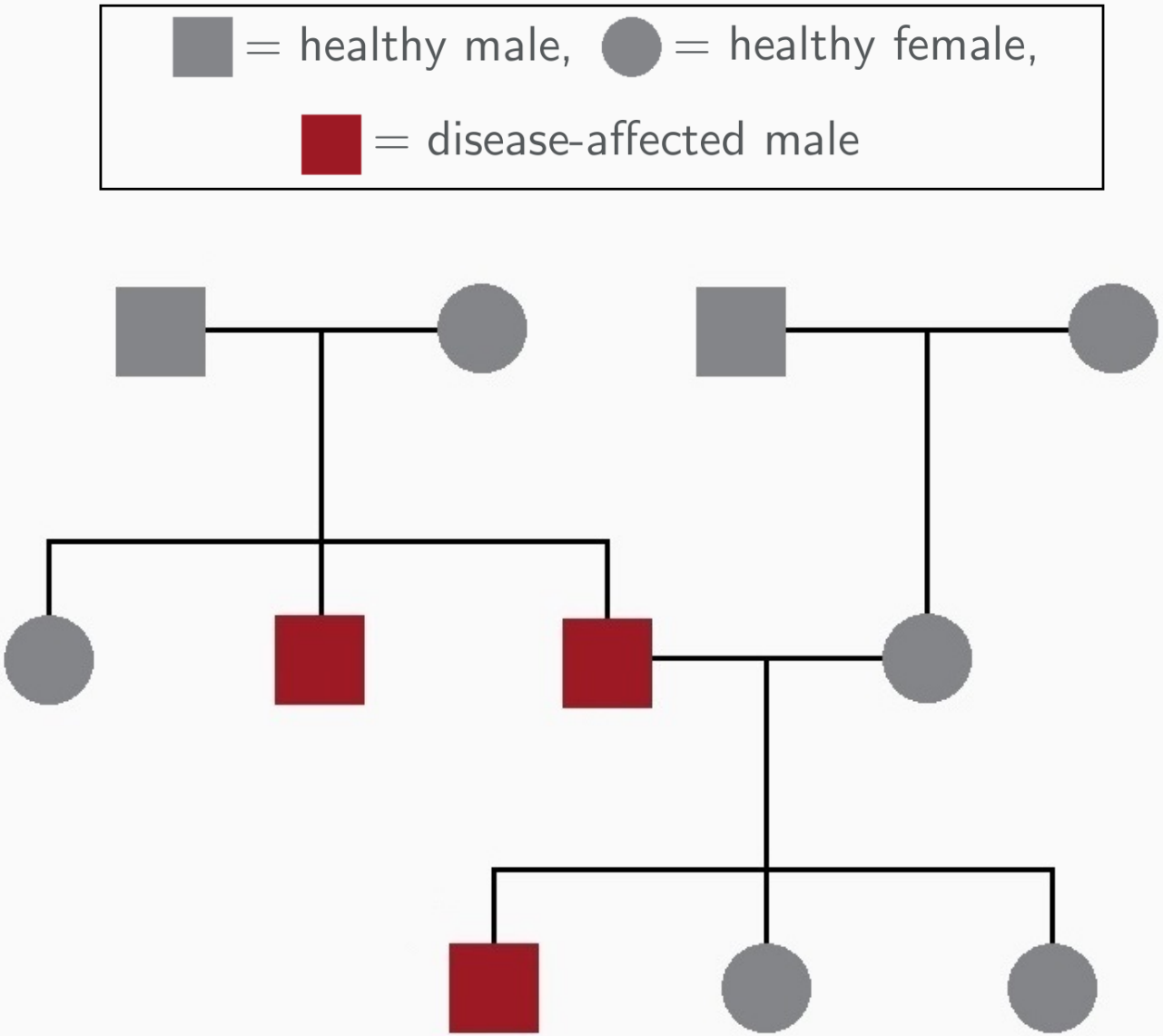
SIMULATING FAMILY STUDIES WITH WHOLE-EXOME SEQUENCING OF MULTIPLE DISEASE-AFFECTED RELATIVES

Christina Nieuwoudt
Supervisor: Jinko Graham

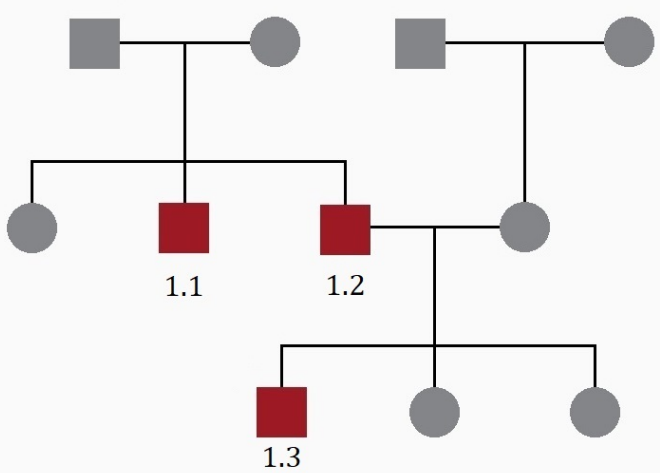
6 August 2018

Simon Fraser University
Department of Statistics

MOTIVATION



SINGLE-NUCLEOTIDE VARIANT (SNV) DATA



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

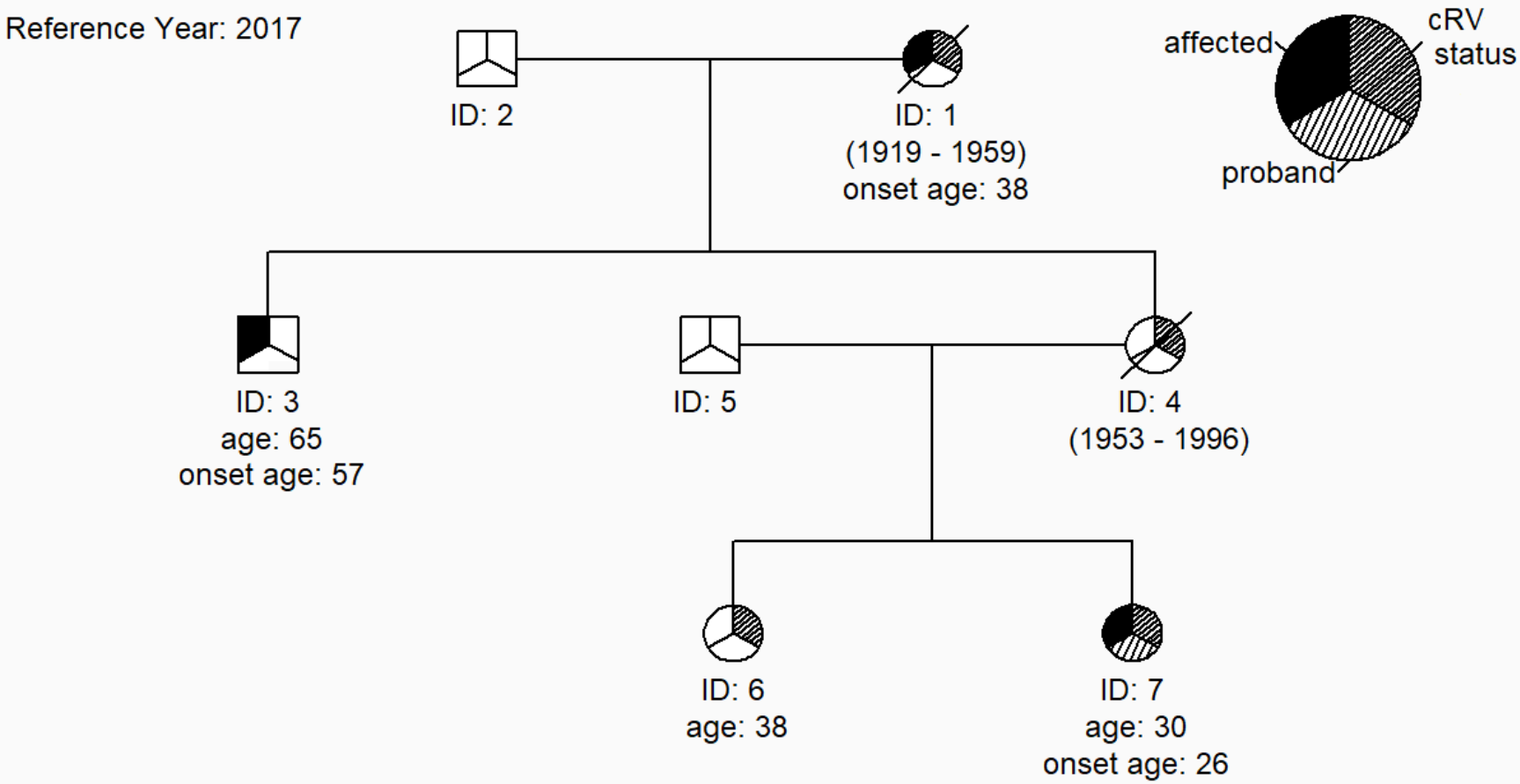
Mutated alleles are coded as 1 and wild type alleles are coded by 0.

To simulate sequence data for a pedigree we require:

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

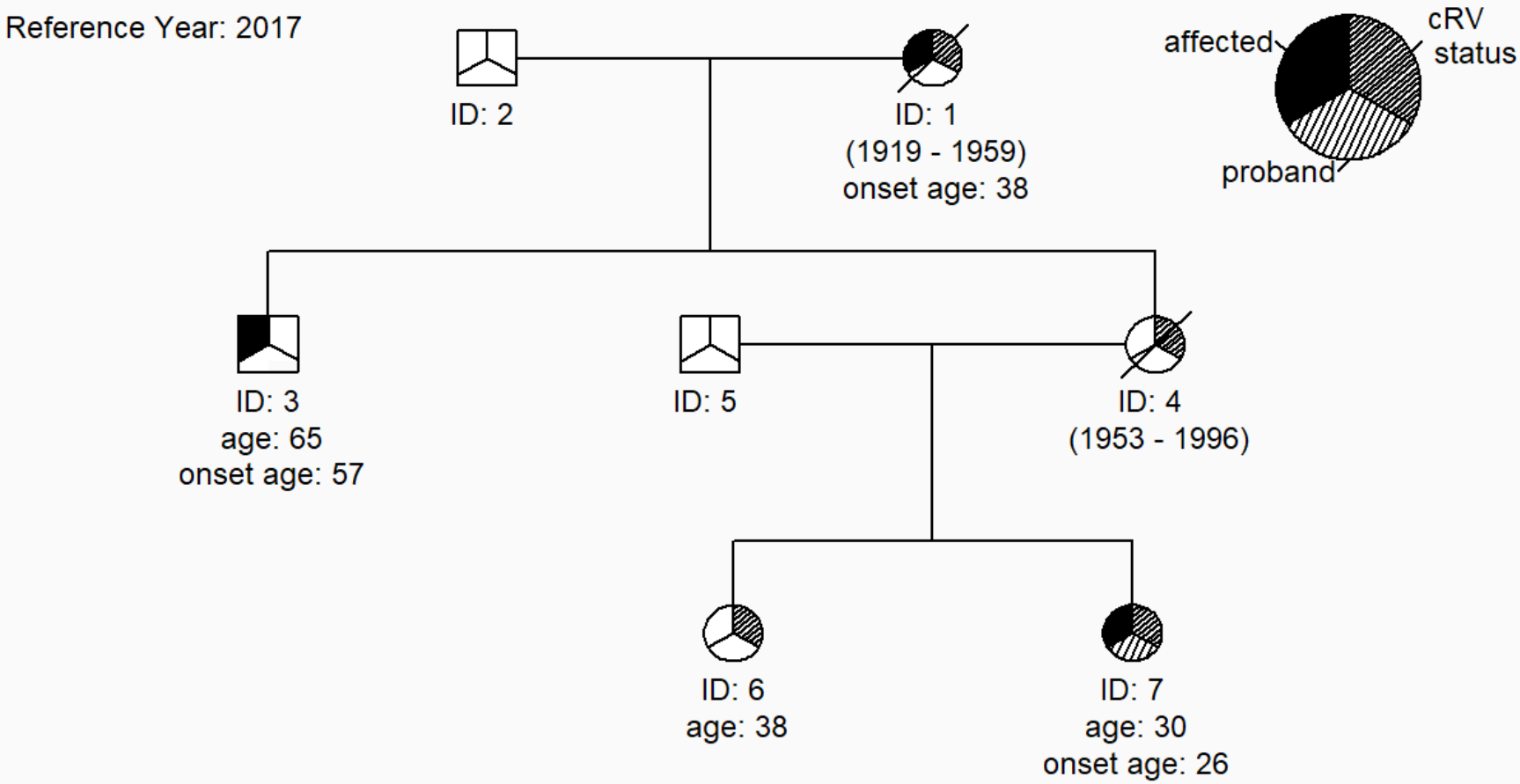
Presently, we streamline the use of exon-only SNV data simulated by SLiM 2.0 and pedigrees simulated by `SimRVPedigree`.

PEDIGREE SIMULATION



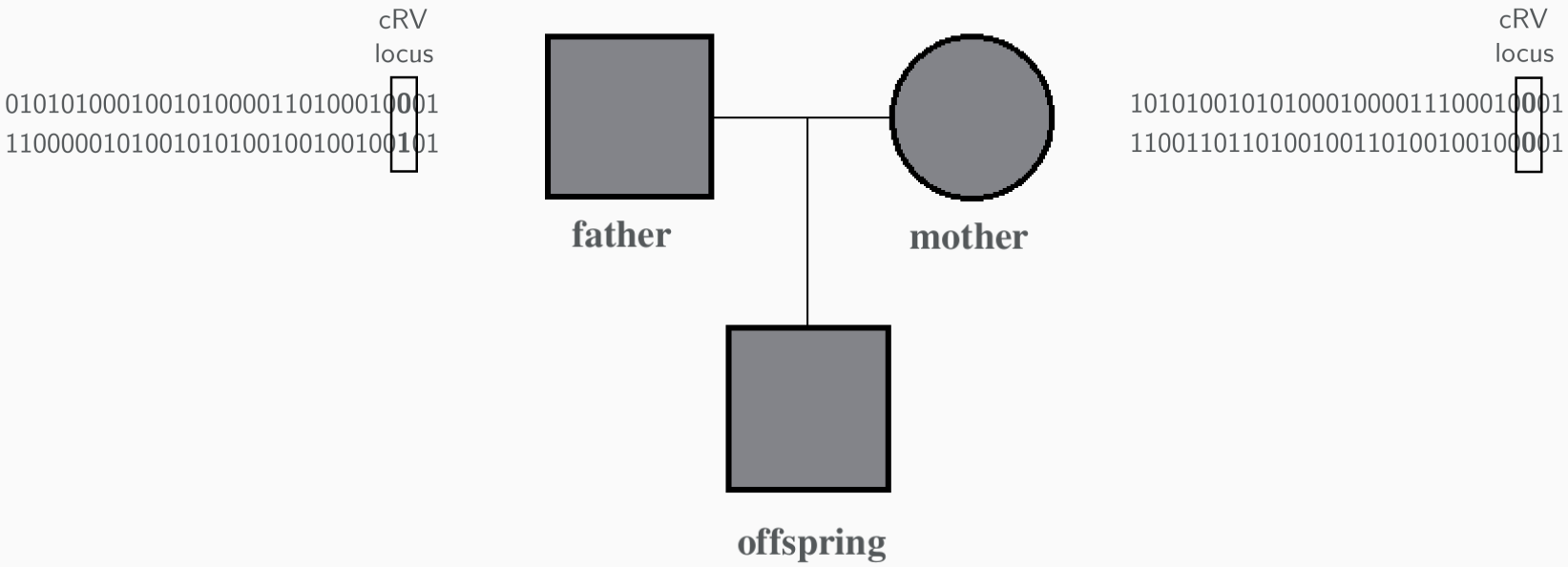
- We simulate life events by way of a competing-risk model which makes use of age-specific hazard rates of disease-onset and death provided by the user.
- Possible life events include: disease-onset, reproduction, and death.

PEDIGREE SIMULATION

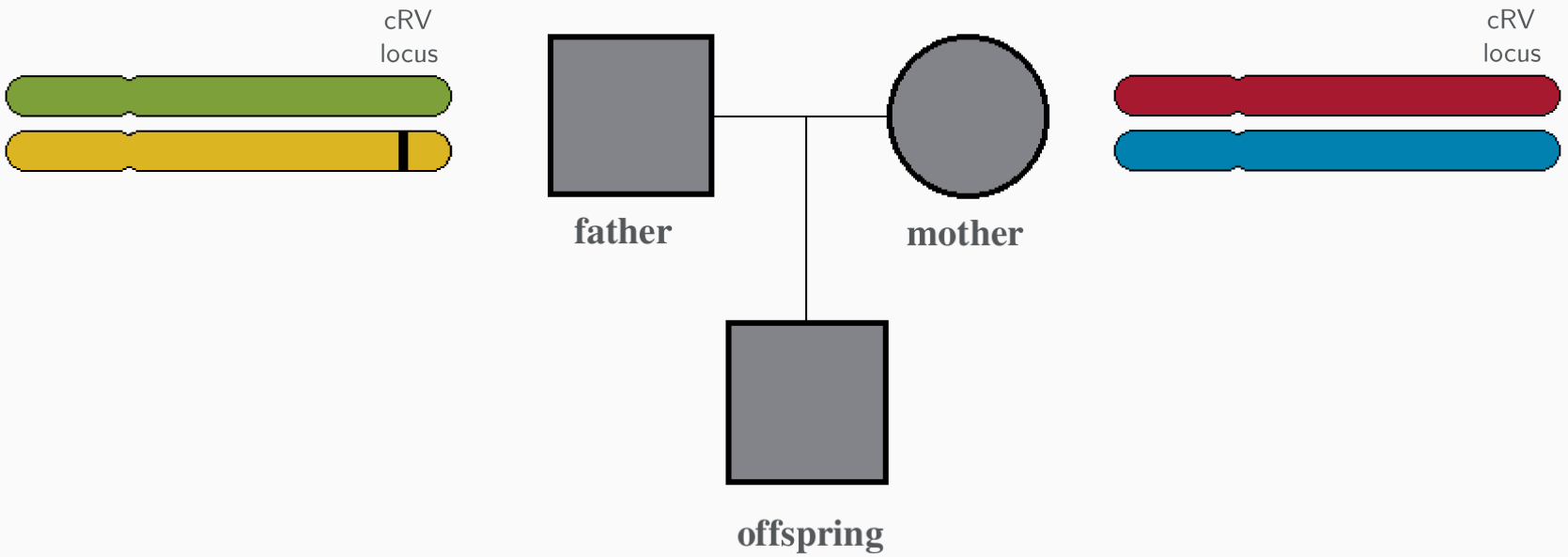


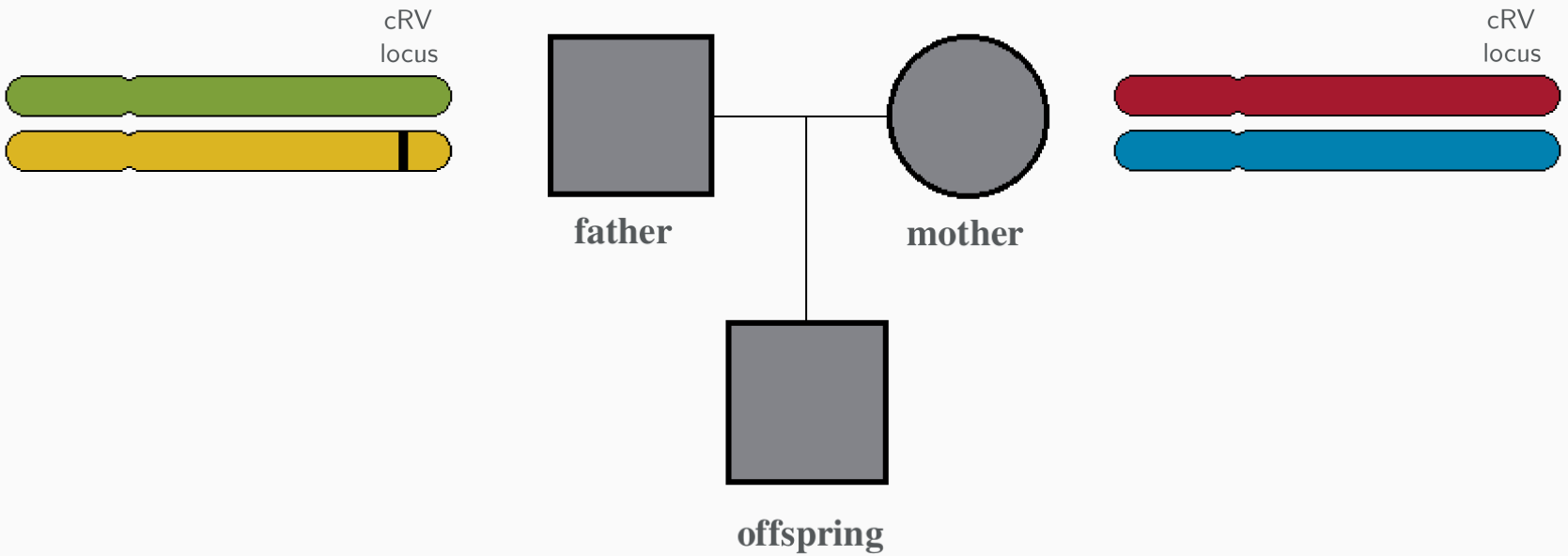
- At the individual level, disease onset is influenced by the presence (or absence of) a **causal rare variant**, or **cRV**.
- We assume the *c*RV has been introduced by no more than one founder and transmit it according to Mendel's laws.

SEQUENCE DATA FOR OFFSPRING



SEQUENCE DATA FOR OFFSPRING

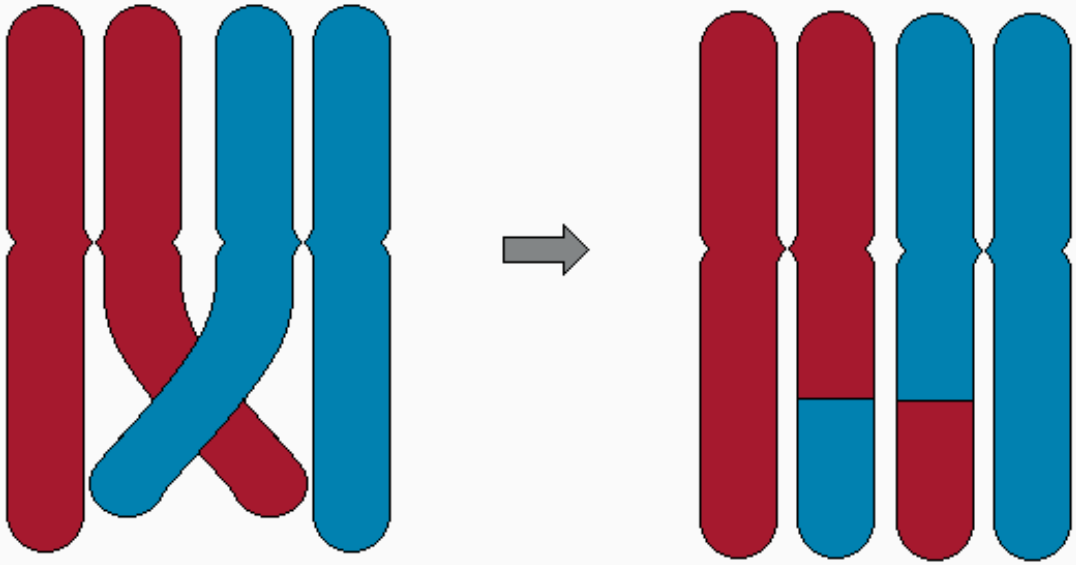




Given the *c*RV status of each pedigree member we perform a conditional gene drop to simulate inheritance.

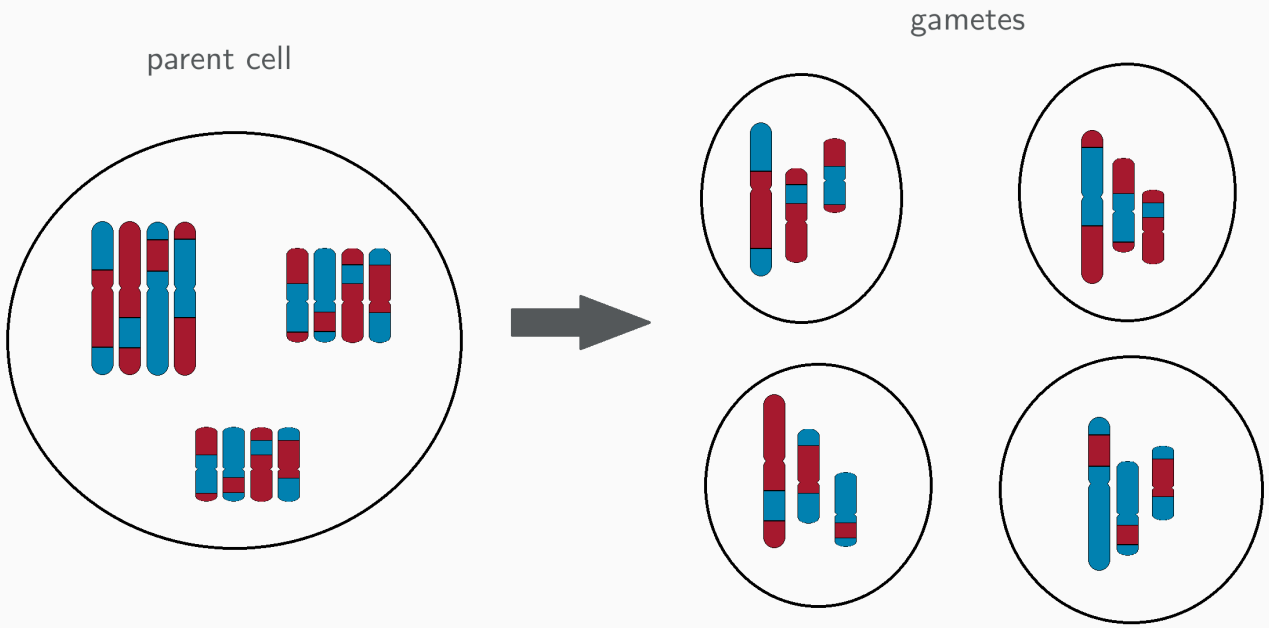
- Simulate genetic recombination among parental haplotypes.
- Sample the inherited gamete conditionally on *c*RV status.

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

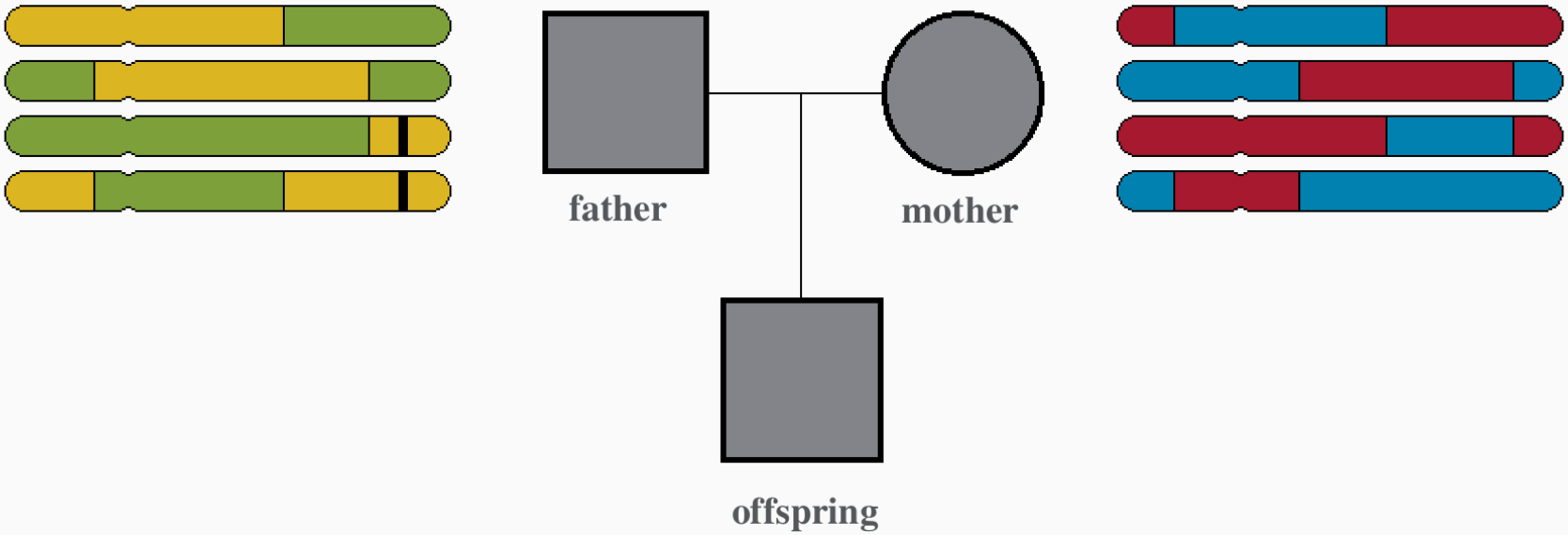


We model the locations of crossover events as stochastic point process with a gamma renewal density [12].

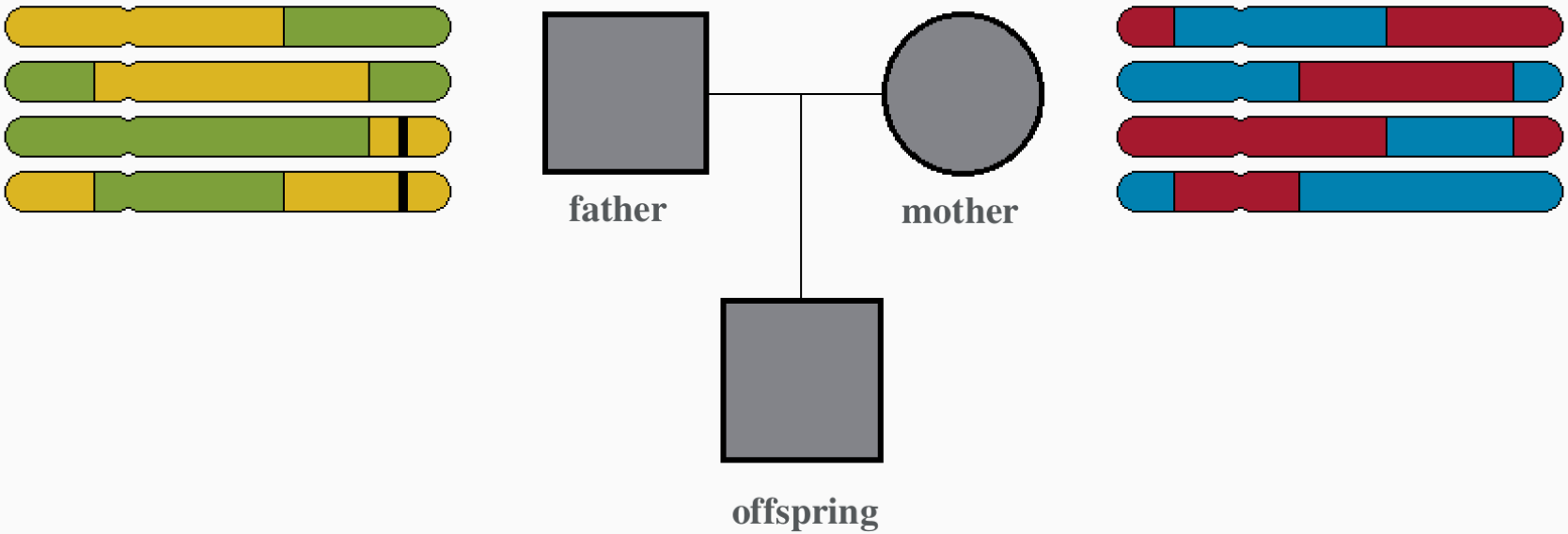
FORMATION OF GAMETES



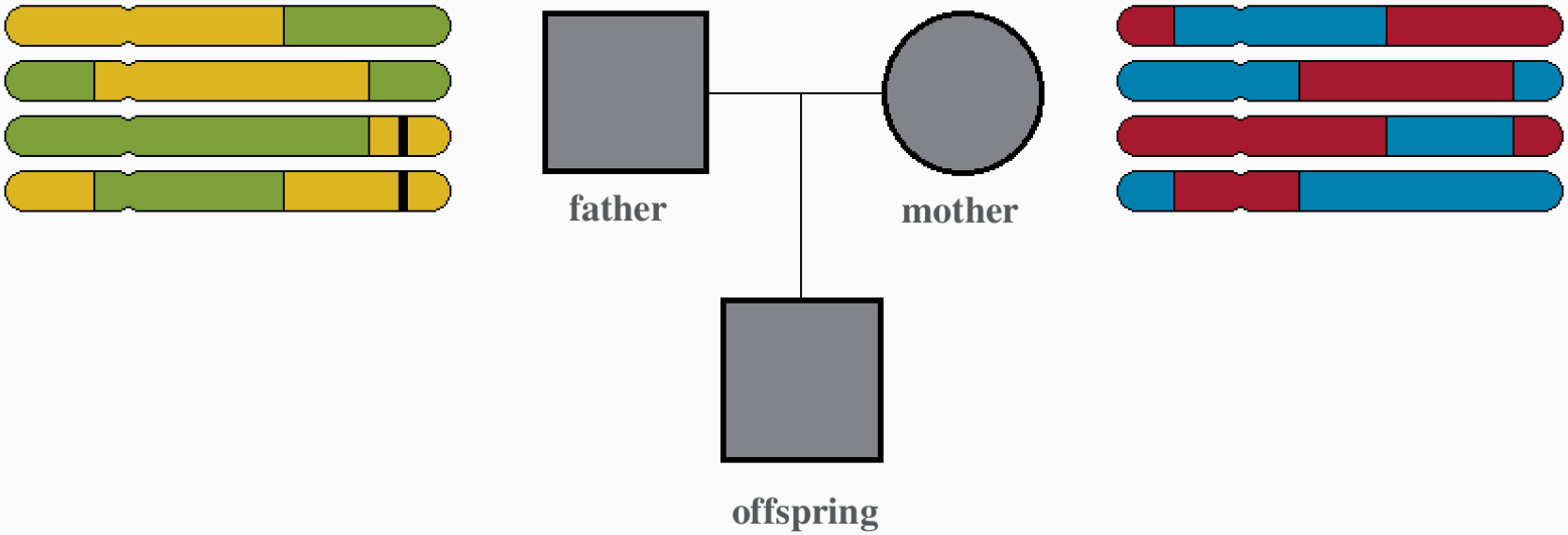
- To simulate the formation of gametes we assume that homologous chromatids are assigned to one of four gamete cells with equal probability.
- This assignment occurs independently for non-homologous chromosomes.



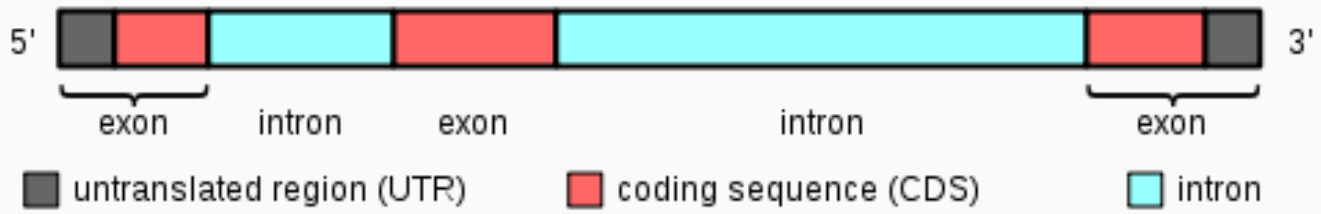
Case 1: If **both the parent and offspring carry the cRV** we sample the inherited gamete from those that carry the *c*RV.



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.



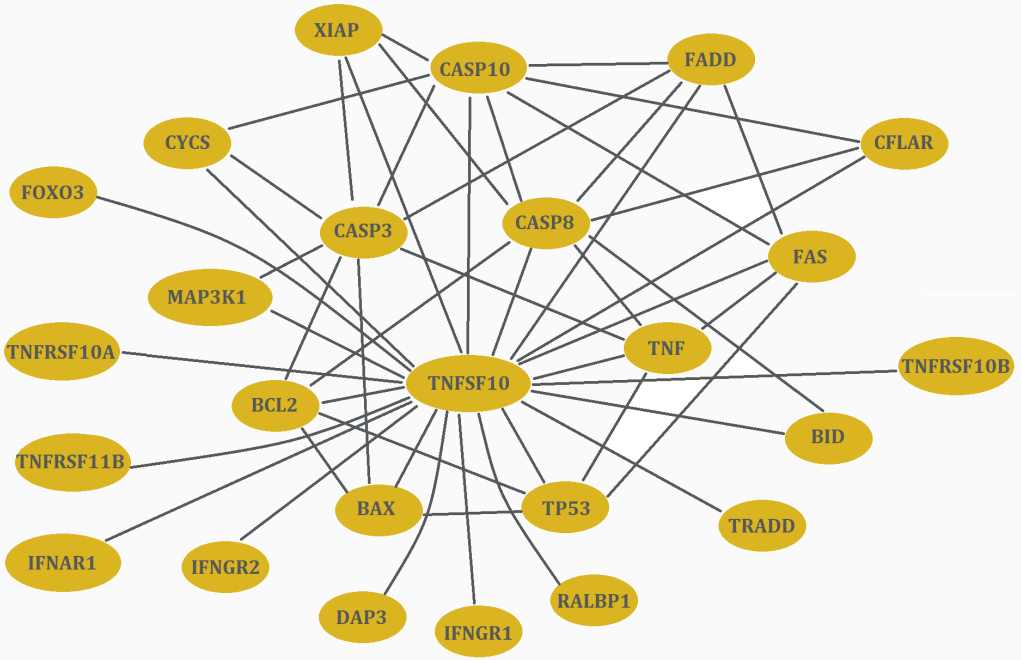
Case 3: If a **parent is not a carrier of the cRV**, we sample the inherited gamete from the four parental gametes.



graphic by Daycd, at the English Wikipedia Project, distributed under a CC-BY 2.0 license

- We assume founders are unrelated and represent a random sample from a global population of individuals.
- Exon-only sequence data may be obtained from:
  - publicly available sequence data (1000 genomes project),
  - a coalescent simulator (msprime or fastsimcoal), or
  - a forward-in-time evolutionary simulator (SLiM 2.0).

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

Task	Time
Simulating exon-only sequence data with SLiM 2.0 • 10,000 individuals • 44,000 generations • recombination rate 1×10^{-8} • mutation rate 1×10^{-8}	$>$ 24 hrs
Simulating pedigrees with `SimRVPedigree`. • 200 pedigrees • 2 or more disease-affected relatives	1 - 90 mins
Simulating genetic data for the disease-affected relatives with `SimRVSequences`.	53 sec

ACKNOWLEDGEMENTS

Supervisor: Jinko Graham

Lymphoid Cancer Families Study
(PI Angela Brooks-Wilson)

REFERENCES

[1] 1000 Genomes Project (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467:1061-1073.

[2] Haller, B. C., Messer P. W., SLiM 2: Flexible, Interactive Forward Genetic Simulations. *Molecular Biology and Evolution*, 34(1):230-240.

[3] Harris, K., Nielsen, R., (2016). The Genetic Cost of Neanderthal Introgression. *Genetics*, 203(2):881-891.

[4] Karolchik, D. *et al.*, (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. doi: 10.1093/nar/gkh103

[5] Kent, W. J., *et al.*, (2002). The human genome browser at UCSC. *Genome Res*, 12(6):996-1006.

[6] Lange, K., *Mathematical and Statistical Methods for Genetic Analysis*. Springer, NewYork. 2^{nd} edition.

[7] Nieuwoudt, C., Graham, J., (2017) *SimRVPedigree: Simulate Pedigrees Ascertained for a Rare Disease*. R package version 0.1.0. https://CRAN.R-project.org/package=SimRVPedigree.

[8] Nieuwoudt, C., *et al.* (2017). Simulating Pedigrees Ascertained for Multiple Disease-Affected Relatives. *bioRxiv* 234153.

[9] Poon, H., *et al.*, (2014). Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics*, 30:2840-2842.

[10] Takahata, N., (1993). Allelic genealogy and human evolution. *Mol Biol Evol*, 10:2-22.

[12] Voorrips, R. E., Maliepaard, C. A, (2012) The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinformatics*, 13:248.