PLINK

A genotype/phenotype analysis tool

Before starting…

- Basic UNIX commands will be needed.
  - cd
  - ls
  - more
- Short cheat sheet can be downloaded.

What is PLINK?

Free, open-source, THE standard command-line program…

to perform basic, large-scale

genotype/phenotype analysis

in a computationally efficient manner.

What can I do with PLINK?

- Manage genomic data with file format conversion
- Filter by quality, genomic location, list of SNPs, missing, allele frequency, and correlation.
- Perform basic statistics
- Calculate population genetics metrics
- (Association analysis)

How to use PLINK?

Always consult the LOG file (printed on console as well as in .log)

PLINK has no memory

Write all commands in one line, or change line with “\”

Consult the web documentation (https://www.cog-genomics.org/plink/1.9/)

Standard plink files

- 2 file types: **.ped** and **.map**

- **.ped** contains information about family, phenotype and genotype

.ped

Family	Sample	Father	Mother	Sex	Phenotype	SNP1A	SNP1B	SNP2A	SNP2A	…
HG01500	HG01500	0	0	0	-9	C	C	G	G	
HG01501	HG01501	0	0	0	-9	C	C	T	G	

- **.map** contains information about marker location

.map

Chr	Marker	cM	Position
20	rs56993397	0	800648
20	rs57400069	0	802019

Compressed plink files

- Compressed file format: .bed, accompanied by **.bim** and **.fam**
- **.bed** is binary **.ped** file about genotype only

Family	Sample	Father	Mother	Sex	Phenotype	SNP1A	SNP1B	SNP2A	SNP2A	…
HG01500	HG01500	0	0	0	-9	1	1	1	1	
HG01501	HG01501	0	0	0	-9	1	1	0	1	

.fam .bed

- **.bim** contains information about marker location and genotype

.bim

Chr	Marker	cM	Position	Allele A	Allele B
20	rs56993397	0	800648	T	C
20	rs57400069	0	802019	T	G

vcf file

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##simulateGenotypeData=1.1
##source=simulateGenotypeDataFrom1000G
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##contig=<ID=20>
##bcftools_viewVersion=1.8+htslib-1.8
##bcftools_viewCommand=view -r 20:800000-3200000 -Oz -o chr20.chunk1.vcf.gz chr20.FINAL.vcf.gz; Date=Sun Apr 22 11:36:10 201
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG01500	HG01501	...
20	800648	rs56993397	C	T	100	PASS	.	GT	0/0	0/0	
20	802019	rs57400069	G	T	100	PASS	.	GT	0/0	0/1	

Set-up

- https://www.cog-genomics.org/plink2
- A OS X (64-bit) version is already downloaded for you in the GroupBio.
- Create a folder in your directory, and name it “my_plink_folder”.
- Move the “plink” executable and “ALL.chr20.chunk1.vcf.gz” file in.

Input file

Requires files of same root name

- **-vcf** chr20.chunk1 (.vcf required)
- **-file** chr20.chunk1 (.ped and .map required)
- **-bfile** chr20.chunk1 (.bed, .bim and .fam required)

Output file

- Default name: plink
- New root name needs to be indicated by: **-out**
- Options to indicate file format:
  - **-make-bed** >> .bed, .bim and .fam
  - **-recode** >> .ped and .map
  - **-recode vcf** >> .vcf

Exercise 1: Conversion between file formats

- vcf
- ped, map
- bed, bim, fam

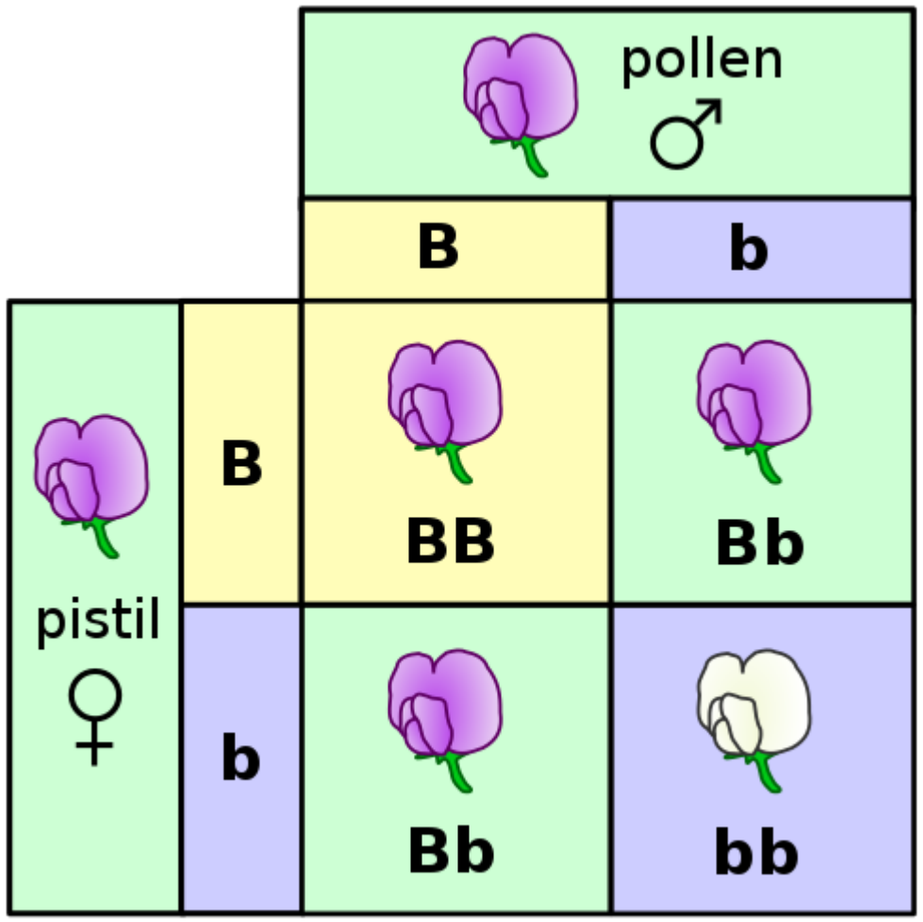
(example usage: plink -vcf xxx -recode)

Exercise 2: filtering (output is data)

- **-extract** mysnps.txt
- how many variants removed?
  - **-maf** 0.05
  - **-geno** 0.05
  - **-hwe** 1e-3
- how many individuals removed?
  - **-mind** 0.01

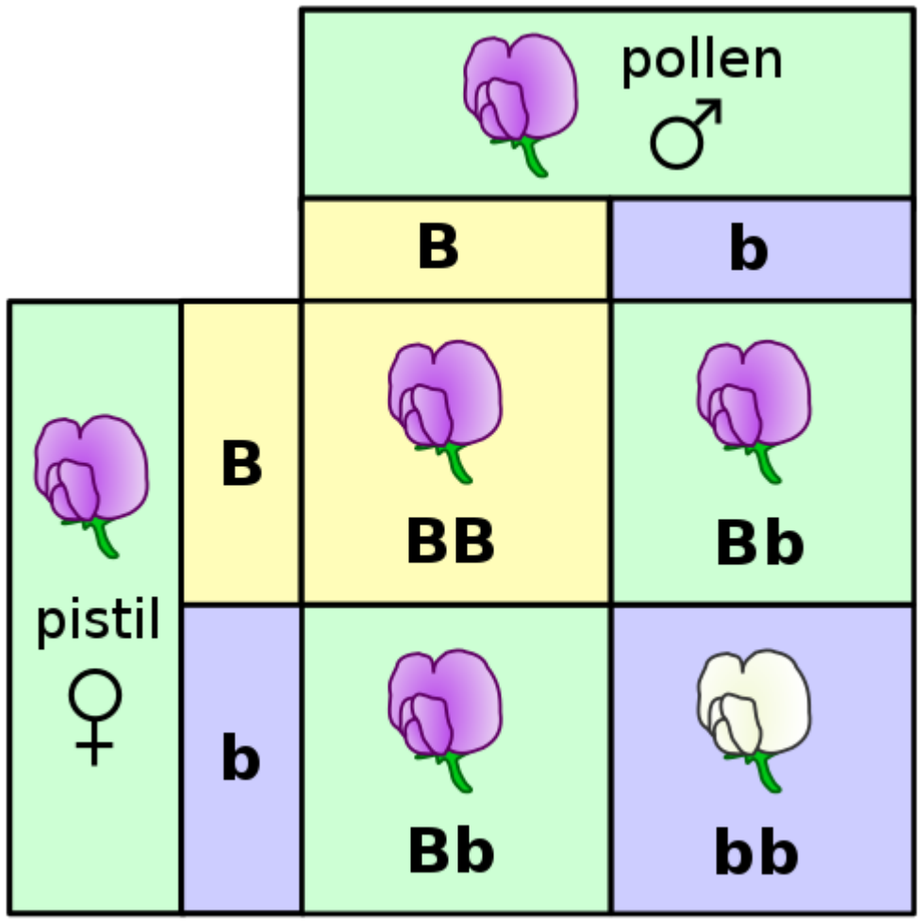
(you have 15min…)

Mendel's Law



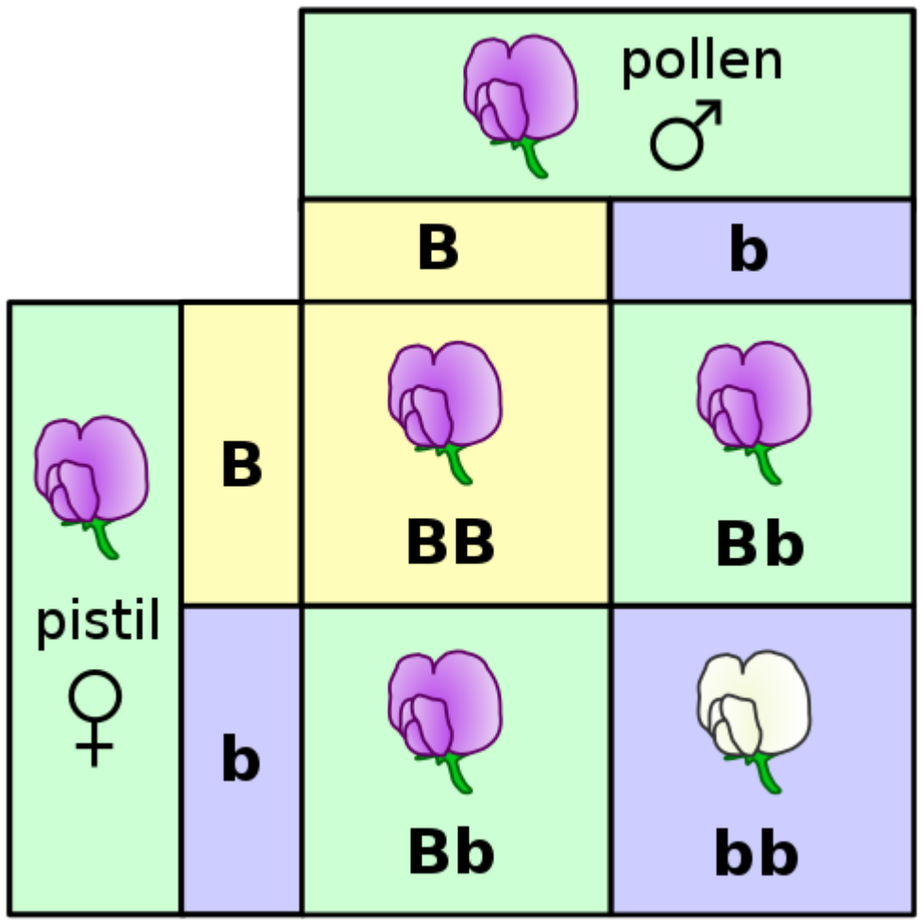
- Law of segregation
- Law of independent assortment
- Law of Dominance

Mendel's Law



- Law of segregation
- *Law of independent assortment*
- Law of Dominance

Hardy Weinberg Equilibrium



✳ *Law of independent assortment*

p(BB) = p(B)^2

p(Bb) = 2 * p(B) * p(b)

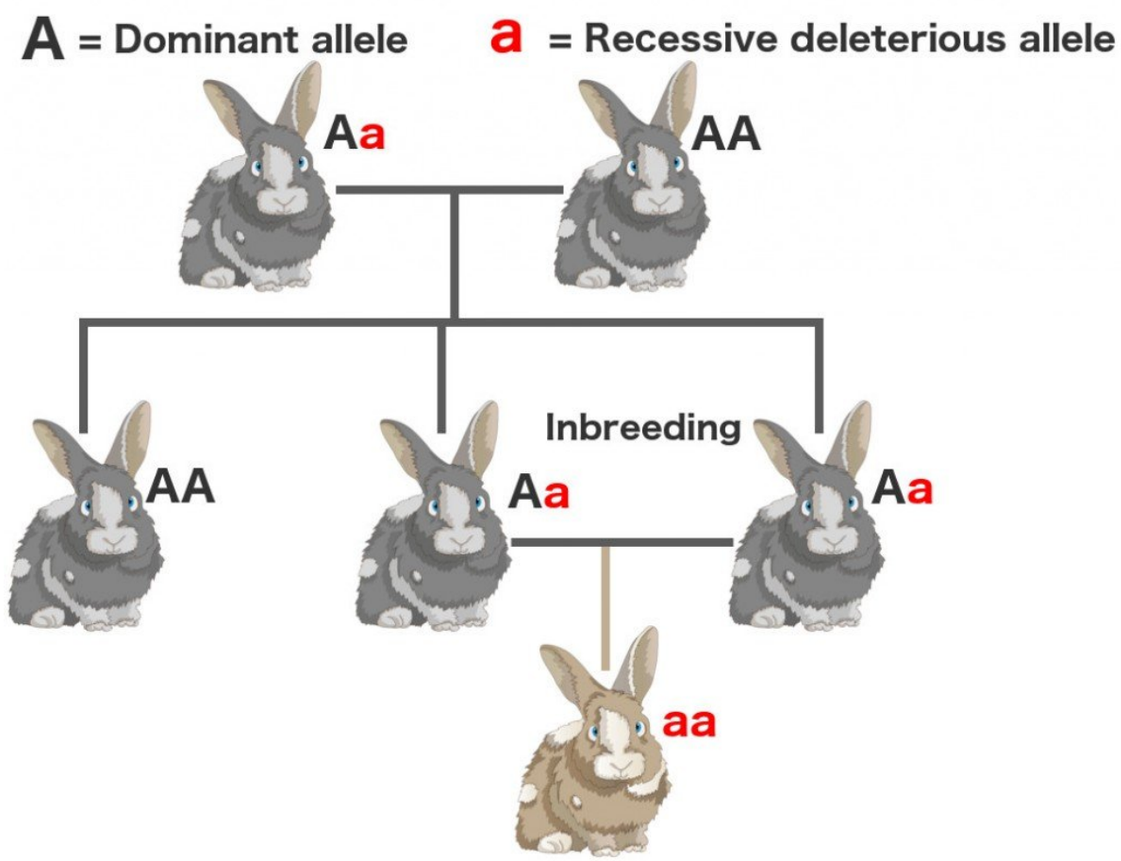
p(bb) = p(b)^2

✳ Deviation from this expectation

can be tested.

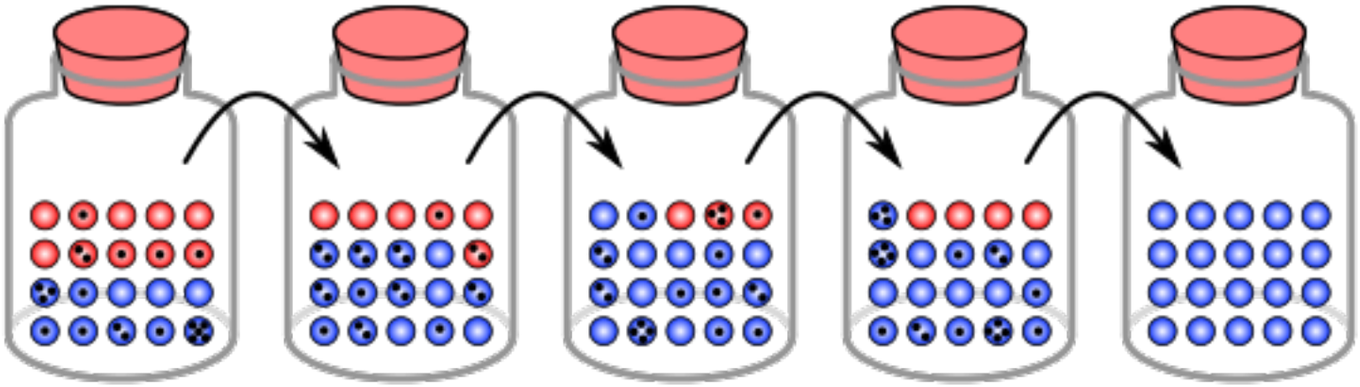
Reasons for HWE deviation

- **Inbreeding**
- Genetic drift
- Population bottleneck/Founder event
- (Genotyping error)



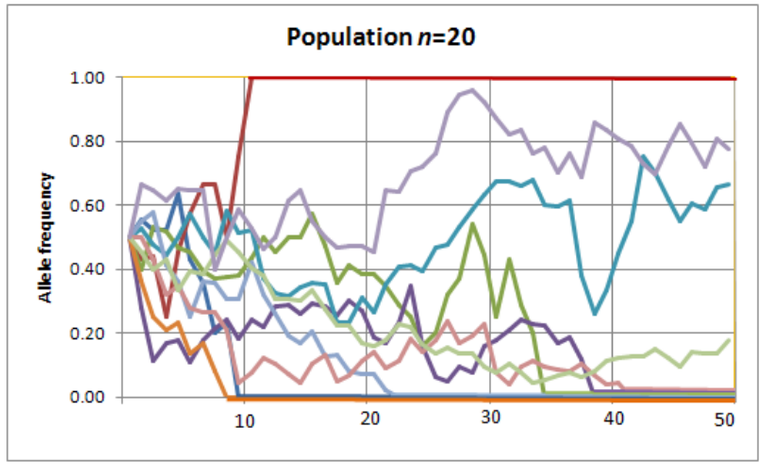
Reasons for HWE deviation

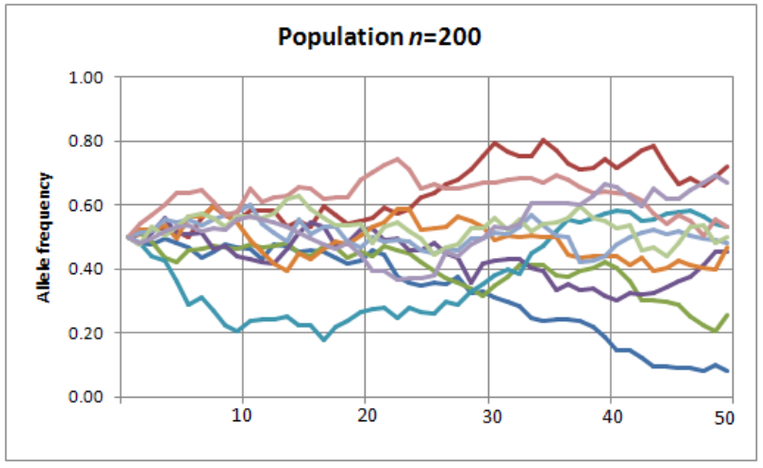
- Inbreeding
- **Genetic drift**
- Population bottleneck/Founder event
- (Genotyping error)

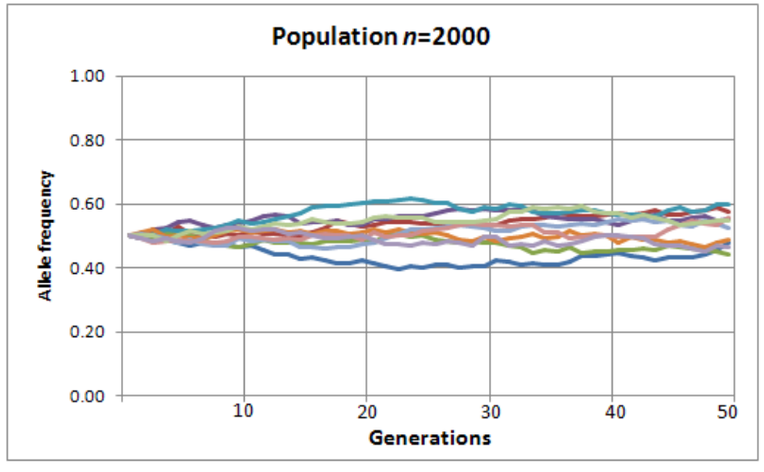


Reasons for HWE deviation

- Inbreeding
- **Genetic drift**
- Population bottleneck/Founder event
- (Genotyping error)

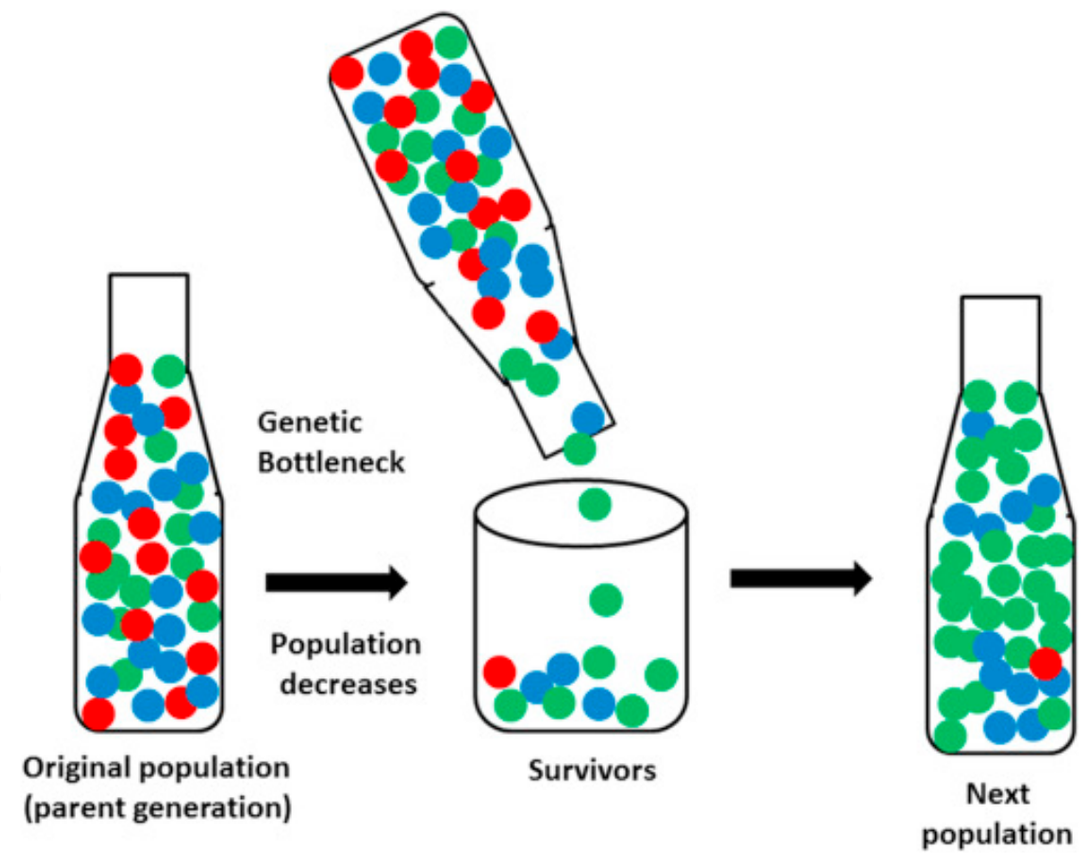




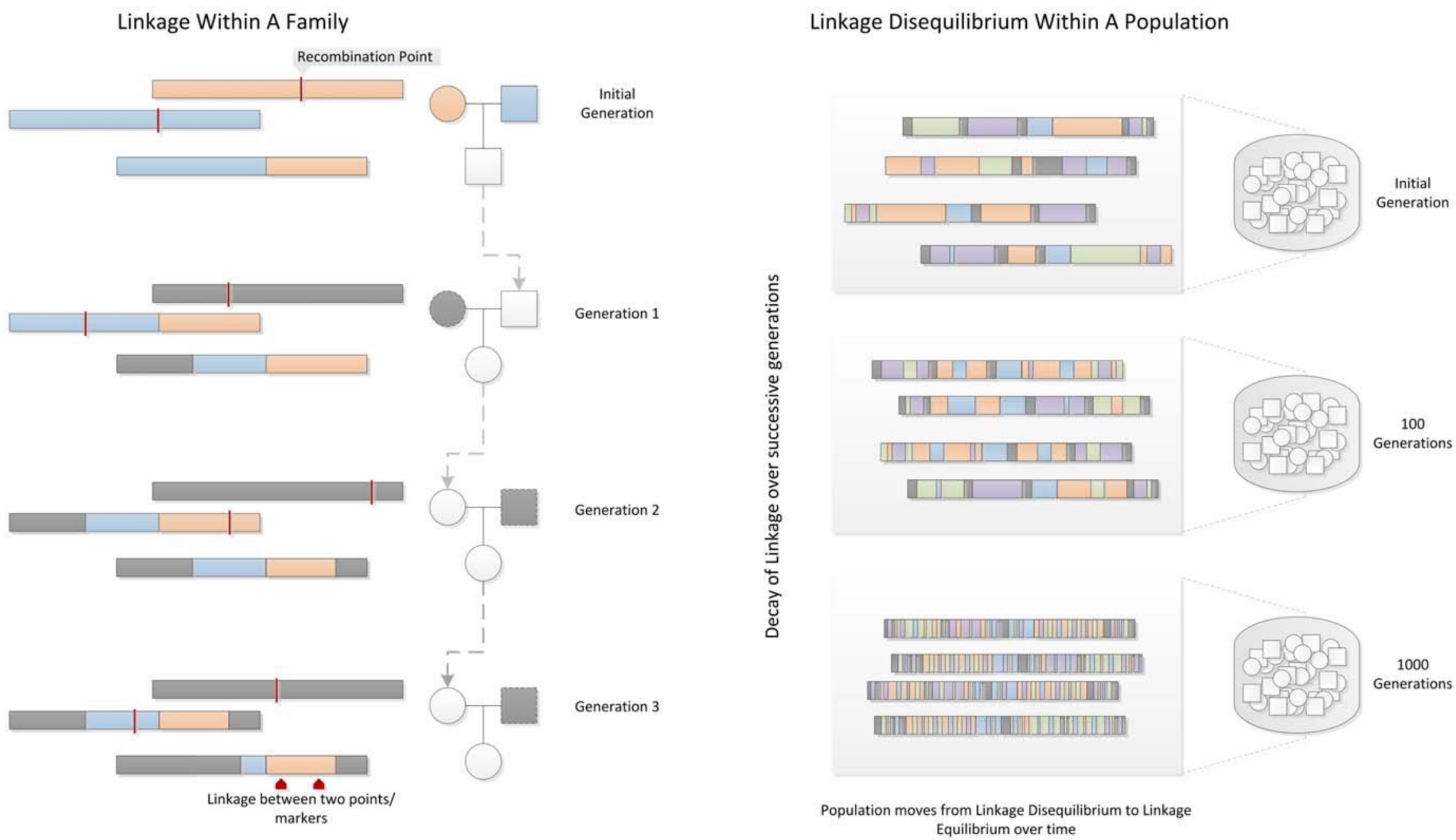


Reasons for HWE deviation

- Inbreeding
- Genetic drift
- **Population bottleneck/Founder event**
- (Genotyping error)



Linkage and Linkage disequilibrium (LD)



doi:10.1371/journal.pcbi.1002822.g002

Linkage and Linkage disequilibrium (LD)

Linked genetic markers are inherited together rather than being broken apart by recombination events

Linkage disequilibrium continuous stretches of founder chromosomes from the initial generation.

LD decay linked blocks sequentially reduced in size by recombination events.

doi:10.1371/journal.pcbi.1002822.g002

Exercise 3: basic statistics (output is stats results)

What output files do you get?

- -freq
- -missing

```
awk 'BEGIN{FS=" +"}{if ($6>0) print }' plink.lmiss
```

- -hardy

(you have 15min…)

Exercise 4 population genetics metrics (output is results)

- **-indep 50 5 2** produce a pruning list of variants
- **-r2** calculate linkage/correlation
- **-blocks no-pheno-req** generate LD blocks