Thursday	8:30-10:00	**Alie**	Population Genetics	Hardy-Weinberg Equilibrium, population structure, admixture mapping.
	10:30-12:00	**Sara**	Family-based Studies	Linkage Analysis, family-based association studies.
	1:30-3:00	**Alie**	Association Studies	Sequencing, genotyping, imputation, association analyses.
	3:30-5	**Sara**	Association Studies	GWAS (including bias), rare variants.

Association Studies

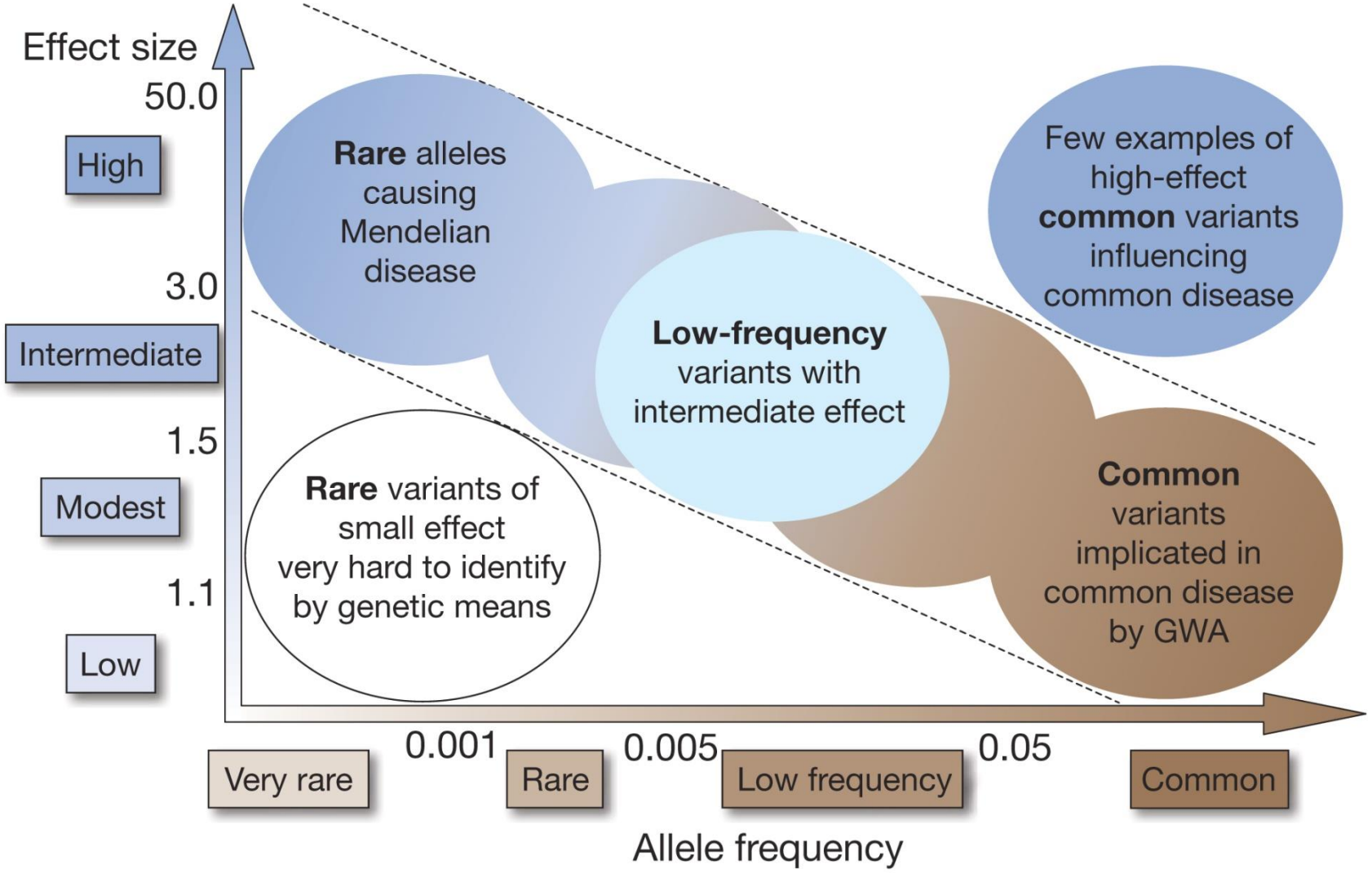
Section 6

(1.5 hours)

Learning objectives

- Describe the differences and the pros and cons of sequencing vs genotyping.
- Calculate and interpret odds ratios in case/control genetic association studies.
- Interpret quantitative trait association studies.
- Understand role for imputation.

Genetic Variation and Disease

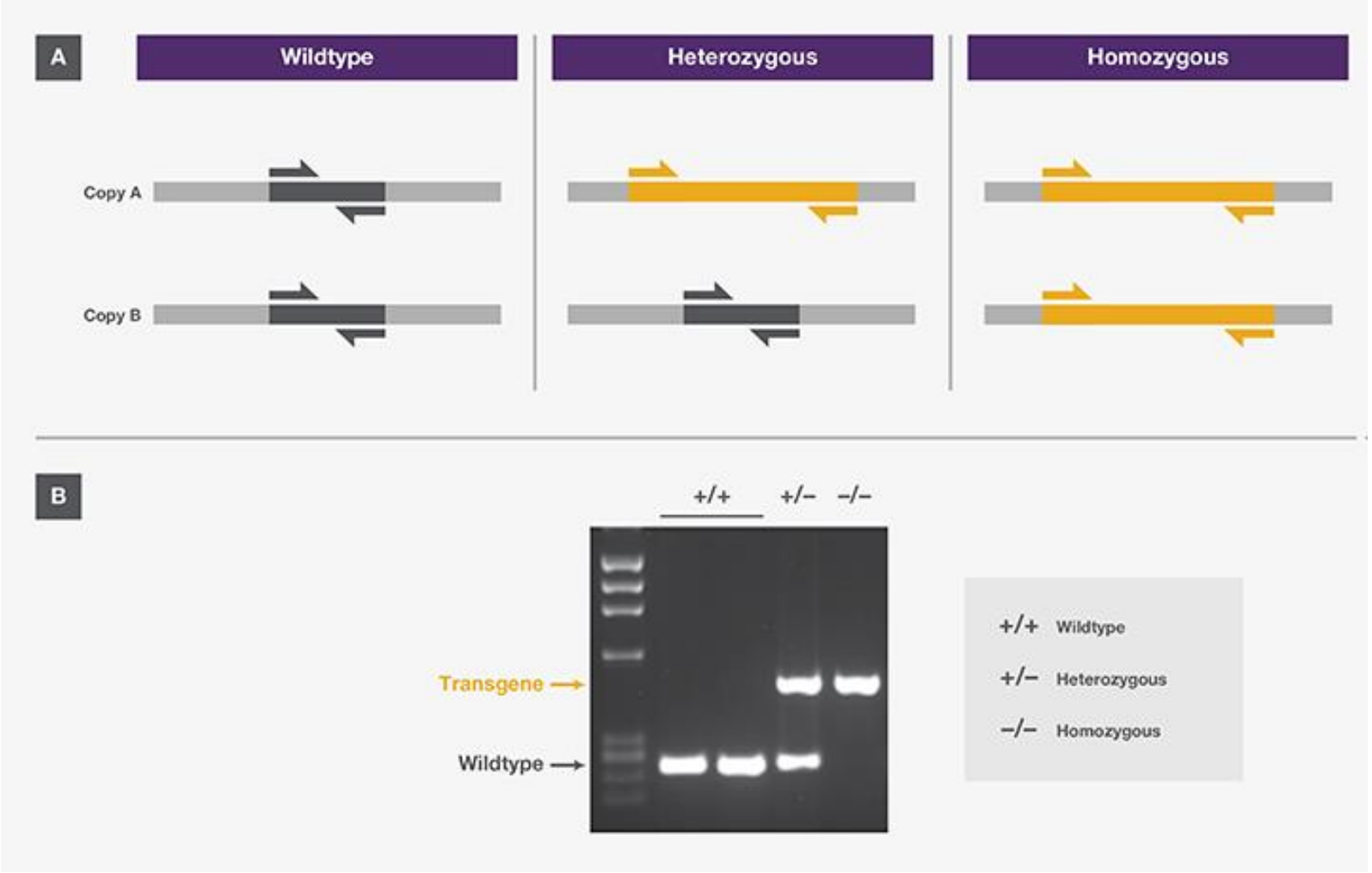


Manolio et al. Nature 2009; 461: 747-753.

Genetic data collection

- TaqMan Polymerase chain reaction (PCR)
  - Targeted, low throughput.
  - Detect deletions and structural variations.
- Genotyping chip
  - Targeted locations, high throughput.
  - Detects single, *a priori* locations.
- Sequencing
  - Collects all bases, increasingly high throughput.
  - Identify novel variants.
  - Analyzing data more intensive

TaqMan PCR to identify variants



ThermoFisher Scientific

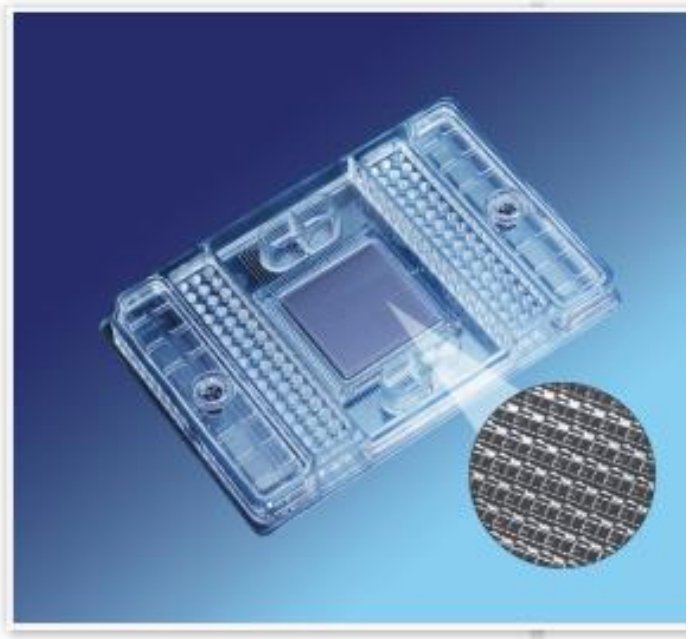
Genotyping technologies (low-throughput)

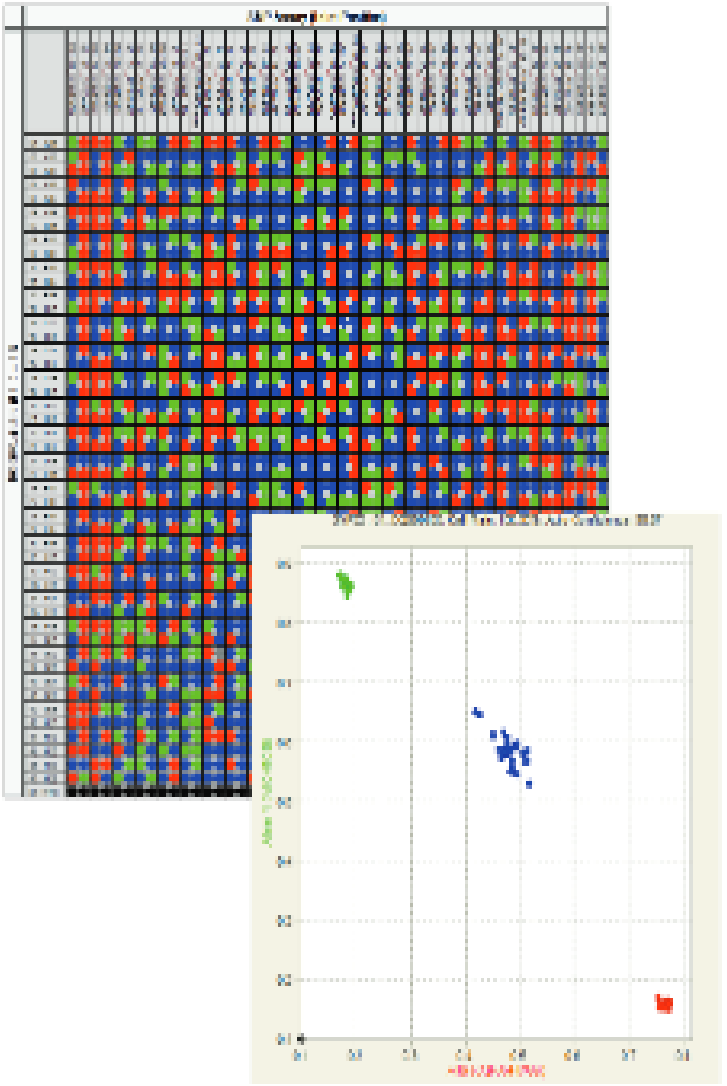
Illumina	SNPlex	Sequenom	TaqMan
			
1500 - 300 SNPs	400 - 40 SNPs	40 - 5 SNPs	10 - 1 SNPs

Chip Genotyping

Why we like SNPs:

- Abundant in the genome
- Easy to measure

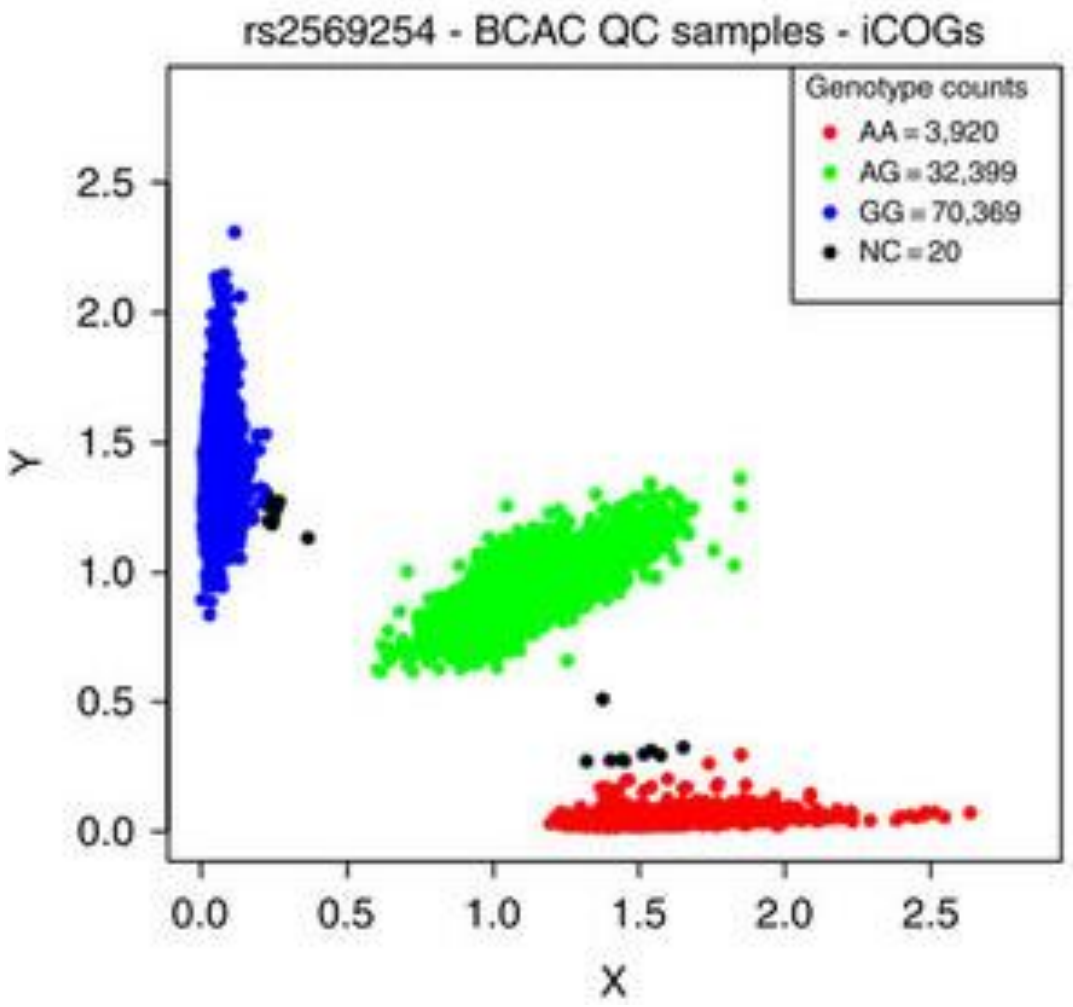


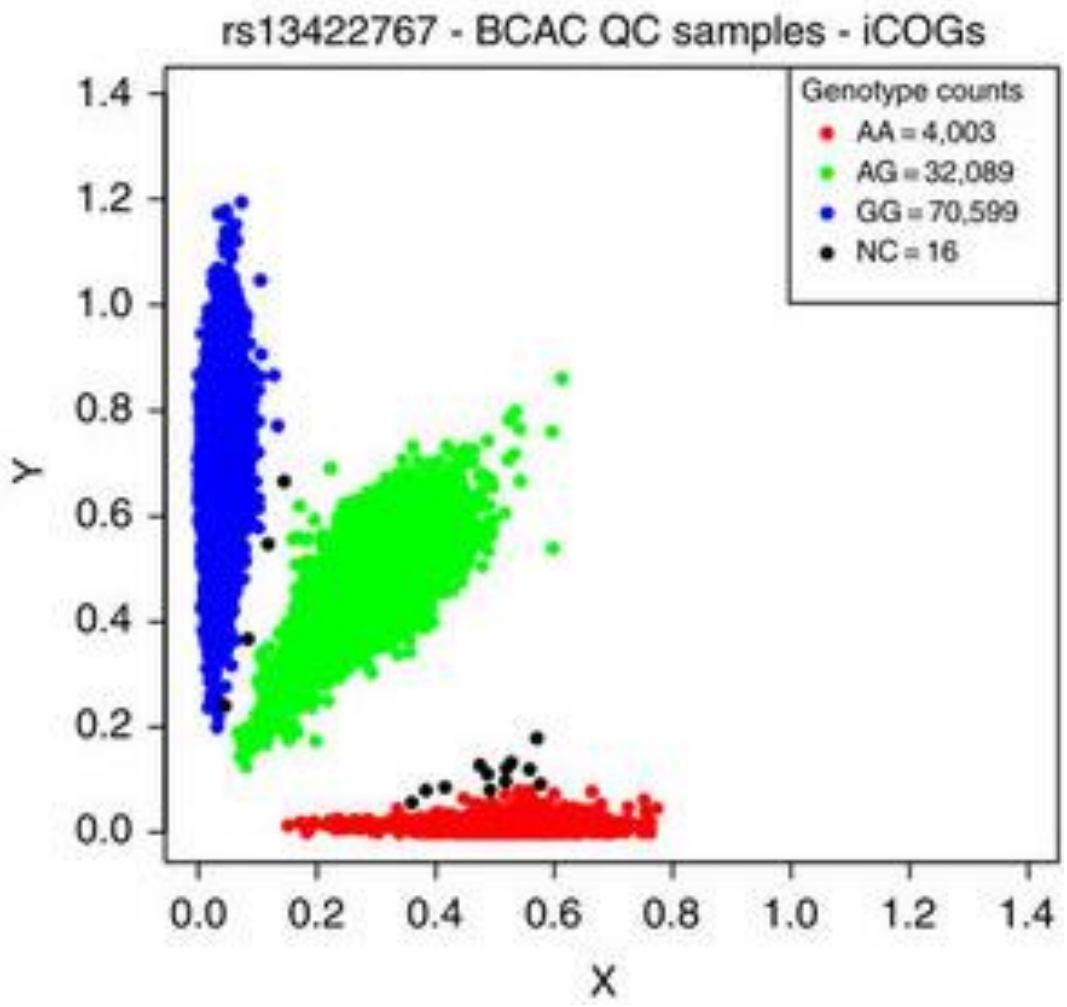


Microfluidics, 96 samples x 96 assays, DNA probes with fluorescent markers.

Fluidigm platform

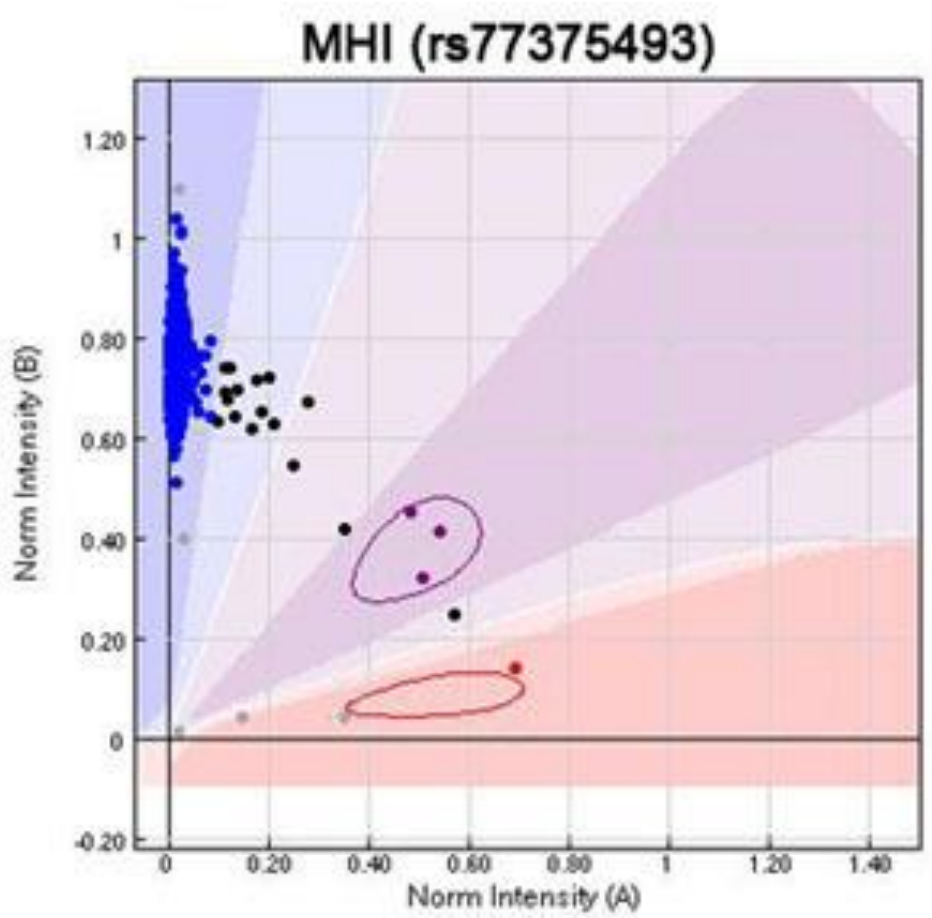
Genotyping Output

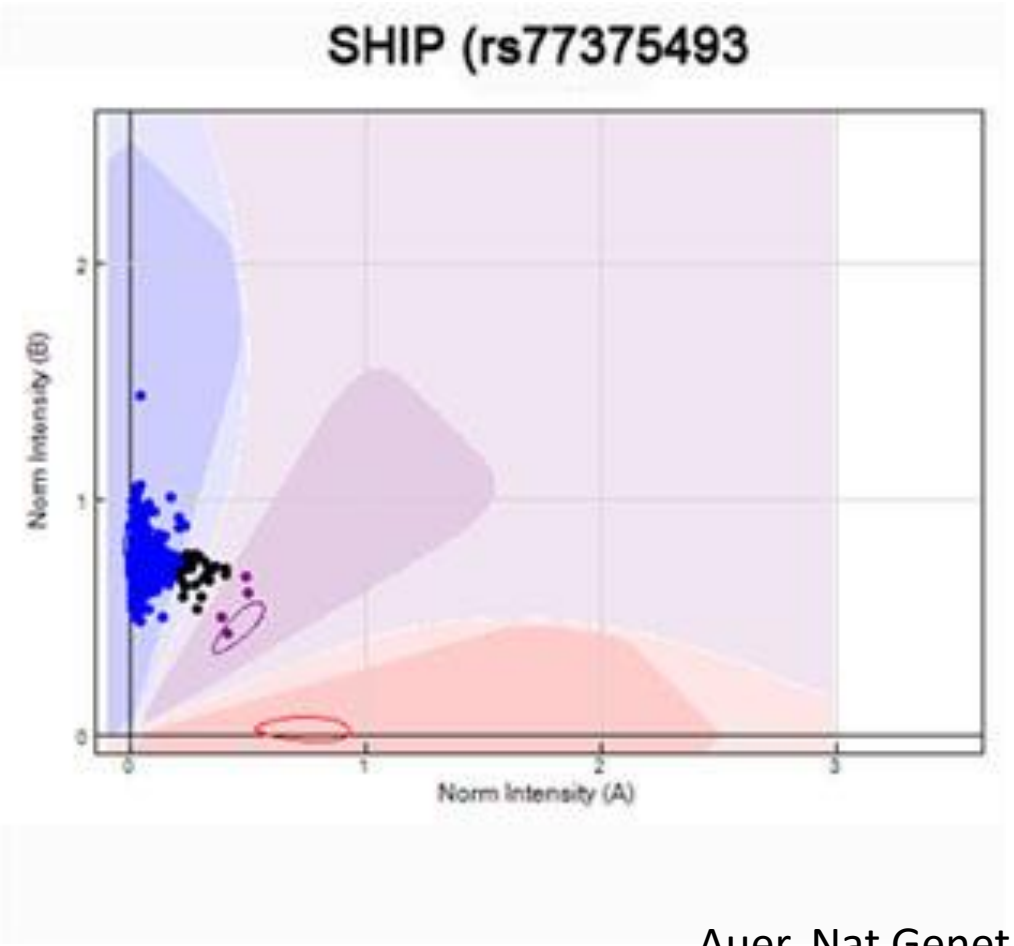




Li, Nat Comm 2014

Genotype cluster plot for rare variants

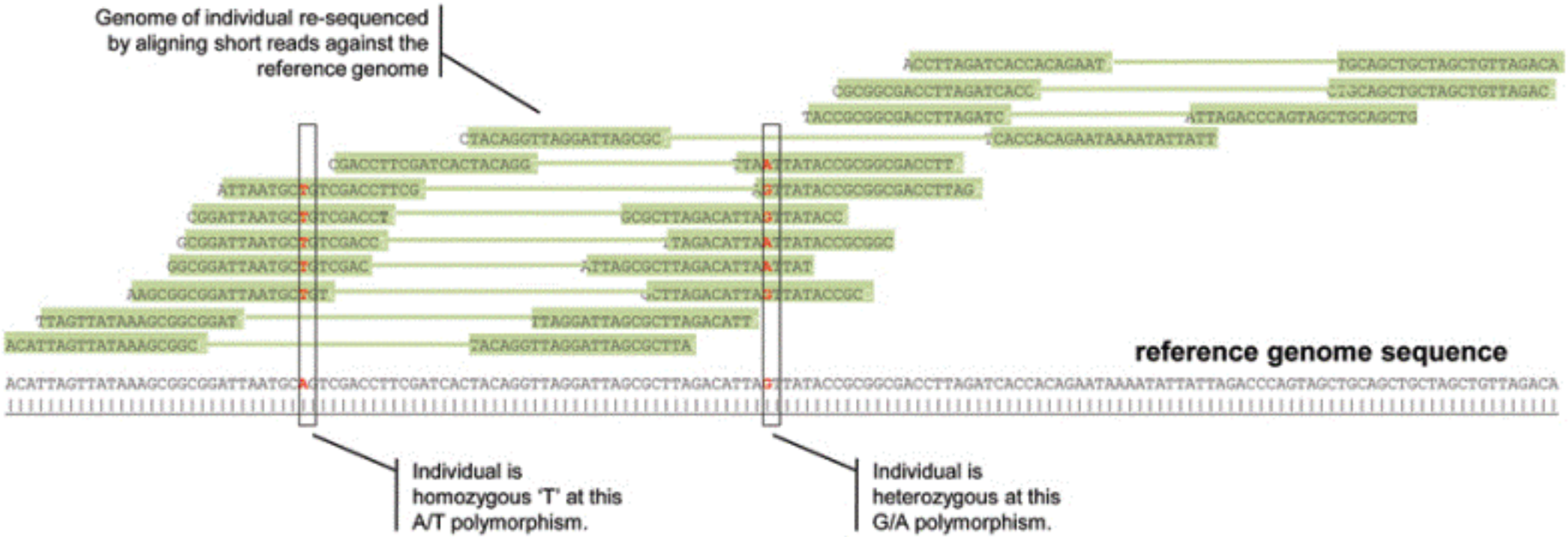




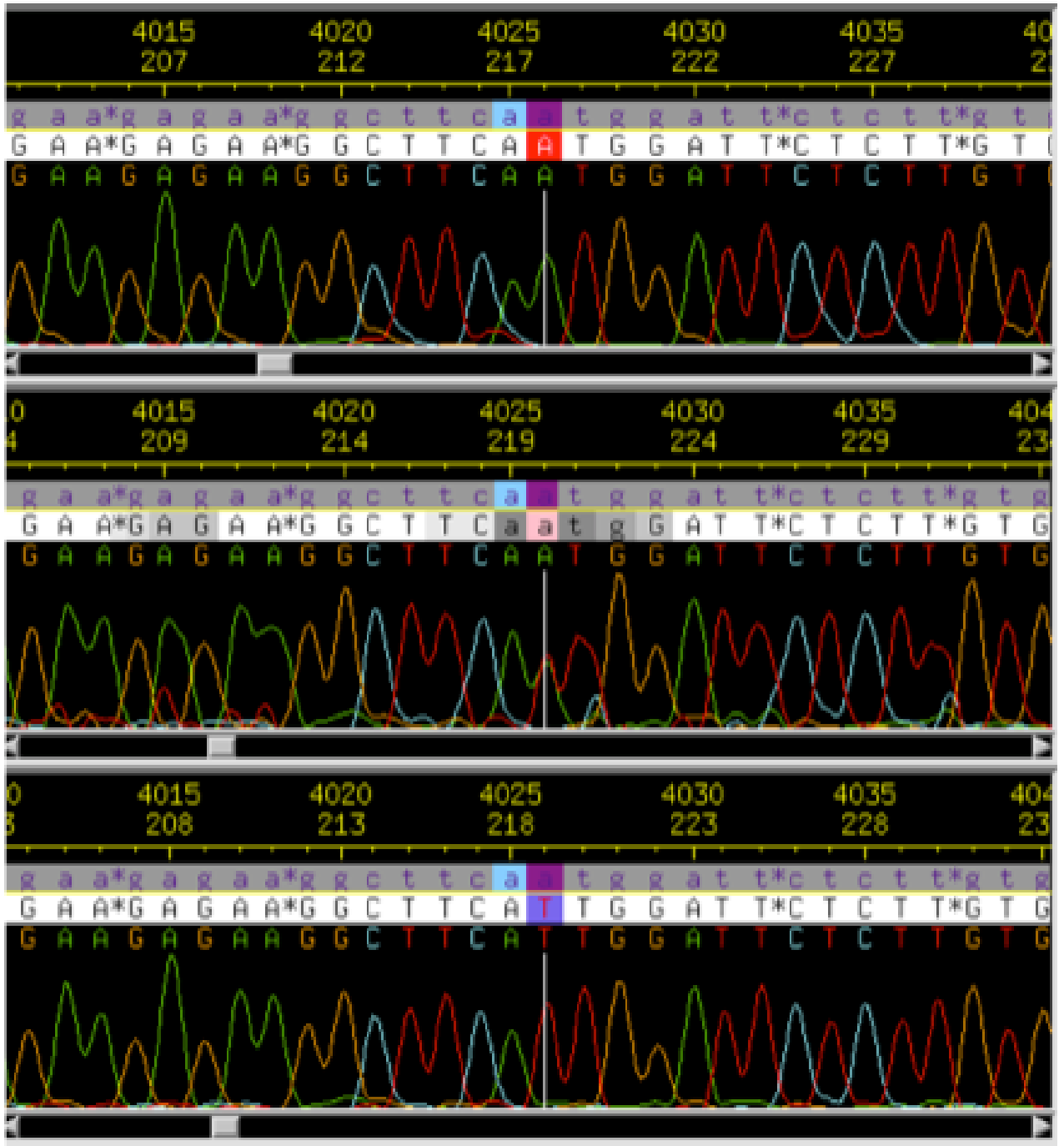
Auer, Nat Genet 2014

Sequencing alignment and depth

Depth: The number of times one basepair is sequenced

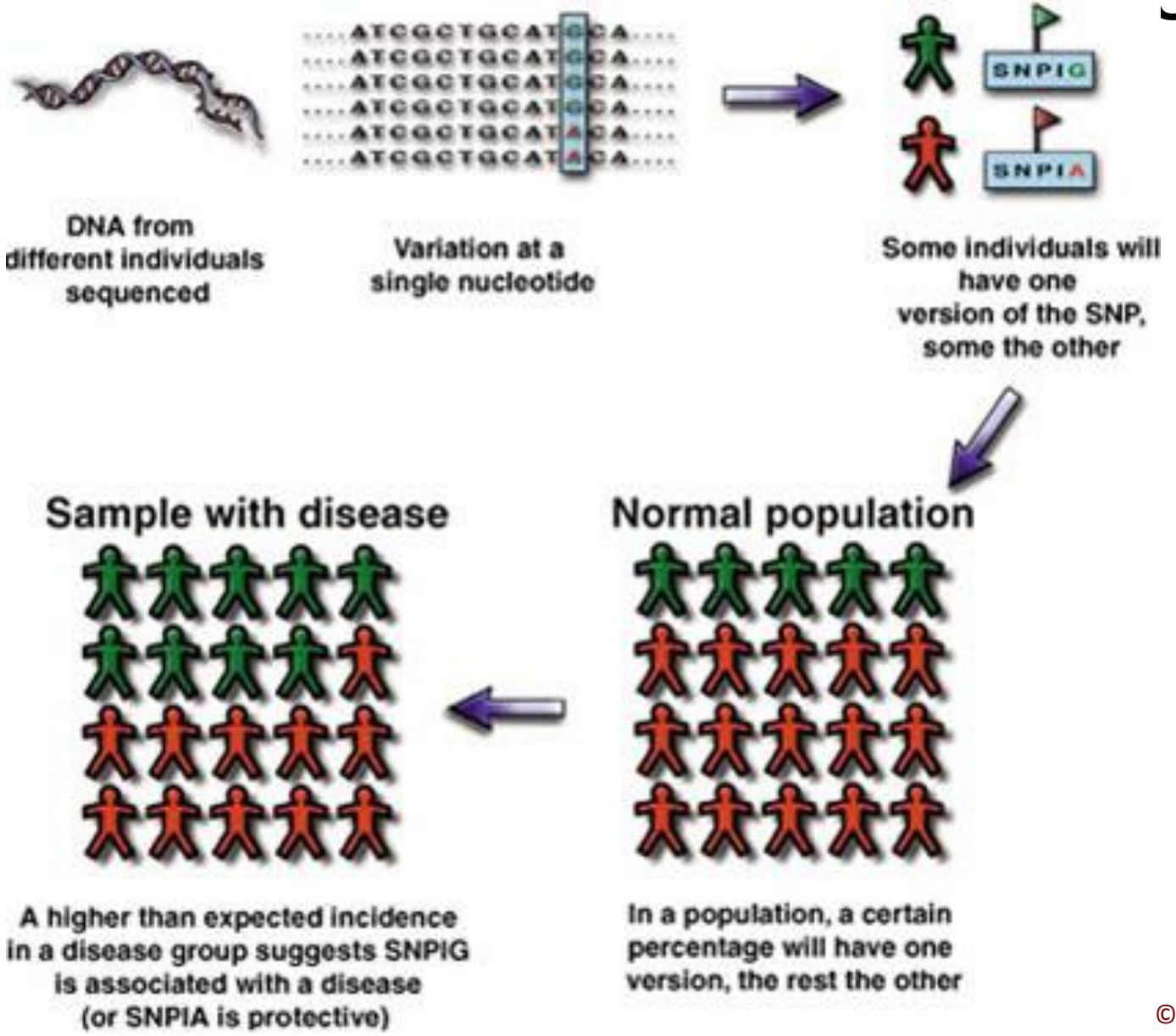


Sequencing output



Fohner 2015

Genetic association studies using SNPs



Association studies

- Determine if a particular genetic feature (exposure) co-occurs with a trait (disease) more often than would be expected by chance.
- Binary: Calculate ‘odds’ of an outcome occurring.
  - Framed as an ‘odds ratio’, the odds of an outcome after an exposure (genotype) in relation to the odds of an outcome without the exposure (reference genotype).
- Continuous: calculate change in an outcome for every unit increase of an exposure.

measure of events out of all possible events (RR) vs ratio of events to non-events (OR)

$$RR = \frac{\text{Risk of event in the Treatment group}}{\text{Risk of event in the Control group}} = \frac{a/(a+b)}{c/(c+d)}$$

$$OR = \frac{\text{Odds of event in Treatment group}}{\text{Odds of event in Control group}} = \frac{a/b}{c/d} \quad :$$

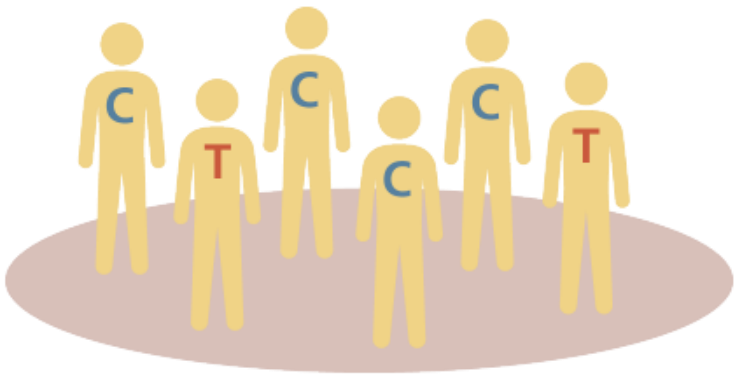
measure of events out of all possible events (Ratio) vs ratio of events to non-events (Odds)

$$RR = \frac{\text{Risk of event in the Treatment group}}{\text{Risk of event in the Control group}} = \frac{a/(a+b)}{c/(c+d)}$$

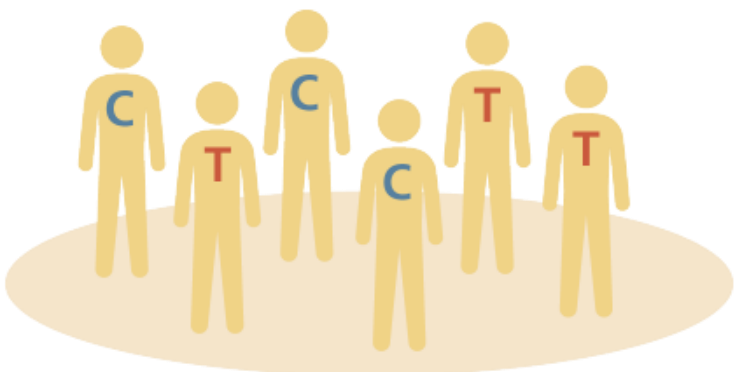
$$OR = \frac{\text{Odds of event in Treatment group}}{\text{Odds of event in Control group}} = \frac{a/b}{c/d} :$$

If an outcome occurs 10 out of 100 times, the risk is 10%
But the odds is 10/90 = 11.1%

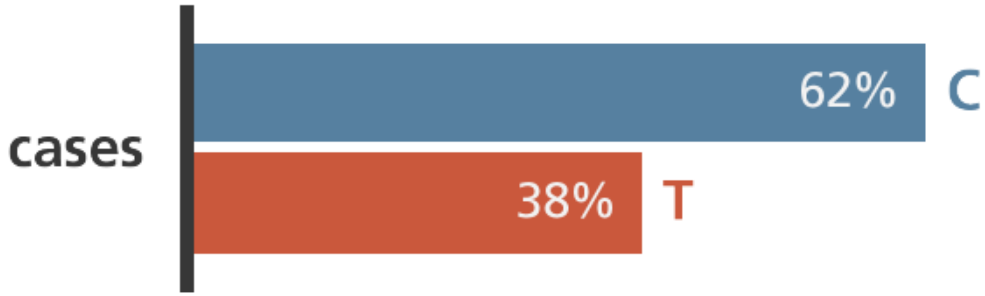
SNP



cases (n=1,000)
people with heart disease



controls (n=1,000)
people without heart disease



Association testing in case-control studies

		Disease status		
		Cases	Controls	Total
Genotype	M	a	b	a+b
	m	c	d	c+d
Total		a+c	b+d	

Association testing in case-control studies

		Disease status		
		Cases	Controls	Total
Genotype	M	a	b	a+b
	m	c	d	c+d
Total		a+c	b+d	

1) Calculate the odds of the disease with the genotype and without the genotype

Odds that the M genotype occurs in a case: $\frac{a/_{a+b}}{b/_{a+b}} = \frac{a}{b}$

Odds that the m genotype occurs in a case: $\frac{c/_{c+d}}{d/_{c+d}} = \frac{c}{d}$

Association testing in case-control studies

		Disease status		
		Cases	Controls	Total
Genotype	M	a	b	a+b
	m	c	d	c+d
Total		a+c	b+d	

2) Calculate Odds Ratio (OR) as the odds that genotype M occurs in a case divided by the odds that genotype m occurs in a case.

$$\left(\frac{a/_{a+b}}{b/_{a+b}}\right)/\left(\frac{c/_{c+d}}{d/_{c+d}}\right) = \frac{a/_b}{c/_d} = \frac{ad}{bc}$$

$$\text{OR} = \frac{ad}{bc}$$

Association testing in case-control studies

		Disease status		
		Cases	Controls	Total
Genotype	M	a	b	a+b
	m	c	d	c+d
Total		a+c	b+d	

Odds that the M allele occurs in a case $= \frac{a}{b}$

Odds that the m allele occurs in a case $= \frac{c}{d}$

H_0: OR = 1 (no association)

OR > 1 indicates increased odds

OR < 1 indicates decreased odds (protective)

The Odds Ratio (OR) is the odds that M occurs in a case divided by the odds that m occurs in a case:

$$\text{OR} = \frac{ad}{bc}$$

Confidence intervals for odds ratios

		Disease status	
		Cases	Controls
Genotype	M	a	b
	m	c	d

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

$$s.e(\log(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Confidence interval: $e^{\log(OR) \pm z_{\alpha/2} \times s.e(\log(OR))}$

Lower limit of 95% confidence interval: $e^{\log(OR) - 1.96 \times s.e}$

Upper limit of 95% confidence interval: $e^{\log(OR) + 1.96 \times s.e}$

Calculate– odds ratio and 95% confidence interval

	Cases	Controls	Total
TT+TC	158	392	550
CC	20	86	106
Total	178	478	1656

$$OR = \frac{ad}{bc}$$

$$s.e(\log(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Odds ratio calculations – odds ratio itself

	Cases	Controls	Total
TT+TC	158	392	550
CC	20	86	106
Total	178	478	1656

$$OR = \frac{158 \times 86}{392 \times 20} = 1.7332$$

$$s.e.(log(OR)) = \sqrt{\frac{1}{158} + \frac{1}{392} + \frac{1}{20} + \frac{1}{86}}$$

Odds ratio calculations – confidence intervals

	Cases	Controls	Total
TT+TC	158	392	550
CC	20	86	106
Total	178	478	1656

$$OR = \frac{158 \times 86}{392 \times 20} = 1.7332$$

$$s.e.(log(OR)) = \sqrt{\frac{1}{158} + \frac{1}{392} + \frac{1}{20} + \frac{1}{86}}$$

lower limit 95% confidence interval:

$$= exp(log(OR) - 1.96 \times s.e.(log(OR)))$$

$$= exp(log(1.7332) - 1.96 \times 0.2665) = 1.03$$

Upper limit 95% confidence interval: 2.92

Let’s practice! Calculate odds ratio

	Thyroid Cancer	No thyroid cancer	Total
AA+AG	50	20	70
GG	300	200	500
Total	350	220	570

$$\text{OR} = \frac{ad}{bc}$$

Let’s practice! Calculate odds ratio

	Thyroid Cancer	No thyroid cancer	Total
AA+AG	50	20	70
GG	300	200	500
Total	350	220	570

Odds ratio: (50*200)/(20*300) = 1.6

Turn this result into a sentence about effect of A allele in thyroid cancer.

Let’s practice! Calculate odds ratio

	Thyroid Cancer	No thyroid cancer	Total
AA+AG	50	20	70
GG	300	200	500
Total	350	220	570

Odds ratio: (50*200)/(20*300) = 1.6

Turn this result into a sentence about effect of A allele in thyroid cancer.

The odds of developing thyroid cancer are 1.6x times greater with an A allele compared to without an A allele.

Often use logistic regression for case-control analyses

Allows you to adjust for relevant factors

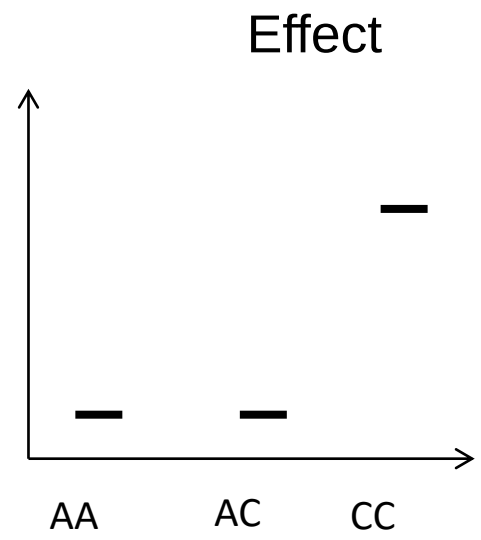
- Population stratification, age, sex, matching variables etc

$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \mathbf{g} + \beta_2 x_1 + \ldots + \beta_{k+1} x_k$ (g is genotype, $x_1, \ldots x_k$ are covariates)

Coefficients are estimated using maximum likelihood estimation (MLE)

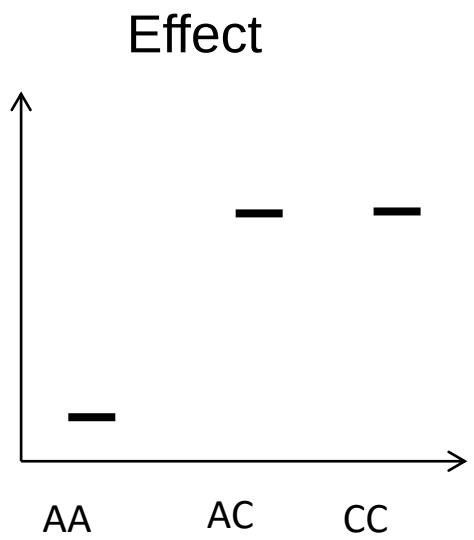
- $\ln\left(\frac{p}{1-p}\right)$ = log odds of an outcome
- Test H_0: $\beta_1 = 0$ (likelihood ratio test, wald test, score test)
- The odds ratio is OR=e^{β_1}
- β_1 = SNP effect (log(OR)) ➔ e^{β_1} = OR

Common models of penetrance



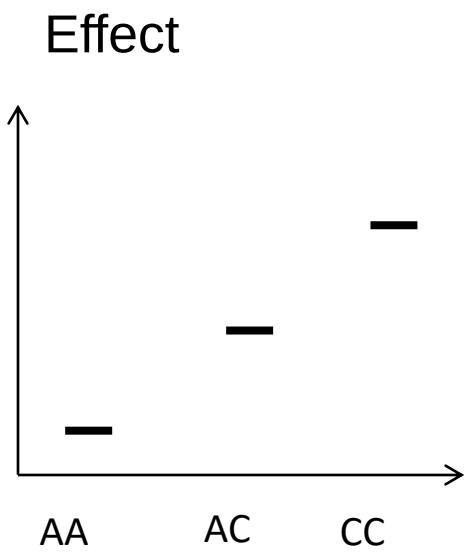
Recessive

Genotype coding: 0,0,1



Dominant

Genotype coding: 0,1,1



Additive

Genotype coding: 0,1,2

Effect = mean of continuous trait or log(OR) of binary trait

Interpret results

$$\log odds\ Disease = 3 + 1.2(\mathbf{A}) - 0.3(\text{Female})$$

Genotypes: GG, GA, AA

Interpret results

$\log odds\ Disease$= 3 + 1.2(**A**) - 0.3(Female)

Genotypes: GG, GA, AA

1) Genotypes are additive (codes 0, 1, 2)
2) Reference gender is male

Interpret results

$$\log odds\ Disease = 3 + 1.2(\mathbf{A}) - 0.3(\text{Female})$$

Genotypes: GG, GA, AA

1) Genotypes are additive (codes 0, 1, 2)
2) Reference gender is male
3) Every A allele increases log odds of disease 1.2
4) OR AG vs GG $e^{1.2} = 3.3$
5) What happens for AA?

Interpret results

$\log odds\ Disease$= 3 + 1.2(**A**) - 0.3(Female)

Genotypes: GG, GA, AA

1) Genotypes are additive (codes 0, 1, 2)
2) Reference gender is male
3) Every A allele increases log odds of disease 1.2
4) OR AG vs GG $e^{1.2} = 3.3$
5) What happens for AA? $e^{1.2*2} = 11$ compared to GG.
6) Being female is protective ($e^{-0.3} = 0.74$)

Continuous outcome genetic association

- Linear regression (instead of logistic)
- Additive coding of SNP (0,1,2) most common

$$Y = \alpha + \beta * SNP + X$$

  - β = SNP effect (for every SNP, unit increase in outcome)
  - SNP = covariate coded (0,1,2)
  - X = additional covariates (e.g. sex, study, age, population stratification)

Continuous outcome genetic association

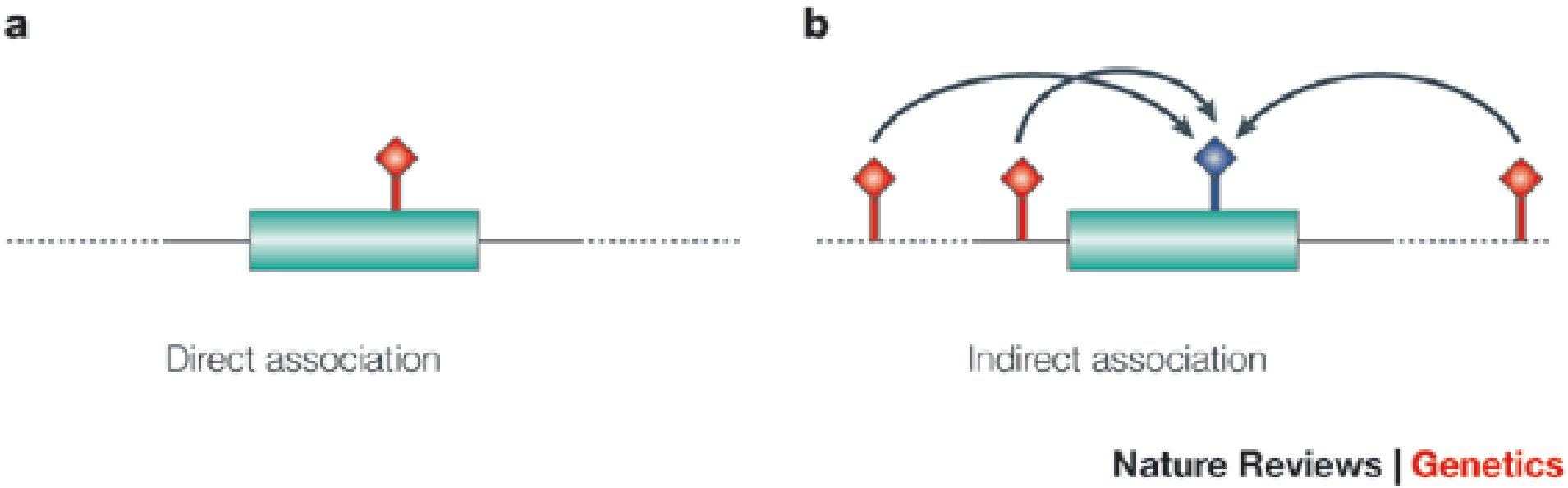
- Linear regression (instead of logistic)
- Additive coding of SNP (0,1,2) most common

$$Y = \alpha + \beta * SNP + X$$

- Y = height in inches
- β = 1.2
- SNP = AA, AC, CC covariate coded (0,1,2)

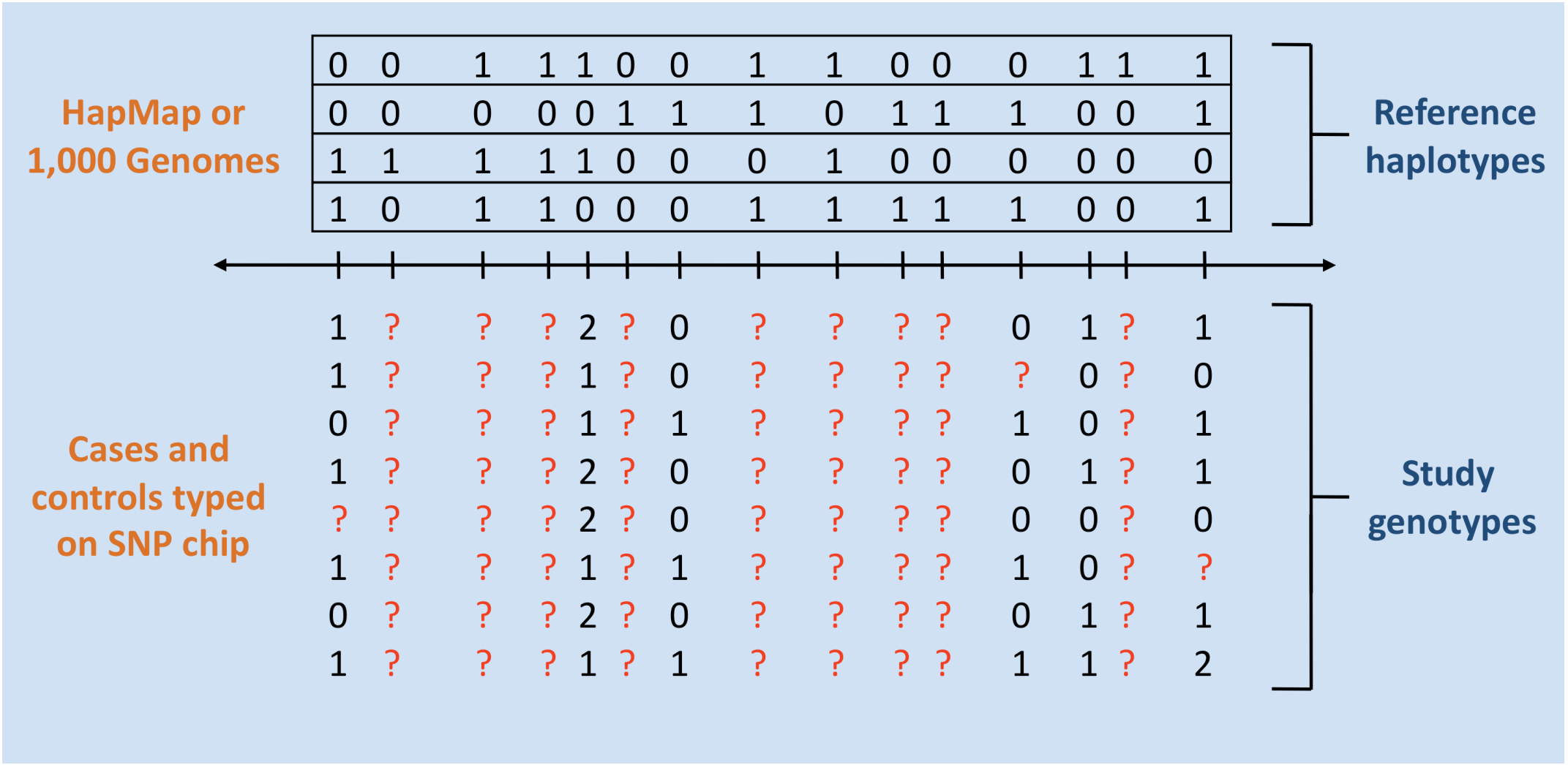
- Interpretation: For every allele C allele, predicted height increases 1.2 inches.

We can use LD in our studies: tagSNPs



Hirschhorn & Daly. *Nature Reviews Genetics 2005,*
http://mathgen.stats.ox.ac.uk/impute/impute_v2.html

We can use LD in our studies: Imputation



Hirschhorn & Daly. *Nature Reviews Genetics 2005,*
http://mathgen.stats.ox.ac.uk/impute/impute_v2.html

Imputation

- Cost efficient
  - Can assess more SNPs than we genotyped (tagSNPs)
- Allows us to keep our sample size
  - Fill in missings for already genotyped SNPs
- Allows us to combine data from existing platforms and different studies that genotype different SNPs

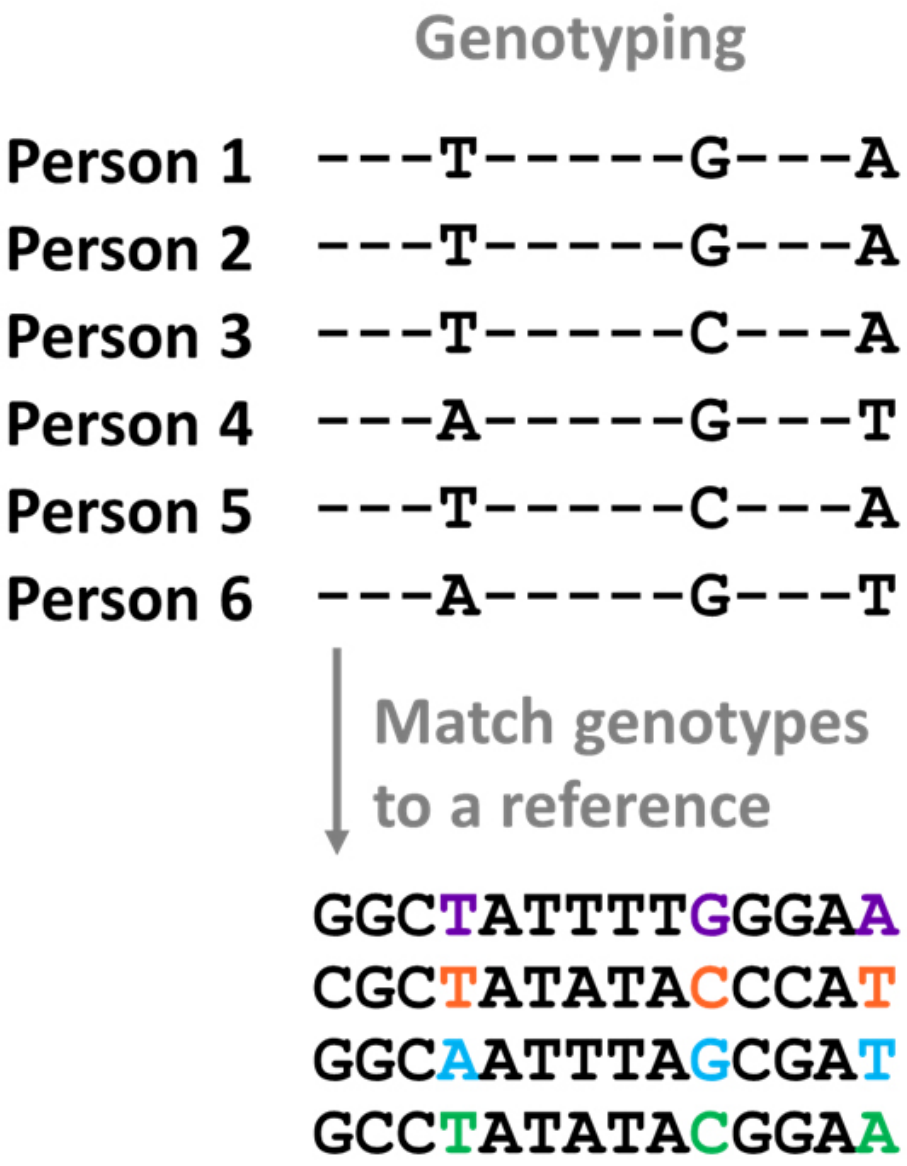
Imputation

Due to LD, we can compare haplotypes between a “reference” panel and our study and thereby guess genotypes

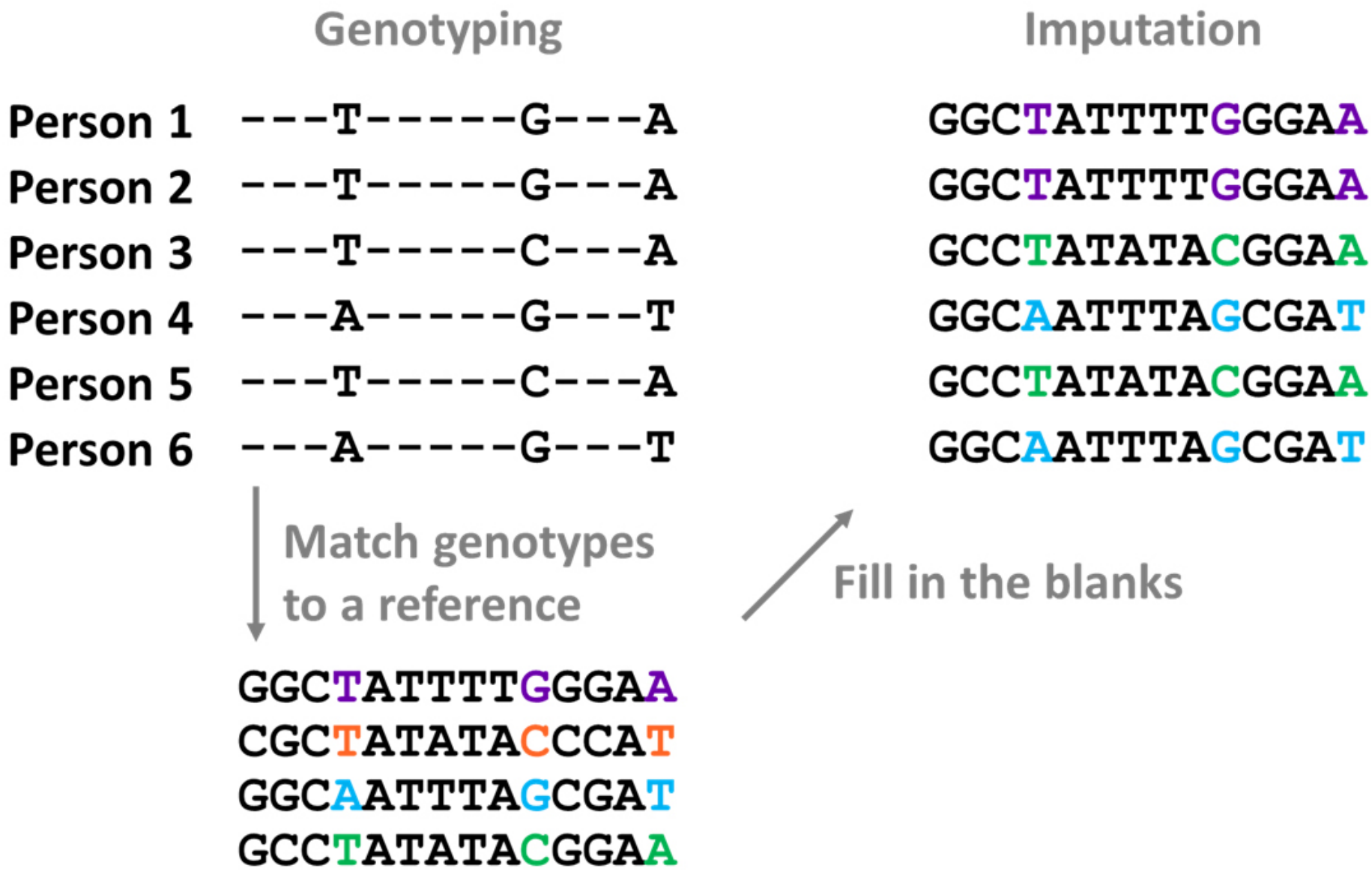
Study Individual: **T A G G T ? T G C C T A ? C G T**

Reference Panel Individual: T A G G T A T G C C T A G C G T

https://mathgen.stats.ox.ac.uk/impute/impute_v2.html



Can you impute the missing bases?


Genotyping
Imputation
Person 1 ---T------G---A
Person 2 ---T------G---A
Person 3 ---T------C---A
Person 4 ---A------G---T
Person 5 ---T------C---A
Person 6 ---A------G---T
GGCTATTTTGGGAA
GGCTATTTTGGGAA
GCCTATATACGGAA
GGCAATTTAGCGAT
GCCTATATACGGAA
GGCAATTTAGCGAT
Match genotypes to a reference
Fill in the blanks
GGCTATTTTGGGAA
CGCTATATACCCAT
GGCAATTTAGCGAT
GCCTATATACGGAA

Imputation

- We can infer genotypes for SNPs we didn’t genotype (or failed in the lab)
  - **Input:** 550,000 SNPs in 10,000 individuals
  - **Reference panel:** 2,504 individuals from the 1000 Genomes project (>80M markers)
  - **Output:** Imputed data for >80M markers for your 10,000 individuals
    - In practice, we exclude markers that were only seen once in 1000Genomes so we end up with ~47M markers)

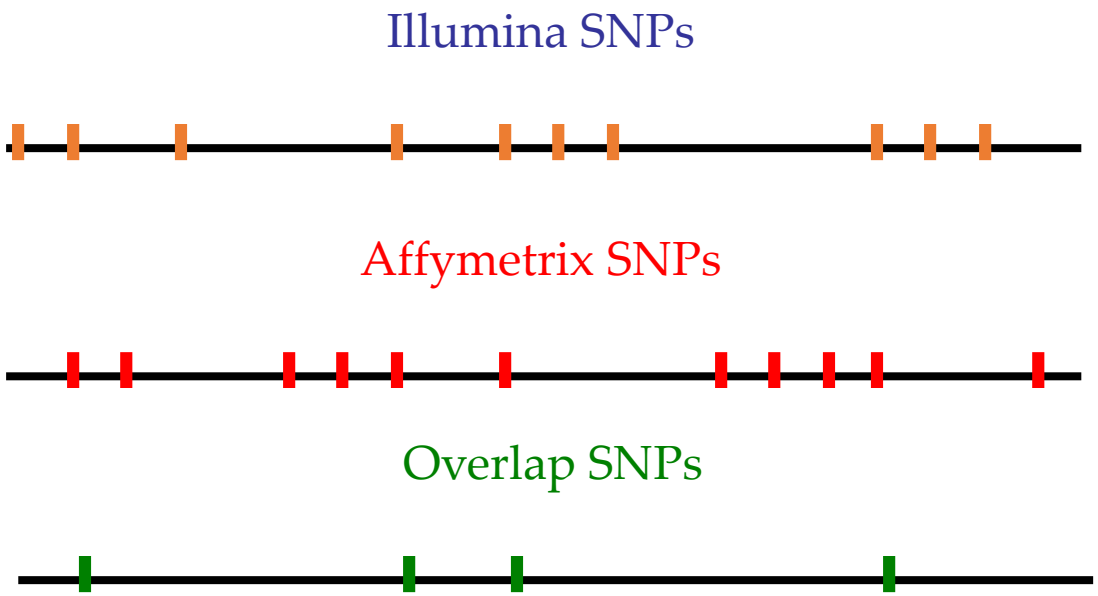
Assessing SNPs across genotyping platforms

	HumanHap	Affy 6.0	OmniExpress
HumanHap	459,999	126,959	260,661
Affy 6.0		668,283	168,223
OmniExpress			565,810

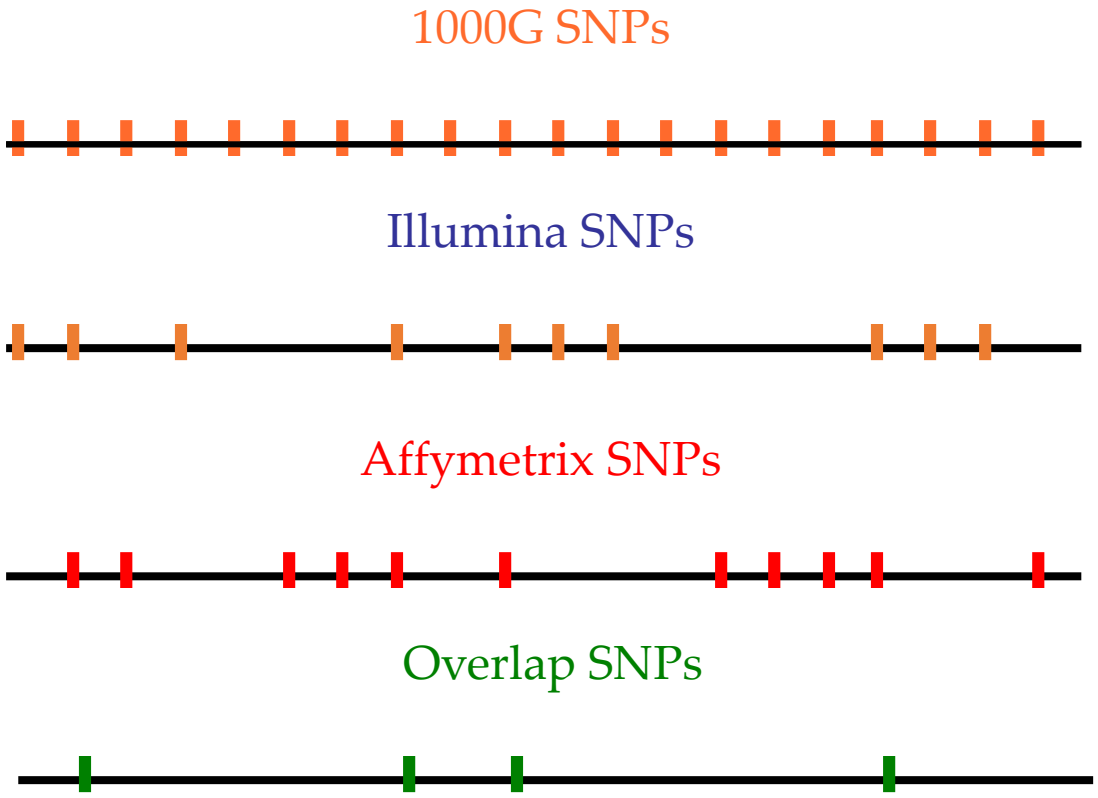
* 75,285 markers are on all 3 platforms

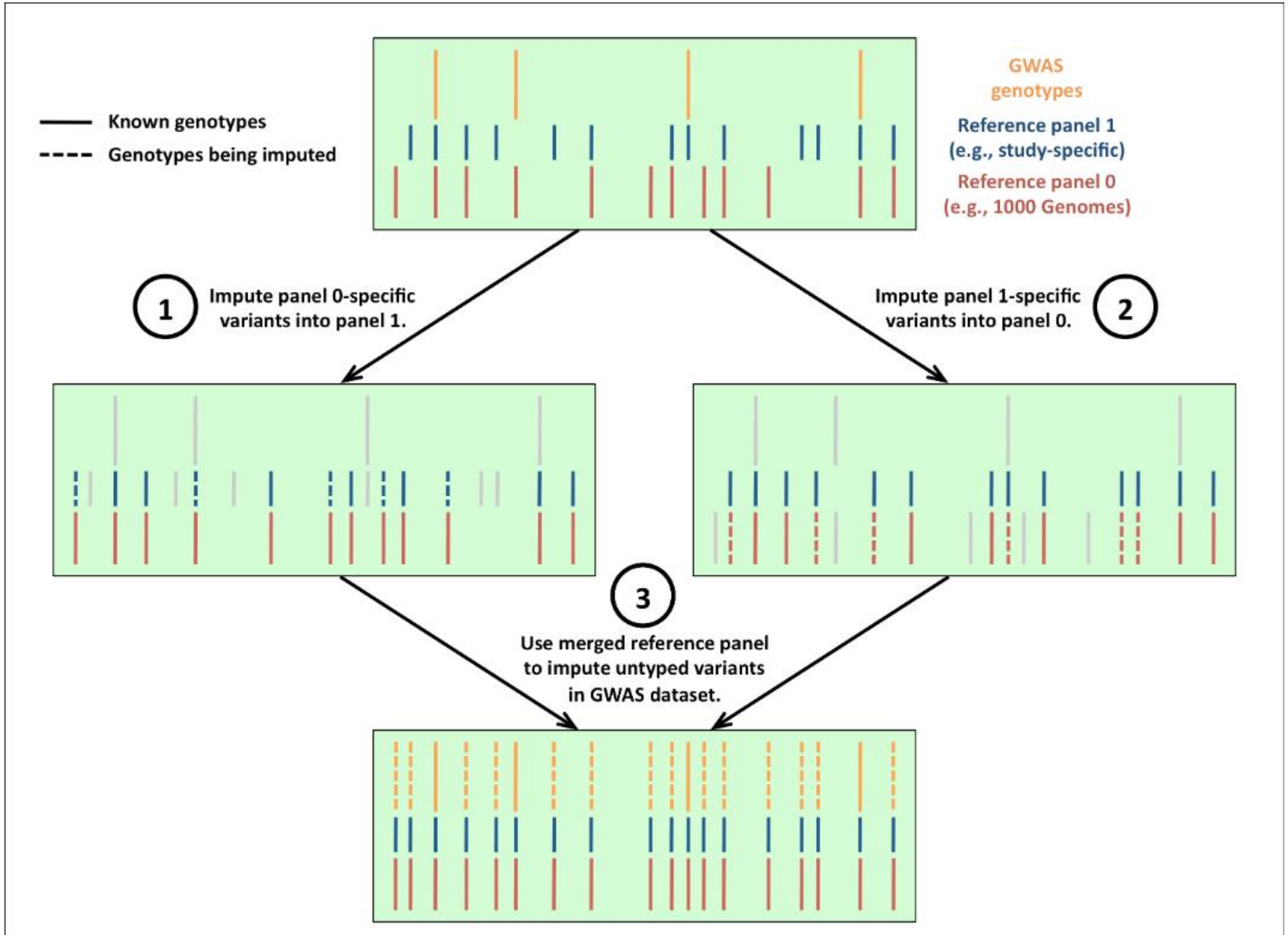
Lindström, PLoS One 2017

Imputation for studying SNPs across platforms



Imputation for studying SNPs across platforms




Known genotypes
Genotypes being imputed
GWAS genotypes
Reference panel 1 (e.g., study-specific)
Reference panel 0 (e.g., 1000 Genomes)
1
Impute panel 0-specific variants into panel 1.
2
Impute panel 1-specific variants into panel 0.
3
Use merged reference panel to impute untyped variants in GWAS dataset.

Imputation

- The imputation quality score r^2 measures how well a SNP was imputed.
  - Ranges between 0 and 1.
  - A quality score of r^2 on a sample of *N* individuals indicates that the amount of data at the imputed SNP is approximately equivalent to a set of perfectly observed genotype data in a sample size of r^2N.
  - Typically, a cut-off of 0.30 or so will flag most of the poorly imputed SNPs, but only a small number (<1%) of well imputed SNPs. Caveat: This is not true for rare SNPs

Imputation

- Factors that affect imputation quality:
  - Number of genotyped SNPs in your data
  - Size of reference panel
  - Similarity in genetic ancestry between reference and study samples
  - Allele frequency

Summary

- Genetic data can be collected through genotyping or sequencing.
- Odds ratios give the odds of an outcome in relation to a reference.
- Linear and logistic regression allow adjustment for other factors.
- Imputation leverages linkage disequilibrium to estimate data not collected.