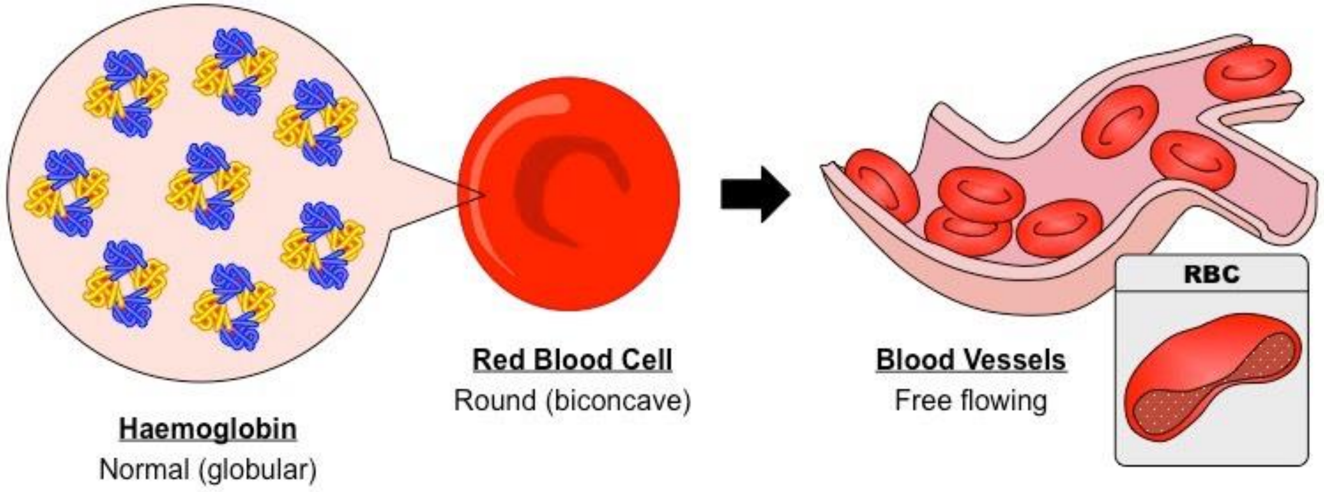
Association Studies

Normal (Wild-Type) Haemoglobin

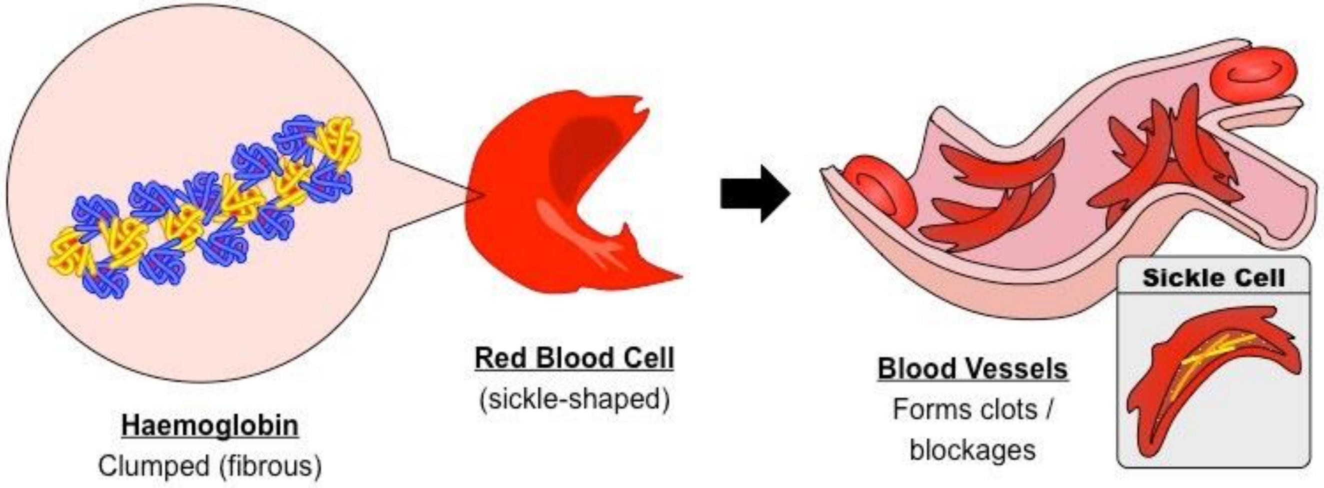


Haemoglobin
Normal (globular)

Red Blood Cell
Round (biconcave)

Blood Vessels
Free flowing

'Sickle Cell' Haemoglobin

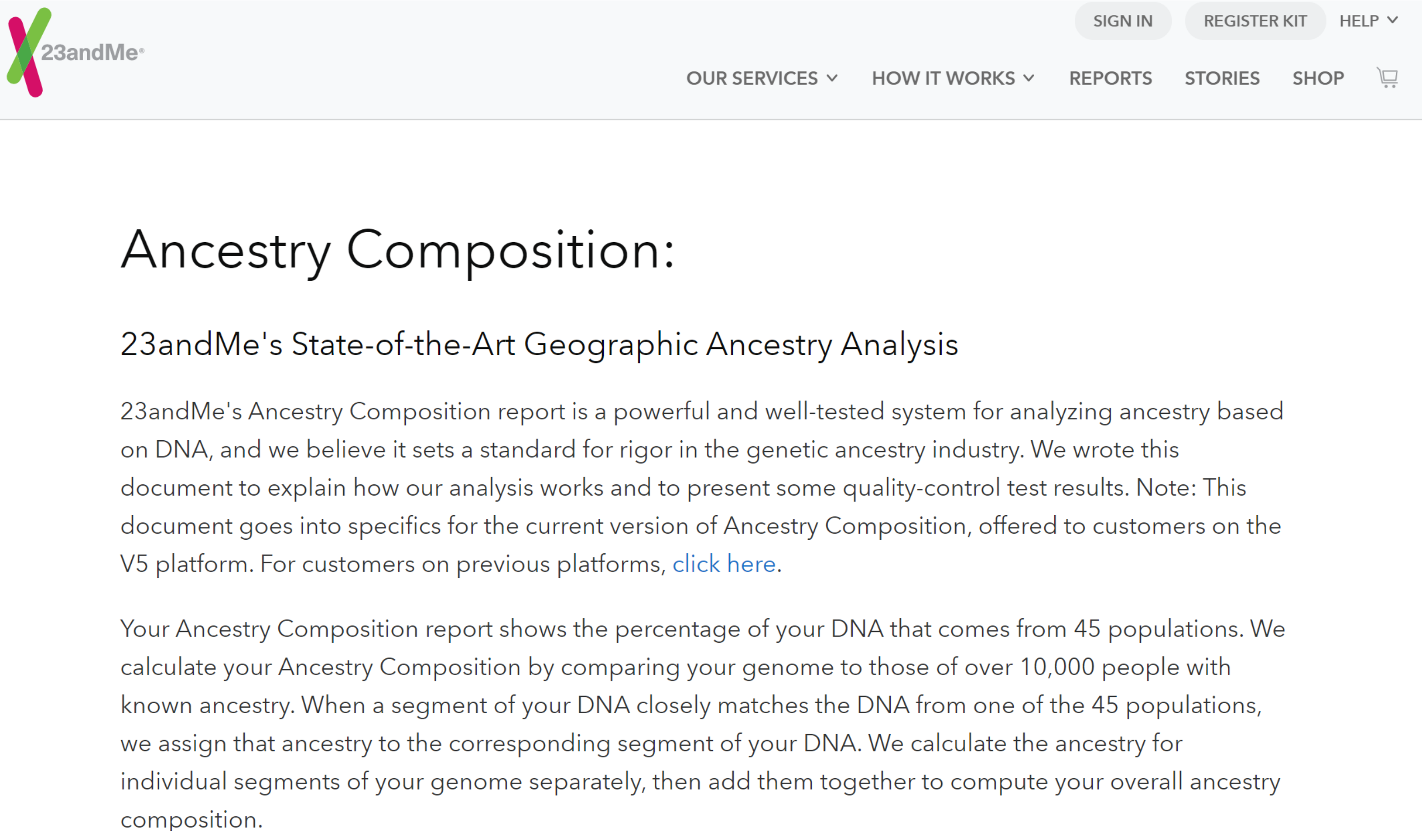


Haemoglobin
Clumped (fibrous)

Red Blood Cell
(sickle-shaped)

Blood Vessels
Forms clots /
blockages

https://www.23andme.com/ancestry-composition-guide/



Ancestry Composition:

23andMe's State-of-the-Art Geographic Ancestry Analysis

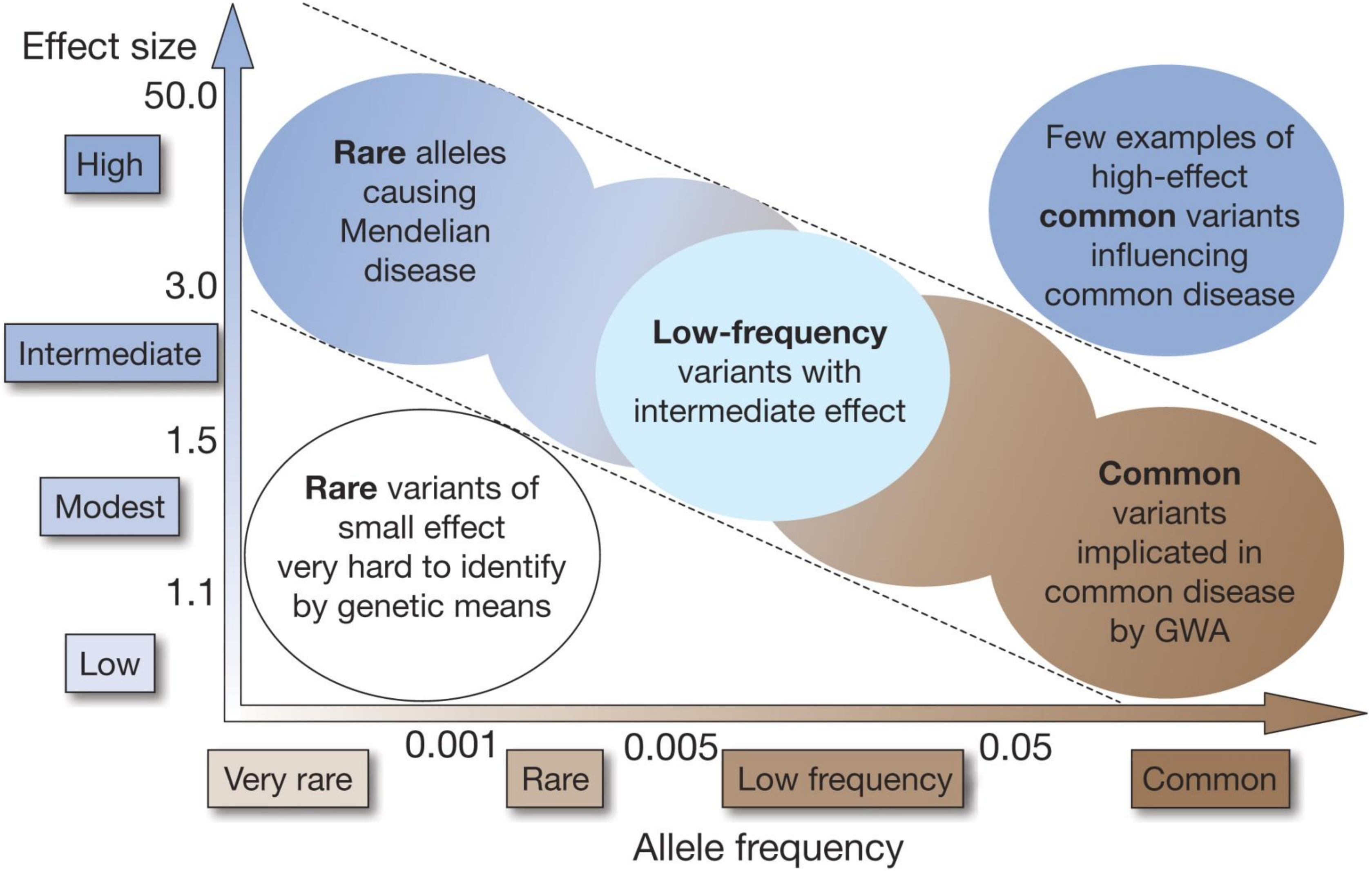
23andMe's Ancestry Composition report is a powerful and well-tested system for analyzing ancestry based on DNA, and we believe it sets a standard for rigor in the genetic ancestry industry. We wrote this document to explain how our analysis works and to present some quality-control test results. Note: This document goes into specifics for the current version of Ancestry Composition, offered to customers on the V5 platform. For customers on previous platforms, click here.

Your Ancestry Composition report shows the percentage of your DNA that comes from 45 populations. We calculate your Ancestry Composition by comparing your genome to those of over 10,000 people with known ancestry. When a segment of your DNA closely matches the DNA from one of the 45 populations, we assign that ancestry to the corresponding segment of your DNA. We calculate the ancestry for individual segments of your genome separately, then add them together to compute your overall ancestry composition.

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



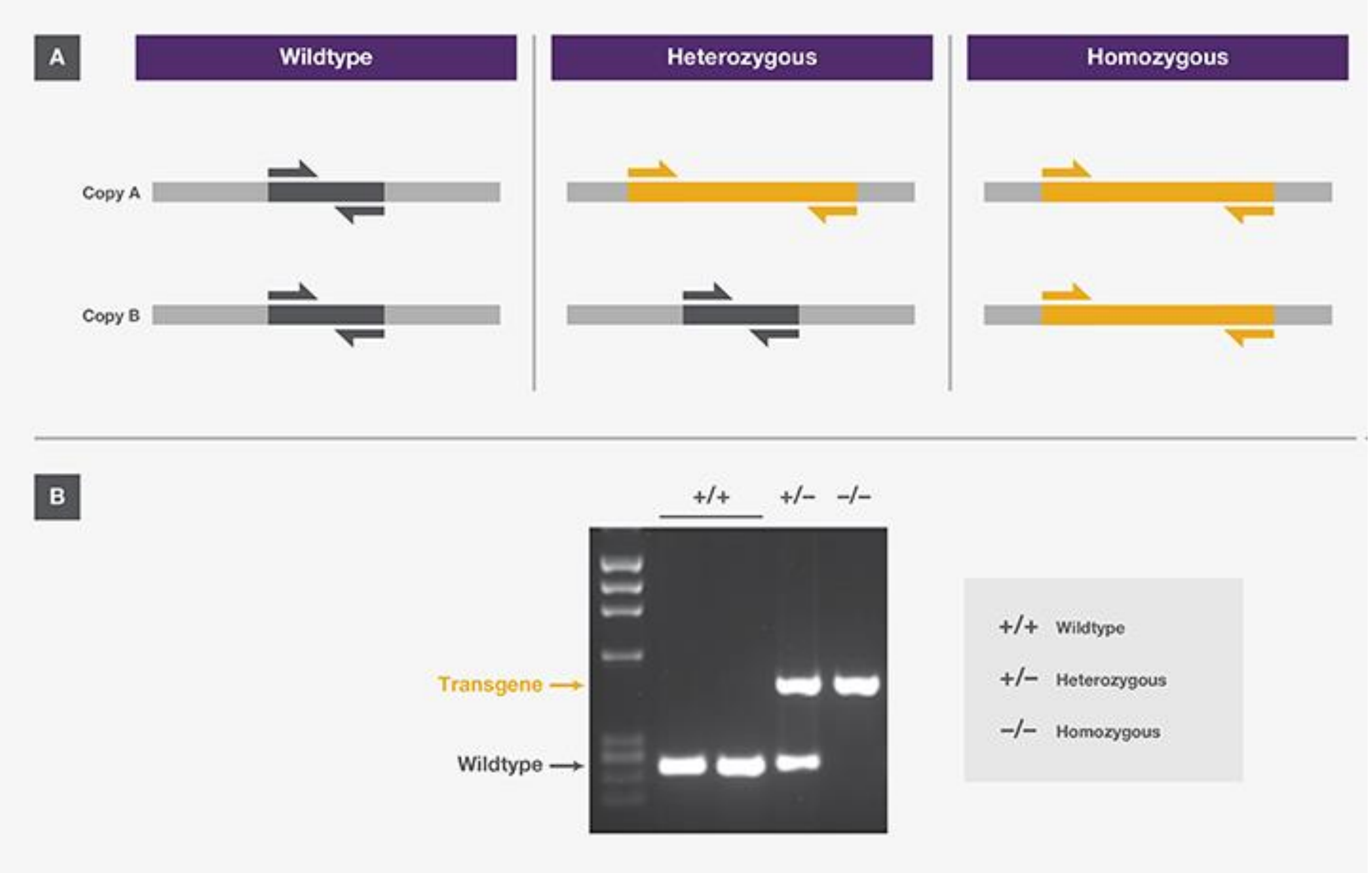
Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect **common** variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

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

1500 - 300 SNPs

SNPlex

400 - 40 SNPs

Sequenom

40 - 5 SNPs

TaqMan

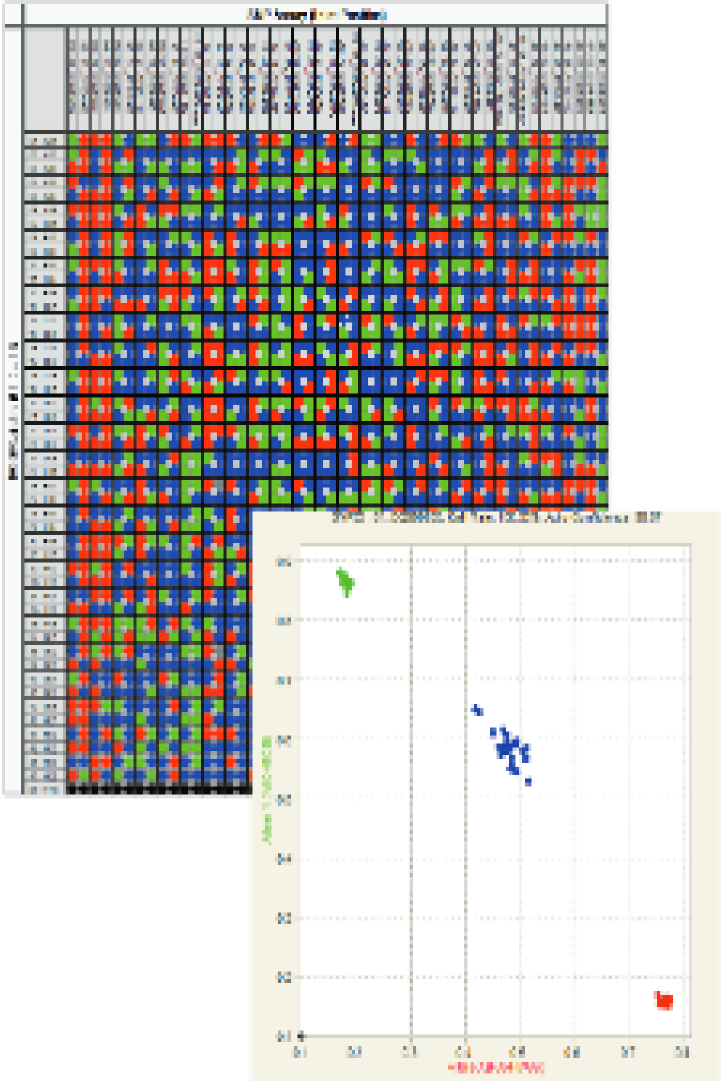
10 - 1 SNPs

Chip Genotyping

Why we like SNPs:

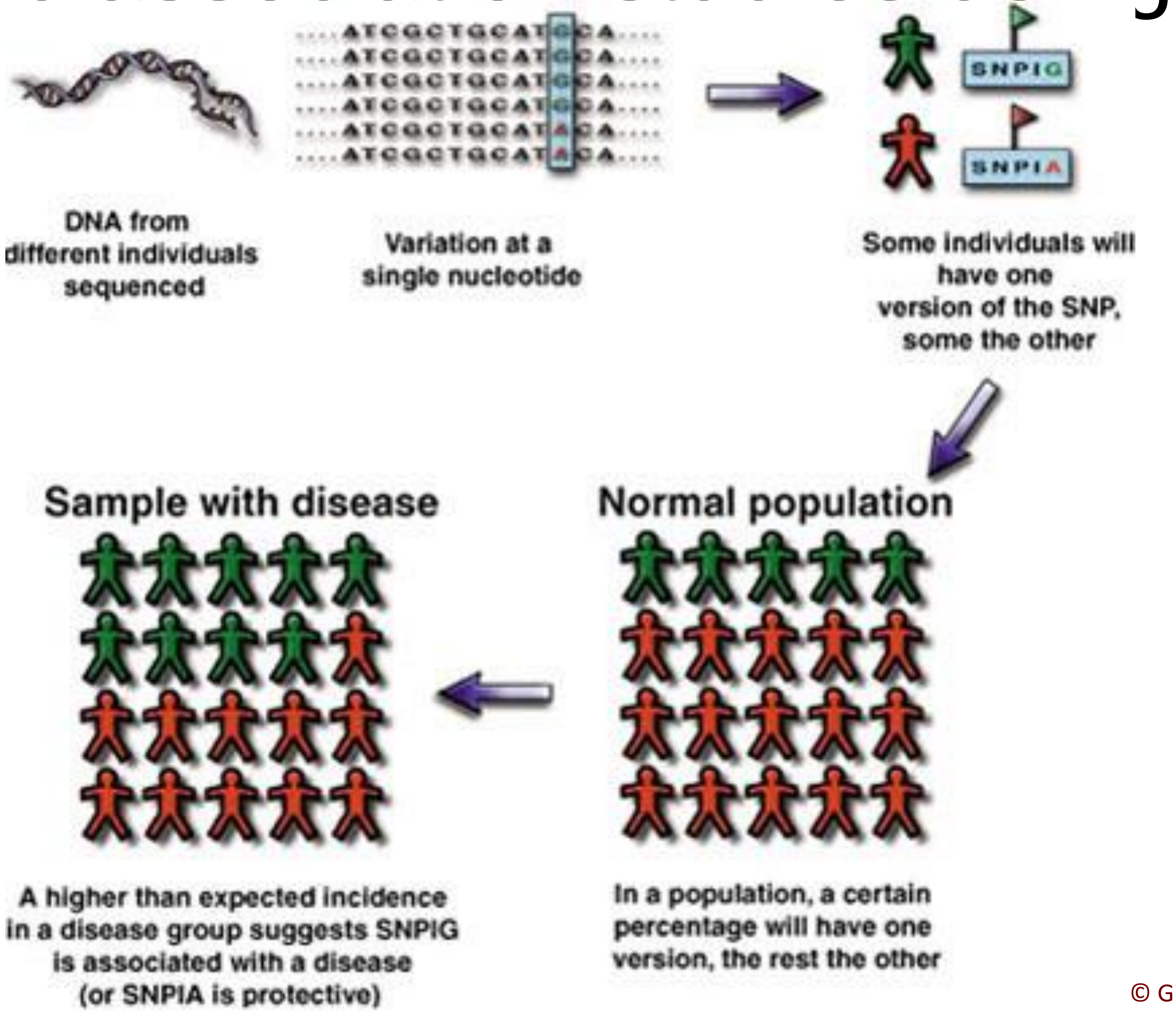
- Abundant in the genome
- Easy to measure

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



Fluidigm platform

Genetic association studies using SNPs



DNA from different individuals sequenced

Variation at a single nucleotide

Some individuals will have one version of the SNP, some the other

Sample with disease

Normal population

A higher than expected incidence in a disease group suggests SNPIG is associated with a disease (or SNPIA is protective)

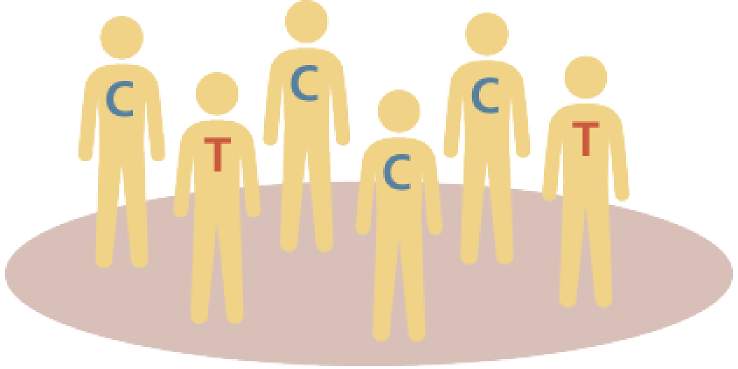
In a population, a certain percentage will have one version, the rest the other

© Gibson & Muse, A Primer of Genome Science

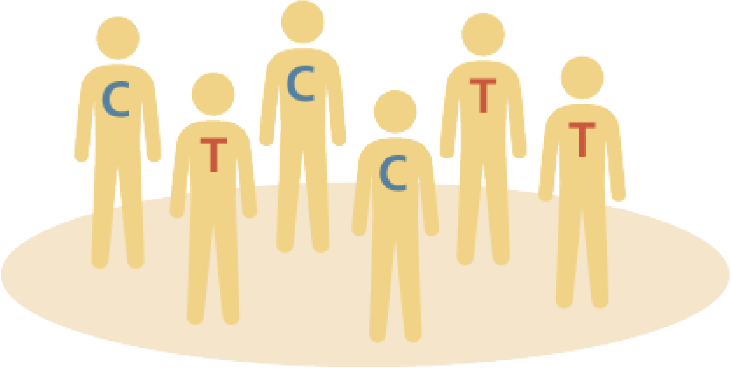
Association studies

- 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.

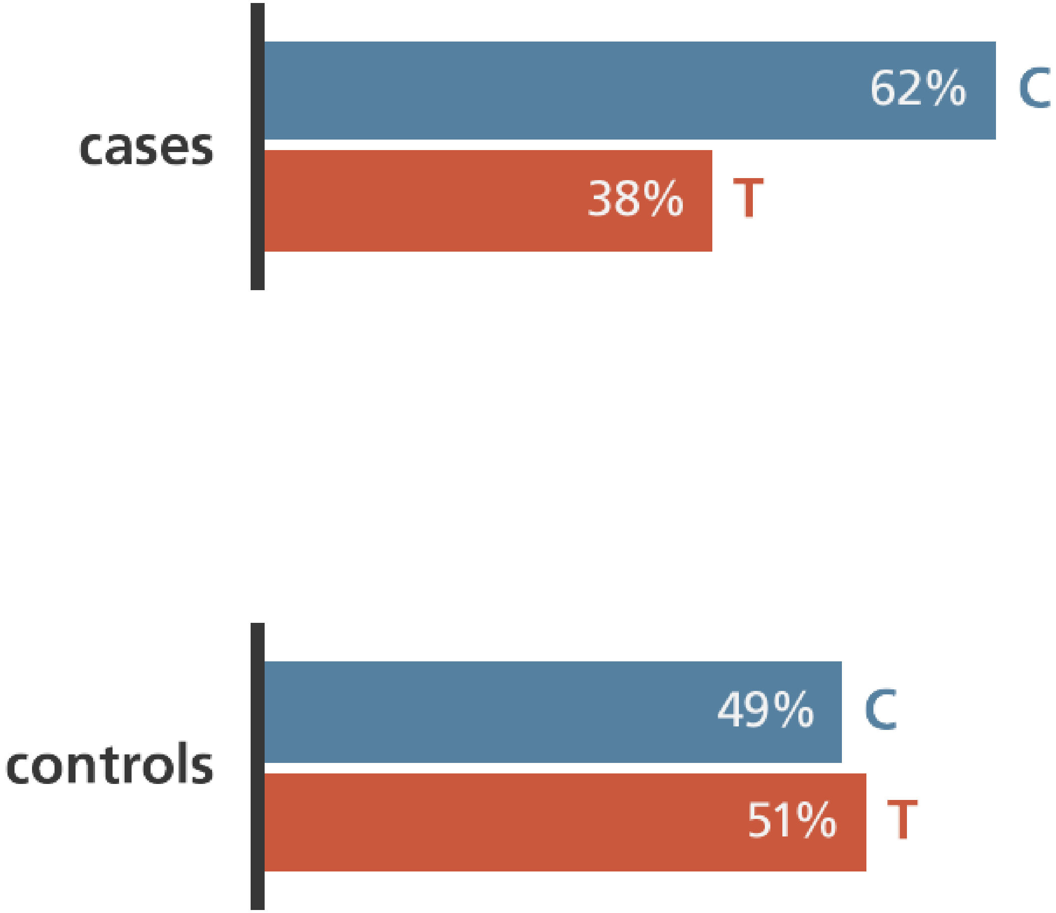
SNP



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



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



Odds ratio

The odds ratio is our measure of association for a case-control study. It tells us whether and how much an exposure increases the likelihood of our outcome of interest. We need to look at two things:

The estimate -- the odds ratio itself. How big in the connection between an exposure and an outcome? Are those with an exposure more likely to have the outcome?

The p-value -- how certain are we that the odds ratio didn’t just happen by chance?

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	

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

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

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

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

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: $= \frac{a}{b}$

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

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.

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

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 $= \frac{c}{d}$

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

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

H_0: OR = 1 (no association)

OR > 1 indicates increased odds

OR < 1 indicates decreased odds (protective)

Confidence intervals for odds ratios

		Disease status	
		Cases	Controls
Genotype	M	a	b
	m	c	d

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

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

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

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

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

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

Odds ratio calculations – odds ratio itself

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

OR = 1.7

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

Odds ratio calculations – odds ratio itself

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

OR = 1.7

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

The odds of developing thyroid cancer are 1.7x times greater with an T allele compared to without an T allele.

Why do we even use odds and odds ratios???

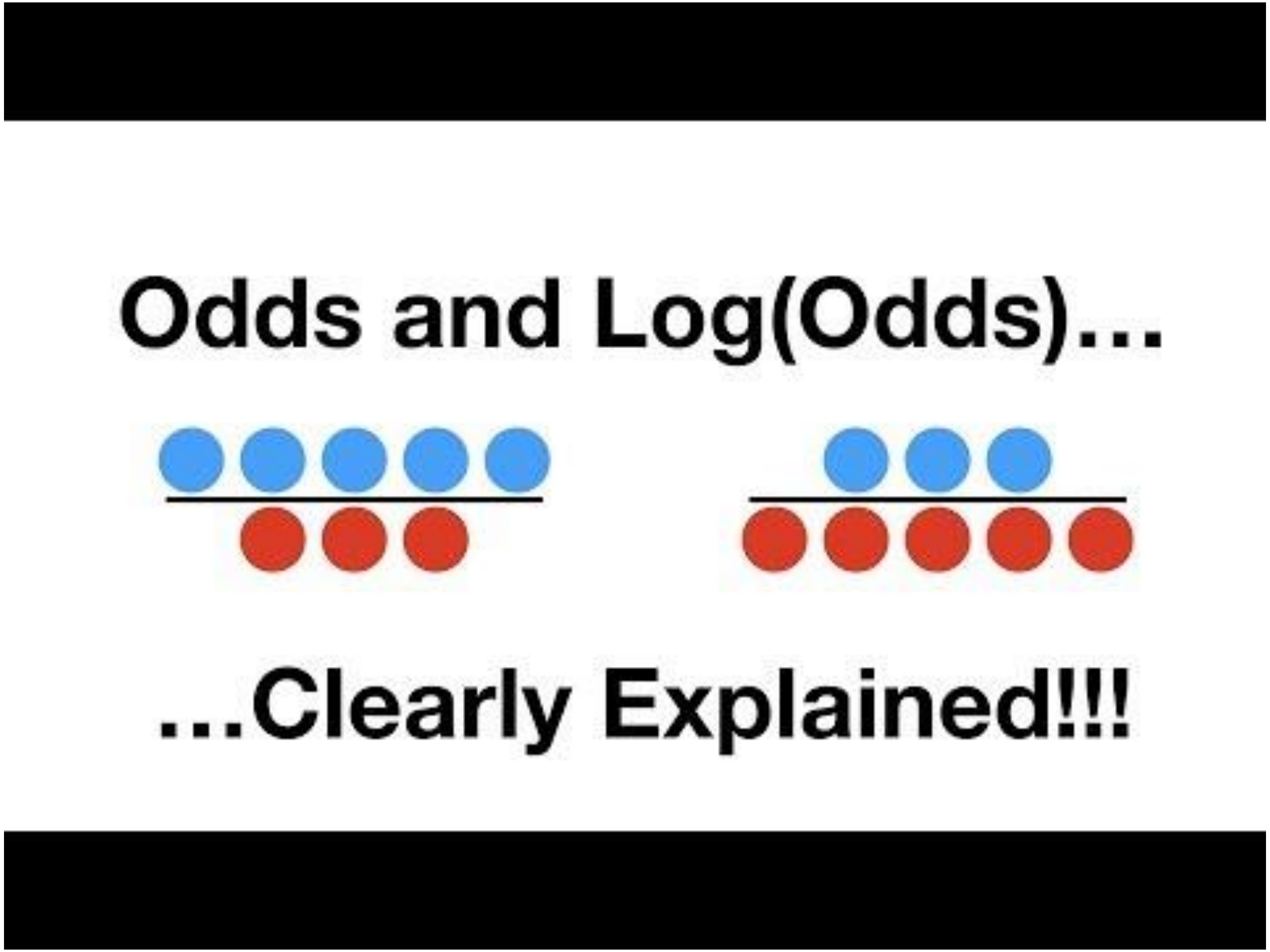
The odds ratio allows us to calculate the associations between an exposure and an outcome without needing the frequency of the exposure in the general population

(very useful to rare exposures, such as rare diseases).

(we'd have to sample A LOT of people to get a true population picture and even pick up one or two cases of the disease)

The log(odds) allows us to transform this weird variable into a linear form, which is easier for us to fit to models and interpret the output.

Why do we use Log odds 5:26 - 8:42



Often use logistic regression for case-control analyses

Allows you to adjust for relevant factors

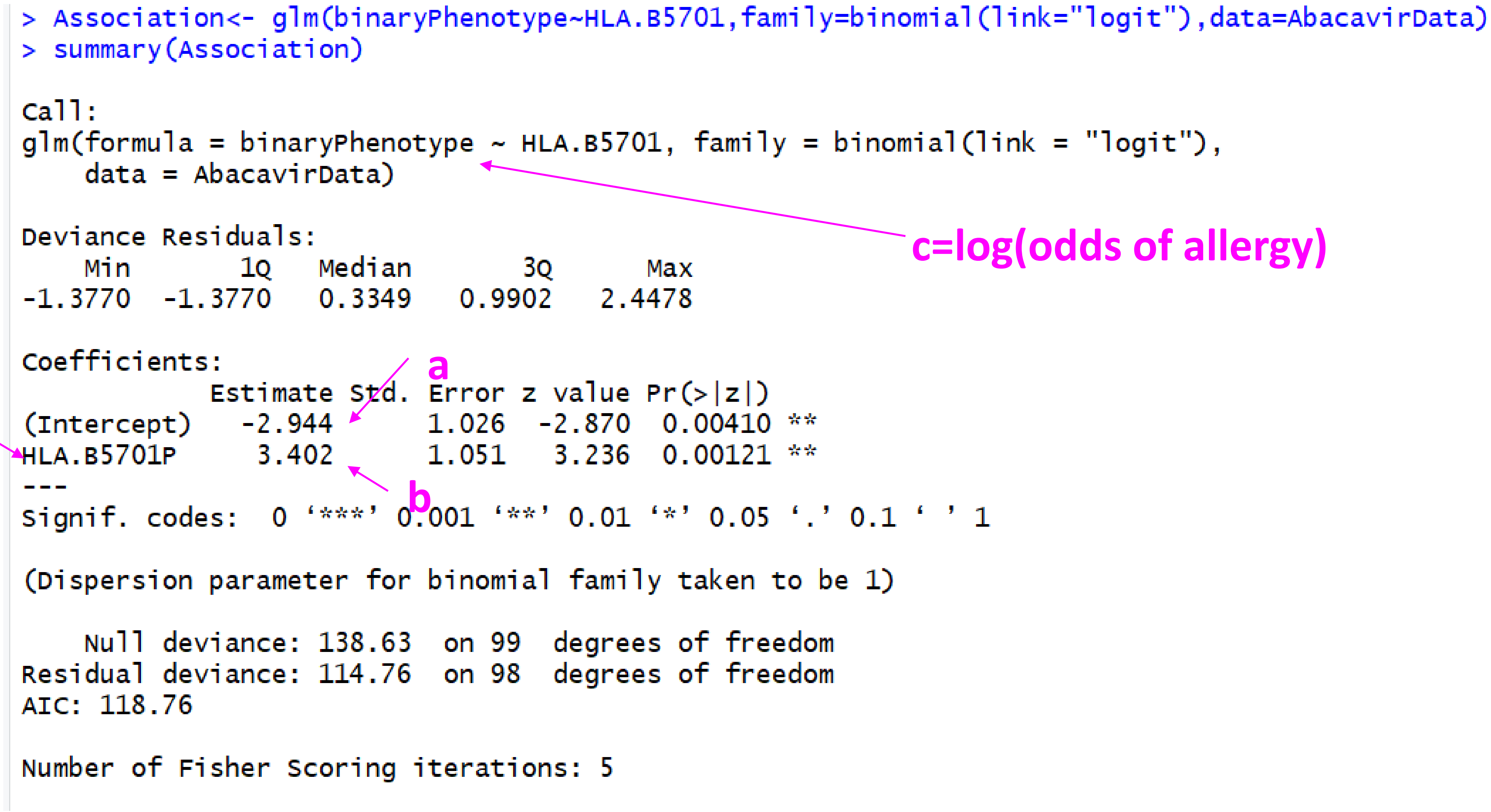
- Population stratification, age, sex, matching variables etc

$\ln\left(\frac{p}{1-p}\right)$= α + β_1**g** + β_2x$_1$ + ….+β_{k+1}x$_k$ (g is genotype, x$_1$,…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$_0$: β_1 =0 (likelihood ratio test, wald test, score test)
- The odds ratio is OR=e^{β_1}
- β_1 = SNP effect (log(OR)) ➔ e $^{\beta_1}$ = OR

Logistic regression output



```
> Association<- glm(binaryPhenotype~HLA.B5701,family=binomial(link="logit"),data=AbacavirData)
> summary(Association)

Call:
glm(formula = binaryPhenotype ~ HLA.B5701, family = binomial(link = "logit"),
    data = AbacavirData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-1.3770  -1.3770   0.3349   0.9902   2.4478  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)   
(Intercept)   -2.944      1.026  -2.870  0.00410 **
HLA.B5701P     3.402      1.051   3.236  0.00121 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138.63  on 99  degrees of freedom
Residual deviance: 114.76  on 98  degrees of freedom
AIC: 118.76

Number of Fisher Scoring iterations: 5
```

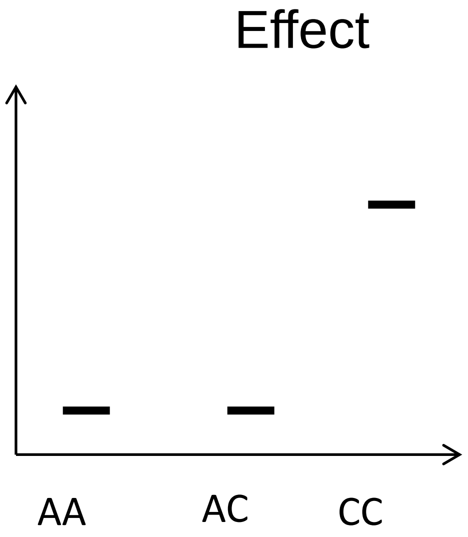
c=log(odds of allergy)

a

x

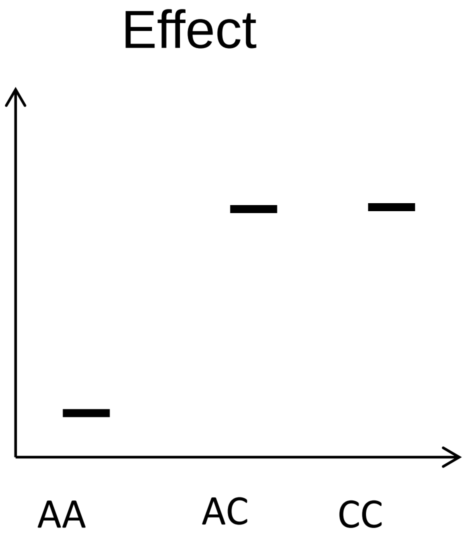
b

Common models of penetrance



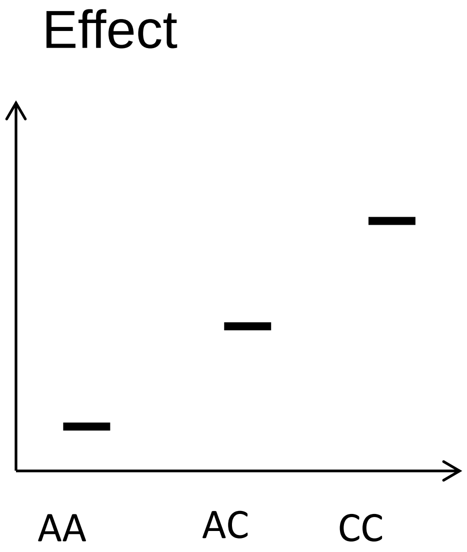
Recessive

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

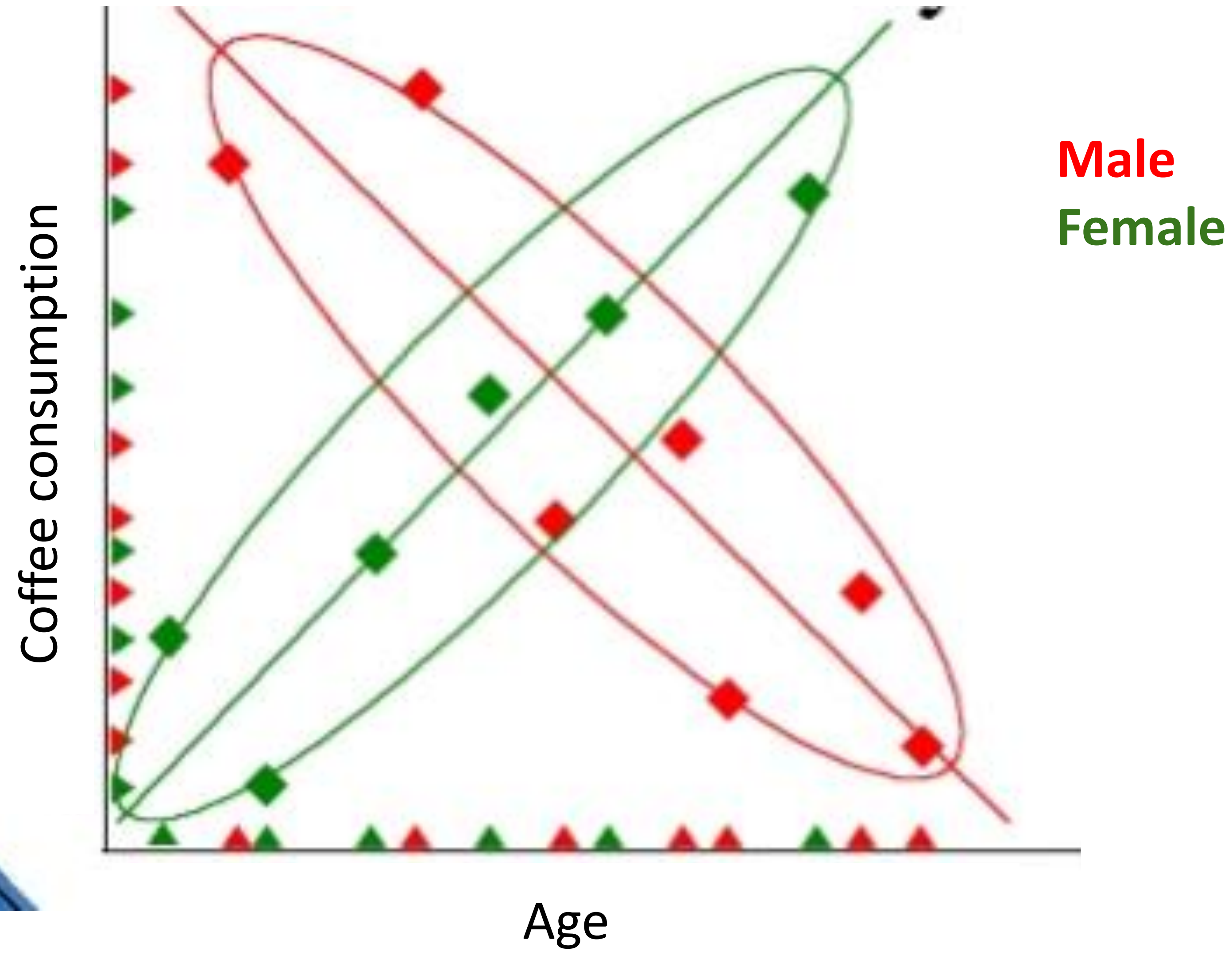
Continuous outcome genetic association

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

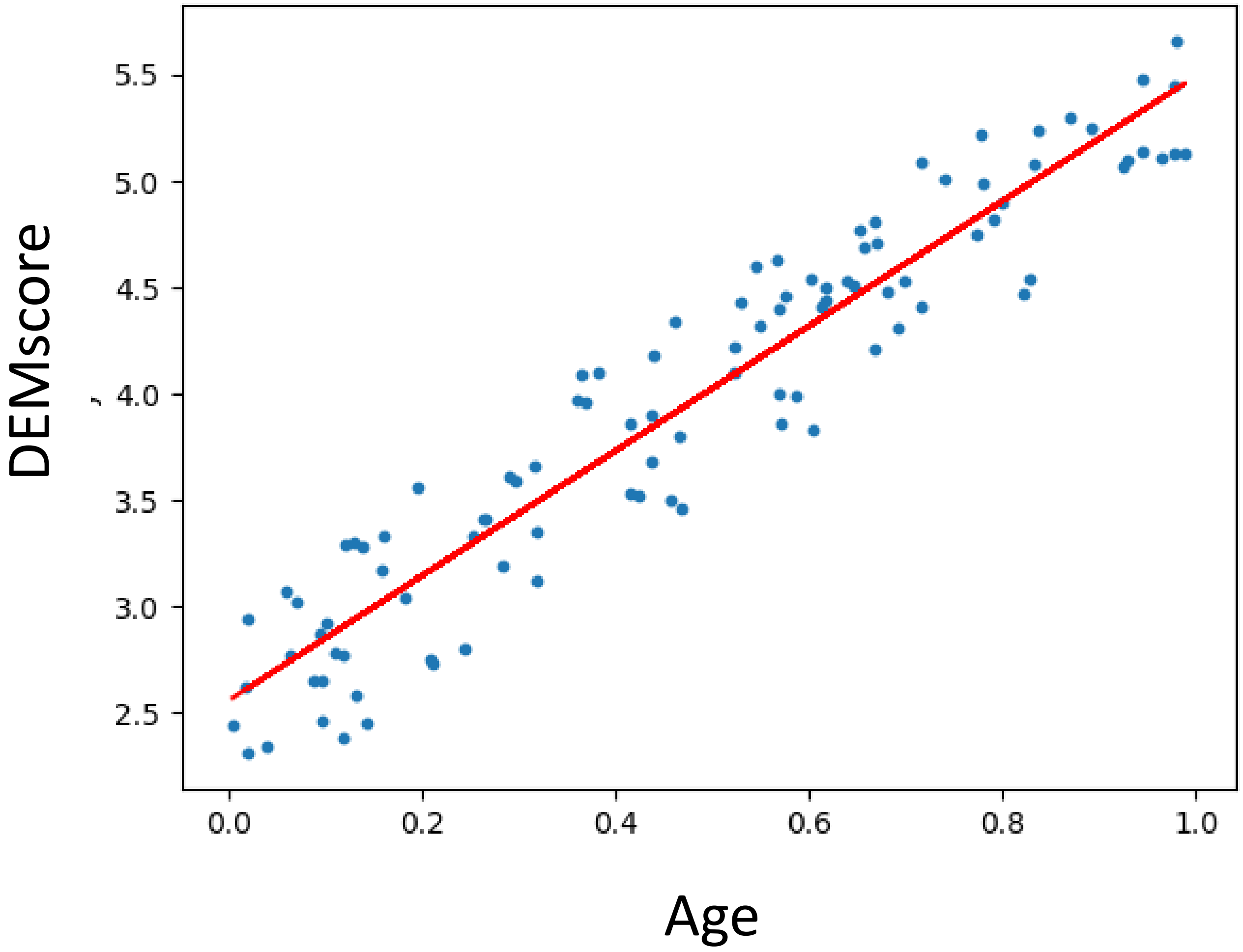
$$Y = \alpha + \beta * SNP + X$$

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

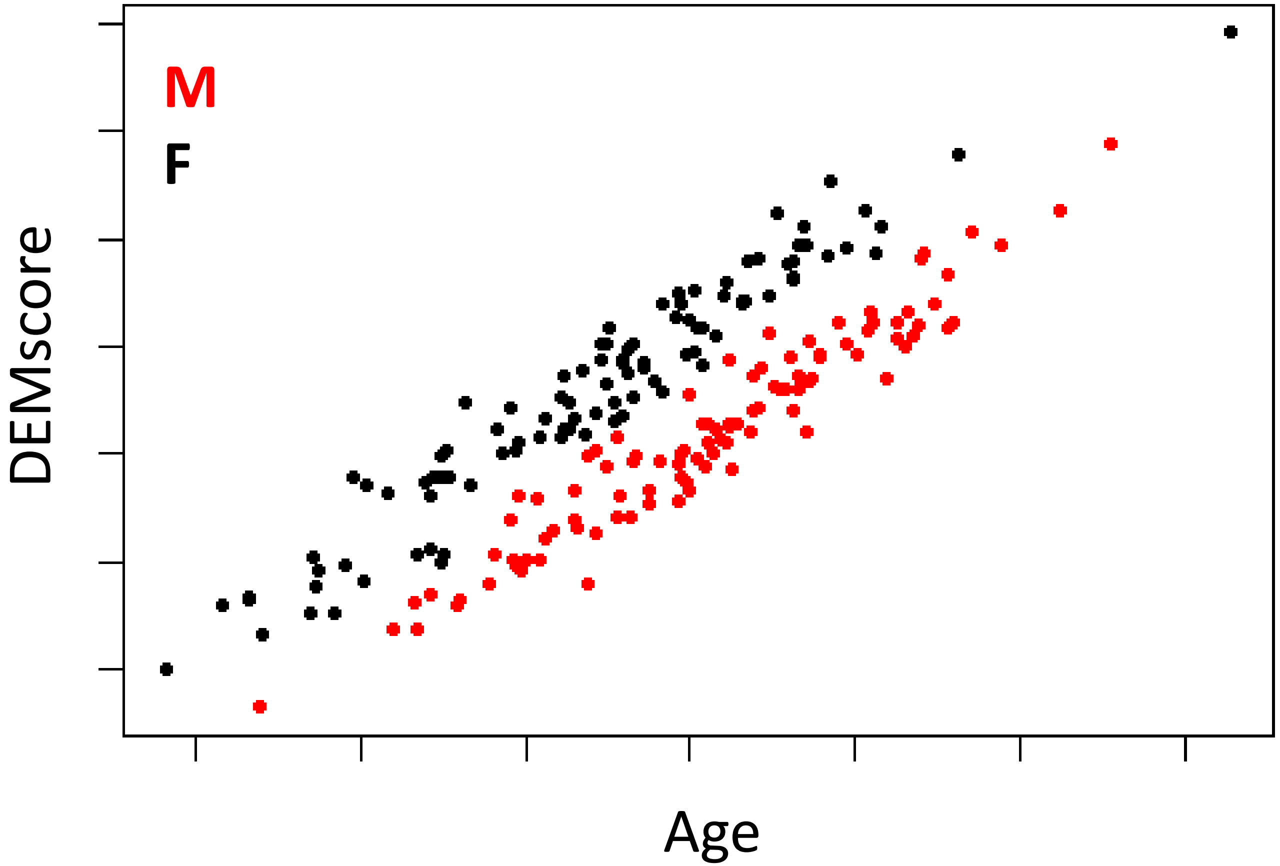
Multivariate analyses



One predictor, one outcome



Multivariate analysis



Importance of setting your reference allele

Odds ratio when AA is reference: ⅔ / ⅓ = ⅔ *3 = 2
The odds of the outcome are 2x more likely among those with CC genotype compared to among those with the AA genotype.

Odds ratio when CC is reference. ⅓ / ⅔ = ⅓ * $^3/_2$ = 0.5
The odds of the outcome are ½ as likely among those with AA genotype compared to among those with the CC genotype.

These are the saying the same thing! But the language matters.

Always know and be purposeful on your reference

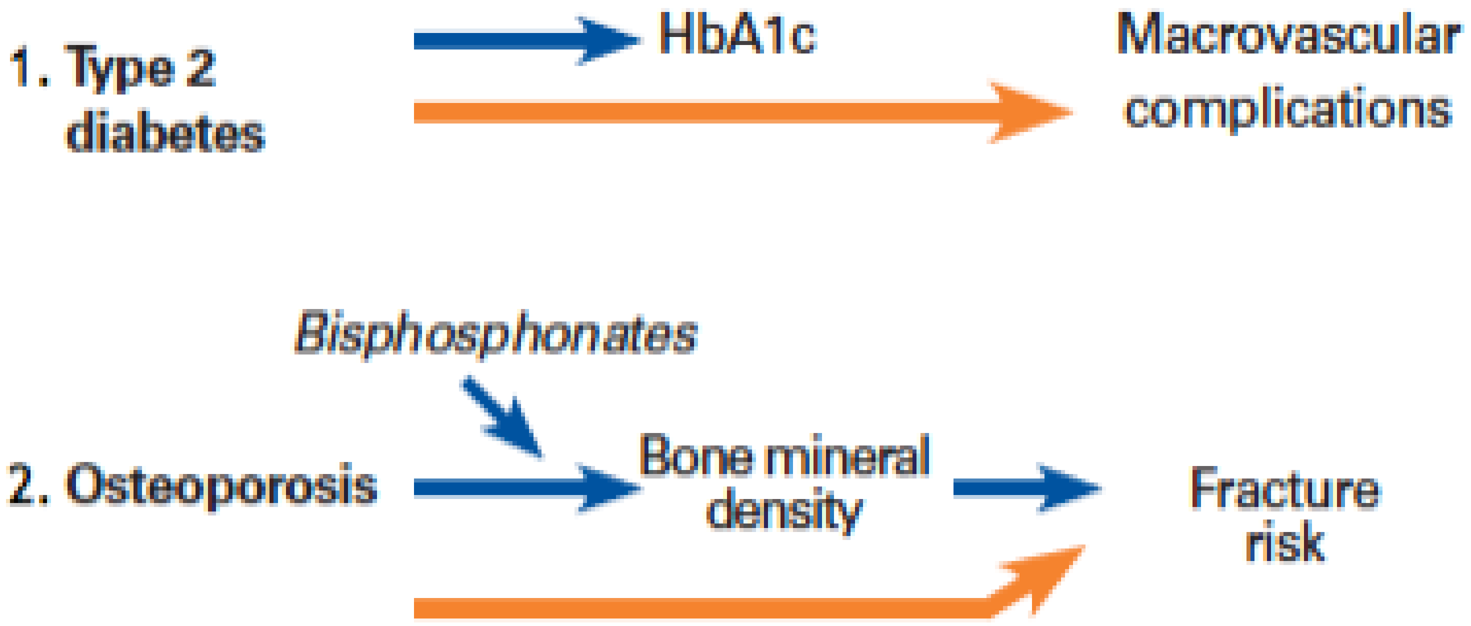
In epidemiology, the reference group always matters.

Exposure (gene allele reference)

Outcome (some outcomes have no “direction”) brown vs black hair

Population (other factors are always involved, i.e. age, diet, access to care).

Surrogate endpoints



1. Type 2 diabetes

HbA1c

Macrovascular complications

Bisphosphonates

2. Osteoporosis

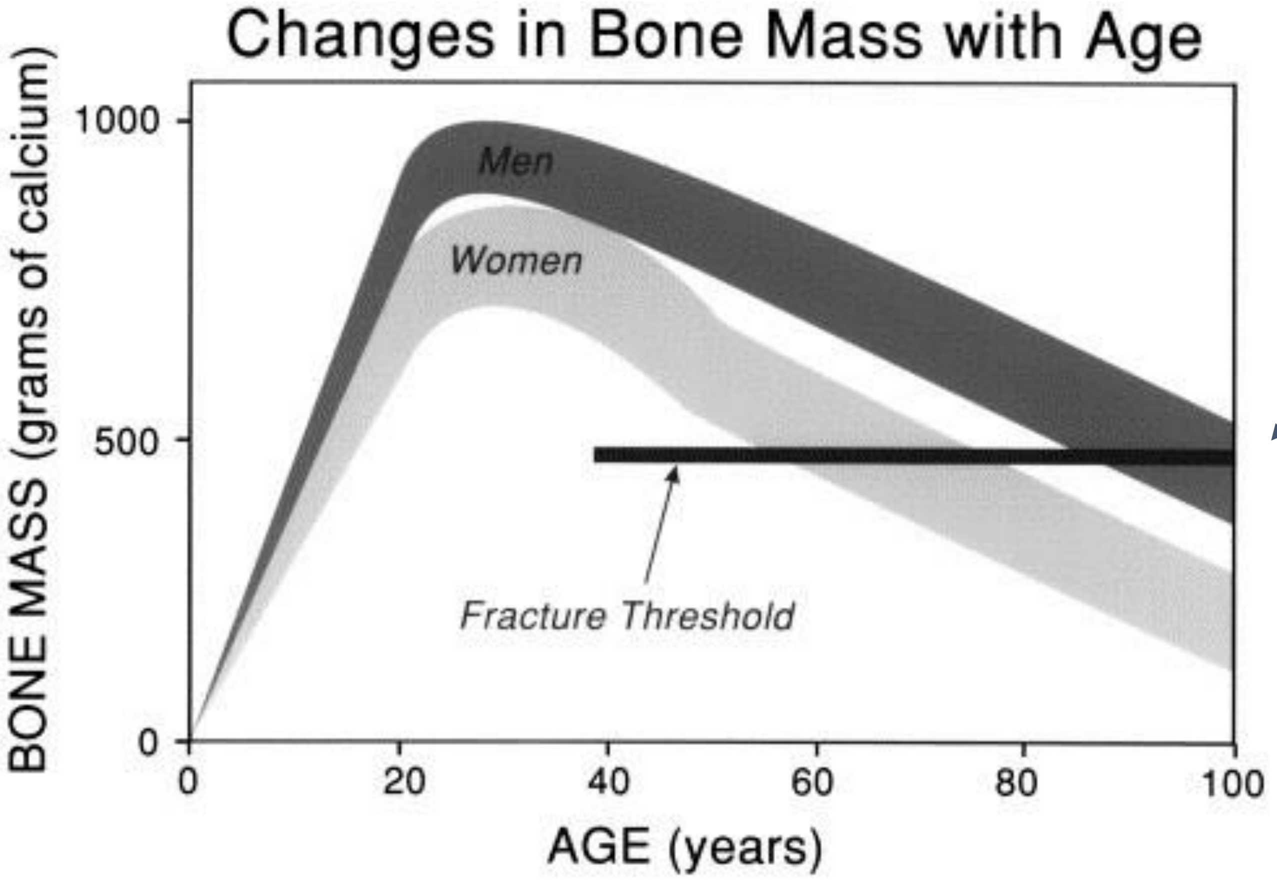
Bone mineral density

Fracture risk

Picking an endpoint

- Surrogate endpoints: an endpoint that in itself means nothing, but gives information about an important endpoint.
    - More proximal in the biological pathway.
        - Time to detect and/or intervene.
        - Can detect earlier and collect more people.
    - Easier to measure, especially if an outcome often results in death.
    - Monitor progress and change in risk.
    - Cheaper to measure and conduct study.
- Problems with surrogate endpoints
    - Misclassification -- loss of precision

We could also have turned Bone mineral density into a binary outcome based on whether the measure was below the threshold for high fracture risk:



Is the BMD more of less than 500g/cm?

Quantitative vs categorical outcomes

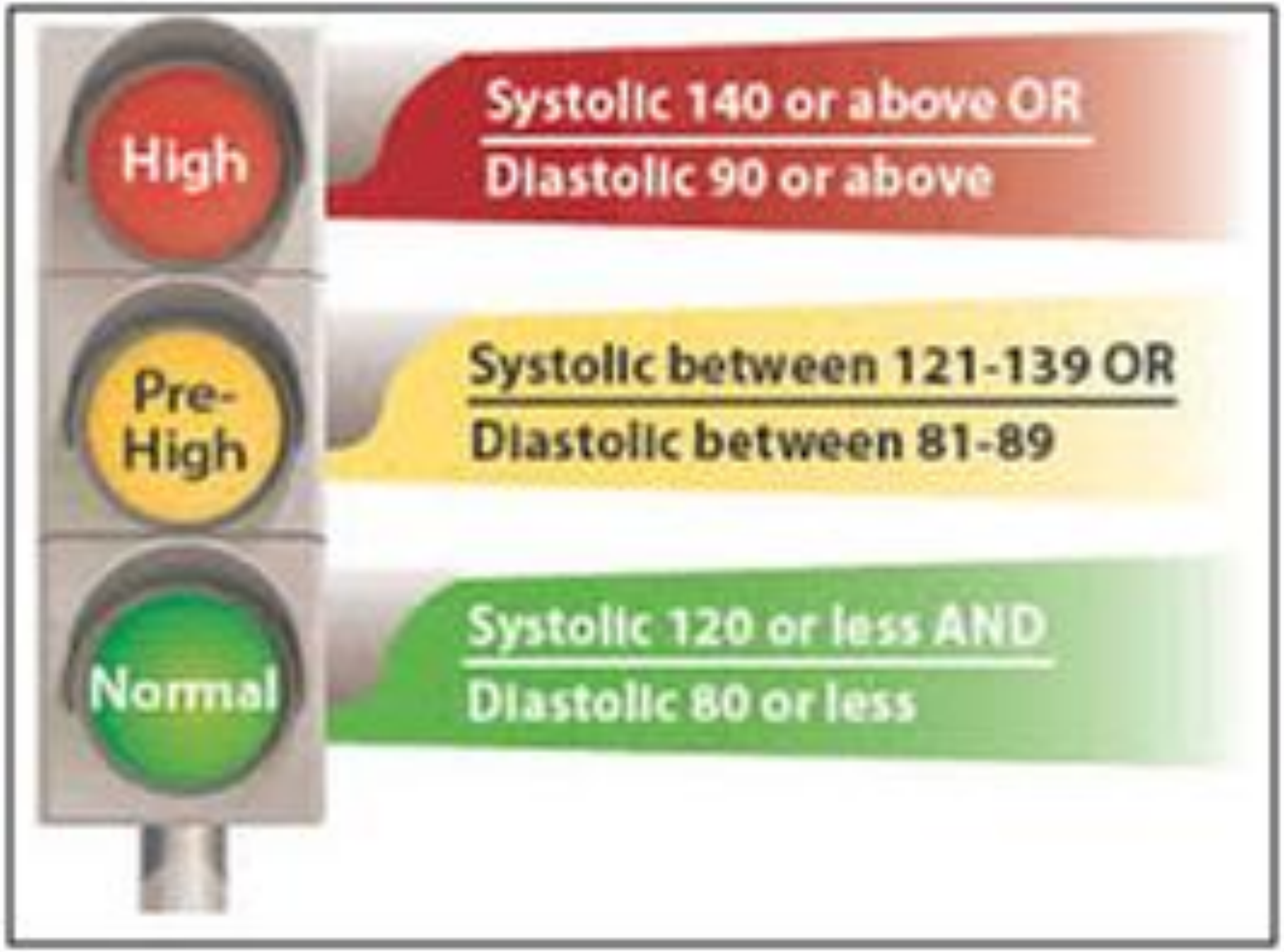
Quantitative

- Does not rely on subjective labels.
- Often more likely to detect differences.
- Interpretation: increase in unit change of phenotype per unit change in risk factor.

Binary

- Must decide cutpoint.
- More straightforward message for action.
- Interpretation: increase in odds of phenotype per unit change in risk factor.

Interpretable cutpoints -> aids policy development



High

Systolic 140 or above OR

Diastolic 90 or above

Pre-High

Systolic between 121-139 OR

Diastolic between 81-89

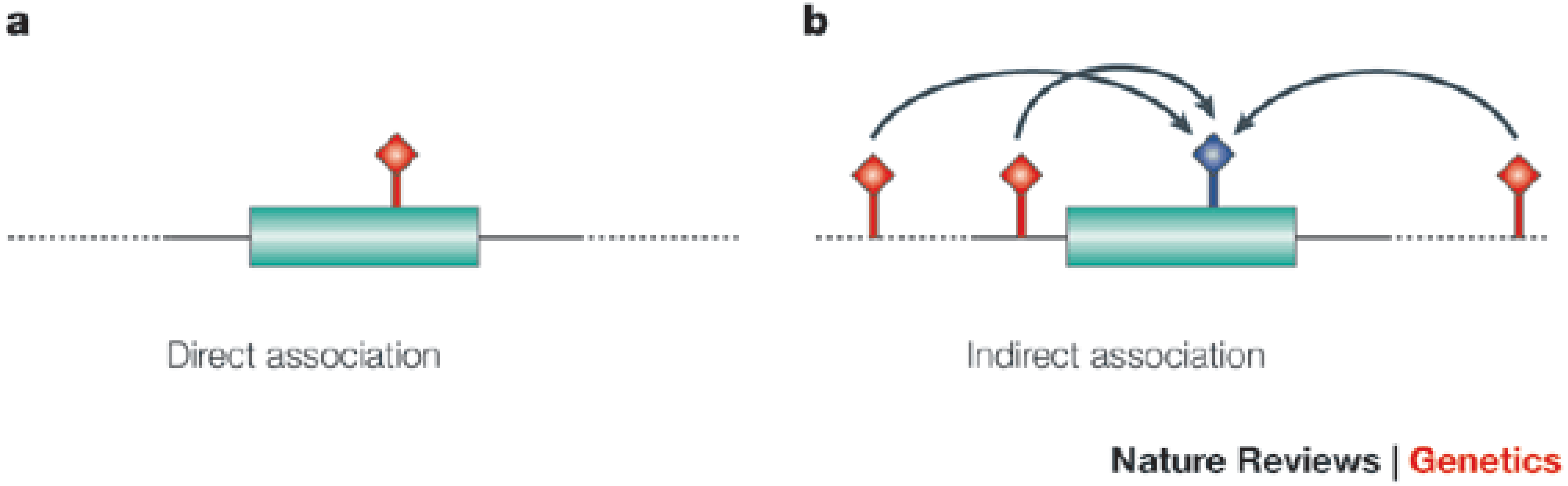
Normal

Systolic 120 or less AND

Diastolic 80 or less

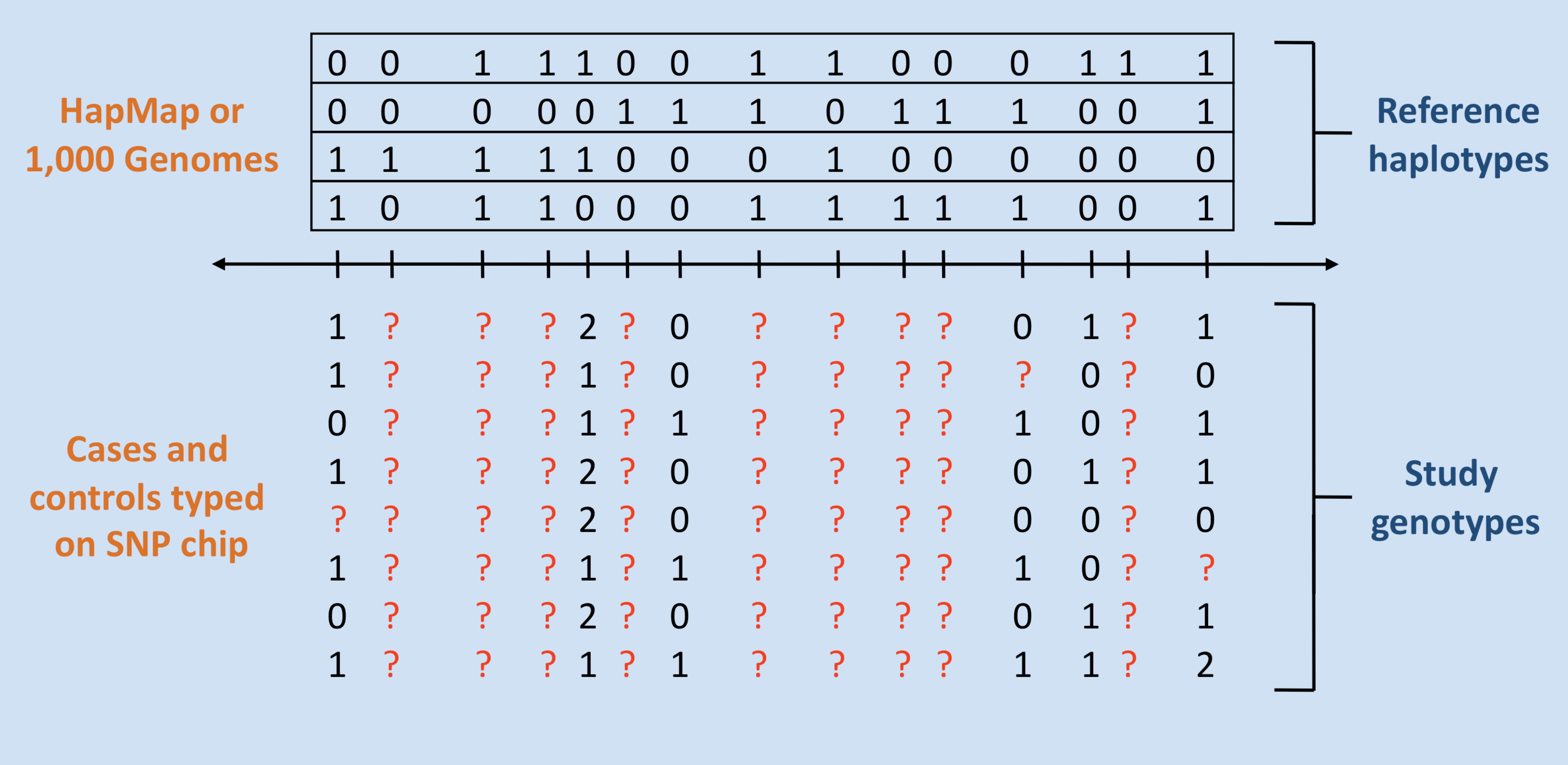
Genotyping platforms can vary by studies, how can we combine data or get more genotyping data than we start with?

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



HapMap or 1,000 Genomes

Reference haplotypes

0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1

Cases and controls typed on SNP chip

Study genotypes

1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	0	?	?	?	?	?	0	?	0
0	?	?	?	1	?	1	?	?	?	?	1	0	?	1
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
?	?	?	?	2	?	0	?	?	?	?	0	0	?	0
1	?	?	?	1	?	1	?	?	?	?	1	0	?	?
0	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	1	?	?	?	?	1	1	?	2

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

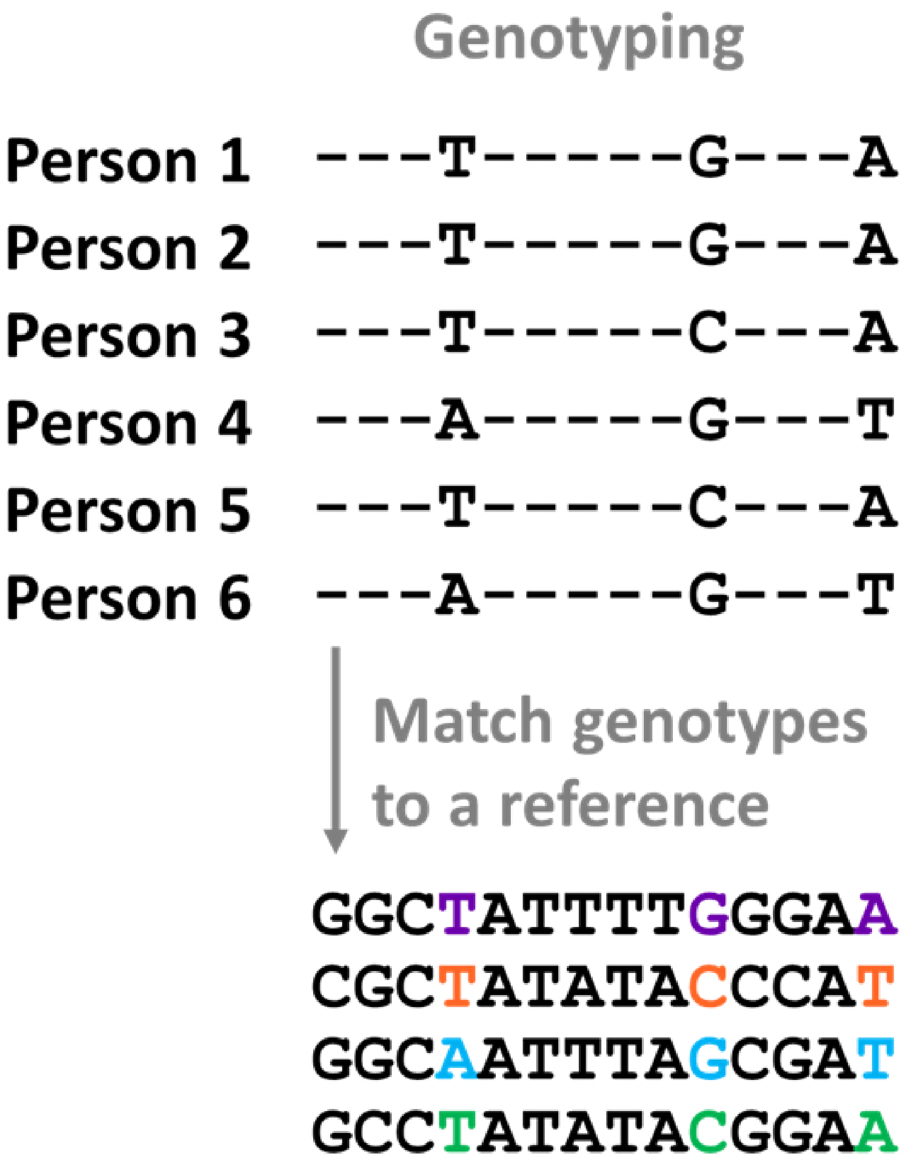
Imputation

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



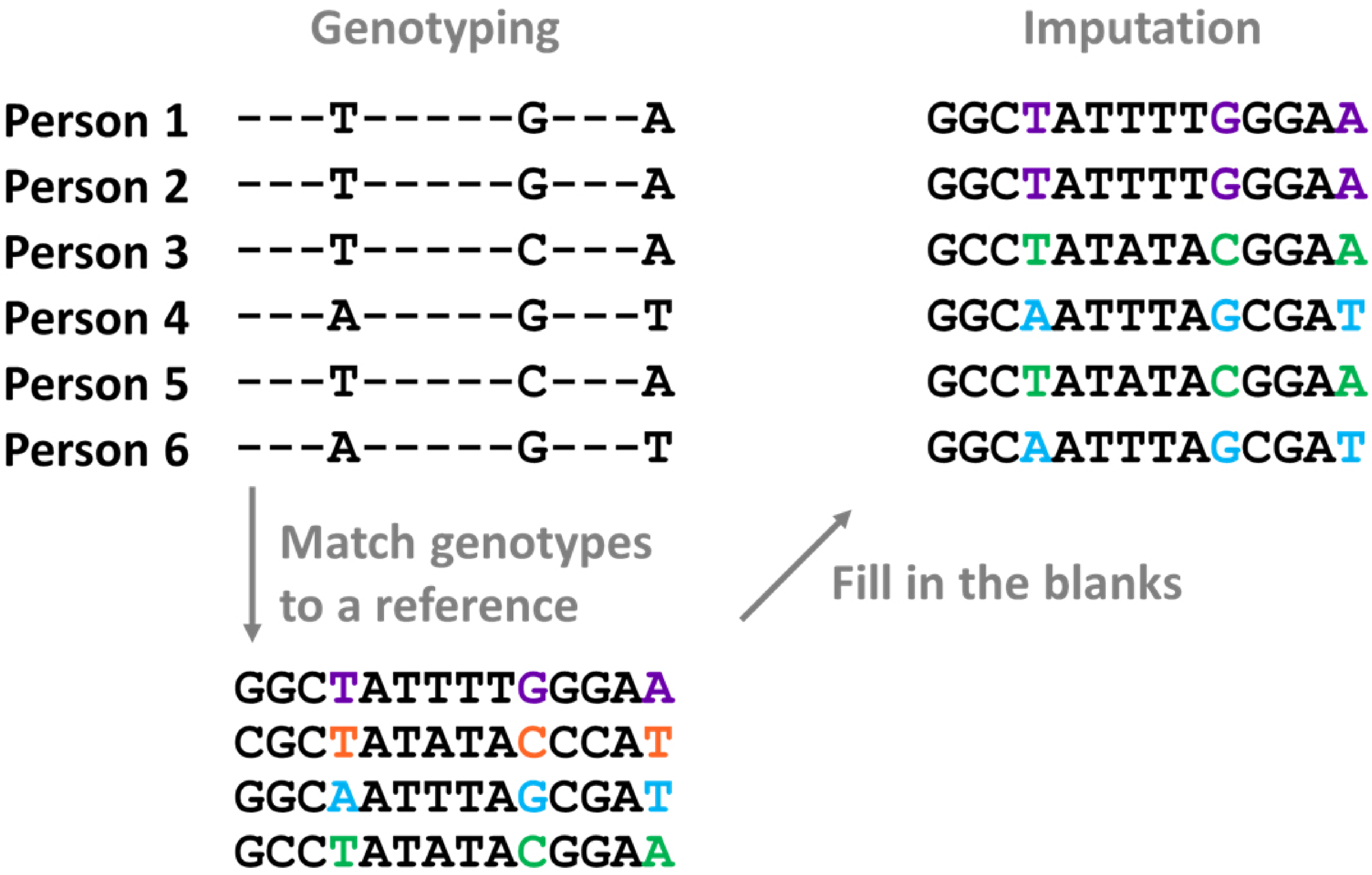
Genotyping

```
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
```

Match genotypes to a reference

```
GGCTATTTTGGAA
CGCTATATACCCAT
GGCAATTTAGCGAT
GCCTATATACGGAA
```

Can you impute the missing bases?



Genotyping

Imputation

	Genotyping	Imputation
Person 1	---T-----G---A	GGCTATTTTGGGAA
Person 2	---T-----G---A	GGCTATTTTGGGAA
Person 3	---T-----C---A	GCCTATATACGGAA
Person 4	---A-----G---T	GGCAATTTAGCGAT
Person 5	---T-----C---A	GCCTATATACGGAA
Person 6	---A-----G---T	GGCAATTTAGCGAT

Match genotypes to a reference

GGCTATTTTGGGAA
CGCTATATACCCAT
GGCAATTTAGCGAT
GCCTATATACGGAA

Fill in the blanks

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

- 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)

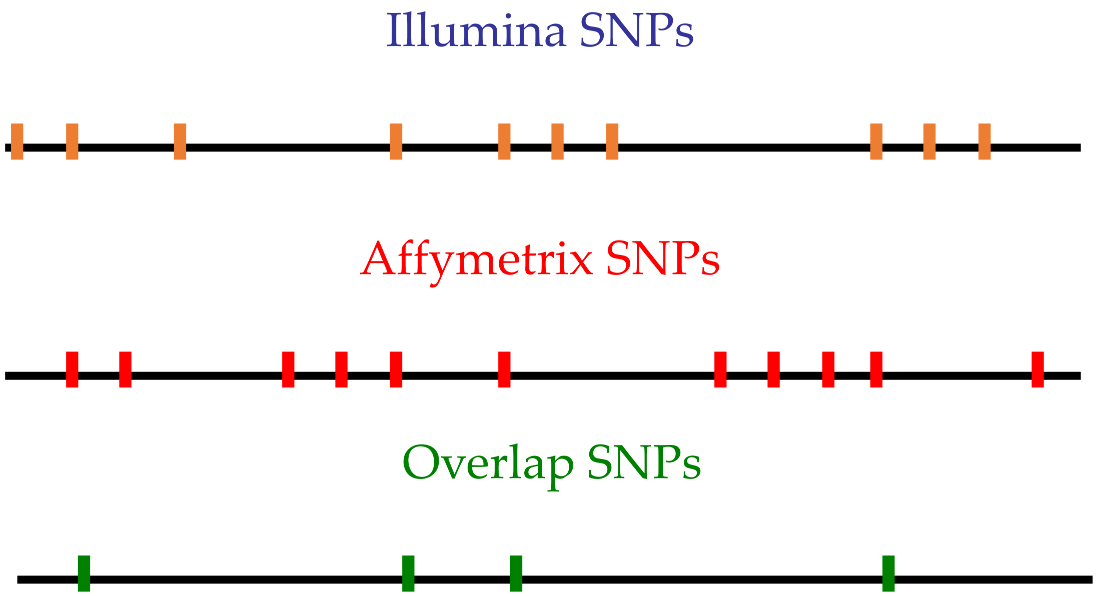
Assessing SNPs across genotyping platforms

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

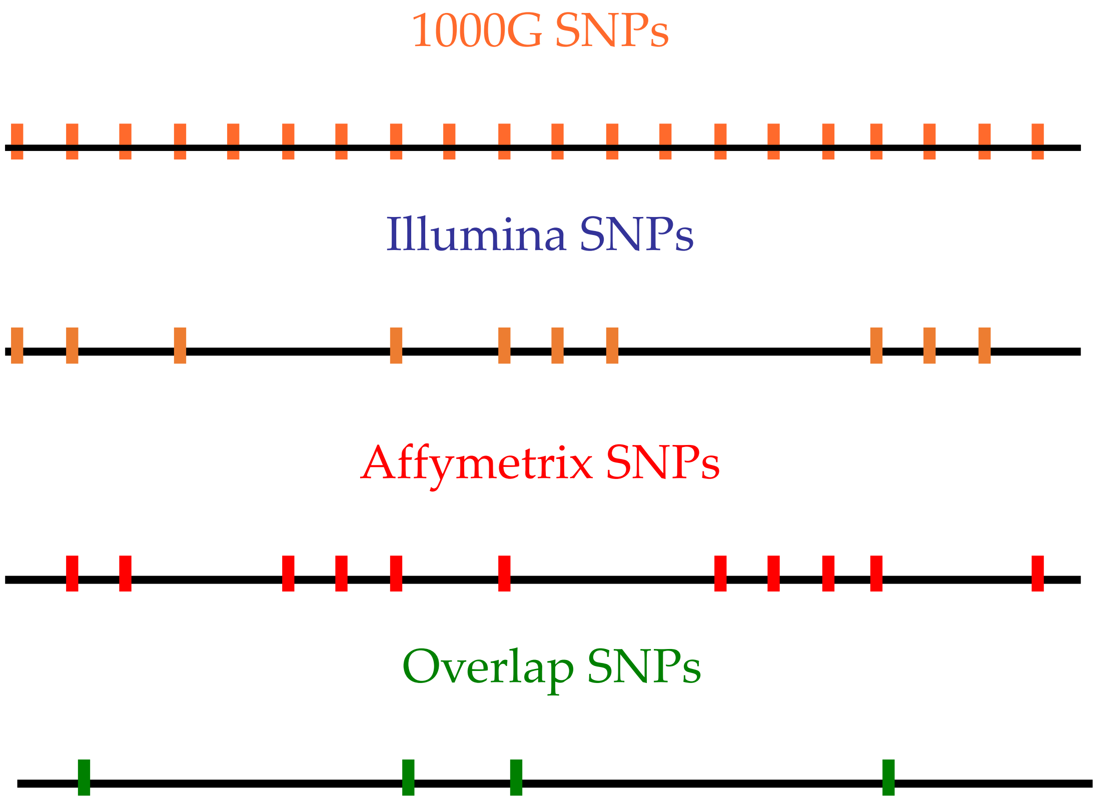
* 75,285 markers are on all 3 platforms

Lindström, PLoS One 2017

Imputation for studying SNPs across platforms



Imputation for studying SNPs across platforms

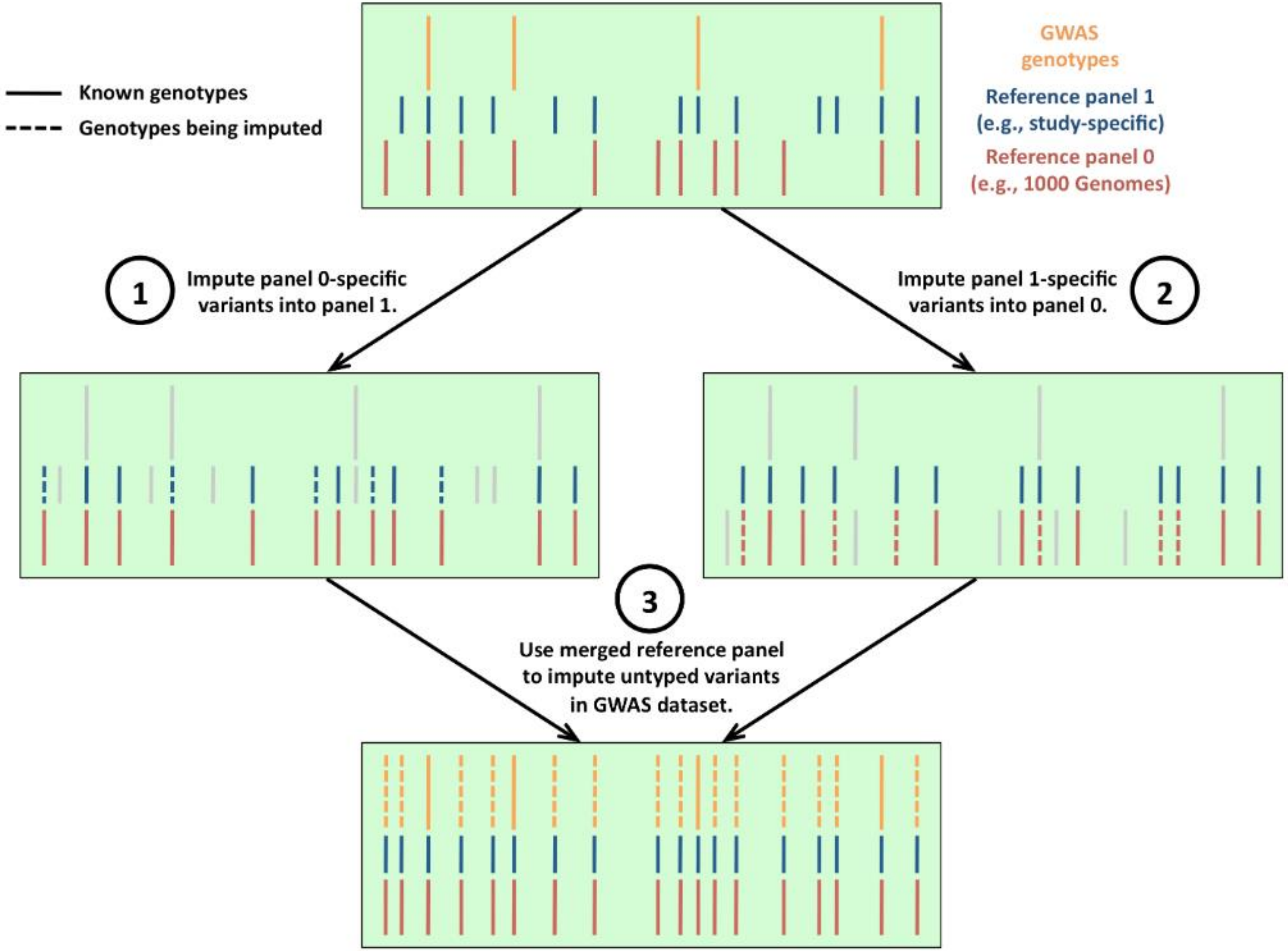


1000G SNPs

Illumina SNPs

Affymetrix SNPs

Overlap SNPs



— 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

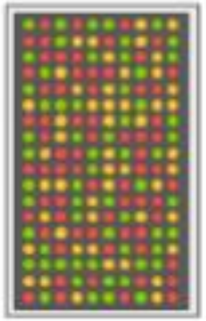
- The imputation quality score r^2 measures how well a SNP was imputed.
  - Ranges between 0 and 1.
  - A quality score of r^2 on a sample of *N* individuals indicates that the amount of data at the imputed SNP is approximately equivalent to a set of perfectly observed genotype data in a sample size 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

What if we don’t know what variants to test or they are too rare to impute?

Sequencing vs Genotyping: Discovery



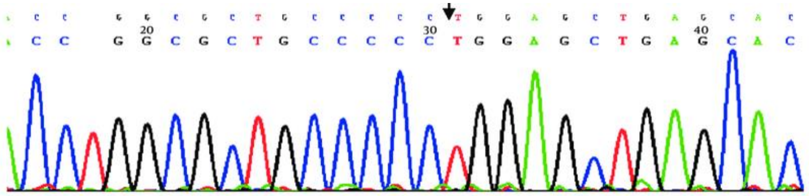
Array imaged to analyze signal at all beads

Genotyping:

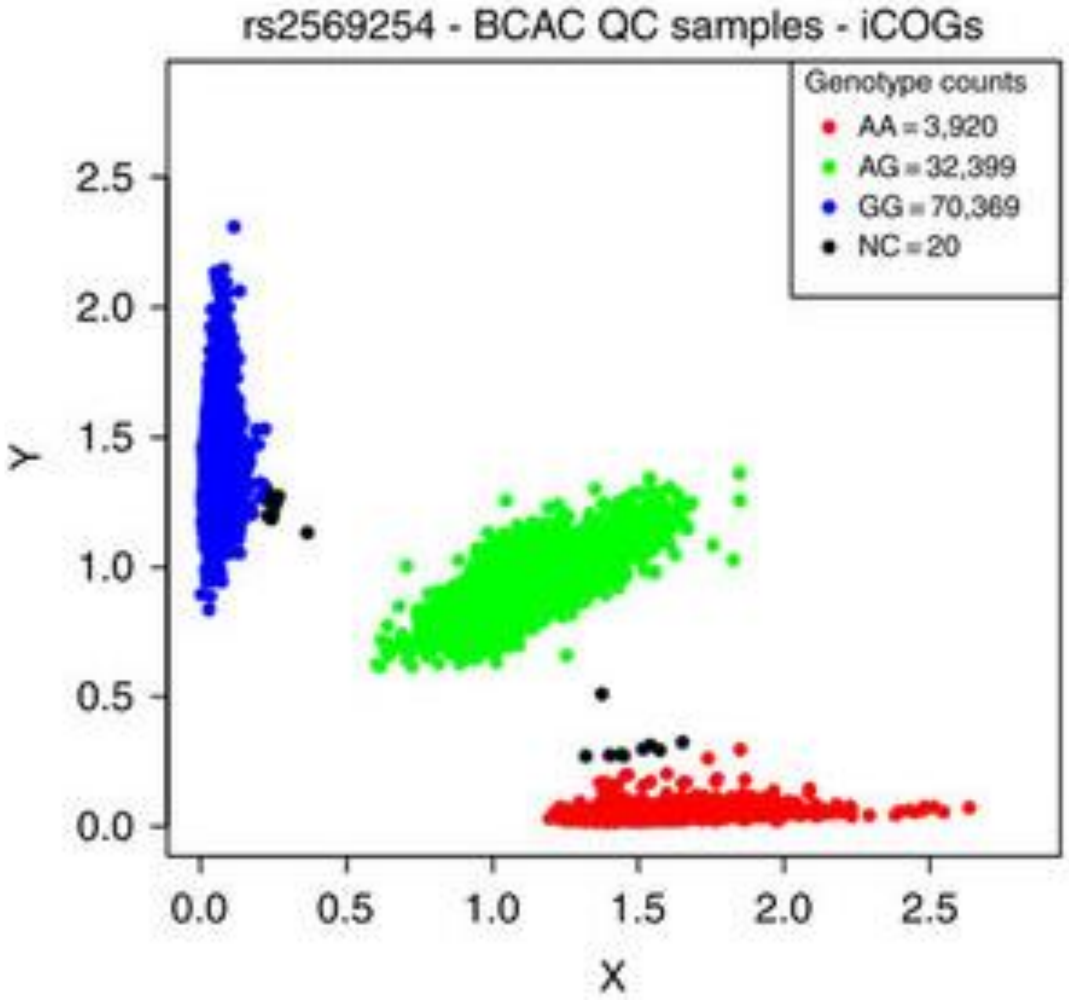
- Common variants (>5% allele frequency)
- large cohorts (cheaper)
- to identify regions of the genome associated with an outcome
- less computationally demanding to get a person’s alleles.

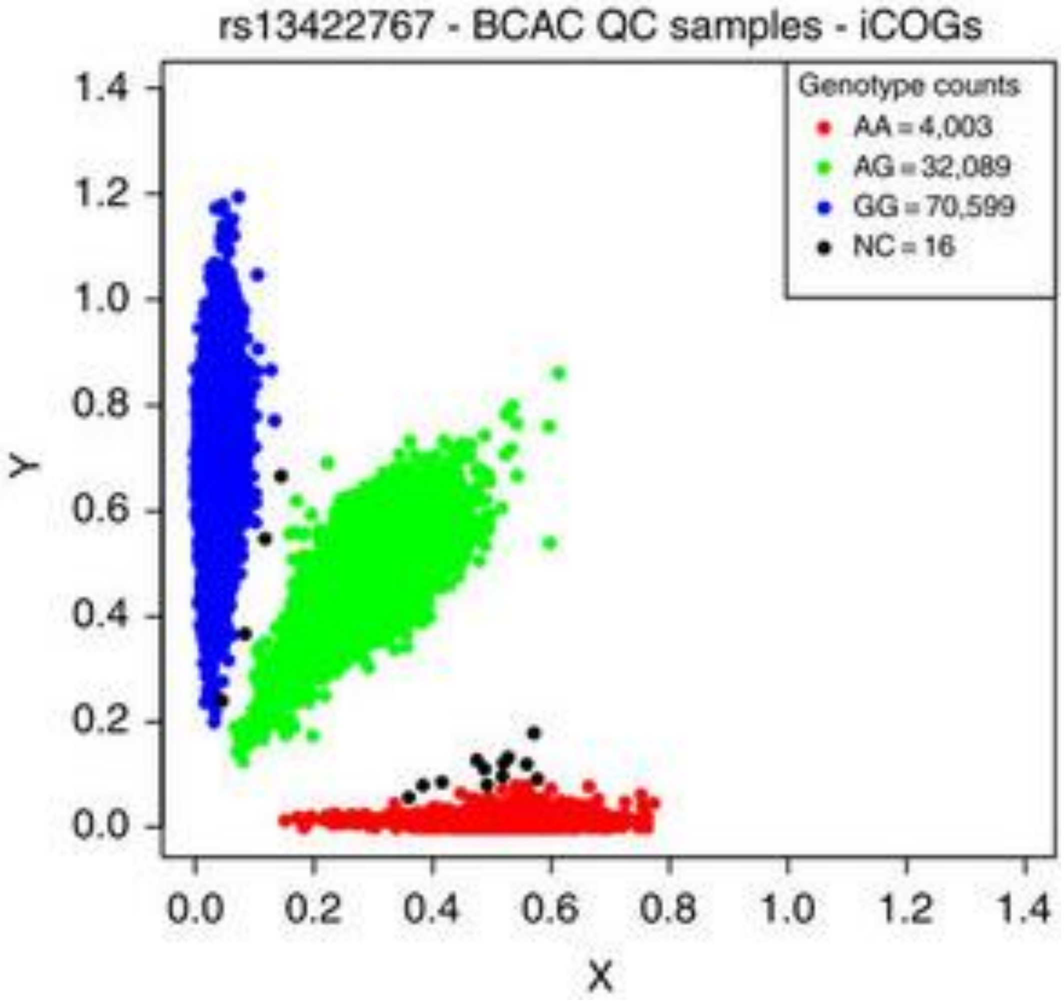
Sequencing:

- Rare variants
- Discover new variants in individuals or small samples (compare children and parents)
- very detailed data
- to add variants across the same gene in studying an effect.



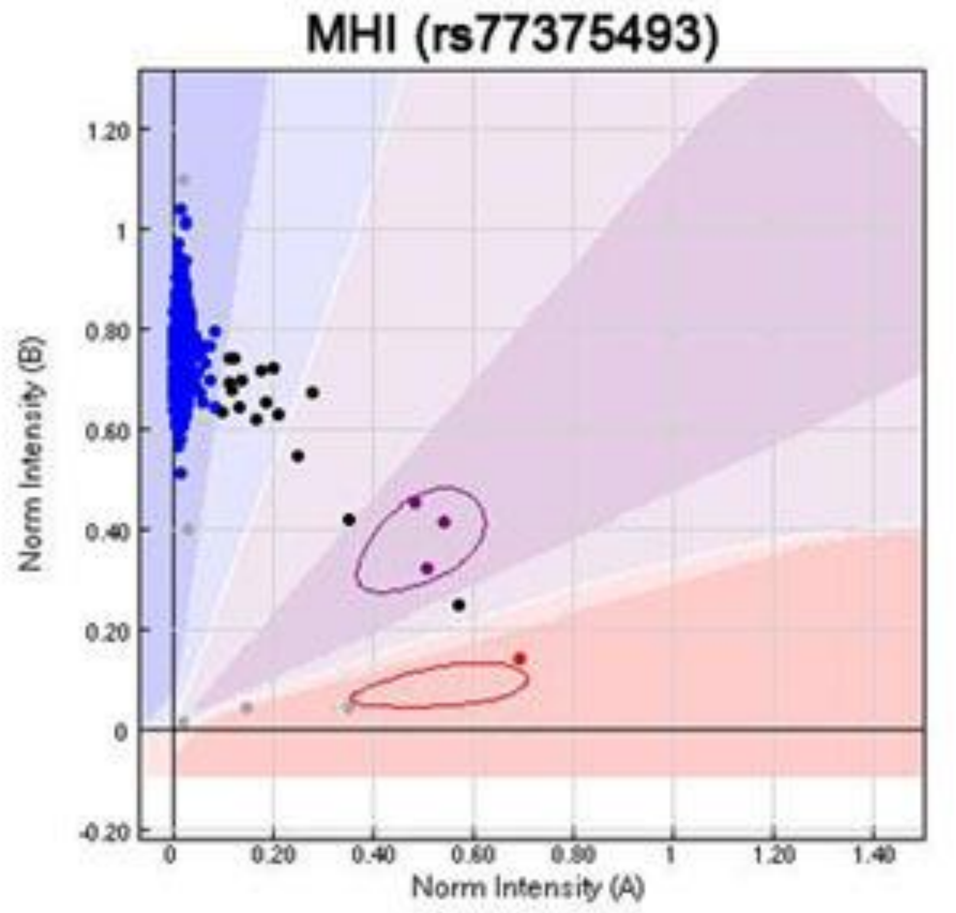
Genotyping Output

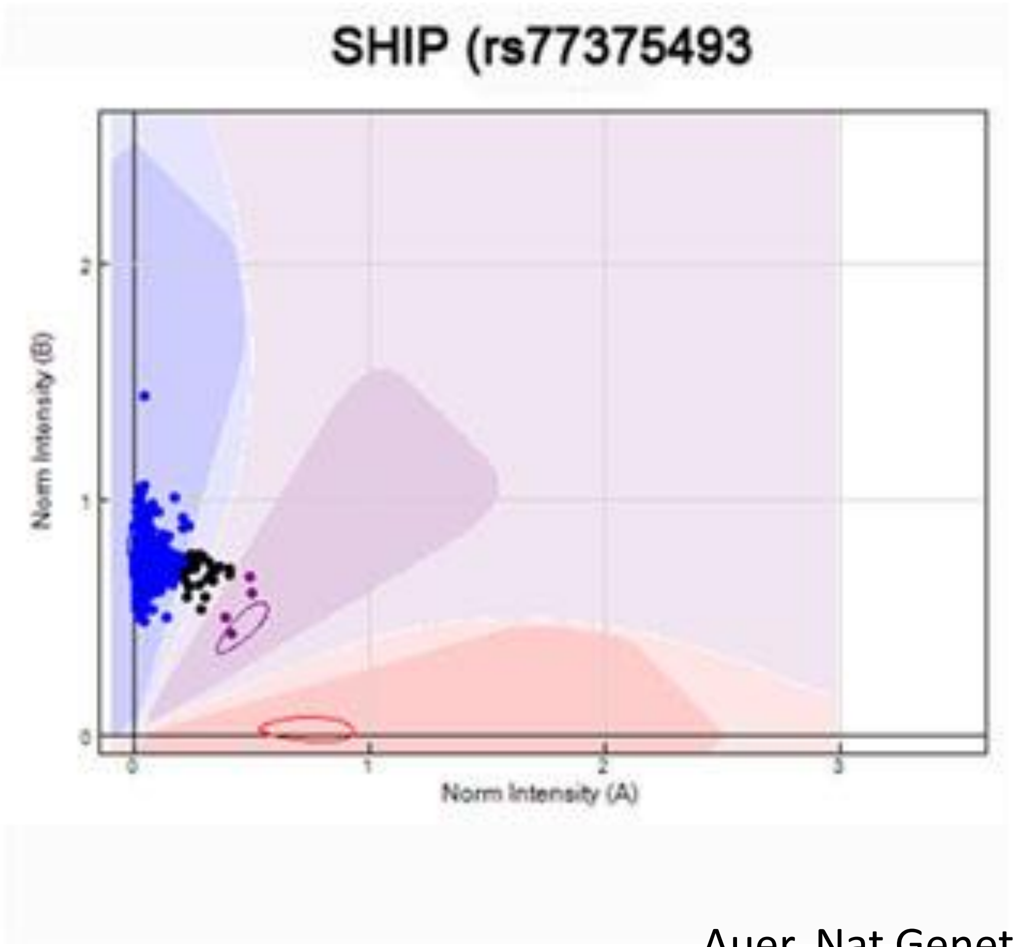




Li, Nat Comm 2014

Genotype cluster plot for rare variants





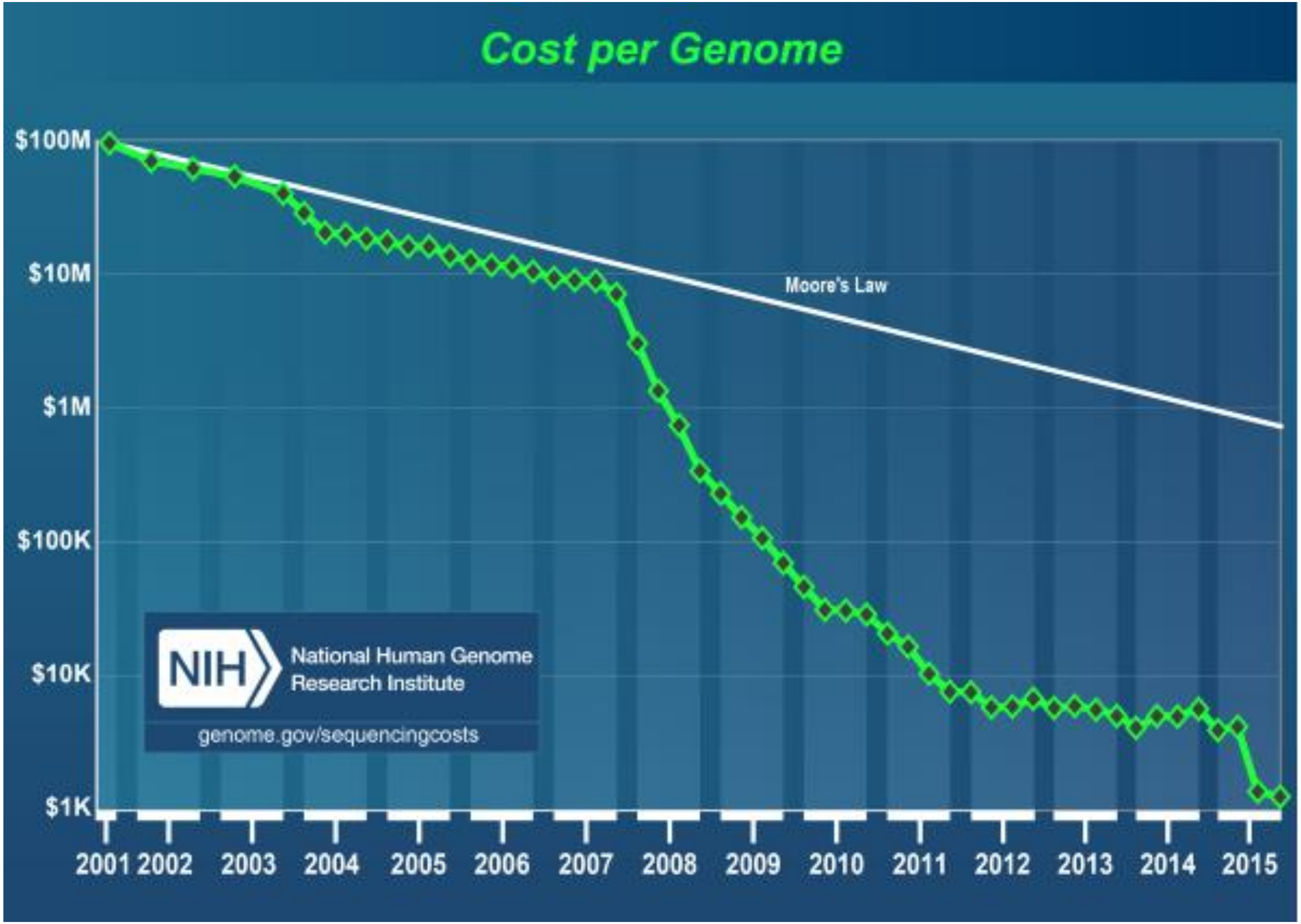
Auer, Nat Genet 2014

Sequencing to identify rare variants.

Same variant shared by individuals in a small group.

Multiple variants in the same gene in individuals with the same condition.

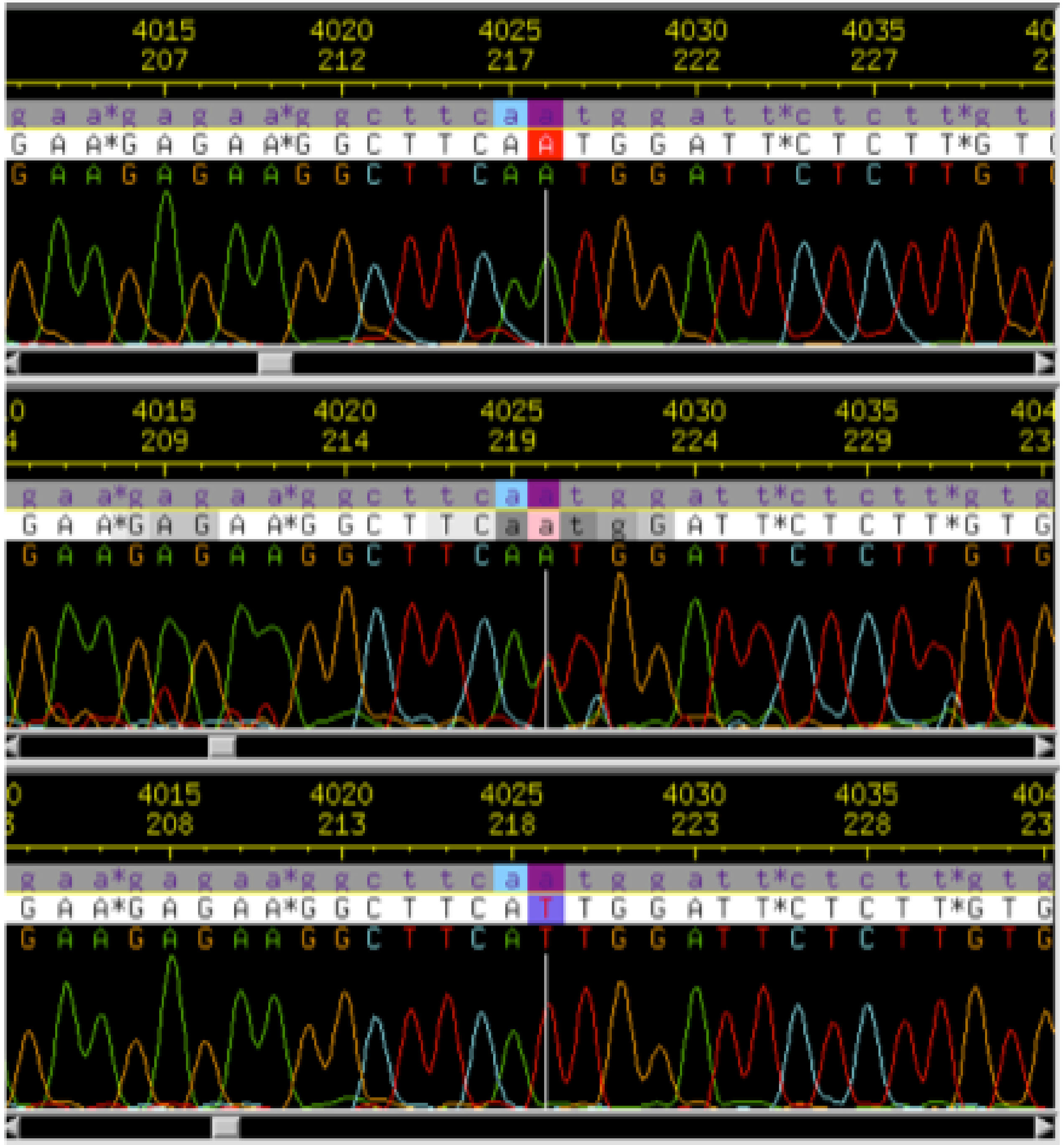
Variants unique to an individual in an important gene.


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts

Sequencing Technologies

- Sanger sequencing uses real time PCR
  - 99.99% accuracy
  - Used for high-accuracy reads of smaller regions
- Next Gen sequencing sequences many segments at once
  - Also called: massively parallel sequencing
  - High throughput
  - Used for multi-gene reads and larger samples

Sequencing output

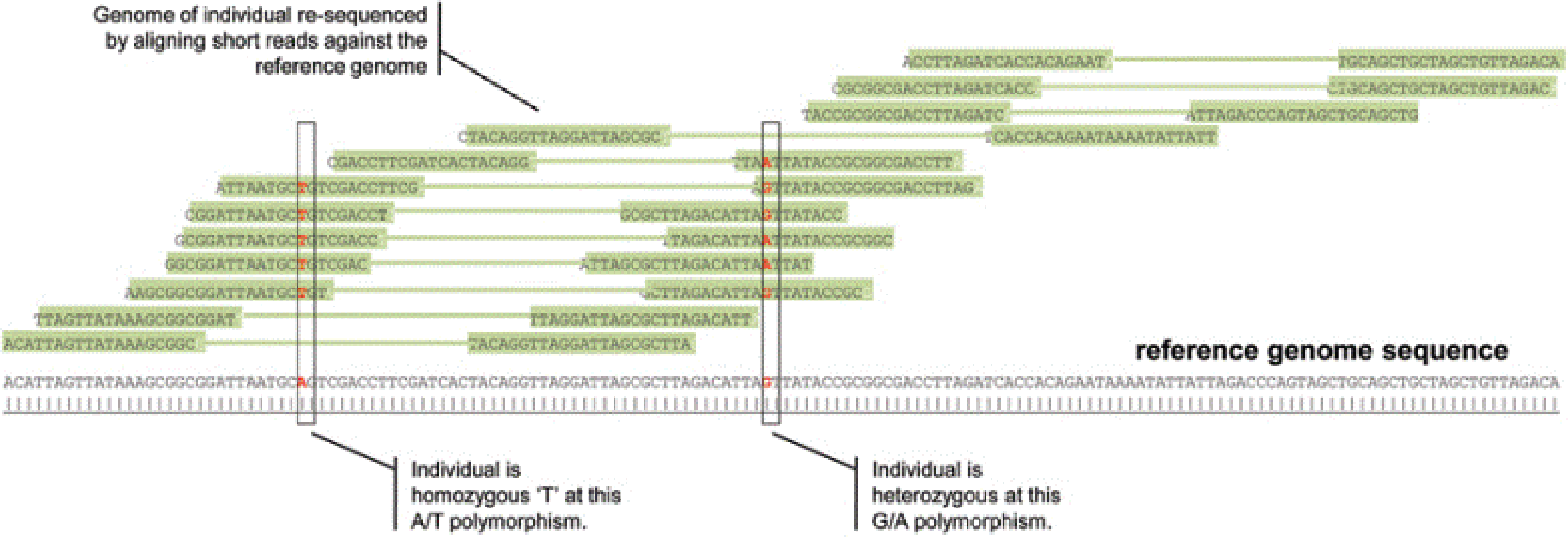


4015 207
4020 212
4025 217
4030 222
4035 227
G A A*G A G A A*G G C T T C A A T G G A T T*C T C T T*G T
4015 209
4020 214
4025 219
4030 224
4035 229
G A A*G A G A A*G G C T T C a a t g G A T T*C T C T T*G T G
4015 208
4020 213
4025 218
4030 223
4035 228
G A A*G A G A A*G G C T T C A T T G G A T T*C T C T T*G T G

Fohner 2015

Sequencing alignment and depth

Depth: The number of times one basepair is sequenced



Genome of individual re-sequenced by aligning short reads against the reference genome

reference genome sequence

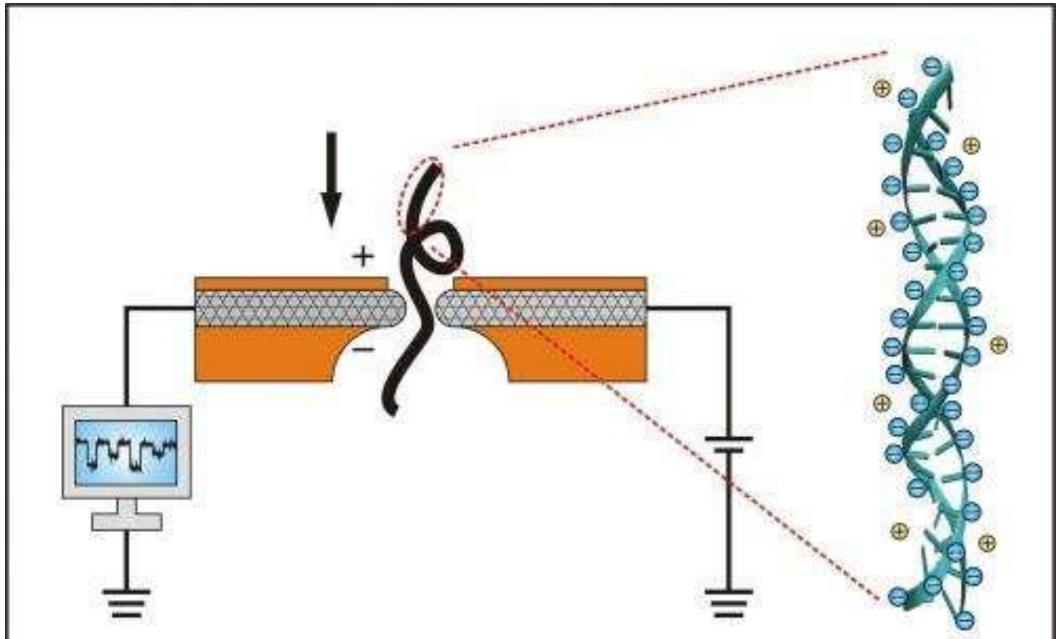
Individual is homozygous 'T' at this A/T polymorphism.

Individual is heterozygous at this G/A polymorphism.

Nanopore sequencing

Works on a single DNA strand. No PCR amplification; No chemical labeling.

Feeds DNA through a *very* tiny hole, sends an electric current, and determines the DNA base based on how the current flows.



Nanopore sequencing of DNA

Sequence Assembly

CTCGCGCGAT

ACCCTCG

GCGATAG

ACTTAATAC

ACCCTCGCGC

GCGATAGACTTA

Sequence Assembly

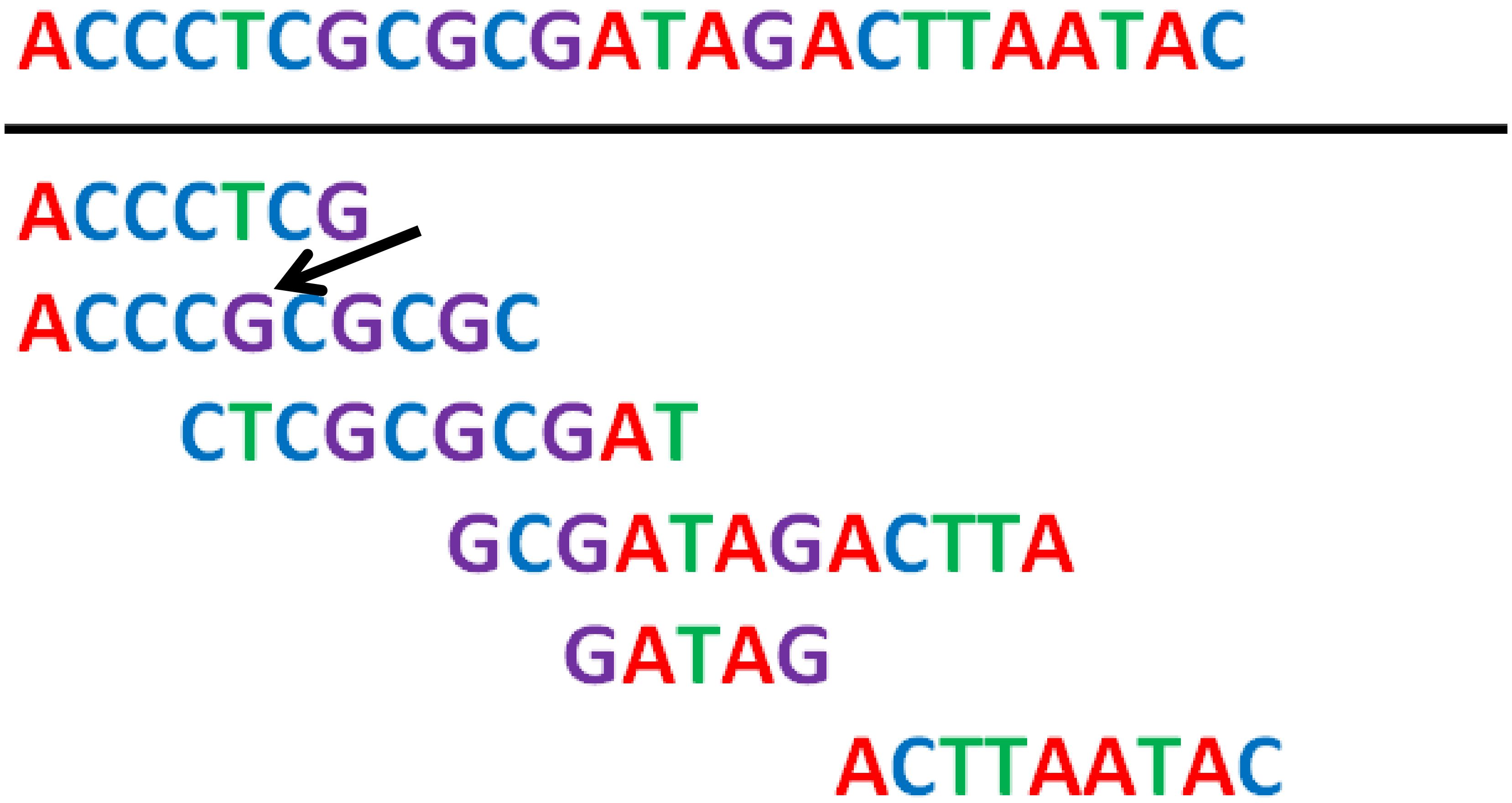
```
ACCCTCGCGCGATAGACTTAATAC

   CTCGCGCGAT
ACCCTCG
        GCGATAG
               ACTTAATAC
ACCCTCGCGC
        GCGATAGACTTA
```

Sequence Assembly: Other Considerations

- Assembly type: De novo or mapping sequence assembly
- Read length: usually 100-700bp
- Read depth: 30 is gold standard

Sequence Assembly: Why Read Depth Matters



ACCCTCGCGCGATAGACTTAATAC

ACCCTCG

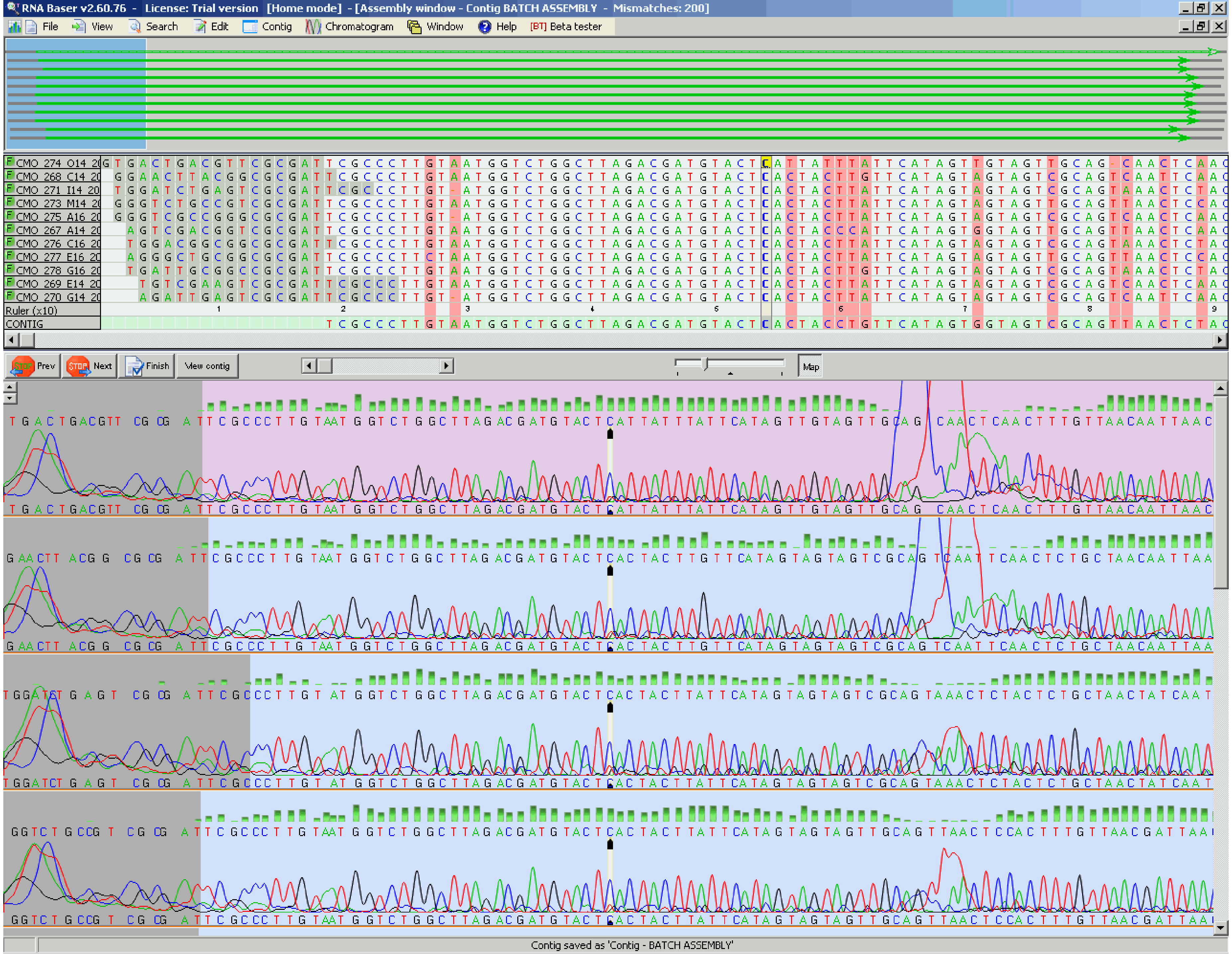
ACCCGCGCGC

CTCGCGCGAT

GCGATAGACTTA

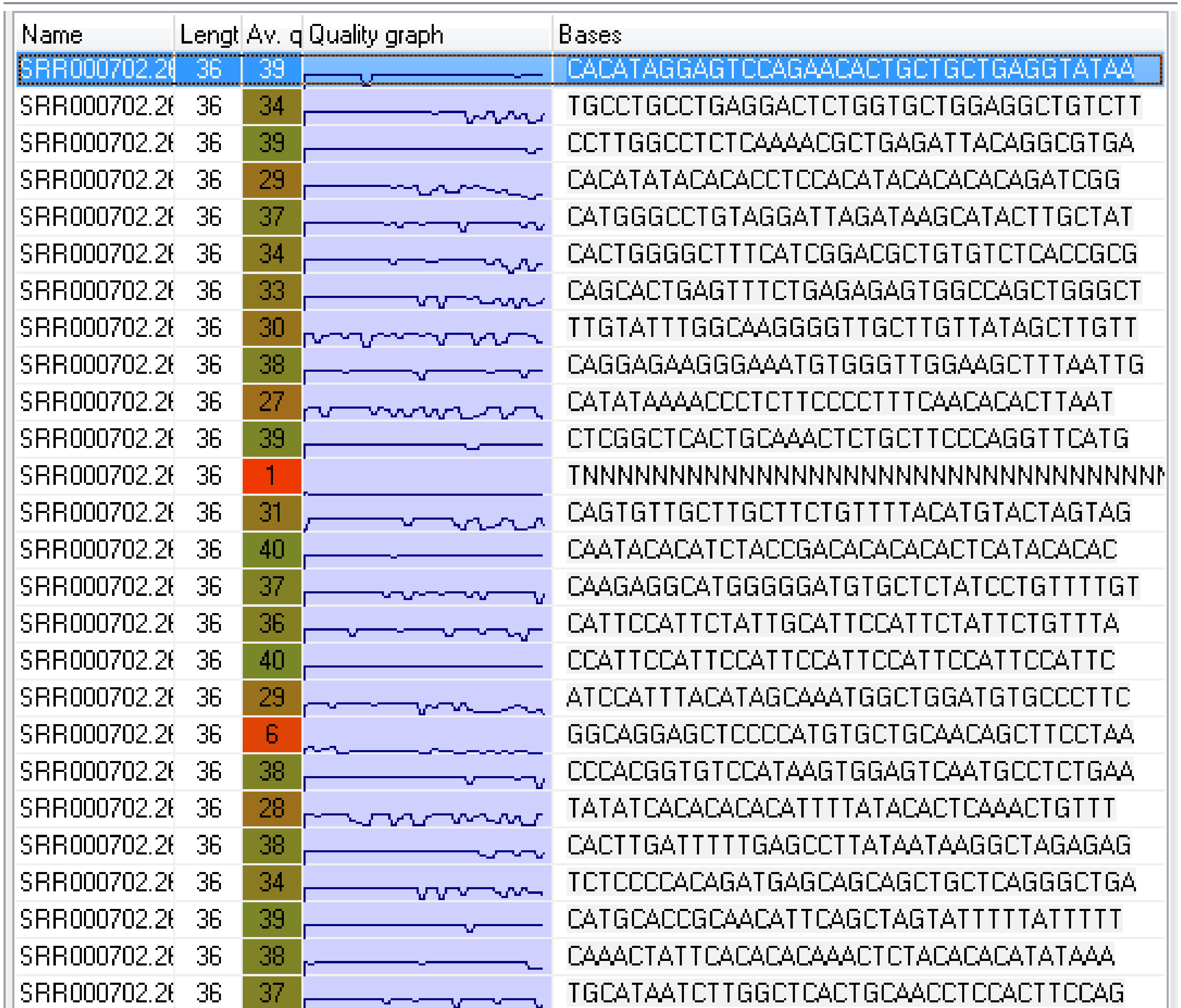
GATAG

ACTTAATAC


RNA Baser v2.60.76 - License: Trial version [Home mode] - [Assembly window - Contig BATCH ASSEMBLY - Mismatches: 200]
File
View
Search
Edit
Contig
Chromatogram
Window
Help
[BT] Beta tester
Prev
Next
Finish
View contig
Map

```
CMO 274 O14 20 G T G A C T G A C G T T C G C G A T T C G C C C T T G T A A T G G T C T G G C T T A G A C G A T G T A C T C A T T A T T T A T T C A T A G T T G T A G T T G C A G - C A A C T C A A C
CMO 268 C14 20   G G A A C T T A C G G C G C G A T T C G C C C T T G T A A T G G T C T G G C T T A G A C G A T G T A C T C A C T A C T T G T T C A T A G T A G T A G T C G C A G T C A A T T C A A C
CMO 271 I14 20   T G G A T C T G A G T C G C G A T T C G C C C T T G T - A T G G T C T G G C T T A G A C G A T G T A C T C A C T A C T T A T T C A T A G T A G T A G T C G C A G T A A A C T C T A C
CMO 273 M14 20   G G G T C T G C C G T C G C G A T T C G C C C T T G T A A T G G T C T G G C T T A G A C G A T G T A C T C A C T A C T T A T T C A T A G T A G T A G T T G C A G T T A A C T C C A C
CMO 275 A16 20   G G G T C G C C G G G C G C G A T T C G C C C T T G T - A T G G T C T G G C T T A G A C G A T G T A C T C A C T A C T T A T T C A T A G T A G T A G T C G C A G T C A A C T C A A C
CMO 267 A14 20     A G T C G A C G G T C G C G A T T C G C C C T T G T A A T G G T C T G G C T T A G A C G A T G T A C T C A C T A C C C A T T C A T A G T G G T A G T T G C A G T T A A C T C A A C
CMO 276 C16 20     T G G A C G G C G G C G C G A T T C G C C C T T G T A A T G G T C T G G C T T A G A C G A T G T A C T C A C T A C T T A T T C A T A G T A G T A G T C G C A G T A A A C T C T A C
CMO 277 E16 20     A G G G C T G C G G C G C G A T T C G C C C T T C T A A T G G T C T G G C T T A G A C G A T G T A C T C A C T A C T T A T T C A T A G T A G T A G T T G C A G T T A A C T C C A C
CMO 278 G16 20     T G A T T G C G G C C G C G A T T C G C C C T T G T A A T G G T C T G G C T T A G A C G A T G T A C T C A C T A C T T G T T C A T A G T A G T A G T C G C A G T A A A C T C T A C
CMO 269 E14 20       T G T C G A A G T C G C G A T T C G C C C T T G T A A T G G T C T G G C T T A G A C G A T G T A C T C A C T A C T T A T T C A T A G T A G T A G T C G C A G T C A A C T C A A C
CMO 270 G14 20       A G A T T G A G T C G C G A T T C G C C C T T G T - A T G G T C T G G C T T A G A C G A T G T A C T C A C T A C T T A T T C A T A G T A G T A G T C G C A G T C A A T T C A A C
Ruler (x10)                      1                   2                   3                   4                   5                   6                   7                   8                   9
CONTIG                               T C G C C C T T G T A A T G G T C T G G C T T A G A C G A T G T A C T C A C T A C C T G T T C A T A G T G G T A G T C G C A G T T A A C T C T A C
```

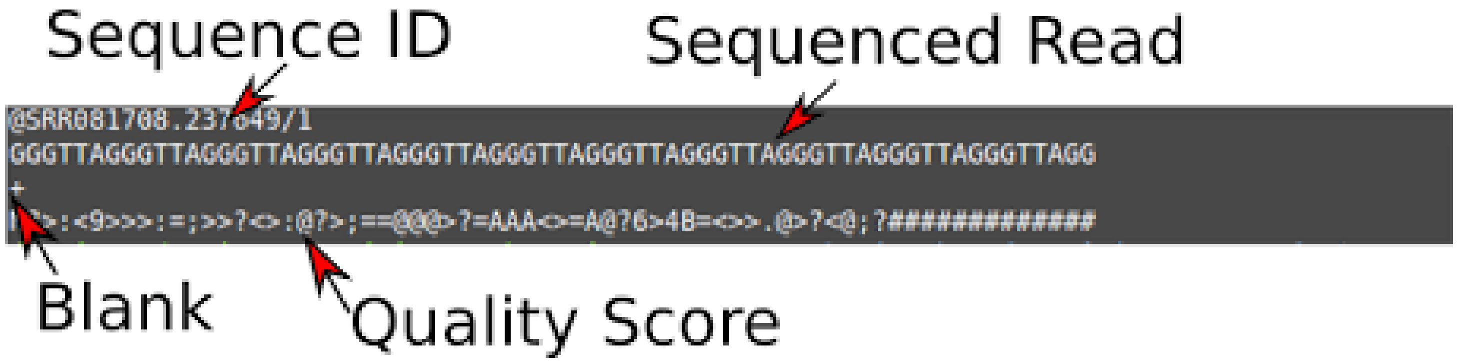
Contig saved as 'Contig - BATCH ASSEMBLY'



Name	Lengt	Av. q	Quality graph	Bases
SRR000702.2	36	39		CACATAGGAGTCCAGAACACTGCTGCTGAGGTATAA
SRR000702.2	36	34		TGCCTGCCTGAGGACTCTGGTGCTGGAGGCTGTCTT
SRR000702.2	36	39		CCTTGGCCTCTCAAAACGCTGAGATTACAGGCGTGA
SRR000702.2	36	29		CACATATACACACCTCCACATACACACACAGATCGG
SRR000702.2	36	37		CATGGGCCTGTAGGATTAGATAAGCATACTTGCTAT
SRR000702.2	36	34		CACTGGGGCTTTCATCGGACGCTGTGTCTCACCGCG
SRR000702.2	36	33		CAGCACTGAGTTTCTGAGAGAGTGGCCAGCTGGGCT
SRR000702.2	36	30		TTGTATTTGGCAAGGGGTTGCTTGTTATAGCTTGTT
SRR000702.2	36	38		CAGGAGAAGGGAAATGTGGGTTGGAAGCTTTAATTG
SRR000702.2	36	27		CATATAAAACCCTCTTCCCCTTTCAACACACTTAAT
SRR000702.2	36	39		CTCGGCTCACTGCAAACTCTGCTTCCCAGGTTCATG
SRR000702.2	36	1		TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
SRR000702.2	36	31		CAGTGTTGCTTGCTTCTGTTTTACATGTACTAGTAG
SRR000702.2	36	40		CAATACACATCTACCGACACACACACTCATACACAC
SRR000702.2	36	37		CAAGAGGCATGGGGGATGTGCTCTATCCTGTTTTGT
SRR000702.2	36	36		CATTCCATTCTATTGCATTCCATTCTATTCTGTTTA
SRR000702.2	36	40		CCATTCCATTCCATTCCATTCCATTCCATTCCATTC
SRR000702.2	36	29		ATCCATTTACATAGCAAATGGCTGGATGTGCCCTTC
SRR000702.2	36	6		GGCAGGAGCTCCCCATGTGCTGCAACAGCTTCCTAA
SRR000702.2	36	38		CCCACGGTGTCCATAAGTGGAGTCAATGCCTCTGAA
SRR000702.2	36	28		TATATCACACACACATTTTATACACTCAAACTGTTT
SRR000702.2	36	38		CACTTGATTTTTGAGCCTTATAATAAGGCTAGAGAG
SRR000702.2	36	34		TCTCCCCACAGATGAGCAGCAGCTGCTCAGGGCTGA
SRR000702.2	36	39		CATGCACCGCAACATTCAGCTAGTATTTTTATTTTT
SRR000702.2	36	38		CAAACTATTCACACACAAACTCTACACACATATAAA
SRR000702.2	36	37		TGCATAATCTTGGCTCACTGCAACCTCCACTTCCAG

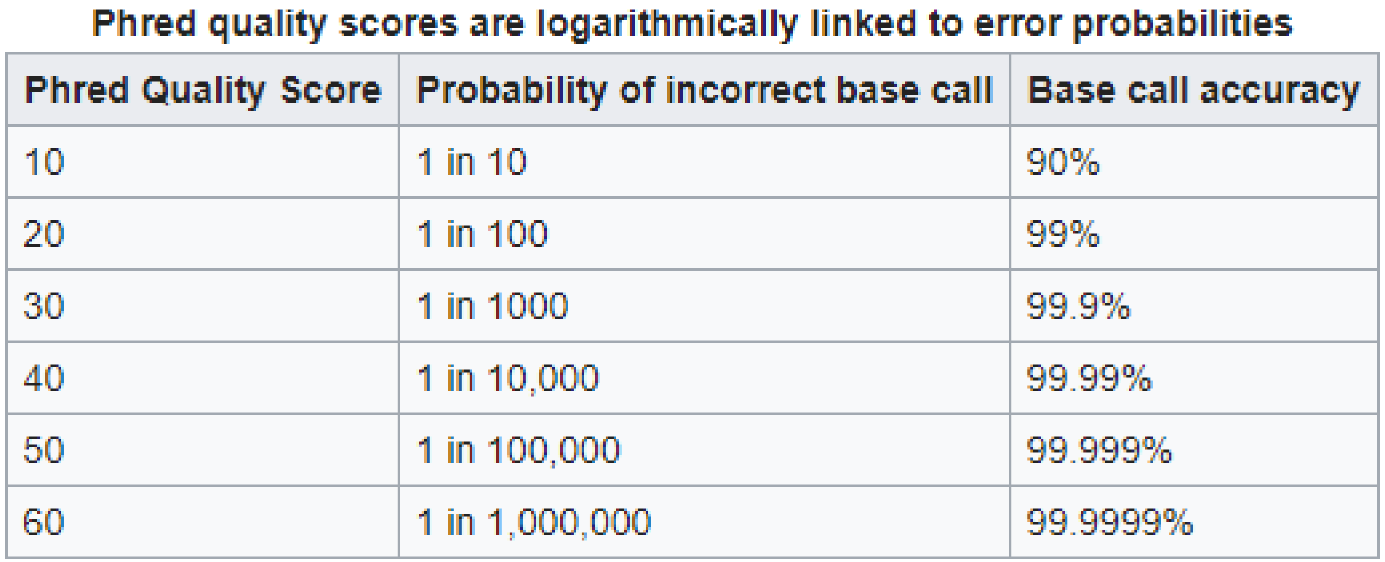
Phred Q scores: probability of incorrect call

p = 10^(-Q/10)



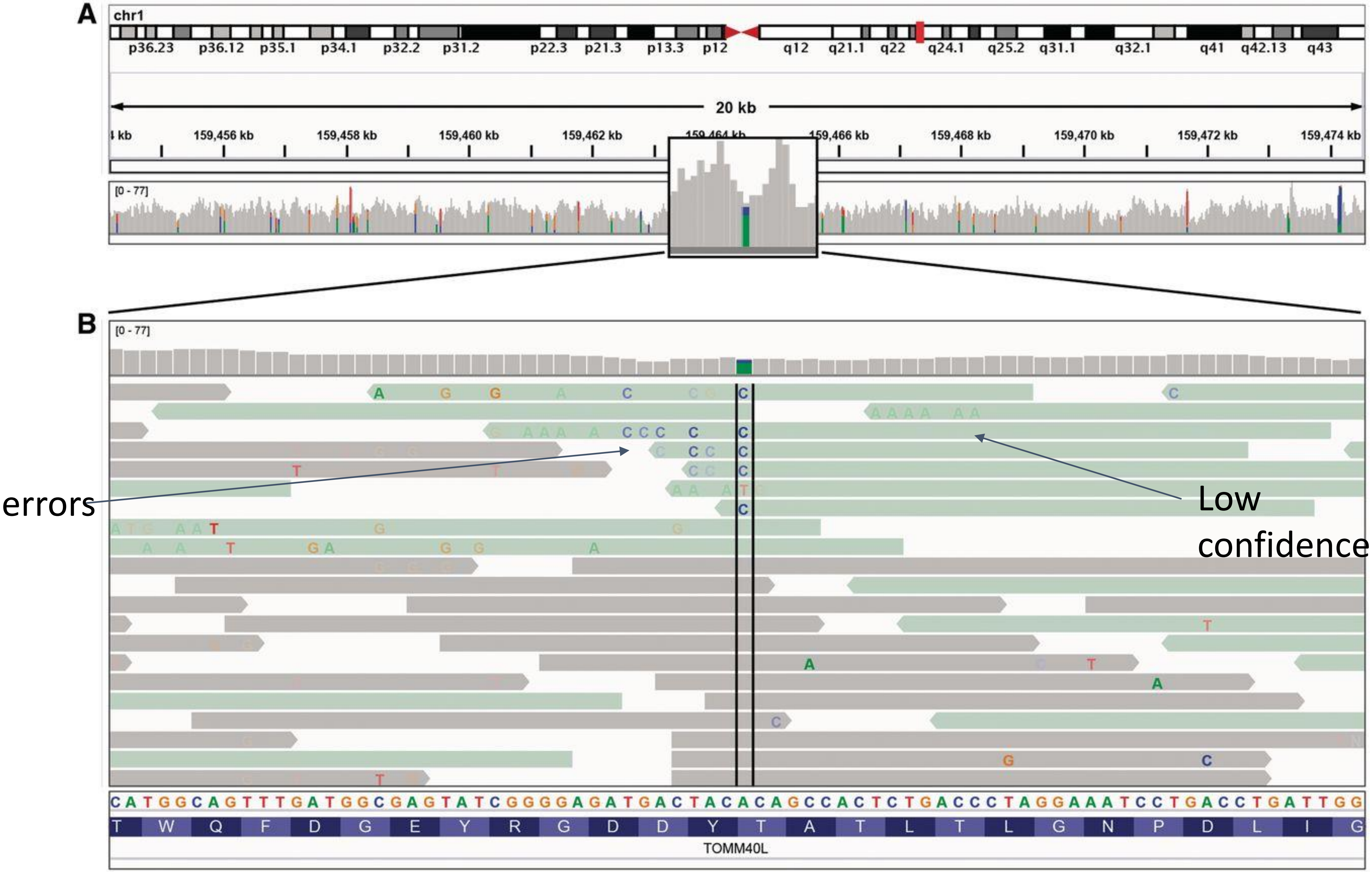
```
@SRR081708.237649/1
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
+
I?>:<9>>>>:=;>>?<>:@?>;==@@@>?=AAA<>=A@?6>4B=<>>.@>?<@;?##############
```

Higher the number, the better



Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%



errors

Low confidence

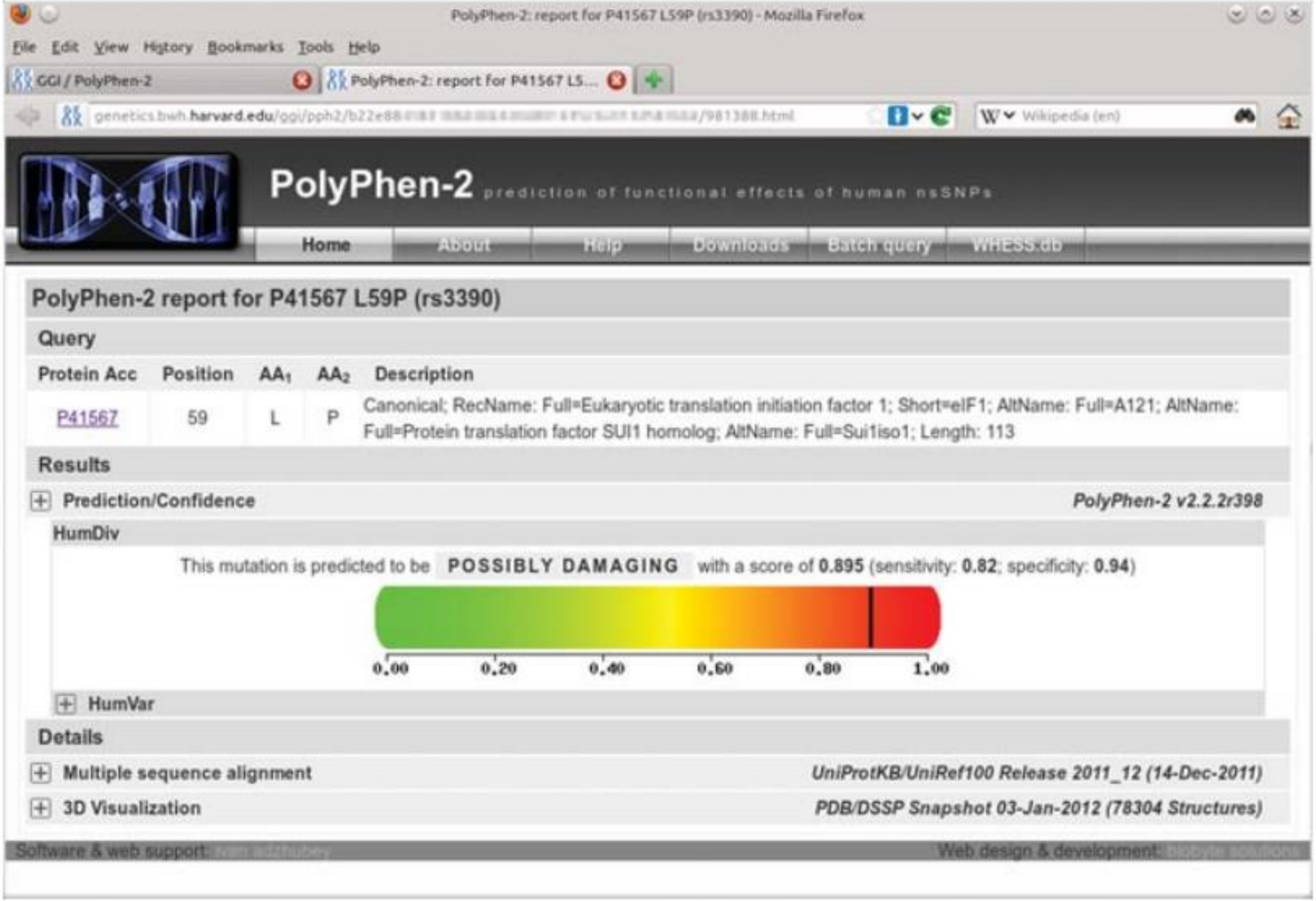
You have a new variant. Now what??

If they change the amino acid structure, algorithms to predict pathogenicity:

SIFT: Sorting Intolerant From Tolerant. Based on sequence homology across species and chemical properties of amino acids. If it is the same, it must be important. Closer to zero = worse.

Polyphen-2: Based on protein structure and function predictions. Such as where in the protein the change occurs. Closer to 1 = worse.

Can have very different answers! Prediction is hard.


PolyPhen-2: report for P41567 L59P (rs3390) - Mozilla Firefox
File Edit View History Bookmarks Tools Help
GGI / PolyPhen-2
PolyPhen-2: report for P41567 L5...
genetics.bwh.harvard.edu/ggi/pph2/b22e88[redacted]/981388.html
Wikipedia (en)
PolyPhen-2 prediction of functional effects of human nsSNPs
Home
About
Help
Downloads
Batch query
WHESS.db

PolyPhen-2 report for P41567 L59P (rs3390)

Query

Protein Acc	Position	AA_1	AA_2	Description
P41567	59	L	P	Canonical; RecName: Full=Eukaryotic translation initiation factor 1; Short=eIF1; AltName: Full=A121; AltName: Full=Protein translation factor SUI1 homolog; AltName: Full=Sui1iso1; Length: 113

Results

⊞ Prediction/Confidence *PolyPhen-2 v2.2.2r398*

HumDiv

This mutation is predicted to be **POSSIBLY DAMAGING** with a score of **0.895** (sensitivity: **0.82**; specificity: **0.94**)

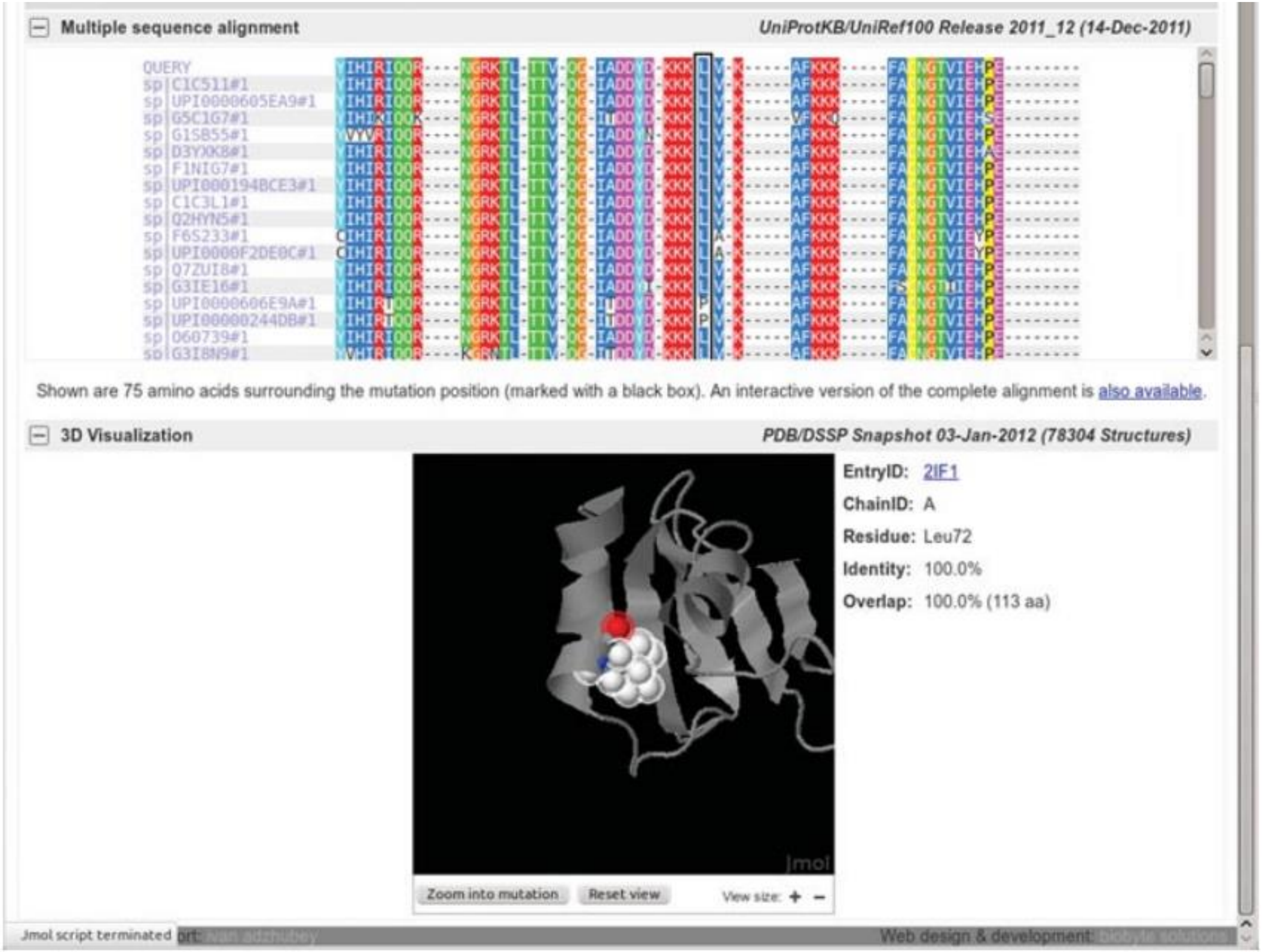
⊞ HumVar

Details

⊞ Multiple sequence alignment *UniProtKB/UniRef100 Release 2011_12 (14-Dec-2011)*

⊞ 3D Visualization *PDB/DSSP Snapshot 03-Jan-2012 (78304 Structures)*

Software & web support: Ivan Adzhubey Web design & development: biobyte solutions



Multiple sequence alignment

UniProtKB/UniRef100 Release 2011_12 (14-Dec-2011)

```
QUERY                 YIHIRIQQR----NGRKTL-ITTV-QG-IADDYD-KKKLV-K-----AFKKK-----FACNGTVIEHPE--------
sp|C1C511#1           YIHIRIQQR----NGRKTL-ITTV-QG-IADDYD-KKKLV-K-----AFKKK-----FACNGTVIEHPE--------
sp|UPI0000605EA9#1    YIHIRIQQR----NGRKTL-ITTV-QG-IADDYD-KKKLV-K-----AFKKK-----FACNGTVIEHPE--------
sp|G5C1G7#1           YIHIKIQQK----NGRKTL-ITTV-QG-ITDDYD-KKKLV-K-----VFKKQ-----FACNGTVIEHSE--------
sp|G1SB55#1           YVYVRIQQR----NGRKTL-ITTV-QG-IADDYN-KKKLV-K-----AFKKK-----FACNGTVIEHPE--------
sp|D3YXK8#1           YIHIRIQQR----NGRKTL-ITTV-QG-IADDYD-KKKLV-K-----AFKKK-----FACNGTVIEHPE--------
sp|F1NIG7#1           YIHIRIQQR----NGRKTL-ITTV-QG-IADDYD-KKKLV-K-----AFKKK-----FACNGTVIEHAE--------
sp|UPI000194BCE3#1    YIHIRIQQR----NGRKTL-ITTV-QG-IADDYD-KKKLV-K-----AFKKK-----FACNGTVIEHPE--------
sp|C1C3L1#1           YIHIRIQQR----NGRKTL-ITTV-QG-IADDYD-KKKLV-K-----AFKKK-----FACNGTVIEHPE--------
sp|Q2HYN5#1           YIHIRIQQR----NGRKTL-ITTV-QG-IADDYD-KKKLV-K-----AFKKK-----FACNGTVIEHPE--------
sp|F6S233#1           CIHIRIQQR----NGRKTL-ITTV-QG-IADDYD-KKKLA-K-----AFKKK-----FACNGTVIEYPE--------
sp|UPI0000F2DE0C#1    CIHIRIQQR----NGRKTL-ITTV-QG-IADDYD-KKKLA-K-----AFKKK-----FACNGTVIEYPE--------
sp|Q7ZUI8#1           YIHIRIQQR----NGRKTL-ITTV-QG-IADDYD-KKKLV-K-----AFKKK-----FACNGTVIEHPE--------
sp|G3IE16#1           YIHIRIQQR----NGRKTL-ITTV-QG-IADDYI-KKKLV-K-----AFKKK-----FSCNGTIIEHPE--------
sp|UPI0000606E9A#1    YIHIRVQQR----NGRKTL-ITTV-QG-IVDDYD-KKKPV-K-----AFKKK-----FACNGTVIEHPE--------
sp|UPI00000244DB#1    YIHIRTQQR----NGRKTL-ITTV-QG-ITDDYD-KKKPV-K-----AFKKK-----FACNGTVIEHPE--------
sp|O60739#1           YIHIRIQQR----NGRKTL-ITTV-QG-IADDYD-KKKLV-K-----AFKKK-----FACNGTVIEHPE--------
sp|G3I8N9#1           YVHIRIQQR----KGRNTL-ITTV-QG-ITDDYD-KKKLV-K-----AFKKK-----FACNGTVIEHPE--------
```

Shown are 75 amino acids surrounding the mutation position (marked with a black box). An interactive version of the complete alignment is also available.

3D Visualization

PDB/DSSP Snapshot 03-Jan-2012 (78304 Structures)

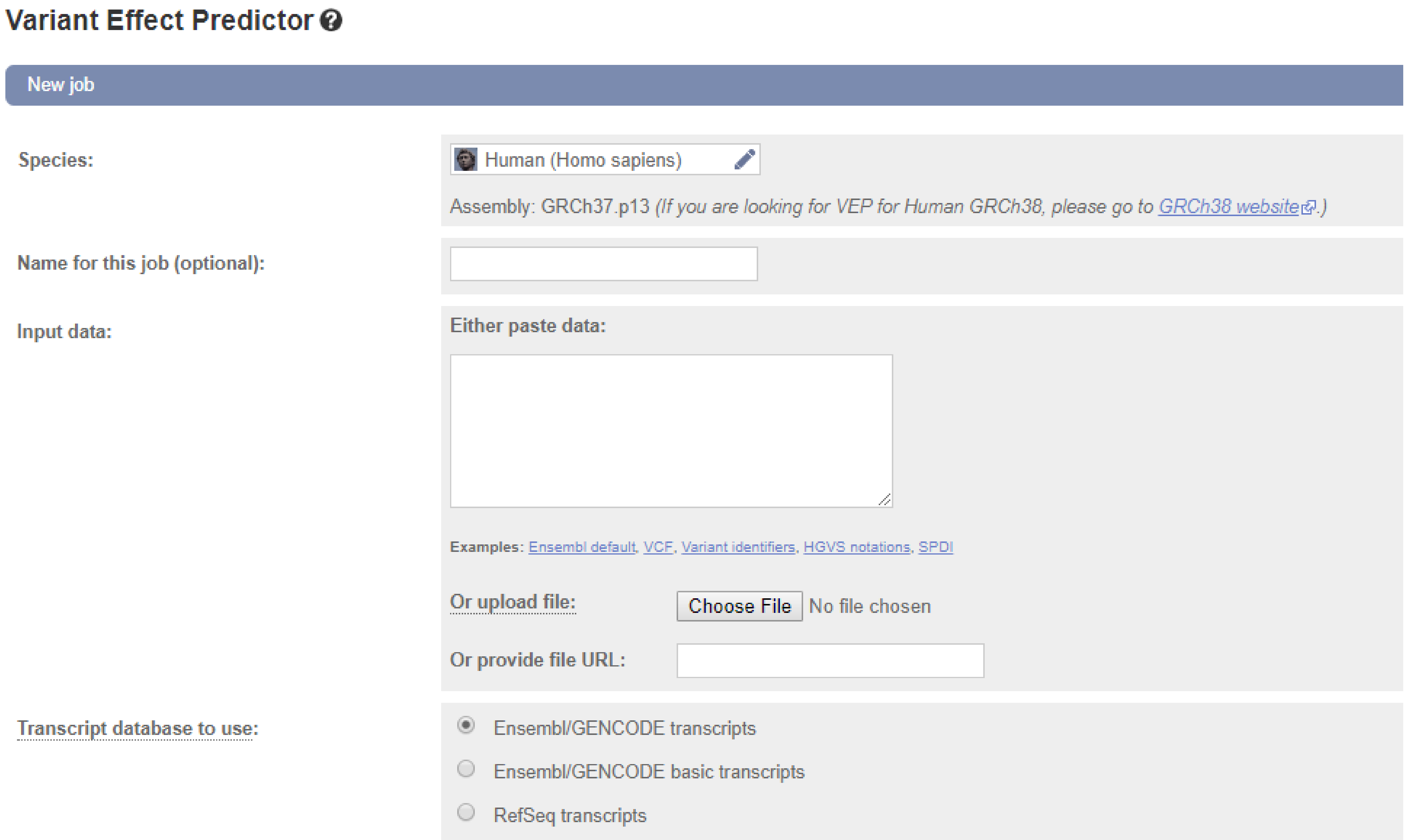
EntryID: 2IF1

ChainID: A

Residue: Leu72

Identity: 100.0%

Overlap: 100.0% (113 aa)



Variant Effect Predictor

New job

Species: Human (Homo sapiens)

Assembly: GRCh37.p13 *(If you are looking for VEP for Human GRCh38, please go to GRCh38 website.)*

Name for this job (optional):

Input data:

Either paste data:

Examples: Ensembl default, VCF, Variant identifiers, HGVS notations, SPDI

Or upload file: Choose File No file chosen

Or provide file URL:

Transcript database to use:

● Ensembl/GENCODE transcripts

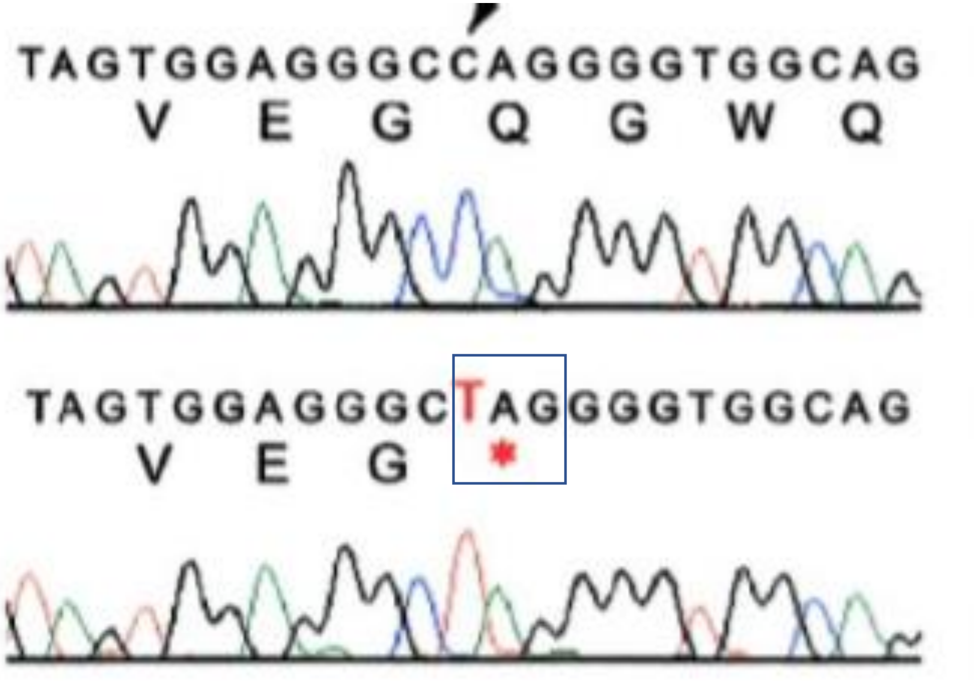
○ Ensembl/GENCODE basic transcripts

○ RefSeq transcripts

Sequencing in Founder Populations

- Osteoporosis is a disease in elderly people, resulting in decreased bone density
  - Many treatments may be carcinogenic
- A treatment for osteoporosis discovered by identifying SOST gene implicated in sclerosteosis in Dutch Afrikaner population
  - Autosomal recessive disorder

All but one patient with sclerosteosis in the 22-family sample shared the same SNP

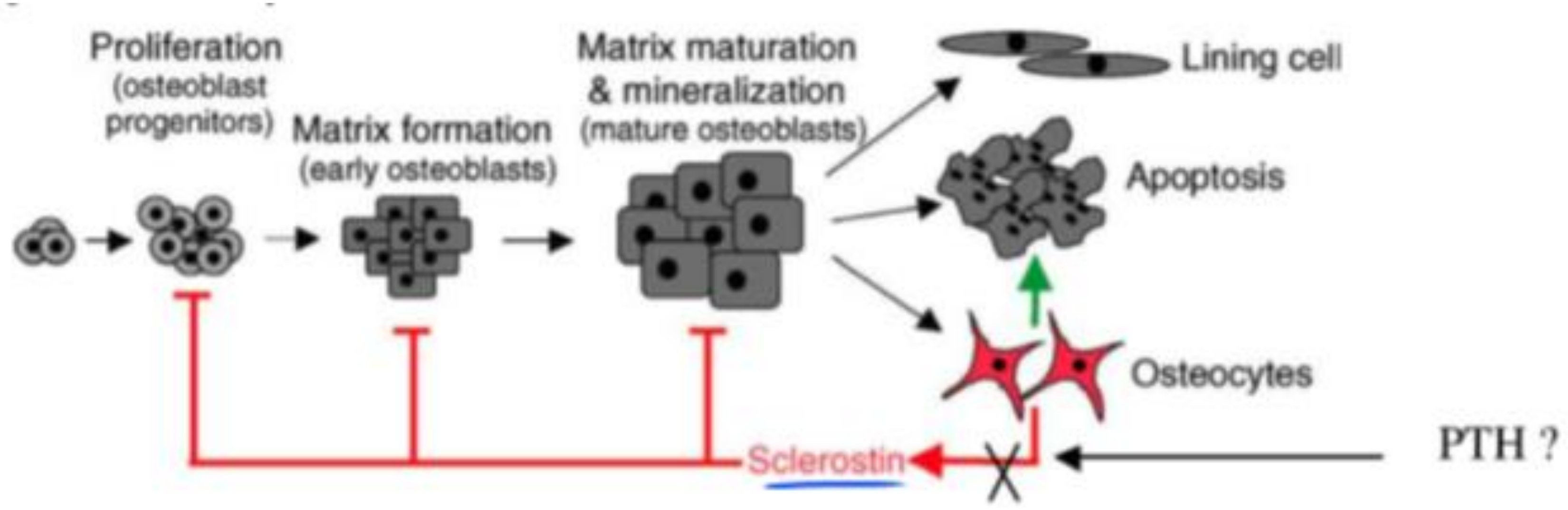


(C69T)

Results in a premature stop codon

-> Drug Development

- A drug was developed to deplete or inhibit sclerostin in people with osteoporosis



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.