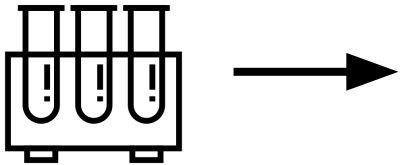
Variant calling, phasing, and imputation

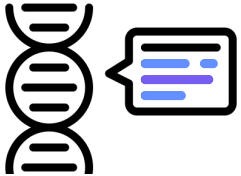
Nik Baya
Oxford Statistical Genomics Summer School
June 19 2023

Overview

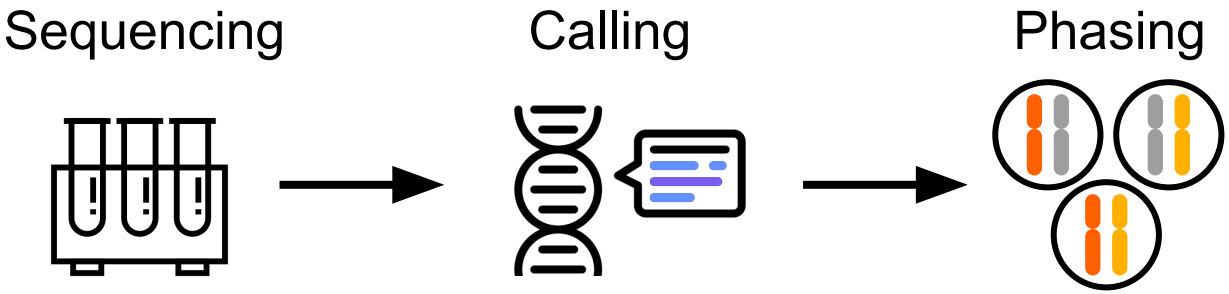
Sequencing



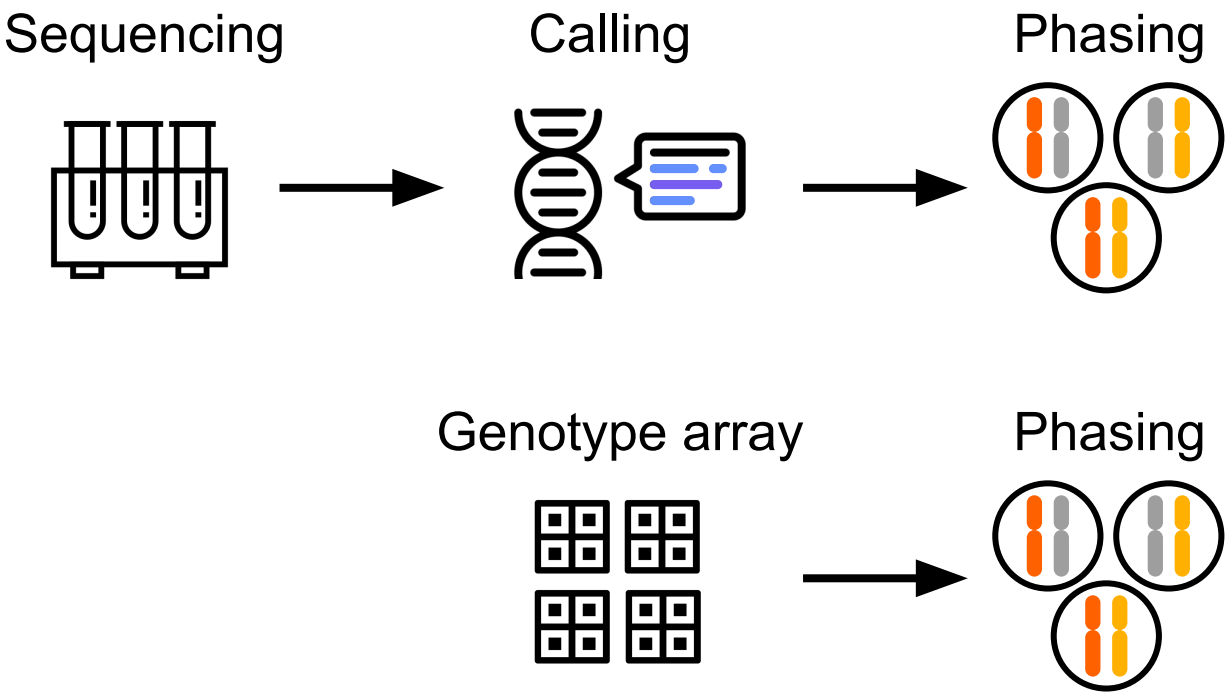
Calling



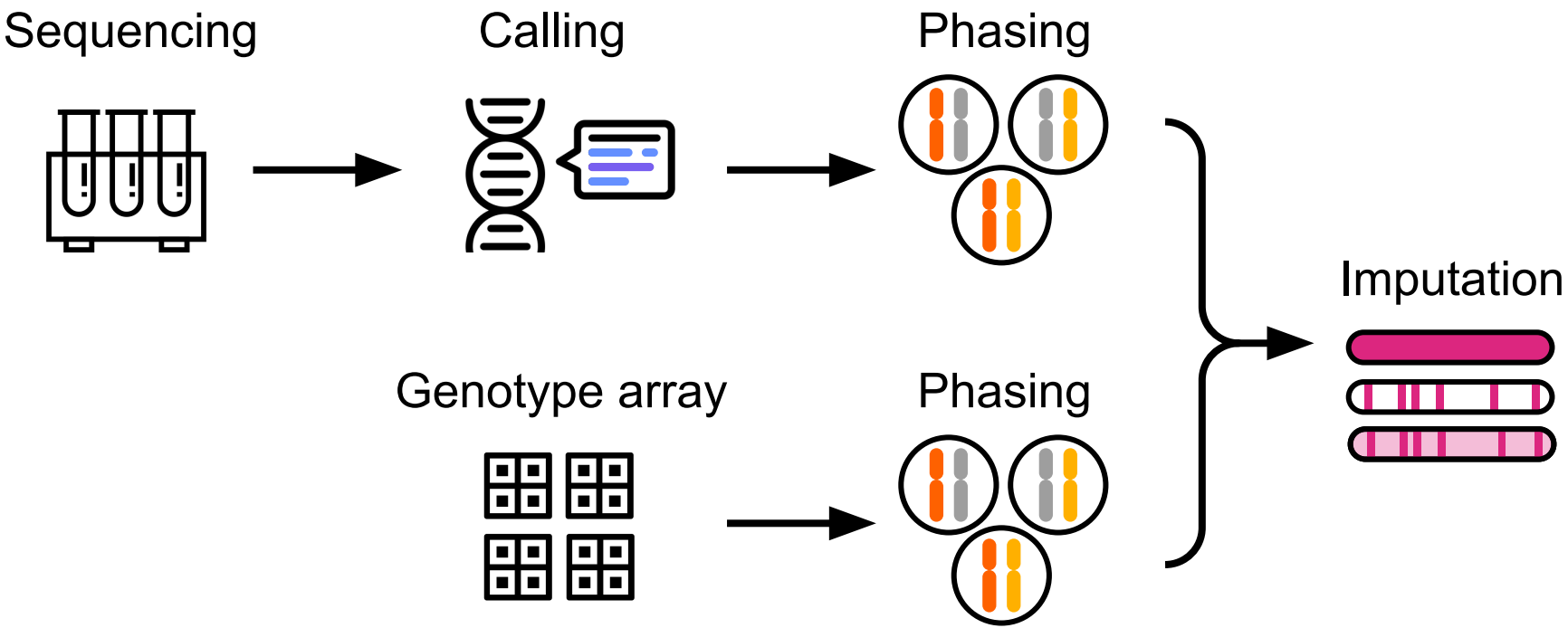
Overview



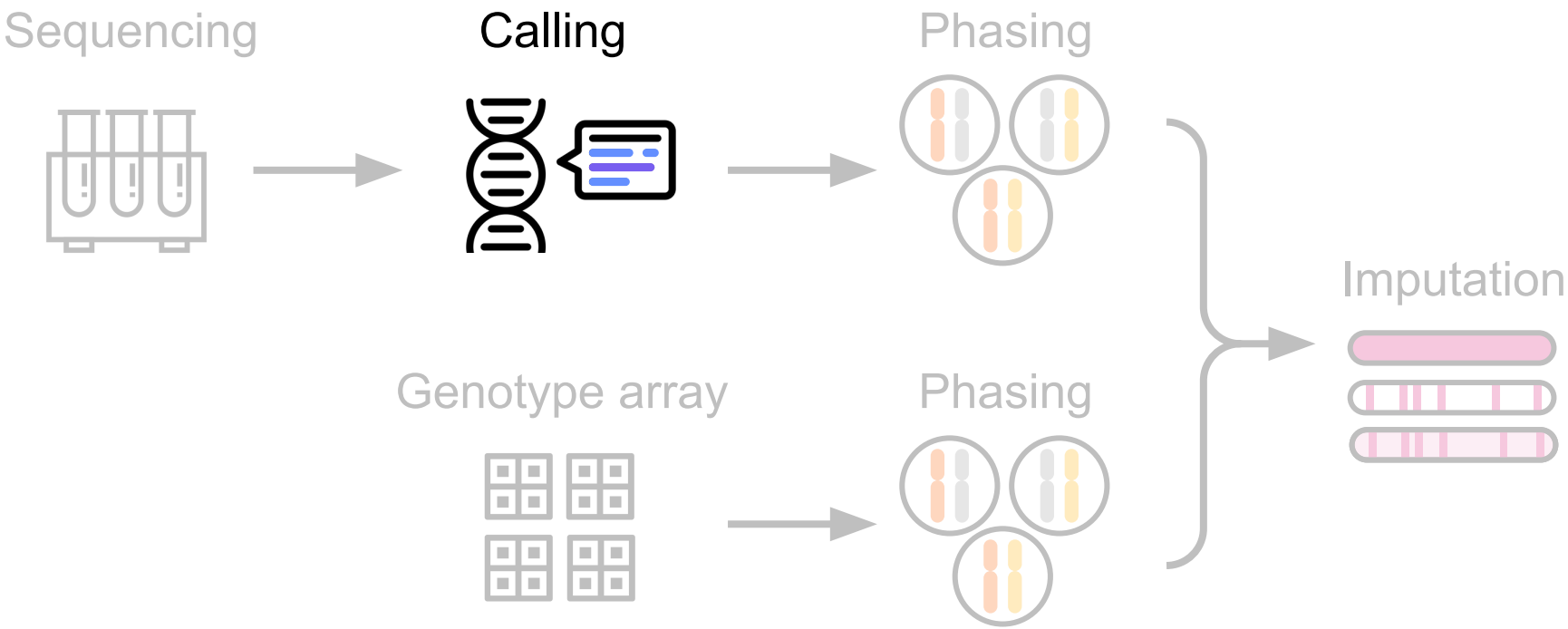
Overview

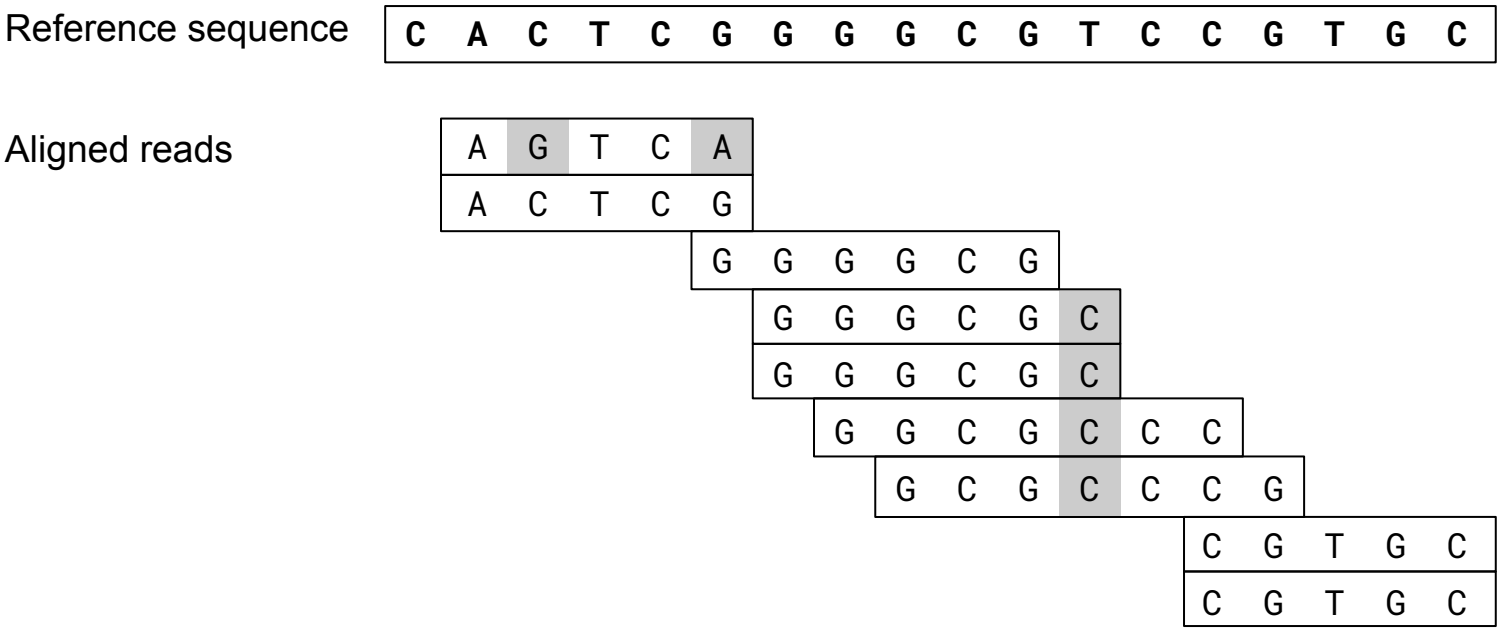


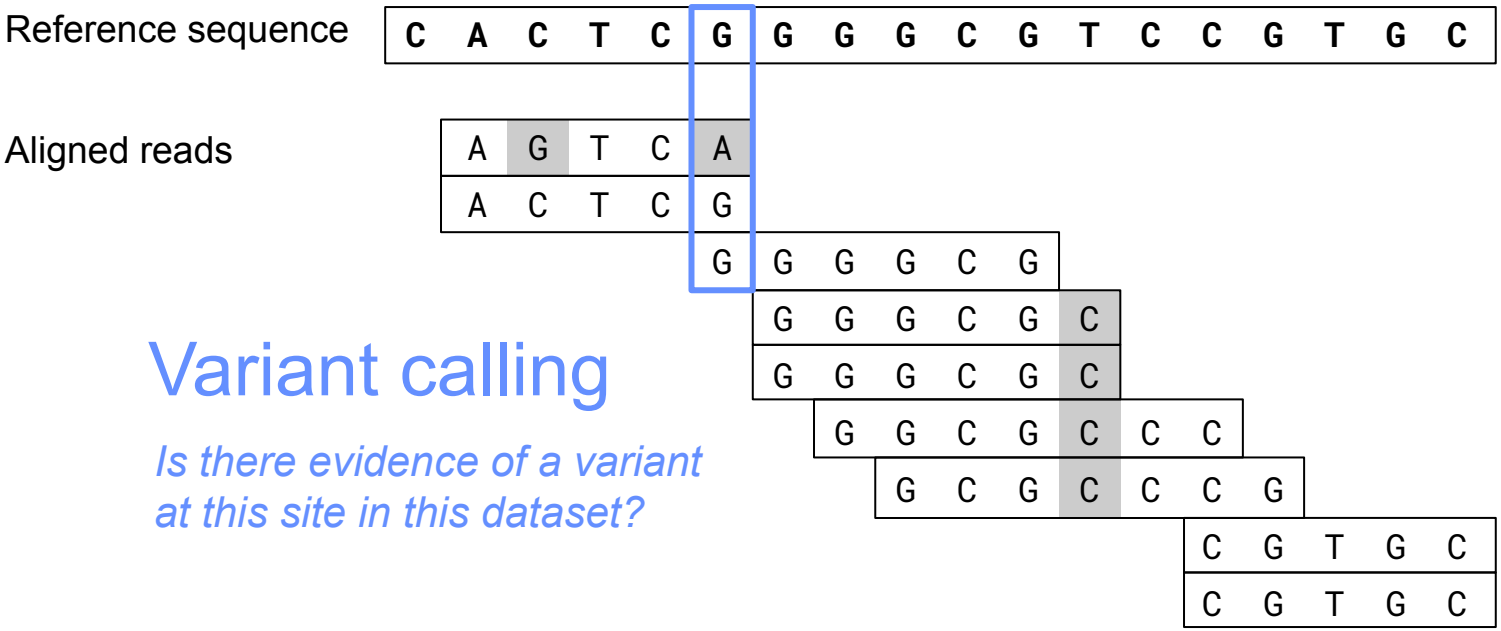
Overview


Sequencing
Calling
Phasing
Imputation
Genotype array
Phasing

Overview


Sequencing
Calling
Phasing
Imputation
Genotype array
Phasing


Reference sequence
C A C T C G G G G C G T C C G T G C
Aligned reads
A G T C A
A C T C G
G G G G C G
G G G C G C
G G G C G C
G G C G C C C
G C G C C C G
C G T G C
C G T G C



Variant calling

Is there evidence of a variant at this site in this dataset?

Variant calling

$$Pr(variable\ site|Data)$$

Probability that this site is variable, given the data

Slide adapted from Luke Jostins-Dean

Variant calling

Bayes' Theorem

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

$$Pr(variable\ site|Data)$$

Probability that this site is variable, given the data

Slide adapted from Luke Jostins-Dean

Variant calling

Bayes' Theorem

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

$$Pr(variable\ site|Data) = \frac{Pr(Data|variable\ site)Pr(variable\ site)}{Pr(Data)}$$

Probability that this site is variable, given the data

Slide adapted from Luke Jostins-Dean

Variant calling

$$Pr(variable\ site|Data) = \frac{Pr(Data|variable\ site)Pr(variable\ site)}{Pr(Data)}$$

Probability that this site is variable, given the data

$$= \frac{prior \prod_i L_x^i}{(1 - prior) \prod_i L_{ref}^i + prior \prod_i L_x^i}$$

Slide adapted from Luke Jostins-Dean

Variant calling

$$Pr(variable\ site|Data) = \frac{Pr(Data|variable\ site)Pr(variable\ site)}{Pr(Data)}$$

Probability that this site is variable, given the data

$$= \frac{prior \prod_i L_x^i}{(1 - prior) \prod_i L_{ref}^i + prior \prod_i L_x^i}$$

"is variable"

"is not variable"

Slide adapted from Luke Jostins-Dean

Variant calling

$$Pr(variable\ site|Data) = \frac{Pr(Data|variable\ site)Pr(variable\ site)}{Pr(Data)}$$

Probability that this site is variable, given the data

$$= \frac{prior \prod_i L_x^i}{(1 - prior) \prod_i L_{ref}^i + prior \prod_i L_x^i}$$

Prior probability that a site is variable (i.e. site is a variant)

Slide adapted from Luke Jostins-Dean

Prior probability of site being a variant

$$prior(N) = 4N_e\mu \sum_{k=1}^{2N-1} 1/k$$

N : Size of sequenced population

N_e : Effective population size

μ : Mutation rate

Slide adapted from Luke Jostins-Dean

Prior probability of site being a variant

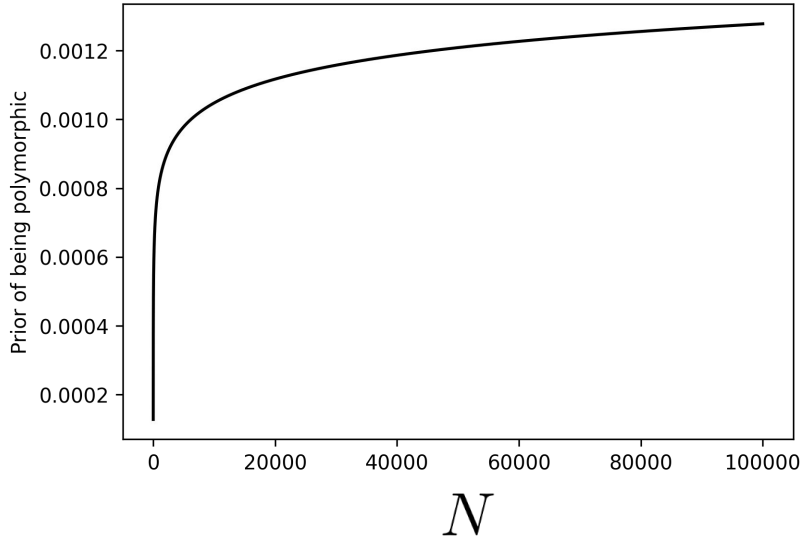
$$prior(N) = 4N_e\mu \sum_{k=1}^{2N-1} 1/k$$

N : Size of sequenced population

N_e : Effective population size

μ : Mutation rate

$$\approx 0.0001 * (1.270147 + \log(N))$$



Slide adapted from Luke Jostins-Dean

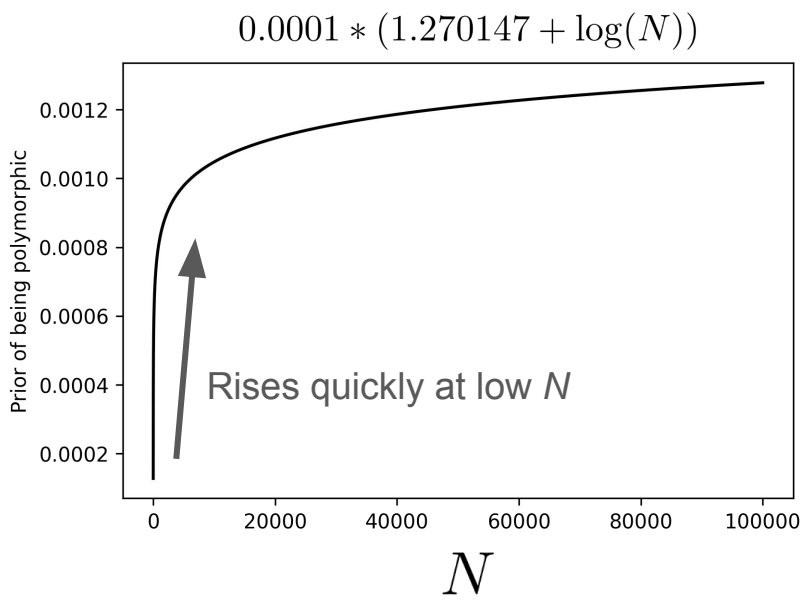
Prior probability of site being a variant

$$prior(N) = 4N_e\mu \sum_{k=1}^{2N-1} 1/k$$

N : Size of sequenced population

N_e : Effective population size

μ : Mutation rate



Slide adapted from Luke Jostins-Dean

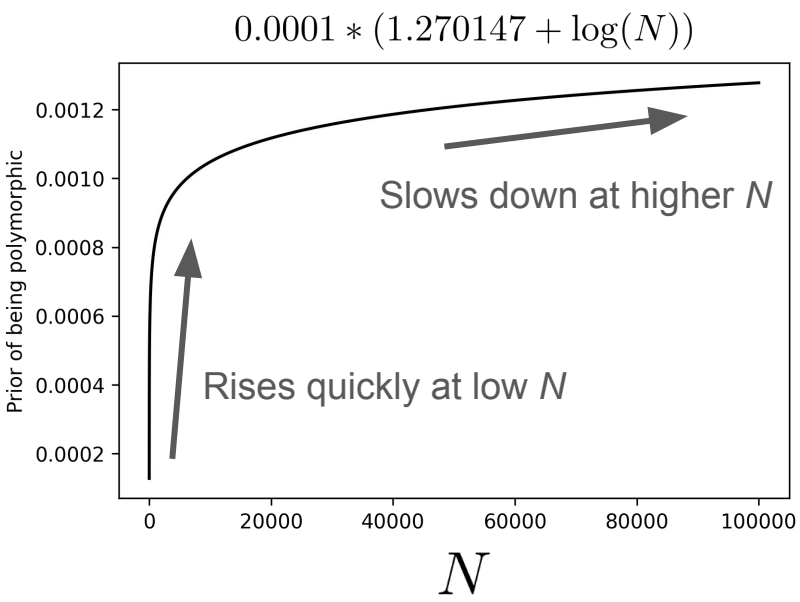
Prior probability of site being a variant

$$prior(N) = 4N_e\mu \sum_{k=1}^{2N-1} 1/k$$

N : Size of sequenced population

N_e : Effective population size

μ : Mutation rate



Slide adapted from Luke Jostins-Dean

Variant calling

$$Pr(variable\ site|Data) = \frac{Pr(Data|variable\ site)Pr(variable\ site)}{Pr(Data)}$$

Probability that this site is variable, given the data

$$= \frac{prior \prod_i L_x^i}{(1 - prior) \prod_i L_{ref}^i + prior \prod_i L_x^i}$$

Slide adapted from Luke Jostins-Dean

Variant calling

$$Pr(variable\ site|Data) = \frac{Pr(Data|variable\ site)Pr(variable\ site)}{Pr(Data)}$$

Probability that this site is variable, given the data

$$= \frac{prior \prod_i L_x^i}{(1 - prior) \prod_i L_{ref}^i + prior \prod_i L_x^i}$$

Likelihood being variable

Likelihood of being invariant

Slide adapted from Luke Jostins-Dean

Allele likelihood

Likelihood of site being **variable**:

$$\mathcal{L}_x^i = 2 * [G_i(g=1)f_x(1-f_x)] + G_i(g=2)f_x^2$$

Likelihood of being **invariant**:

$$\mathcal{L}_{ref}^i = G_i(g=0)(1-f_x)^2$$

Slide adapted from Luke Jostins-Dean

Allele likelihood

f_x : Frequency of allele x

Likelihood of site being **variable**:

$$\mathcal{L}_x^i = 2 * [G_i(g=1) f_x (1 - f_x)] + G_i(g=2) f_x^2$$

Likelihood of being **invariant**:

$$\mathcal{L}_{ref}^i = G_i(g=0)(1 - f_x)^2$$

Slide adapted from Luke Jostins-Dean

Allele likelihood

Likelihood of site being **variable**:

$$\mathcal{L}_x^i = 2 * [G_i(g=1) f_x(1-f_x)] + G_i(g=2) f_x^2$$

Likelihood of being **invariant**:

$$\mathcal{L}_{ref}^i = G_i(g=0)(1-f_x)^2$$

Genotype likelihoods

Slide adapted from Luke Jostins-Dean

Genotype likelihood

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} \Big[(m-g)\epsilon_j + g(1-\epsilon_j) \Big] \prod_{j=l+1}^{k} \Big[(m-g)(1-\epsilon_j) + g\epsilon_j) \Big]$$

g : genotype expressed as number of ref alleles

Li 2011

Genotype likelihood

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} \left[(m-g)\epsilon_j + g(1-\epsilon_j) \right] \prod_{j=l+1}^{k} \left[(m-g)(1-\epsilon_j) + g\epsilon_j \right]$$

E[# of true reference reads]

E[# of true alternate reads]

g : genotype expressed as number of ref alleles

Li 2011

Genotype likelihood

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} \Big[(m-g)\epsilon_j + g(1-\epsilon_j) \Big] \prod_{j=l+1}^{k} \Big[(m-g)(1-\epsilon_j) + g\epsilon_j) \Big]$$

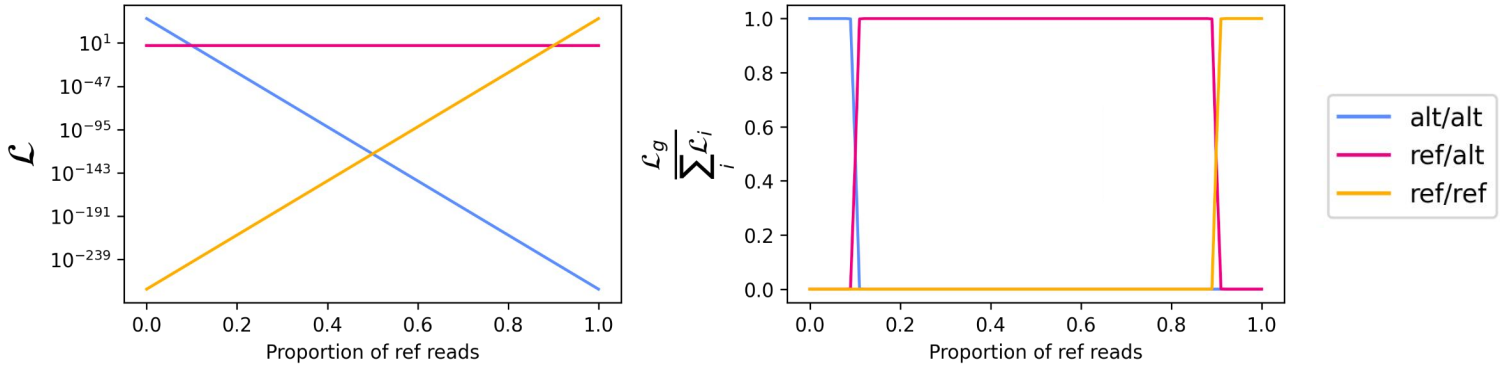
Example: Varying proportion of reference reads

Li 2011

Genotype likelihood

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} \Big[(m-g)\epsilon_j + g(1-\epsilon_j) \Big] \prod_{j=l+1}^{k} \Big[(m-g)(1-\epsilon_j) + g\epsilon_j) \Big]$$

Example: Varying proportion of reference reads



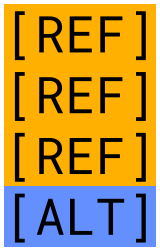
Total reads = 100, error=1e-3

Li 2011

Genotype likelihood

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} \Big[(m-g)\epsilon_j + g(1-\epsilon_j) \Big] \prod_{j=l+1}^{k} \Big[(m-g)(1-\epsilon_j) + g\epsilon_j \Big]$$

Example: Increasing error rate



3 ref reads, 1 alt read

Li 2011

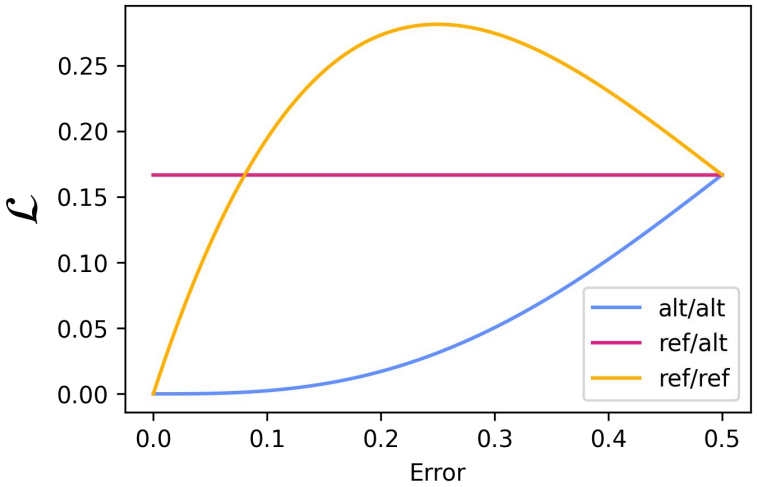
Genotype likelihood

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} \Big[(m-g)\epsilon_j + g(1-\epsilon_j) \Big] \prod_{j=l+1}^{k} \Big[(m-g)(1-\epsilon_j) + g\epsilon_j \Big]$$

Example: Increasing error rate

[REF]
[REF]
[REF]
[ALT]

3 ref reads, 1 alt read



error rate constant across reads (Var(e_j)=0)

Li 2011

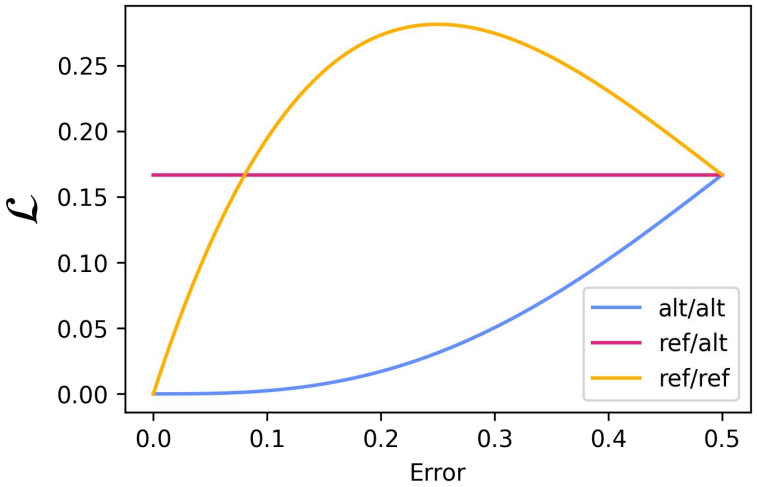
Genotype likelihood

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} \Big[(m-g)\epsilon_j + g(1-\epsilon_j) \Big] \prod_{j=l+1}^{k} \Big[(m-g)(1-\epsilon_j) + g\epsilon_j \Big]$$

Example: Increasing error rate

[REF]
[REF]
[REF]
[ALT]

3 ref reads, 1 alt read



error → 0.5
(i.e. random)

all likelihoods converge

error rate constant across reads (Var(e_j)=0)

Li 2011

Summary: Variant calling

Slide adapted from Luke Jostins-Dean

Summary: Variant calling

Probability that this site is variable, given the data

Depends on:

- Prior probability of variant
- Allele likelihood
  - Allele frequency
  - Genotype likelihoods

Slide adapted from Luke Jostins-Dean

Summary: Variant calling

Probability that this site is variable, given the data

Depends on:

- Prior probability of variant
- Allele likelihood
  - Allele frequency
  - Genotype likelihoods

QUAL score

(or site quality score)

#CHROM	POS	ID	REF	ALT	QUAL
chr19	48693829	.	C	T	3357.45
chr19	48693956	.	A	G	6806.98
chr19	48693992	.	G	A	1802.53
chr19	48694122	.	C	G	125.986

Slide adapted from Luke Jostins-Dean

Summary: Variant calling

Probability that this site is variable, given the data

Depends on:

- Prior probability of variant
- Allele likelihood
  - Allele frequency
  - Genotype likelihoods

Usually expressed on the **phred scale**

QUAL score

(or site quality score)

#CHROM	POS	ID	REF	ALT	QUAL
chr19	48693829	.	C	T	3357.45
chr19	48693956	.	A	G	6806.98
chr19	48693992	.	G	A	1802.53
chr19	48694122	.	C	G	125.986

Slide adapted from Luke Jostins-Dean

Summary: Variant calling

Probability that this site is variable, given the data

Depends on:

- Prior probability of variant
- Allele likelihood
  - Allele frequency
  - Genotype likelihoods

Usually expressed on the **phred scale**

$$Q = -10 \cdot \log_{10}(\text{Probability of error})$$

QUAL score

(or site quality score)

#CHROM	POS	ID	REF	ALT	QUAL
chr19	48693829	.	C	T	3357.45
chr19	48693956	.	A	G	6806.98
chr19	48693992	.	G	A	1802.53
chr19	48694122	.	C	G	125.986

Slide adapted from Luke Jostins-Dean

Summary: Variant calling

Probability that this site is variable, given the data

Depends on:

- Prior probability of variant
- Allele likelihood
  - Allele frequency
  - Genotype likelihoods

QUAL score

(or site quality score)

Usually expressed on the **phred scale**

$$Q = -10 \cdot \log_{10}(\text{Probability of error})$$

Error = 1	Error = 1/100	Error = 1/1000
Q = 0	Q = 20	Q = 30

#CHROM	POS	ID	REF	ALT	QUAL
chr19	48693829	.	C	T	3357.45
chr19	48693956	.	A	G	6806.98
chr19	48693992	.	G	A	1802.53
chr19	48694122	.	C	G	125.986

Slide adapted from Luke Jostins-Dean

Summary: Variant calling

Probability that this site is variable, given the data

Depends on:

- Prior probability of variant
- Allele likelihood
  - Allele frequency
  - Genotype likelihoods

Usually expressed on the **phred scale**

$$Q = -10 \cdot \log_{10}(\text{Probability of error})$$

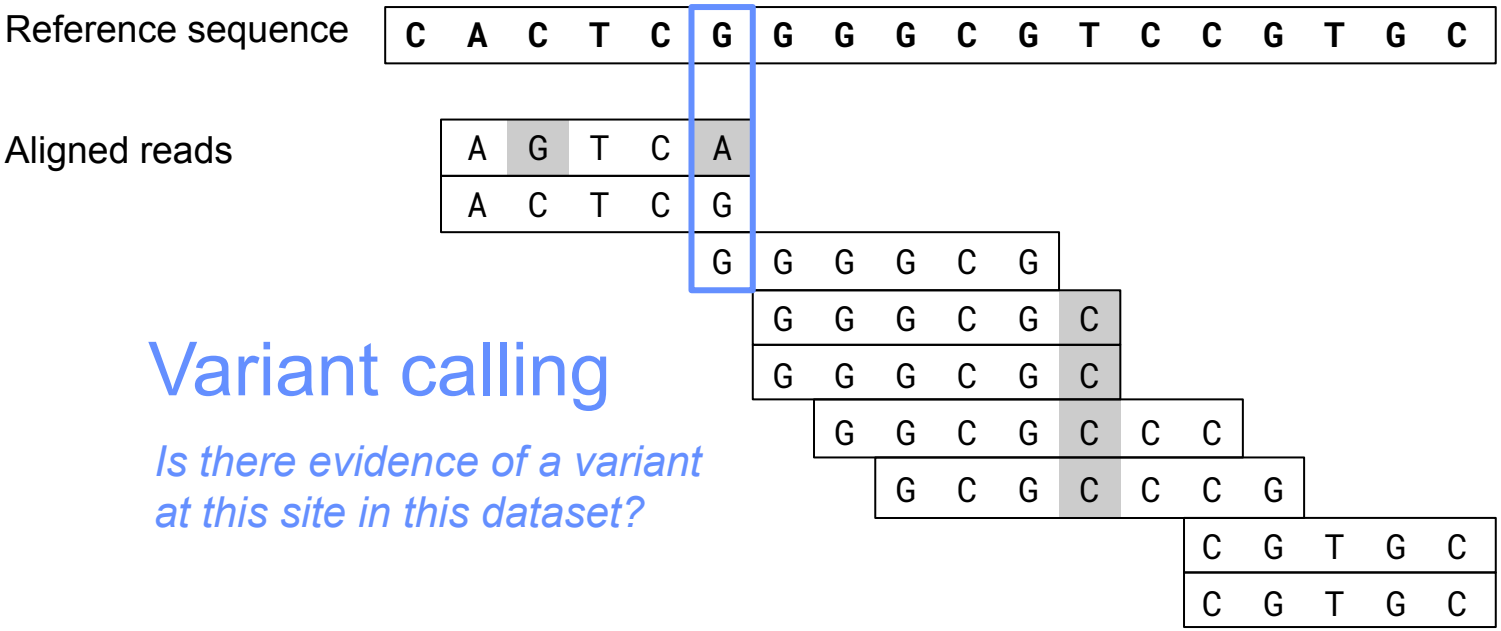
higher score = higher quality

QUAL score

(or site quality score)

#CHROM	POS	ID	REF	ALT	QUAL
chr19	48693829	.	C	T	3357.45
chr19	48693956	.	A	G	6806.98
chr19	48693992	.	G	A	1802.53
chr19	48694122	.	C	G	125.986

Slide adapted from Luke Jostins-Dean



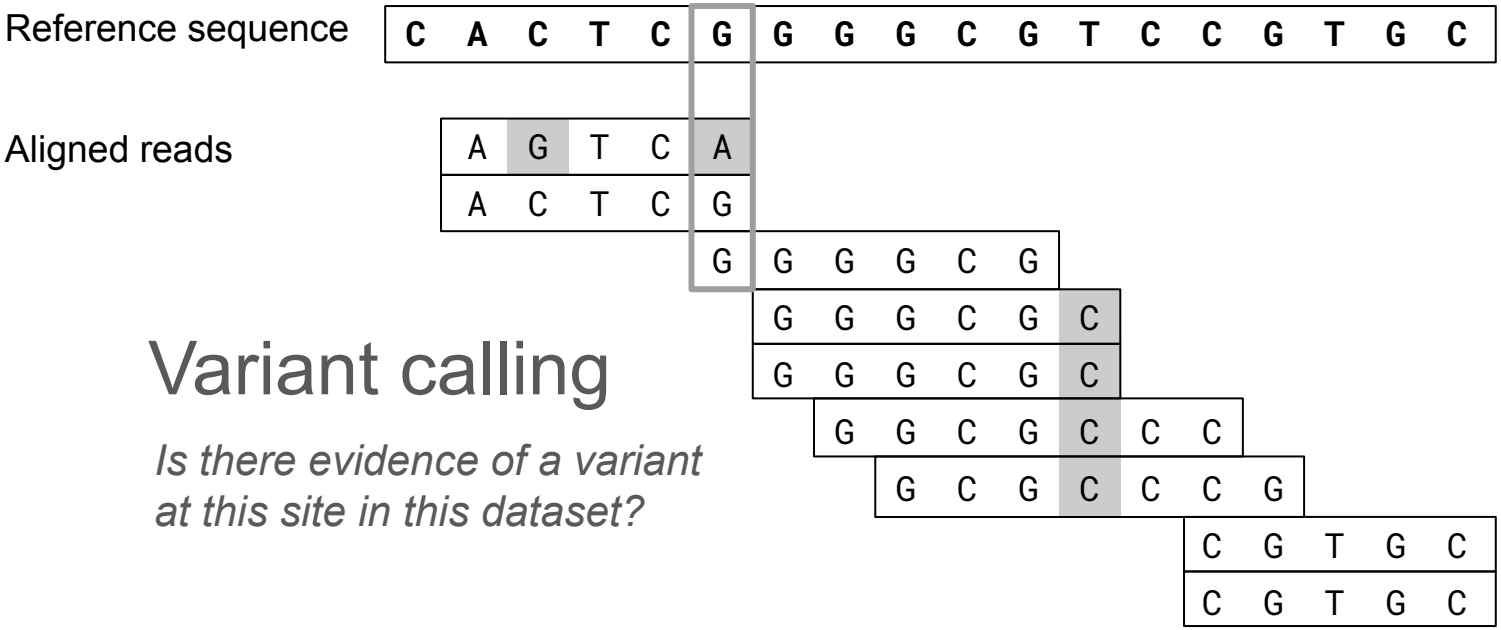
Variant calling

Is there evidence of a variant at this site in this dataset?

G/A?

Genotype calling

What are the genotypes for each individual?



Variant calling

Is there evidence of a variant at this site in this dataset?

Genotype calling

Genotype calling

$$Pr(g|Data_i) = \frac{Pr(Data_i|g)Pr(g|f_x)}{Pr(Data_i)}$$

Probability of genotype g, given data for sample i

Genotype calling

$$Pr(g|Data_i) = \frac{Pr(Data_i|g)Pr(g|f_x)}{Pr(Data_i)}$$

Probability of genotype g, given data for sample i

$$= \frac{G_i(g)Pr(g|f_x)}{\sum_{g'} G_i(g')Pr(g'|f_x)}$$

Genotype calling

Genotype posterior (**GP**)

$$Pr(g|Data_i) = \frac{Pr(Data_i|g)Pr(g|f_x)}{Pr(Data_i)}$$

Probability of genotype g, given data for sample i

$$= \frac{G_i(g)Pr(g|f_x)}{\sum_{g'} G_i(g')Pr(g'|f_x)}$$

Genotype priors

Genotype likelihoods
(**GL**, or **PL** if phred-scaled)

Genotype calling

Genotype posterior (**GP**)

$$Pr(g|Data_i) = \frac{Pr(Data_i|g)Pr(g|f_x)}{Pr(Data_i)}$$

Probability of genotype g, given data for sample i

$$= \frac{G_i(g)Pr(g|f_x)}{\sum_{g'} G_i(g')Pr(g'|f_x)}$$

Genotype priors

Genotype likelihoods
(**GL**, or **PL** if phred-scaled)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	FORMAT	HG02461	HG02462
chr19	48693829	.	C	T	3357.45	.	GT:PL:GP	0/0:0,9,75:0.810234,0.189766,2.21701e-08	1/1:85,9,0:2.87613e-09,0.213018,0.786982
chr19	48693956	.	A	G	6806.98	.	GT:PL:GP	0/0:0,99,255:1,3.1458e-11,4.93632e-28	0/0:0,105,255:1,7.9019e-12,4.93632e-28
chr19	48693992	.	G	A	1802.53	.	GT:PL:GP	0/0:0,96,255:1,1.43086e-11,2.56527e-29	0/0:0,105,255:1,1.80134e-12,2.56527e-29

Genotype calling

Choose genotype with **highest probability**

Genotype posterior (**GP**)

$$Pr(g|Data_i) = \frac{Pr(Data_i|g)Pr(g|f_x)}{Pr(Data_i)}$$

$$= \frac{G_i(g)Pr(g|f_x)}{\sum_{g'} G_i(g')Pr(g'|f_x)}$$

Probability of genotype g, given data for sample i

Genotype priors

Genotype likelihoods
(**GL**, or **PL** if phred-scaled)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	FORMAT	HG02461	HG02462
chr19	48693829	.	C	T	3357.45	.	GT:PL:GP	0/0:0,9,75:0.810234,0.189766,2.21701e-08	1/1:85,9,0:2.87613e-09,0.213018,0.786982
chr19	48693956	.	A	G	6806.98	.	GT:PL:GP	0/0:0,99,255:1,3.1458e-11,4.93632e-28	0/0:0,105,255:1,7.9019e-12,4.93632e-28
chr19	48693992	.	G	A	1802.53	.	GT:PL:GP	0/0:0,96,255:1,1.43086e-11,2.56527e-29	0/0:0,105,255:1,1.80134e-12,2.56527e-29

A few notes:

- Types of variants
- Single-sample vs. joint calling
- Types of errors

Variant types

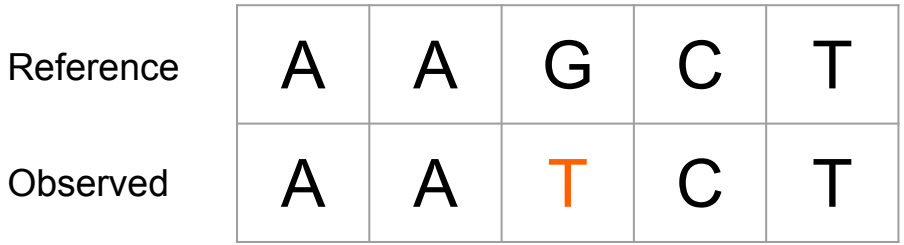
Variant types

Single-nucleotide polymorphism (SNP)

Insertion-deletion (indel)

Structural variant (SV)

Copy number variant (CNV)



Variant types

Single-nucleotide polymorphism (SNP)

Insertion-deletion (indel)

Structural variant (SV)

Copy number variant (CNV)

Insertion

A	A	G	C	T
A	A	TCGA	C	T

Deletion

A	A	GATT	C	T
A	A	T	C	T

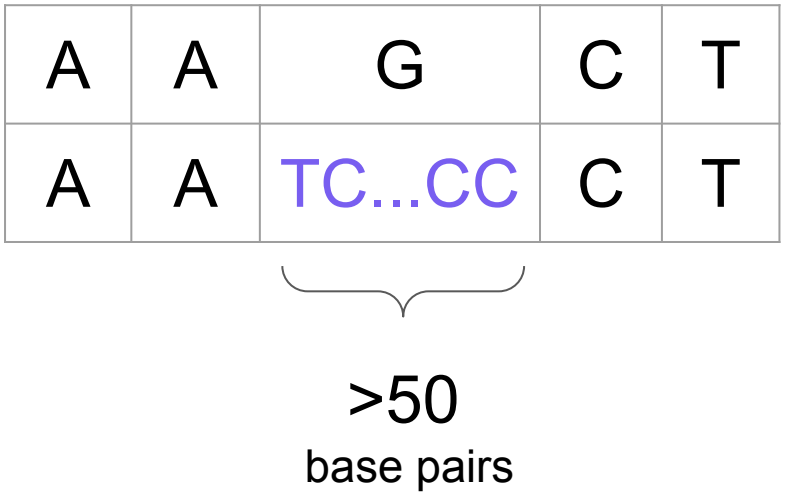
Variant types

Single-nucleotide polymorphism (SNP)

Insertion-deletion (indel)

Structural variant (SV)

Copy number variant (CNV)



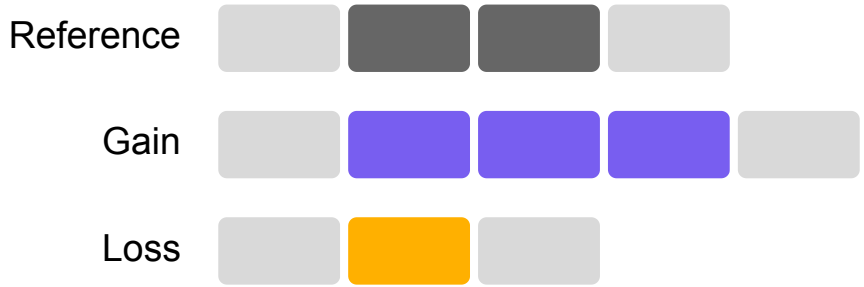
Variant types

Single-nucleotide polymorphism (SNP)

Insertion-deletion (indel)

Structural variant (SV)

Copy number variant (CNV)

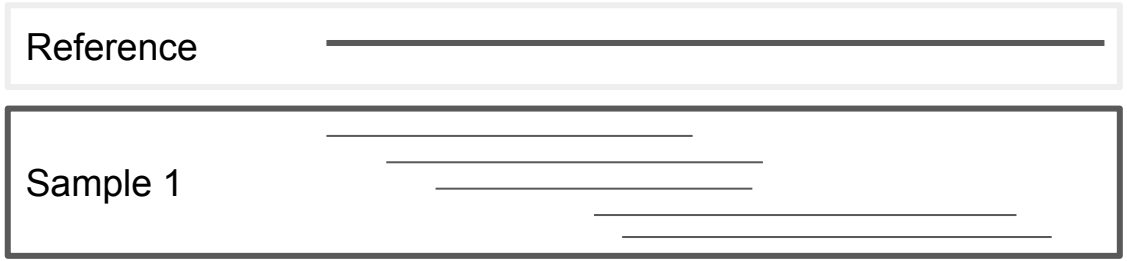


Calling types

Calling types

Single-sample calling

Joint calling

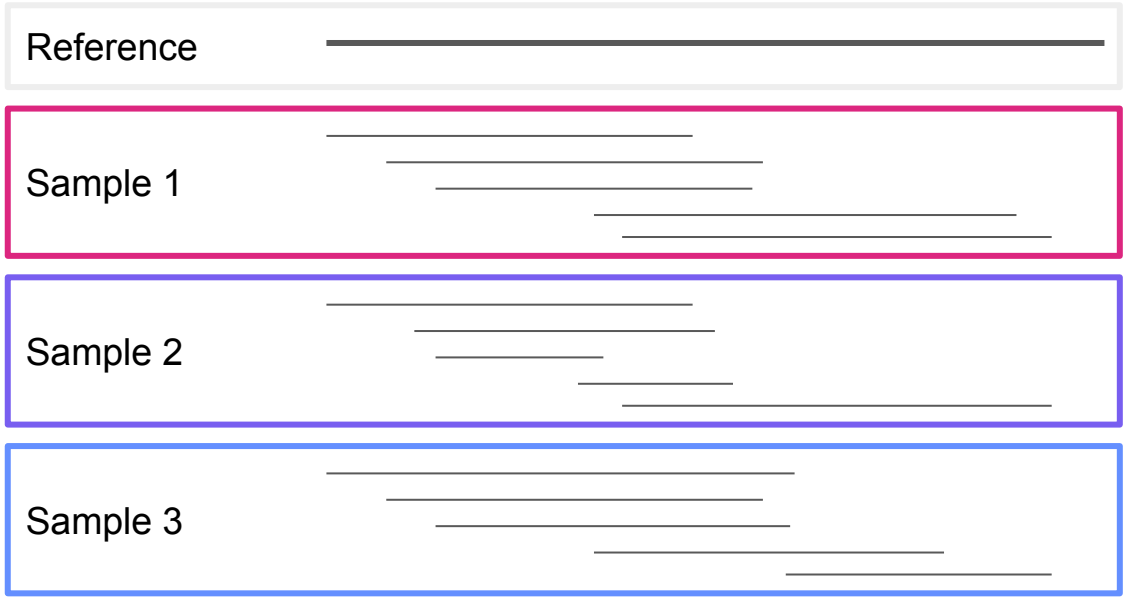


Calling types

Single-sample calling

Joint calling

Pros: Better at identifying rare variants
Cons: More computationally expensive



Error types

Error types

Almost all variant callers rely on probabilistic models

(i.e. truly at random)

These models make assumptions – the biggest of which being that errors are **independent**

Slide adapted from Luke Jostins-Dean

Error types

Almost all variant callers rely on probabilistic models

(i.e. truly at random)

These models make assumptions – the biggest of which being that errors are **independent**

Example: Independent errors

If there's a 0.01 chance of seeing G when truth is a C, the chance of seeing it 5 times is 0.01^5

Slide adapted from Luke Jostins-Dean

Error types

Almost all variant callers rely on probabilistic models

(i.e. truly at random)

These models make assumptions – the biggest of which being that errors are **independent**

> Example: Independent errors
>
> If there's a 0.01 chance of seeing G when truth is a C, the chance of seeing it 5 times is 0.01^5

But in practice, systematic (non-random) errors exist!

Slide adapted from Luke Jostins-Dean

Example: Context-specific errors

Truth	AGGGCAA
Sequencing	AGGGCAA
	AGGG**G**AA
	AGGGCAA
	AGGG**G**AA
	AGGG**G**AA

“GGG” motif

Slide adapted from Luke Jostins-Dean

Example: Context-specific errors

Truth	AGGGCAA
Sequencing	AGGGCAA
	AGGG**G**AA
	AGGGCAA
	AGGG**G**AA
	AGGG**G**AA

"GGG" motif

Solution: Strand bias test

Compare enrichment of alternate alleles in forward vs. reverse strands.

Slide adapted from Luke Jostins-Dean

Example: Context-specific errors

Truth AGGGCAA

Sequencing
AGGGCAA
AGGG**G**AA
AGGGCAA
AGGG**G**AA
AGGG**G**AA

"GGG" motif

Solution: Strand bias test

Compare enrichment of alternate alleles in forward vs. reverse strands.

Solution: Read position bias test

Correct for bias of sequencing errors typically occurring at the ends of reads.

Slide adapted from Luke Jostins-Dean

Example: Systematic mapping errors

Reference TCCACATGAACCACGTCTTGAAATC

Truth TCCACATGAACCACGT---GAAATC

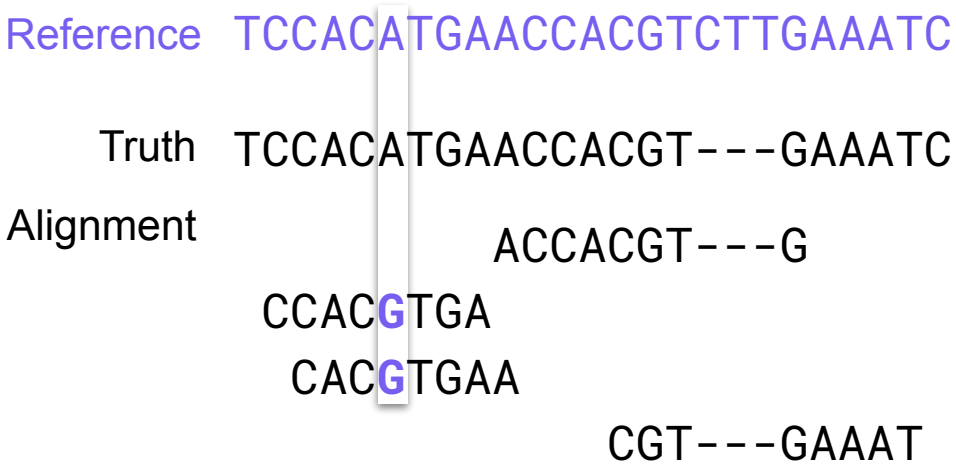
Alignment

```
ACCACGT---G
 CCACGT---GA
  CACGT---GAA
    CGT---GAAAT
```

This is the **true mapping**
(CTT deletion)

Slide adapted from Luke Jostins-Dean

Example: Systematic mapping errors

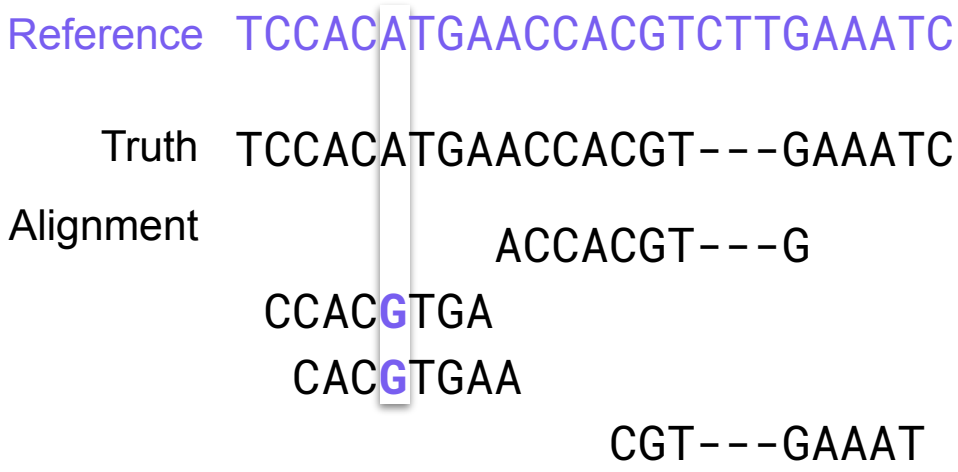


However, the alignment algorithm may lead to **incorrect mapping**.

This looks like a new SNP but the SNP doesn't exist!

Slide adapted from Luke Jostins-Dean

Example: Systematic mapping errors



Solution:
Mapping quality bias test

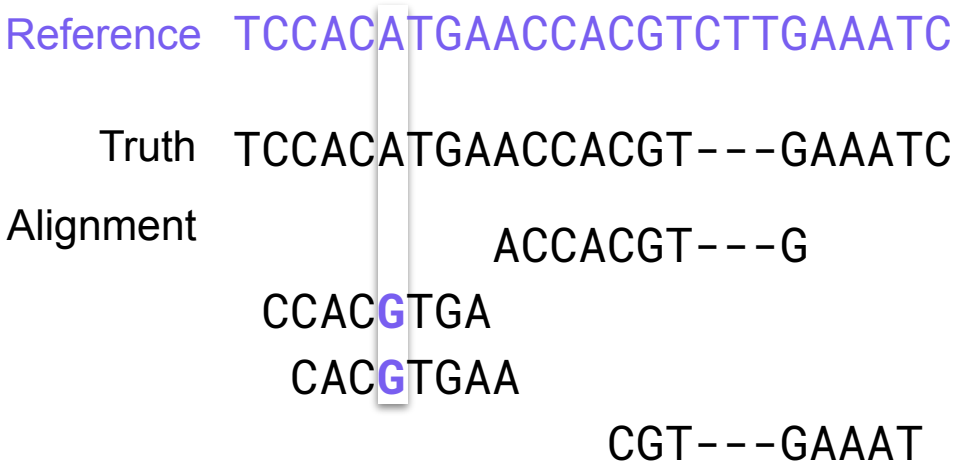
Test whether non-reference reads have lower quality.

However, the alignment algorithm may lead to **incorrect mapping**.

This looks like a new SNP but the SNP doesn't exist!

Slide adapted from Luke Jostins-Dean

Example: Systematic mapping errors



However, the alignment algorithm may lead to **incorrect mapping**.

This looks like a new SNP but the SNP doesn't exist!

Solution:
Mapping quality bias test

Test whether non-reference reads have lower quality.

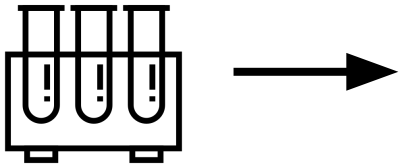
Solution: mapQ0

Remove variants with high counts of Q=0 reads

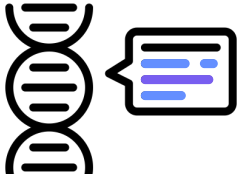
Slide adapted from Luke Jostins-Dean

Overview

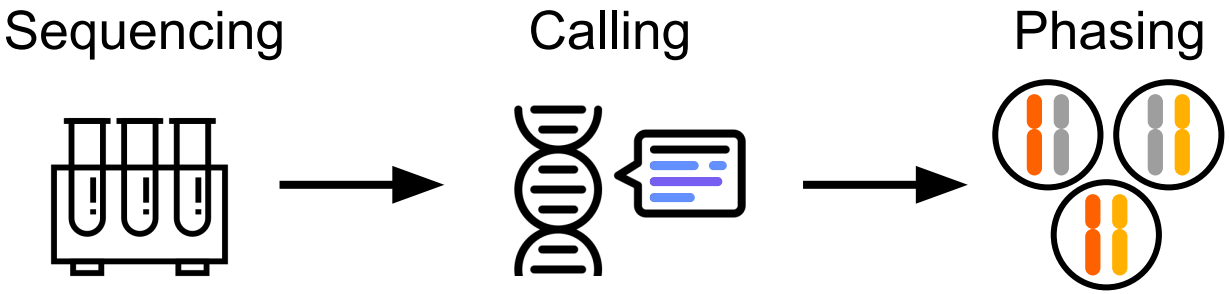
Sequencing



Calling

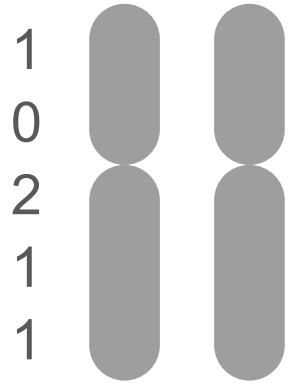


Overview

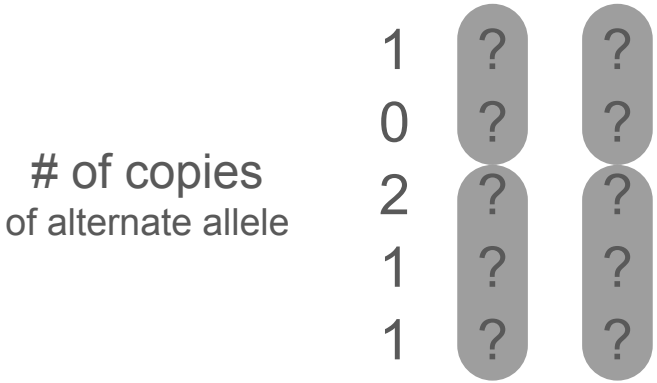



A/G
T/T
G/G
A/T
T/G

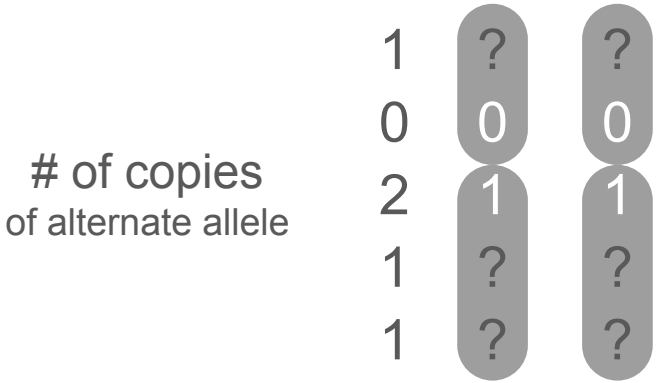
# of copies
of alternate allele



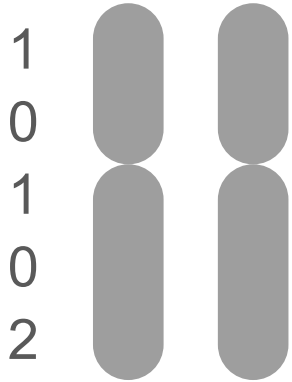
How to separate genotype into haplotypes?

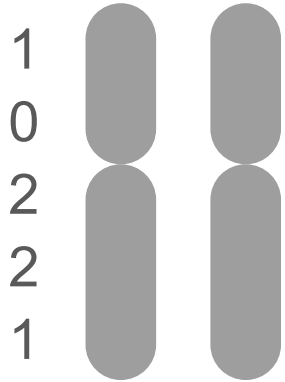


How to separate genotype into haplotypes?



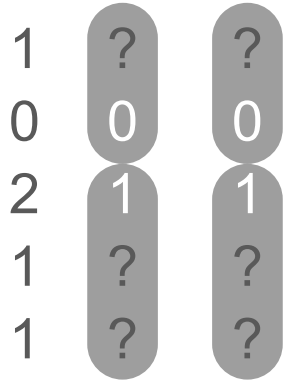
Mother



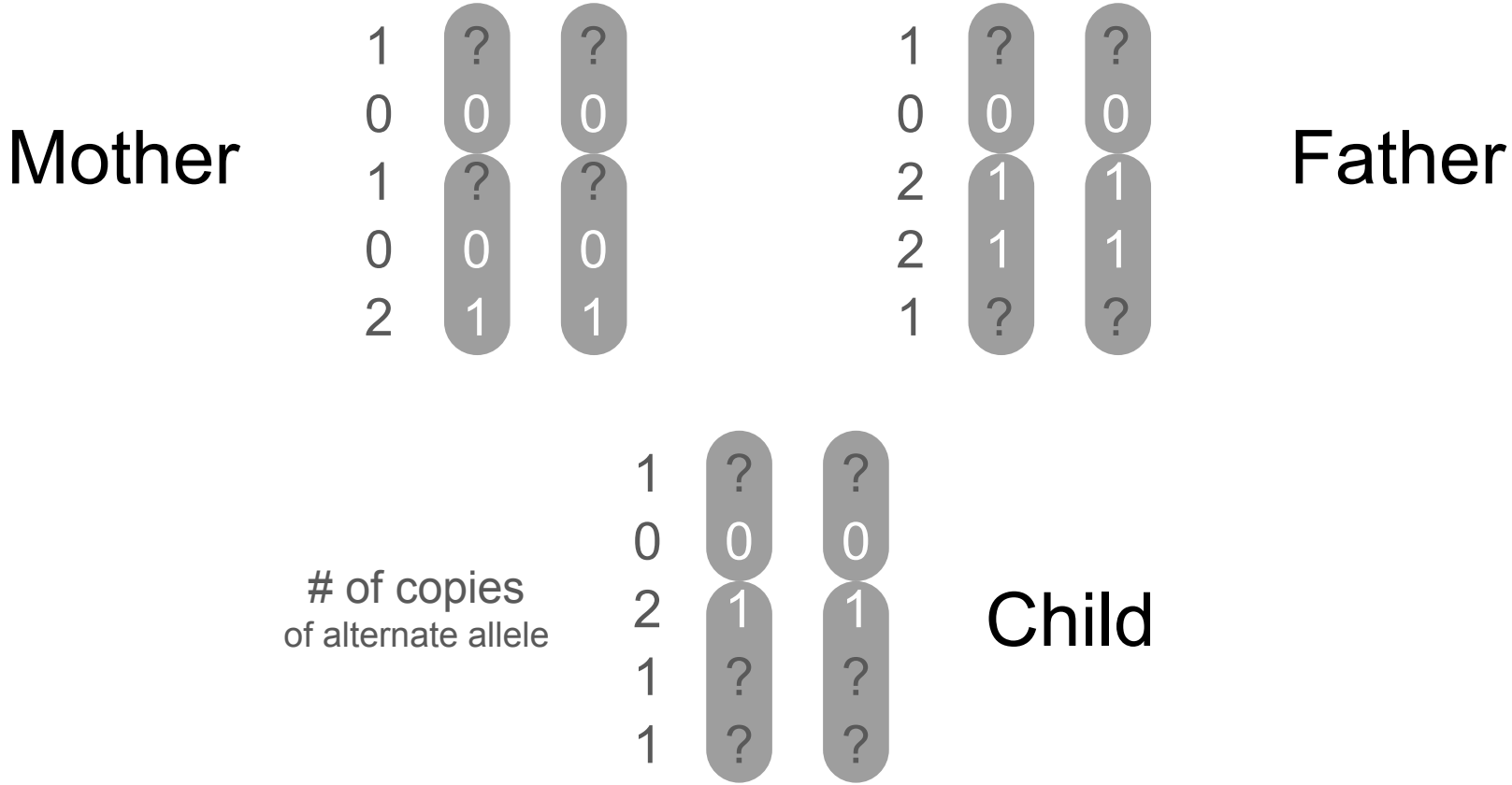


Father

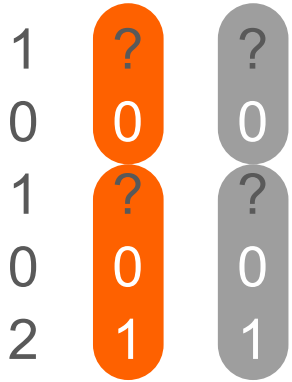
# of copies
of alternate allele

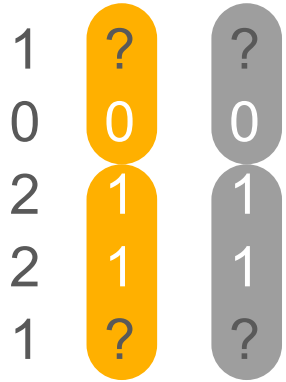


Child


Mother
1 ? ?
0 0 0
1 ? ?
0 0 0
2 1 1
1 ? ?
0 0 0
2 1 1
2 1 1
1 ? ?
Father
# of copies
of alternate allele
1 ? ?
0 0 0
2 1 1
1 ? ?
1 ? ?
Child

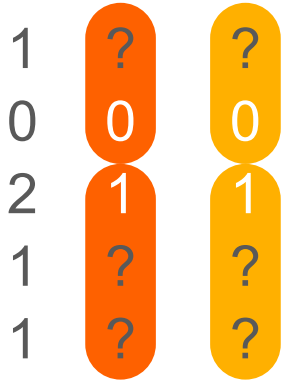
Mother





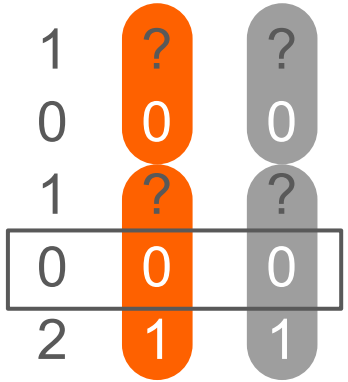
Father

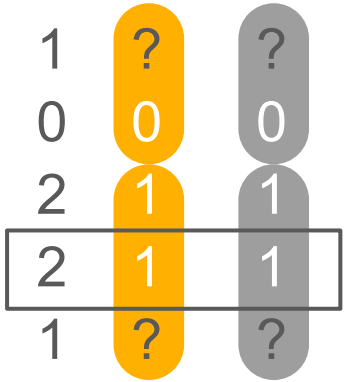
# of copies
of alternate allele



Child

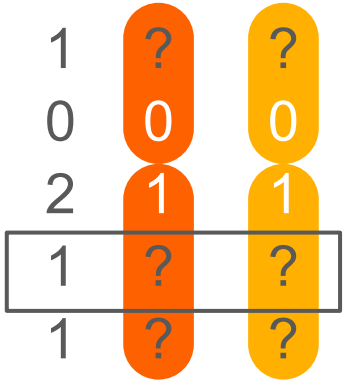
Mother





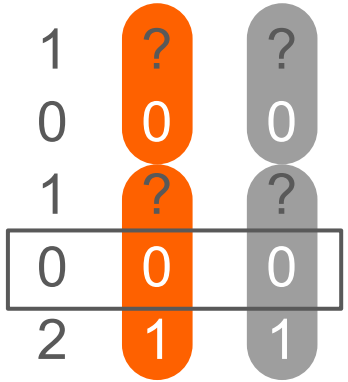
Father

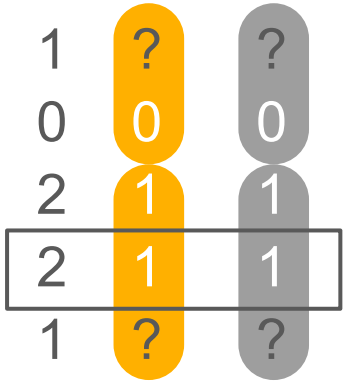
# of copies
of alternate allele



Child

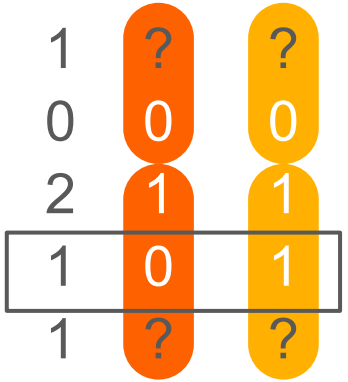
Mother





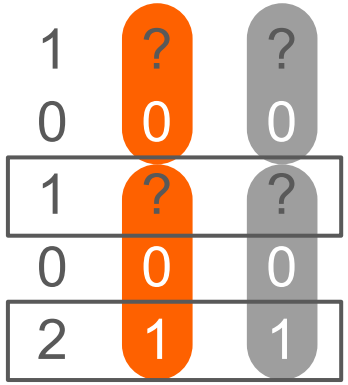
Father

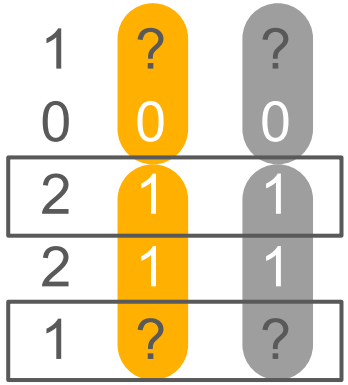
# of copies
of alternate allele



Child

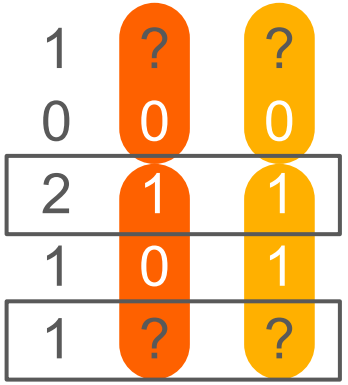
Mother





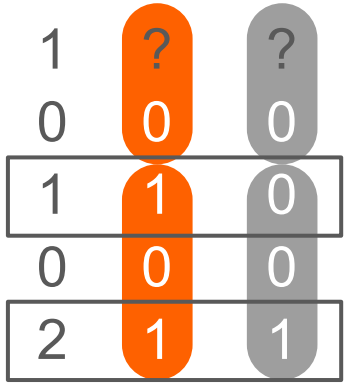
Father

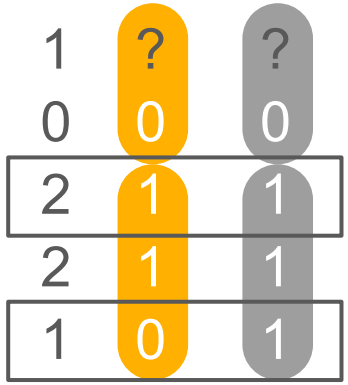
# of copies
of alternate allele



Child

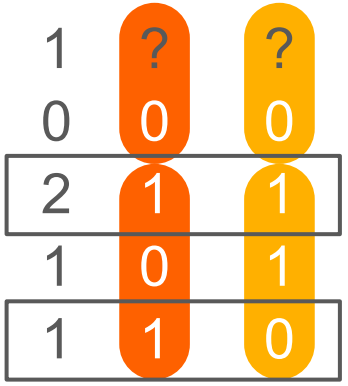
Mother





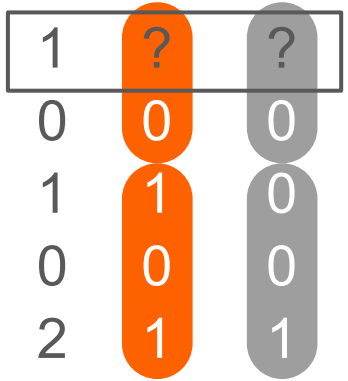
Father

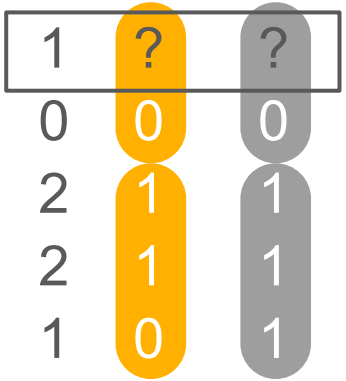
# of copies
of alternate allele



Child

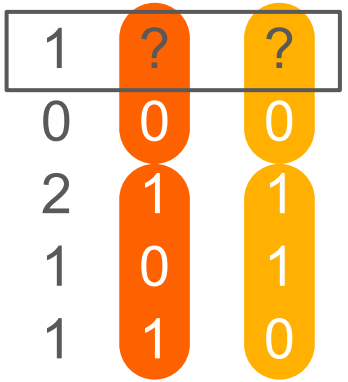
Mother





Father

# of copies
of alternate allele



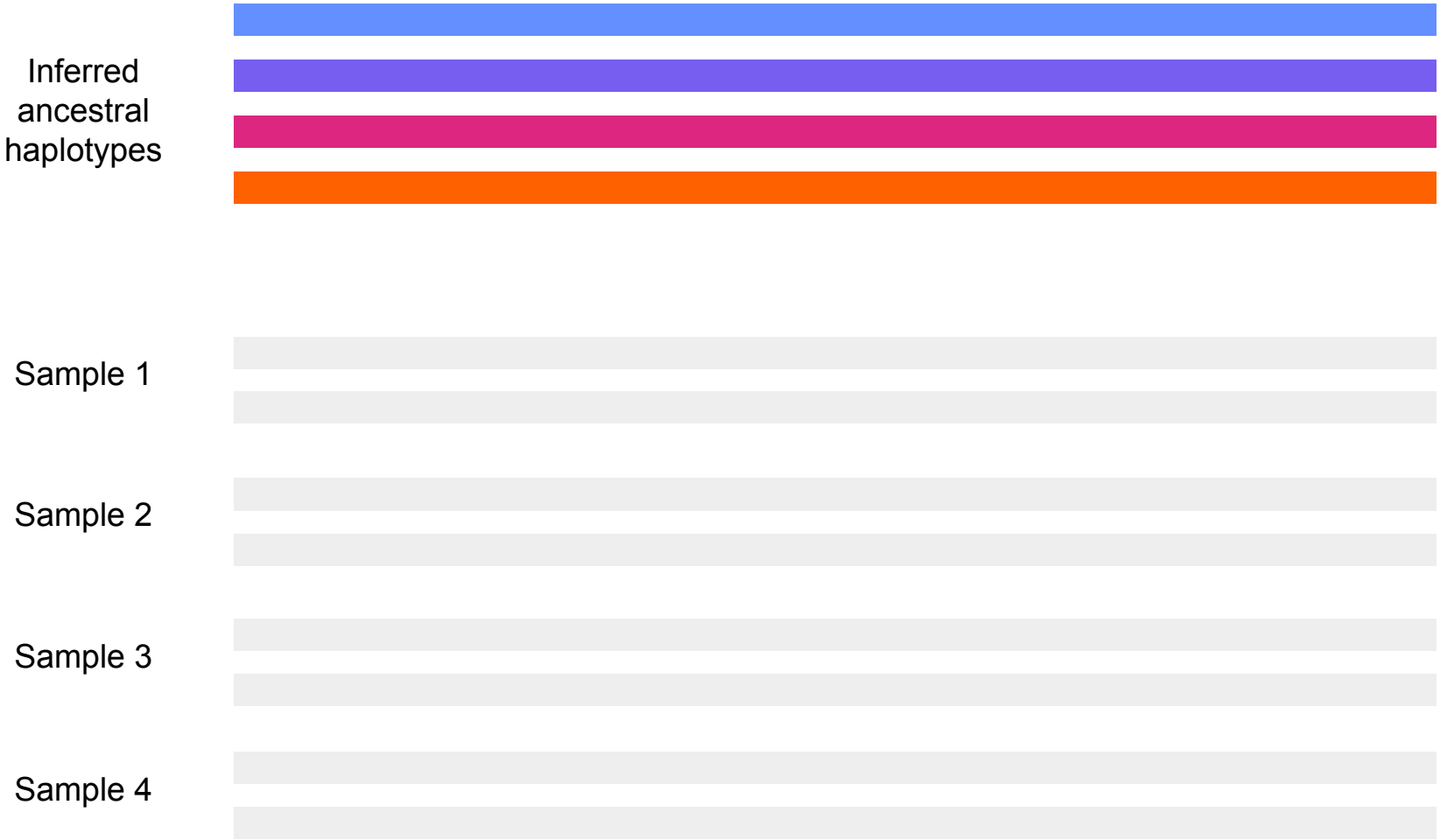
Child

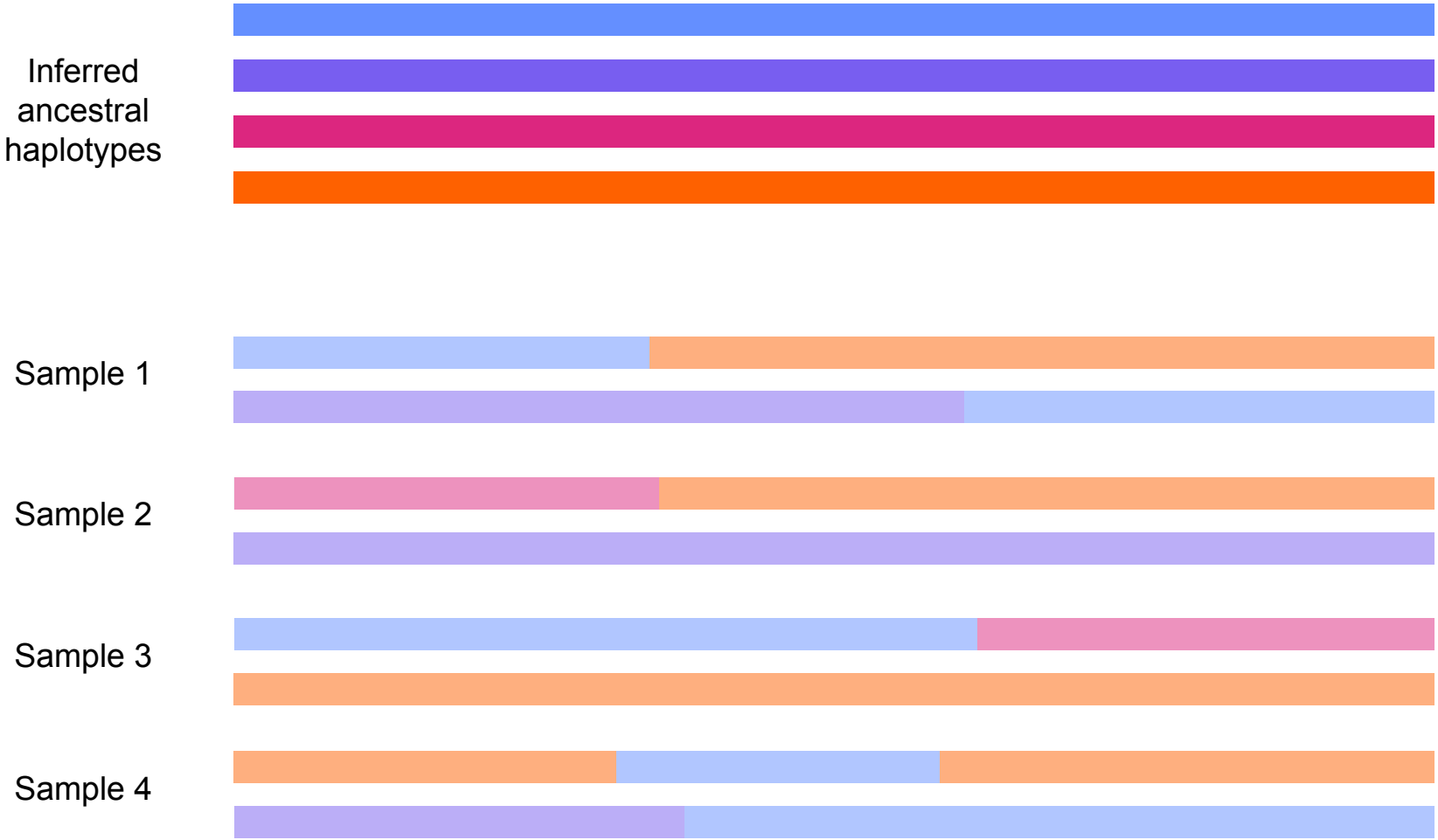
Sample 1

Sample 2

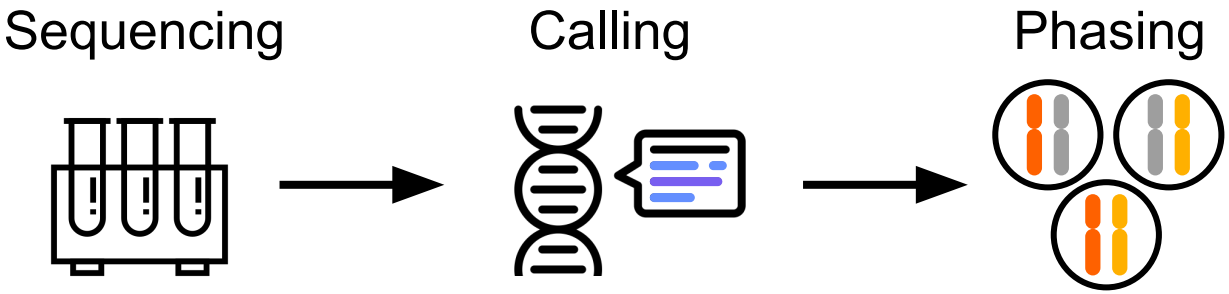
Sample 3

Sample 4

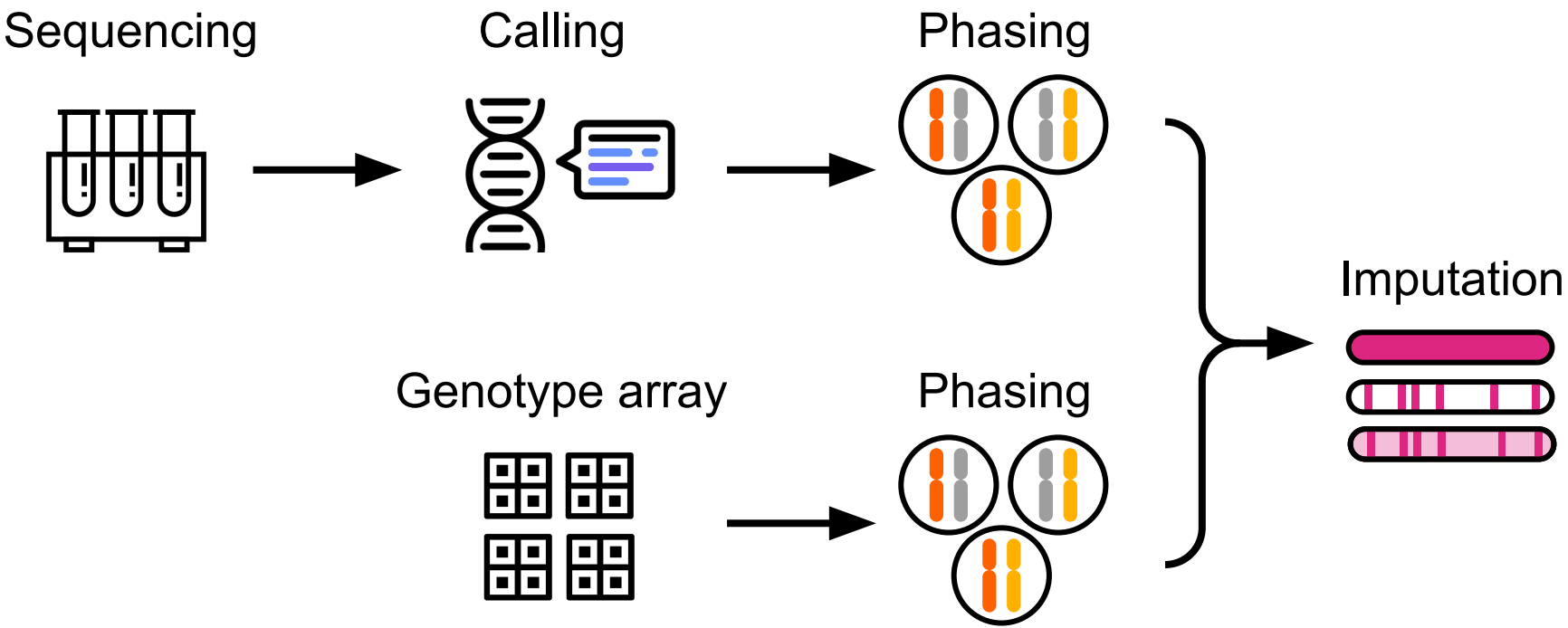

Inferred ancestral haplotypes
Sample 1
Sample 2
Sample 3
Sample 4


Inferred ancestral haplotypes
Sample 1
Sample 2
Sample 3
Sample 4

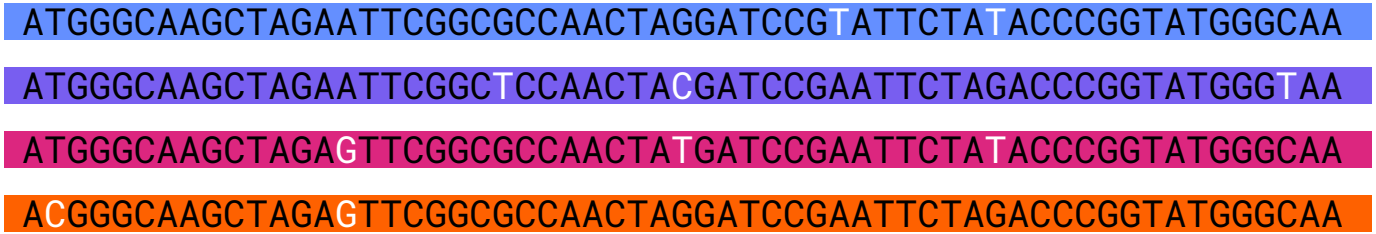
Overview



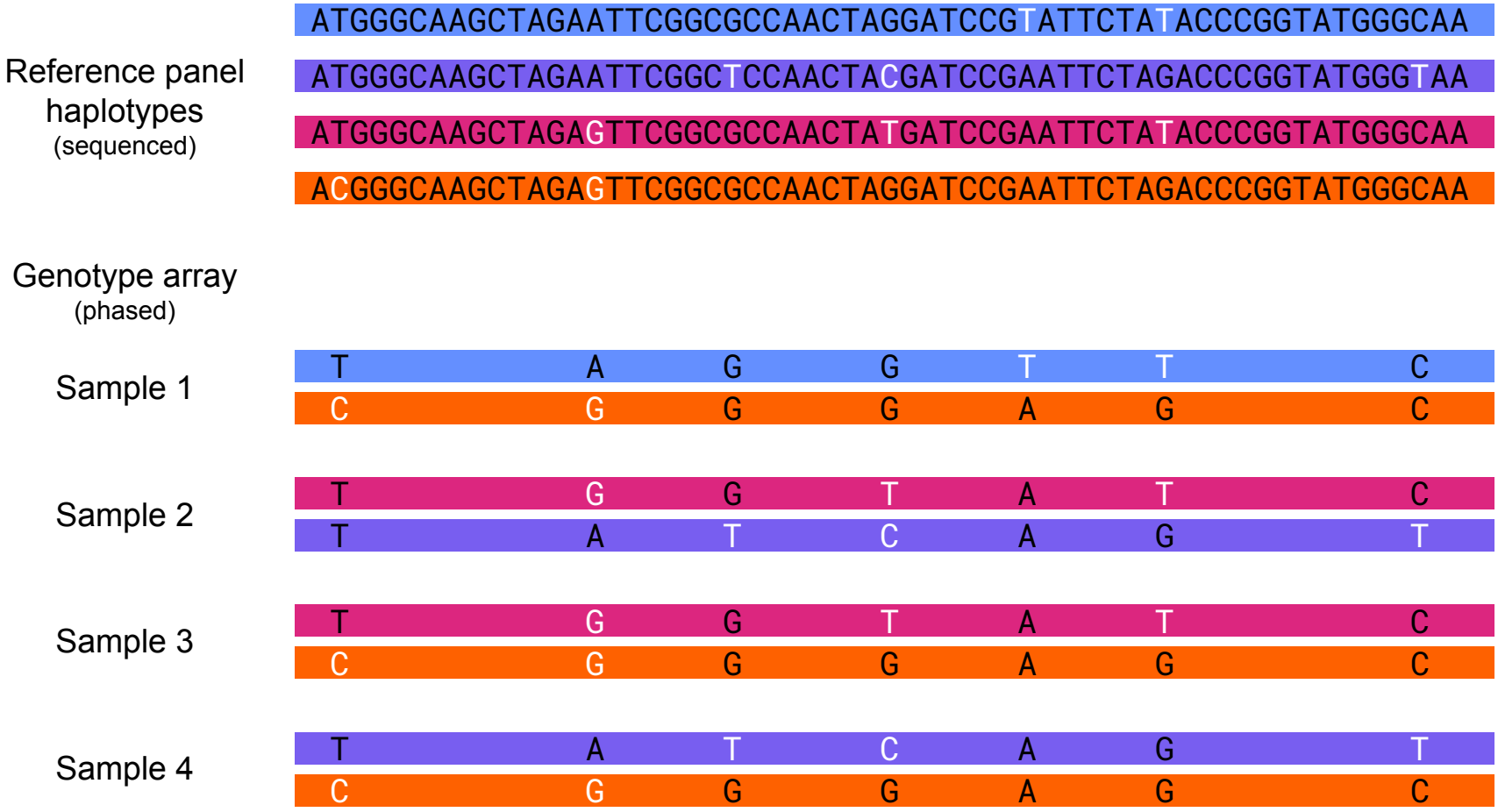
Overview


Sequencing
Calling
Phasing
Imputation
Genotype array
Phasing

Reference panel haplotypes
(sequenced)

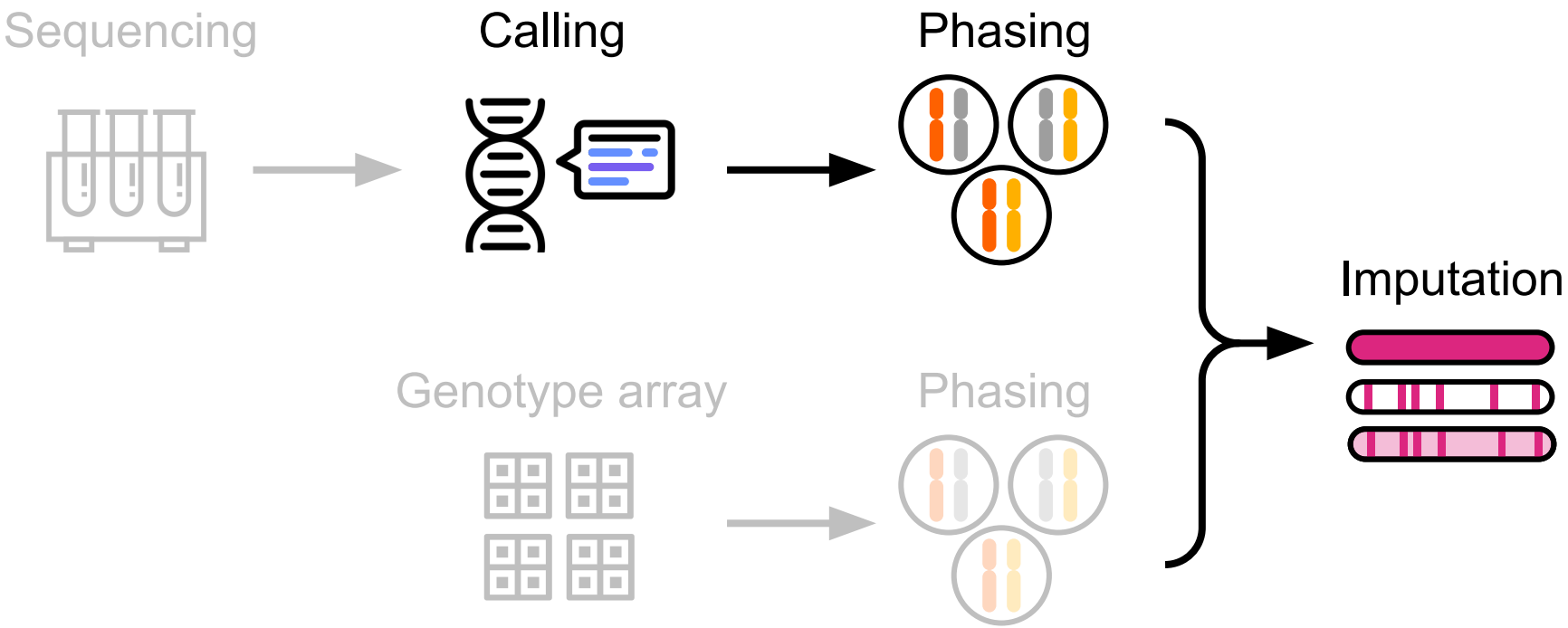



Reference panel haplotypes
(sequenced)
ATGGGCAAGCTAGAATTCGGCGCCAACTAGGATCCGTATTCTATACCCGGTATGGGCAA
ATGGGCAAGCTAGAATTCGGCTCCAACTACGATCCGAATTCTAGACCCGGTATGGGTAA
ATGGGCAAGCTAGAGTTCGGCGCCAACTATGATCCGAATTCTATACCCGGTATGGGCAA
ACGGGCAAGCTAGAGTTCGGCGCCAACTAGGATCCGAATTCTAGACCCGGTATGGGCAA
Genotype array
(phased)
Sample 1
T A G G T T C
C G G G A G C
Sample 2
T G G T A T C
T A T C A G T
Sample 3
T G G T A T C
C G G G A G C
Sample 4
T A T C A G T
C G G G A G C


Reference panel haplotypes
(sequenced)
ATGGGCAAGCTAGAATTCGGCGCCAACTAGGATCCGTATTCTATACCCGGTATGGGCAA
ATGGGCAAGCTAGAATTCGGCTCCAACTACGATCCGAATTCTAGACCCGGTATGGGTAA
ATGGGCAAGCTAGAGTTCGGCGCCAACTATGATCCGAATTCTATACCCGGTATGGGCAA
ACGGGCAAGCTAGAGTTCGGCGCCAACTAGGATCCGAATTCTAGACCCGGTATGGGCAA
Genotype array
(phased)
Sample 1
T A G G T T C
C G G G A G C
Sample 2
T G G T A T C
T A T C A G T
Sample 3
T G G T A T C
C G G G A G C
Sample 4
T A T C A G T
C G G G A G C


Reference panel haplotypes (sequenced)
ATGGGCAAGCTAGAATTCGGCGCCAACTAGGATCCGTATTCTATACCCGGTATGGGCAA
ATGGGCAAGCTAGAATTCGGCTCCAACTACGATCCGAATTCTAGACCCGGTATGGGTAA
ATGGGCAAGCTAGAGTTCGGCGCCAACTATGATCCGAATTCTATACCCGGTATGGGCAA
ACGGGCAAGCTAGAGTTCGGCGCCAACTAGGATCCGAATTCTAGACCCGGTATGGGCAA
Genotype array (phased)
Sample 1
ATGGGCAAGCTAGAATTCGGCGCCAACTAGGATCCGTATTCTATACCCGGTATGGGCAA
ACGGGCAAGCTAGAGTTCGGCGCCAACTAGGATCCGAATTCTAGACCCGGTATGGGCAA
Sample 2
ATGGGCAAGCTAGAGTTCGGCGCCAACTATGATCCGAATTCTATACCCGGTATGGGCAA
ATGGGCAAGCTAGAATTCGGCTCCAACTACGATCCGAATTCTAGACCCGGTATGGGTAA
Sample 3
ATGGGCAAGCTAGAGTTCGGCGCCAACTATGATCCGAATTCTATACCCGGTATGGGCAA
ACGGGCAAGCTAGAGTTCGGCGCCAACTAGGATCCGAATTCTAGACCCGGTATGGGCAA
Sample 4
ATGGGCAAGCTAGAATTCGGCTCCAACTACGATCCGAATTCTAGACCCGGTATGGGTAA
ACGGGCAAGCTAGAGTTCGGCGCCAACTAGGATCCGAATTCTAGACCCGGTATGGGCAA

Practical


Sequencing
Calling
Phasing
Imputation
Genotype array
Phasing