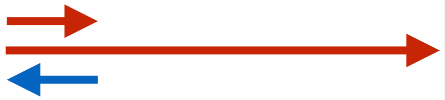
Read Orientation Artifact Filter for Somatic Variants

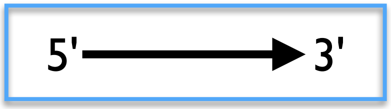
Takuto Sato

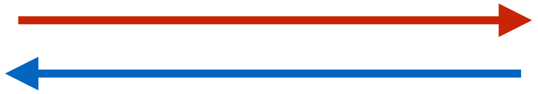
DSDE Methods Meeting
9/22/17

What is read orientation

How top strand and bottom strand map to the reference



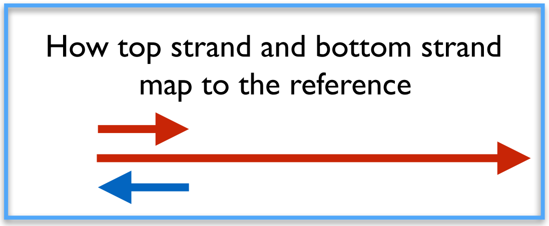


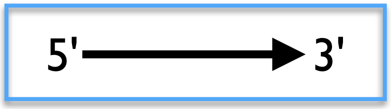


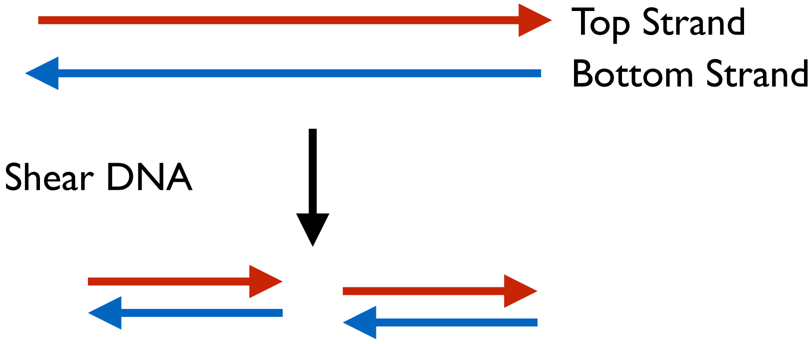
Top Strand

Bottom Strand

What is read orientation


How top strand and bottom strand map to the reference


5' 3'

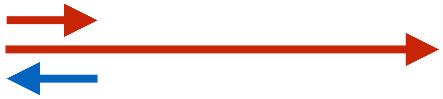

Top Strand

Bottom Strand

Shear DNA

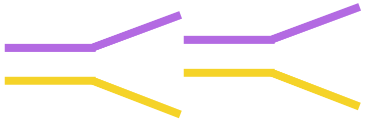
What is read orientation

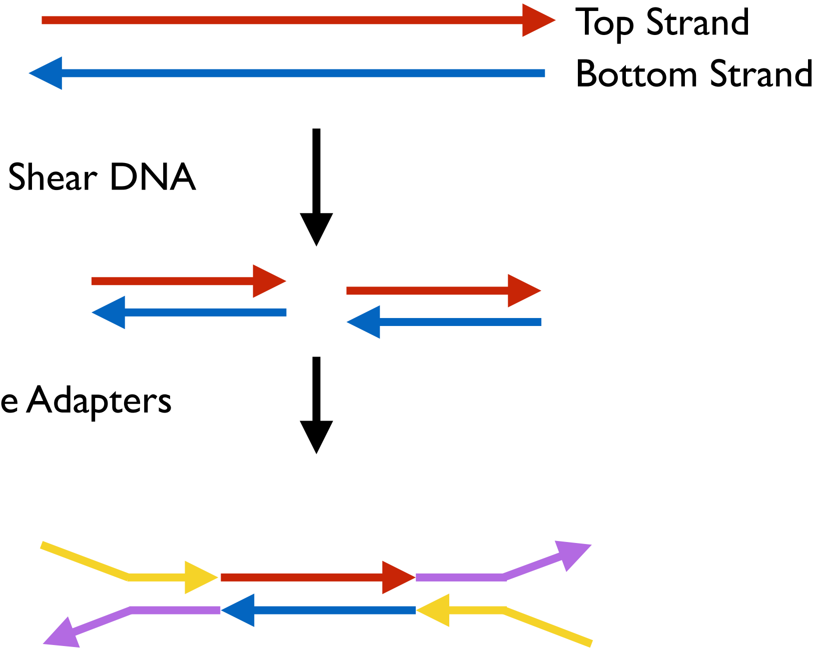
How top strand and bottom strand map to the reference



5' → 3'

Adapters





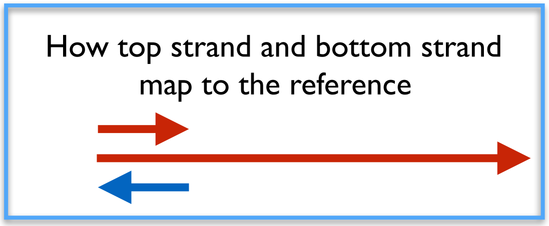
Top Strand

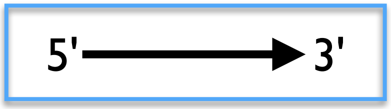
Bottom Strand

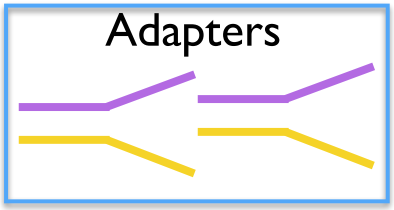
Shear DNA

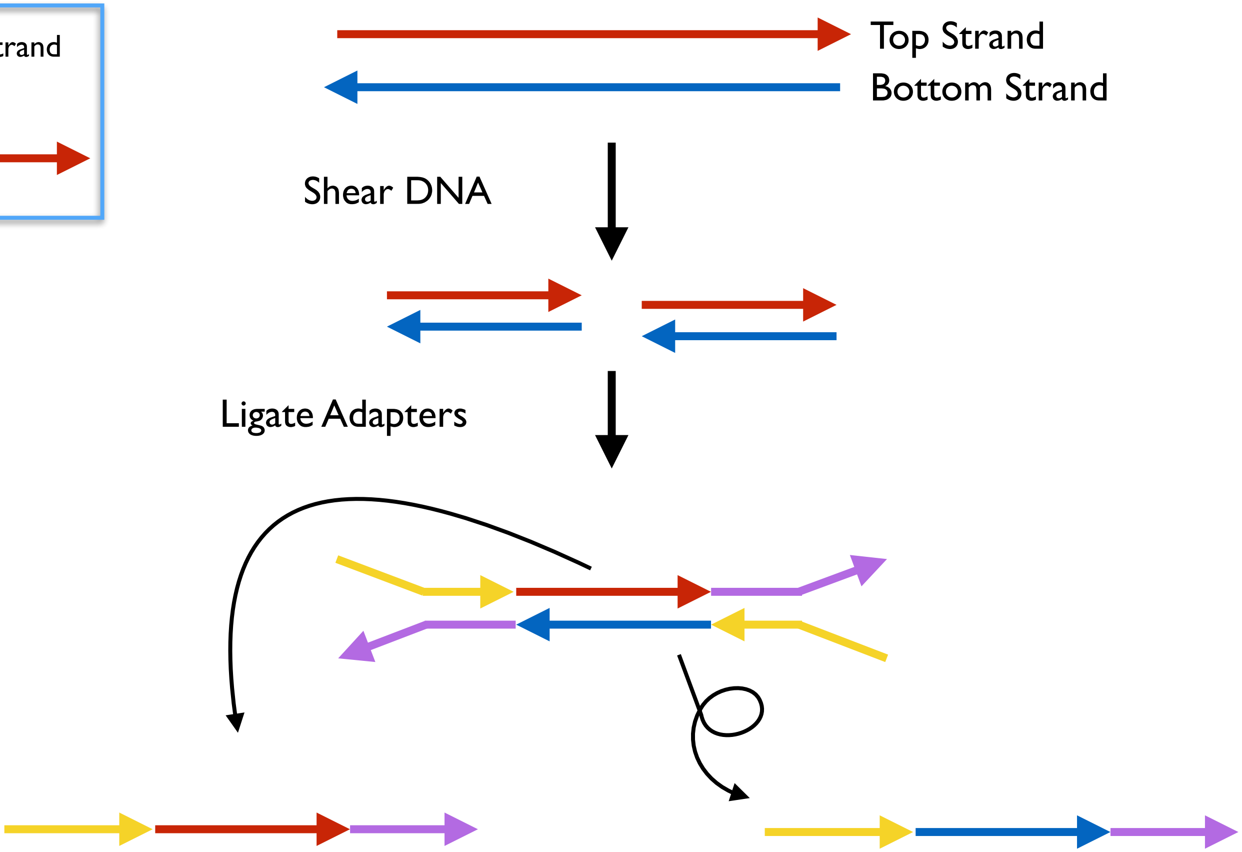
Ligate Adapters

What is read orientation


How top strand and bottom strand map to the reference



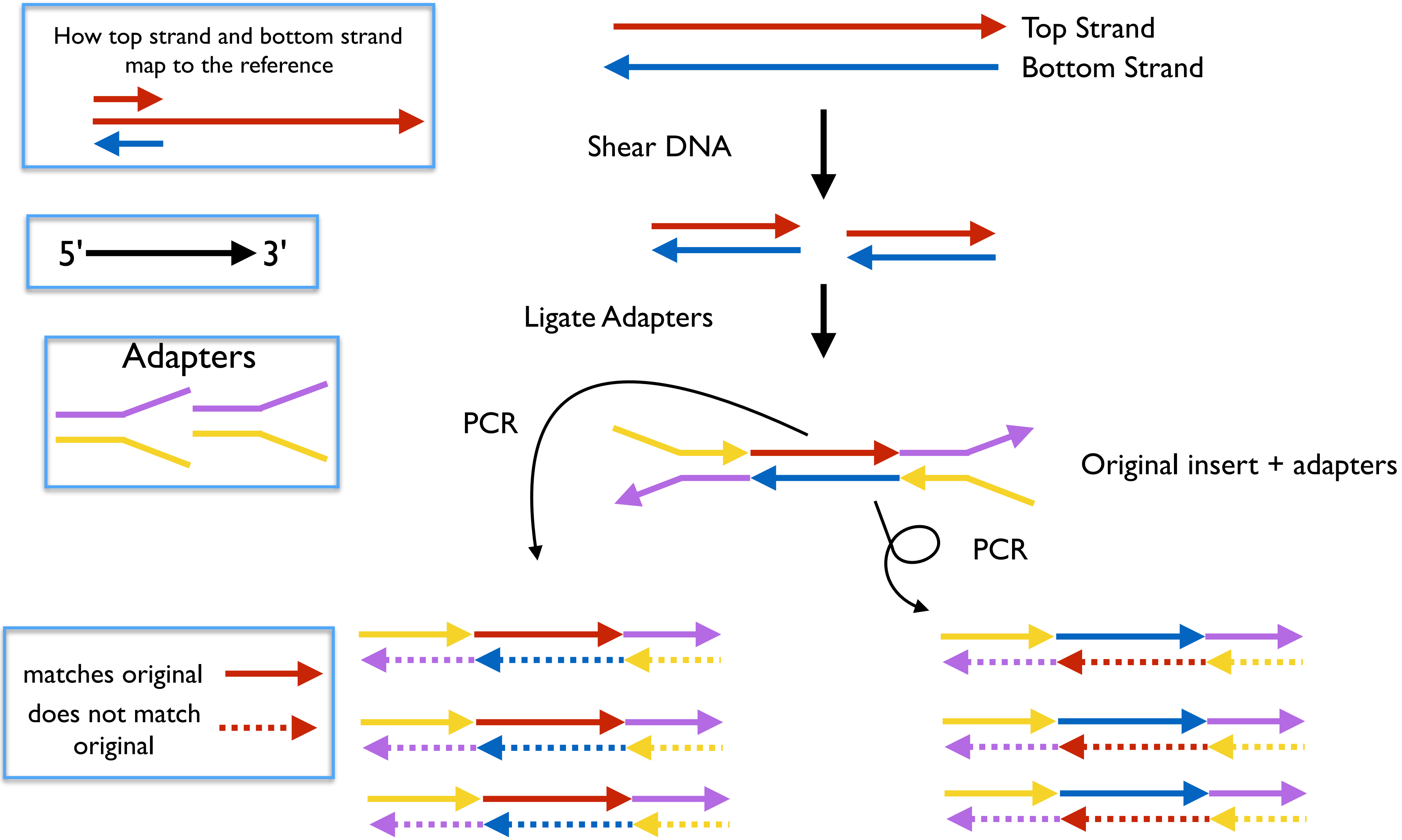

Adapters


Top Strand
Bottom Strand

Shear DNA

Ligate Adapters

What is read orientation



How top strand and bottom strand map to the reference

5' 3'

Adapters

matches original

does not match original

Top Strand

Bottom Strand

Shear DNA

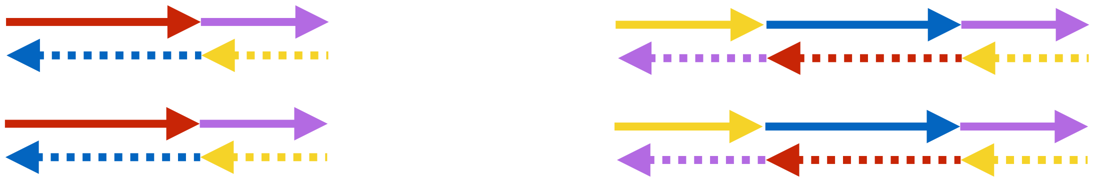
Ligate Adapters

PCR

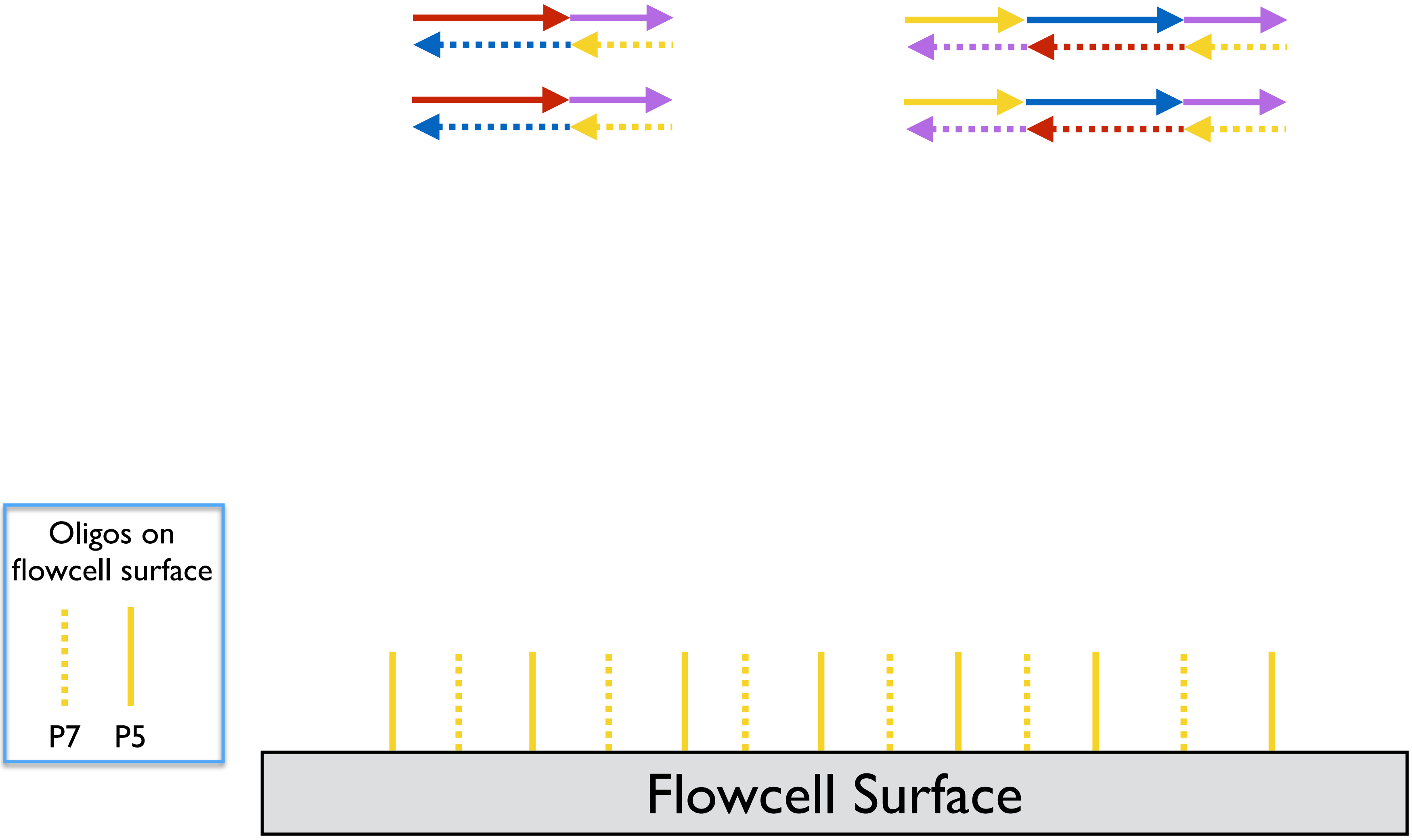
Original insert + adapters

PCR

Illumina Sequencing



Illumina Sequencing

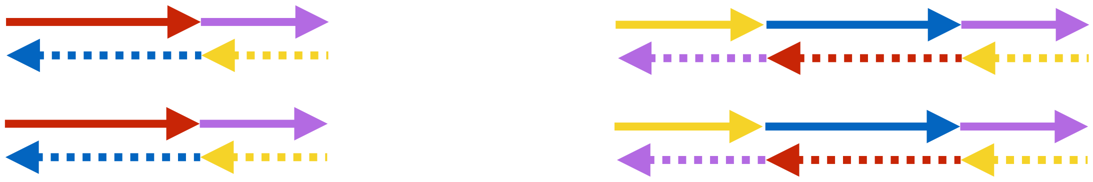


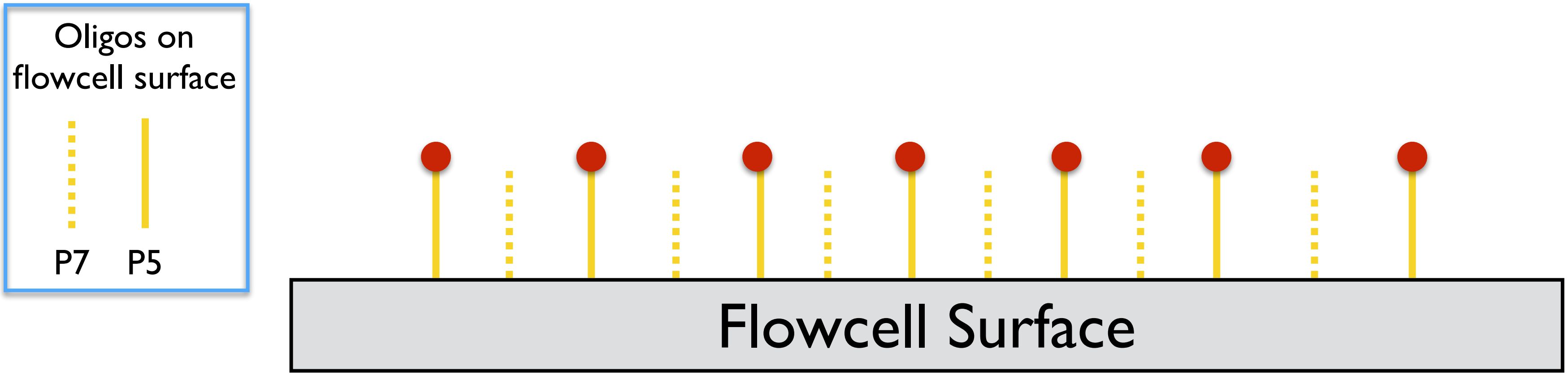
Oligos on flowcell surface

P7 P5

Flowcell Surface

Illumina Sequencing





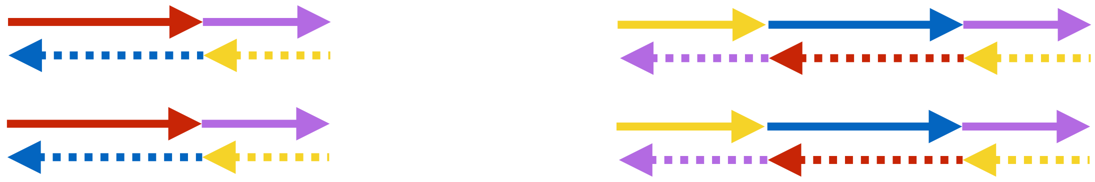
Oligos on flowcell surface

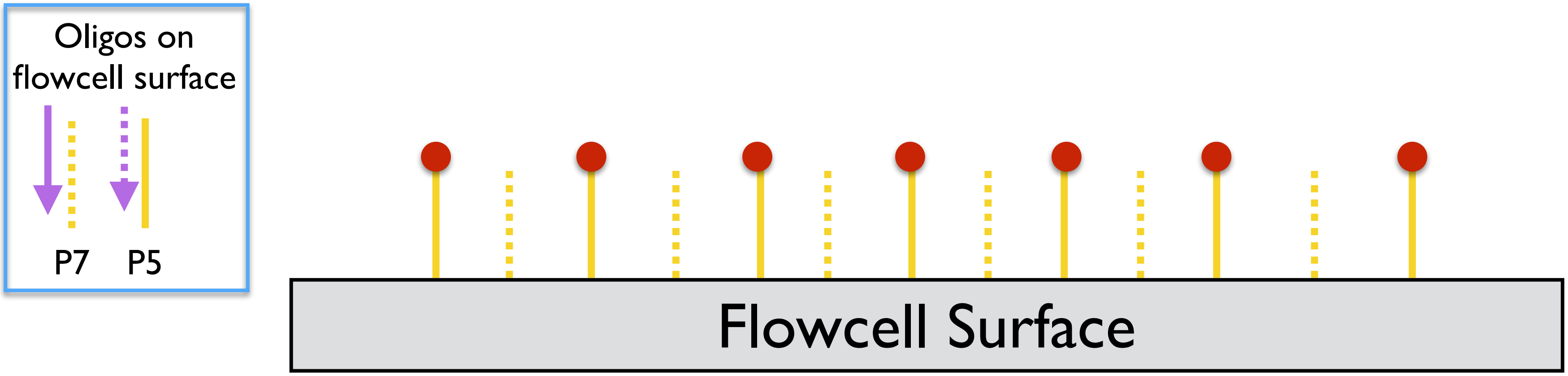
P7 P5

Flowcell Surface

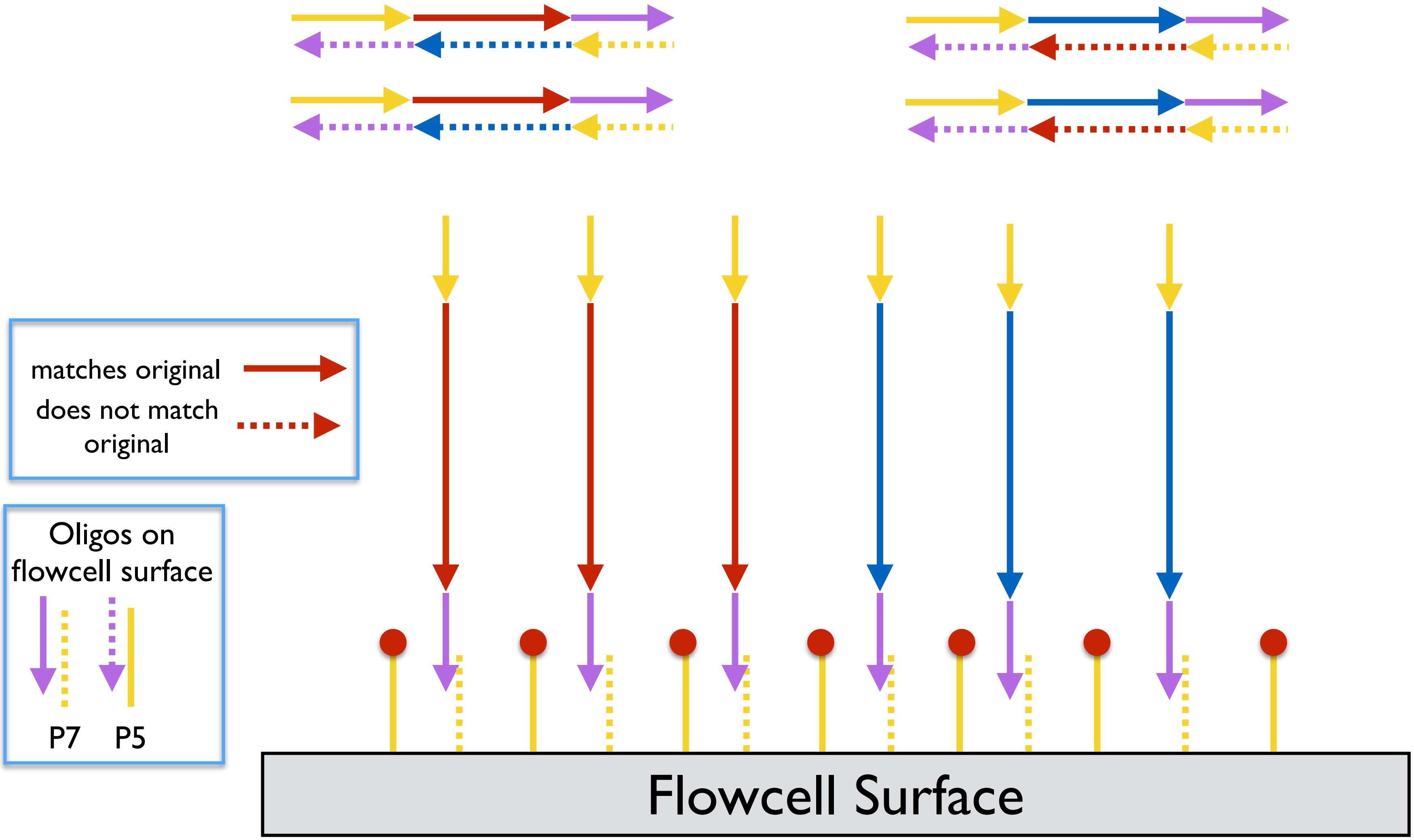
Block the p5 oligos

Illumina Sequencing





Illumina Sequencing



matches original

does not match original

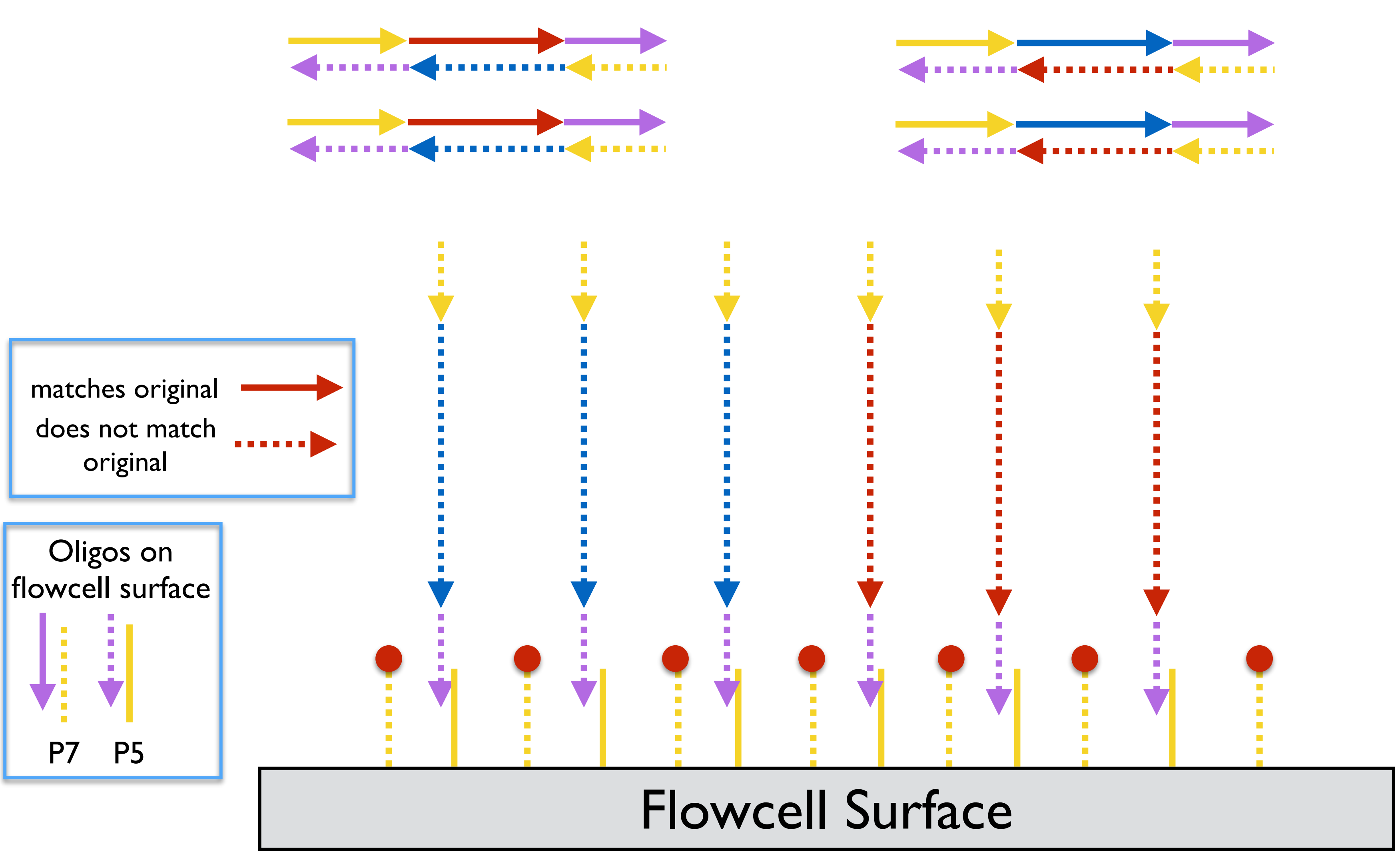
Oligos on flowcell surface

P7 P5

Flowcell Surface

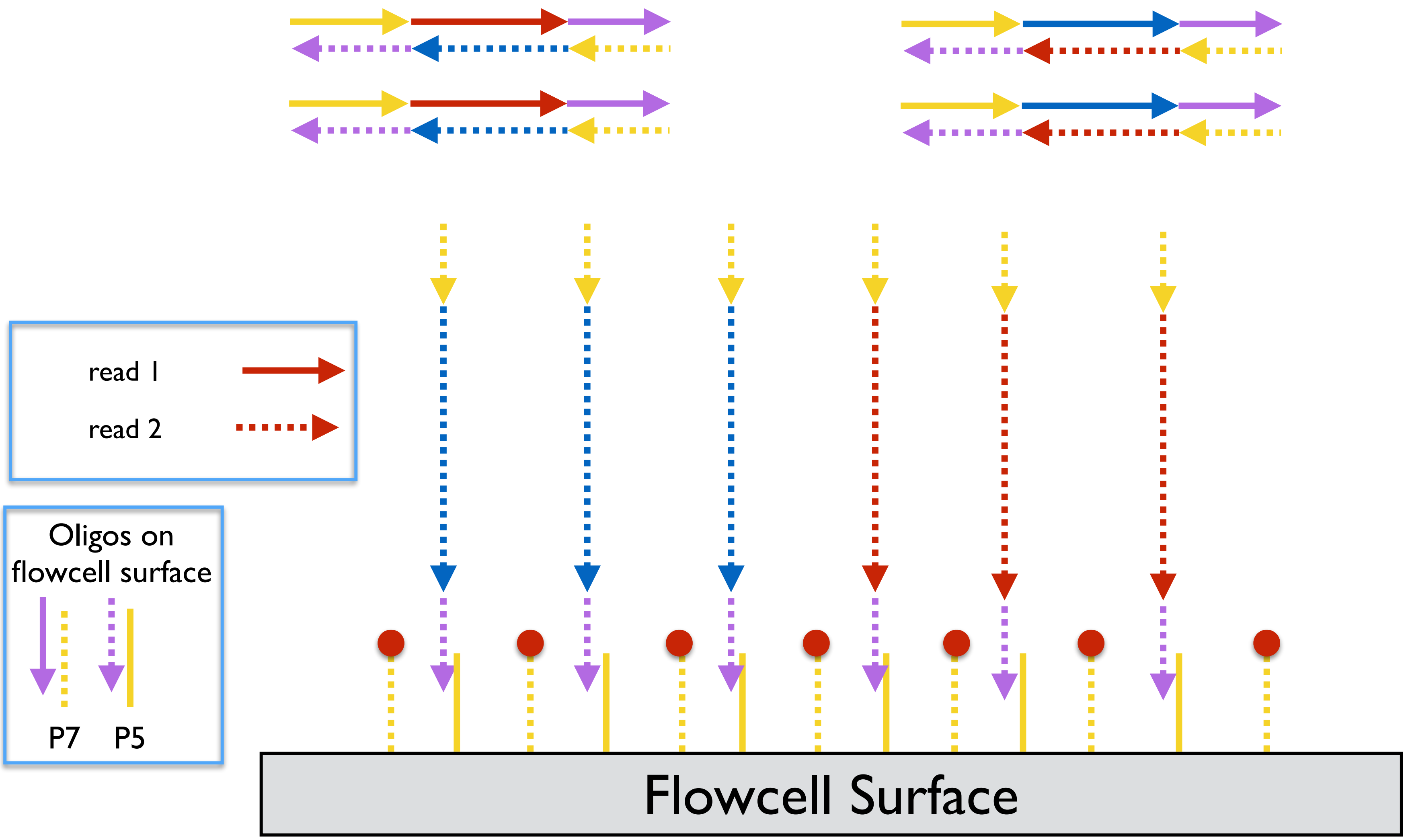
Read 1: use solid adapter

Illumina Sequencing



Read 2: use dotted adapters (not original adapter seq)

Illumina Sequencing



read 1

read 2

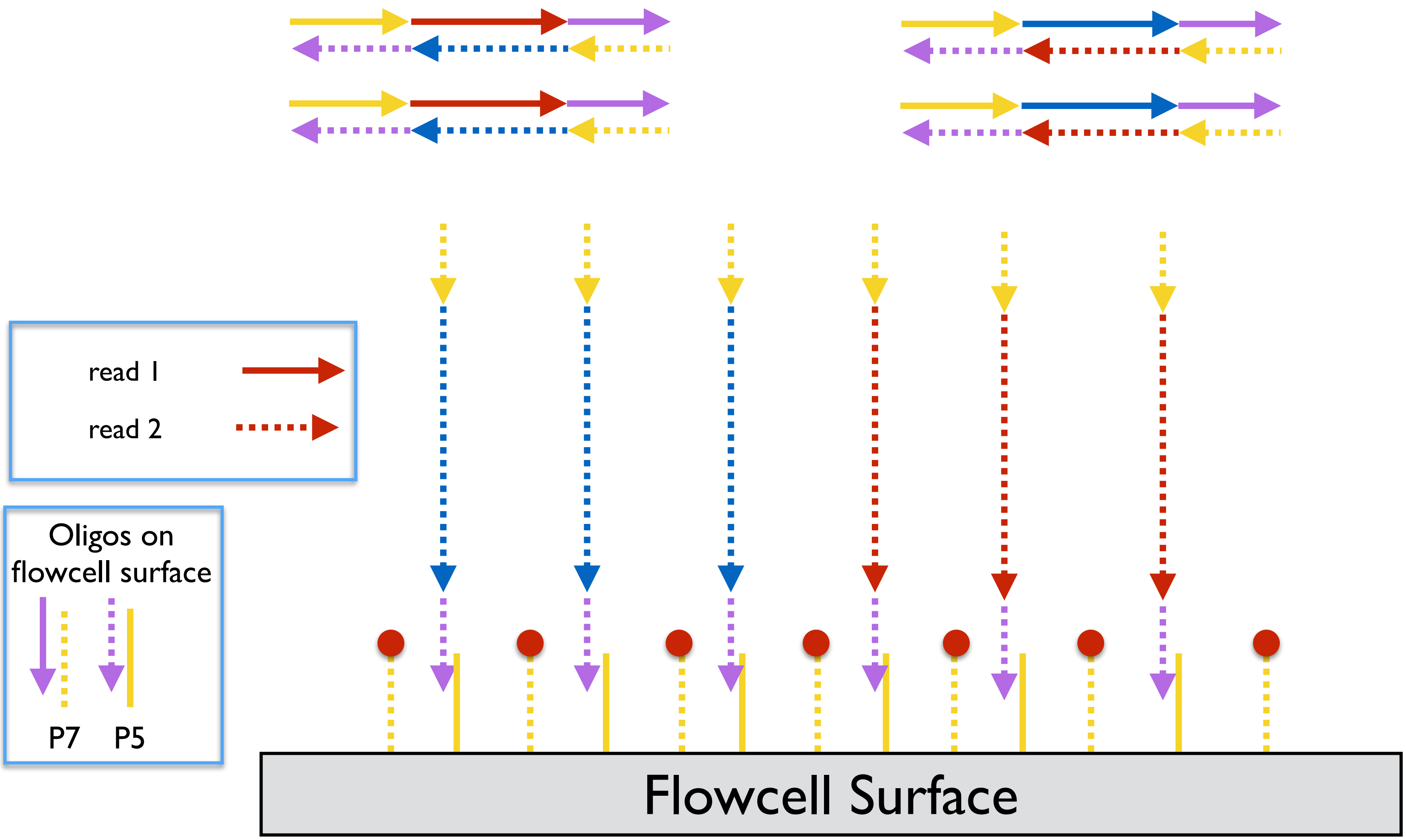
Oligos on flowcell surface

P7 P5

Flowcell Surface

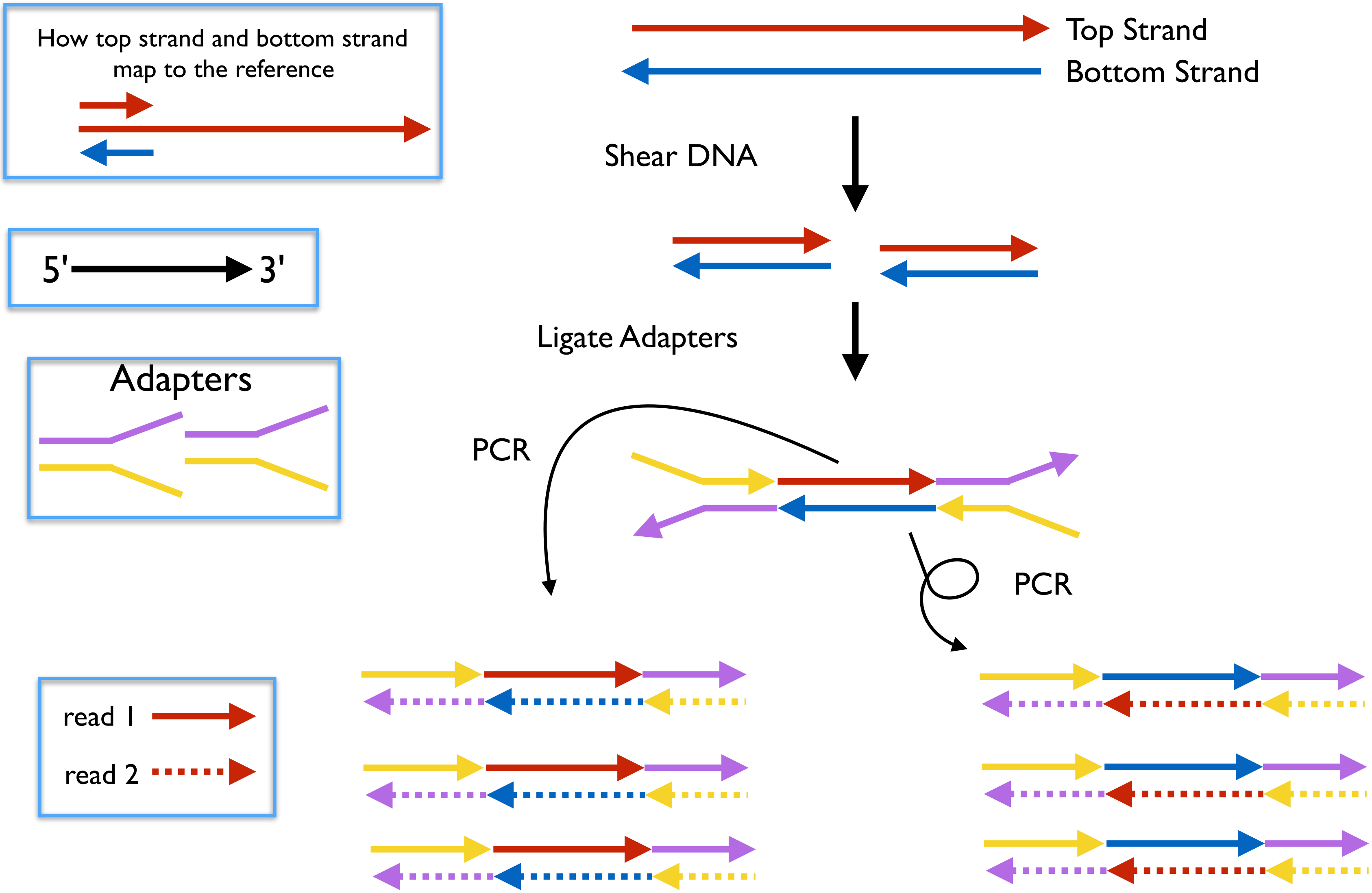
Read 2: use dotted adapters (not original adapter seq)

Illumina Sequencing



Aren't ↓ and ⇣ identical?

Recall...



How top strand and bottom strand map to the reference

5' 3'

Adapters

read 1

read 2

Top Strand

Bottom Strand

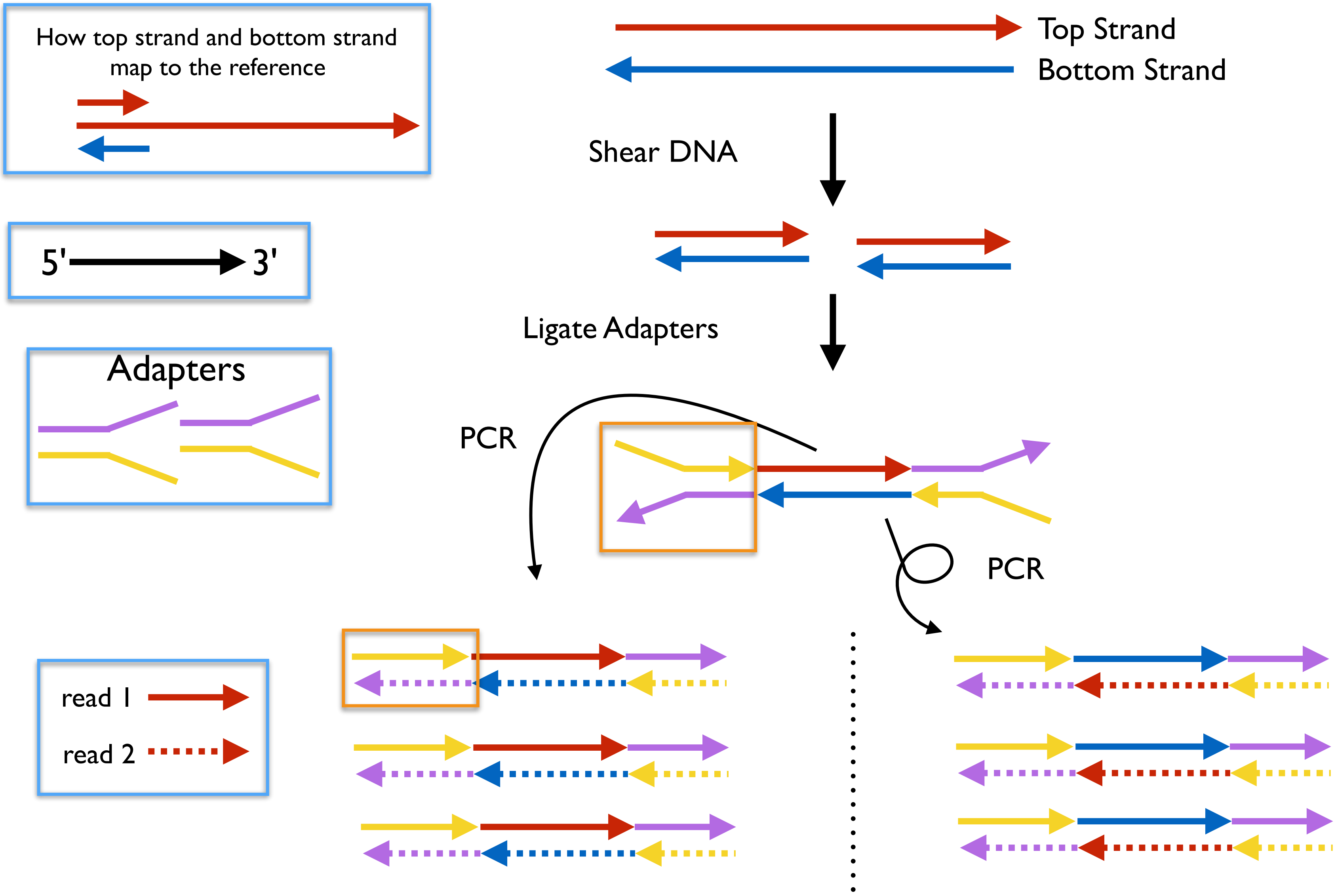
Shear DNA

Ligate Adapters

PCR

PCR

Recall...forked adapters



How top strand and bottom strand map to the reference

5' → 3'

Adapters

read 1

read 2

Top Strand

Bottom Strand

Shear DNA

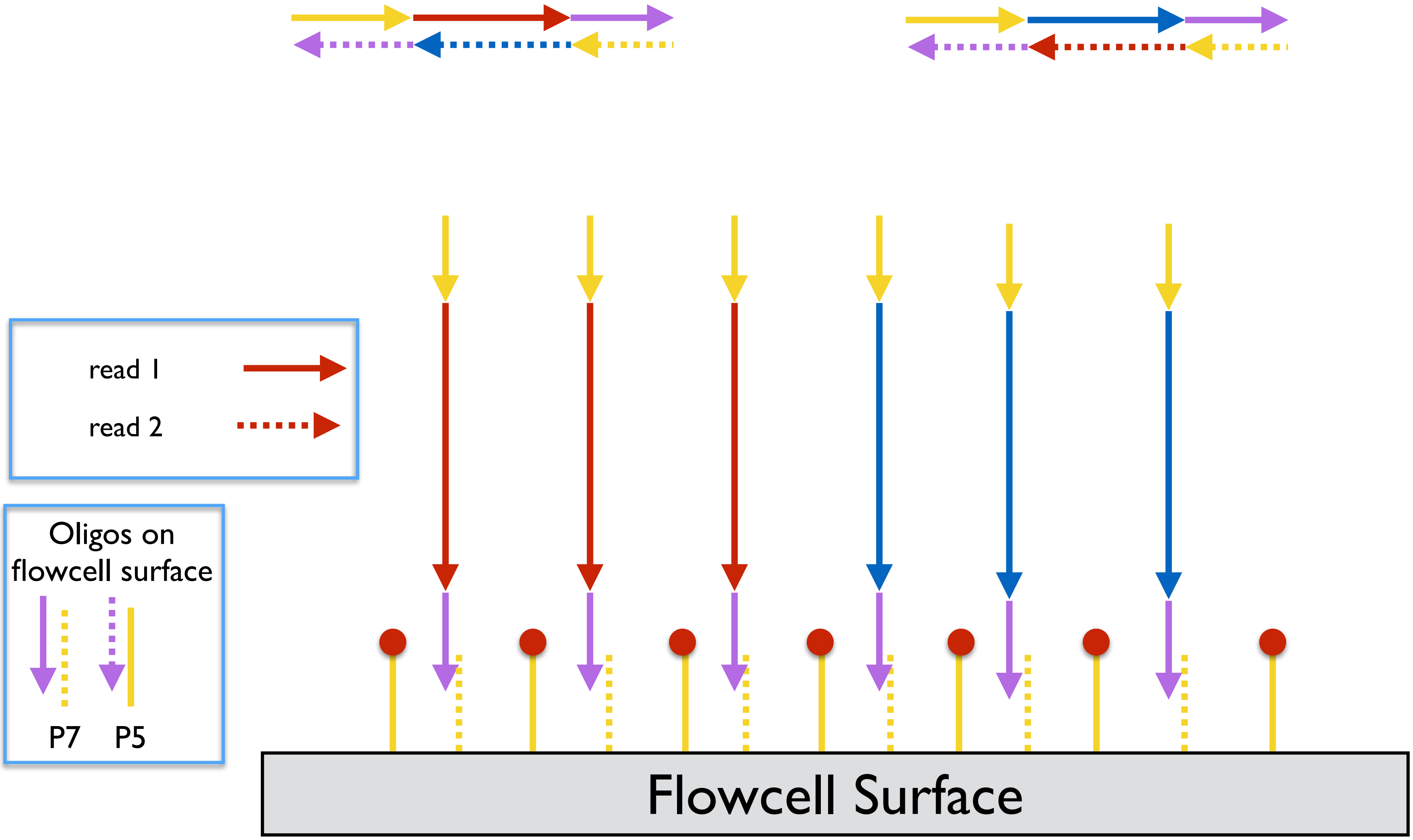
Ligate Adapters

PCR

PCR

Forked adapters: → and ⇠ are different

Illumina Sequencing



read 1

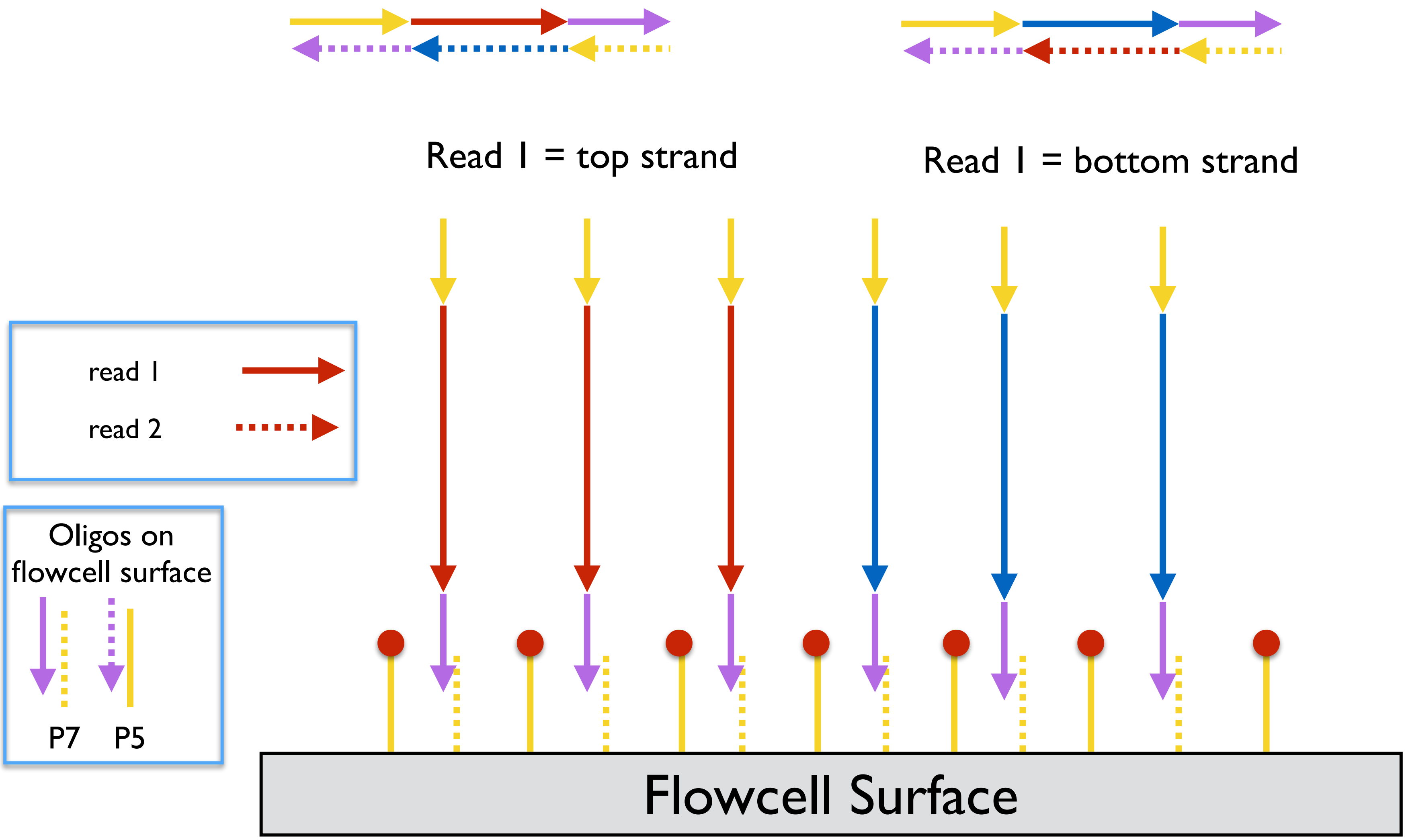
read 2

Oligos on flowcell surface

P7 P5

Flowcell Surface

Illumina Sequencing



Read 1 = top strand

Read 1 = bottom strand

read 1

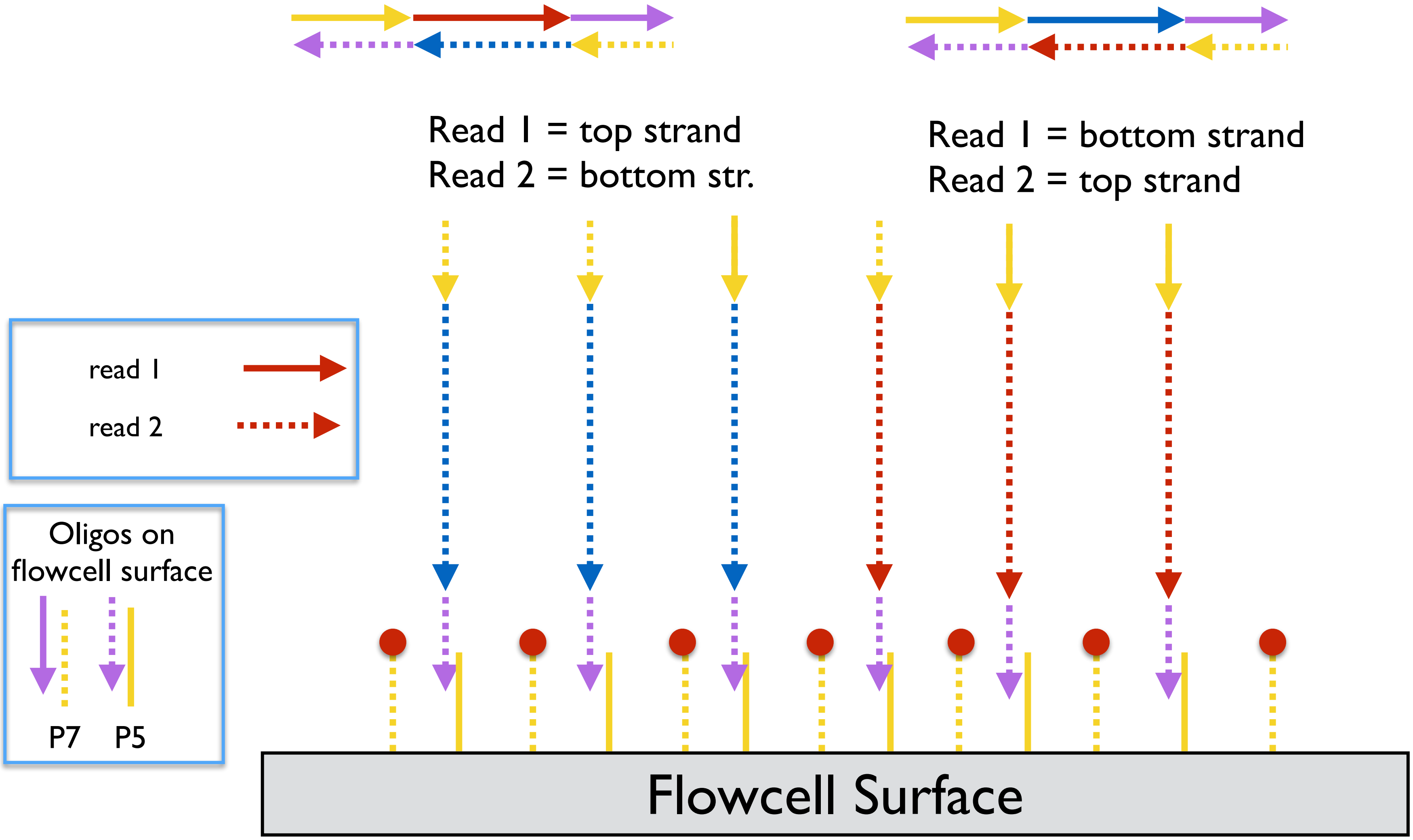
read 2

Oligos on flowcell surface

P7 P5

Flowcell Surface

Illumina Sequencing



Read 1 = top strand
Read 2 = bottom str.

Read 1 = bottom strand
Read 2 = top strand

read 1

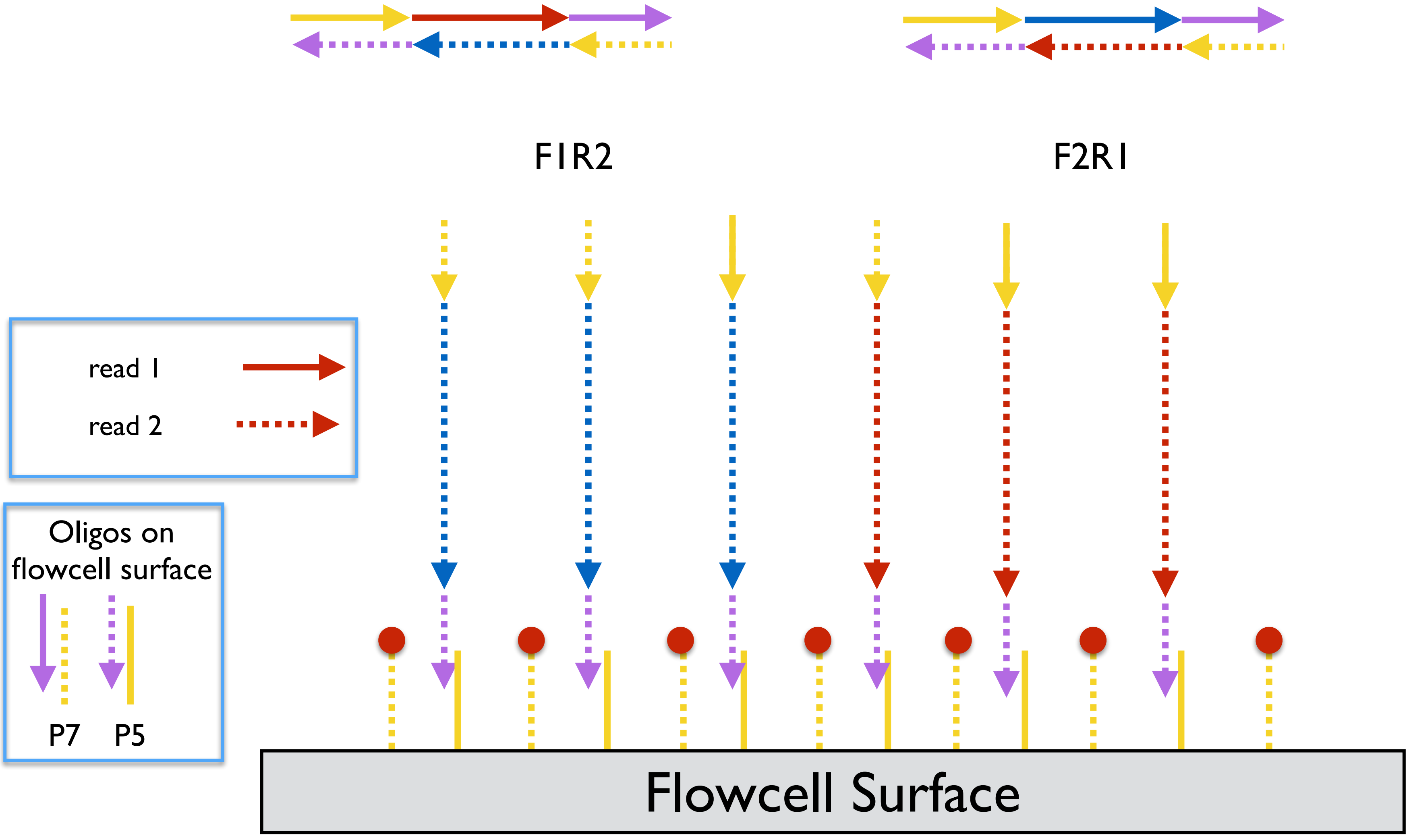
read 2

Oligos on flowcell surface

P7 P5

Flowcell Surface

Illumina Sequencing



F1R2

F2R1

read 1

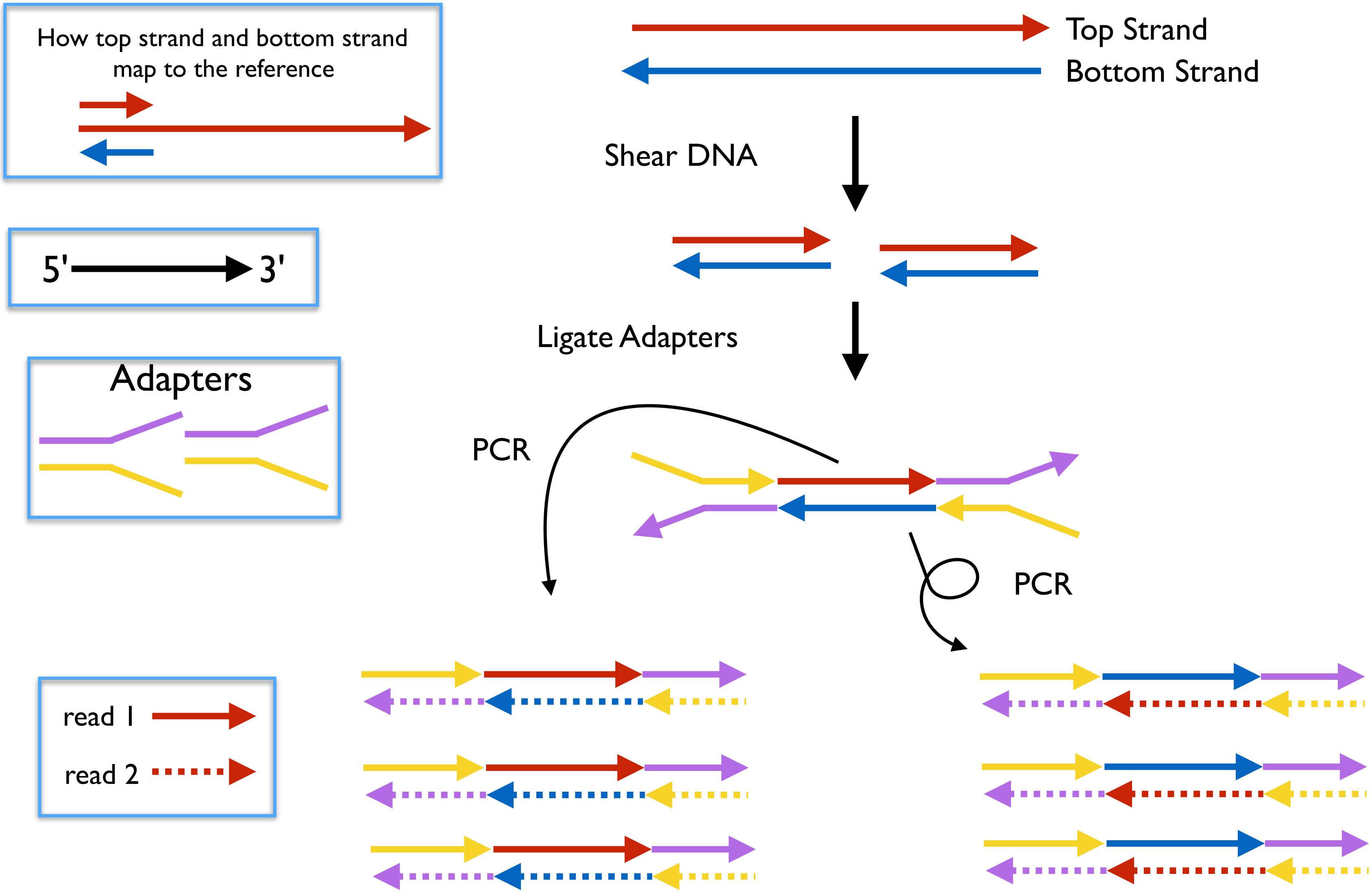
read 2

Oligos on flowcell surface

P7 P5

Flowcell Surface

What is read orientation



How top strand and bottom strand map to the reference

5' 3'

Adapters

read 1

read 2

Top Strand

Bottom Strand

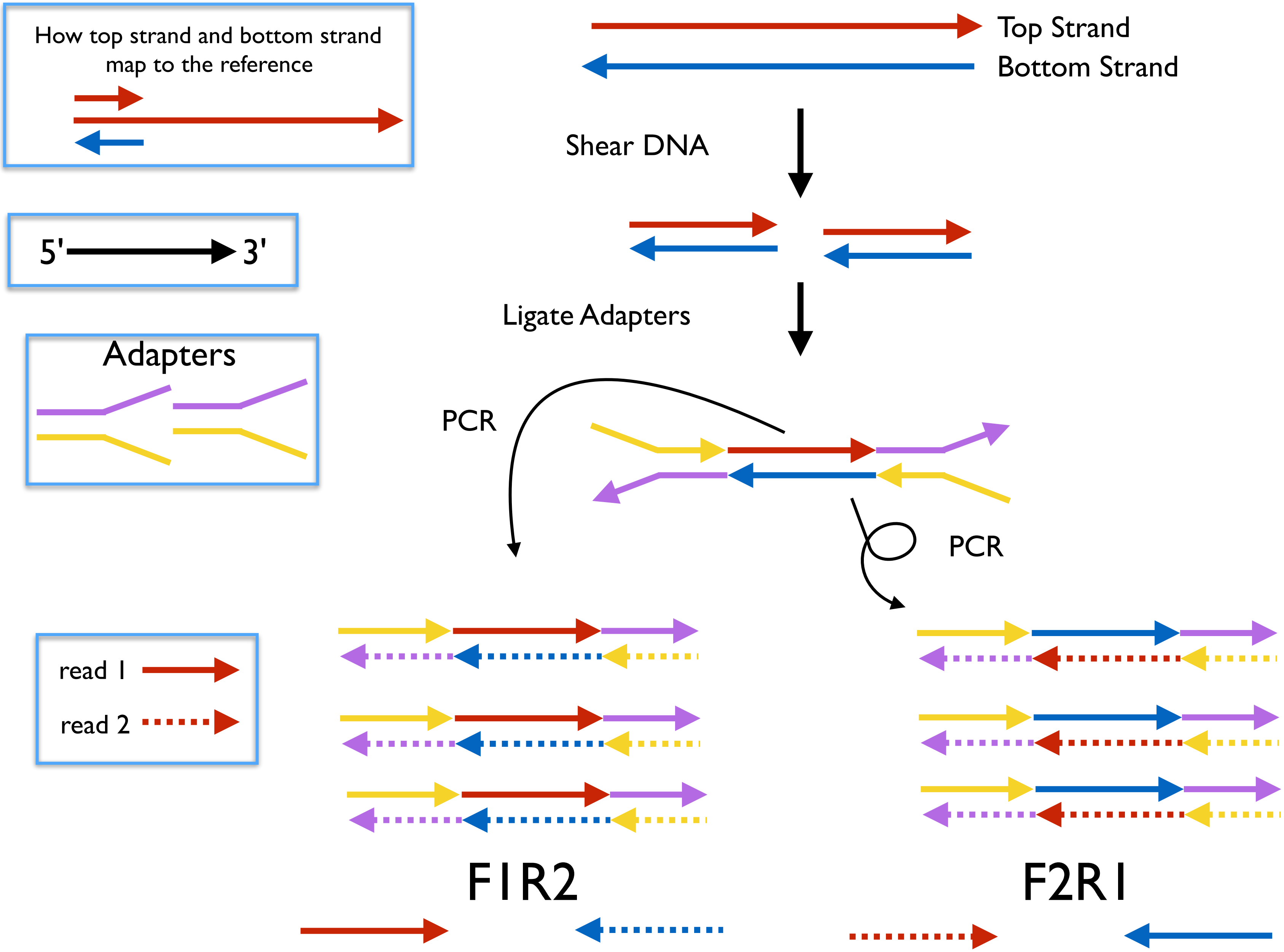
Shear DNA

Ligate Adapters

PCR

PCR

What is read orientation



How top strand and bottom strand map to the reference

5' 3'

Adapters

read 1

read 2

Top Strand

Bottom Strand

Shear DNA

Ligate Adapters

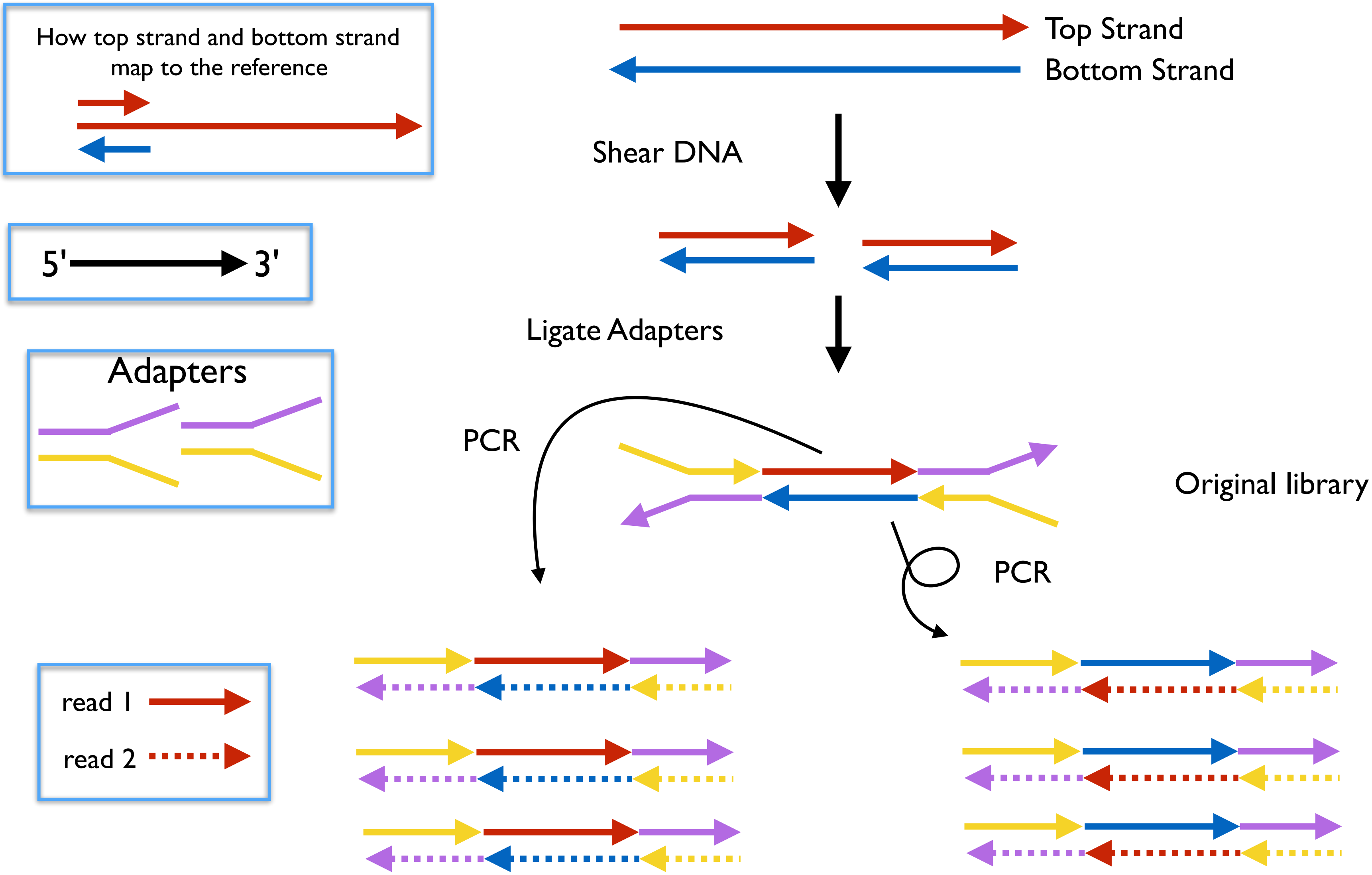
PCR

PCR

F1R2

F2R1

What is read orientation



How top strand and bottom strand map to the reference

5' 3'

Adapters

read 1

read 2

Top Strand

Bottom Strand

Shear DNA

Ligate Adapters

PCR

Original library

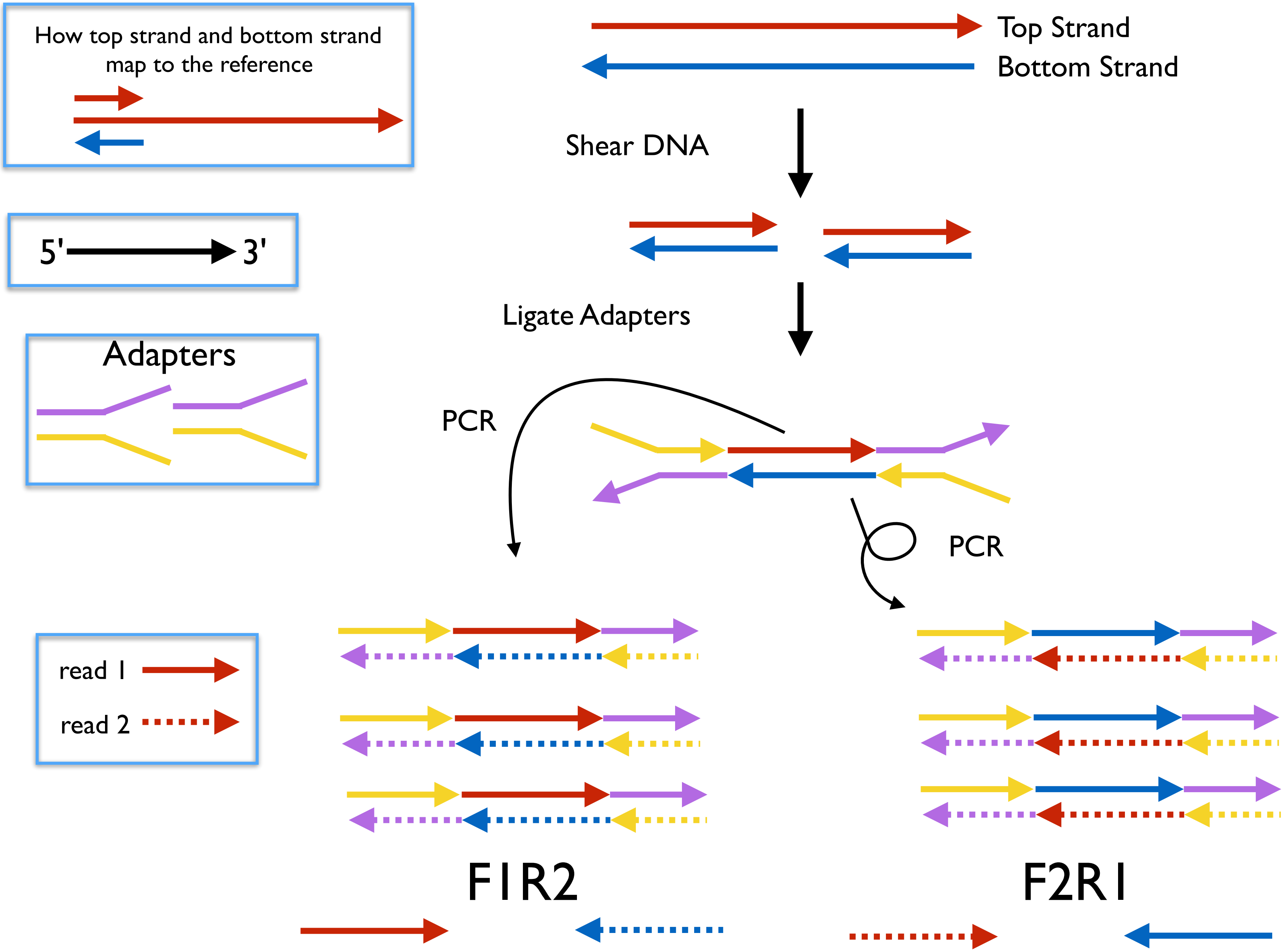
PCR

Read orientation tells us which strand (top or bottom) in the read came from original library

Read orientation artifact

- Alt bases are *only* in F1R2 (i.e. ones on the left)
- or *only* in F2R1 (ones on the right)
- When might we see an artifact like this?

What is read orientation



How top strand and bottom strand map to the reference

5' 3'

Adapters

read 1

read 2

Top Strand

Bottom Strand

Shear DNA

Ligate Adapters

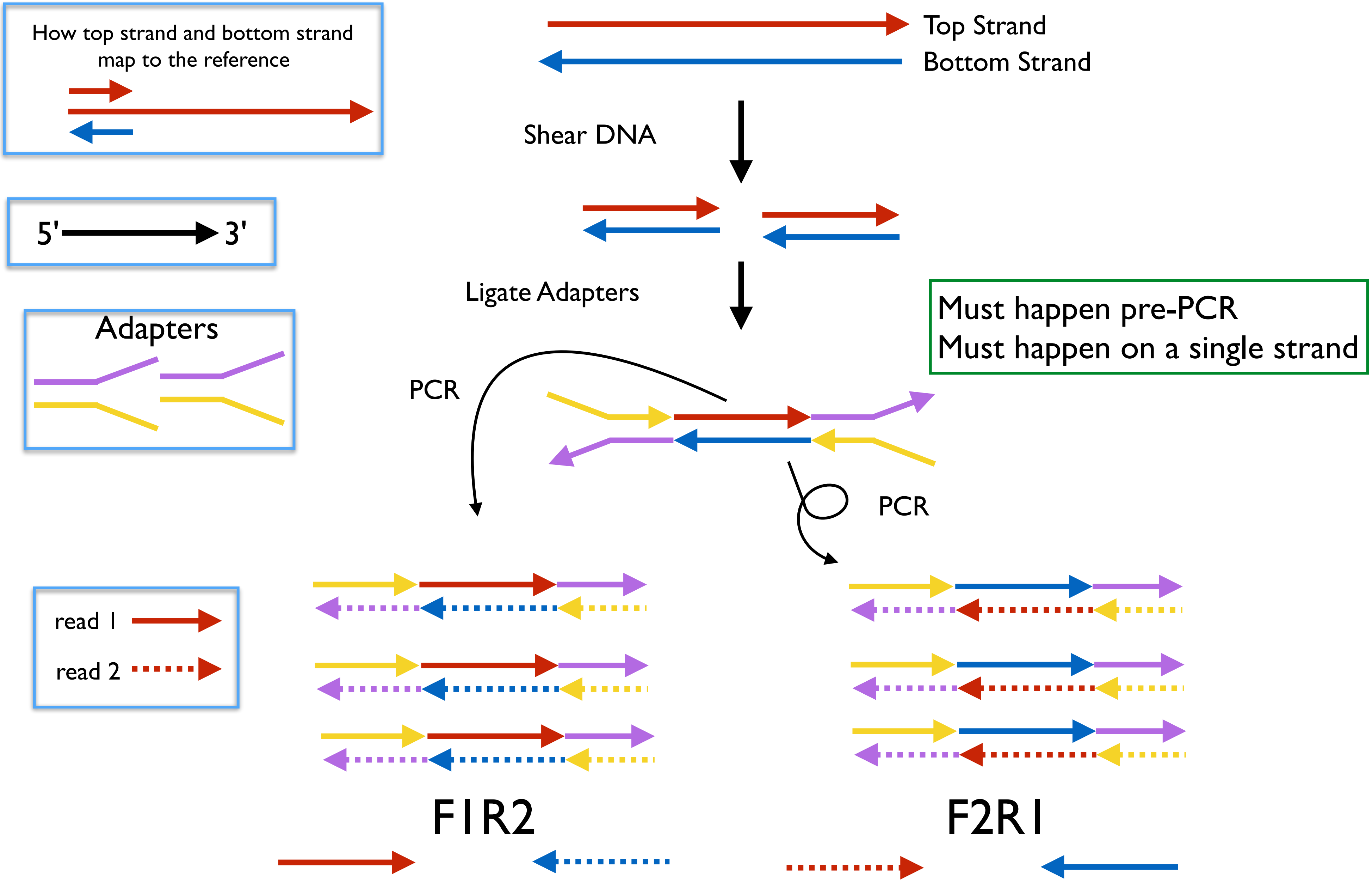
PCR

PCR

F1R2

F2R1

What is read orientation



How top strand and bottom strand map to the reference

5' 3'

Adapters

read 1

read 2

Top Strand

Bottom Strand

Shear DNA

Ligate Adapters

PCR

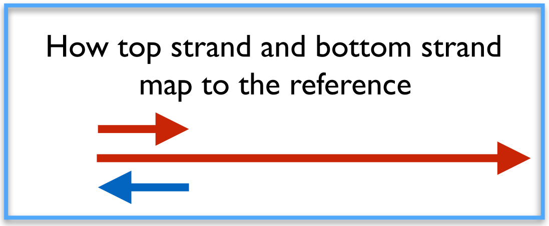
Must happen pre-PCR
Must happen on a single strand

PCR

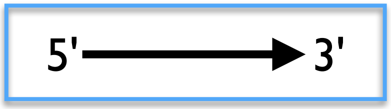
F1R2

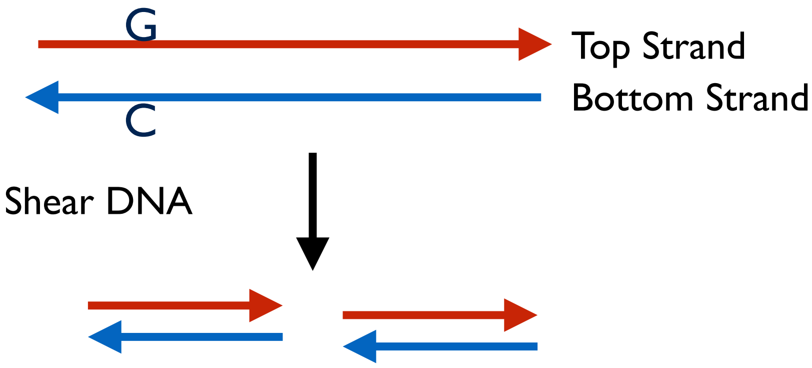
F2R1

Example: G -> T single-stranded artifact

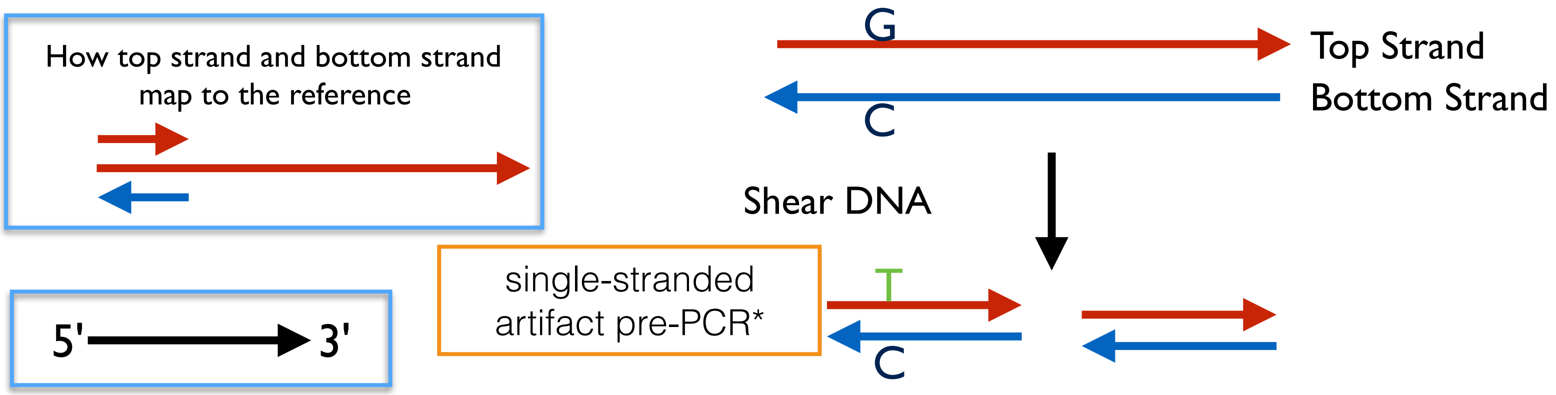


How top strand and bottom strand map to the reference





Example: G -> T single-stranded artifact

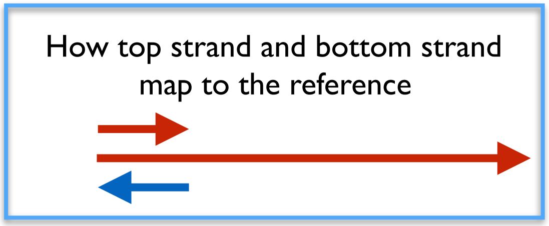


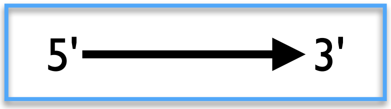
How top strand and bottom strand map to the reference

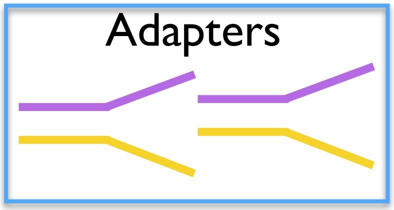
single-stranded artifact pre-PCR*

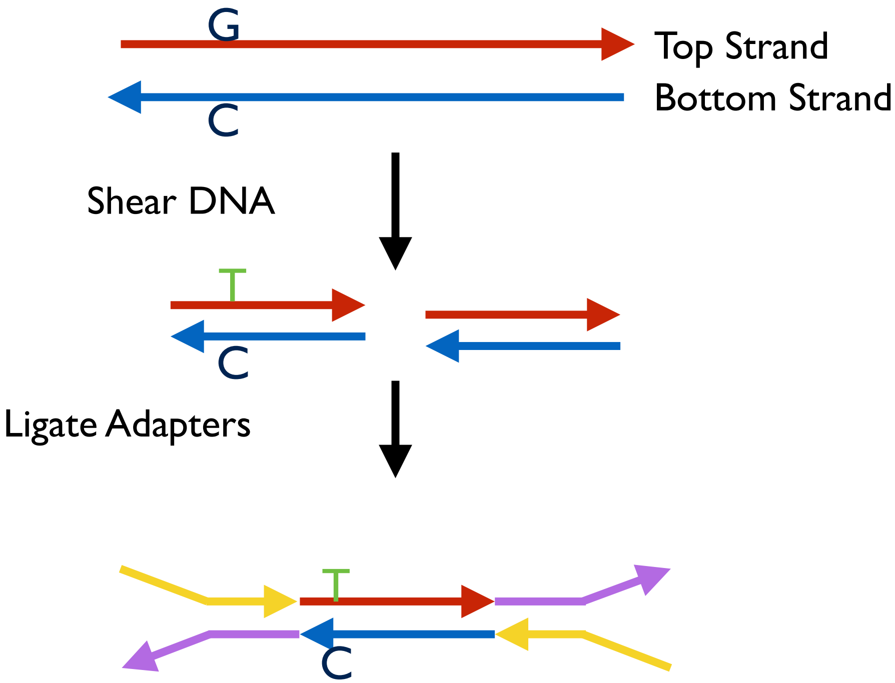
More precisely, G oxidizes and becomes oxo-G (G), which has high affinity for A i.e. G* acts like a T

Example: G -> T single-stranded artifact


How top strand and bottom strand map to the reference




Adapters

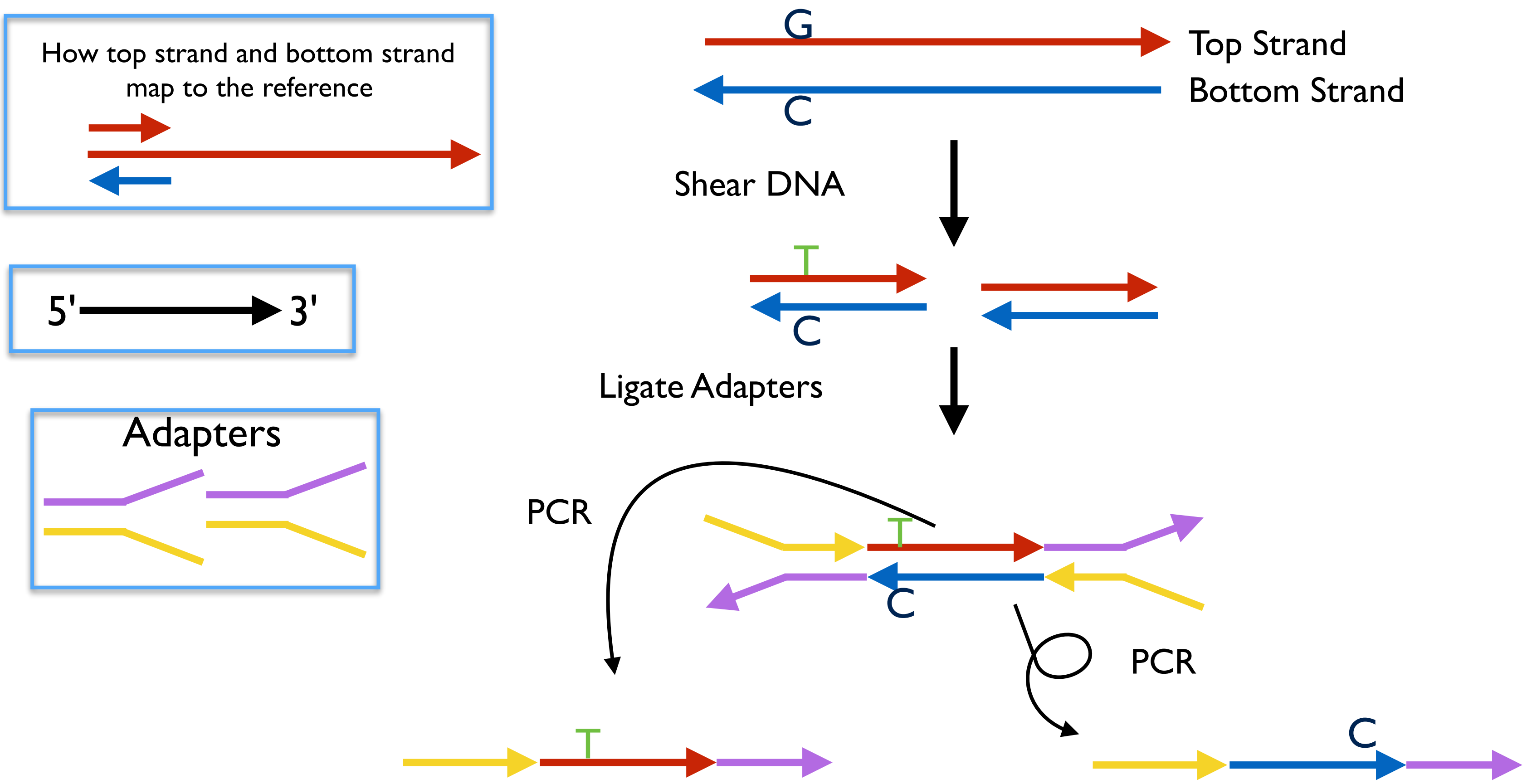

Top Strand

Bottom Strand

Shear DNA

Ligate Adapters

Example: G -> T single-stranded artifact



How top strand and bottom strand map to the reference

5' 3'

Adapters

Top Strand

Bottom Strand

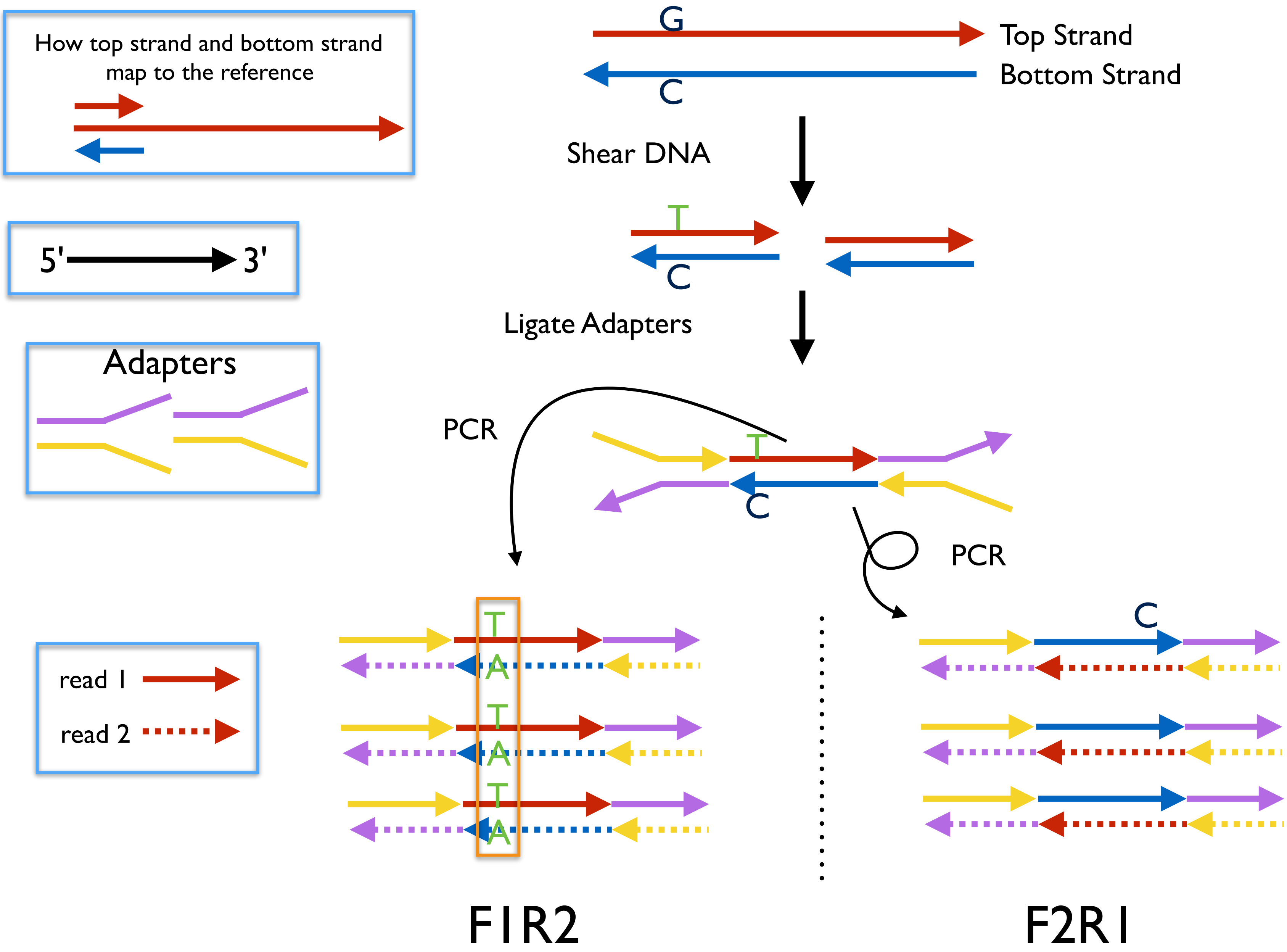
Shear DNA

Ligate Adapters

PCR

PCR

Example: G -> T single-stranded artifact



How top strand and bottom strand map to the reference

Adapters

read 1

read 2

Top Strand

Bottom Strand

Shear DNA

Ligate Adapters

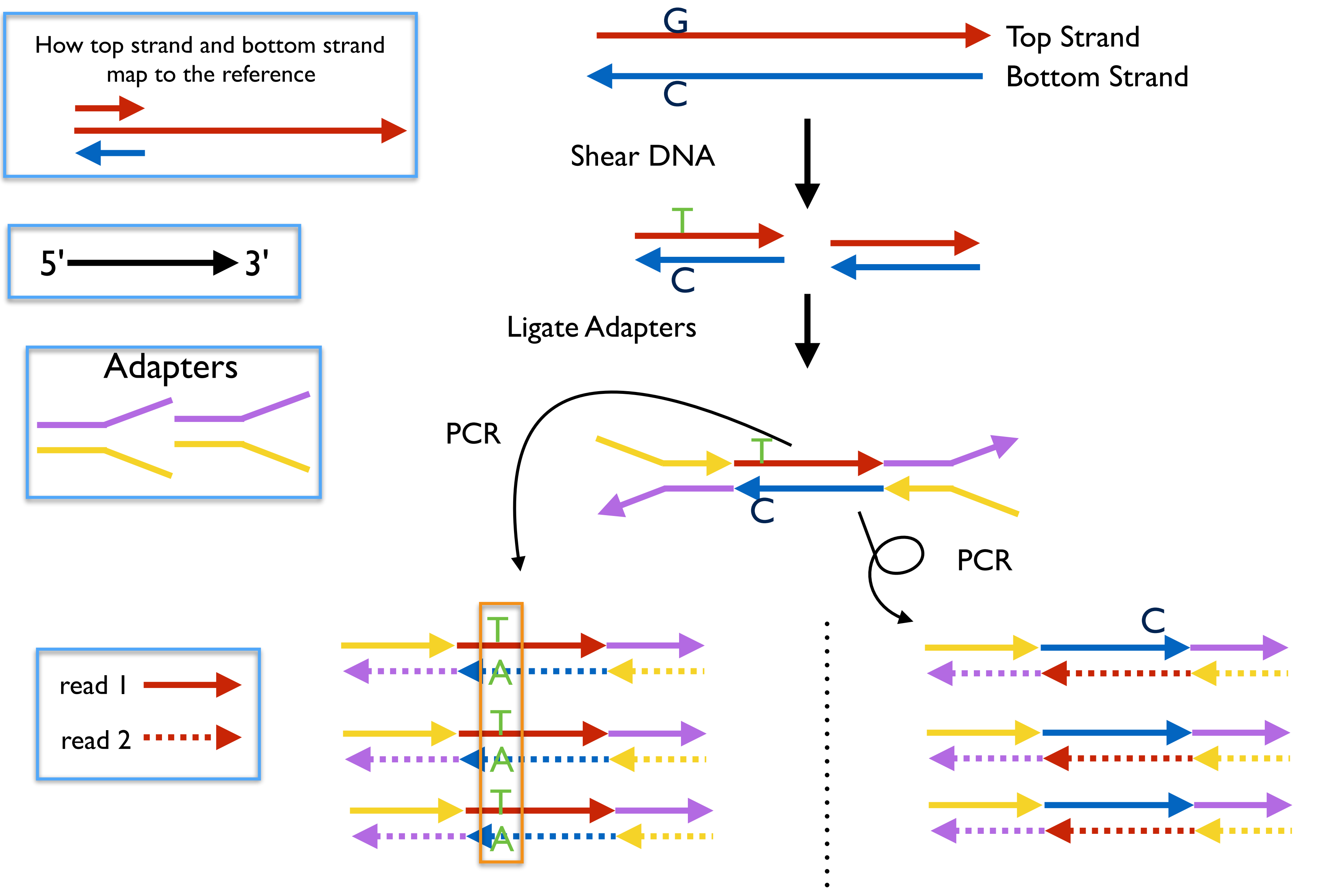
PCR

PCR

F1R2

F2R1

Example: G -> T single-stranded artifact



How top strand and bottom strand map to the reference

5' 3'

Adapters

read 1

read 2

Top Strand

Bottom Strand

Shear DNA

Ligate Adapters

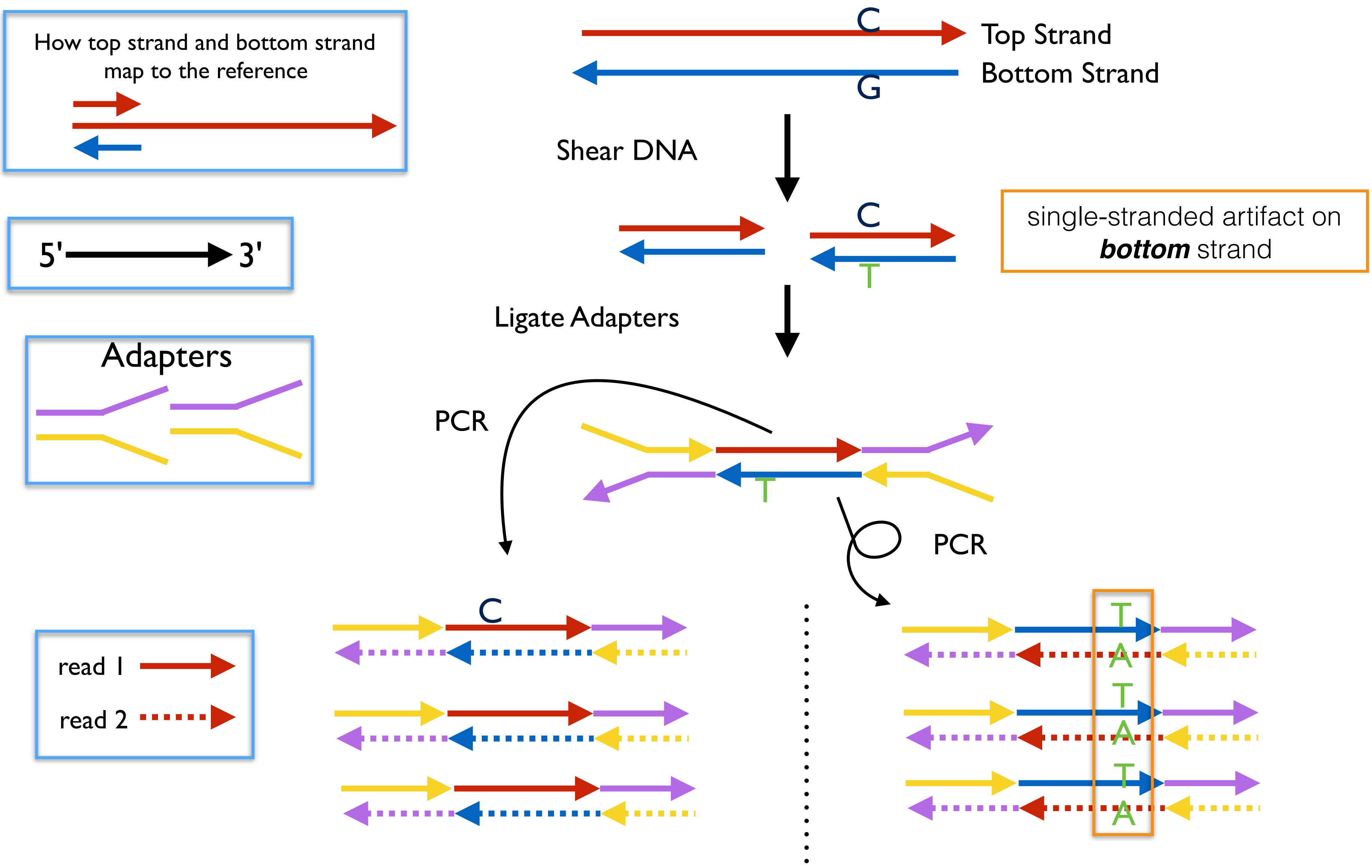
PCR

PCR

F1R2, G to T

F2R1, ref

Example: G -> T single-stranded artifact



How top strand and bottom strand map to the reference

Top Strand

Bottom Strand

Shear DNA

single-stranded artifact on ***bottom*** strand

Ligate Adapters

Adapters

PCR

PCR

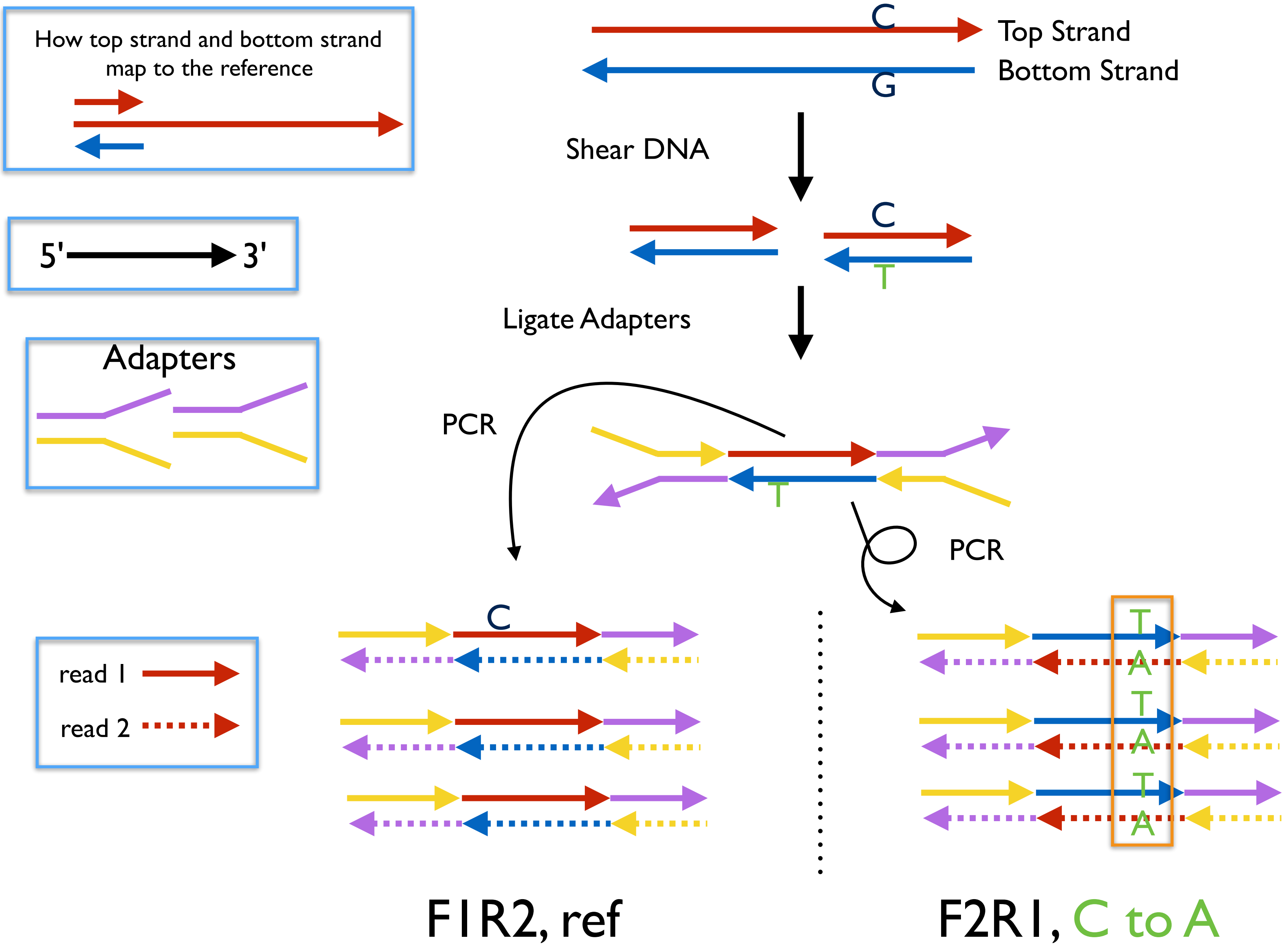
read 1

read 2

F1R2, ref

F2R1 ???

Example: G -> T single-stranded artifact



How top strand and bottom strand map to the reference

5' 3'

Adapters

read 1

read 2

Top Strand

Bottom Strand

Shear DNA

Ligate Adapters

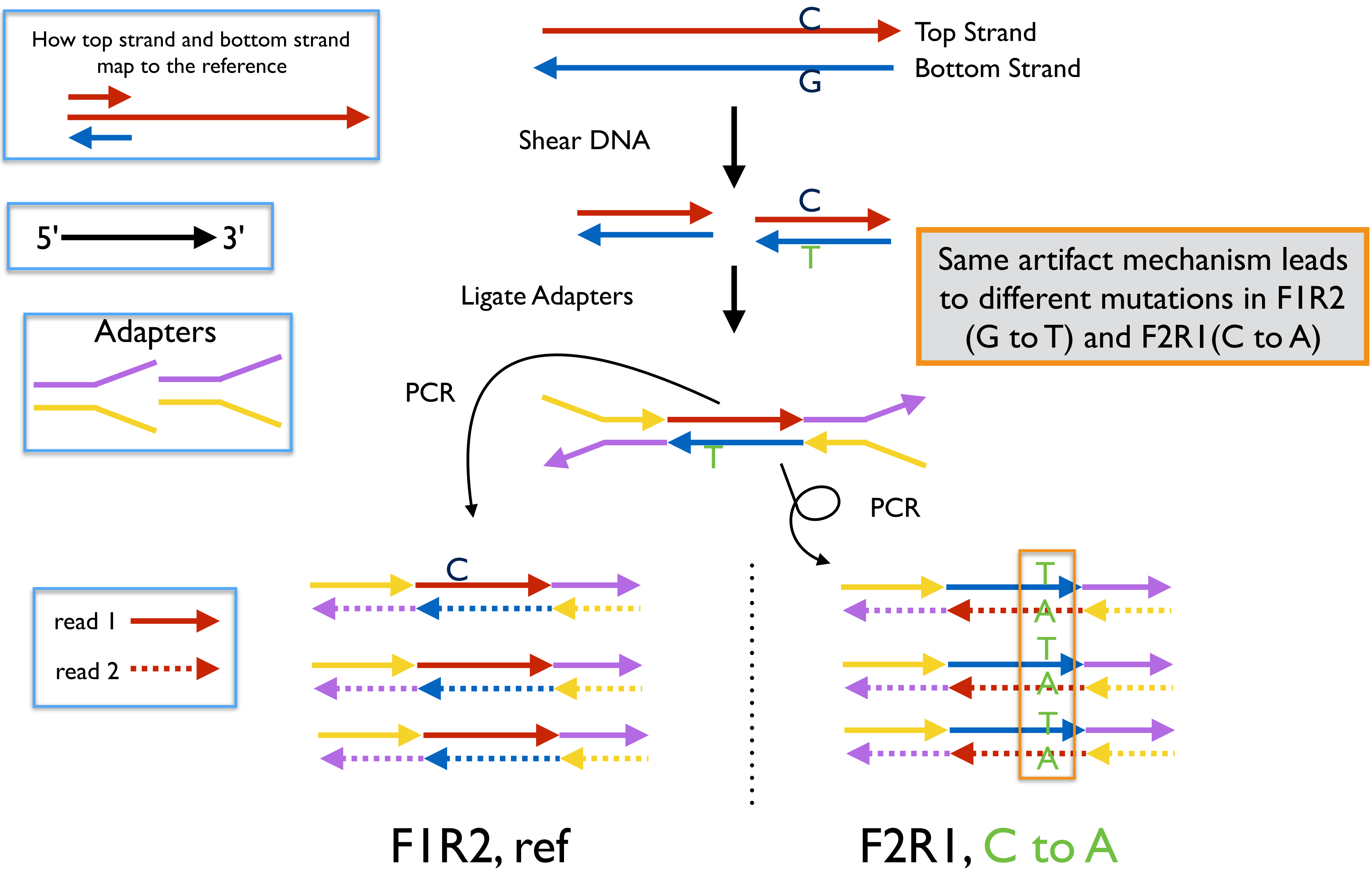
PCR

PCR

F1R2, ref

F2R1, C to A

Example: G -> T single-stranded artifact



How top strand and bottom strand map to the reference

5' 3'

Adapters

read 1

read 2

Top Strand

Bottom Strand

Shear DNA

Ligate Adapters

PCR

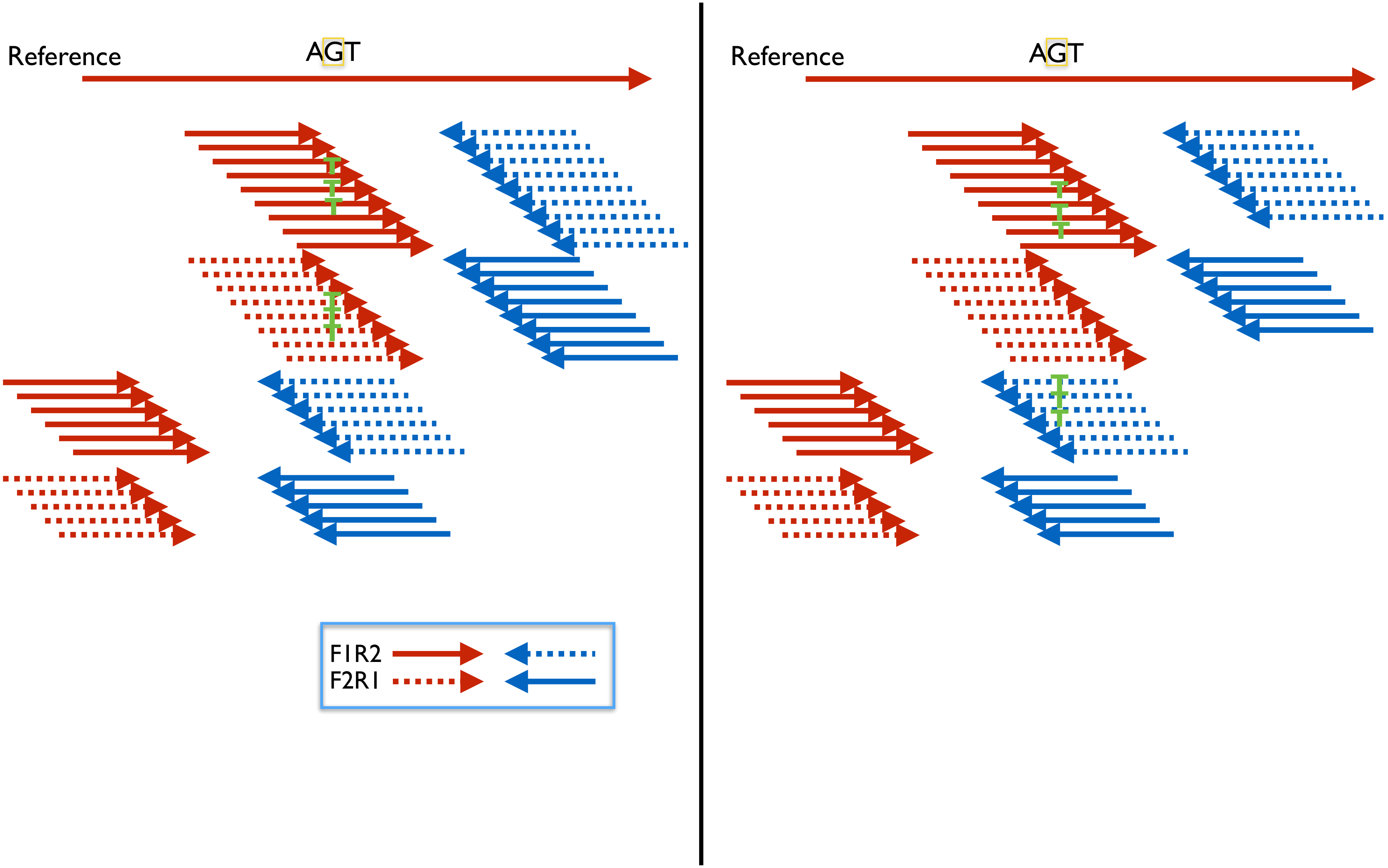
PCR

Same artifact mechanism leads to different mutations in F1R2 (G to T) and F2R1(C to A)

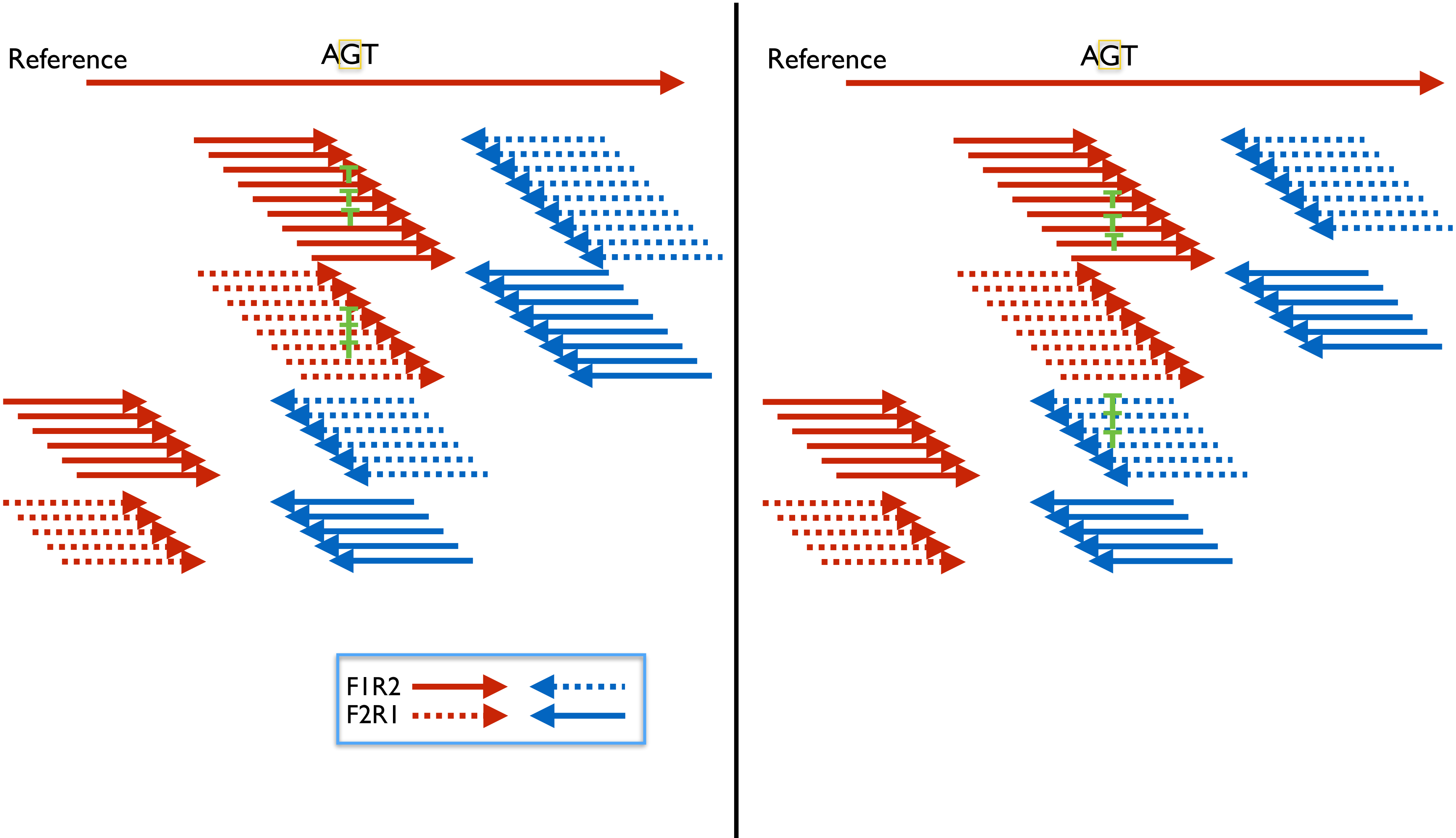
F1R2, ref

F2R1, C to A

strand artifact != read orientation artifact



strand artifact != read orientation artifact

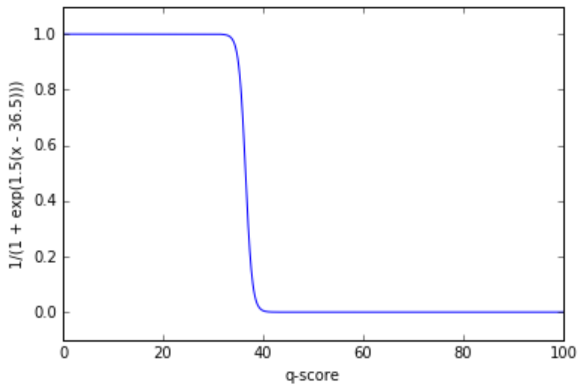


Strand Artifact

Read Orientation Artifact

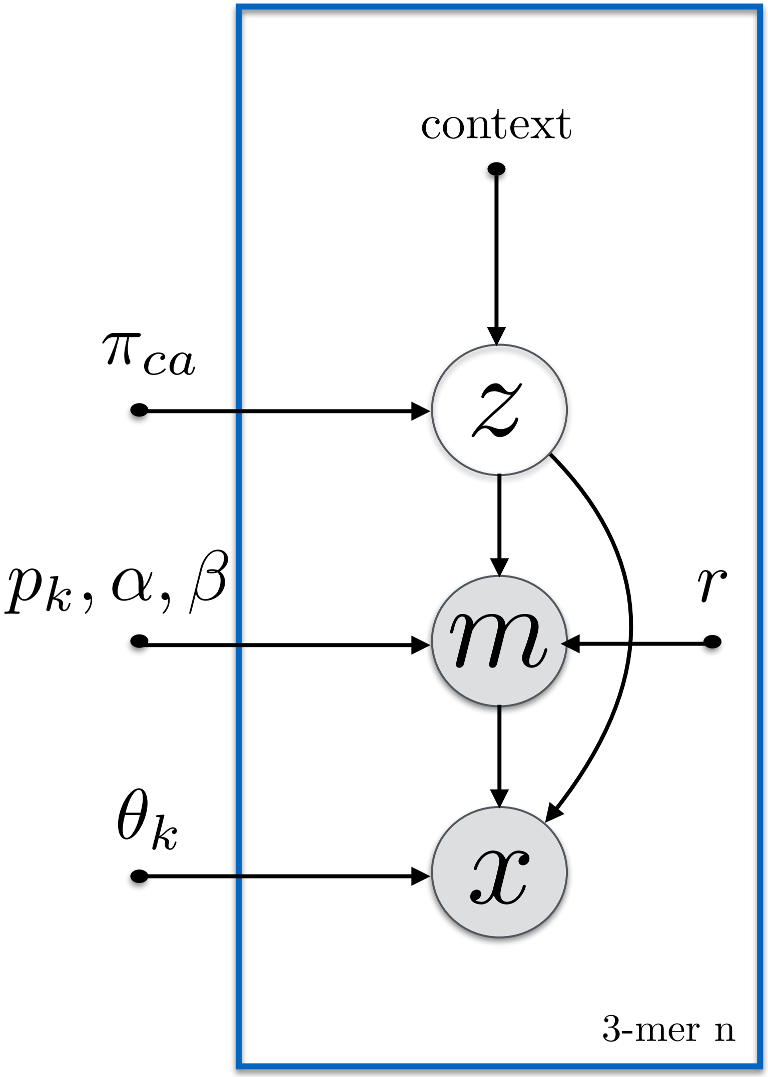
Existing filter

- Walk through bam, for each (ref context*, allele) pair, Q = $\text{qscore} = \text{phred}(\frac{\sum_{\text{sites}} \#\ \text{Alt F1R2} - \#\ \text{Alt F2R1}}{\sum_{\text{sites}} \text{Alt depth}})$
- Higher Q = no artifact
- For each variant,
  - compute p-value with null hypothesis: # alt F1R2 reads ~ Binom(alt depth, 0.96)
  - False discovery = falsely call an artifact as non-artifact...
  - Reject non-artifact sites, Benjamini-Hochberg procedure to control FDR...or something
- Multiply # variants to filter by $\text{supressor} = \frac{1}{1 + e^{1.5(Q - 36.5)}}$
  - does not generalize
  - cannot detect rare events
- Must specify the artifact mode e.g. A -> T by hand
  - Requires manual inspection of collect sequencing metrics file
- Upshot - we need a new model



*ref context = 3-mer in reference

New Read Orientation Filter Model v4

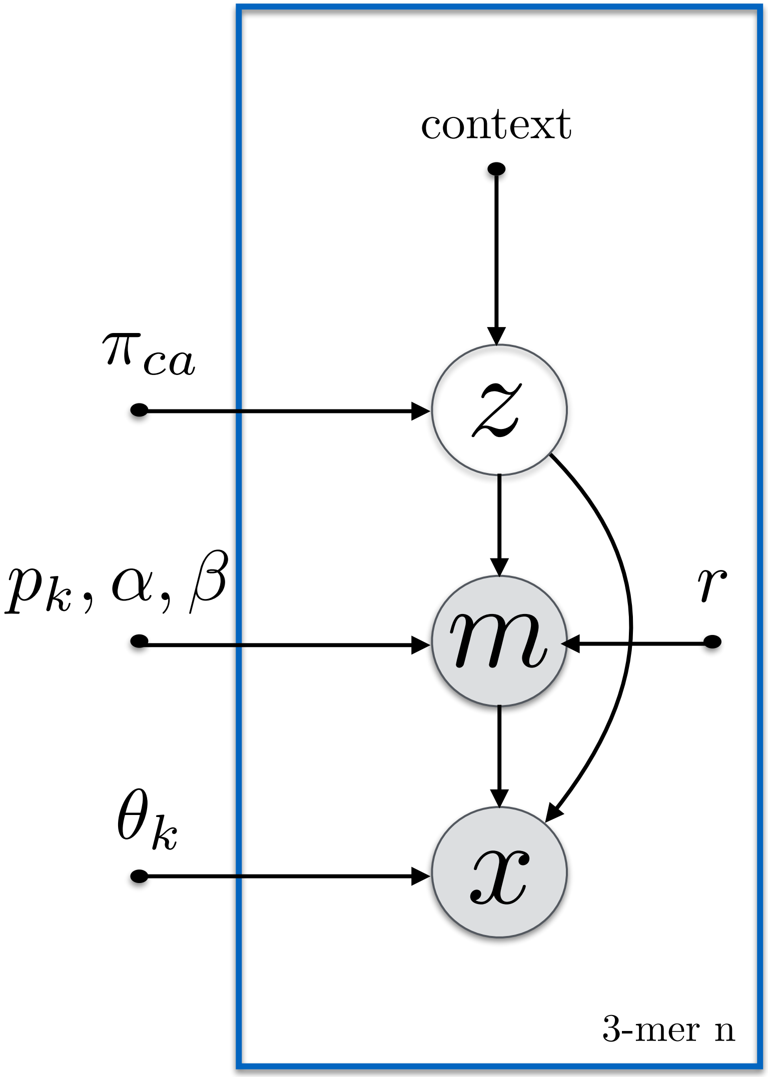


$z \in \{\text{F1R2}_a, \text{ F2R1}_a, \text{ Hom Ref, Somatic Het, Germline Het Hom Var}\}$

where $a \in \mathbb{A}$ and $\mathbb{A}$ is a set of possible alt alleles in context c

$$
\begin{aligned}
z &\sim \text{Categorical}(\pi_{ca}) \\
r &\equiv \text{depth} \\
m &\equiv \text{alt depth} \\
m|z &\sim \begin{Bmatrix} \text{BetaBinom}(p_k, r, \alpha, \beta), & z = \text{Somatic Het} \\ \text{Binom}(p_k, r) & \text{otherwise} \end{Bmatrix} \\
x &\equiv \text{alt F1R2 depth} \\
x|z, m &\sim \text{Binom}(\theta_k, m)
\end{aligned}
$$

New Read Orientation Filter Model v4



artifact non artifact (F1R2 & F2R1 balanced)

$z \in \{\text{F1R2}_a, \text{ F2R1}_a, \text{ Hom Ref, Somatic Het, Germline Het Hom Var}\}$

where $a \in \mathbb{A}$ and $\mathbb{A}$ is a set of possible alt alleles in context c

$$z \sim \text{Categorical}(\pi_{ca})$$

$$r \equiv \text{depth}$$

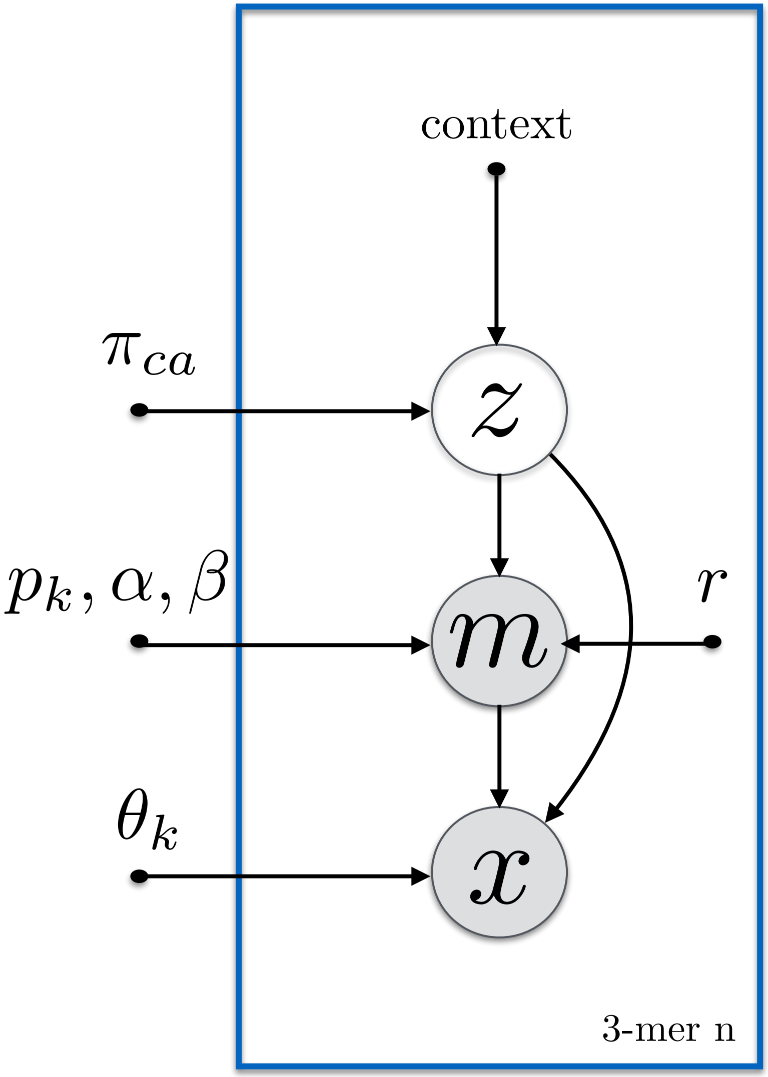
$$m \equiv \text{alt depth}$$

$$m|z \sim \left\{\begin{matrix} \text{BetaBinom}(p_k, r, \alpha, \beta), & z = \text{Somatic Het} \\ \text{Binom}(p_k, r) & \text{otherwise} \end{matrix}\right\}$$

$$x \equiv \text{alt F1R2 depth}$$

$$x|z, m \sim \text{Binom}(\theta_k, m)$$

New Read Orientation Filter Model v4



- Aware of the relative frequency of the artifact under each context
  - e.g. artifact in 1 in 1000 sites, 1 in 10 sites,
- can detect rare events
- No need to manually specify transitions you're looking for
- Simpler interpretation - posterior probabilities of z
- Replace CollectSequencingArtifact Metrics (pending Megan's approval)

New Read Orientation Filter Model v4

1. Learning Step (Name not ready for public announcement)
    1. estimate hyperparameters with EM
2. Inference Step (FilterMutectCalls)
    1. compute p(z=F1R2|data), p(z=F2R1|data)

New Read Orientation Filter Model v5

1. Data Collection Step (Java, LocusWalker)
   1. write out the design matrix to a file
2. Learning Step
   1. estimate hyperparameters with EM, powered by PyMC
3. Inference Step (FilterMutectCalls/Java)
   1. for each variant, compute the posterior probabilities of z, and filter