Transcriptome Assembly

James Pease
14 July 2014

The Most Important Slide

[1] RNA-Seq is an extremely active Research area, stay current!

[2] The principles are general, the specifics will change in less than ~~two~~ one year~~s~~.

[3] Everyone worries about processing, most errors are bad data or analysis

Basic Steps

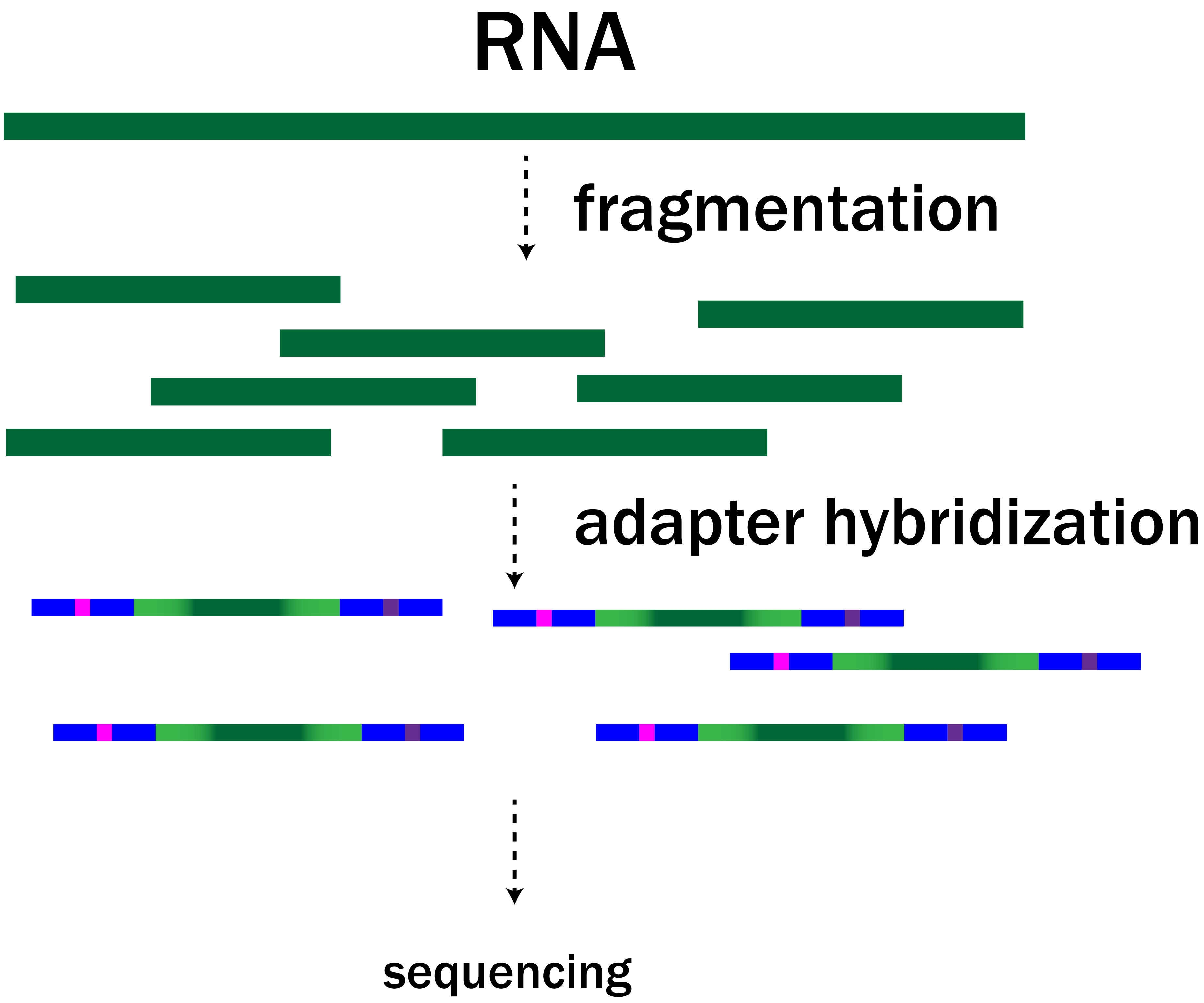
[1] Sequence data quality control

[2] Read trimming and filtering

[3] Assembly

[4] Assembly Quality Control

[1] Sequence data QC



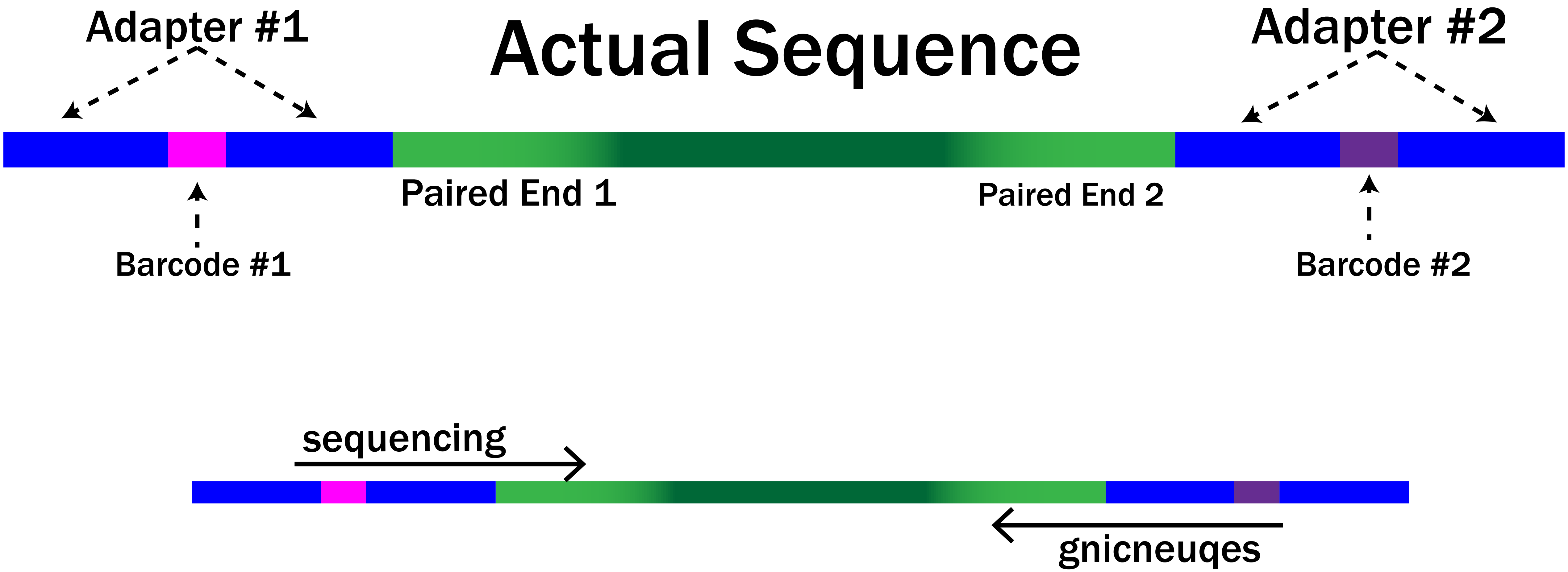
RNA

fragmentation

adapter hybridization

sequencing

[1] Sequence data QC



[1] Sequence data QC

FastQ file(s)

```
@SEQNAME
DNA SEQUENCE
+
QUALITY SCORES
```

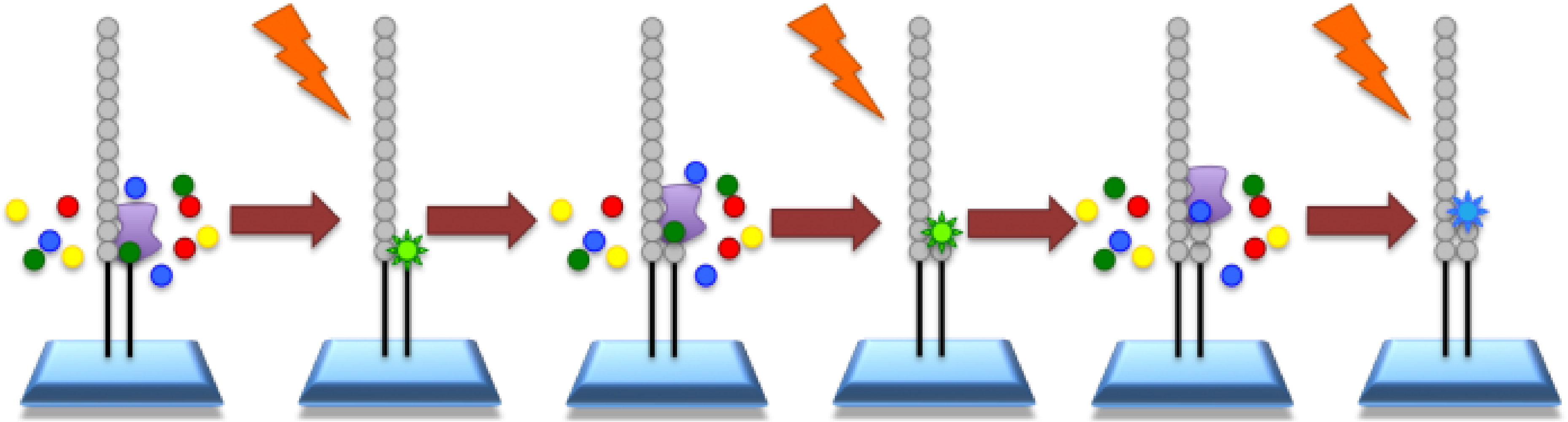
Pairs can come in two files or interleaved

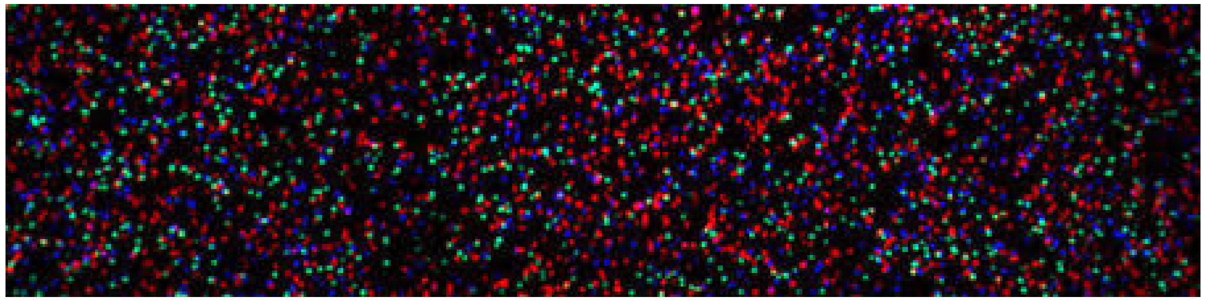
```
@SEQNAME1
GGACGGAGACGACATGATGTGCTGACTGTACTGTNNNNNN
+
F+A42292<A<?1**:??DC@9))8??FDFF*8C######

@SEQNAME2
GATACAACGTACACAA.....
```

[1] Sequence data QC

What is quality?





→ ATGGGCATATTATGCC

What is the probability the wrong base was read by the laser?

[1] Sequence data quality controls

Goal

check that sequencing is both good quality and unbiased

Software: FastQC, et al.

- Sequencing quality
- Contamination
- Sequencing bias

[2] Read trimming and filtering

Goal

remove all non-genomic and erroneous sequence data

Software: Scythe, Trimmomatic, CutAdapt, et al.

- Trim adapters
- Trim low-quality bases
- Filter low-quality reads

[2] Read trimming and filtering

Trim Adapters

CTTCTCCTTCCTGCGACGTCGCGGGCACCGCCCACGTCGCCGCGATCCGAACAGATCGGAAGAGCACA

CTTTGGTCGCTTGAACGACCCACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGCTCATTATC

↓

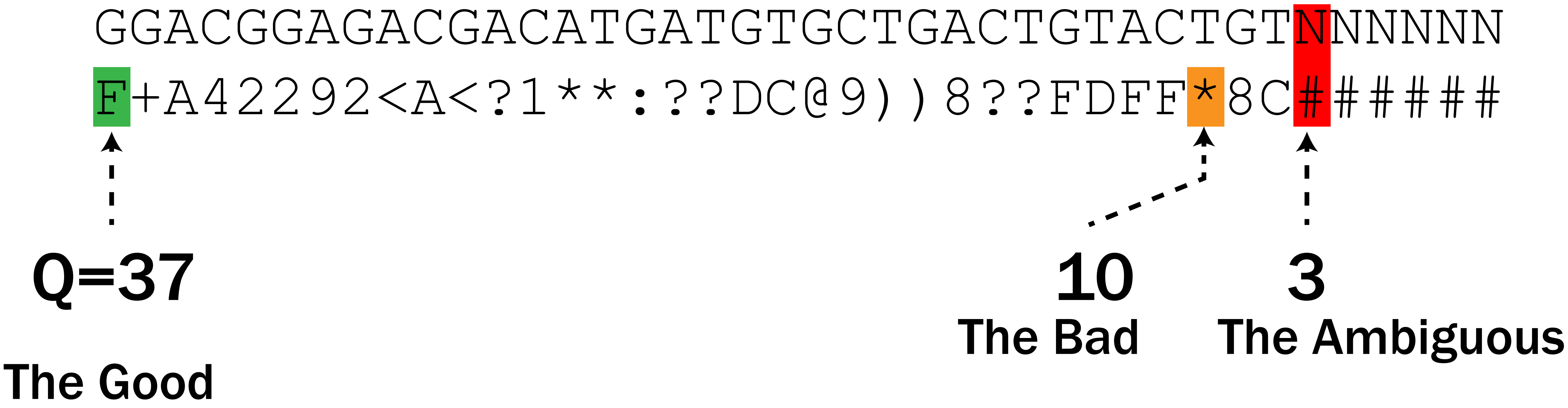
CTTCTCCTTCCTGCGACGTCGCGGGCACCGCCCACGTCGCCGCGATCCGAAC

CTTTGGTCGCTTGAACGACCCAC

[2] Read trimming and filtering

Trim low-quality bases

Quality scale 3-40 (20 - 30 is usual cutoff)



```
GGACGGAGACGACATGATGTGCTGACTGTACTGTNNNNNN
F+A42292<A<?1**:??DC@9))8??FDFF*8C######
```

Q=37

The Good

10

The Bad

3

The Ambiguous

[2] Read trimming and filtering

Filter out short reads or reads w/ low avg. quality

```
GGACGGAGACGACATGATGTGCTGACTGTACTGT
F+A42292<A<?1**:??DC@9))8??FDFF*8C
```

Average quality = 24.8

Standard Cutoff = 30bp

Minimum Length = 50bp

[2] Read trimming and filtering

The most important question:

How much can you afford to lose?

Raw coverage = 5X vs. 50X

[3a] *de novo* assembly

Goal

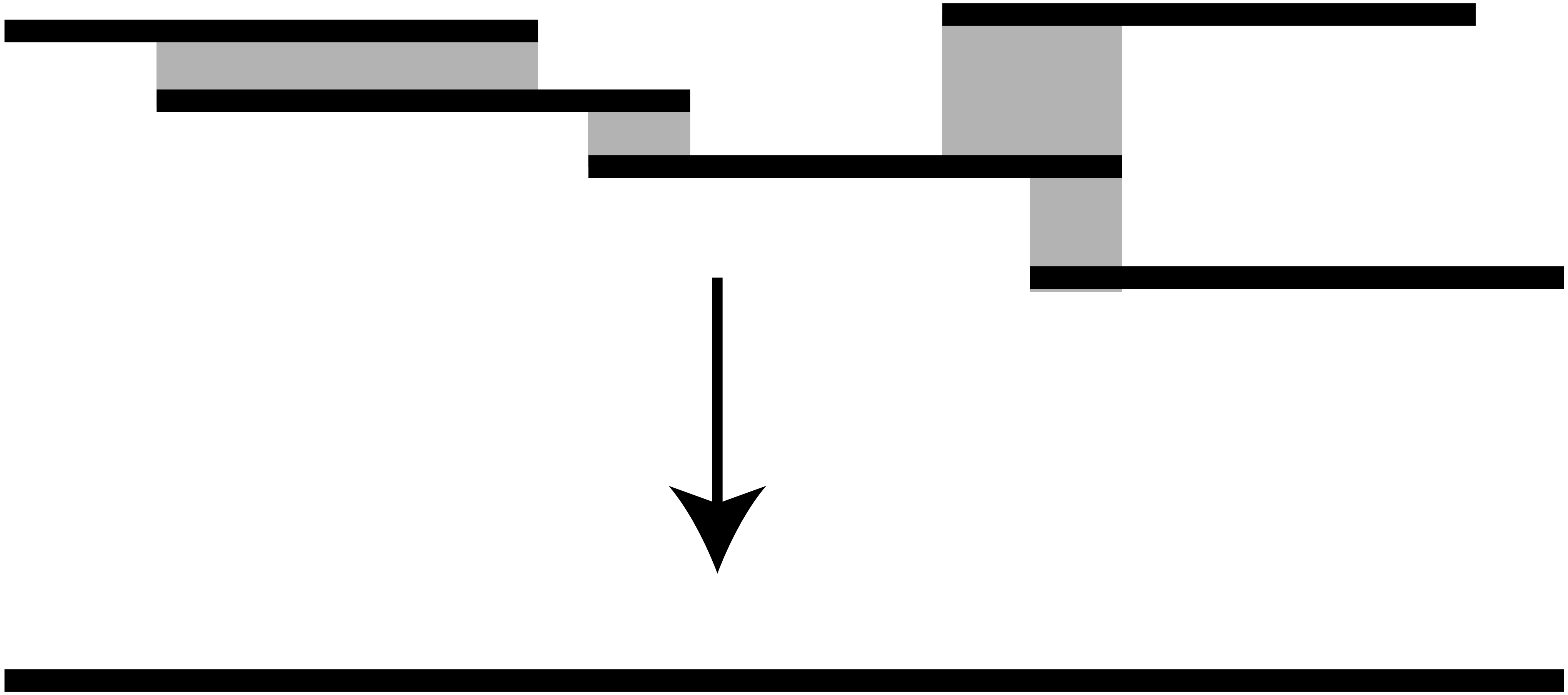
assemble the maximum number of accurate full-length transcripts

Software: Trinity*,
SoapDenovo, TransAbyss

Trimming is ESSENTIAL,
before *de novo* assembly

[3a] *de novo* assembly

Overlapping Reads

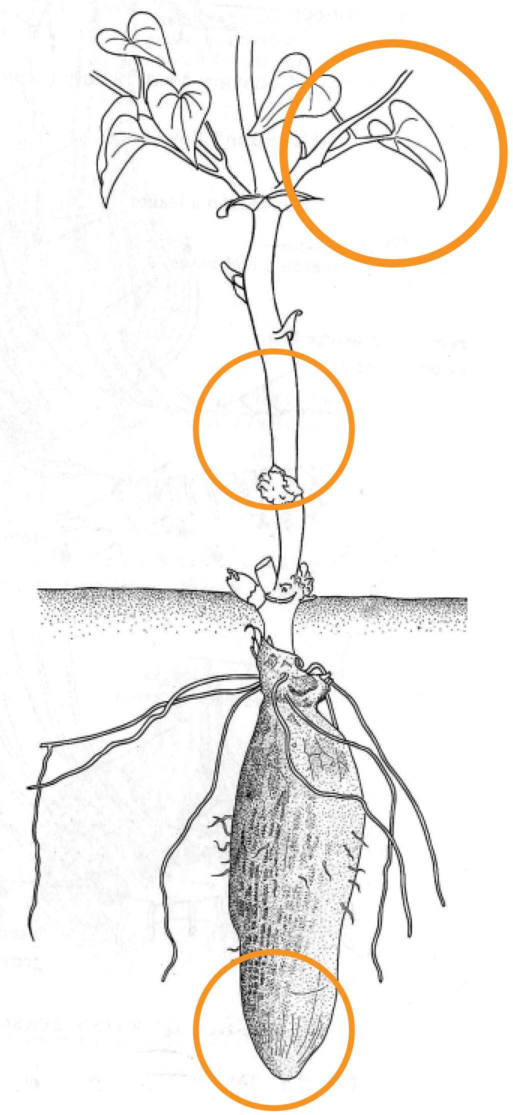


Assembled Transcript

[3a] *de novo* assembly

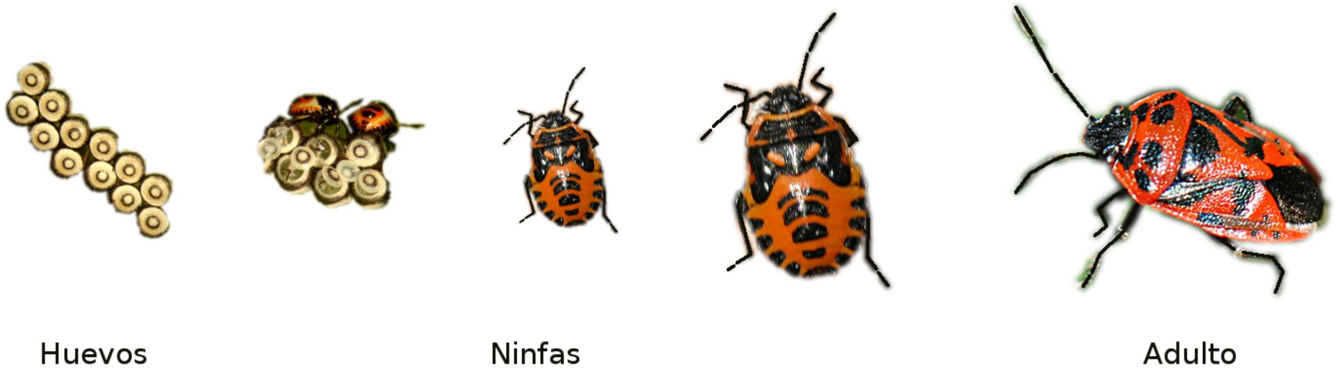
Maximizing assembly coverage:

Not all genes are expressed all the time.

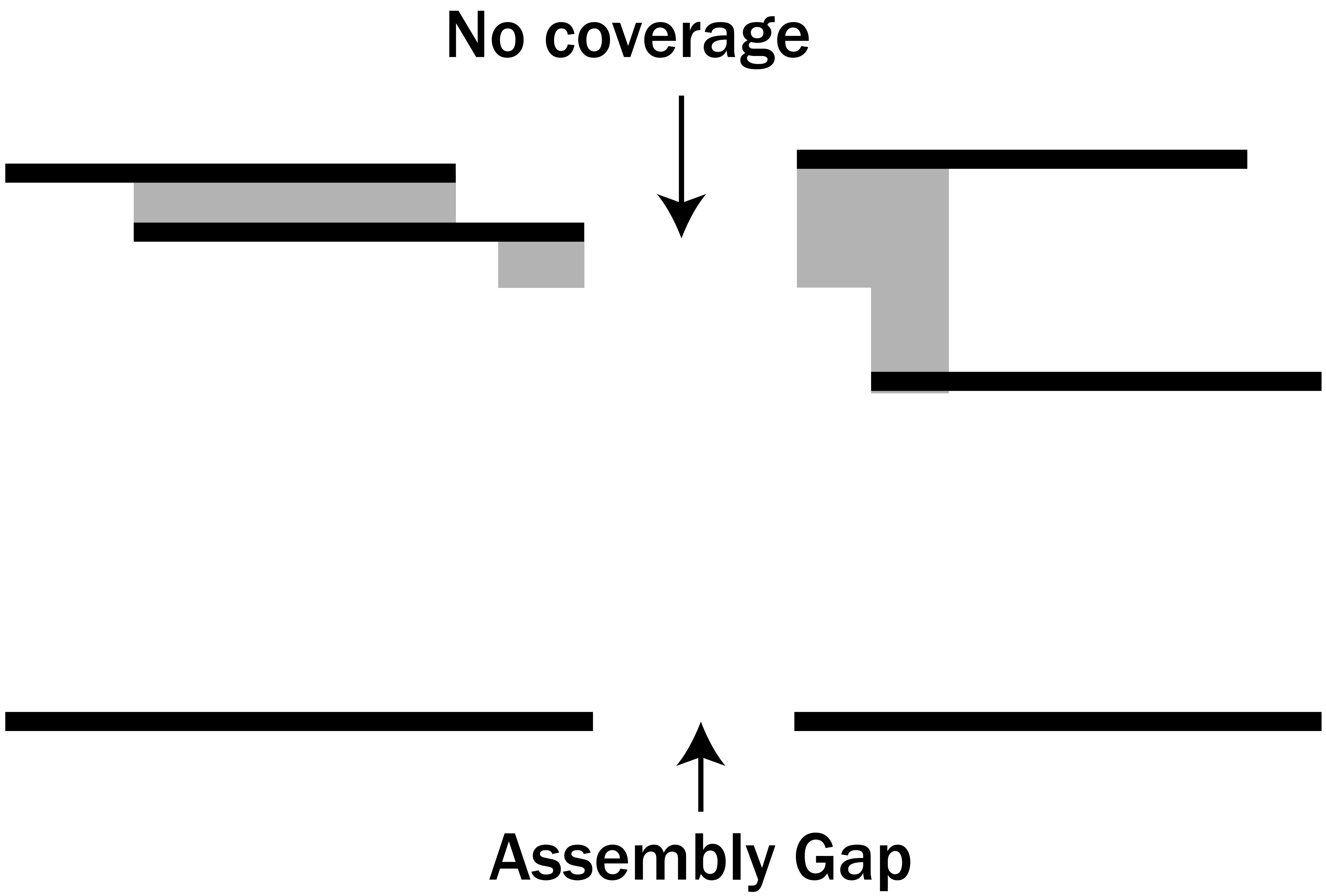


Multiple tissues

Multiple developmental stages

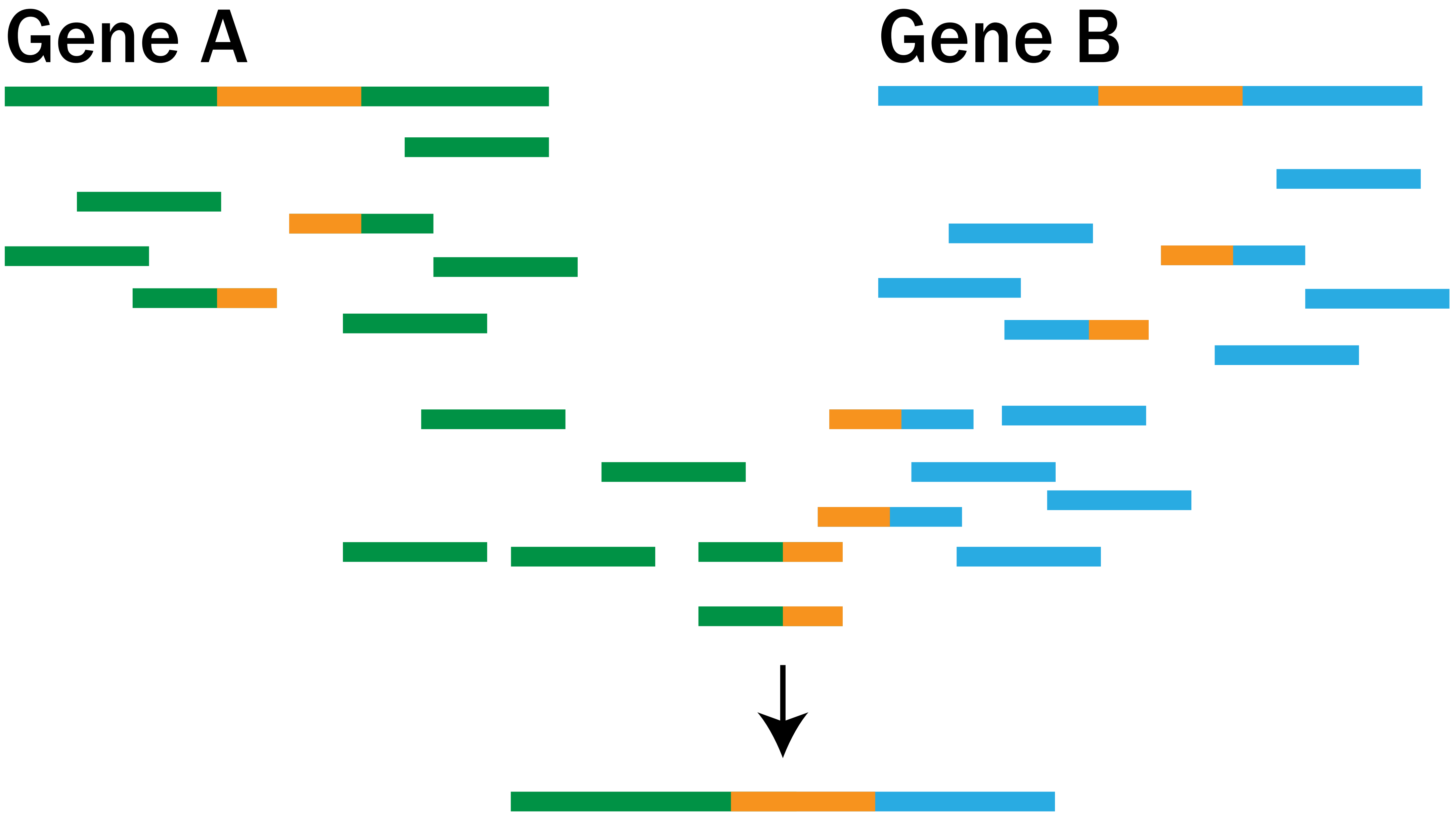


[3a] *de novo* assembly



[3a] *de novo* assembly

Chimeric transcripts



[3a] *de novo* assembly

Annotating *de novo* transcripts

- BLAST search against related organisms

- Gene Ontology search

- Protein domain and motifs

[3a] *de novo* assembly

If you have a reference genome
or closely-related genome,

you can map assembled transcripts to
the genome to “fill in the gaps”

(i.e. join transcript fragments that failed
to assemble into a full-length CDS)

[3b] Mapping

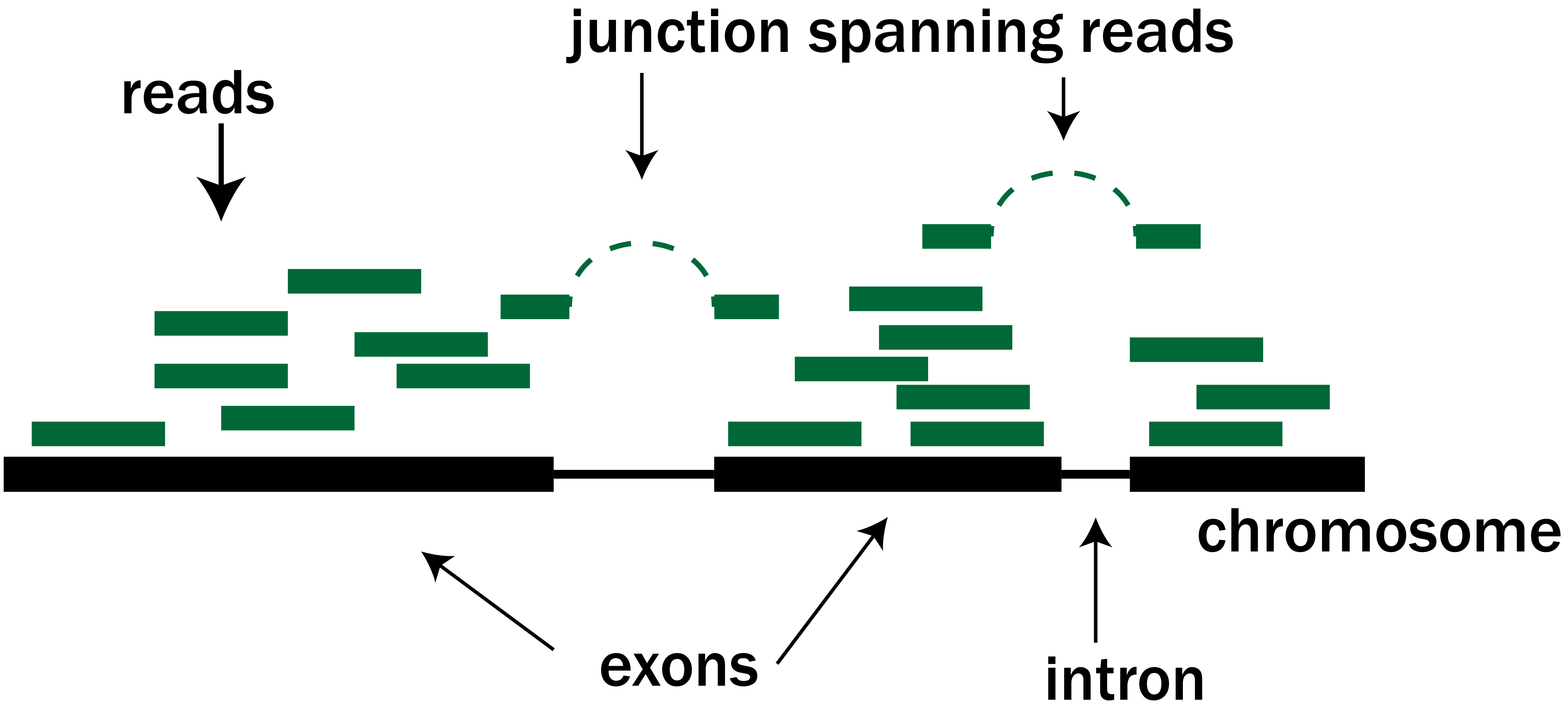
Goal

accurately align the maximum number of reads to the reference

Software: BWA, STAR, Bowtie2, Novalign, RMAP, et al.

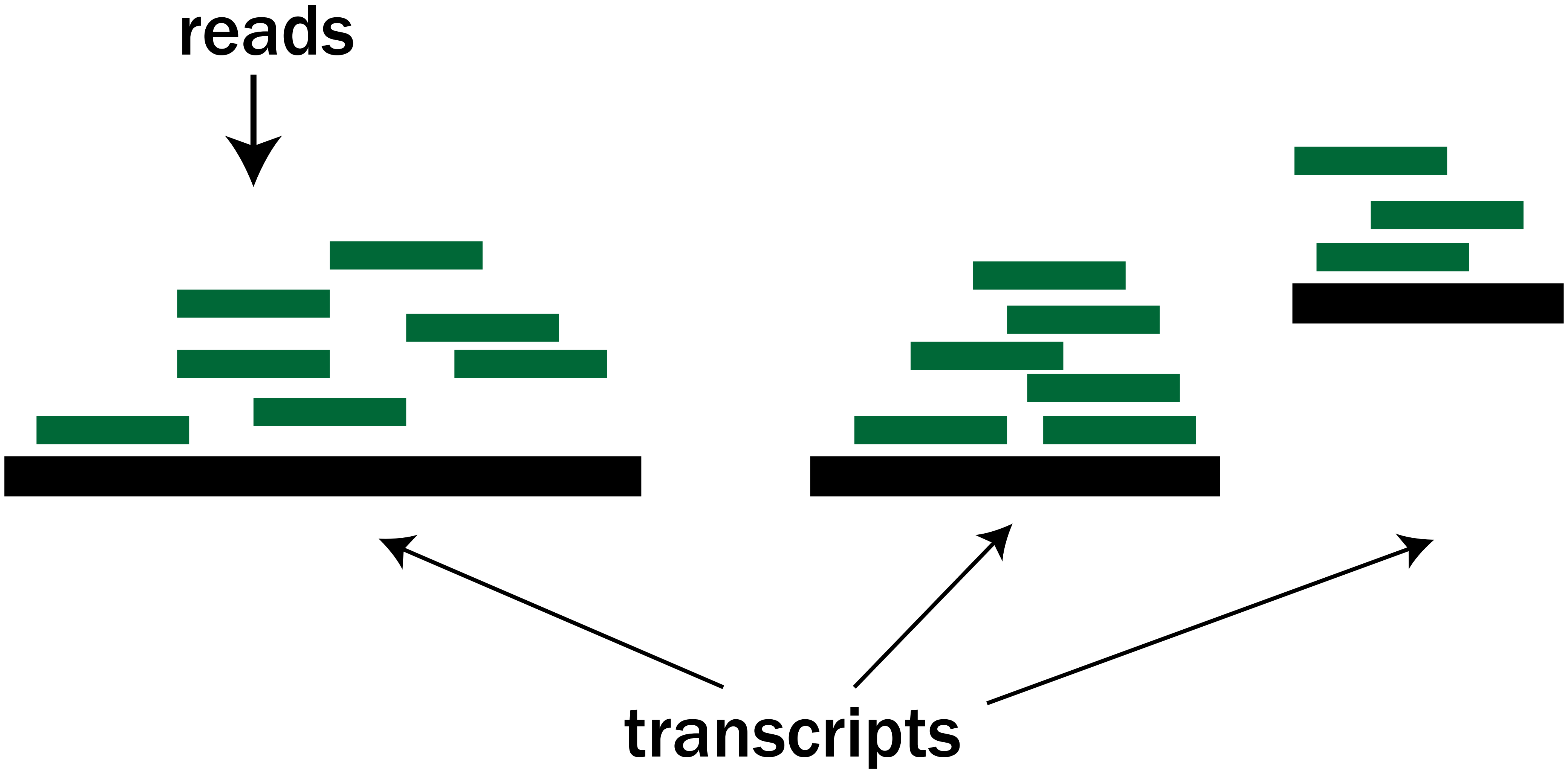
[3b] Mapping

When you map to a genome (with introns)
use a "Spliced Mapper" with gene annotation*



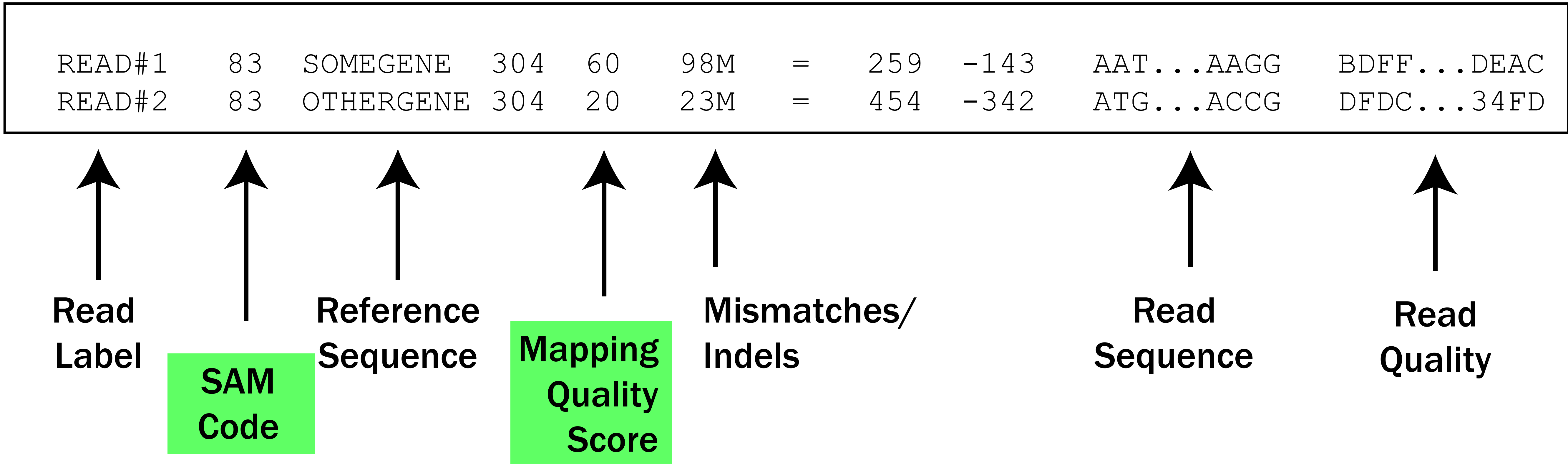
[3b] Mapping

When you map to a transcriptome (no introns), can use a direct (unspliced) mapper



[3b] Mapping

SAM Files



```
READ#1    83  SOMEGENE   304  60   98M   =   259  -143   AAT...AAGG   BDFF...DEAC
READ#2    83  OTHERGENE  304  20   23M   =   454  -342   ATG...ACCG   DFDC...34FD
```

Read Label

SAM Code

Reference Sequence

Mapping Quality Score

Mismatches/ Indels

Read Sequence

Read Quality

[3b] Mapping

Why would you get a low mapping score?

- Mapping one read to more than one gene (with similar sequence regions)
- Read alignment has too many mismatches or indels (insertions/deletions)

- Did this read map?
- Did both reads in a pair map to the same gene?

[3b] Mapping

Read mappers vary a lot:

- alignment algorithm
- quality filtering
- ability to use/infer reference gene annotations
- specifics of their output

[3b] Mapping

Where do I get the genome annotation?

Generic Feature Format (GFF)
Gene Transfer Format (GTF)

WARNING: Columns not always standard

```
AB000381 Twinscan  CDS          380   401   .   +   0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          501   650   .   +   2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          700   707   .   +   2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  start_codon  380   382   .   +   0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  stop_codon   708   710   .   +   0  gene_id "001"; transcript_id "001.1";
.....
```

[4] Assembly QC

Goal

assess the quality and accuracy of your assembly

Software: FASTX Toolkit, SAMstats

[4] Assembly QC

Sequence quantity:

- Total amount of sequence

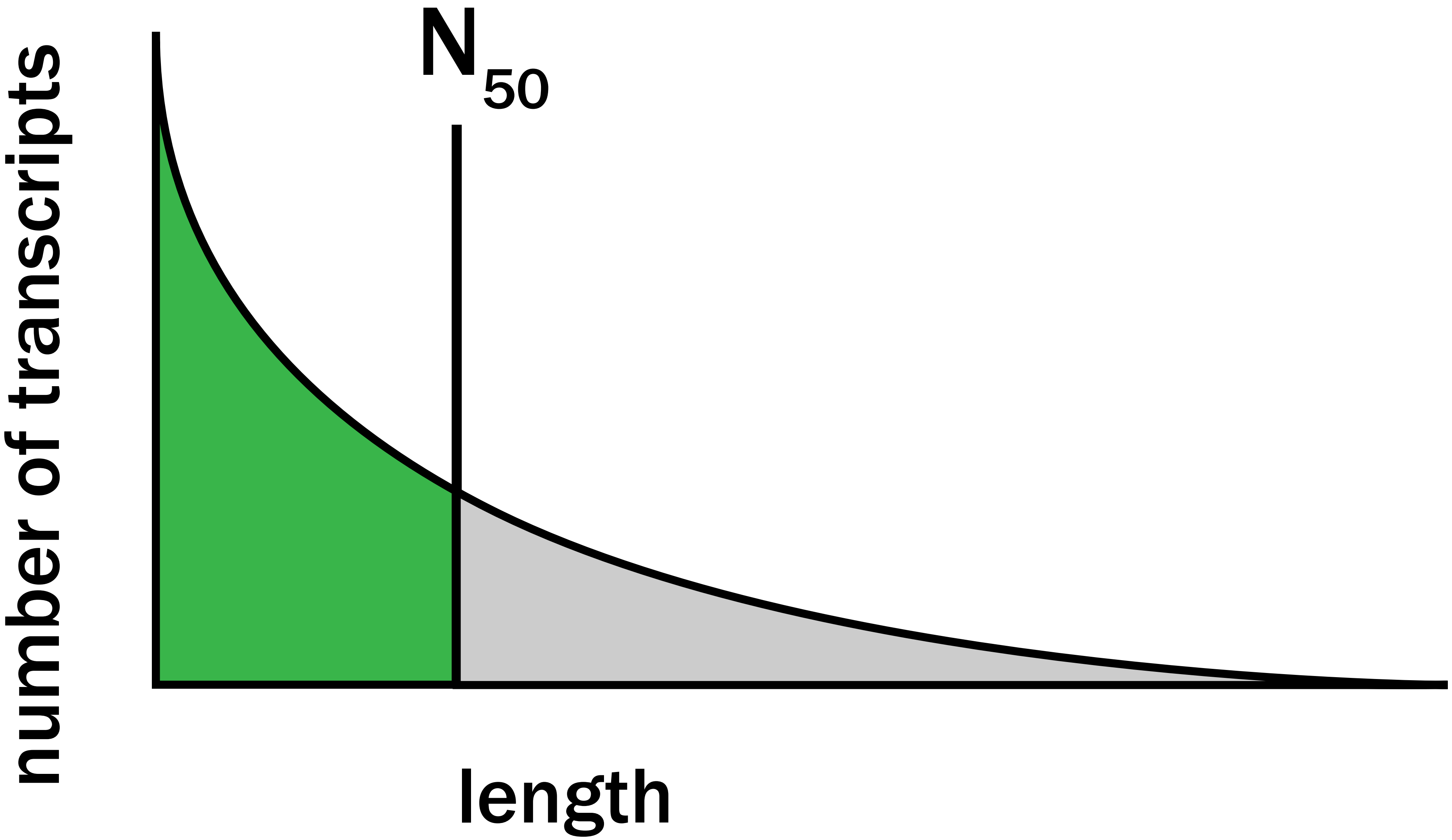
  Does it seem reasonable for your organism?

- Mean transcript length

  For eukaryotes ~ 700-1000bp is common

- Do you recover long genes (several kb)?

[4] Assembly QC



[4] Assembly QC

Data use efficiency:

- What proportion of reads map back to your assembly?
- What proportion of reads map to the reference?

[4] Assembly QC

Mapping coverage:

- What proportion of annotated transcripts are covered by reads?
- Is there mapping bias?

The Most Important Slide

[1] RNA-Seq is an extremely active research area, stay current!

[2] The principles are general, the specifics will change in less than ~~two~~ one year~~s~~.

[3] Everyone worries about processing, most errors are bad data or analysis

Questions?