High-Throughput Sequencing: A Brief Introduction
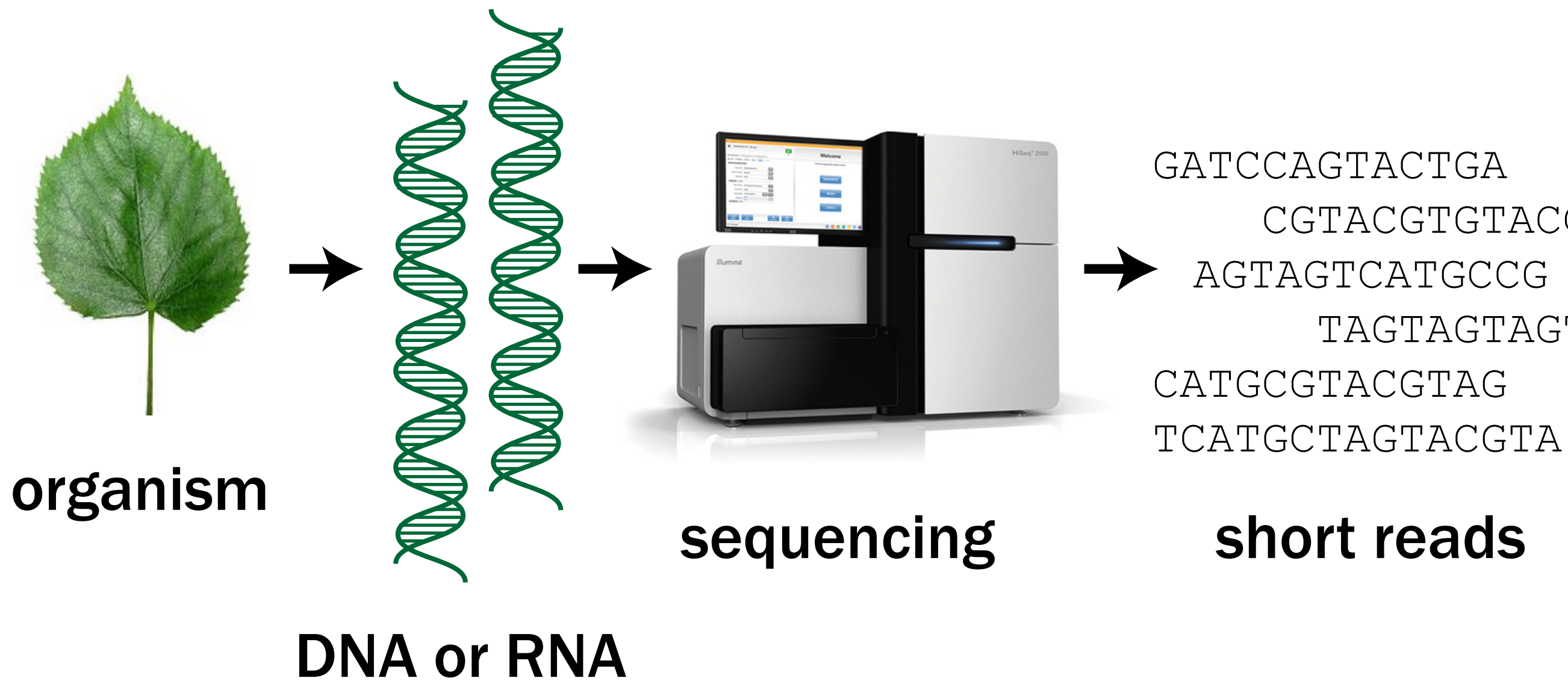
James Pease

14 July 2014

High throughput sequencing
("Next Generation Sequencing")

organism

DNA or RNA

sequencing

```
GATCCAGTACTGA
  CGTACGTGTAC
AGTAGTCATGCCG
    TAGTAGTAGT
CATGCGTACGTAG
TCATGCTAGTACGTA
```

short reads

Many sequencing technologies

illumina®

http://www.youtube.com/watch?v=l99aKKHcxC4

PACIFIC BIOSCIENCES®

http://youtu.be/v8p4ph2MAvI?t=52s

Millions (billions) of short reads

Paired end sequencing

sequence each fragment
from both ends

secuencia

aicneuces

Normally, data will come as two files
(one for each end, left/right)

DNA or RNA?

DNA

the whole genome

non-coding regions
(introns, etc.)

costs more

RNA

only transcripts
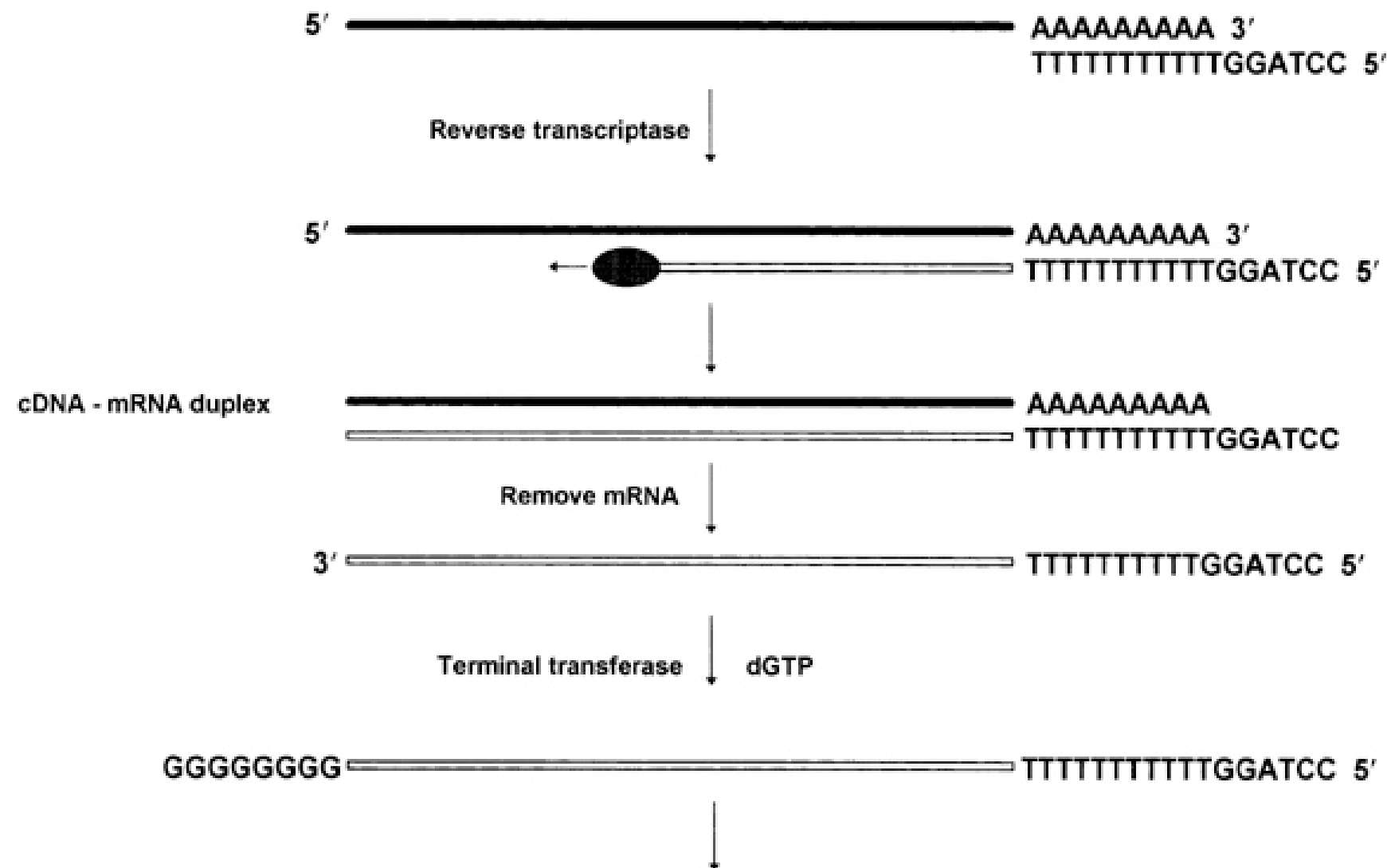
expression data

alternatively spliced
isoforms

costs less

RNA-Seq is actually cDNA-Seq

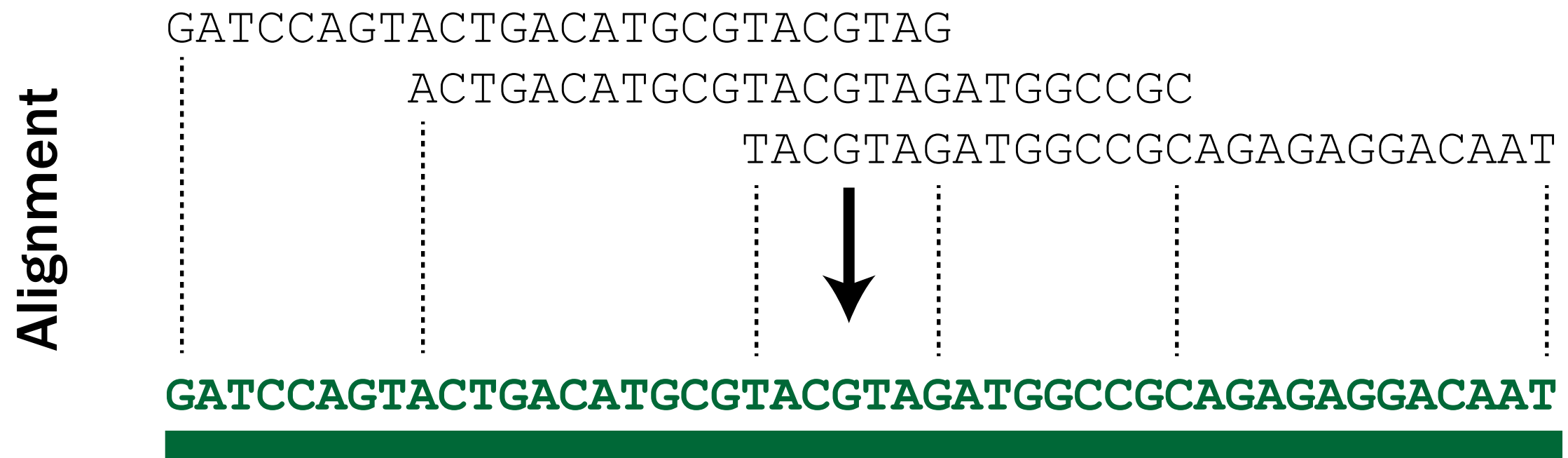Reverse Transcriptase linked to Polymerase Chain Reaction (RT-PCR)   1

mRNA + primer 1 (BamHI)

```
5' ━━━━━━━━━━━━━━━━━━  AAAAAAAAA 3'
                       TTTTTTTTTTTGGATCC 5'
```

        Reverse transcriptase ↓

```
5' ━━━━━━━━━━━━━━━━━━  AAAAAAAAA 3'
              ←  ●━━━━  TTTTTTTTTTTGGATCC 5'
```

                            ↓

cDNA - mRNA duplex
```
━━━━━━━━━━━━━━━━━━  AAAAAAAAA
══════════════════  TTTTTTTTTTTGGATCC
```

        Remove mRNA ↓

```
3' ══════════════════  TTTTTTTTTTTGGATCC 5'
```

        Terminal transferase ↓ dGTP

```
GGGGGGGG ══════════════════════  TTTTTTTTTTTGGATCC 5'
```

                            ↓

Short Read Assembly

DNA Short Reads

Alignment

```
GATCCAGTACTGACATGCGTACGTAG
          ACTGACATGCGTACGTAGATGGCCGC
                    TACGTAGATGGCCGCAGAGAGGACAAT
```

GATCCAGTACTGACATGCGTACGTAGATGGCCGCAGAGAGGACAAT

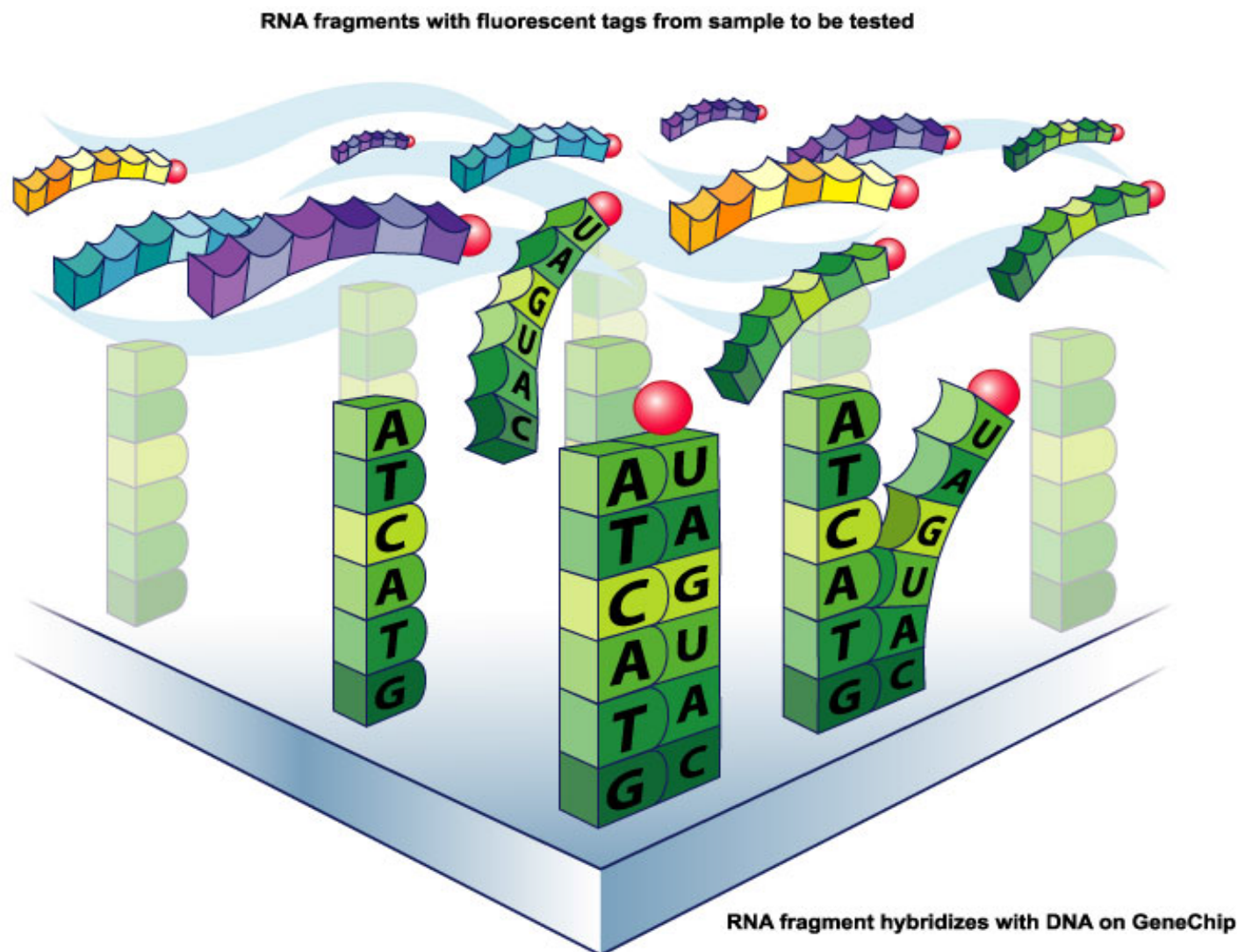## Contiguous sequence "contig"

contigs

scaffold

chromosome

Mapping reads

Alignment of DNA (or RNA) short reads
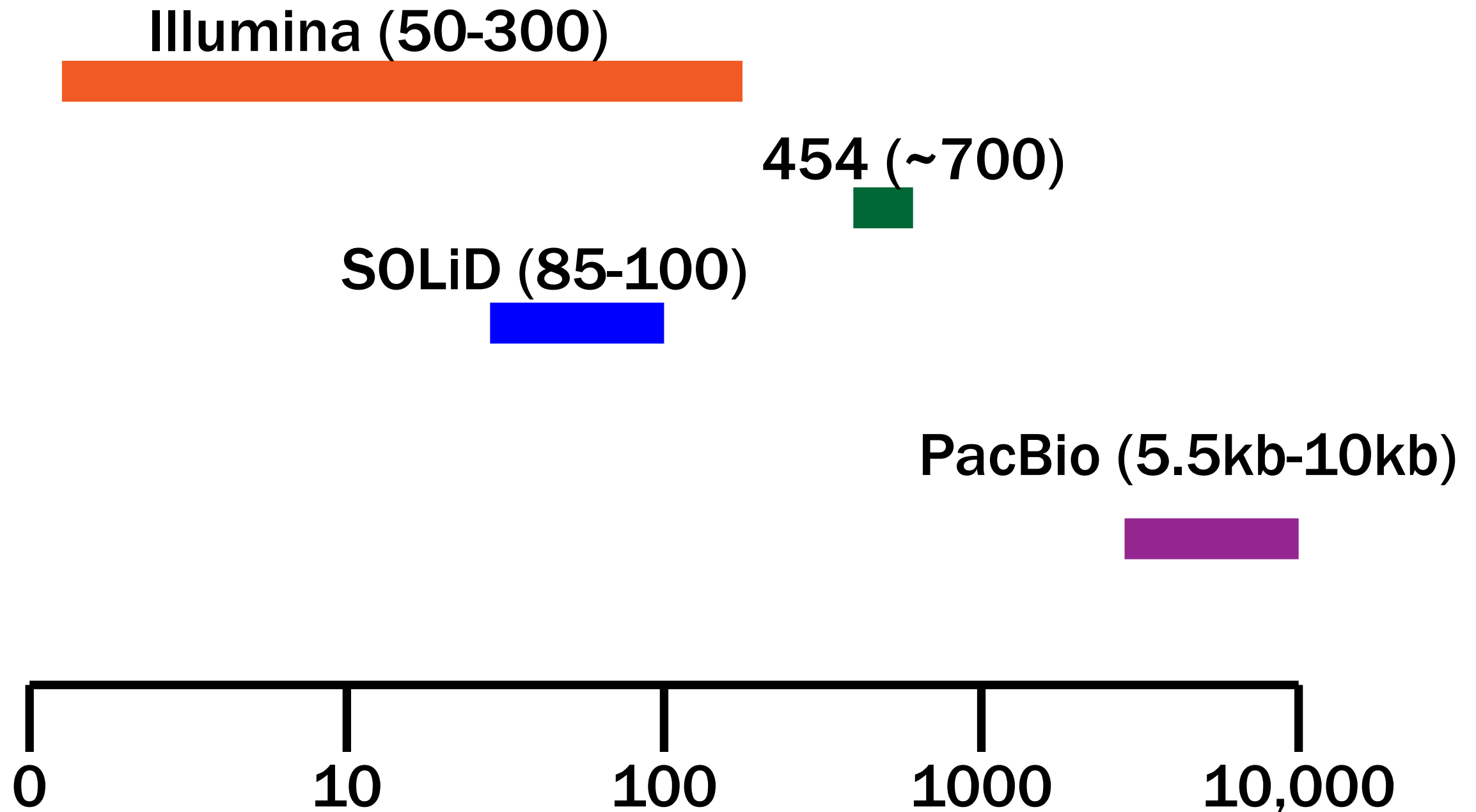to chromsomes (or transcripts)
from a reference genome

Microarray

Requires a pre-designed panel of <u>known</u> transcript sequences

RNA fragments with fluorescent tags from sample to be tested

RNA fragment hybridizes with DNA on GeneChip

	RNA-Seq	microarray
Annotate genes	yes	no
All transcripts	yes*	no
Differential Expression	yes	yes
Genotyping	yes	no*

DNA-Seq Options

Whole-genome

RAD-Seq:
Restriction site Associated DNA Markers

Stranded/unstranded

RNA-Seq Options

Whole-transcriptome

Small RNAs

Stranded/unstranded

Normalized/non-normalized

Choosing a Sequencing Technology

- Read Length

- Error Rates

- Software Options

- Cost (time and money)

Researchers often spend months or years painstakingly designing a study and collecting data, and then throw their data into a computer and try to analyse it in minutes. The data deserve better. The quick and reckless approach to data analysis often fails to identify important aspects of the data. You should become friends with them! Conducting data analysis is like drinking a fine wine. It is important to swirl and sniff the wine, to unpack the complex bouquet and to appreciate the experience.
Gulping the wine doesn't work.

Daniel Wright (2003) BJEP