HTS Analysis

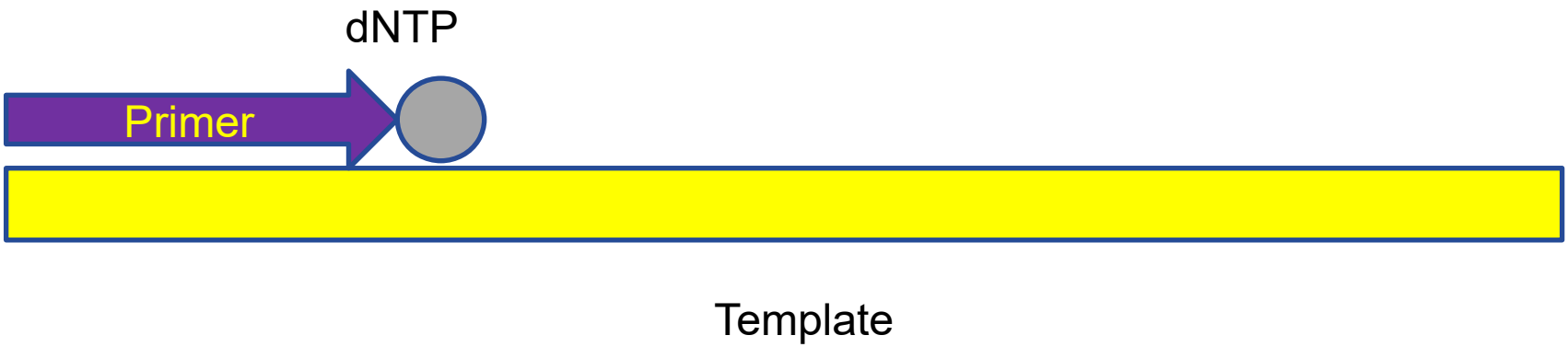
Technology and analysis

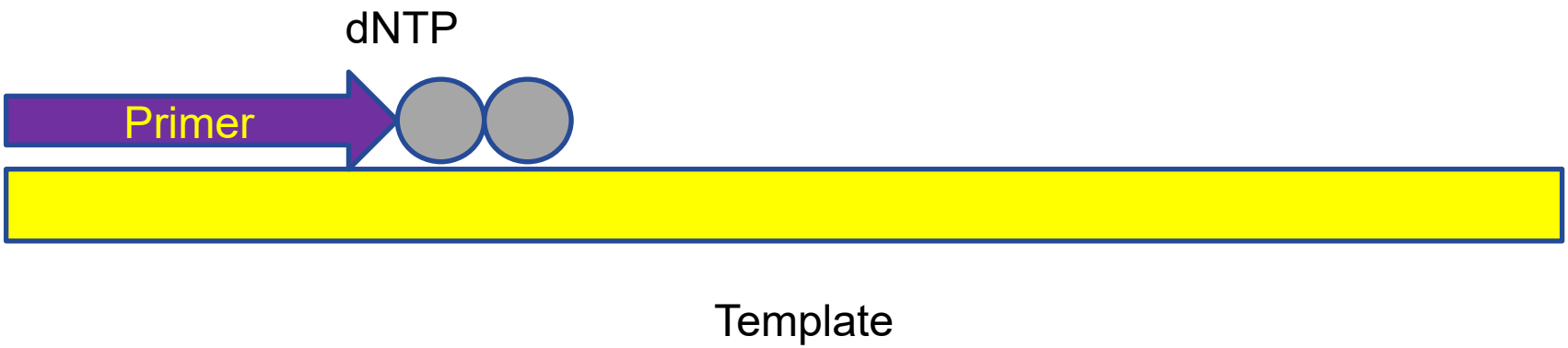
All slides Copyright 2026 by Michael V. Osier. All rights reserved.

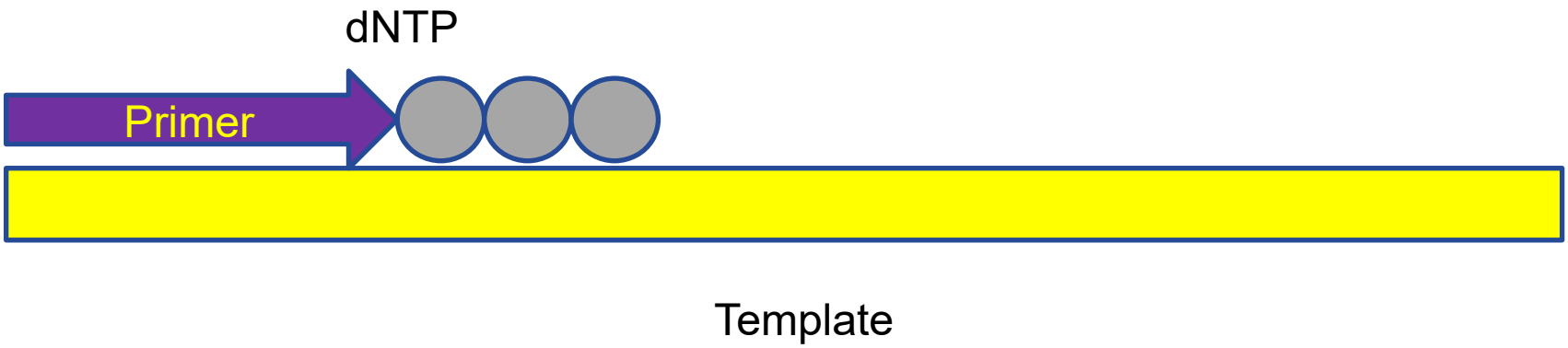
Technology

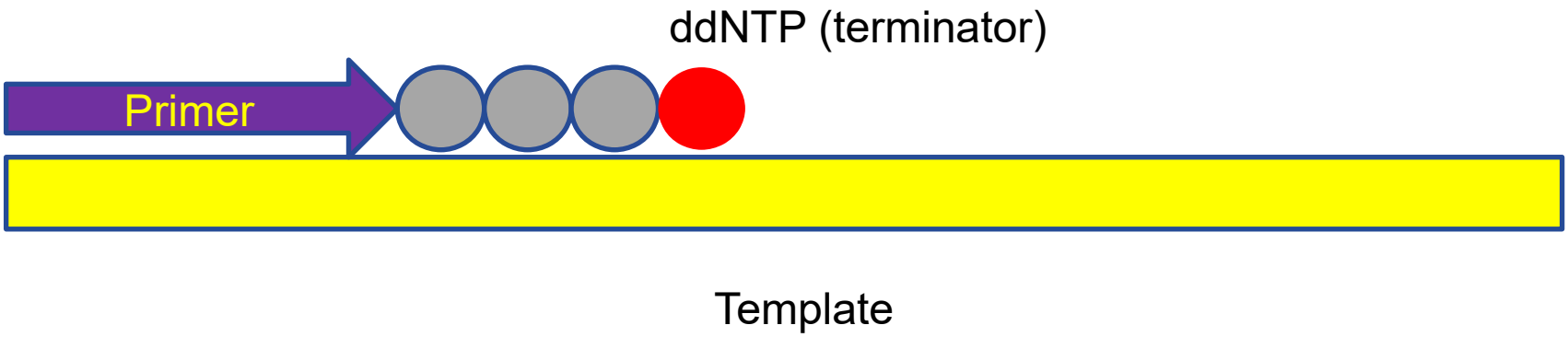
Sanger sequencing

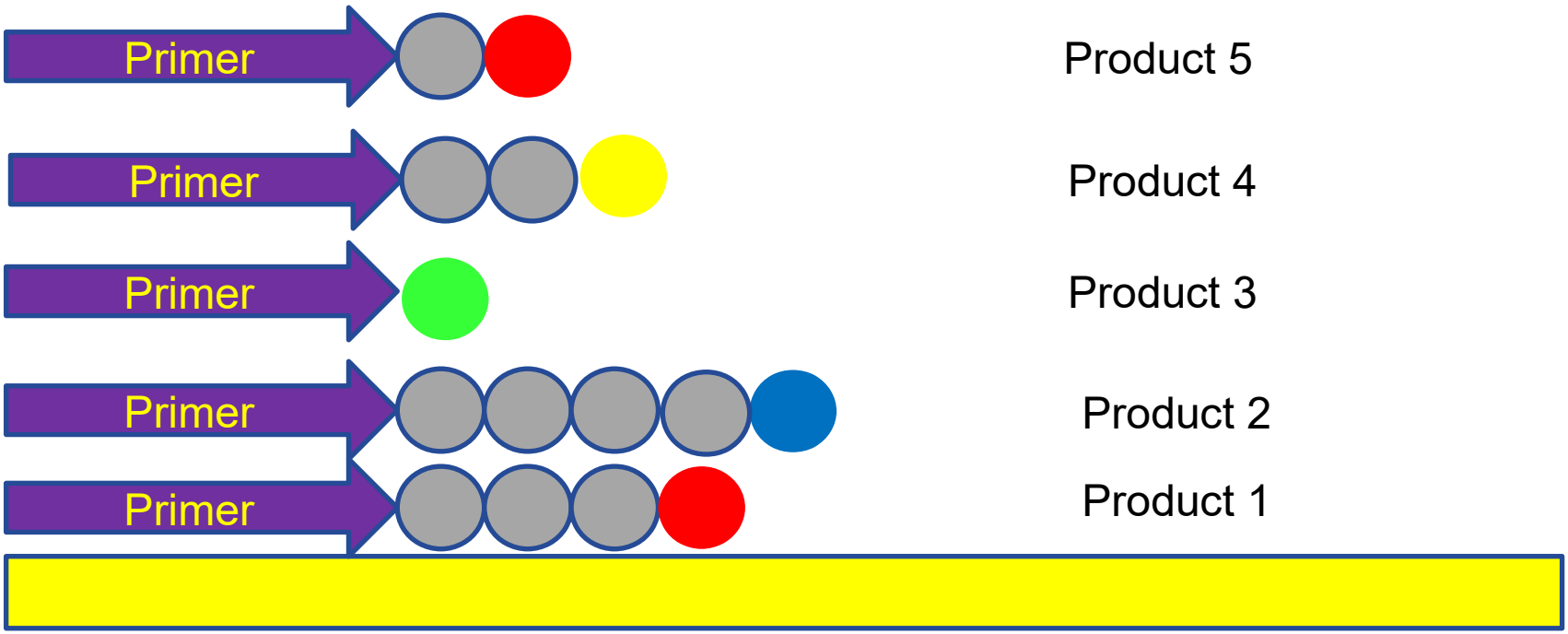
Template

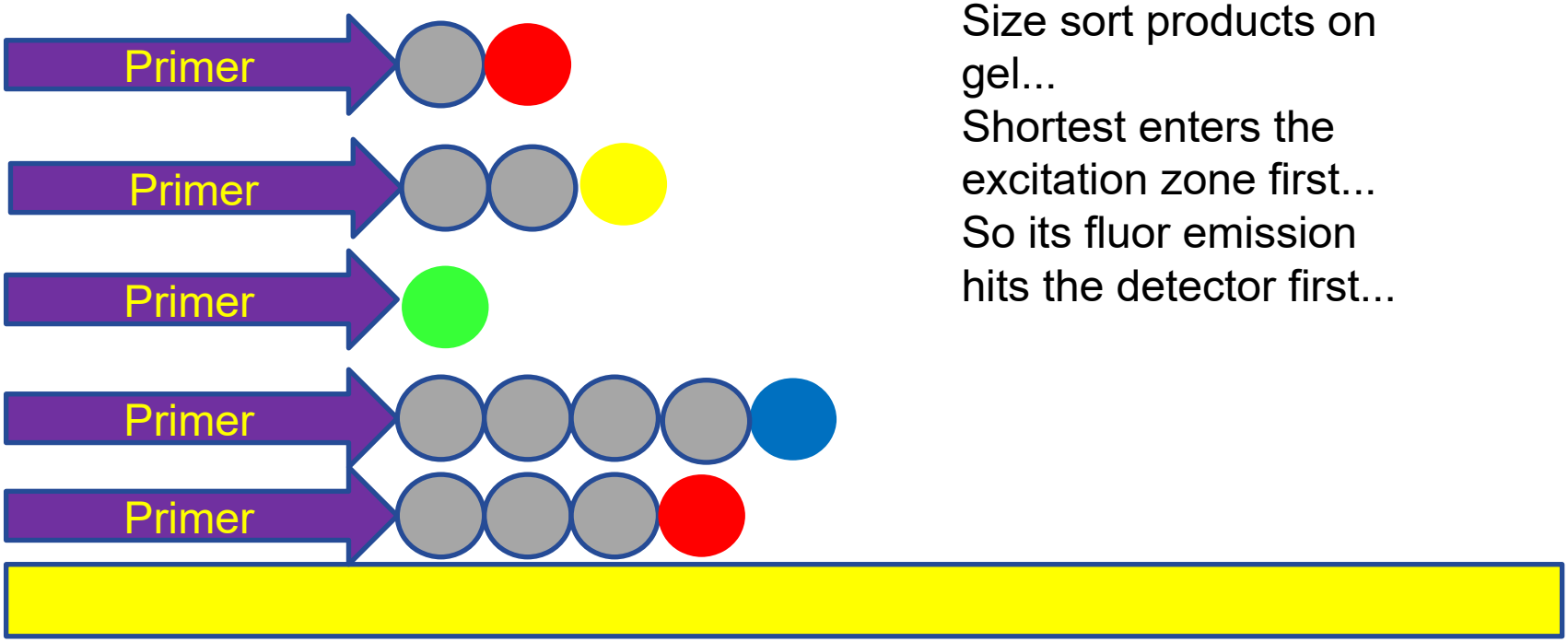

dNTP
Primer
Template


dNTP
Primer
Template


dNTP
Primer
Template


ddNTP (terminator)
Primer
Template


Primer
Primer
Primer
Primer
Primer
Product 5
Product 4
Product 3
Product 2
Product 1



Size sort products on gel...
Shortest enters the excitation zone first...
So its fluor emission hits the detector first...

454

Pyrosequencing

Pyrosequencing

- Instead of using chain terminators…
- …pyrosequencing detects as each base is added by polymerase and identifies which base it was.

Pyrosequencing

Template bound to bead in a microwell

Pyrosequencing

Polymerase

+

One dNTP 

+

Template

Pyrosequencing

Polymerase

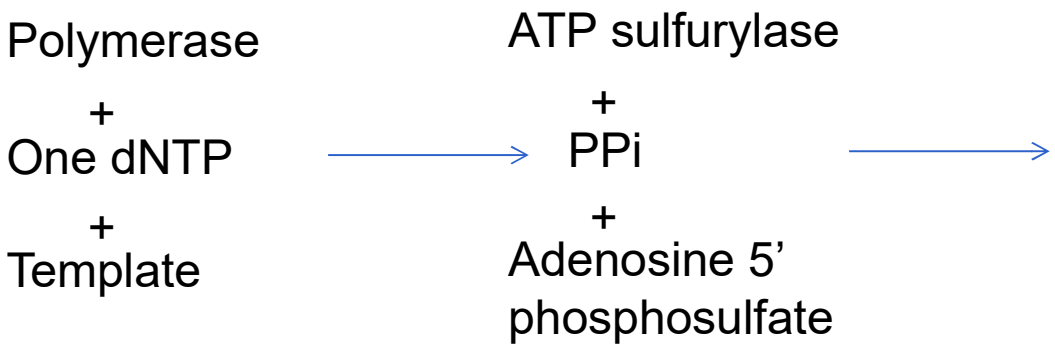
+

One dNTP  PPi

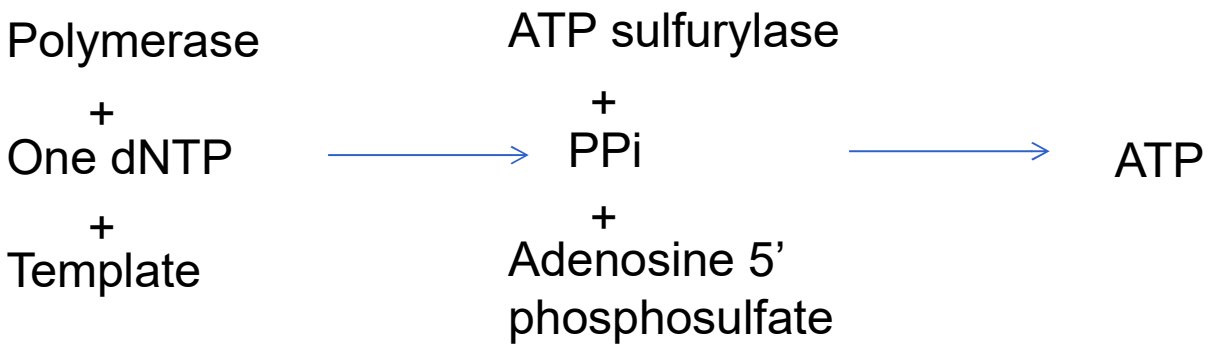
+

Template

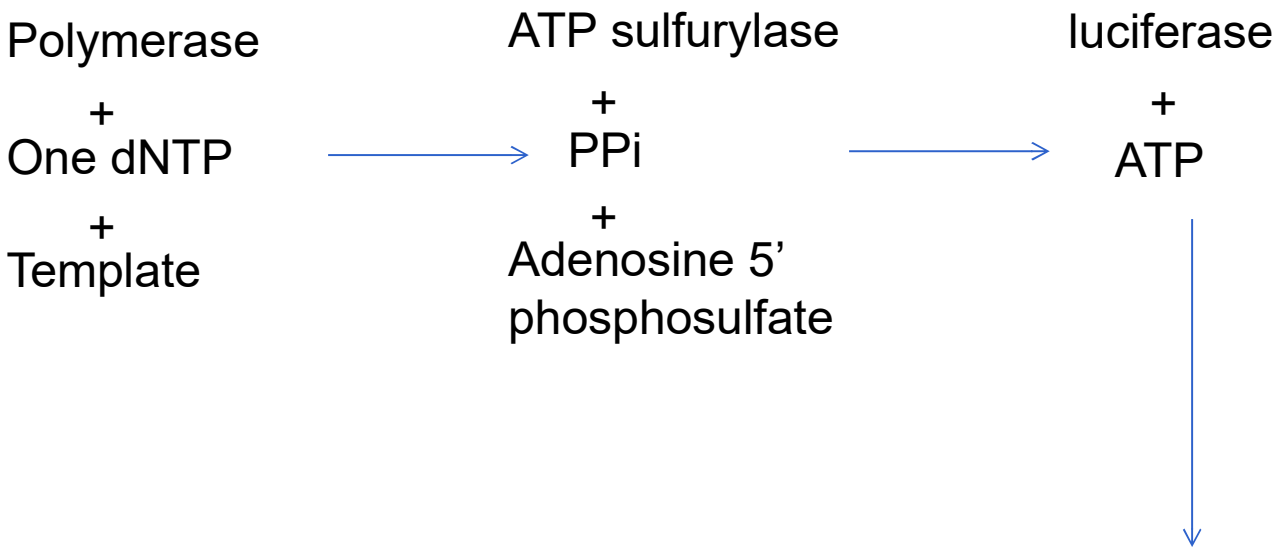
Pyrosequencing



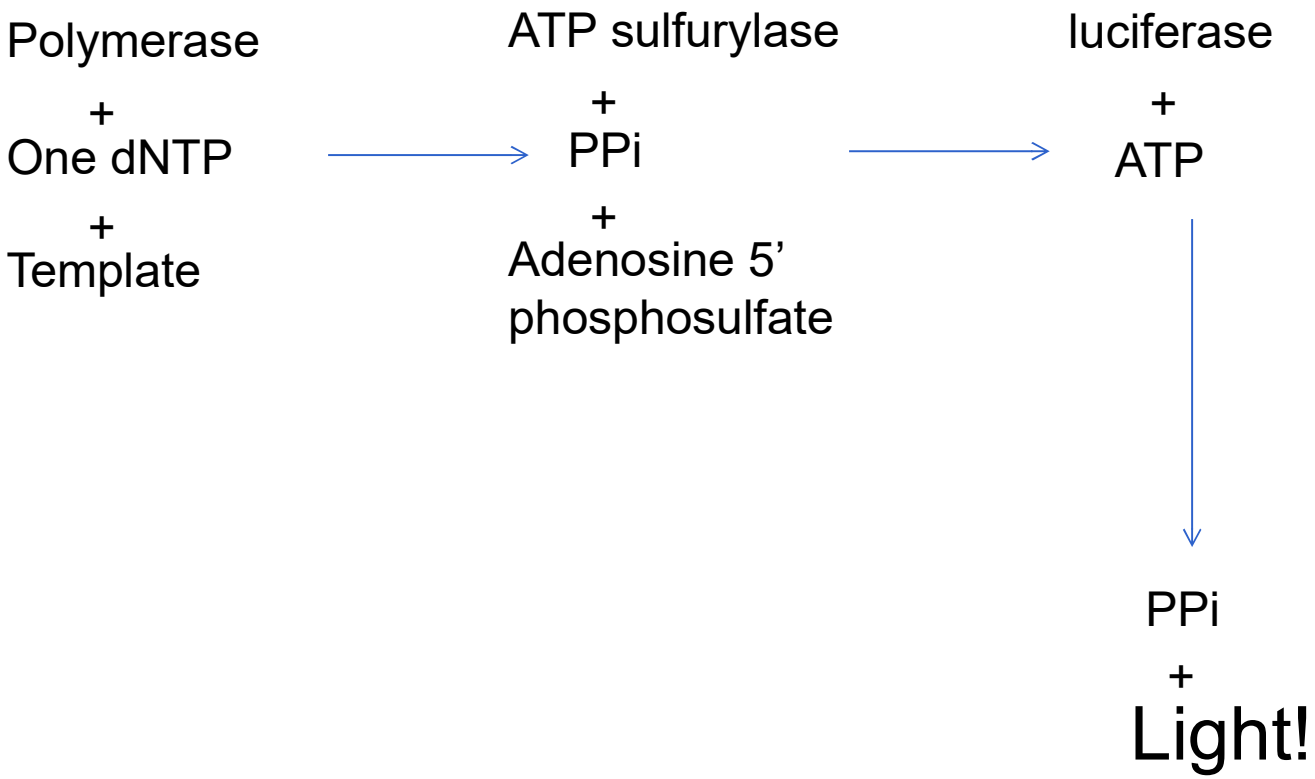
Pyrosequencing



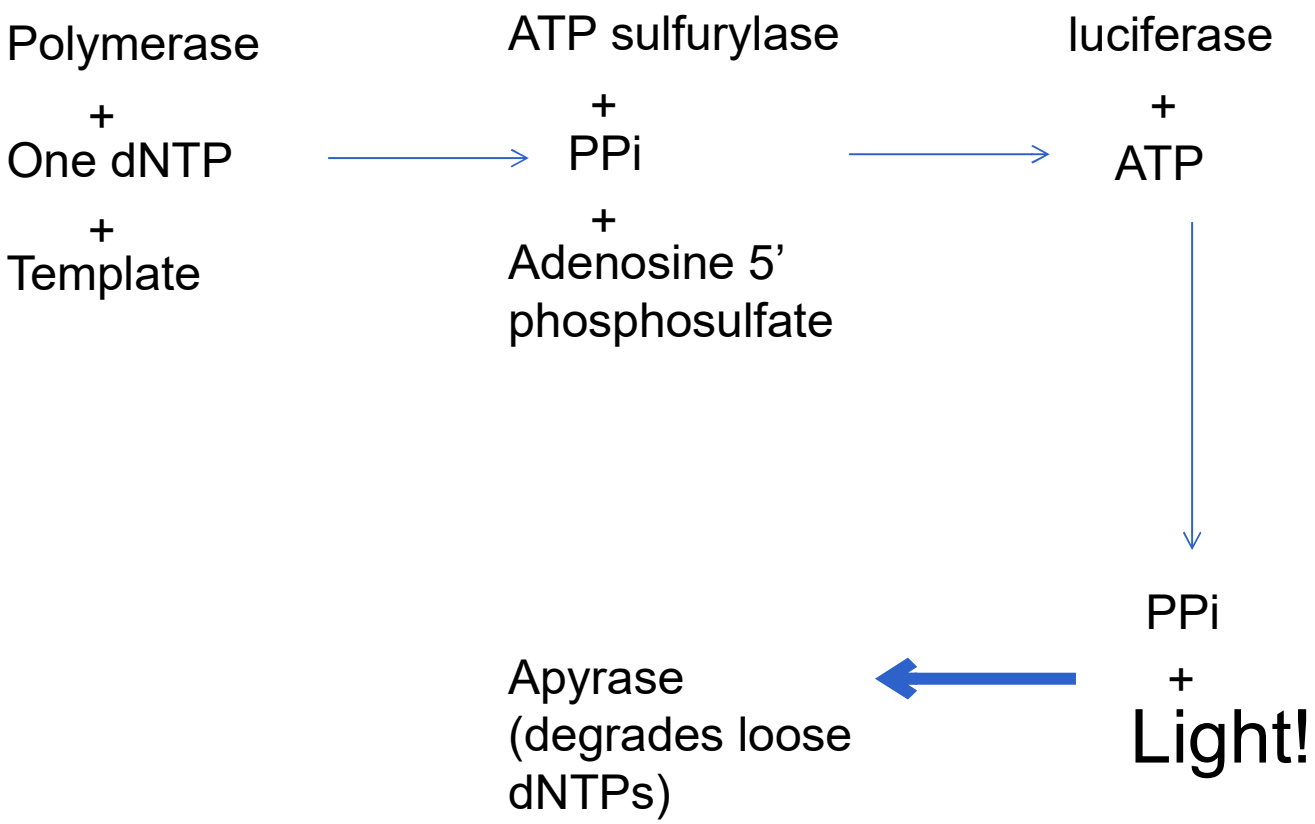
Pyrosequencing



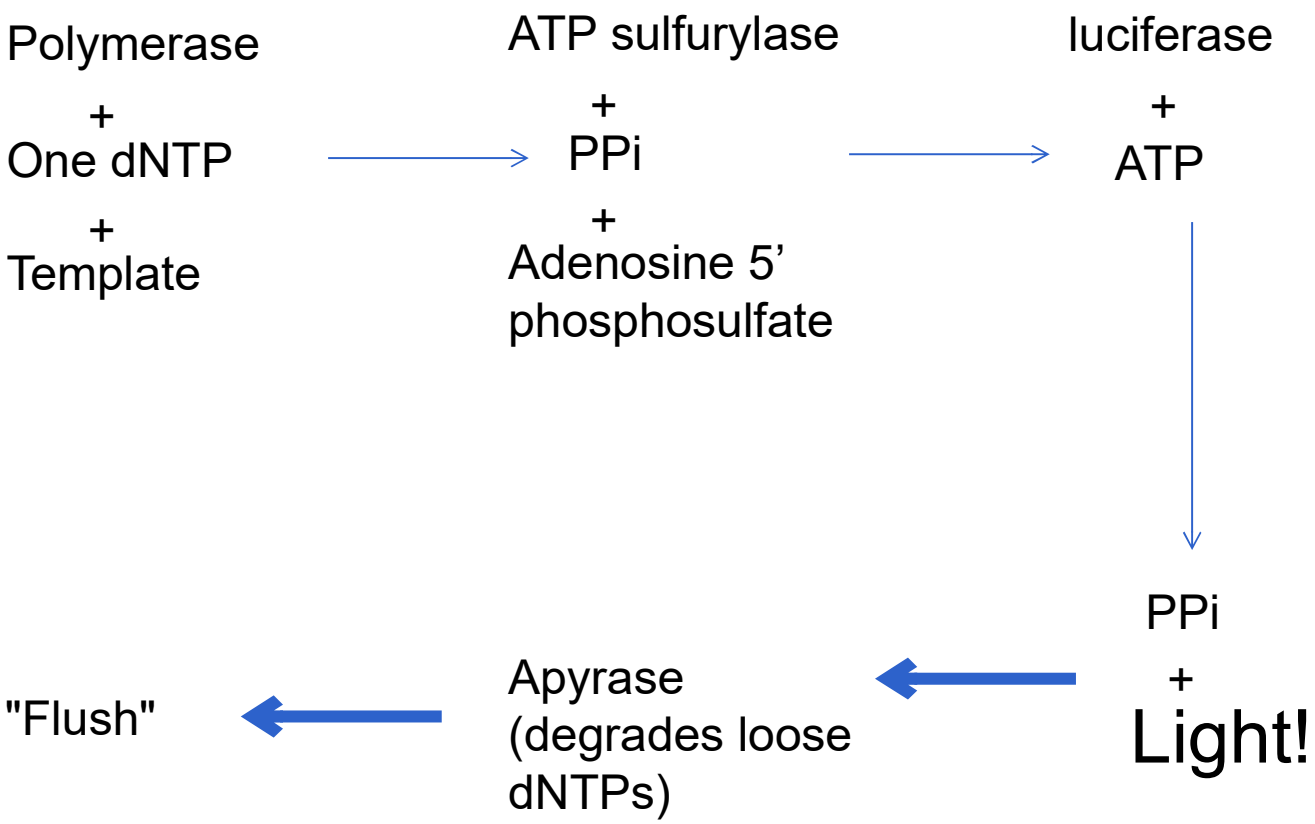
Pyrosequencing



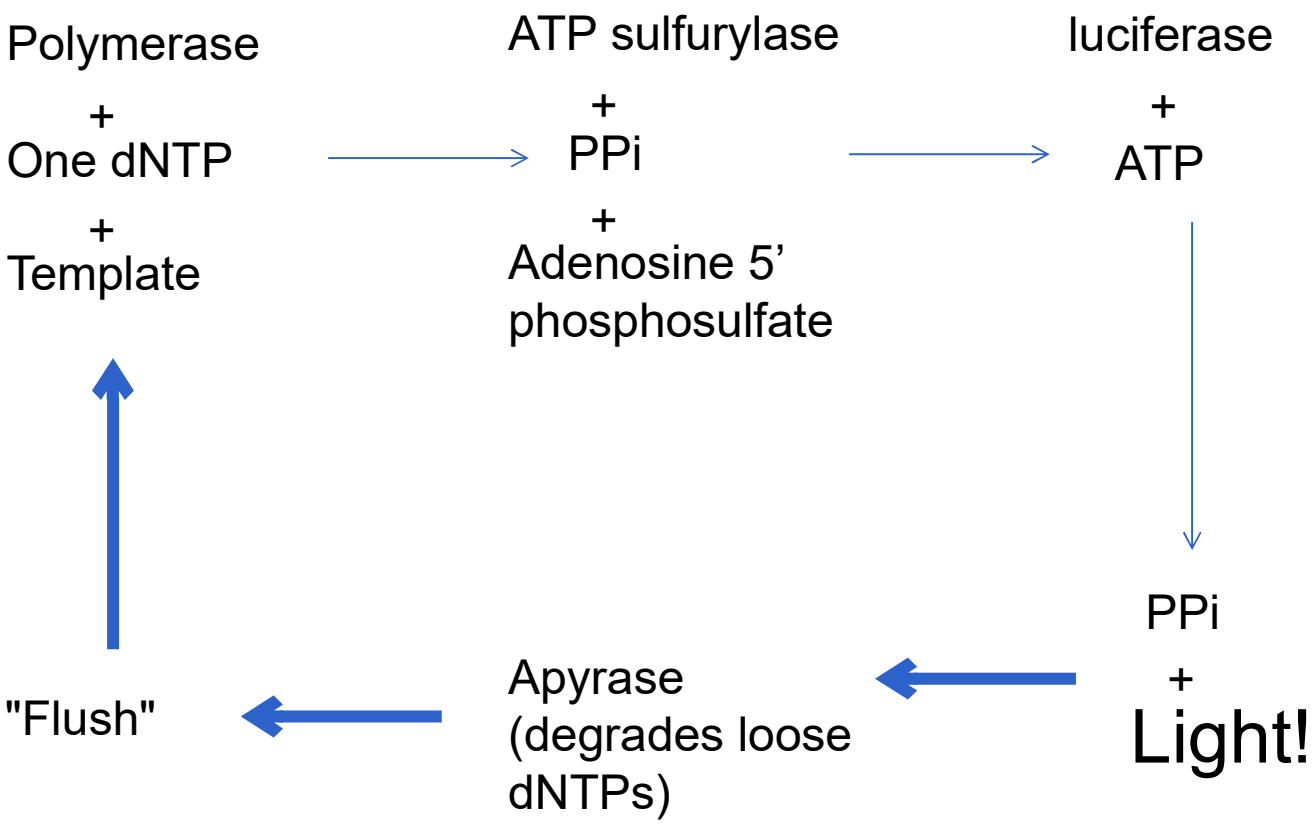
Pyrosequencing



Pyrosequencing



Pyrosequencing



Pyrosequencing

Pros

- Fast!
  - 400-600 Mbp per 10 hours!
- Cheaper
  - About $5000-7000 per run
- Good in GC-rich regions

Pyrosequencing

Pros

- Fast!
  - 400-600 Mbp per 10 hours!
- Cheaper
  - About $5000-7000 per run
- Good in GC-rich regions

Cons

- Smaller sequence length
  - 400-500 bp per read
  - Sanger is 800-1000 bp per read
  - 454 releasing new kit with up to 1000 bp per read
- Error prone
  - Esp. poly-nucleotides (… AAAA..., …CCCCC…, etc.)
  - Higher misincorporation

GS FLX+ System

- 85% of reads > 500bp
- Mode (most common length) 700bp
- 23 hour run
  - ~1,000,000 reads
  - ~700 Mb

- As reported by Roche

- Good summary of pyrosequencing in:
- Ronaghi M. (2001) Pyrosequencing sheds light on DNA sequencing. Genome Research 11:3-11.

RIP 454...

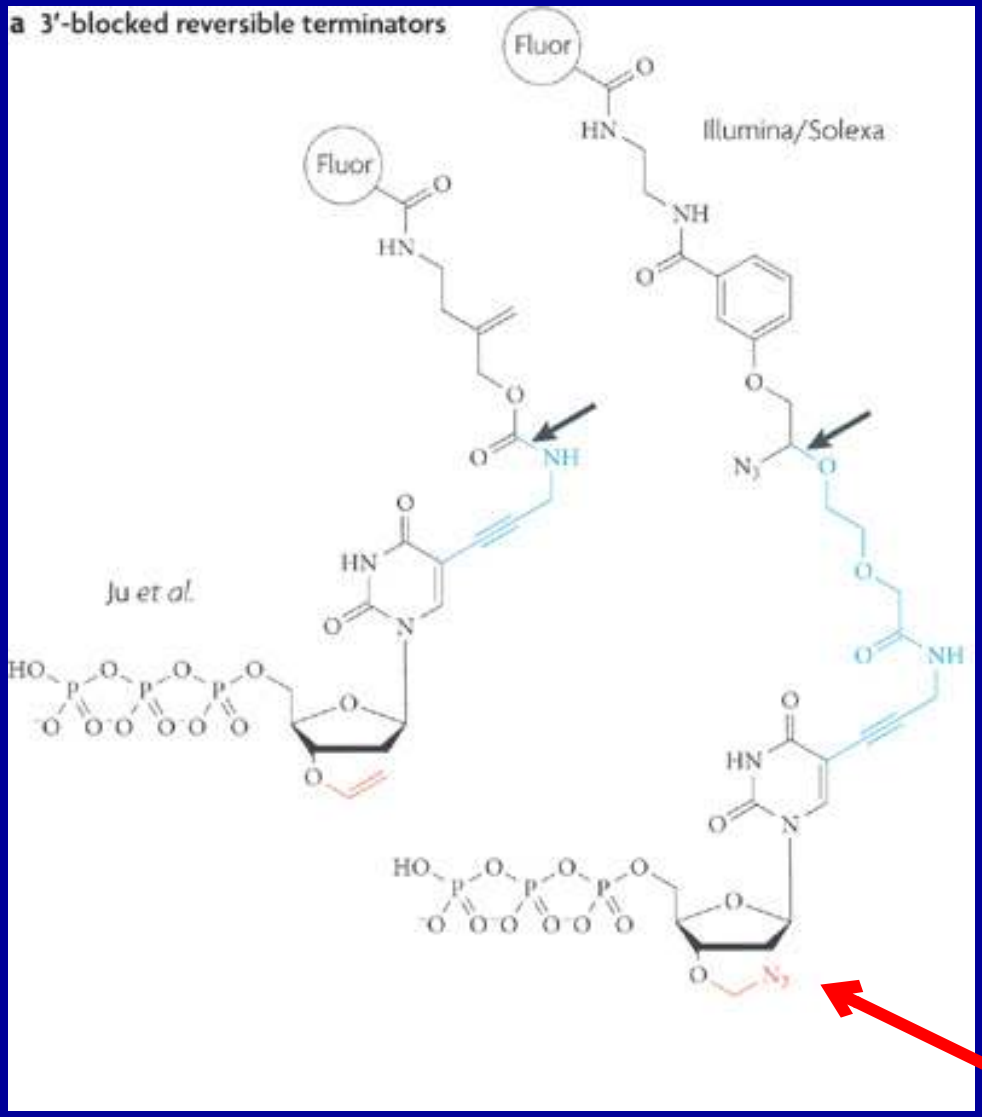
Illumina

Sequencing by Synthesis (SBS)

SBS

- Similar to pyrosequencing…
- Instead of using light to identify when a nucleotide was added…
- …add all four dNTPs, each with a different fluorescent probe terminator!
  - Prevents further elongation
  - Named “reversible terminators”.

Reversible terminators

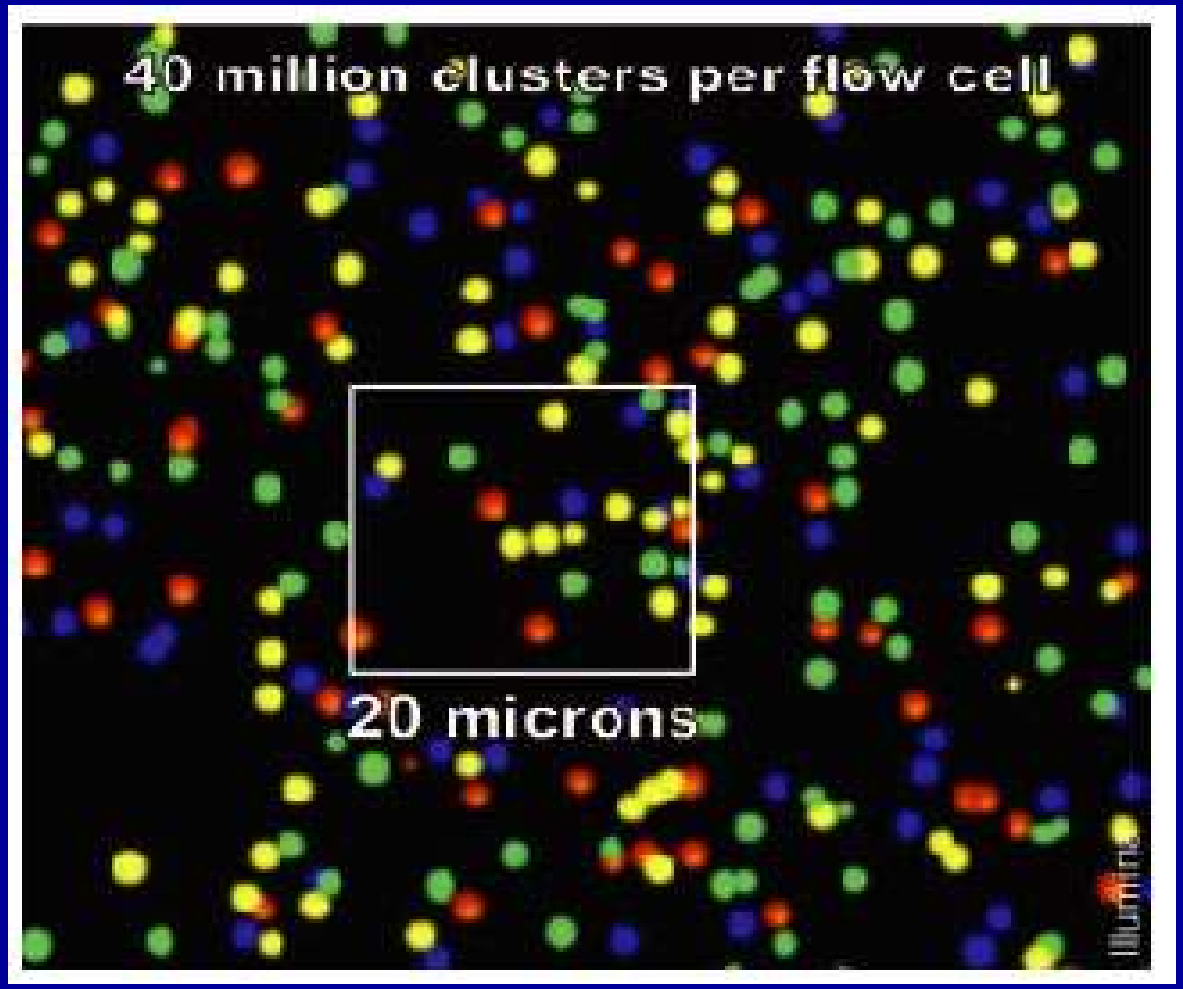


Metzker (2010) Sequencing technologies — the next generation. Nature Reviews Genetics 11, 31-46.

SBS

- Similar to pyrosequencing…
- Instead of using light to identify when a nucleotide was added…
- …add all four dNTPs with, each with a different fluorescent probe terminator!
  - Prevents further elongation
- Remove unincorporated dNTPs.
- *Click*…identify which fluor is present.
- Cleave fluor off and repeat the process.

SBS plate



Chi, KR. (2008) The year of sequencing. Nature Methods, 5:11-15.

Illumina SBS

- Actually a little more complicated.
- Uses something called bridge amplification” on a matrix to form clusters of replicants of each fragment.
- Imaging happens for all clusters simultaneously.

SBS

- Good description on the Illumina Web site:
- http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Illlumina SBS

Pros

- Fast!
  - 50 Gbp per 36 hours
- Cheaper than pyrosequencing
- More accurate than pyrosequencing?
- Good in GC-rich regions

Illlumina SBS

Pros

- Fast!
  - 50 Gbp per 36 hours
- Cheaper than pyrosequencing
- More accurate than pyrosequencing?
- Good in GC-rich regions

Cons

- Smaller sequence length
  - 30 bp per read
- 1.5 to 9 days per run
- Slightly error prone

Illumina HiSeq 2500/1500

- With read length 1 x 36, dual flow cell
  - 2 days (twice 454 example)
  - 95-105 Gb
- Rapid Run Mode
  - 1 sample, human genome, 30x coverage
  - 27 hours
- As reported by Illumina

HiSeq X Ten

- 2 x 150bp
- 1.8 Gb run
- Up to 6 billion reads
- $1000 human genome

HiSeq X Ten

- 2 x 150bp
- 1.8 Gb run
- Up to 6 billion reads
- $1000 human genome
- Requires purchasing 10 machines...

Pyro vs. SBS

- Both are prone to errors, especially alignment-based resulting in deletions and fragmentation.
- See Ye et al., 2011. (reading for next week)

Ion Torrent

- Similar to 454
- Instead of detecting PPi and light given off enzymatically...
- ...detect the hydrogen ion given off by the nucleotide addition.
- Super sensitive ion (pH) detector.

Ion Torrent

- Similar to 454
- Instead of detecting PPi and light given off enzymatically...
- ...detect the hydrogen ion given off by the nucleotide addition.
- Super sensitive ion (pH) detector.
- Same homopolymer issues as 454.
- Shorter reads than 454.

Ion Torrent

- Similar to 454
- Instead of detecting PPi and light given off enzymatically...
- ...detect the hydrogen ion given off by the nucleotide addition.
- Super sensitive ion (pH) detector.
- Same homopolymer issues as 454.
- Shorter reads than 454.
- About 20x cheaper than 454 (2011 data).

Ion S5

- 60-80 million reads
- Up to 15 Gb per run
- 2.5 hours per run

Long-fragment technologies

MinION

- Oxford Nanopore
- Microfluidics and strand sequencing
- Claimed to be
  - $900
  - USB powered
  - Capable of genome sequencing one sample
- Announced in Feb '12 to be on the market “this year”.

MinION

- Oxford Nanopore
- Strand sequencing
- Claimed to be
  - $900
  - USB powered
  - Capable of genome sequencing one sample
- Announced in Feb '12 to be on the market “this year”.
- On market since May 2015…

MinION

- Up to 200 Kb reads
- Tiny chip...no machine
- Up to 48 hours runtime
- Up to 2.2M reads at 10Kb at standard speed
  - 4.4M in fast mode
- Up to 42 Gb in 48 hour run

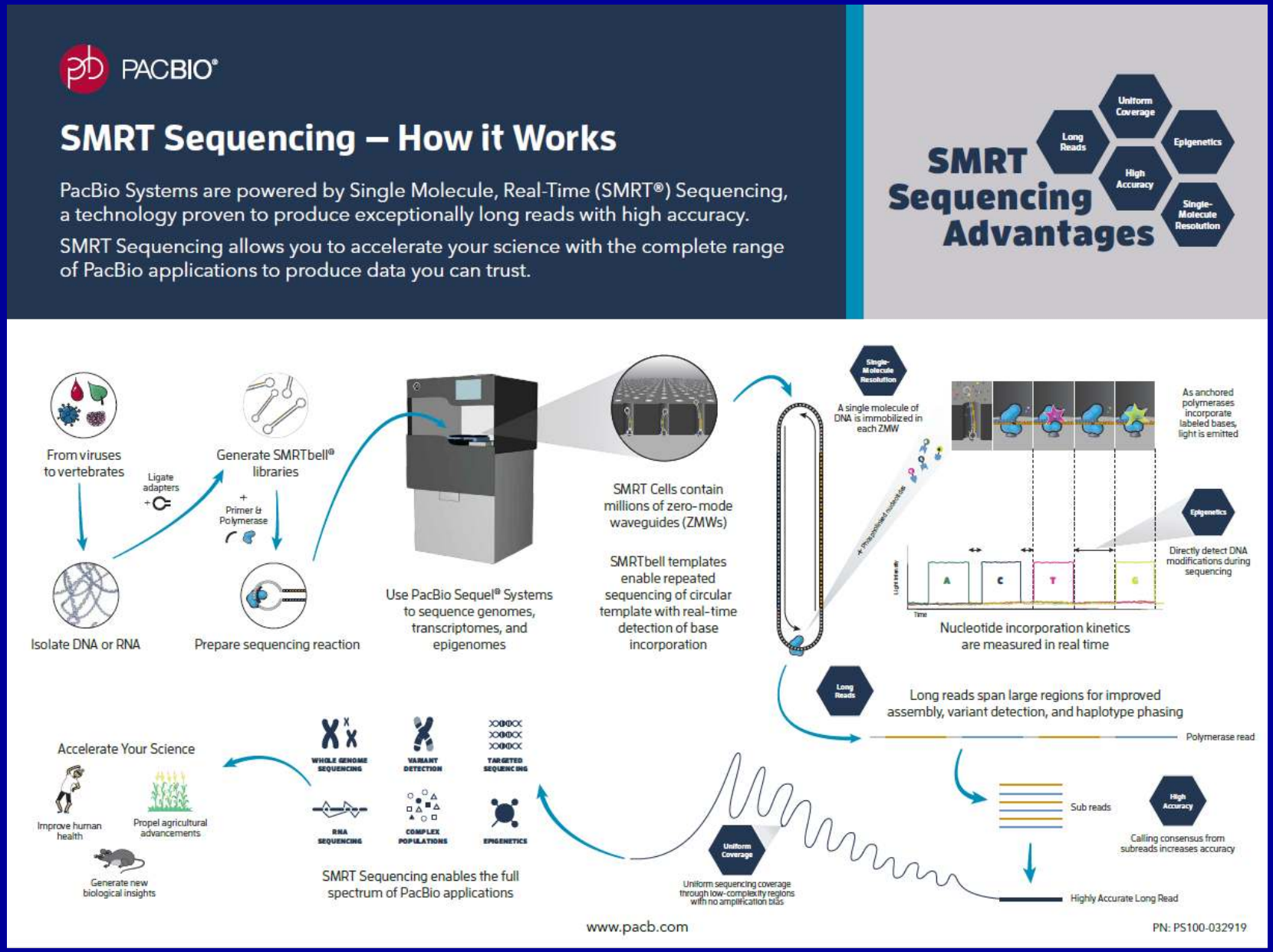
MinION

- Up to 200 Kb reads
- Tiny chip…no machine
- Up to 48 hours runtime
- Up to 2.2M reads at 10Kb at standard speed
  - 4.4M in fast mode
- Up to 42 Gb in 48 hour run
- Costs $270-900 per run
- High error rate (<99% accuracy)

PacBio SMRT

- https://www.pacb.com/technology/hifi-sequencing/how-it-works/

PacBio


PACBIO®
SMRT Sequencing – How it Works
PacBio Systems are powered by Single Molecule, Real-Time (SMRT®) Sequencing, a technology proven to produce exceptionally long reads with high accuracy.
SMRT Sequencing allows you to accelerate your science with the complete range of PacBio applications to produce data you can trust.
SMRT Sequencing Advantages
Uniform Coverage
Long Reads
Epigenetics
High Accuracy
Single-Molecule Resolution
From viruses to vertebrates
Isolate DNA or RNA
Ligate adapters
Generate SMRTbell® libraries
Primer & Polymerase
Prepare sequencing reaction
Use PacBio Sequel® Systems to sequence genomes, transcriptomes, and epigenomes
SMRT Cells contain millions of zero-mode waveguides (ZMWs)
SMRTbell templates enable repeated sequencing of circular template with real-time detection of base incorporation
A single molecule of DNA is immobilized in each ZMW
As anchored polymerases incorporate labeled bases, light is emitted
Directly detect DNA modifications during sequencing
Nucleotide incorporation kinetics are measured in real time
Long reads span large regions for improved assembly, variant detection, and haplotype phasing
Polymerase read
Sub reads
Calling consensus from subreads increases accuracy
Highly Accurate Long Read
Uniform sequencing coverage through low-complexity regions with no amplification bias
Accelerate Your Science
Improve human health
Propel agricultural advancements
Generate new biological insights
WHOLE GENOME SEQUENCING
VARIANT DETECTION
TARGETED SEQUENCING
RNA SEQUENCING
COMPLEX POPULATIONS
EPIGENETICS
SMRT Sequencing enables the full spectrum of PacBio applications
www.pacb.com
PN: PS100-032919

Long-read error rates

- 90-95% is an A-, right?
  - That's great!

Long-read error rates

- 90-95% is an A-, right?
  - That's great!
- Is it so great?

Long-read error rates

- 90-95% is an A-, right?
  - That's great!
- Is it so great?

- Now getting 98%+ accuracy

RNASeq and Analysis

Why RNASeq?

- Which genes are "up/down regulated"?
- Microarrays were king for over a decade

Why RNASeq?

- Which genes are "up/down regulated"?
- Microarrays were king for over a decade
- Microarrays have low reproducibility
- Microarrays also have limits of detection at extreme high/low expression

RNASeq pro/con

- Has improved reproducibility
- No limit to high/low expression
  - Quantitative
  - Low solved by increased sequencing depth

RNASeq pro/con

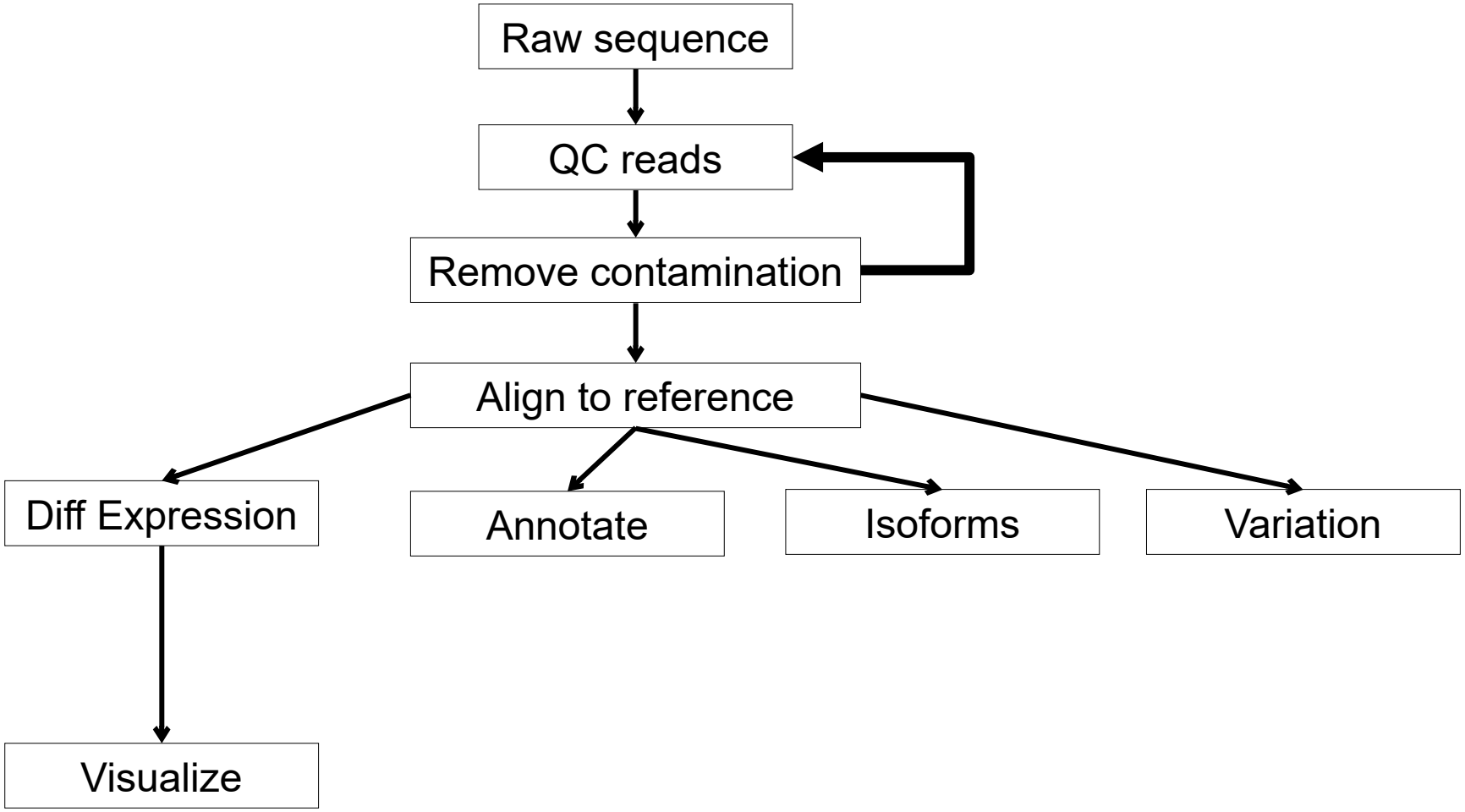
- Has improved reproducibility
- No limit to high/low expression
  - Quantitative
  - Low solved by increased sequencing depth
- Some issues with assigning fragments to specific loci (later with multireads)
- More expensive *

* As of early 2012, almost even...swapped since.

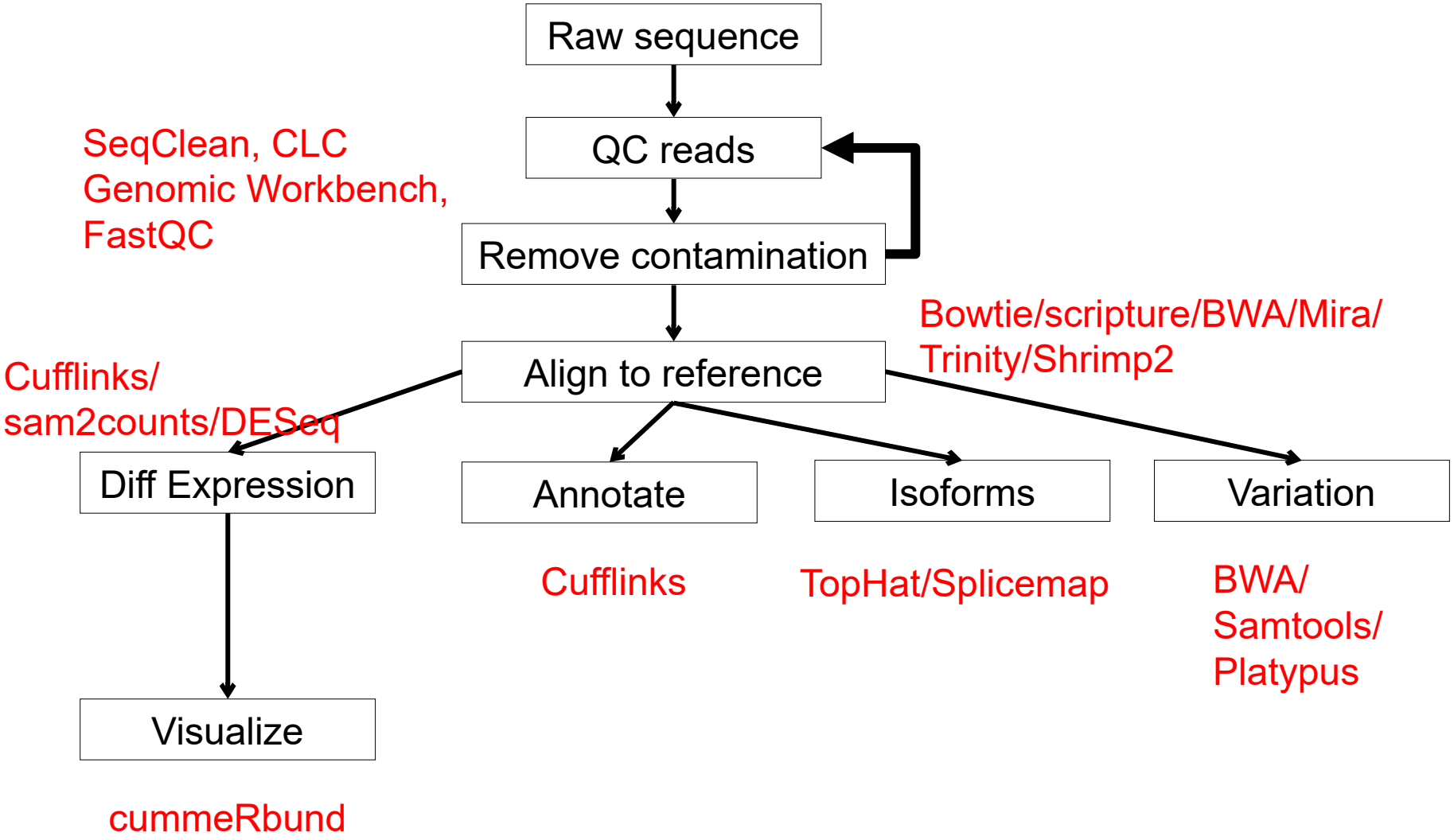
General concerns

- Removing contamination
  - host, vector, other species etc.
- Genome/Transcriptome reference sequence?
- Isoforms of genes
- Multireads
- Variation or sequence error?

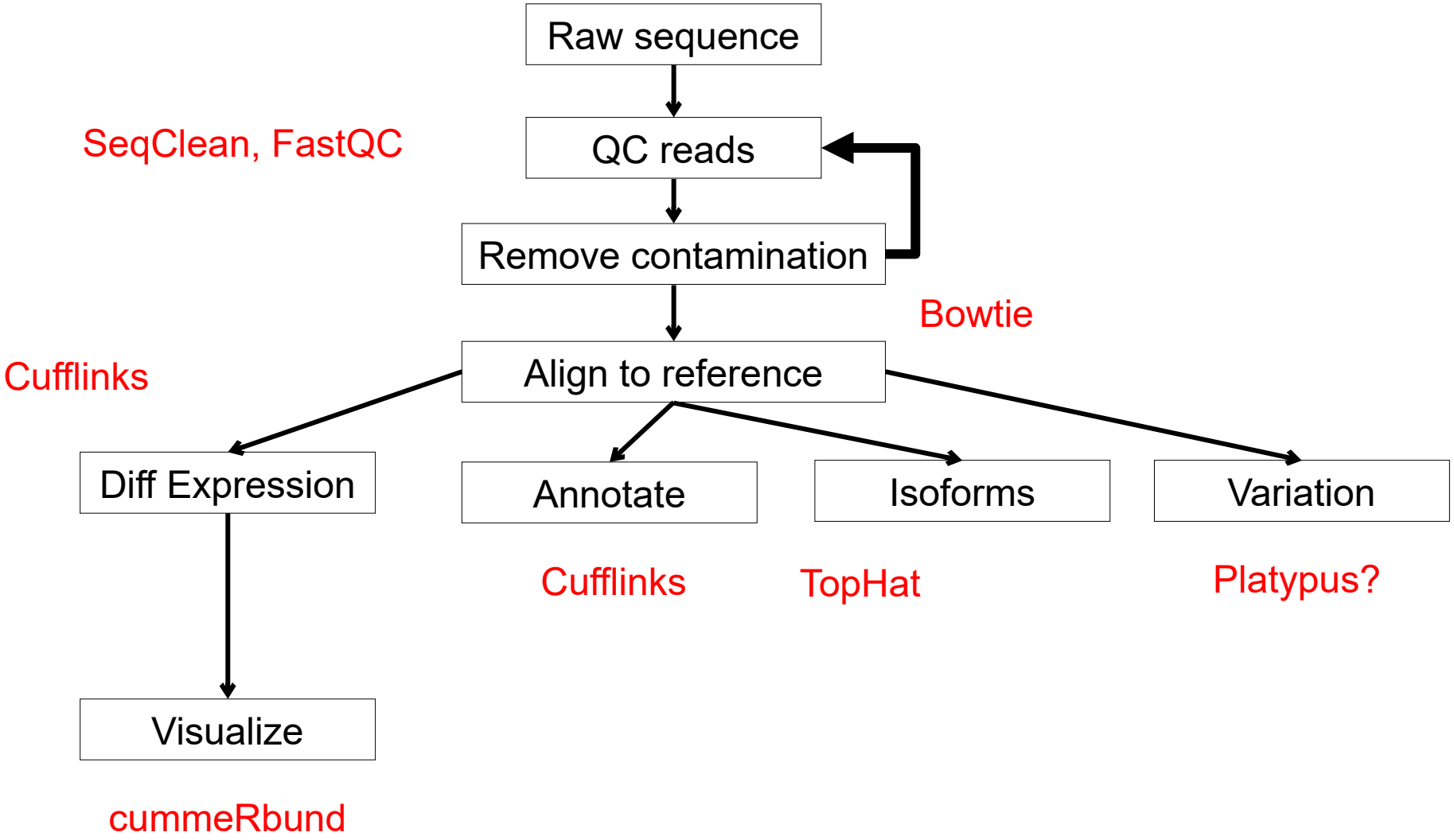
Typical pipeline



Typical pipeline



Typical pipeline

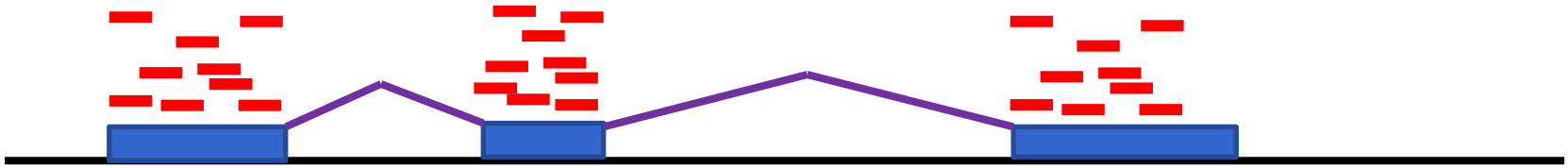


Remove contaminants

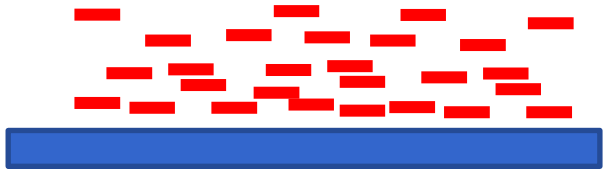
- Vector
- Mixed samples
  - Host/pathogen

Align to reference

Genomic reference

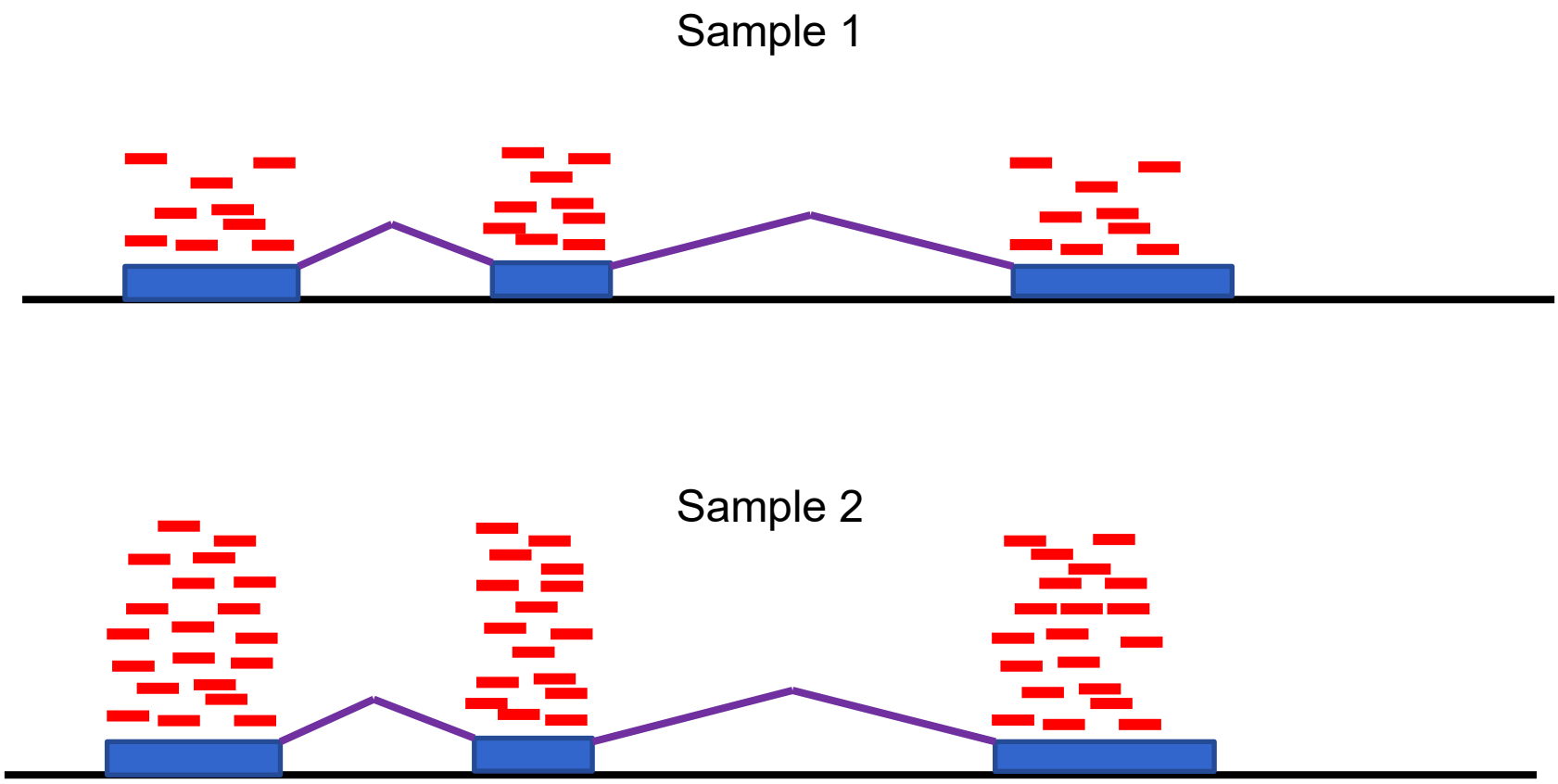


Transcriptome reference

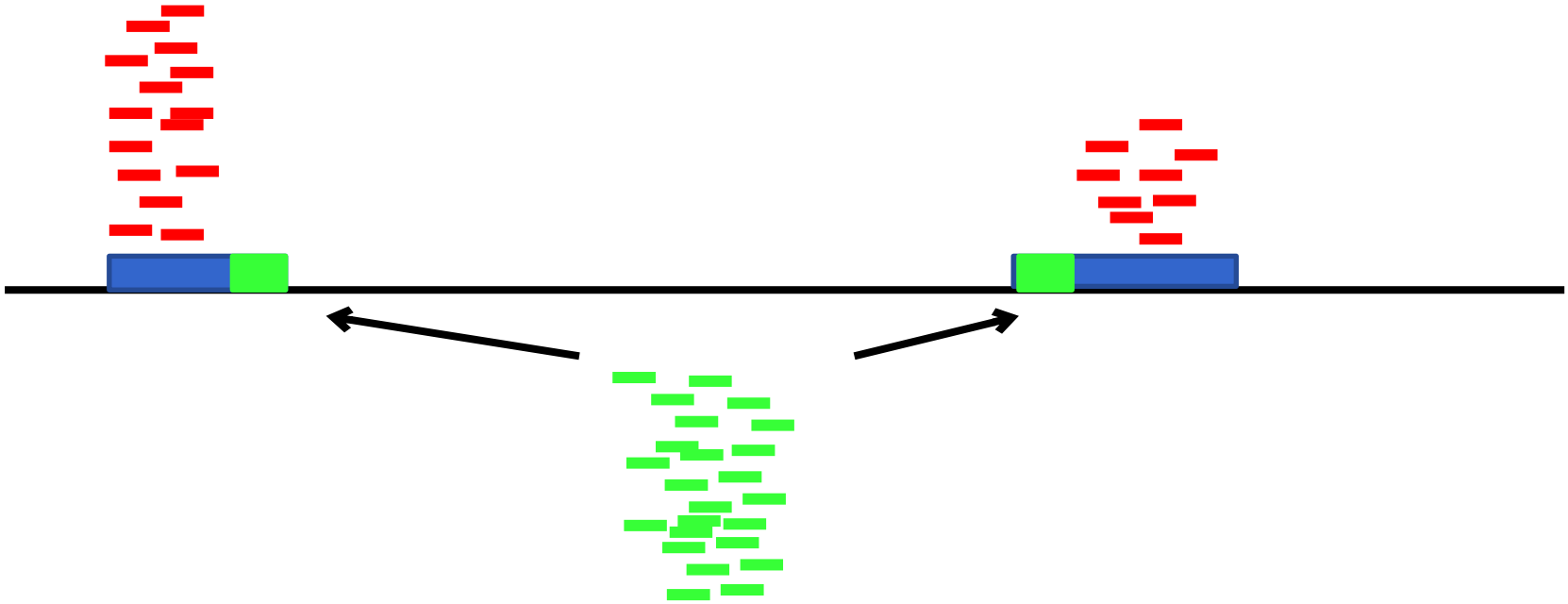


De novo assembly...

Differential Expression

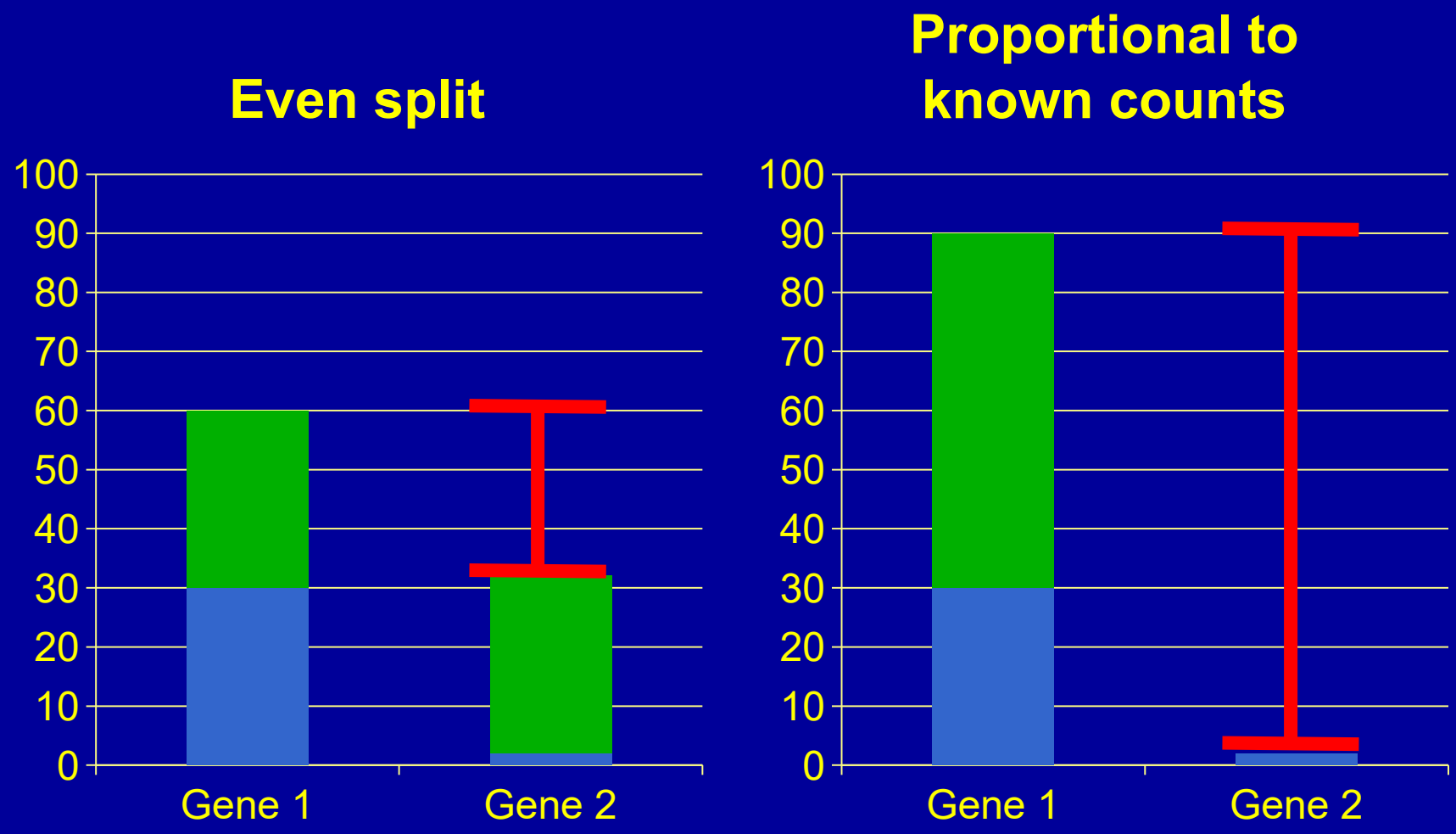


Multireads

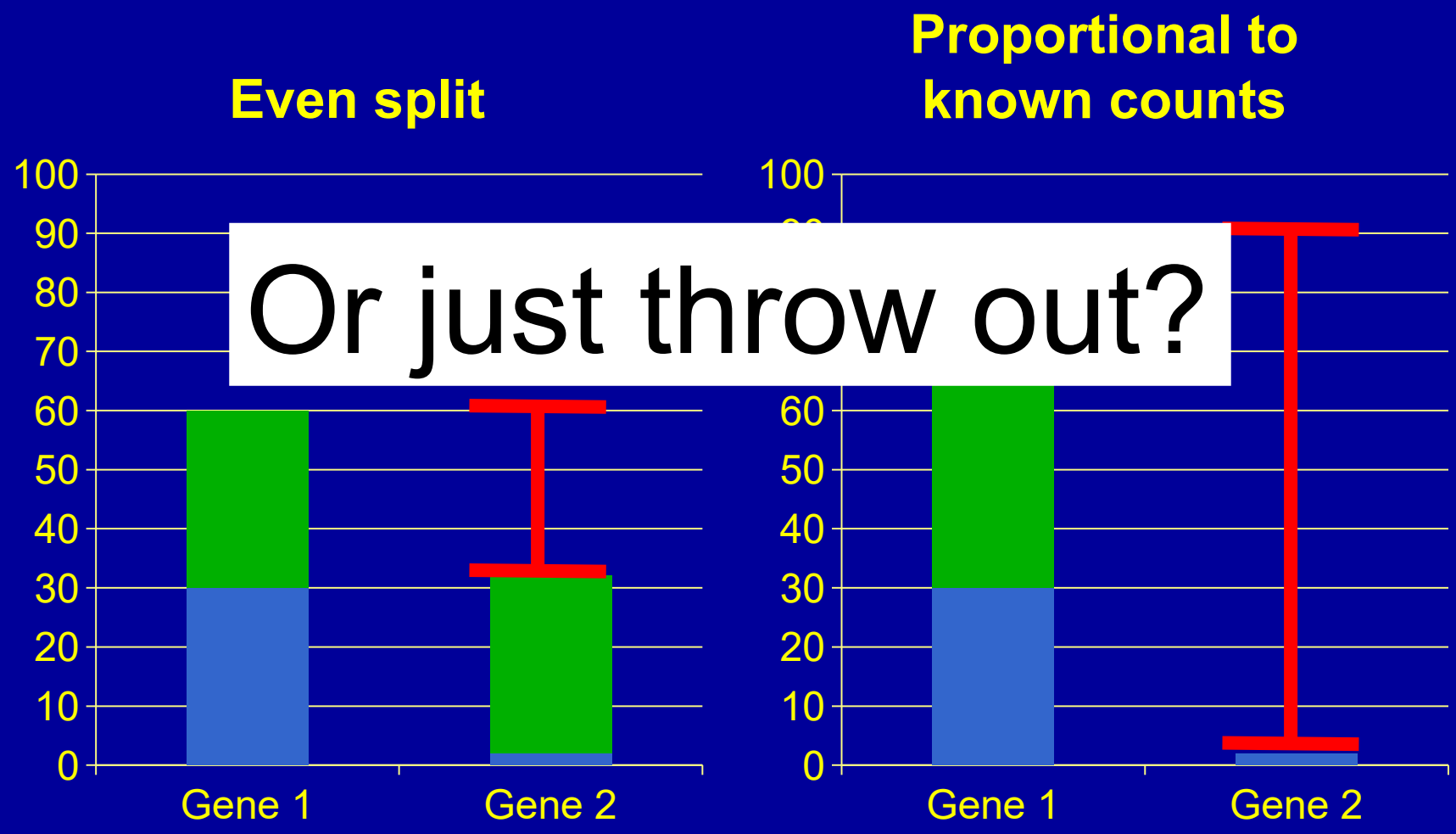


Could be assigned to either gene or split between them!

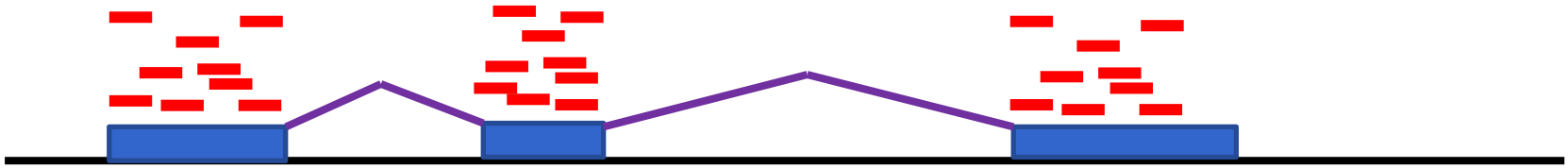
Multireads


Even split
Proportional to known counts
100
90
80
70
60
50
40
30
20
10
0
Gene 1
Gene 2
100
90
80
70
60
50
40
30
20
10
0
Gene 1
Gene 2

Multireads

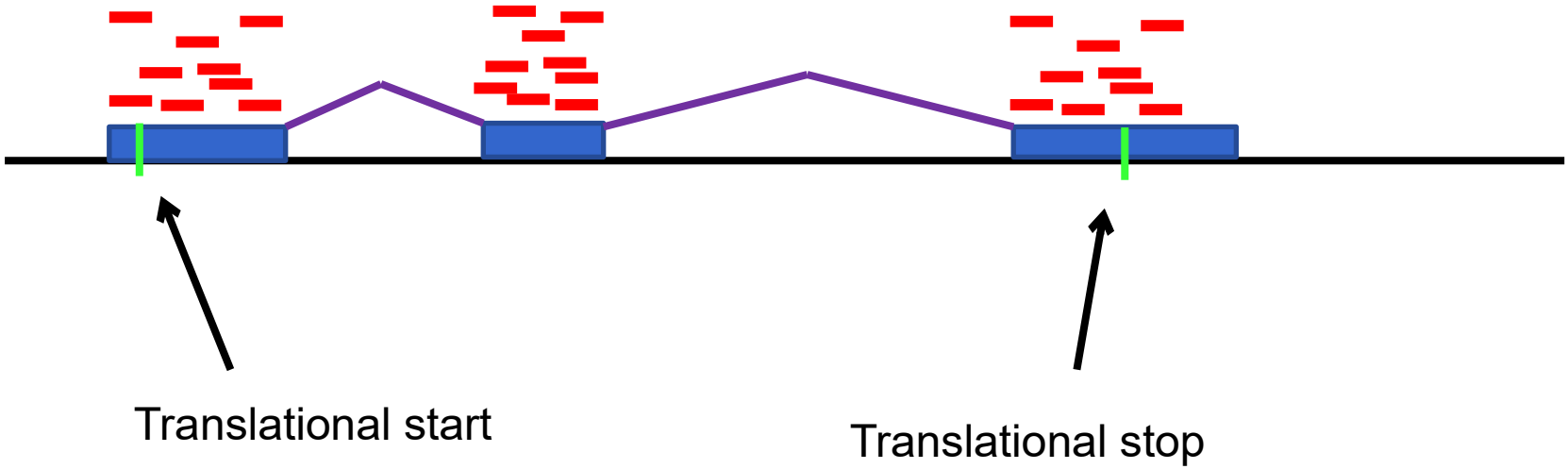

Even split
Proportional to known counts
100
90
80
70
60
50
40
30
20
10
0
Gene 1
Gene 2
Or just throw out?

Annotation

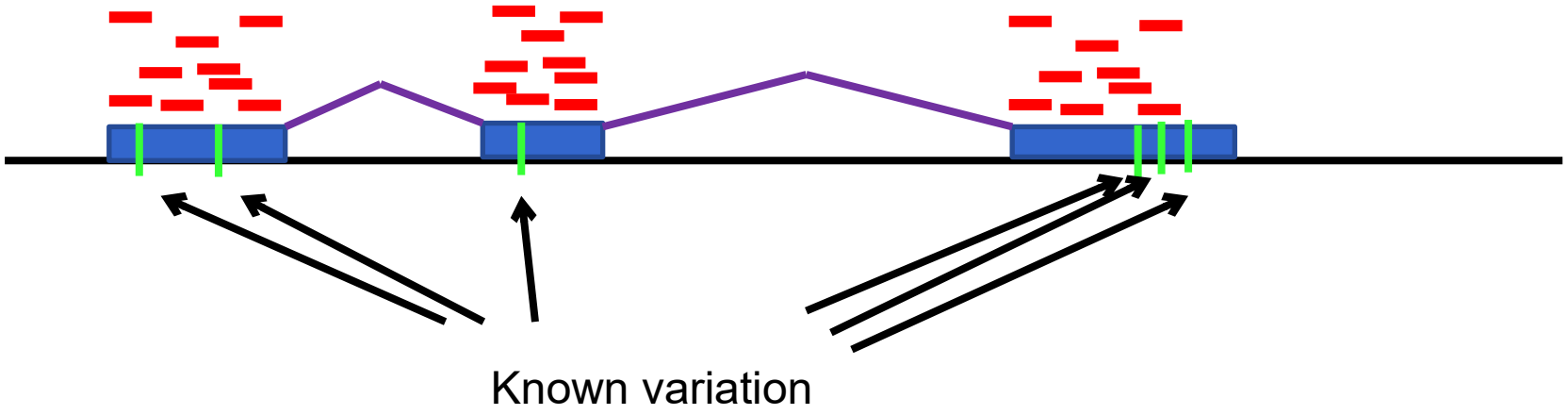


Gene name
Gene product function/location (GO)
Metabolic pathways (KEGG)
Etc...

Annotation

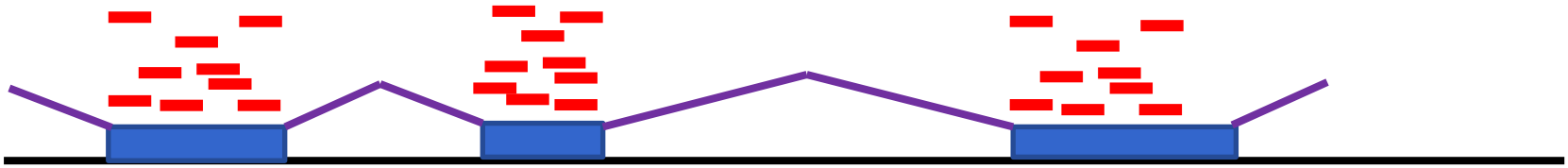


Annotation



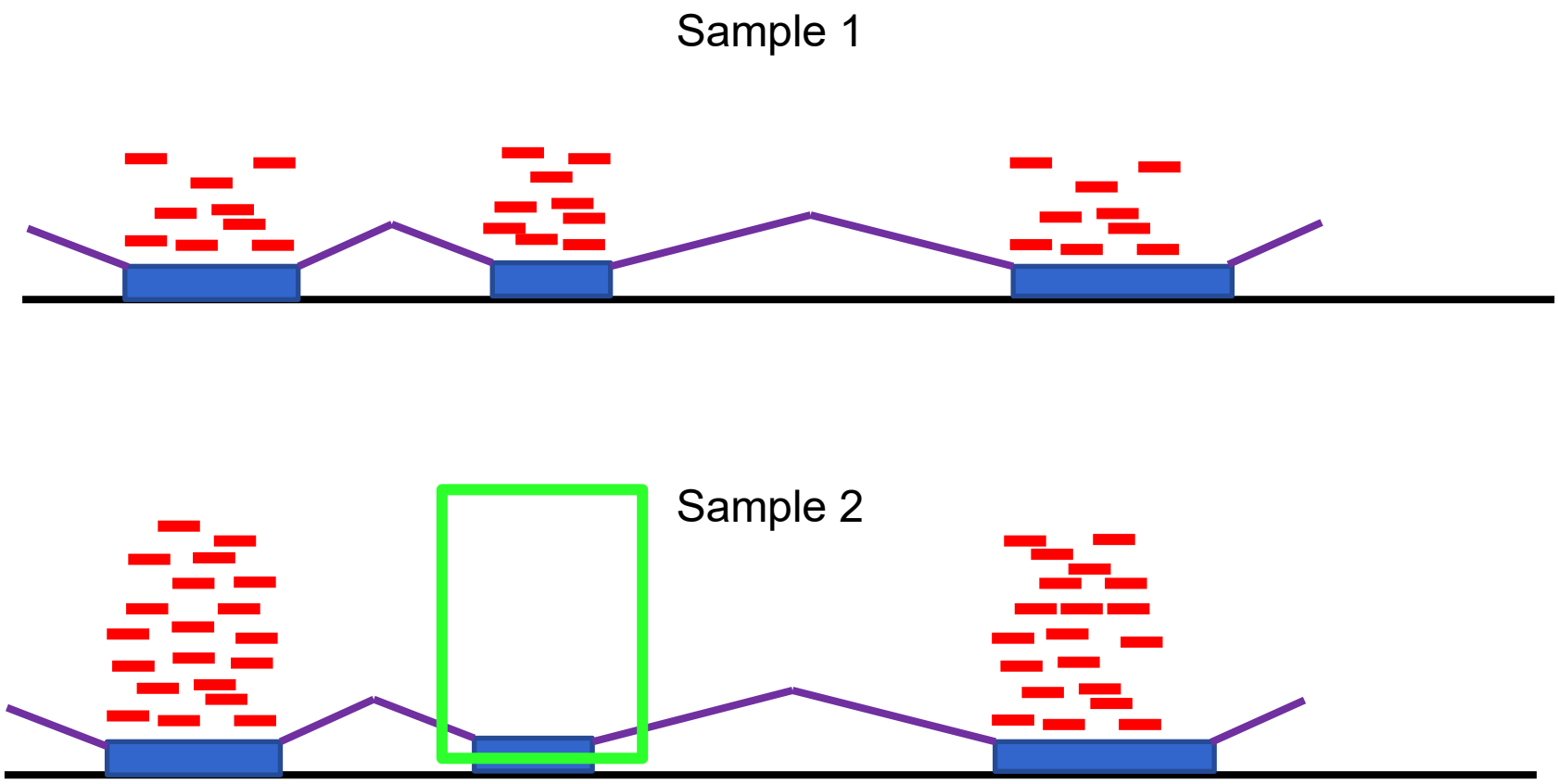
Isoforms

Sample 1

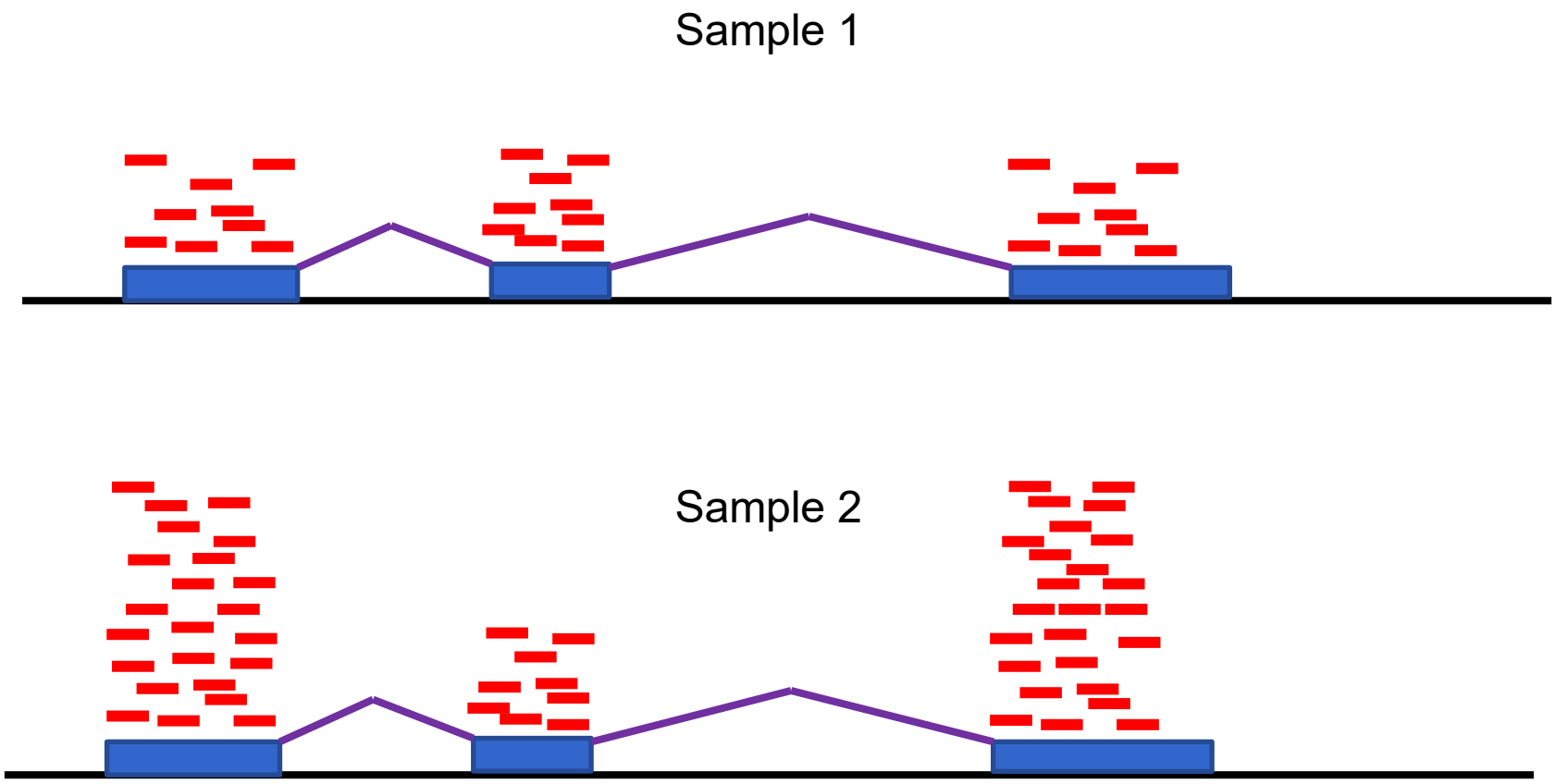


Sample 2

Isoforms



Isoforms? Error?



Variation or error?

Summary

- RNASeq analysis is sophisticated.
- Need to adjust for many potential sources of erroneous data.
- This is not the only method of using HTS.
  - Whole genome sequencing
  - Metagenomics
  - eDNA
  - scRNASeq
  - epigenetics