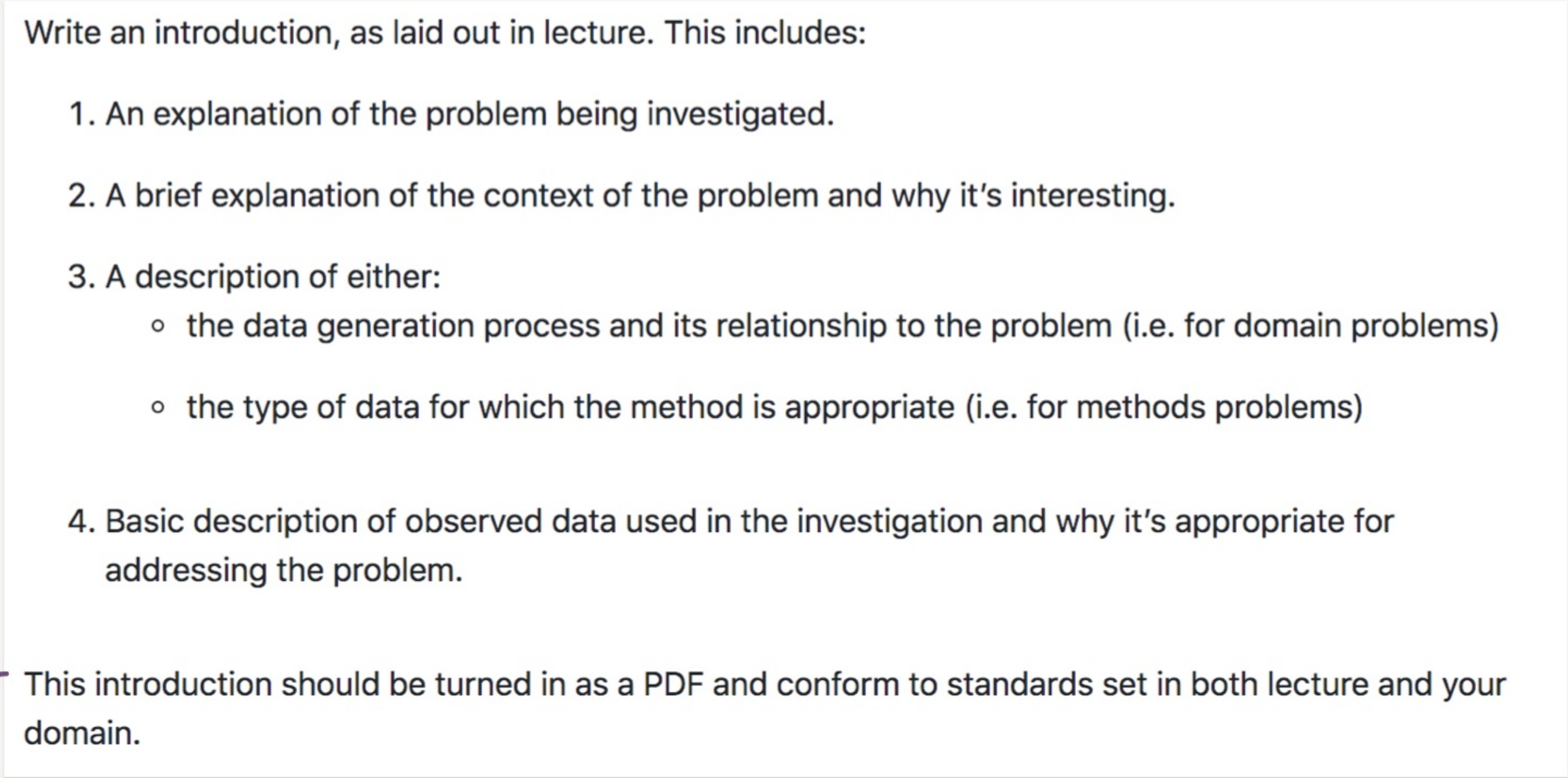
Checkpoint #1 Expectations

Report



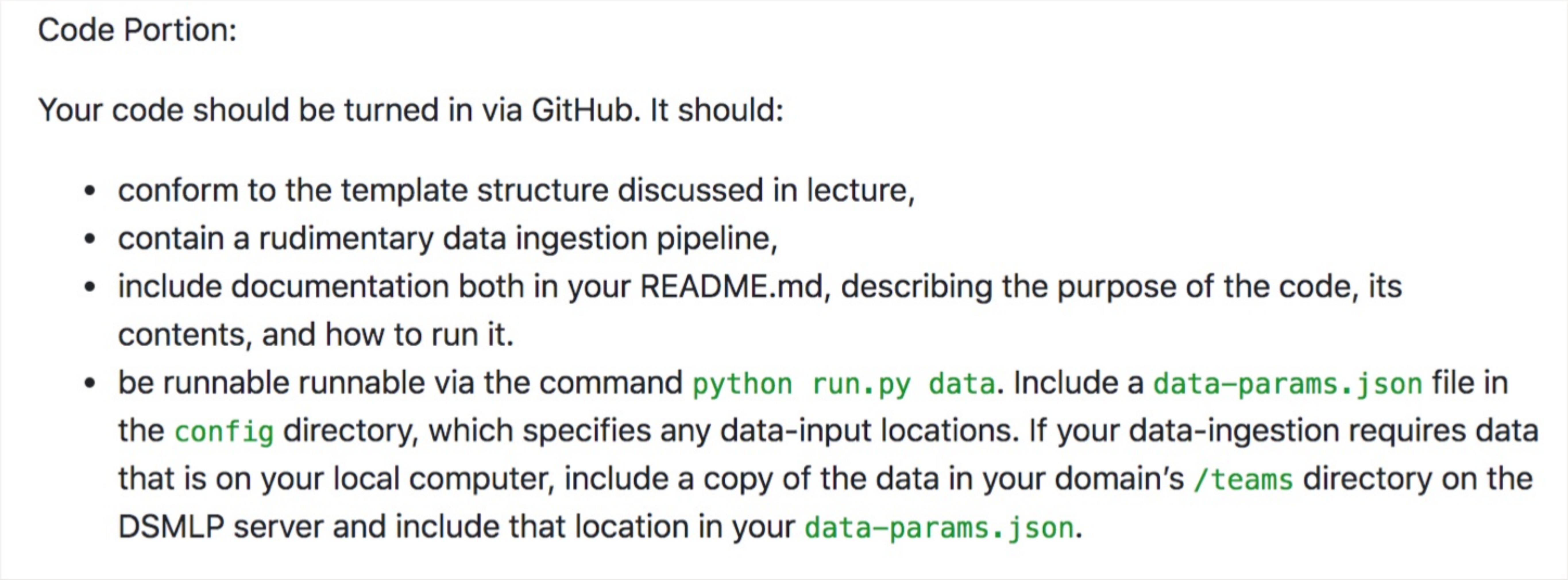
Write an introduction, as laid out in lecture. This includes:

1. An explanation of the problem being investigated.
2. A brief explanation of the context of the problem and why it's interesting.
3. A description of either:
    - the data generation process and its relationship to the problem (i.e. for domain problems)
    - the type of data for which the method is appropriate (i.e. for methods problems)
4. Basic description of observed data used in the investigation and why it's appropriate for addressing the problem.

This introduction should be turned in as a PDF and conform to standards set in both lecture and your domain.

__Note__: track decisions

Code



Code Portion:

Your code should be turned in via GitHub. It should:

- conform to the template structure discussed in lecture,
- contain a rudimentary data ingestion pipeline,
- include documentation both in your README.md, describing the purpose of the code, its contents, and how to run it.
- be runnable runnable via the command `python run.py data`. Include a `data-params.json` file in the `config` directory, which specifies any data-input locations. If your data-ingestion requires data that is on your local computer, include a copy of the data in your domain's `/teams` directory on the DSMLP server and include that location in your `data-params.json`.

ligation + library prep

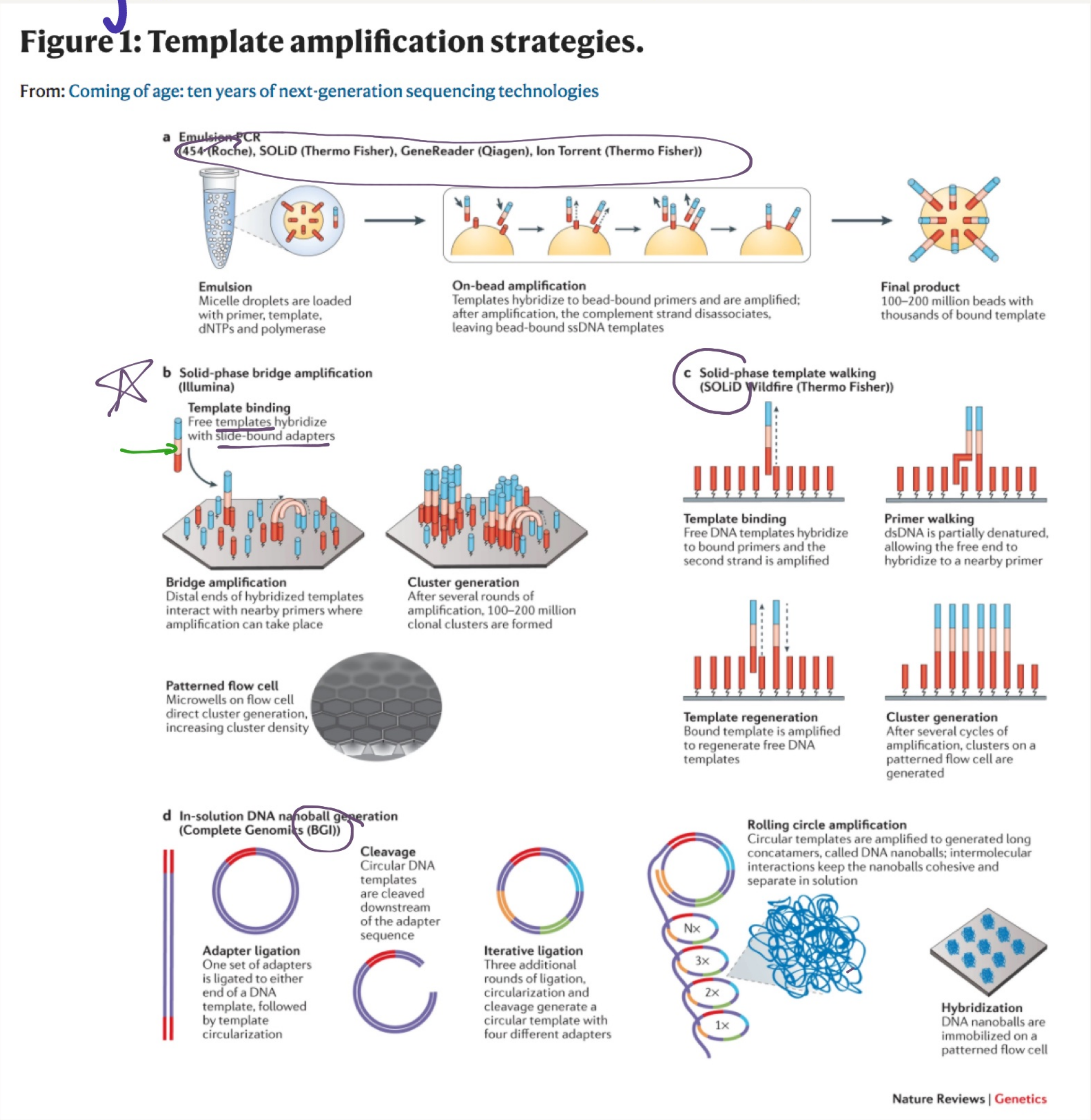


Figure 1: Template amplification strategies.

From: Coming of age: ten years of next-generation sequencing technologies

a Emulsion PCR
(454 (Roche), SOLiD (Thermo Fisher), GeneReader (Qiagen), Ion Torrent (Thermo Fisher))

Emulsion
Micelle droplets are loaded with primer, template, dNTPs and polymerase

On-bead amplification
Templates hybridize to bead-bound primers and are amplified; after amplification, the complement strand disassociates, leaving bead-bound ssDNA templates

Final product
100–200 million beads with thousands of bound template

b Solid-phase bridge amplification
(Illumina)

Template binding
Free templates hybridize with slide-bound adapters

Bridge amplification
Distal ends of hybridized templates interact with nearby primers where amplification can take place

Cluster generation
After several rounds of amplification, 100–200 million clonal clusters are formed

Patterned flow cell
Microwells on flow cell direct cluster generation, increasing cluster density

c Solid-phase template walking
(SOLiD Wildfire (Thermo Fisher))

Template binding
Free DNA templates hybridize to bound primers and the second strand is amplified

Primer walking
dsDNA is partially denatured, allowing the free end to hybridize to a nearby primer

Template regeneration
Bound template is amplified to regenerate free DNA templates

Cluster generation
After several cycles of amplification, clusters on a patterned flow cell are generated

d In-solution DNA nanoball generation
(Complete Genomics (BGI))

Adapter ligation
One set of adapters is ligated to either end of a DNA template, followed by template circularization

Cleavage
Circular DNA templates are cleaved downstream of the adapter sequence

Iterative ligation
Three additional rounds of ligation, circularization and cleavage generate a circular template with four different adapters

Rolling circle amplification
Circular templates are amplified to generated long concatamers, called DNA nanoballs; intermolecular interactions keep the nanoballs cohesive and separate in solution

Hybridization
DNA nanoballs are immobilized on a patterned flow cell

attach == ligate

adapter RNA fragment

attach

template
(sample)

↳ cutadapt

known sequences
(ATCGs)

① Read length

*Illumina
~100 - (1,000)

② Cost

③ Accuracy

2nd/next-gen sequencing = Illumina "reads"

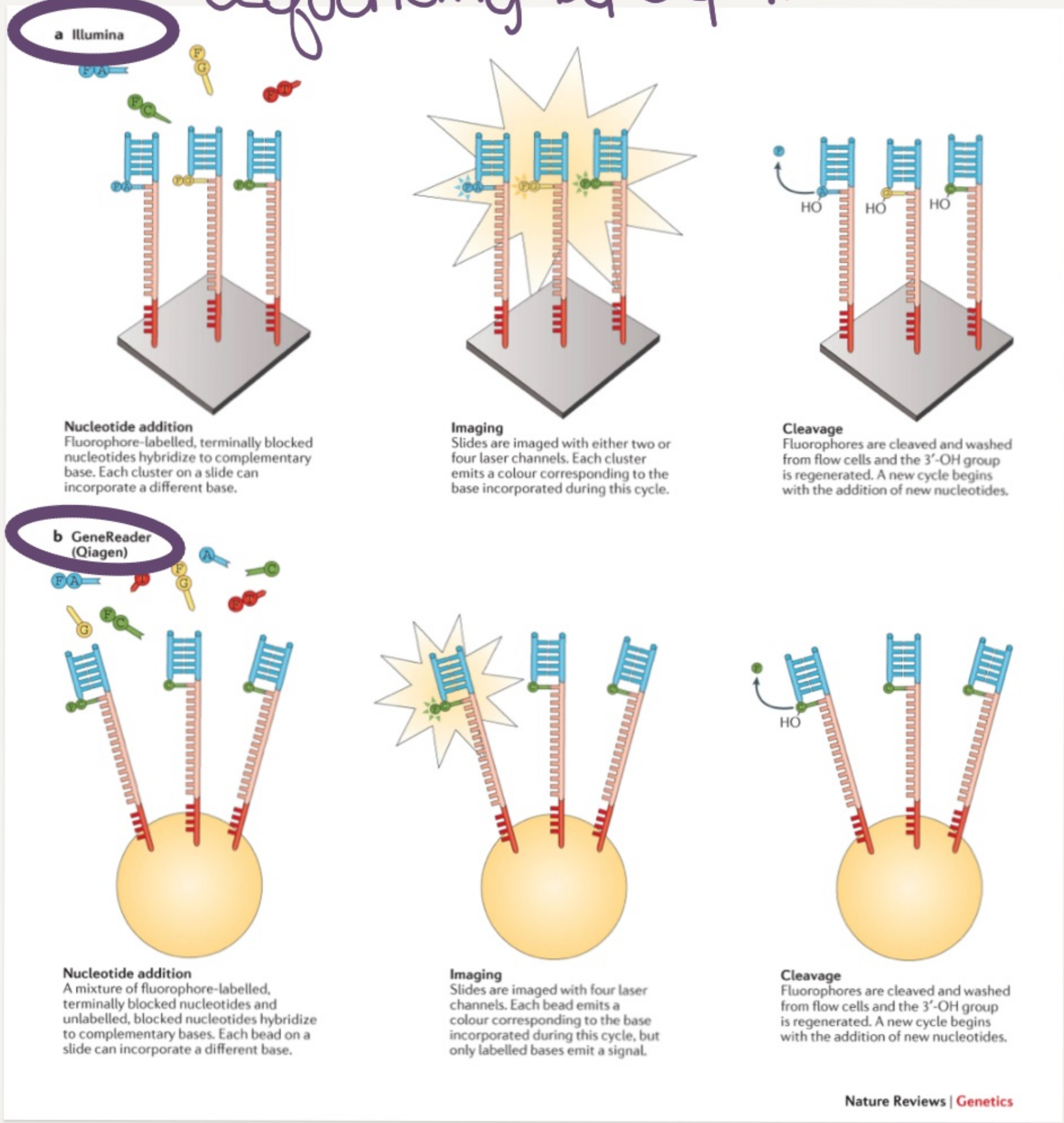
454 vs. Illumina (sequencing technologies)

90s (Sanger sequencing)

Illumina + others Next-gen

Sequencing-by-synthesis

colour-space



Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.

Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.

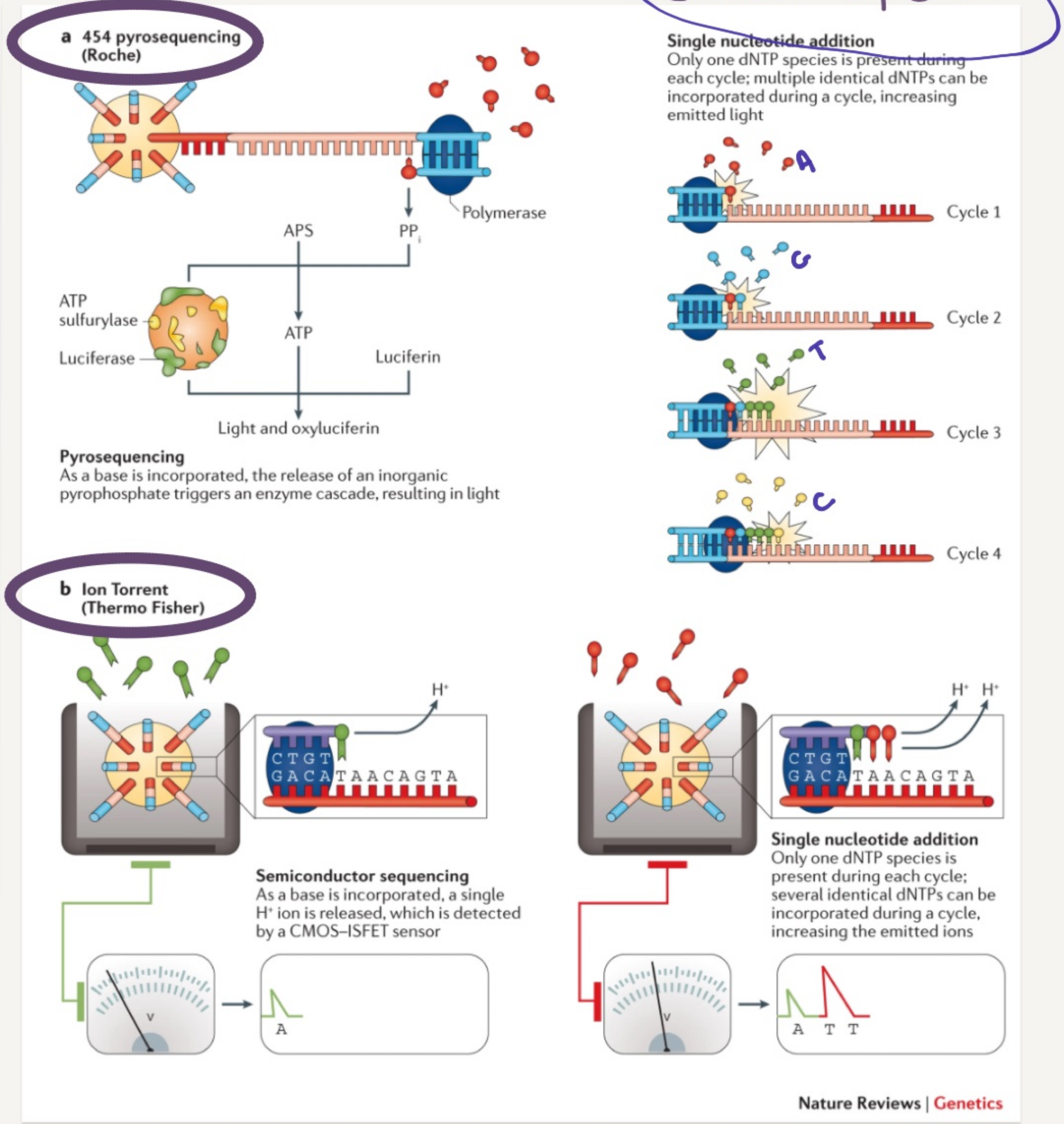
Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

Nucleotide addition
A mixture of fluorophore-labelled, terminally blocked nucleotides and unlabelled, blocked nucleotides hybridize to complementary bases. Each bead on a slide can incorporate a different base.

Imaging
Slides are imaged with four laser channels. Each bead emits a colour corresponding to the base incorporated during this cycle, but only labelled bases emit a signal.

Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

Nature Reviews | Genetics



Pyrosequencing
As a base is incorporated, the release of an inorganic pyrophosphate triggers an enzyme cascade, resulting in light

Single nucleotide addition
Only one dNTP species is present during each cycle; multiple identical dNTPs can be incorporated during a cycle, increasing emitted light

Semiconductor sequencing
As a base is incorporated, a single H+ ion is released, which is detected by a CMOS–ISFET sensor

Single nucleotide addition
Only one dNTP species is present during each cycle; several identical dNTPs can be incorporated during a cycle, increasing the emitted ions

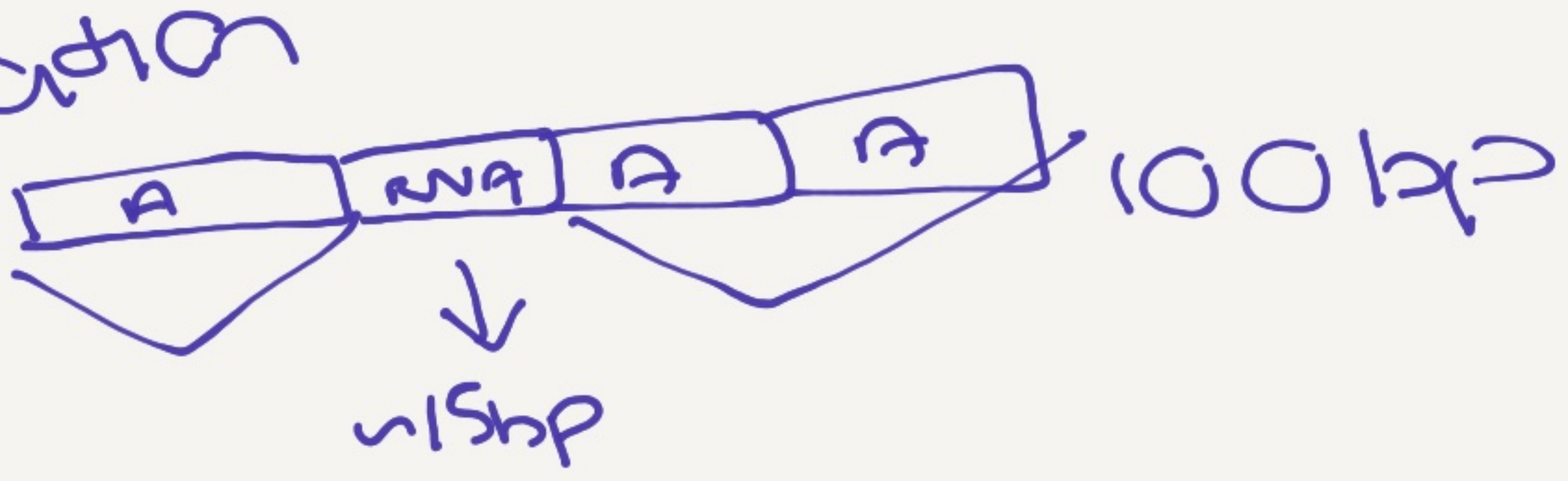
Nature Reviews | Genetics

Crummy seq. data + what to do...

pre-processing

- *no base calls ~~NNNNNNNNATCGNNNN~~ ~100bp read
- *RNA fragmentation (bad library prep) <20bp

- *adapter contamination



- *GC Bias
  (FASTQC)
  fasta → fastq. quality control

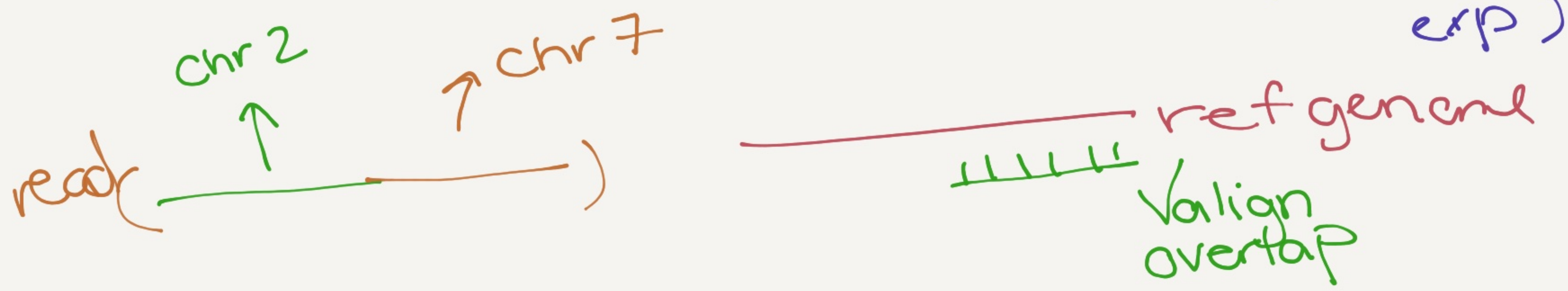
✓ remove reads

alignment
(where in genome?)

- *overlap → ambiguos mapping

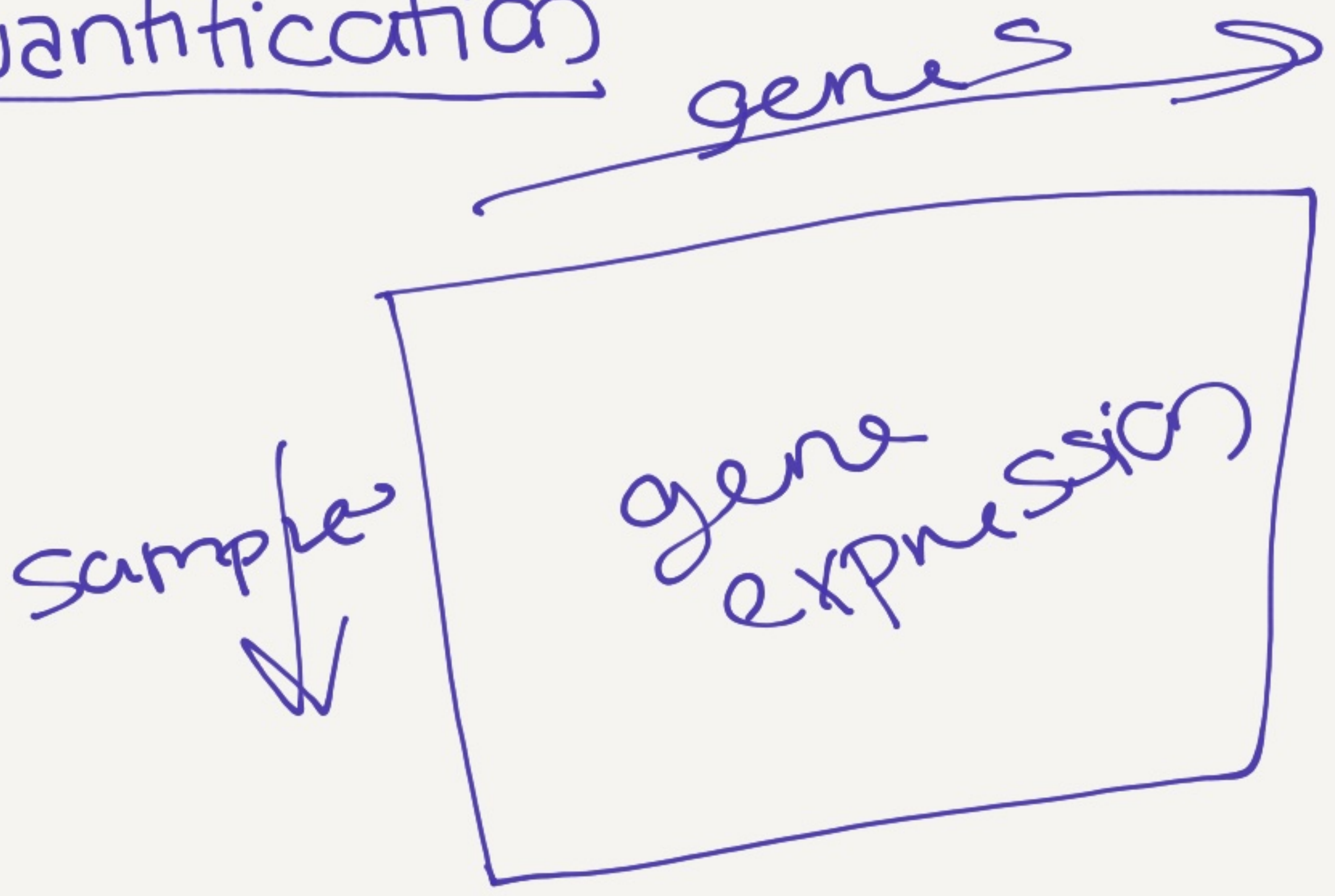
✓ remove sample (PCA genes exp)

*



- *align to multiple locations

quantification



*normalization

cutadapt

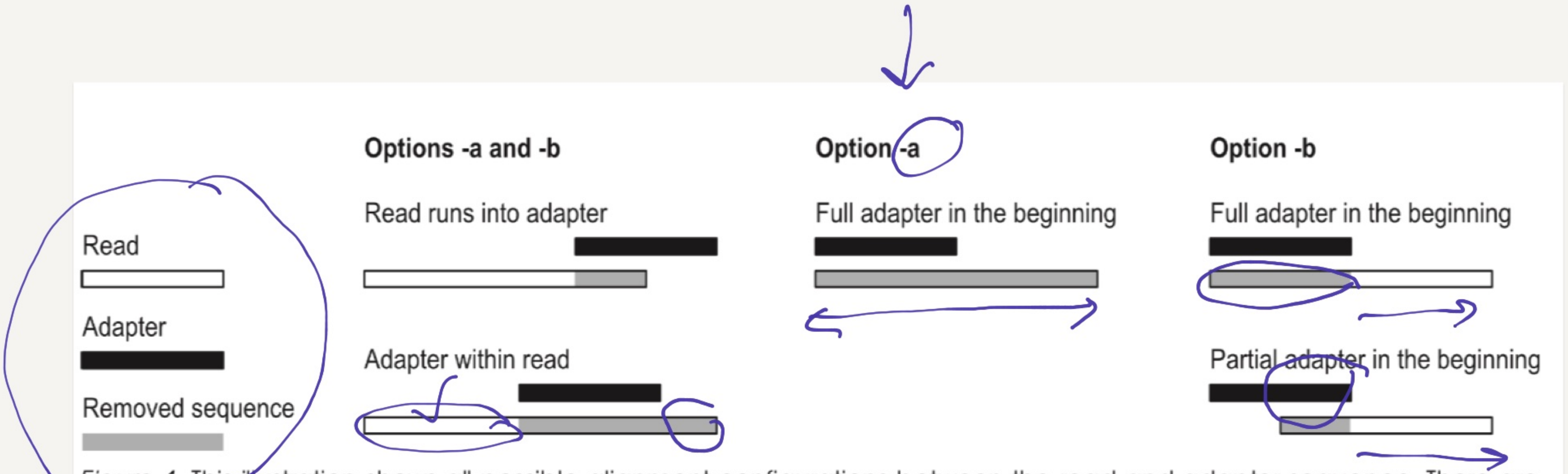
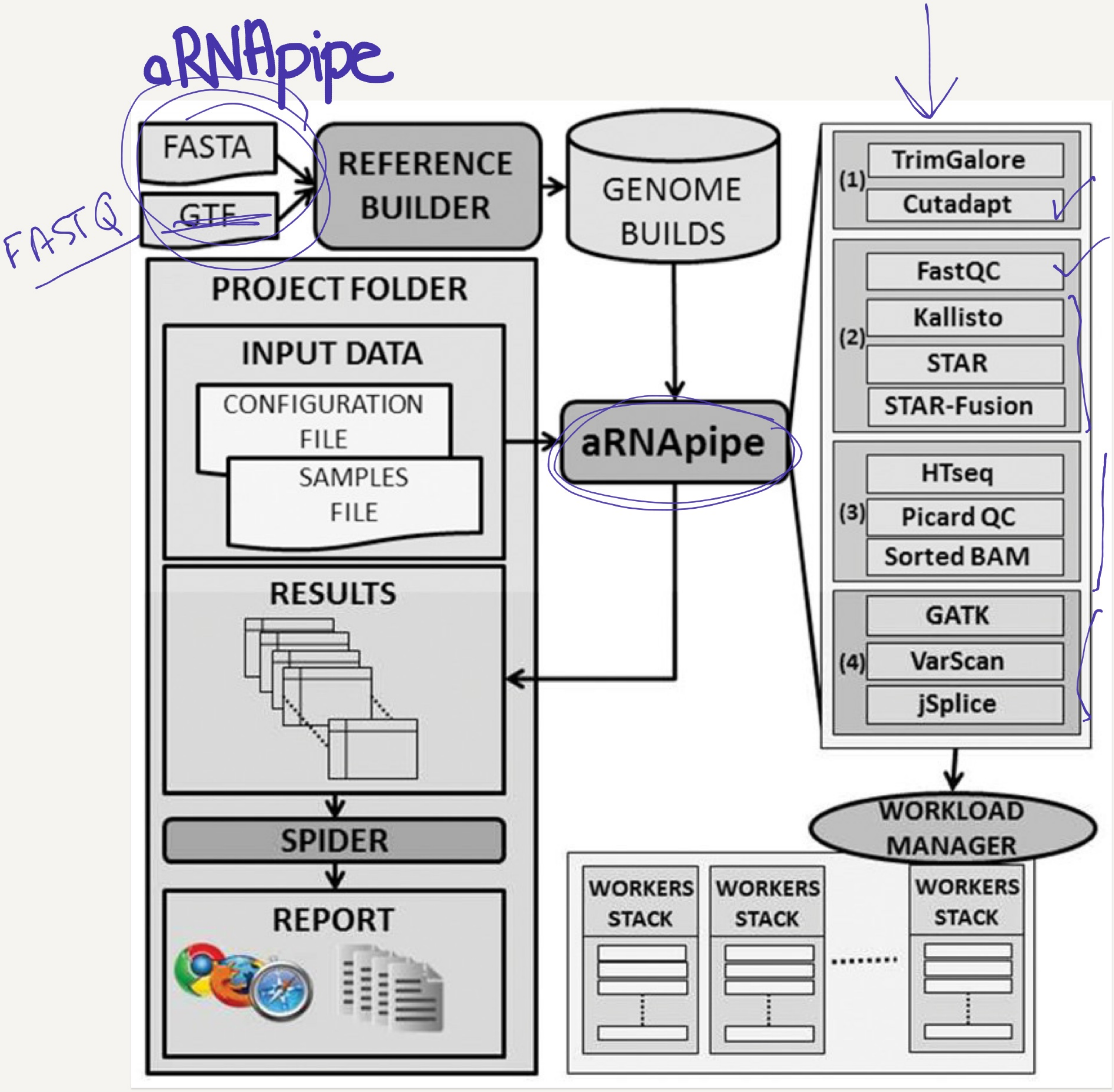


Figure 1. This illustration shows all possible alignment configurations between the read and adapter sequence. There are two different trimming behaviours, triggered by whether option "-a" or "-b" is used to provide the adapter sequence. Note that the case "Partial adapter in the beginning" is not possible with option "-a", as the alignment algorithm prevents it.


aRNApipe
FASTA
~~GTF~~
FASTQ
REFERENCE BUILDER
GENOME BUILDS
PROJECT FOLDER
INPUT DATA
CONFIGURATION FILE
SAMPLES FILE
RESULTS
SPIDER
REPORT
aRNApipe
(1)
TrimGalore
Cutadapt
(2)
FastQC
Kallisto
STAR
STAR-Fusion
(3)
HTseq
Picard QC
Sorted BAM
(4)
GATK
VarScan
jSplice
WORKLOAD MANAGER
WORKERS STACK
WORKERS STACK
WORKERS STACK