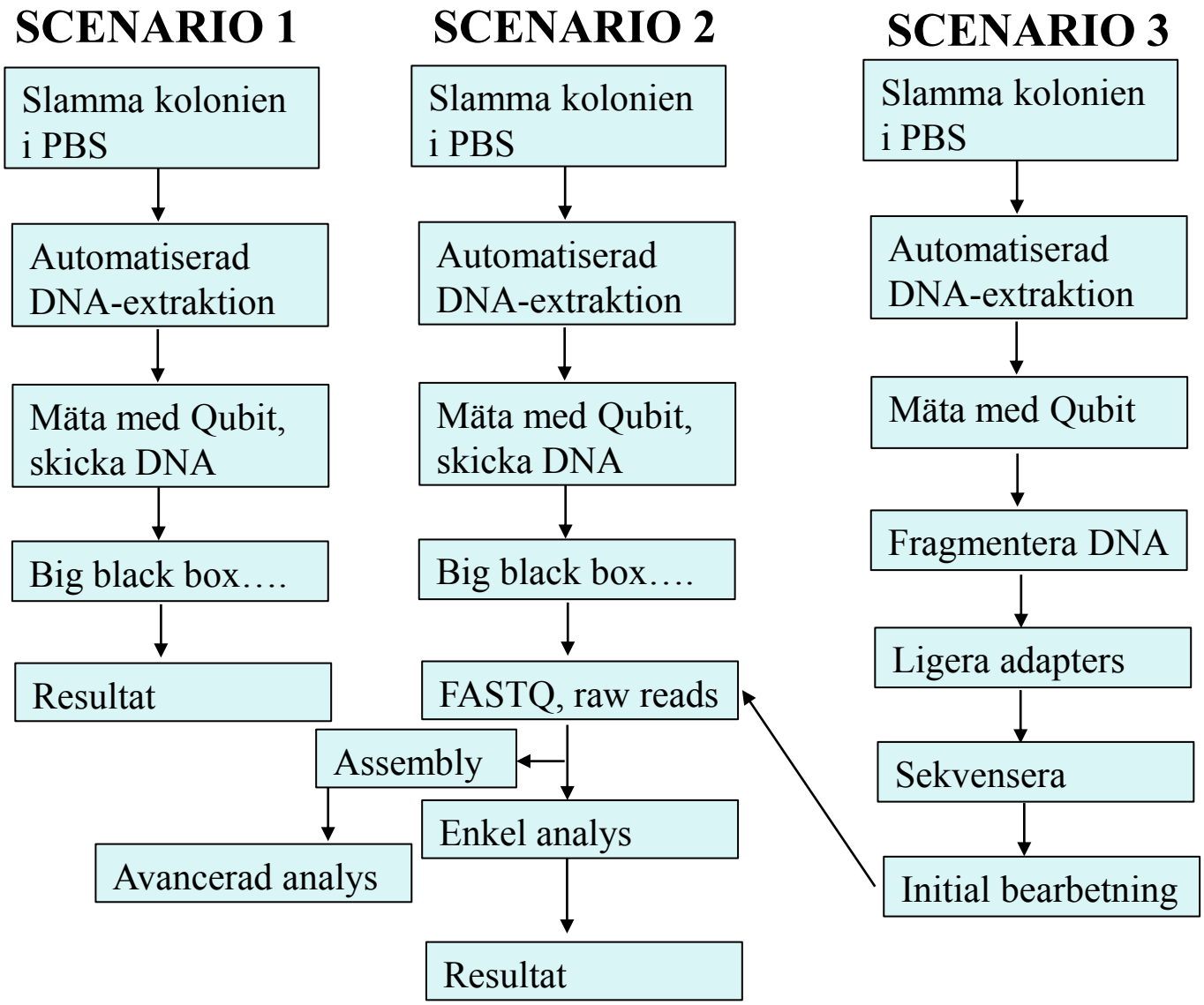
NGS – preanalys och sekvensering

Christian G. Giske, MD/PhD
Docent / Överläkare
Karolinska Institutet och
Karolinska Universitetssjukhuset
23 Maj 2017

Några viktiga ord och begrepp

- Read=korta eller långa DNA sekvenser från körningen
- Contig=samling överlappande DNA segment – framkommer när man sätter ihop (gör assembly) av reads mha bioinformatiska dataprogram
- Sekvensdjup/coverage: genomsnittlig antal gånger man sekvenserar varje bas (ska oftast vara 30-50 x)
- FASTQ: textbaserad fil som innehåller både DNA sekvens och information om kvalitet
  - → Kommer ofta som komprimerad fil (.fastq.gz) – många program kan hantera analys med denna typ, så inte alltid nödvändigt att dekomprimera
  - → Sekvensdata från bara ett isolat kan ofta vara runt 1 GB

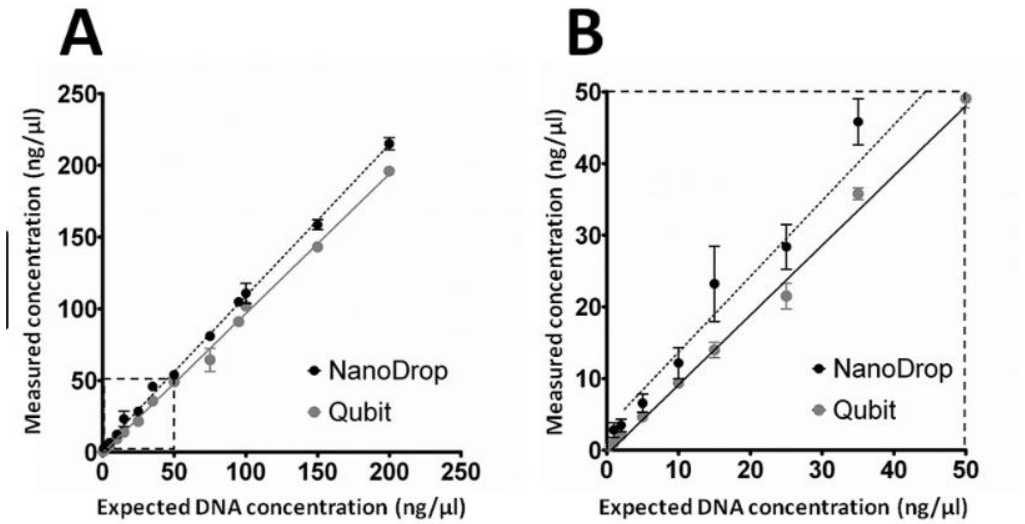
Tänk dig att du har en koloni….



DNA extraktion

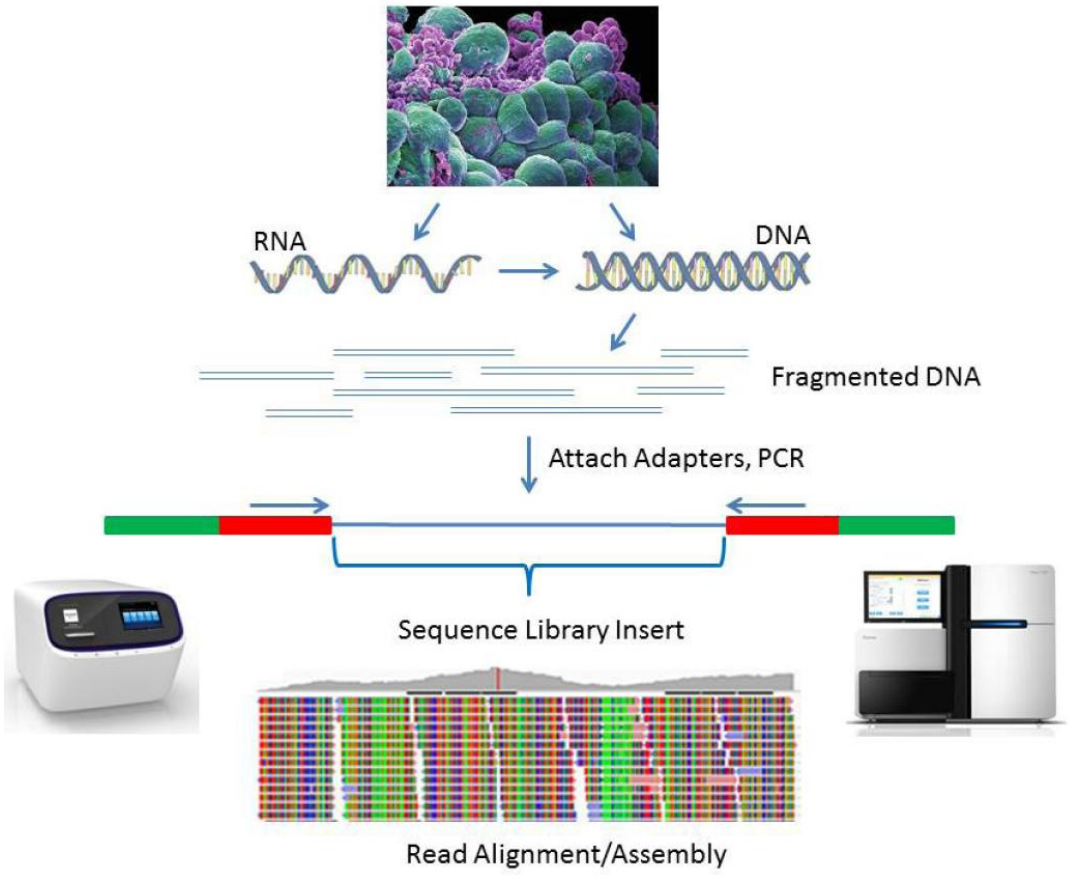
- Slamma några färska kolonier i PBS
- Utför automatiserad extraktion (t ex MP96, Biorobot M48 etc.)
- För vissa arter (t ex *S. aureus*) krävs förbehandling innan extraktion (lysozym, lyostaphin och proteinas K)
- Generellt är svårighetsgraden: svamp > grampositiva bakterier > gramnegativa bakterier
- The proof is in the pudding – mät vad du får ut och modifiera protokollet vid behov
- Väldigt okomplicerat med vanliga gramnegativer

Mäta DNA koncentration: Qubit fluorometer



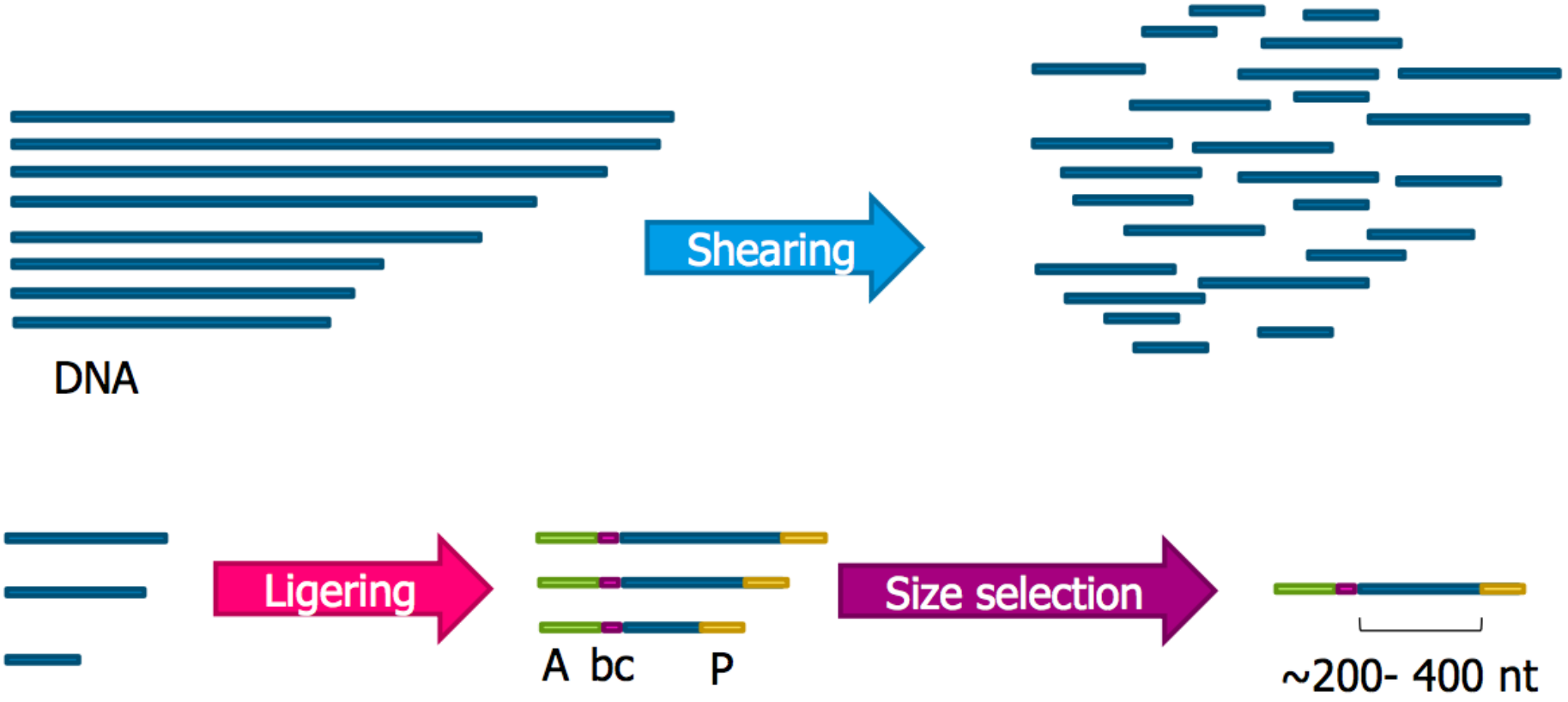
Baserat på tillsats av färgämnen specifika för olika molekyler: DNA, RNA eller protein
Färgämnen har låg fluorescens innan de binder till target (DNA, RNA, protein)
När de binder: intens fluorescens som kan detekteras av fluorometerteknologin

Library preparation



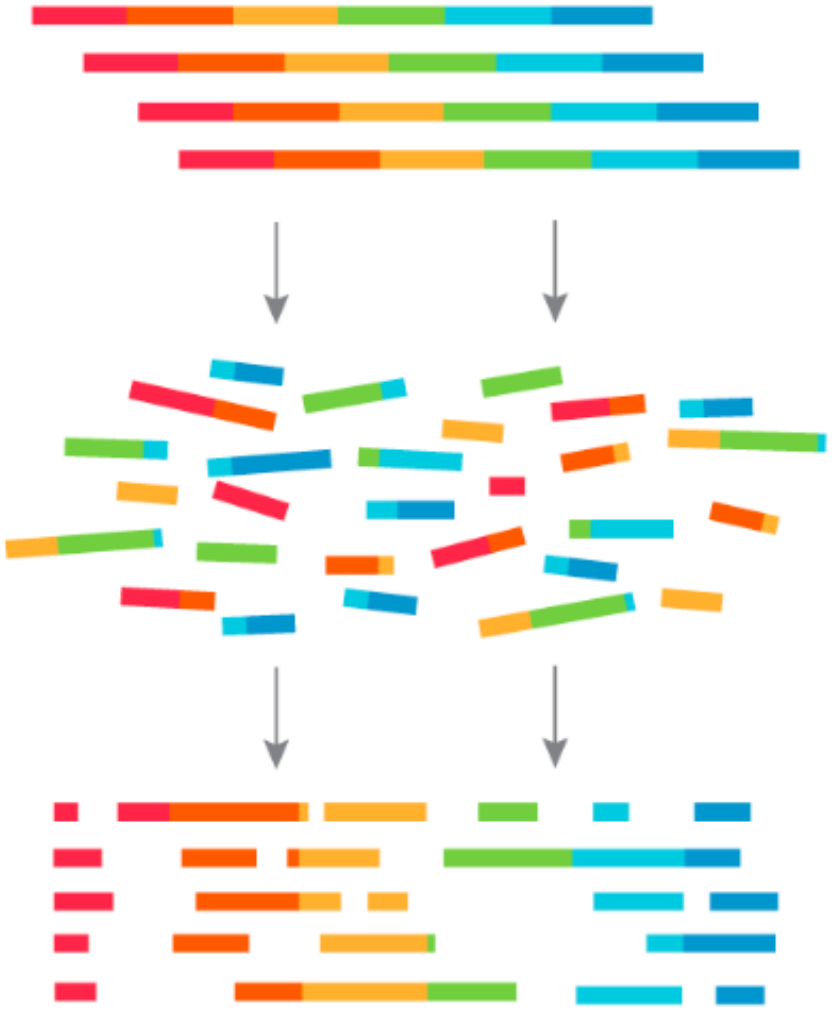
1) Fragmentering: oftast enzymatisk (ofta kit-baserat)
2) I vissa fall: RNA till cDNA eller amplifiering av target
3) Ligering till sekvenseringsadapters – innehåller specifika sekvenser som ska interagera med NGS-plattformen (flow cell – Illumina, beads – Ion Torrent)

Library preparation (forts.)



A= sekvenseringadapter forward
Bc=barcode (sekvens specifik för provet
P=sekvenseringsadapter

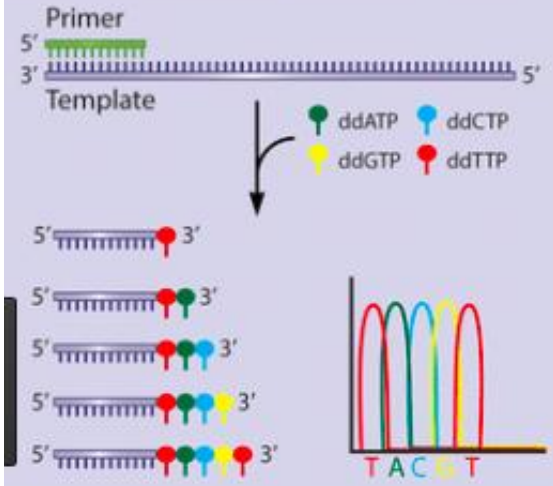
Vad är det vi försöker göra?



ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAA

Tre generationer av sekvensering

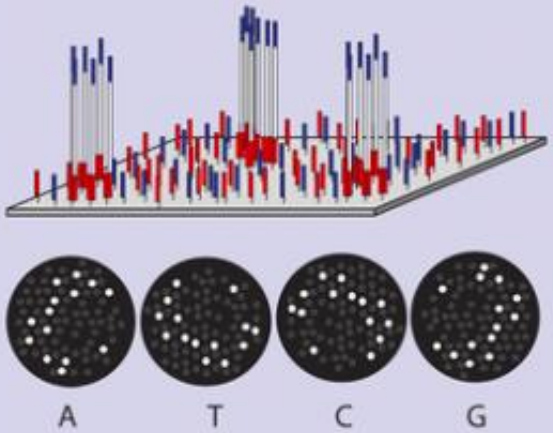
First Generation
Shotgun Sequencing



- Sequencing by synthesis
- High accuracy
- Long read lengths
- Relatively small amount of data generated

e.g., ABI capillary sequencer (ABI)

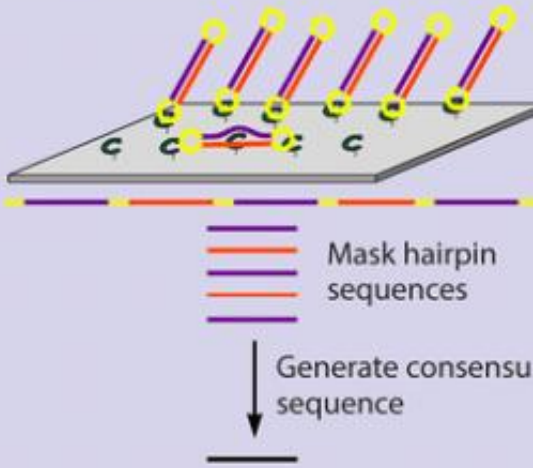
Second Generation
Massively Parallel Sequencing



- Sequencing by synthesis
- Amplified templates are generated during sequencing, reducing the requirements for starting material
- High accuracy
- Short read lengths

e.g., MiSeq (Illumina), Ion Torrent (Thermo Fisher Scientific)

Third Generation
Single-molecule Sequencing



- Single-molecule templates
- Low accuracy
- Long read lengths

e.g., Single-Molecule Real-Time (SMRT) – Sequencing (Pacific Biosciences), MinION (Oxford Nanopore Technologies)

Vilken sekvenseringsplattform???

- Big black box – cannot be bothered…
- Jag vill att det ska gå snabbt (Ion Torrent)
- Jag vill att det ska bli ekonomiskt (Illumina, HiSeq)
- Jag vill ha långa fragment (PacBio eller Nanopore)
- Benchtop (Ion Torrent, MiSeq) vs monster (HiSeq) vs chip (Nanopore)
- Ibland begränsar sig valet av vad du har tillgång till
- Priset kan vara väldigt variabelt (500-flera 1000)

Varför är read length viktigt?

1

Short reads: ~150 nucleotides long (from 2nd gen technology)

Long reads: 100s-1000s nucleotides long (from 3rd gen technology)

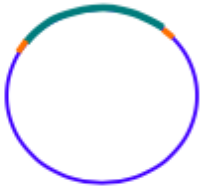
2

Ambiguity in sequence assembly

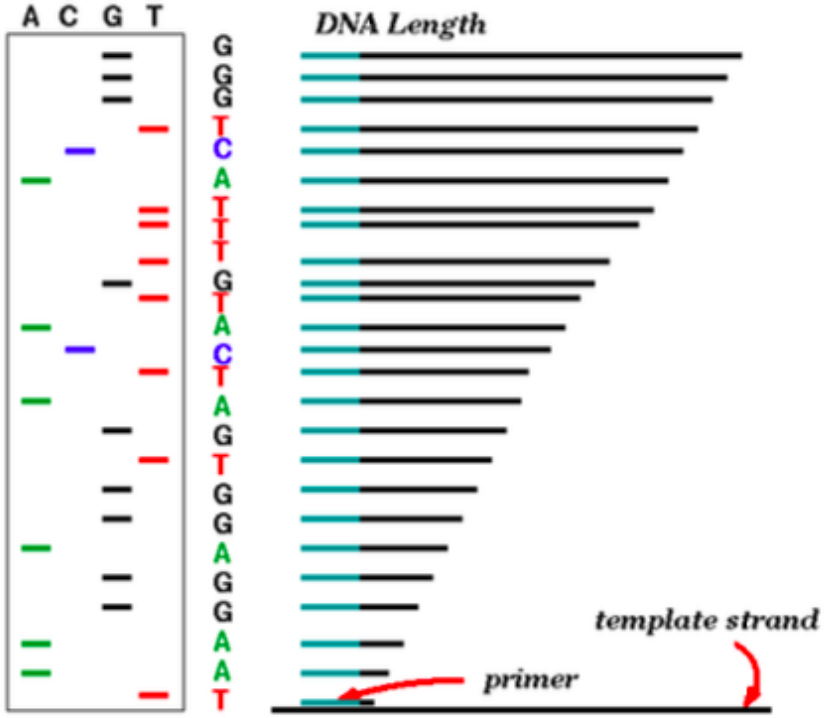
3

Hybrid assembly helps resolve ambiguities with higher coverage and differing read lengths

Sanger sekvensering (first generation)

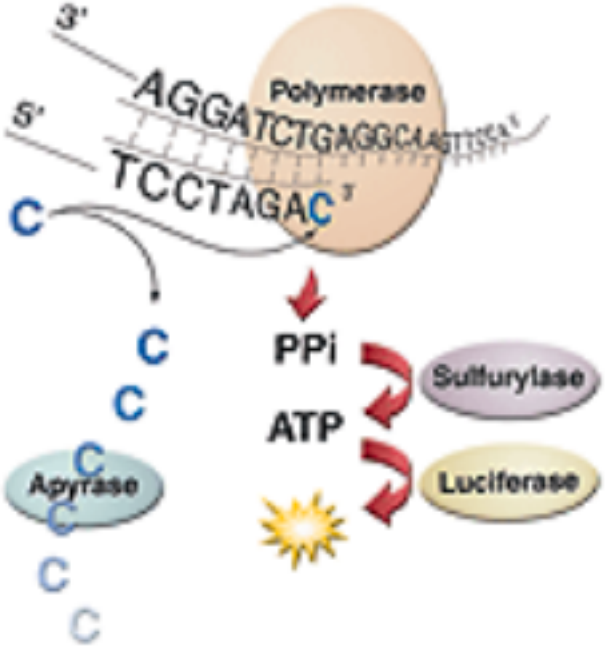


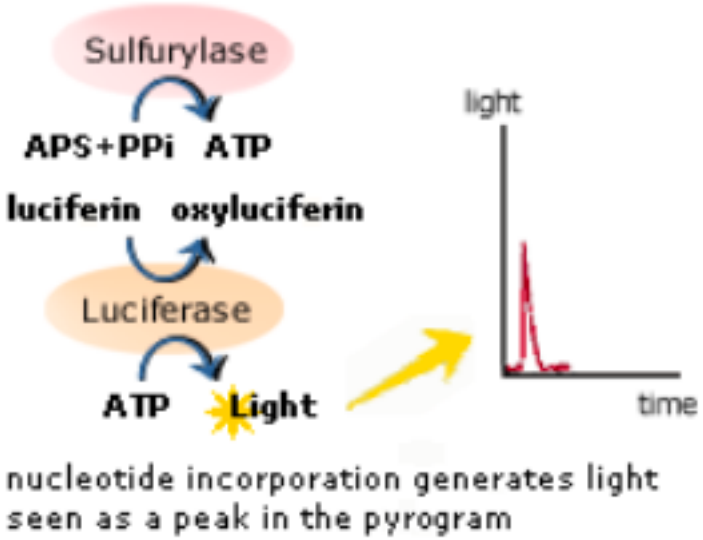
1. Start at primer (restriction site)
2. Grow DNA chain
3. Include ddNTPs
4. Stops reaction at all possible points
5. Separate products by length, using gel electrophoresis



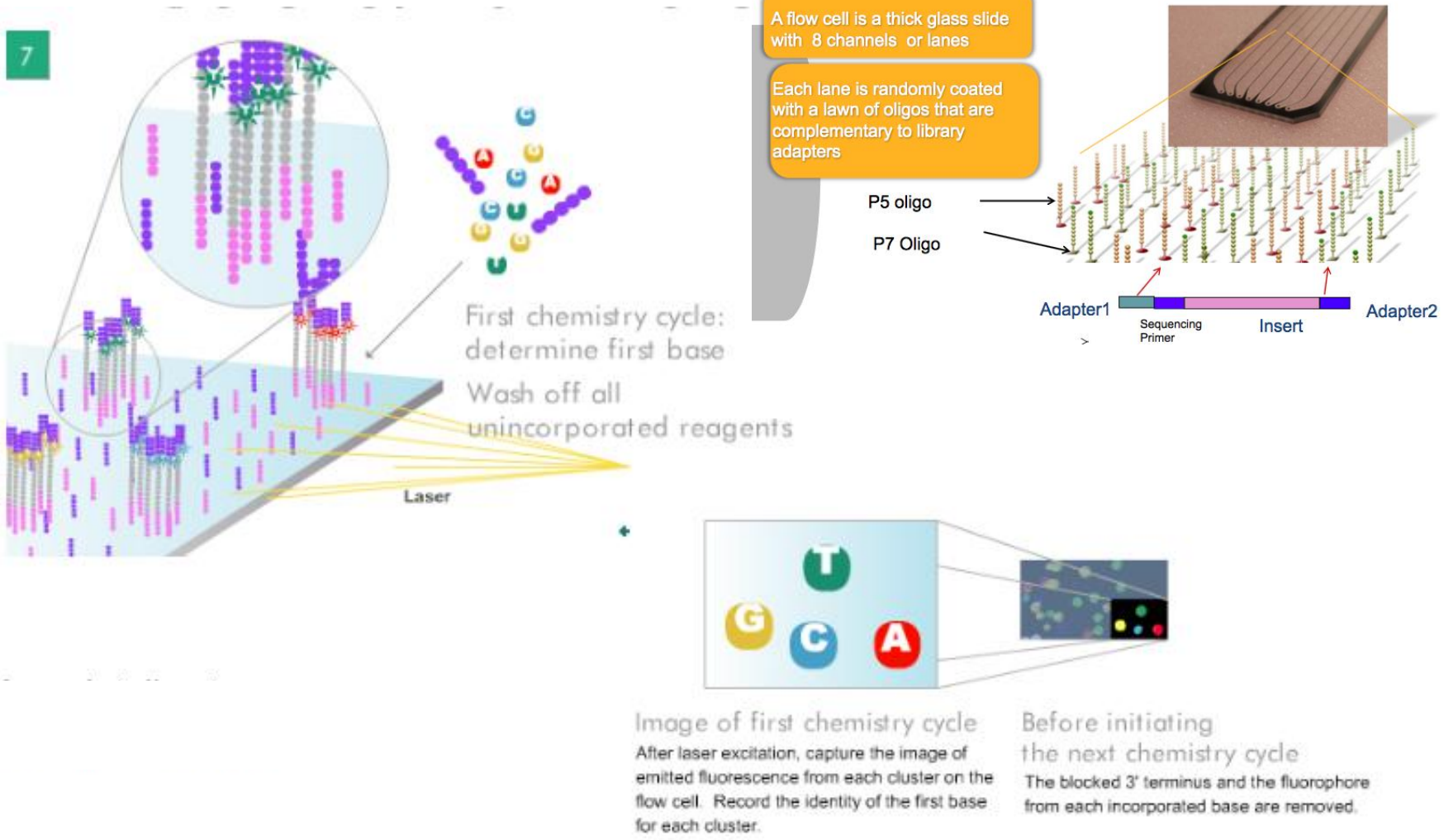
Chain terminators:
dideoxynucleotides triphosphates (ddNTPs)

454: pyrosekvensering: emission av ljus för varje bas som inkorporeras (2nd gen)

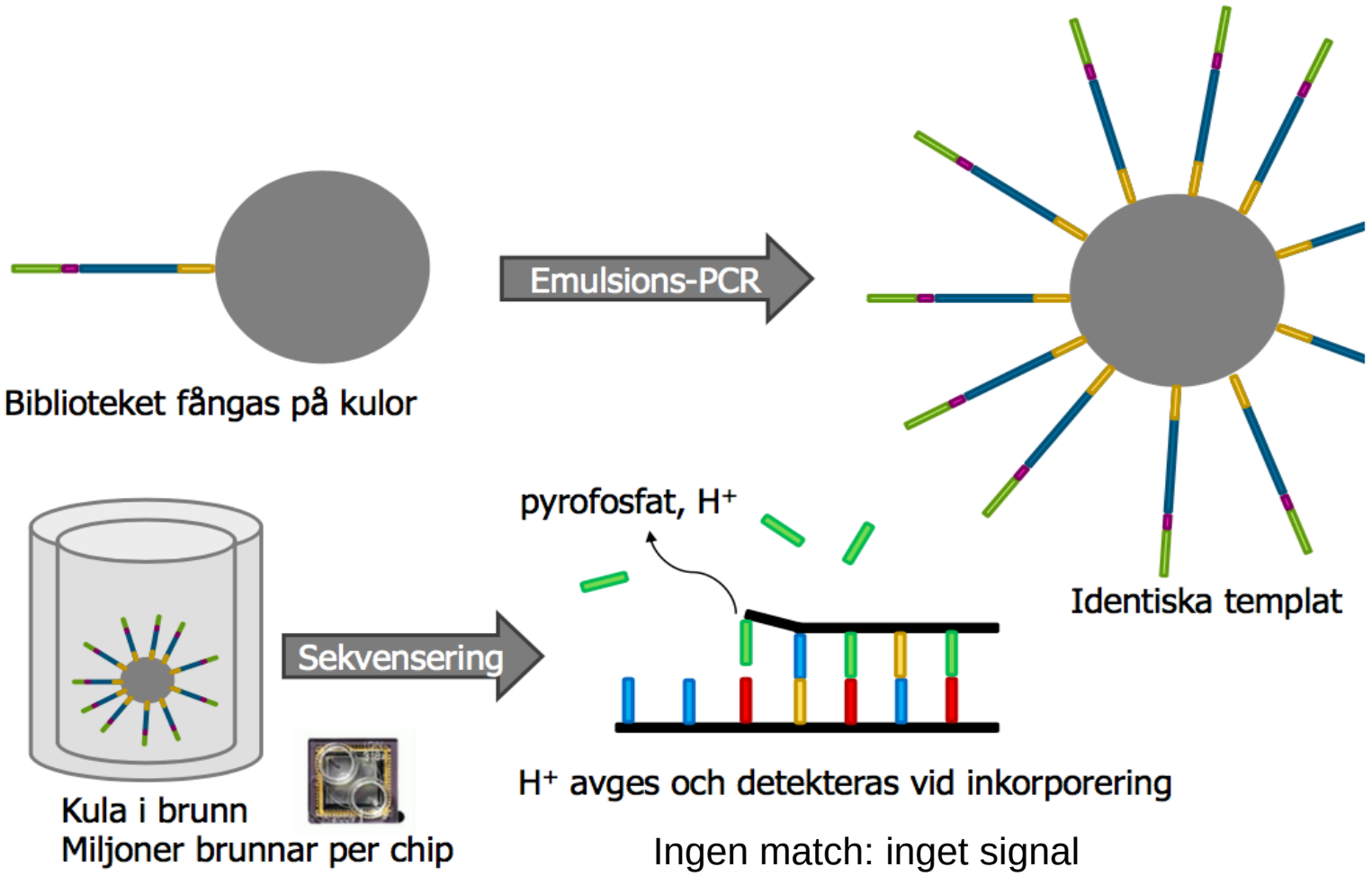




Illumina sekvensering (2nd): flow cell



Ion Torrent (2nd): semikonduktionschip detekterar H^+ - kemisk till digital



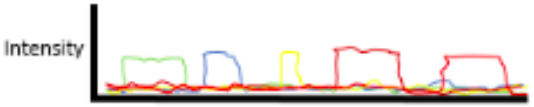
Long read (3rd gen) sekvensering

Karolinska Institutet

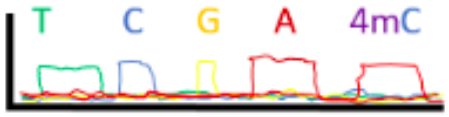
PacBio SMRT seq

DNA passes thru polymerase in an illuminated volume

Raw output is fluorescent signal of the nucleotide incorporation, specific to each nucleotide

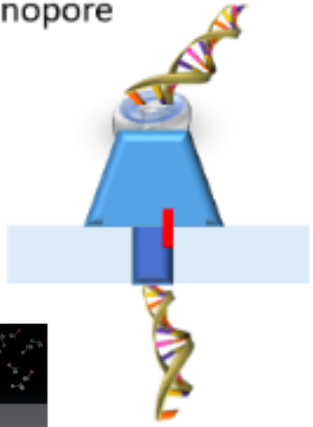


A,C,T,G have known pulse durations, which are used to infer methylated nucleotides



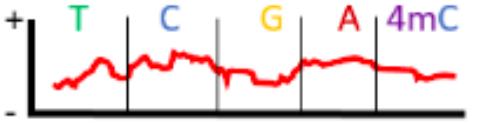
Oxford Nanopore

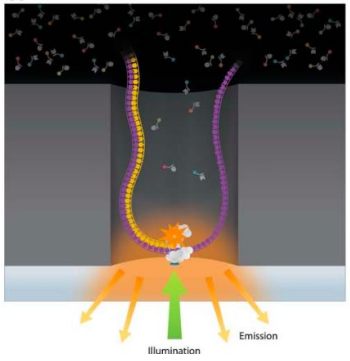
DNA passes thru nanopore

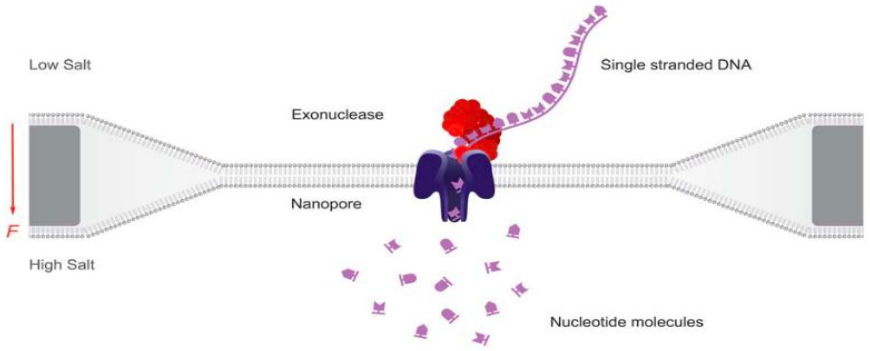


Raw output is electrical signal caused by nucleotide blocking ion flow in nanopore

Each nucleotide has a specific electric "signature"







Dags för dataanalys!