Introduction to NGS Analysis

Anna Battenhouse

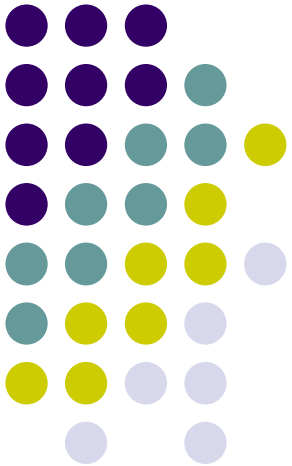
abattenhouse@utexas.edu

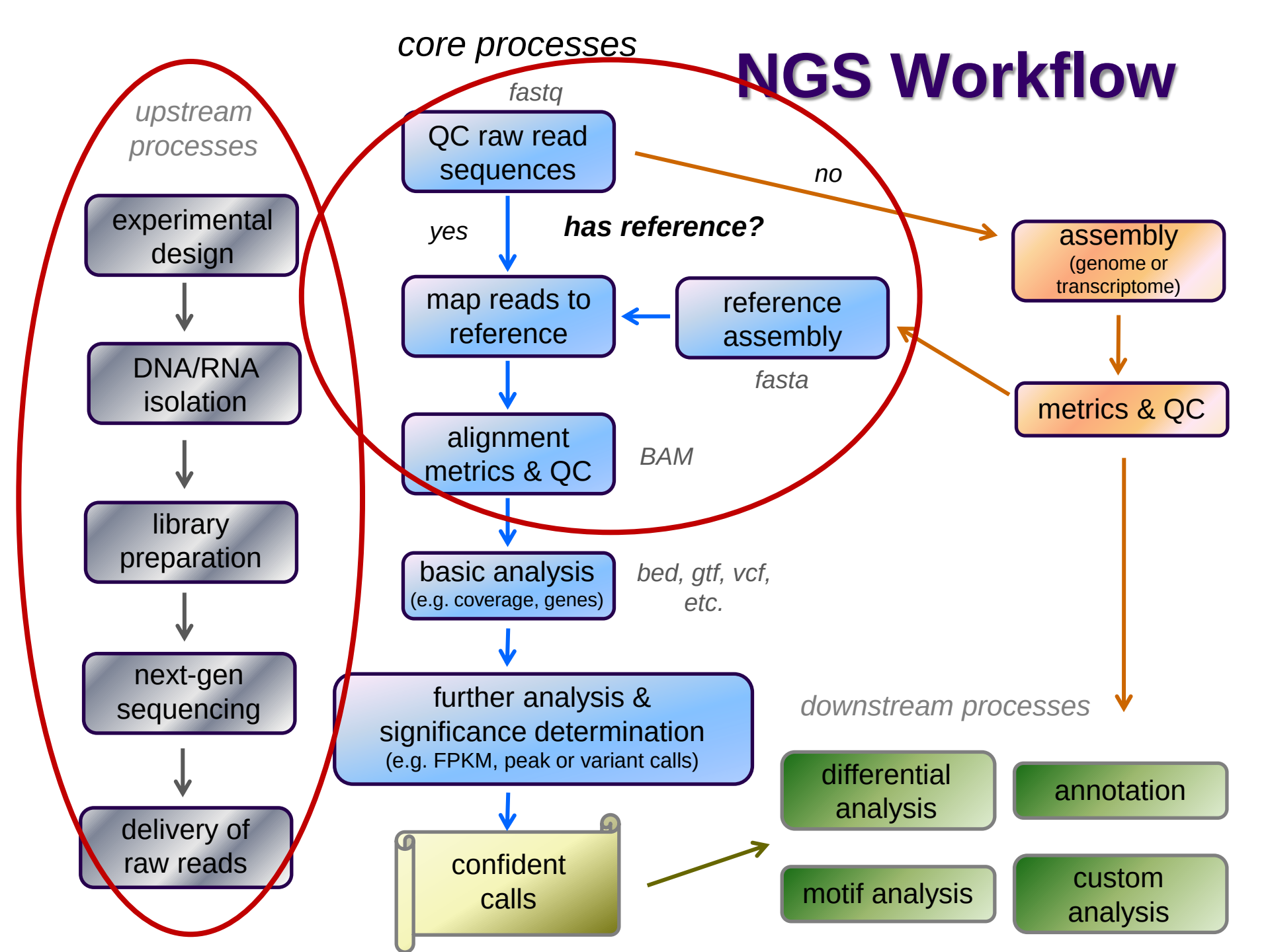
February 27, 2025

Associate Research Scientist

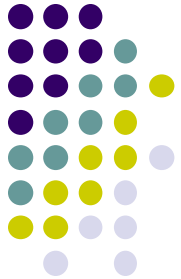
Center for Biomedical Research Support (CBRS)
member, Bioinformatics Consulting Group (BCG)
manager, Biomedical Research Computing Facility (BRCF)
Genome Sequencing & Analysis Facility (GSAF)

Center for Systems and Synthetic Biology (CSSB)
staff, Ed Marcotte & Vishwanath Iyer labs




core processes
NGS Workflow
upstream processes
fastq
QC raw read sequences
no
yes
has reference?
experimental design
map reads to reference
reference assembly
assembly
(genome or transcriptome)
fasta
DNA/RNA isolation
metrics & QC
library preparation
alignment metrics & QC
BAM
basic analysis
(e.g. coverage, genes)
bed, gtf, vcf, etc.
next-gen sequencing
further analysis & significance determination
(e.g. FPKM, peak or variant calls)
downstream processes
differential analysis
annotation
delivery of raw reads
confident calls
motif analysis
custom analysis

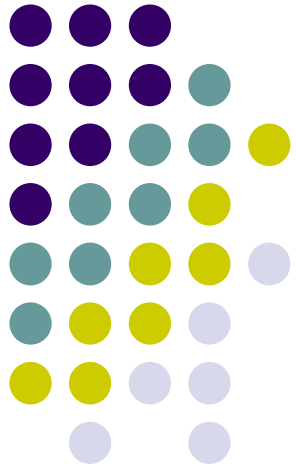
Outline

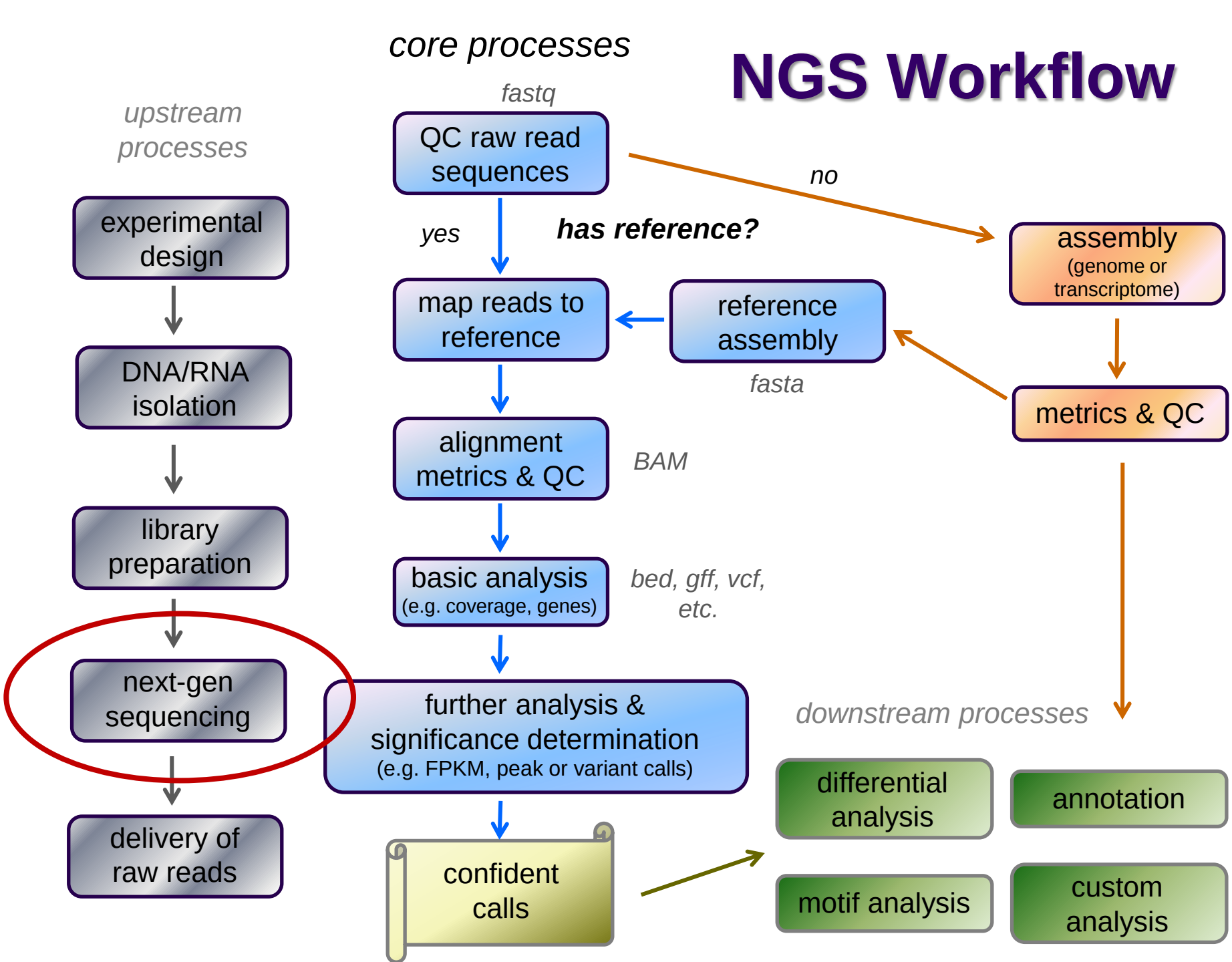


1. Overview of sequencing technologies
2. NGS concepts and terminology
3. The FASTQ format and
   raw data QC & preparation
4. Alignment to a reference

Part 1: Overview of Sequencing Technologies

- High-throughput (“next gen”) sequencing
- Illumina short-read sequencing
- Long read sequencing




core processes
NGS Workflow
upstream processes
fastq
QC raw read sequences
no
experimental design
yes
has reference?
assembly
(genome or transcriptome)
map reads to reference
reference assembly
fasta
DNA/RNA isolation
metrics & QC
alignment metrics & QC
BAM
library preparation
basic analysis
(e.g. coverage, genes)
bed, gff, vcf, etc.
next-gen sequencing
further analysis & significance determination
(e.g. FPKM, peak or variant calls)
downstream processes
differential analysis
annotation
delivery of raw reads
confident calls
motif analysis
custom analysis

“Next Generation” sequencing

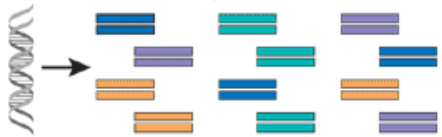
- Massively parallel
  - simultaneously sequence “library” of ***millions*** of different DNA fragments
- ***PCR colony clusters*** generated
  - individual template DNA fragments titrated onto a flowcell to achieve inter-fragment separation
  - PCR “bridge amplification” creates ***clusters*** of identical molecules
- ***Sequencing by synthesis***
  - fluorescently-labeled dNTs added
  - incorporation generates signal
  - flowcell image captured after each cycle
  - images computationally converted to base calls (including a quality score)
  - results in 30 – 300 base “reads”
    - much shorter than Sanger sequencing

Sanger

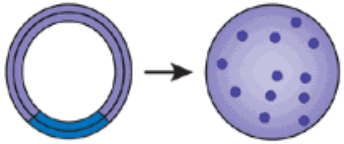
→ *Single type of molecule*

a

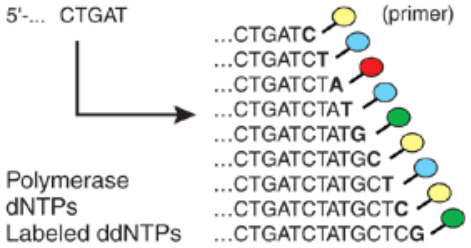
DNA fragmentation



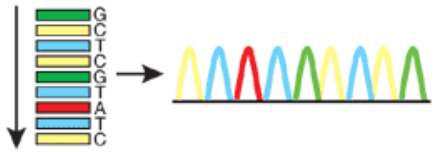
In vivo cloning and amplification



Cycle sequencing



Electrophorsesis
(1 read/capillary)



NGS

→ *Many different molecules*

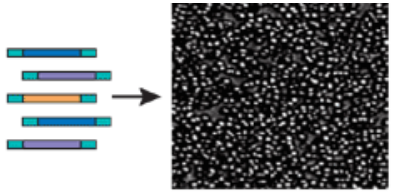
b

DNA fragmentation

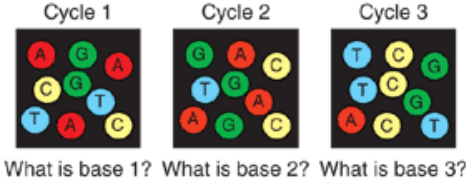
In vitro adaptor ligation
(sequencing adapter)

polony = PCR colony cluster

Generation of polony array



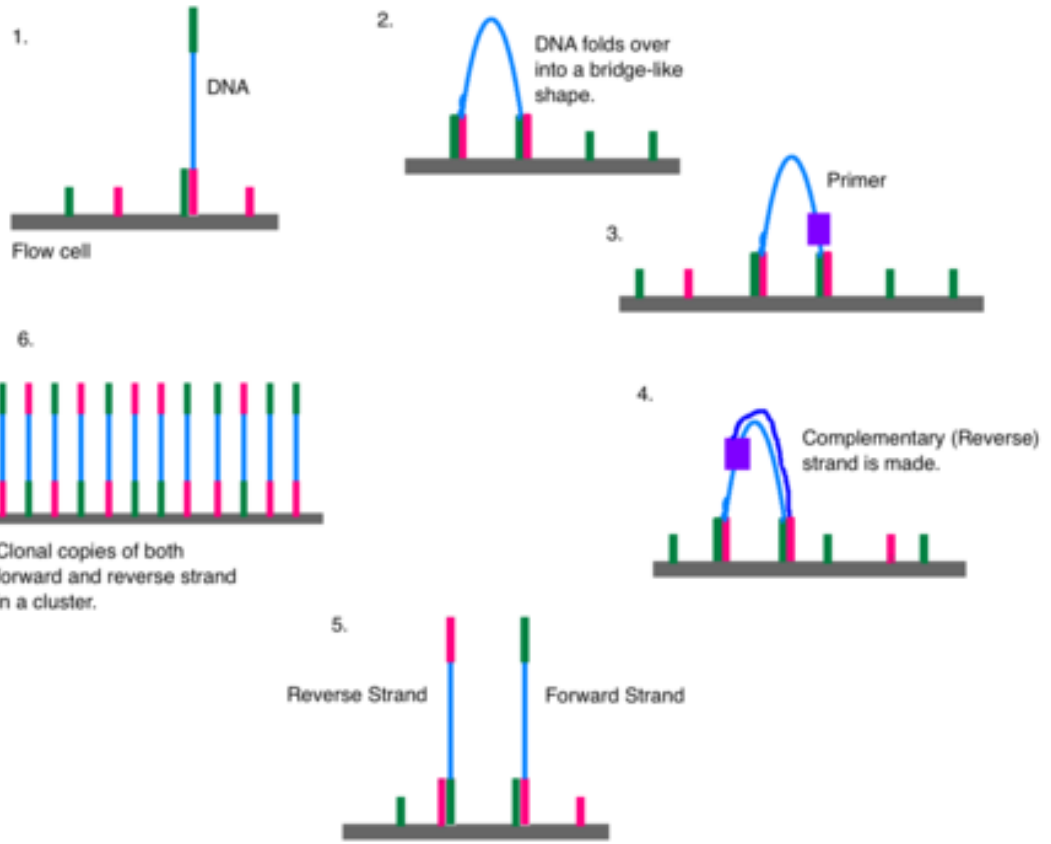
Cyclic array sequencing
(>10^6 reads/array)



Shendure et al, Nature Biotechnology. 2008.
https://www.nature.com/articles/nbt1486

Illumina sequencing

1. Library preparation
2. ***Cluster generation via bridge amplification***
3. Sequencing by synthesis
4. Image capture
5. Convert to base calls



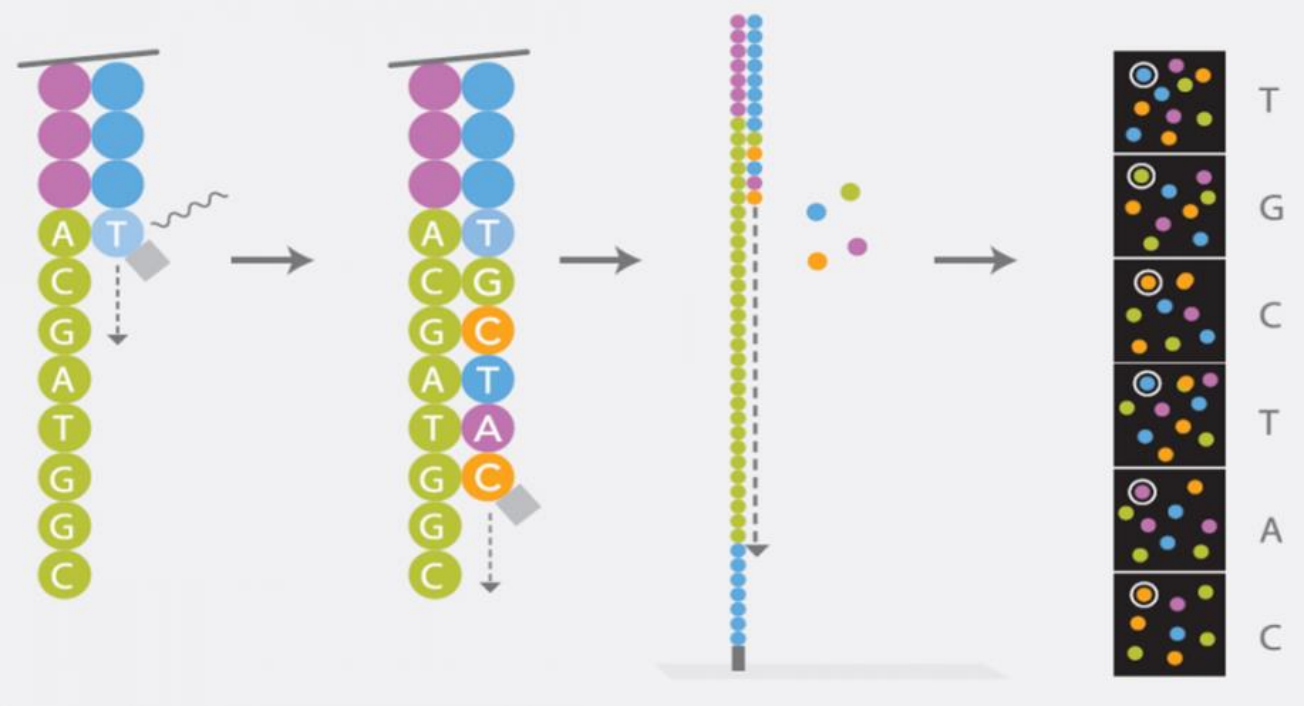
Short Illumina video
(https://tinyurl.com/hvnmwjb)

- Note
  - 2 PCR amplifications performed
    1. during ***library preparation***
    2. during ***cluster generation***
  - ***amplification always introduces bias!***

https://en.wikipedia.org/wiki/Illumina_dye_sequencing

Illumina sequencing

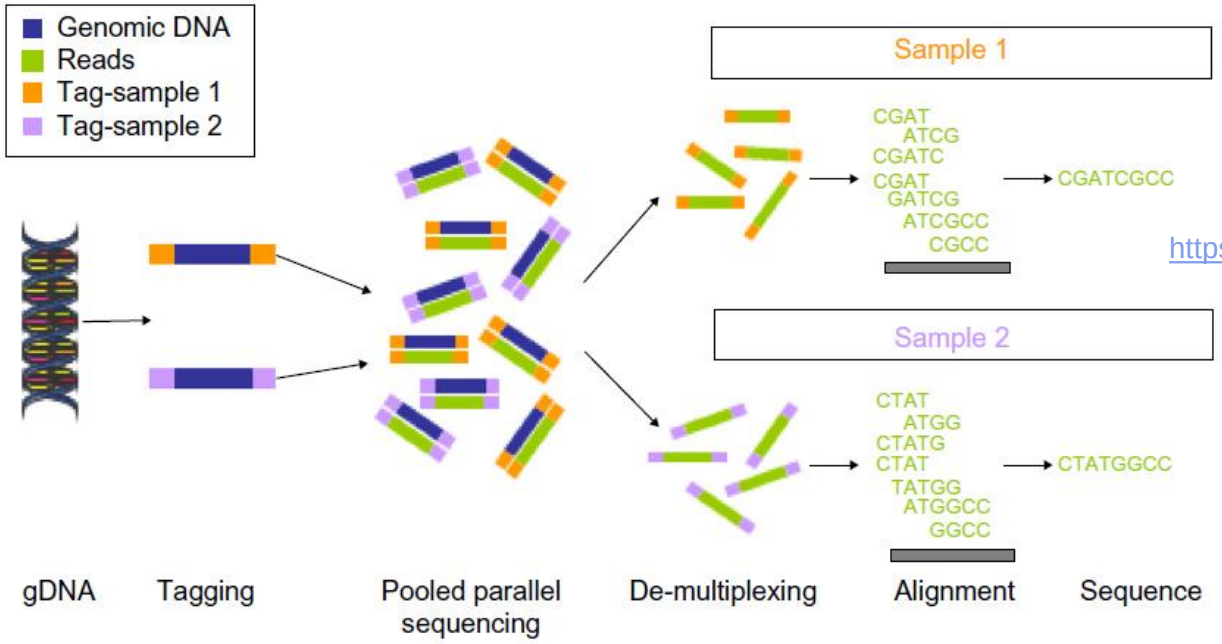
1. Library preparation
2. Cluster generation via bridge amplification
3. ***Sequencing by synthesis***
4. ***Image capture***
5. ***Convert to base calls***



http://www.cegat.de/

Multiplexing

- Illumina sequencers have one or more flowcell “lanes”, each of which can generate ***millions*** of reads
  - ~20**M** reads/lane for MiSeq, ~10**G** reads/lane for NovaSeq
- When less than a full flowcell lane is needed, multiple samples with different ***barcodes*** (a.k.a. ***indexes***) can be run on the same lane
  - 6-8 bp ***library barcode*** attached to DNA library fragments
  - data from sequencer must be ***demultiplexed*** to determine which reads belong to which library



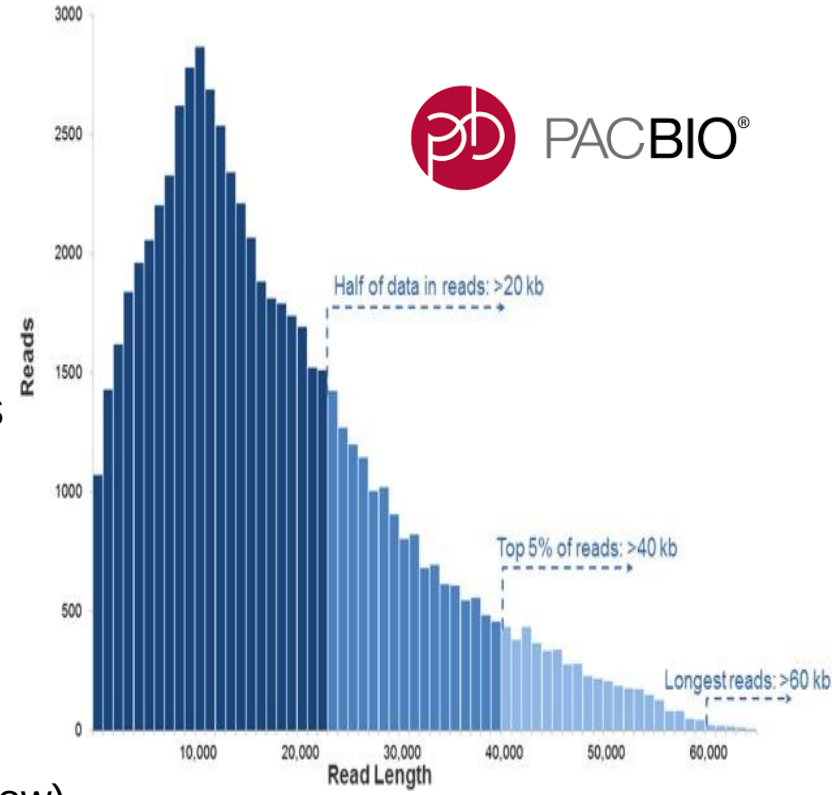
https://doi.org/10.2147/BLCTT.S51503

Long read sequencing

- Short read technology limitations
  - 30 – 300 base reads (150 typical)
  - PCR amplification bias
  - short reads are difficult to assemble
    - e.g., too short to span a long repeat region
- Newer "***single molecule***" sequencing
  - sequences ***single molecules***, not clusters
  - allows for ***much*** longer reads (multi-Kb!)
    - no signal wash-out due to lack of synchronization among cluster molecules

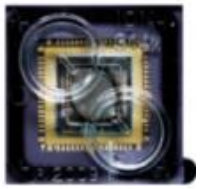
but:

  - weaker signal leads to high error rate
    - 10% *vs* <1% for Illumina (but improving now)
    - and fewer reads may be generated
      - ~ 1 million vs 10s to 100s of millions w/short reads



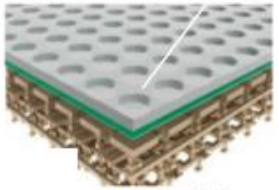
Long read sequencing

- Oxford Nanopore ION technology systems (e.g. MinION)
  - https://nanoporetech.com/
  - DNA “spaghetti’s” through tiny protein pores
  - Addition of different bases produces different pH changes
    - measured as different changes in electrical conductivity
  - MinION is hand-held, starter kit costs ~$1,000 – including reagents!
    - inexpensive, but high error rates (~10%, but improving)



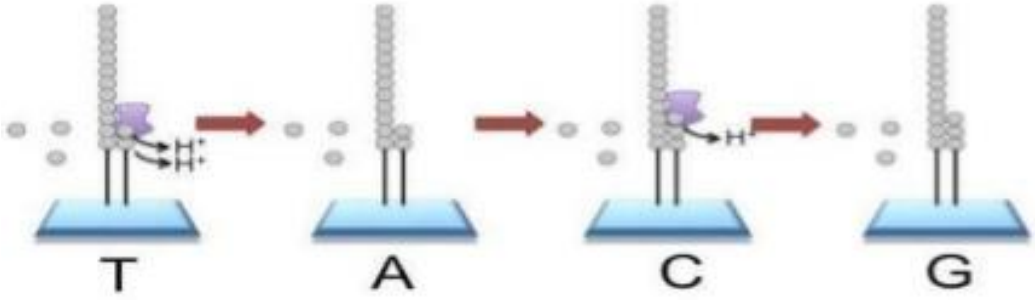
Chip

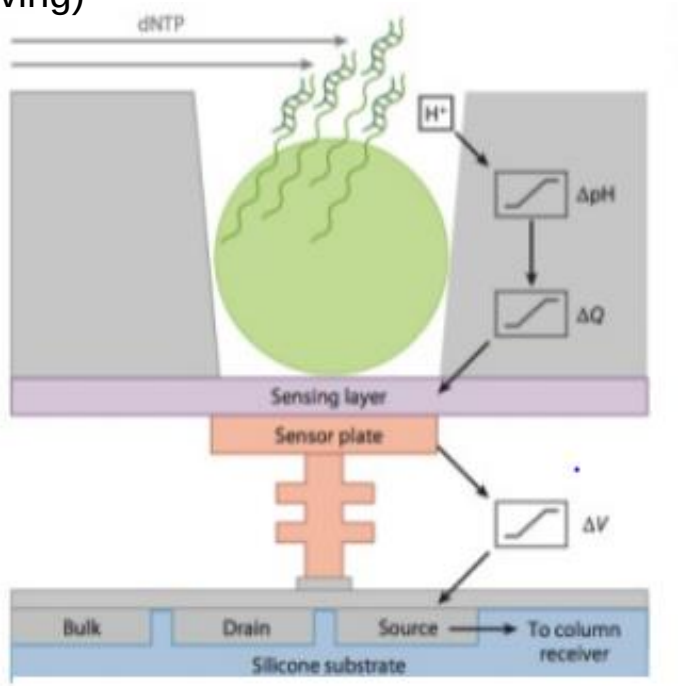
Semicor.ductor Packaging



Millions of Sensors

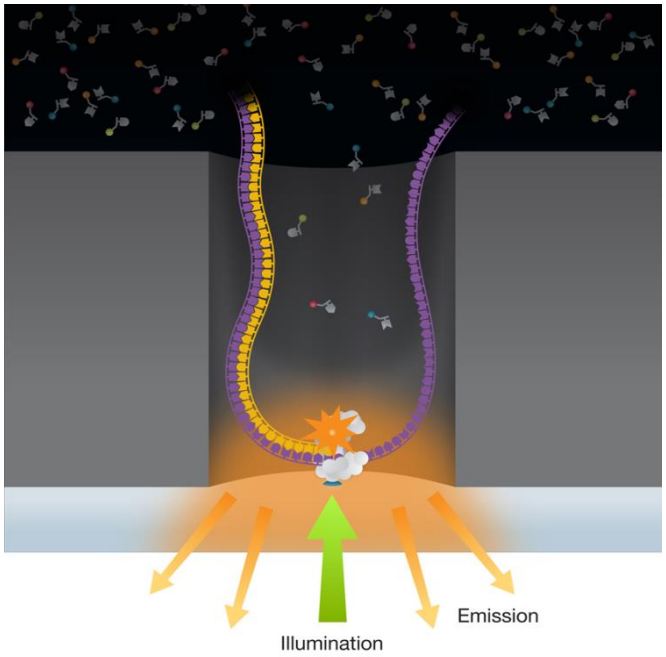
Semiconductor Design

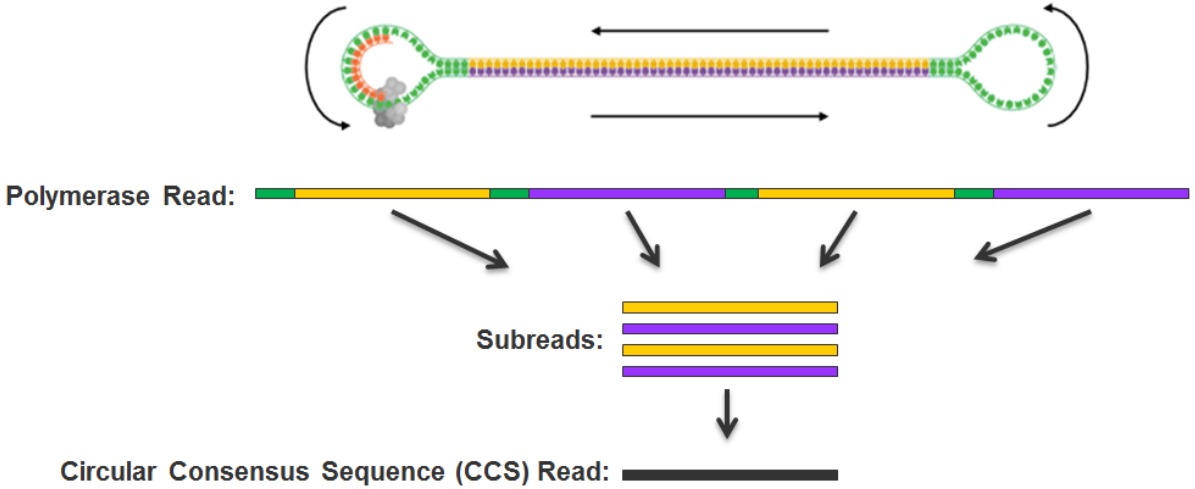




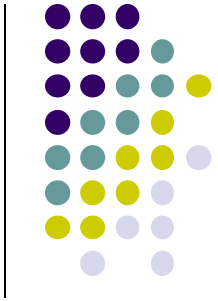
Long read sequencing

- PacBio SMRT system
  - http://www.pacb.com/smrt-science/smrt-sequencing/
  - Sequencing by synthessis in ***Zero-Mode Waveguide*** (ZMW) wells
  - DNA is circularized then repeatedly sequenced to achieve “consensus”
    - reduces error rate (~1-2%), but equipment ***quite*** expensive
  - Now the preferred technology for assembly of large eukaryotic genomes
    - especially polyploid species (e.g. many plants)





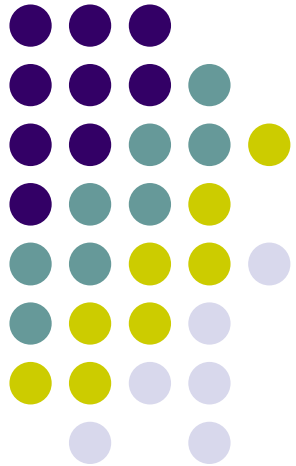
Part 1 summary

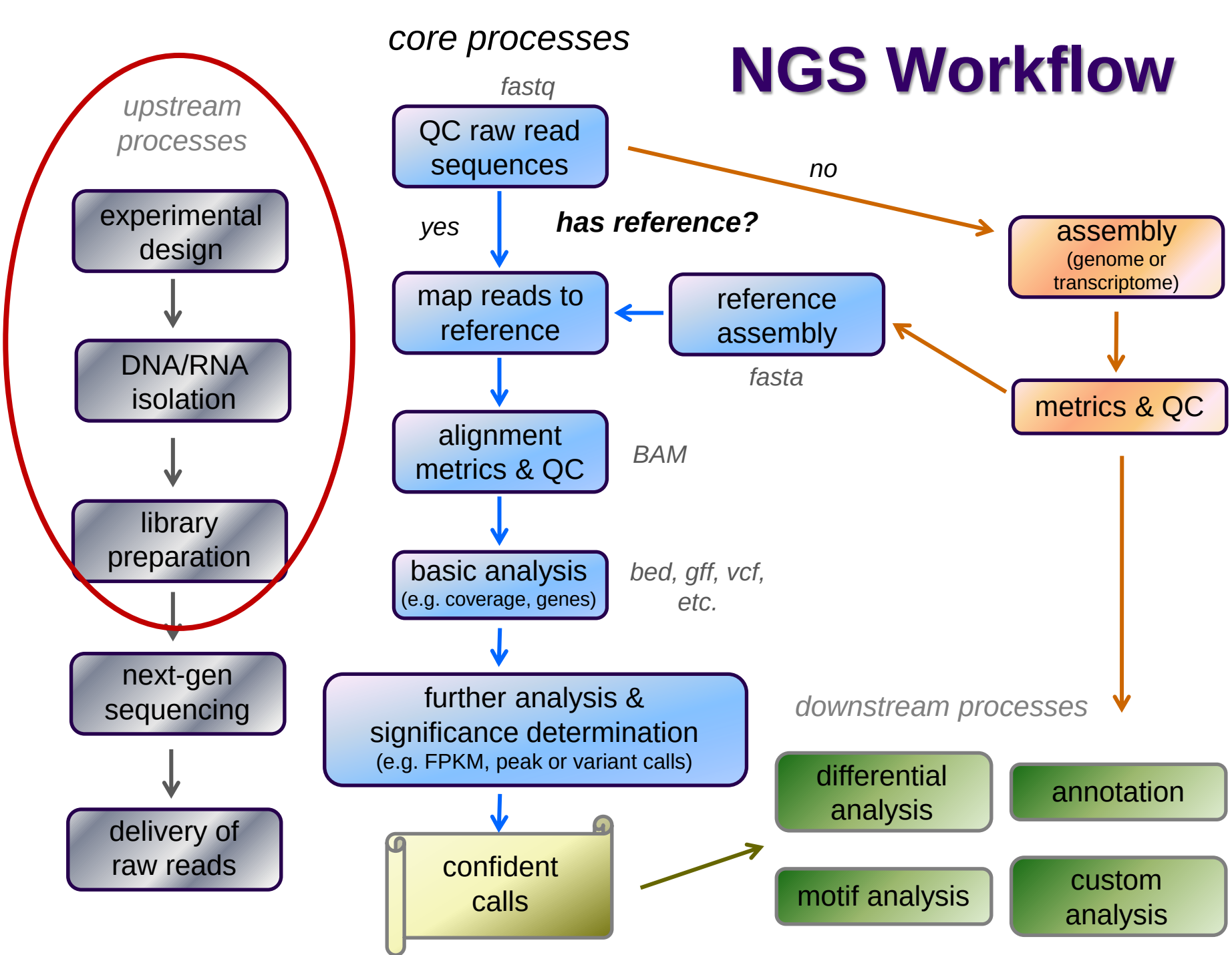


- "Next Generation" sequencing operates on a ***library*** of ***millions of different DNA fragments*** vs a single purified molecular species
- Illumina platforms are dominant for ***short reads*** (30-300 base)
  - Use ***sequencing by synthesis*** on ***clusters*** of identical molecules (clones); de-synchronization of signal limits read length
  - Since instruments are so high throughput, multiple ***barcoded*** sample libraries are pooled for sequencing, then ***demultiplexed***
- ***Single molecule*** technologies produce ***long reads*** (multi-Kilobase)
  - Very low signal from single molecule readout presents accuracy challenges
  - Oxford Nanopore has low-cost models, but with high error rates
  - PacBio has lower error rates but cost of both equipment & sequencing are high

Part 2: NGS Concepts & Terminology

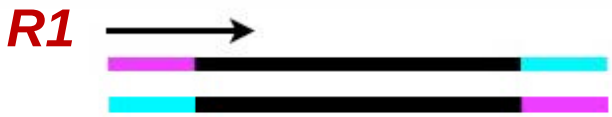
- Sequencing terminology
- Experiment types & library complexity
- Sequence duplication issues




core processes
NGS Workflow
upstream processes
experimental design
DNA/RNA isolation
library preparation
next-gen sequencing
delivery of raw reads
fastq
QC raw read sequences
no
yes
has reference?
assembly
(genome or transcriptome)
map reads to reference
reference assembly
fasta
metrics & QC
alignment metrics & QC
BAM
basic analysis
(e.g. coverage, genes)
bed, gff, vcf, etc.
further analysis & significance determination
(e.g. FPKM, peak or variant calls)
downstream processes
confident calls
differential analysis
annotation
motif analysis
custom analysis

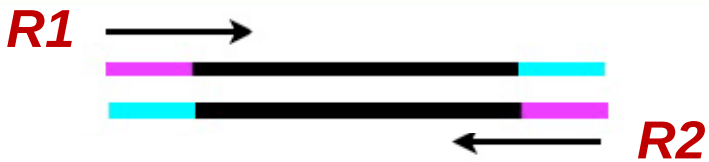
Read types

single-end



independent reads

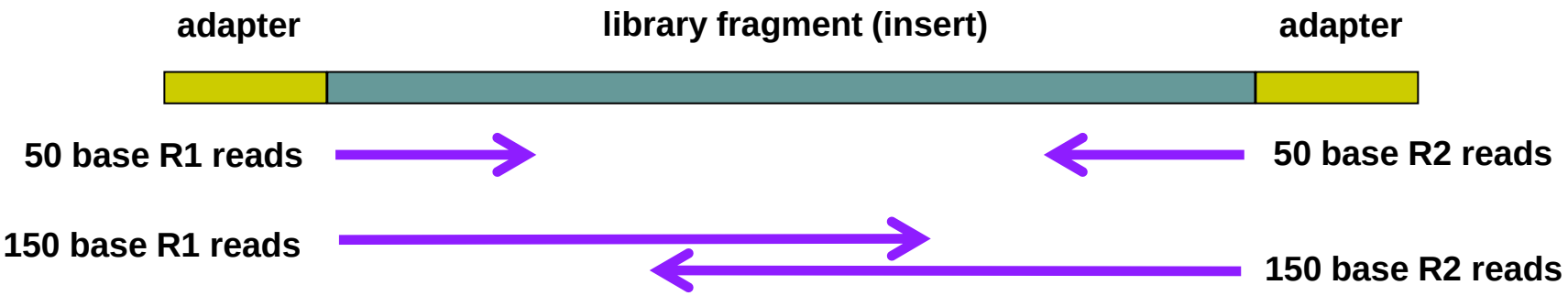
paired-end



two inwardly oriented reads separated by ~200 nt

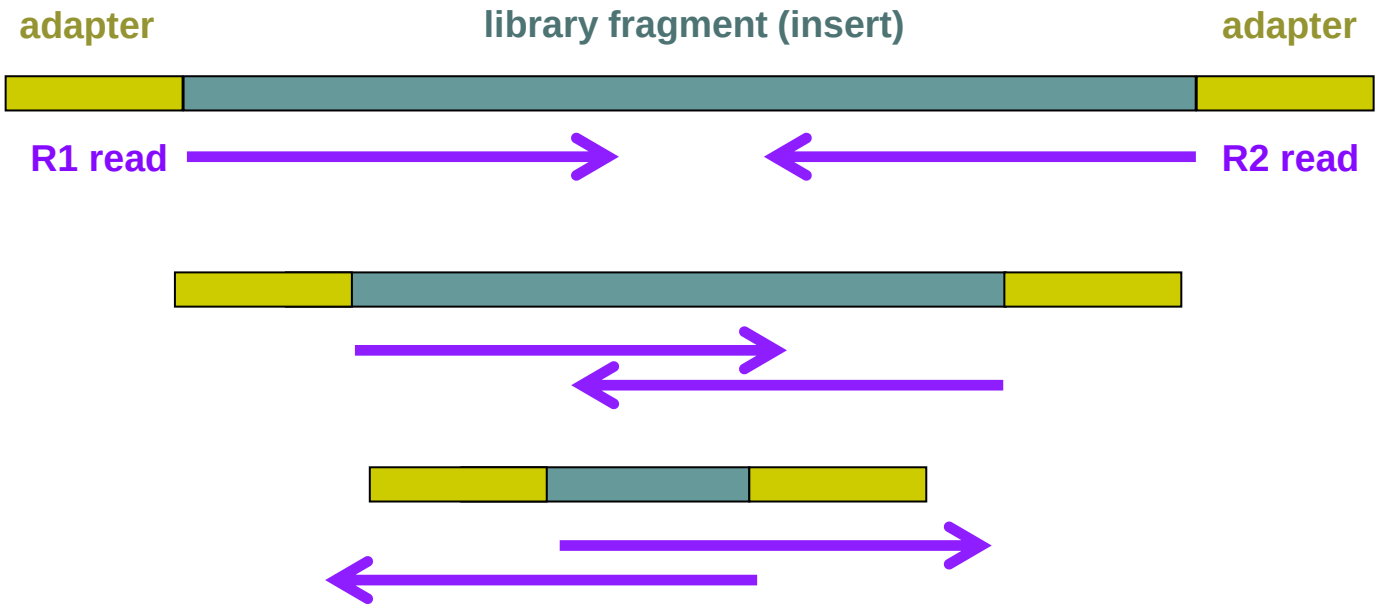
Reads and Fragments

- With ***paired-end*** sequencing, keep in mind the distinction between
  - the library ***fragment*** from your library that was sequenced
    - also called ***inserts***
  - the ***sequence reads*** (R1s & R2s) you receive
- An **R1** and its associated **R2** form a ***read pair***
  - a readout of part (or all) of the fragment molecule



Library fragment distribution

- What is fixed size in your sequencing library:
  - the adapter region (including all barcodes)
  - the read length (e.g. 50, 100, 150)
- But the ***insert fragments are of variable length***
  - due to random shearing during library preparation
  - bioanalyzer provides an *estimate* of the library's fragment distribution



Library Complexity

Library complexity (diversity) is a measure of the number of ***distinct molecular species*** in the library.

Many different molecules → ***high complexity***

Few different molecules → ***low complexity***

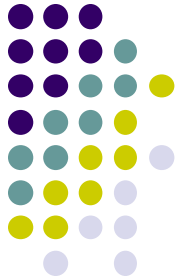
The number of different molecules in a library depends on ***enrichment*** performed during ***library construction***.

Popular Experiment Types

- ***Whole Genome sequencing (WGS)***
  - ***applications***: genome assembly
  - ***library***: all genomic DNA (no enrichment)
  - ***complexity***: ***high*** (fragments must cover the entire genome)
- ***Exome sequencing (WXS)***
  - ***applications***: polymorphism/SNP detection; genotyping
  - ***library***: DNA from eukaryotic exons (uses special kits)
  - ***complexity***: ***high/med*** (only ~5% of eukaryotic genome is in exons)
- ***RNA-seq***
  - ***applications***: differential gene expression between 2 or more conditions
  - ***library***: extracted RNA converted to cDNA
  - ***complexity***: ***med/high*** (only a subset of genes are expressed in any given tissue)
- ***Amplicon panels (targeted sequencing)***
  - ***applications***: genetic screening panels; metagenomics; mutagenesis
  - ***library***: DNA from a set of PCR-amplified regions using custom primers
  - ***complexity***: ***very low*** (only a few hundred-to-thousand different library molecules)

Type	Library construction	Applications	Complexity
Whole genome (WGS)	• extract genomic DNA & fragment	• Genome assembly • Variant detection, genotyping	high
Bisulfite sequencing	• bisulfite treatment converts C → U but not 5meC	• Methylation profiling (CpG)	high
RAD-seq, ddRAD	• restriction-enzyme digest DNA & fragment	• Variant detection (SNPs) • Population genetics, QTL mapping	high
Exome (WXS)	• capture DNA from exons only (manufacturer kits)	• Variant detection, genotyping	high-medium
ATAC-seq	• high-activity transposase cuts DNA & ligates adapters	• Profile nucleosome-free regions (“open chromatin”)	medium-high
RNA-seq, Tag-seq	• extract RNA & fragment • convert to cDNA	• Differential gene or isoform expression • Transcriptome assembly (RNA-seq only)	medium, medium-low for Tag-seq
Transposon seq (Tn-seq)	• create library of transposon-mutated genomic DNA • amplify mutants via Tn-PCR	• Characterize genotype/phenotype relationships w/high sensitivity	medium
ChIP-seq	• cross-link proteins to DNA • pull-down proteins of interest w/ specific antibody, reverse cross-links	• Genome-wide binding profiles of transcription factors, epigenetic marks & other proteins	medium (but variable)
GRO-seq	• isolate actively-transcribed RNA	• Characterize transcriptional dynamics	medium-low
RIP-seq	• like ChIP-seq, but with RNA	• Characterize protein-bound RNAs	low-medium
miRNA-seq	• isolate 15-25bp RNA band	• miRNA profiling	low
Amplicons	• amplify 1-1000+ genes/regions	• genotyping, metagenomics, mutagenesis	low

Library complexity is primarily a function of experiment type



higher complexity

Less enrichment for specific sequences

genomic

bisulfite-seq

exon capture

RNA-seq

ChIP-seq

amplicons

More enrichment for specific sequences

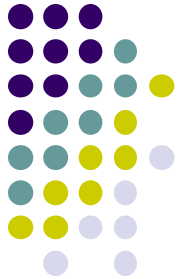
lower complexity

Higher diversity of library molecules
Lower sequence duplication expected
More sequencing depth required

… as well as…
- *genome size & sequencing depth*
- *library construction skill & luck!*

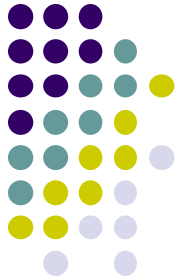
Lower diversity of library molecules
Higher sequence duplication expected
Less sequencing depth required

Sequence Duplication



- The set of sequences you receive can contain ***exact duplicates***
- Duplication can arise from:
  1. sequencing of species ***enriched*** in your library (***biological – good!***)
     - each read comes from a different DNA molecule
  2. sequencing of ***artifacts*** (***technical – bad!***)
     - differentially amplified PCR species (***PCR duplicates***)
       - recall that 2 PCR amplifications are performed w/Illumina sequencing
     - ***cannot tell which using standard sequencing methods!***
- Different experiment types have different *expected* duplication

Expected sequence duplication is primarily a function of experiment type



higher complexity

Less enrichment for specific sequences

genomic

bisulfite-seq

exon capture

RNA-seq

ChIP-seq

amplicons

More enrichment for specific sequences

lower complexity

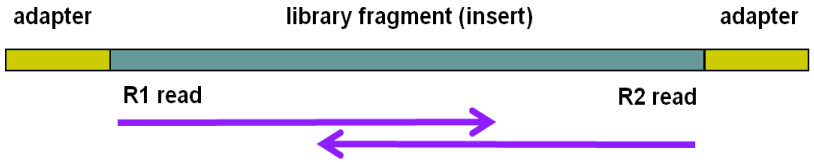
Higher diversity of library molecules
Lower sequence duplication expected
More sequencing depth required

… as well as…
- *genome size & sequencing depth*
- *library construction skill & luck!*

Lower diversity of library molecules
Higher sequence duplication expected
Less sequencing depth required

Single end *vs* Paired end

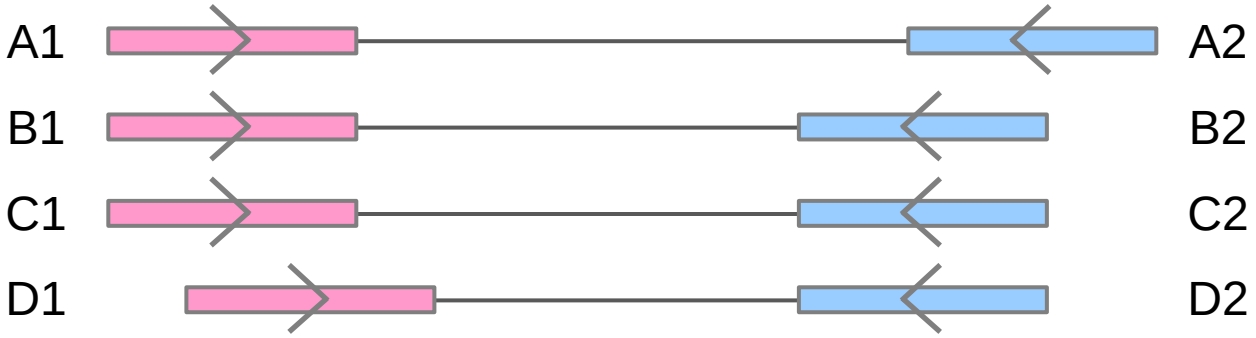
- ***Single End*** (***SE***) reads are less expensive
- ***Paired End*** (***PE***) reads:
  - provide more bases around a locus
    - e.g. for analysis of polymorphisms
  - actual fragment sizes can be easily determined
  - helps distinguish the true complexity of a library
    - by clarifying which ***fragments*** are duplicates (vs ***sequence*** duplicates)
  - ***but*** PE reads are more expensive – and larger
    - more storage space and processing time required



- General guidelines
  - use **PE** for high location accuracy and/or base-level sensitivity
  - use **SE** for lower-complexity experiment types

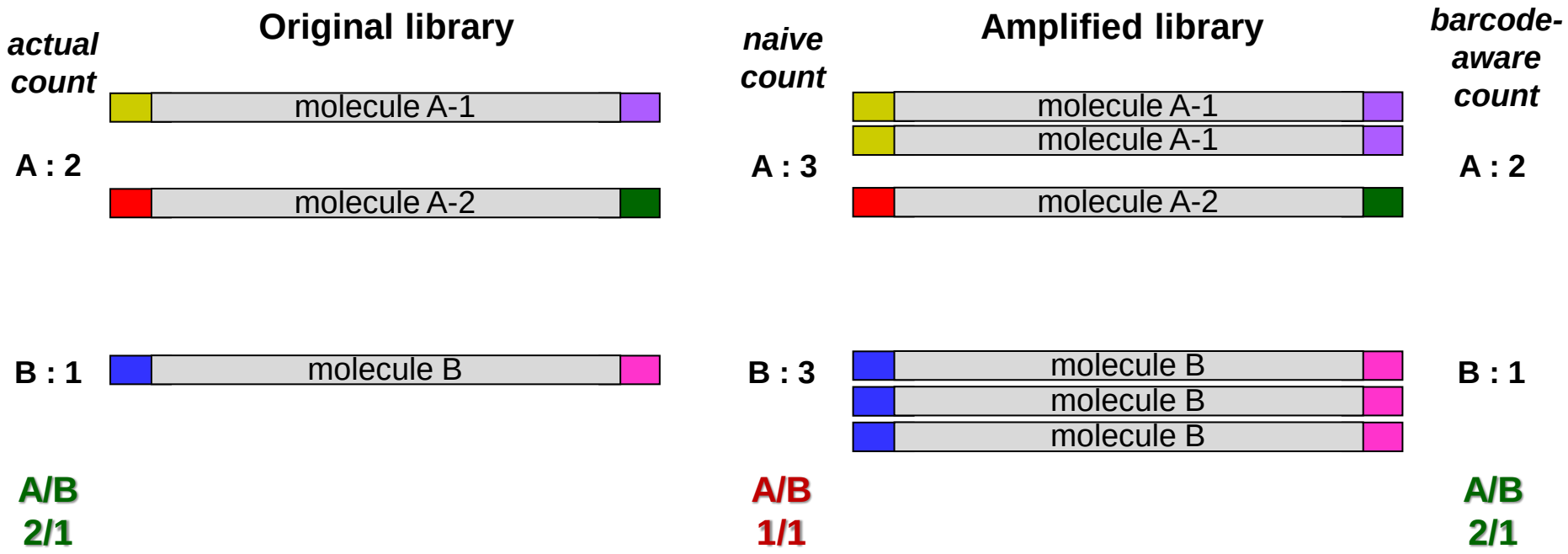
Read vs Fragment duplication

- Consider the 4 “aligned” fragments below
  - 4 R1 reads (pink), 4 R2 reads (blue)
- Duplication when only 1 end considered
  - A1, B1, C1 have identical start locations, D1 different
    - 2 unique + 2 duplicates = 50% duplication rate
  - B2, C2, D2 have identical start locations, A2 different
    - 2 unique + 2 duplicates = 50% duplication rate
- Duplication when both ends considered
  - only fragments B and C are duplicates (same external locations)
    - 3 unique + 1 duplicate = 25% duplication rate



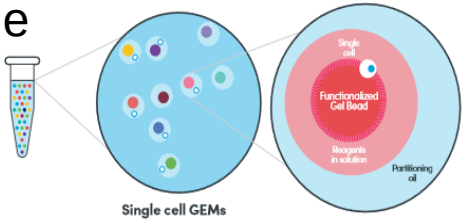
Molecular Barcoding

- Resolves ambiguity between biological and technical (PCR amplification) duplicates
  - adds secondary, ***internal barcodes*** to ***pre-PCR*** molecules
    - a.k.a ***UMI****s (****U****nique* ***M****olecular* ***I****ndexes)*
  - combination of barcodes + insert sequence provides accurate quantification
  - but requires specialized pre- and post-processing

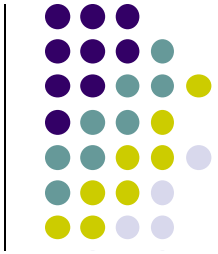


Single *Cell* sequencing

- Standard sequencing library starts with ***millions*** of cells
  - will be in different states unless synchronized
  - a heterogeneous “ensemble” with (possibly) high cell-to-cell variability
- ***Single cell sequencing*** technologies aim to capture this variability
  - e.g: cells in different tissue layers/regions or different areas of a tumor
  - essentially a very sophisticated library preparation technique
- Typical protocol (RNA-seq)
  1. isolate a few thousand cells (varying methods)
  2. the single-cell platform partitions each cell into an emulsion droplet
     - e.g. 10x Genomics (https://www.10xgenomics.com/solutions/single-cell/)
  3. a different barcode is added to the RNA in each cell
  4. resulting library submitted for standard Illumina short-read sequencing
  5. custom downstream analysis links results to their cell (barcode) of origin



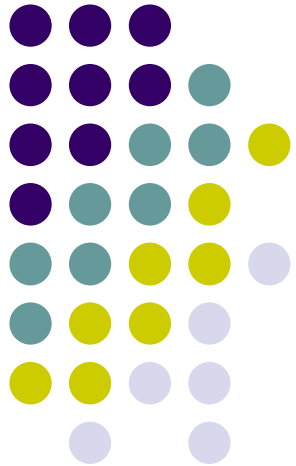
Part 2 summary

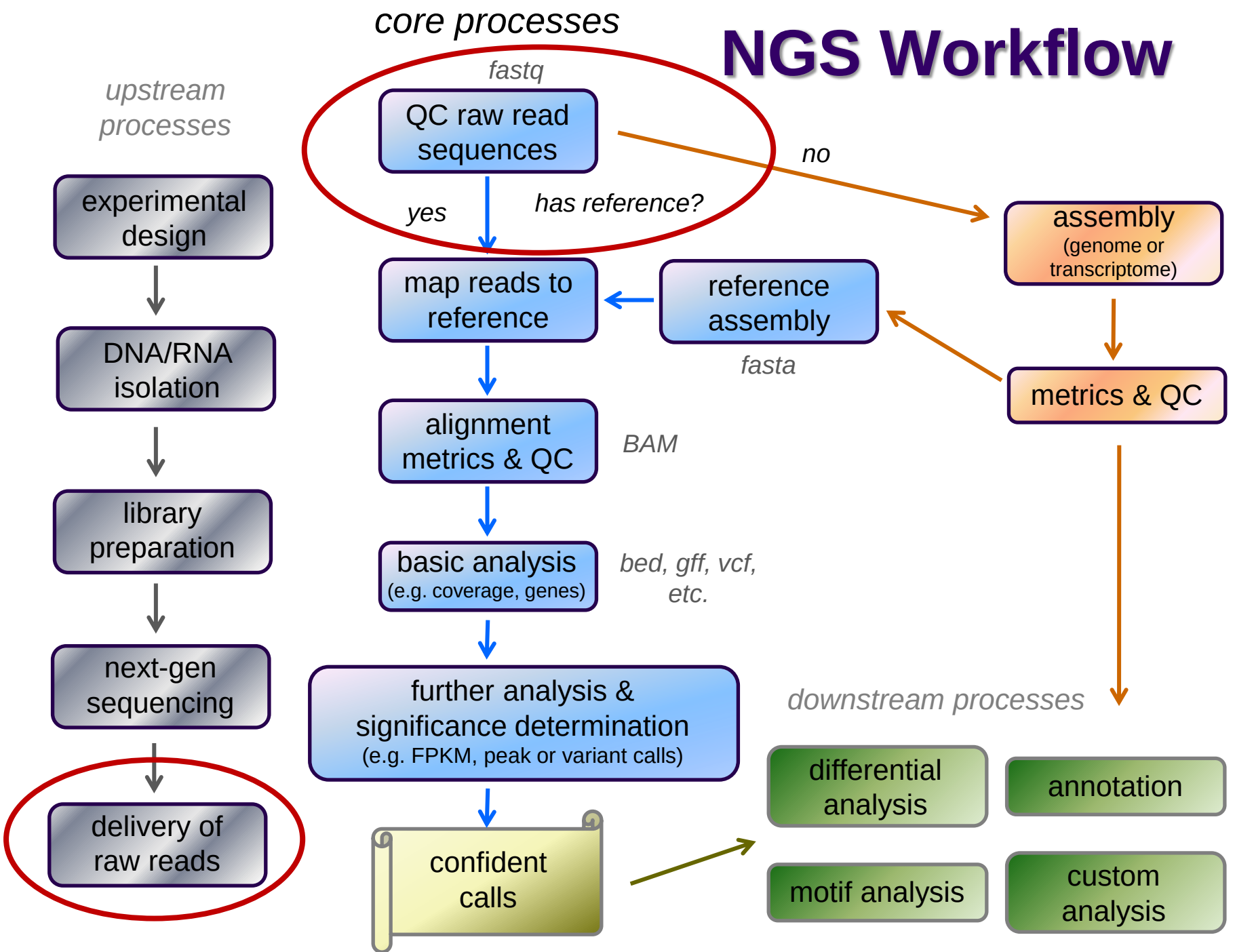


- Read terminology:
  - ***Single-end*** (SE) reads sequence from one end only (**R1**)
  - ***Paired-end*** (PE) reads sequence inwardly from both ends (**R1/R2 pair**)
  - ***Adapters*** (including primers & library barcodes) are fixed size & added to both ends
  - ***DNA library fragments*** have a ***distribution of insert sizes*** between adapters (~200 bases)
- ***Library complexity*** describes the number of ***different molecular species***
  - ***Many*** distinct molecules → ***high complexity***; ***Few*** → ***low complexity***
  - Primarily a function of the experiment type's ***enrichment profile***
    - ***Less*** enrichment → ***higher complexity***; ***More*** enrichment → ***lower complexity***
  - Popular experiment types have different ***expected*** library complexity
    - Whole genome sequencing → ***high complexity***; Amplicons → ***low complexity***
- ***Expected sequence duplication*** is also a function of library complexity
  - Also due to ***desired enrichment*** (***biological/good***) or ***PCR duplicates*** (***techincal/bad***)
  - Only addition of ***Unique Molecular Indexes*** (***UMI***s) can properly distinguish
- ***Single cell sequencing*** is designed to capture cell-to-cell variability
  - Essentially a sophisticated library prep technique followed by short-read sequencing

Part 3: The FASTQ format, Data QC & preparation

- **FASTA** and **FASTQ** formats
- QC of raw sequences with **FastQC** tool
- Dealing with adapters




core processes
NGS Workflow
upstream processes
fastq
QC raw read sequences
no
yes
has reference?
experimental design
assembly (genome or transcriptome)
DNA/RNA isolation
map reads to reference
reference assembly
fasta
metrics & QC
library preparation
alignment metrics & QC
BAM
next-gen sequencing
basic analysis (e.g. coverage, genes)
bed, gff, vcf, etc.
further analysis & significance determination (e.g. FPKM, peak or variant calls)
downstream processes
differential analysis
annotation
delivery of raw reads
confident calls
motif analysis
custom analysis

FASTQ format

- Text format for storing sequence and quality data
  - http://en.wikipedia.org/wiki/FASTQ_format
- 4 lines per sequence:
  1. ***@read name*** *(plus extra information after a space)*
     - *R1 and R2 reads have the same read name*
  2. ***called base sequence*** **(ACGTN)**
     always 5' to 3'; *usually* excludes 5' adapter
  3. ***+optional information***
  4. ***base quality scores encoded as text characters***

- FASTQ representation of a single, 50 base R2 sequence

```
@HWI-ST1097:97:D0WW0ACXX:4:1101:2007:2085 2:N:0:ACTTGA
ATTCTCCAAGATTTGGCAAATGATGAGTACAATTATATGCCCCAATTTACA
+
?@@?DD;?;FF?HHBB+:ABECGHDHDCF4?FGIGACFDFH;FHEIIIB9?
```

FASTQ quality scores

- Base quality ***probabilities*** expressed as ***Phred*** scores
  - ***Phred*** scores are log scaled, ***higher = better***
  - ***Quality*** **20** = $1.0e^{-2}$ = 1/100 errors, **30** = $1.0e^{-3}$ = 1/1000 errors

$$\textbf{Probability of Error} = 10^{-Q/10}$$

- Integer ***Phred*** score converted to ***Ascii*** text (add 33)

https://www.asciitable.com/

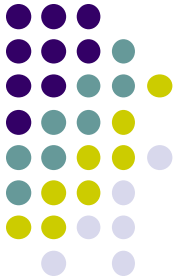
```
Quality character   !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
                    |          |          |          |          |
ASCII Value         33         43         53         63         73
Base Quality (Q)    0          10         20         30         40
```

```
?@@?DD;?;FF?HHBB+:ABECGHDHDCF4?FGIGACFDFH;FHEIIIB9?
```

Raw sequence quality control

- Critical step! Garbage in = Garbage out
  - general sequence quality
    - base quality distributions
    - initial sequence duplication rate
  - trim 3' bases with poor quality?
    - important for *de novo* assembly
  - trim 3' adapter sequences?
    - important for RNA-seq
  - other contaminents?
    - biological – rRNA in RNA-seq
    - technical – samples sequenced w/other barcodes

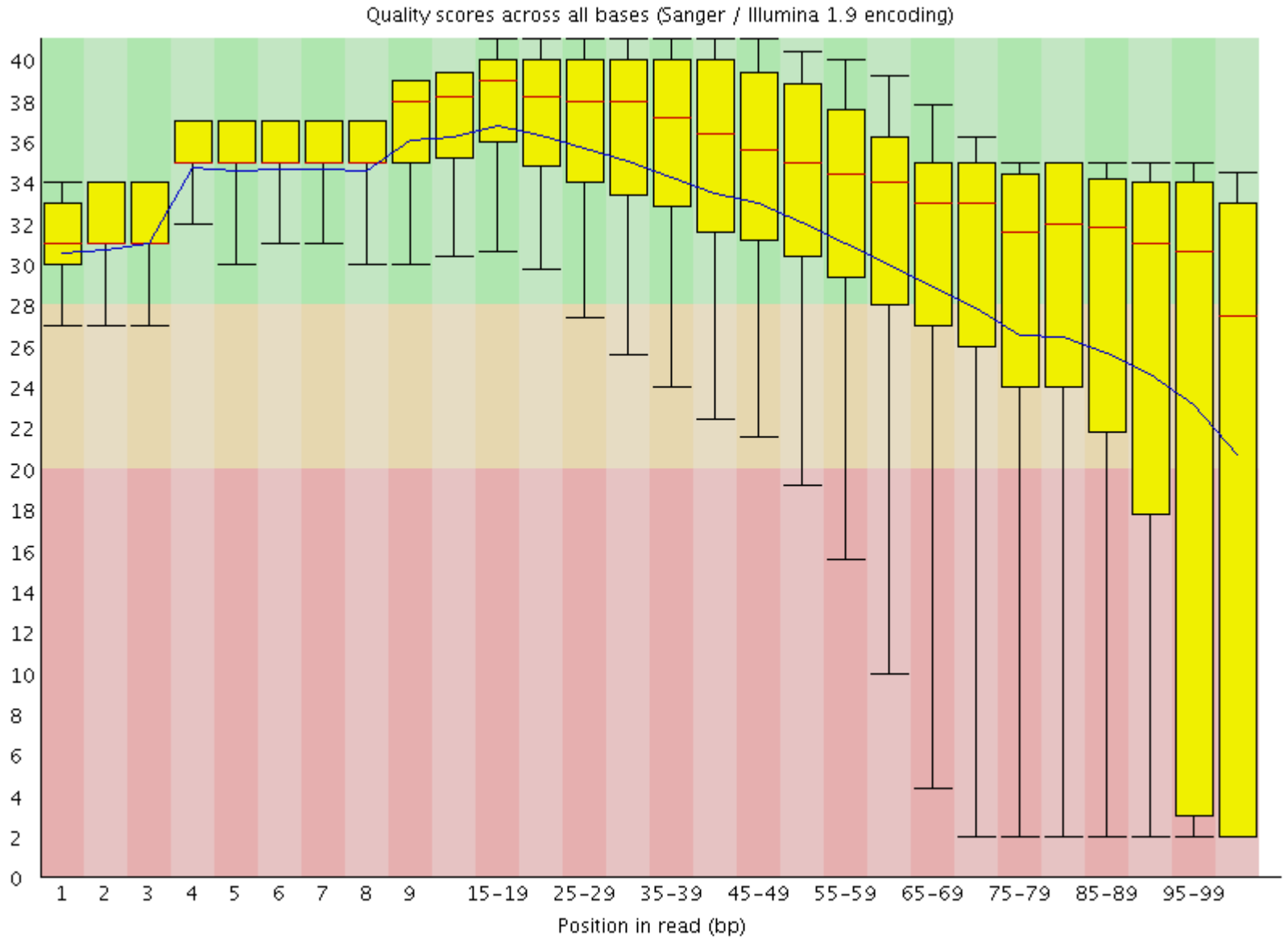
FastQC



- Quality Assurance tool for FASTQ sequences
  http://www.bioinformatics.babraham.ac.uk

- Most useful reports:
  1. Should I trim low quality bases?
     - *Per-base sequence quality Report*
  2. How complex is my sequence library?
     - *Sequence duplication levels Report*
  3. Do I need to remove adapter sequences?
     - *Overrepresented sequences Report*

1. FastQC Per-base sequence quality report



2. FastQC Sequence duplication report

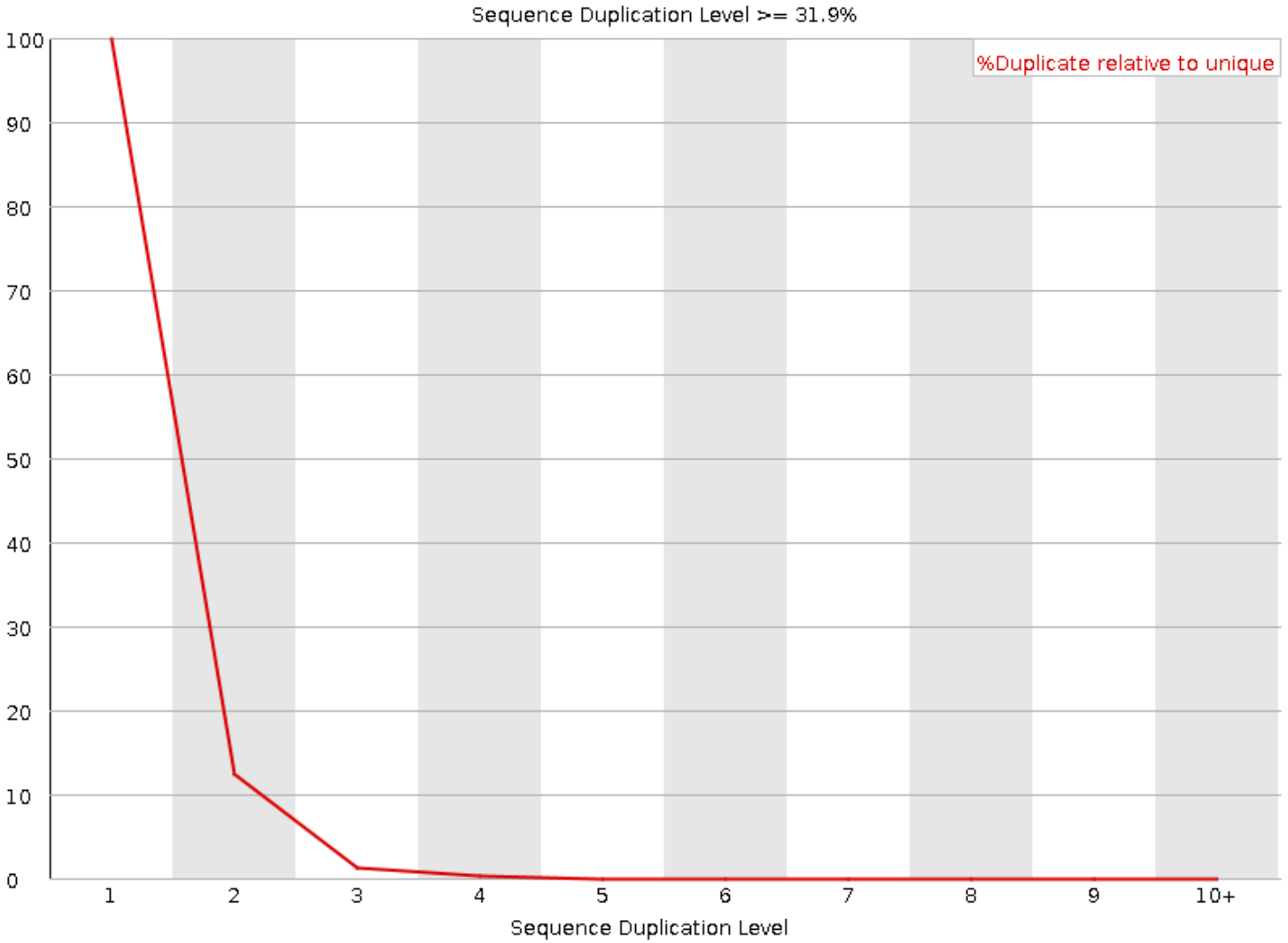
Yeast ChIP-seq

For every 100 unique sequences there are:

~12 sequences w/2 copies

~1-2 with 3 copies

Ok – Some duplication expected due to IP enrichment



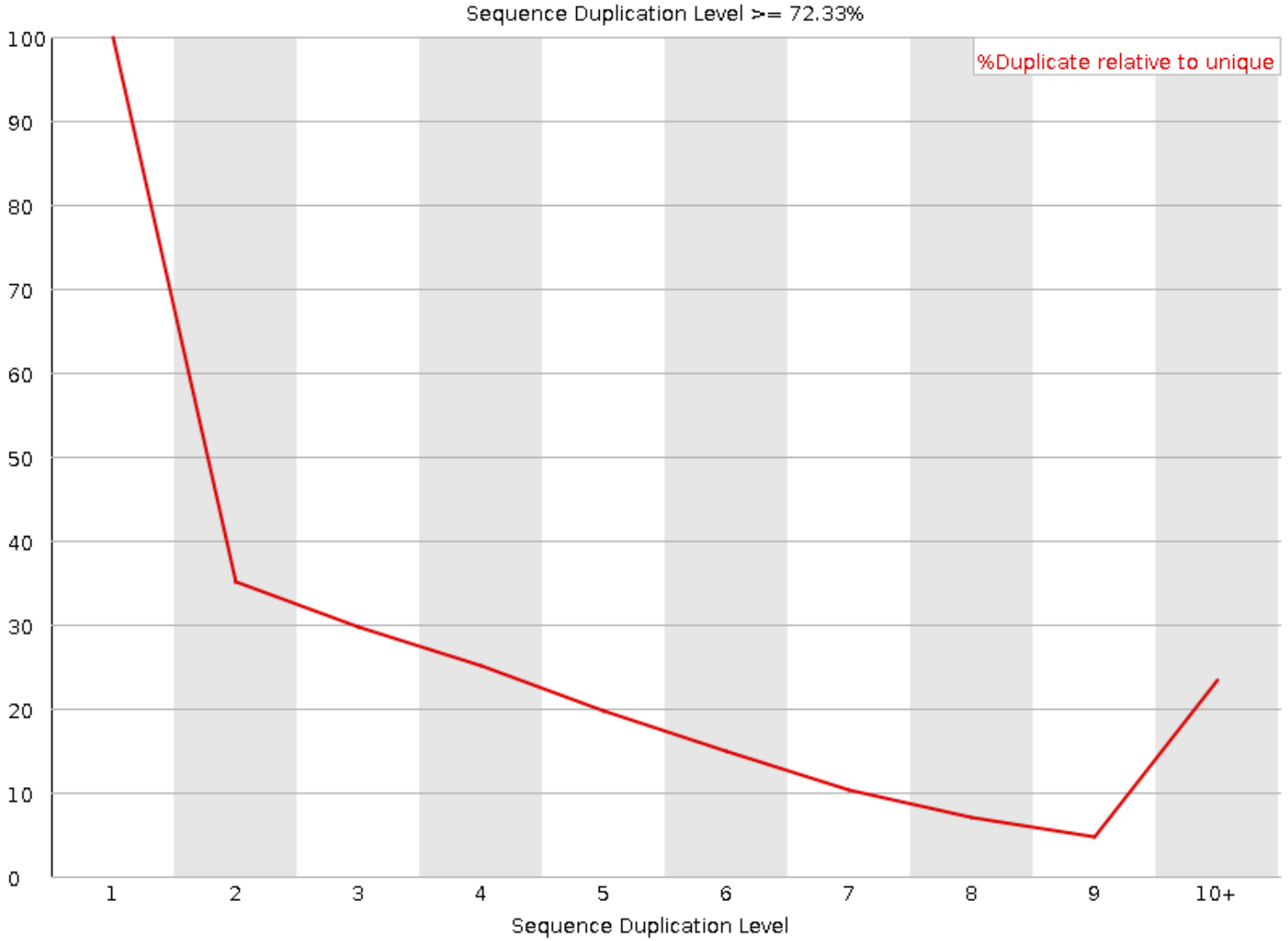
2. Sequence duplication report

Yeast ChIP-exo

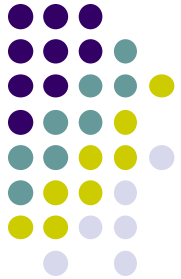
For every 100 unique sequences there are:

- ~35 sequences w/2 copies
- ~22 with 10+ copies

Success! Protocol expected to have high duplication



Expected sequence duplication is primarily a function of experiment type



higher complexity

Less enrichment for specific sequences

genomic

bisulfite-seq

exon capture

RNA-seq

ChIP-seq

ChIP-exo

amplicons

More enrichment for specific sequences

lower complexity

Higher diversity of library molecules
Lower sequence duplication expected
More sequencing depth required

… as well as…
- *genome size & sequencing depth*
- *library construction skill & luck!*

Lower diversity of library molecules
Higher sequence duplication expected
Less sequencing depth required

3. FastQC Overrepresented sequences report

- **FastQC** knows Illumina adapter sequences
- Here ~9-10% of sequences contain adapters
  - calls for adapter removal or trimming

Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATG	60030	5.01369306977828	TruSeq Adapter, Index 1 (97% over 37bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATGC	42955	3.5875926338884896	TruSeq Adapter, Index 1 (97% over 37bp)
CACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATGCCGTCTTCTGCT	3574	0.29849973398946483	RNA PCR Primer, Index 40 (100% over 41bp)
CAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTAT	2519	0.2103863542024236	TruSeq Adapter, Index 1 (97% over 37bp)
GAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTAT	1251	0.10448325887543942	TruSeq Adapter, Index 1 (97% over 37bp)

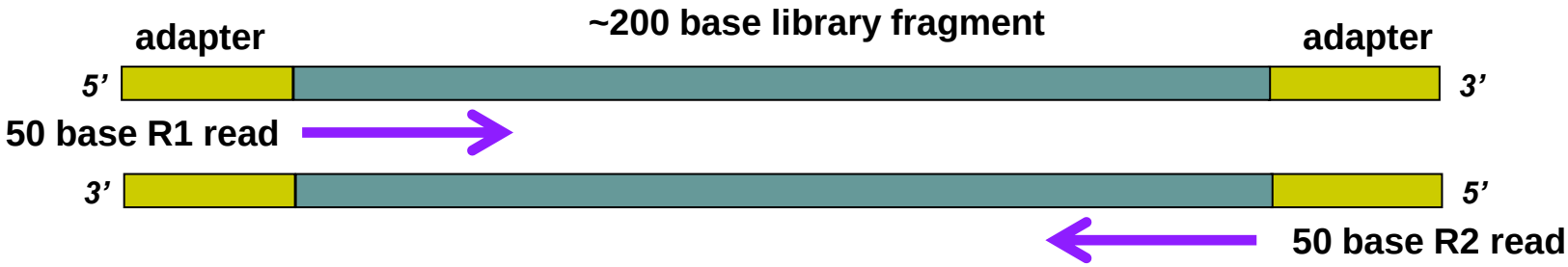
3. Overrepresented sequences

- Here nearly 1/3 of sequences some type of non-adapter contamination!
  - **BLAST** the sequence to identify it

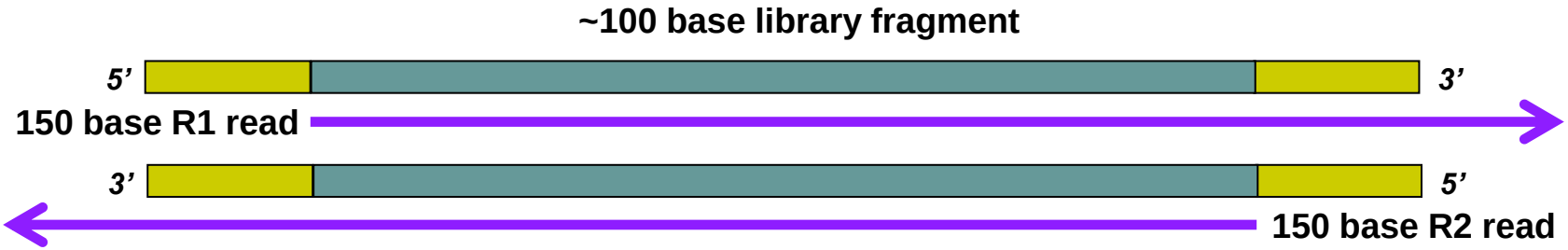
Sequence	Count	Percentage	Possible Source
GAAGGTCACGGCGAGACGAGCCGTTTATCATTACGATAGGTGTCAAGTGG	5632816	32.03026785752871	No Hit
TATTCTGGTGTCCTAGGCGTAGAGGAACAACACCAATCCATCCCGAACTT	494014	2.8091456822607364	No Hit
TCAAACGAGGAAAGGCTTACGGTGGATACCTAGGCACCCAGAGACGAGGA	446641	2.539765344040083	No Hit
TAAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAAGAAC	179252	1.0192929387357474	No Hit
GAAGGTCACGGCGAGACGAGCCGTTTATCATTACGATAGGGGTCAAGTGG	171681	0.9762414422996221	No Hit
AACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAAGAACGTA	143415	0.8155105483274229	No Hit
AGAACATGAAACCGTAAGCTCCCAAGCAGTGGGAGGAGCCCTGGGCTCTG	111584	0.6345077504066322	No Hit
AAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAAGAACG	111255	0.6326369351474214	No Hit
ATTACGATAGGTGTCAAGTGGAAGTGCAGTGATGTATGCAGCTGAGGCAT	73682	0.41898300890326096	No Hit
GAAGGTCACGGCGAGACGAGCCGTTTATCATTACGATAGGTGTCAAGGGG	71661	0.4074908580252516	No Hit
GGATGCGATCATACCAGCACTAATGCACCGGATCCCATCAGAACTCCGCA	69548	0.3954755612388914	No Hit
ATATTCTGGTGTCCTAGGCGTAGAGGAACAACACCAATCCATCCCGAACT	54017	0.30716057099328803	No Hit

3’ Adapter contamination

A. reads short compared to fragment size (no contamination)



B. Reads long compared to library fragment (3’ adapter contamination)



The presence of the 3’ adapter sequence in the read can cause problems during alignment, because it does not match the genome.

Dealing with 3’ adapters

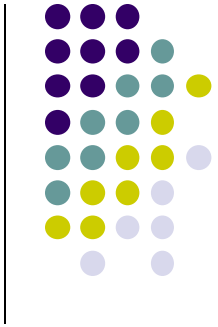
- Three main options:

1. ***Hard trim*** all sequences by specific amount
   - e.g. trim 100 base reads to 50 bases
   - *Pro*: fast & easy to perform; trims low-quality 3’ bases
   - *Con*: removes information (bases) you might want

2. ***Remove adapters*** specifically
   - e.g. using specific tools *(always needed for RNA-seq alignment)*
   - *Pro*: removes adapter contamination without losing sequenced bases
   - *Con*: requires knowledge of insert fragment structure & adapters

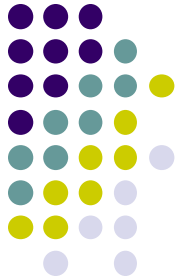
3. Perform a ***local alignment*** (vs ***global***)
   - e.g. **bowtie2 --local** or **bwa mem**
   - *Pro*: mitigates adapter contamination while retaining full query sequence
   - *Con*: limited aligner support

FASTQ trimming and adapter removal



- Tools:
  - **cutadapt** – https://code.google.com/p/cutadapt/
  - **trimmomatic** – http://www.usadellab.org/cms/?page=trimmomatic
  - FASTX-Toolkit – http://hannonlab.cshl.edu/fastx_toolkit/
- Features:
  - hard-trim specific number of bases
  - trimming of low quality bases
  - specific trimming of adapters
  - support for trimming paired end read sets (except FASTX-Toolkit)
  - **cutadapt** has protocol for separating reads based on internal barcode

Local vs. Global alignment



- ***Global*** alignment
  - requires query sequence to map ***fully*** (end-to-end) to reference
- ***Local*** alignment
  - allows a ***subset*** of the query sequence to map to reference
    - "untemplated" 5' and 3' sequences will be "soft clipped" (ignored)

global *(end-to-end) alignment of query*

CACAAGTACAATTATACAC

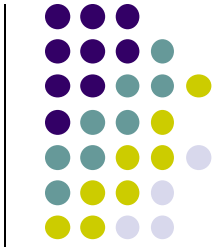
local *(subsequence) alignment of query*

CTAGCTTATCGCCCTGAAGGACT

TACATACACAAGTACAATTATACACAGACATTAGTTCTTATCGCCCTGAAAATTCTCC

reference sequence

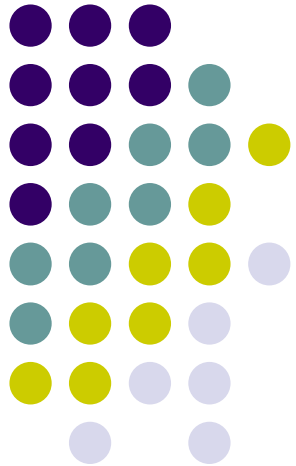
Part 3 summary

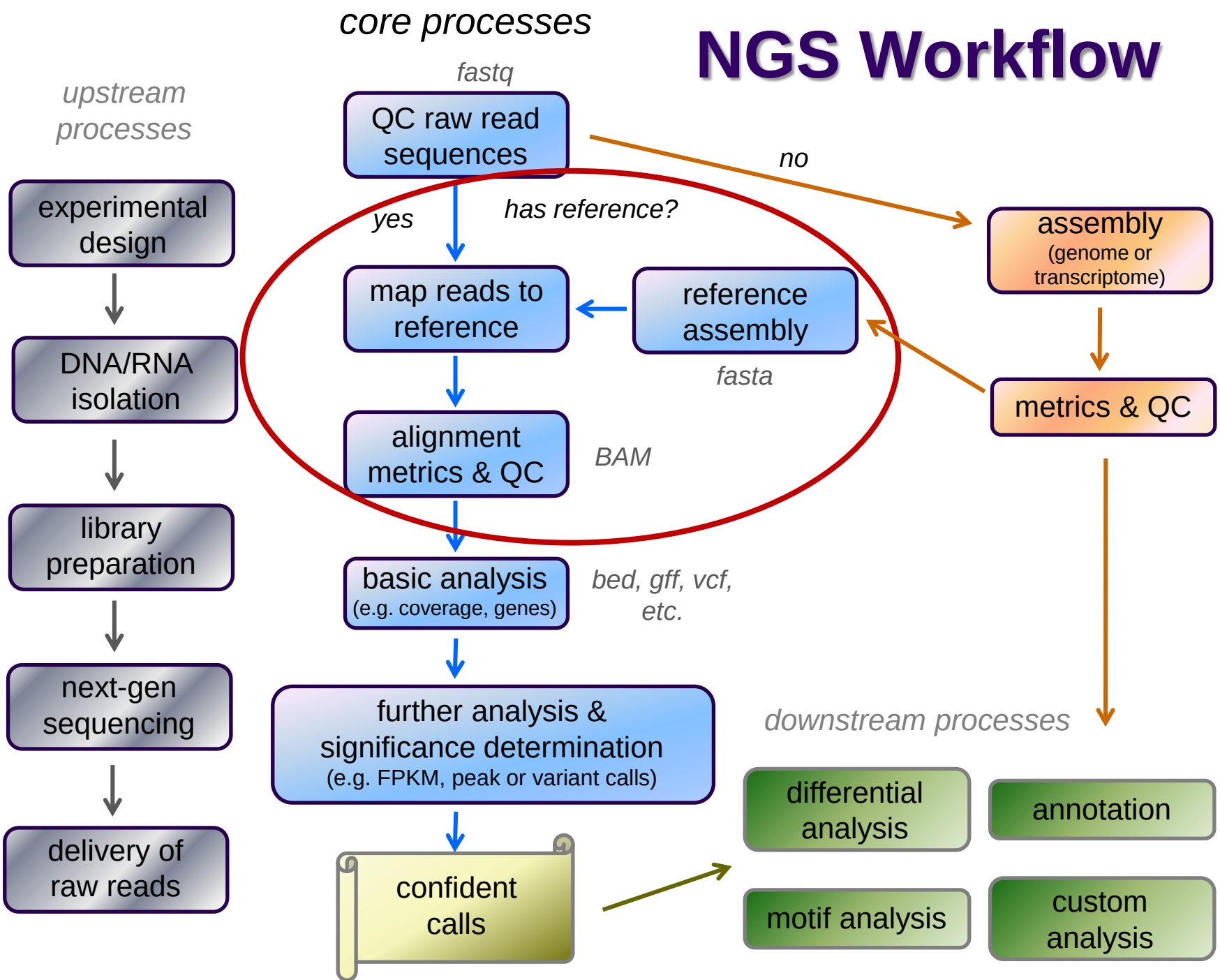


- Sequenced reads are delivered as **FASTQ** files
  - R1 read file(s) and R2 file(s) if paired end, with 4 lines per read:
    - read name; called sequence; optional info; ASCII-encoded quality scores
- The **FastQC** tool generates quality reports for each **FASTQ** file
  1. ***Per-base sequence quality*** → trim low quality bases?
  2. ***Sequence duplication level*** → sequence complexity estimate
  3. ***Overrepresented sequences*** → adapter trimming needed?
- ***3' adapter contamination*** can prevent reads from aligning
  - Dealing with adapter contamination:
    1. Hard trim a specific number of bases (simple, but information is lost)
    2. Remove adapters specifically (required for RNA-seq analysis)
       - Tools: **cutadapt**, **trimmomatic**
    3. Perform ***local alignment*** (versus standard ***global alignment***)
       - Not suitable for RNA-seq reads since splice-junction information is missed

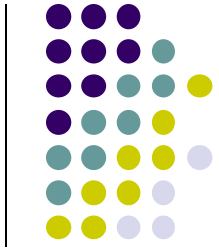
Part 4: Alignment to a reference assembly

- Alignment overview & concepts
- Preparing a reference genome
- Alignment workflow steps




core processes
NGS Workflow
upstream processes
fastq
QC raw read sequences
no
experimental design
yes
has reference?
assembly
(genome or transcriptome)
map reads to reference
reference assembly
fasta
DNA/RNA isolation
metrics & QC
alignment metrics & QC
BAM
library preparation
basic analysis
(e.g. coverage, genes)
bed, gff, vcf, etc.
next-gen sequencing
further analysis & significance determination
(e.g. FPKM, peak or variant calls)
downstream processes
differential analysis
annotation
delivery of raw reads
confident calls
motif analysis
custom analysis

Short Read Aligners



- Short read mappers determine placement of ***query sequences*** (your reads) against a known ***reference***
  - **BLAST**:
    - ***one*** query sequence (or a few)
    - many matches for each
  - short read aligners
    - many ***millions*** of query sequences
    - want only one “best” mapping (or a few) for each
- Many aligners available! Two of the most popular
  - **bwa** (Burrows Wheeler Aligner) by Heng Li
    http://bio-bwa.sourceforge.net/
  - **bowtie2** – part of the Johns Hopkins Tuxedo suite of tools
    http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml
  - Given similar input parameters, they produce similar alignments
    - and both run relatively quickly

Mapping *vs* Alignment

- ***Mapping*** determines one or more ***positions*** (a.k.a. ***seeds*** or ***hits***) where a read shares a *short* sequence with the reference
- ***Alignment*** starts with the seed and determines how read bases are best ***matched***, base-by-base, around the seed
- Mapping quality and alignment scores are both reported
  - High mapping quality ≠ High alignment score
  - ***mapping quality*** describes ***positioning***
    - reflects the probability that the read is ***incorrectly*** mapped to the reported location
    - is a Phred score: $P(\text{incorrectly mapped}) = 10^{-\text{mappingQuality}/10}$
  - ***alignment score*** describes ***fit***
    - reflects the correspondence between the read and the reference sequence

• Maps to one location
high mapping quality

• Has 2 mismatches
low alignment score

Read 1

```
GCGTAGTCTGCC
||  ||||  ||||
TAGCCTAGTGTGCCGC
```

Read 2

```
ATCGGGAGATCC
||||||||||||
TAATCGGGAGATCCGC
```

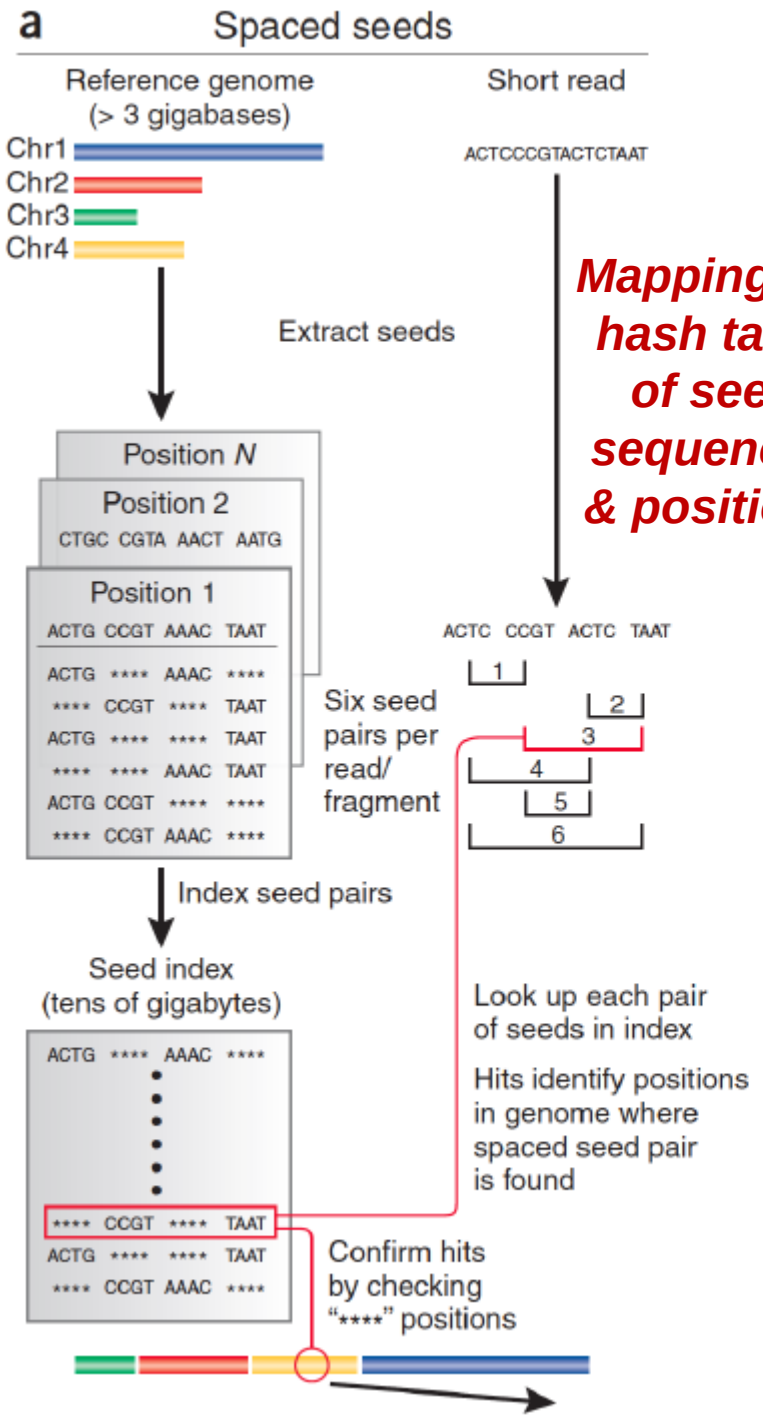
or

```
ATCGGGAGATCC
||||||||||||
TTATCGGGAGATCCGC
```

• Maps to 2 locations
low mapping quality

• Matches perfectly
high alignment score

reference sequence



Mapping via hash table of seed sequences & positions

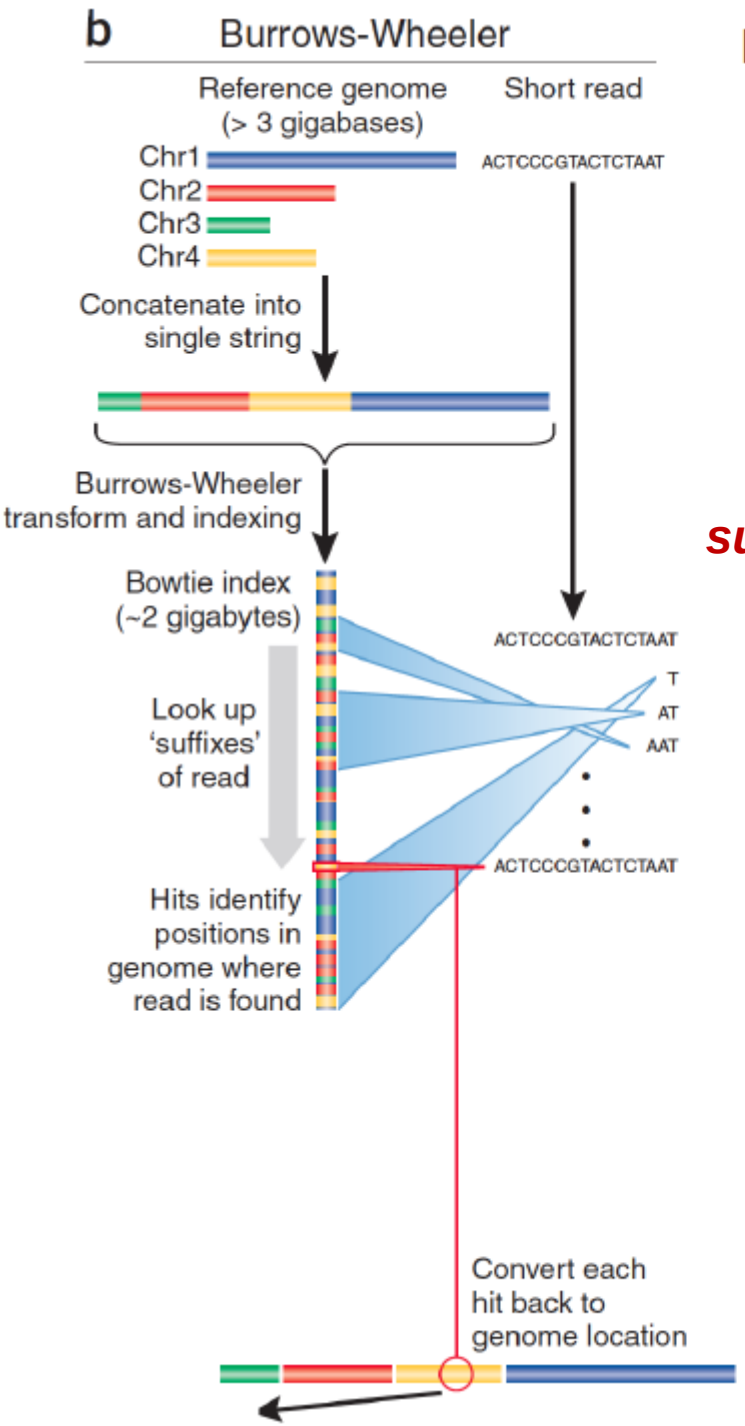
Hash table enables lookup of exact matches.

Subsequence	Reference Positions
ATAGCTAATCCAAA	2341, 2617264
ATAGCTAATCCAAT	
ATAGCTAATCCAAC	134, 13311, 732661,
ATAGCTATCCAAAG	
ATAGCTAATCCATA	
ATAGCTAATCCATT	3452
ATAGCTAATCCATC	
ATAGCTATCCAATG	234456673

Table is sorted and complete so you can jump immediately to matches. (But this can take a lot of memory.)

May include N bases, skip positions, etc.

Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech.* **27**, 455–457 (2009).

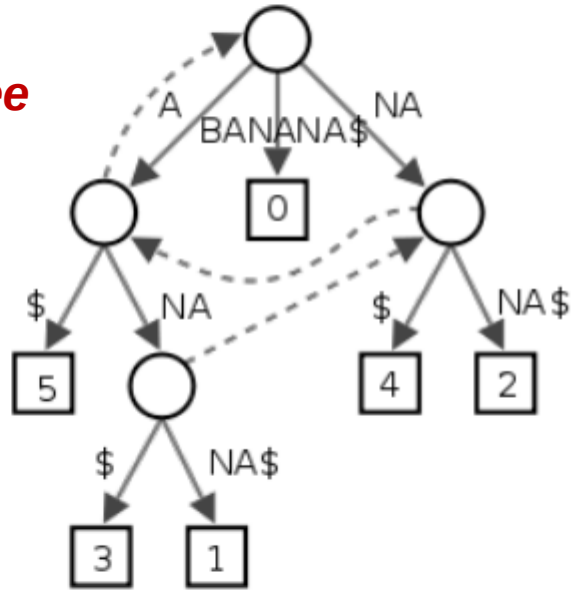


Burrows-Wheeler transform compresses sequence.

Input	SIX.MIXED.PIXIES.SIFT.SIXTY.PIXIE.DUST.BOXES
Output	TEXYDST.E.IXIXIXXSSMPPS.B..E.S.EUSFXDIIOIIIT

Suffix tree enables fast lookup of subsequences.

Mapping via suffix array tree



http://en.wikipedia.org/wiki/Suffix_tree

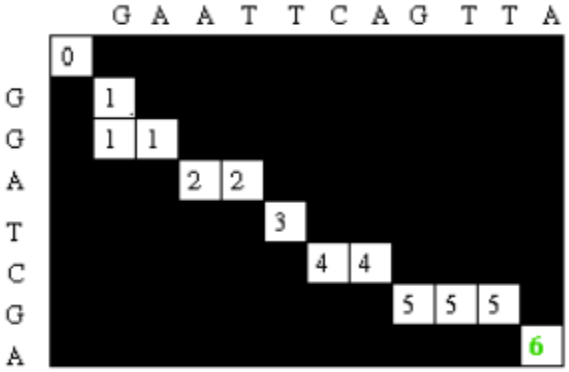
Exact matches at all positions below a node.

Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech.* **27**, 455–457 (2009).

Alignment via dynamic programming

- Dynamic programming algorithm (Smith-Waterman | Needleman-Wunsch)

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	1	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5	5
A	0	1	2	3	3	3	4	5	5	5	5	6



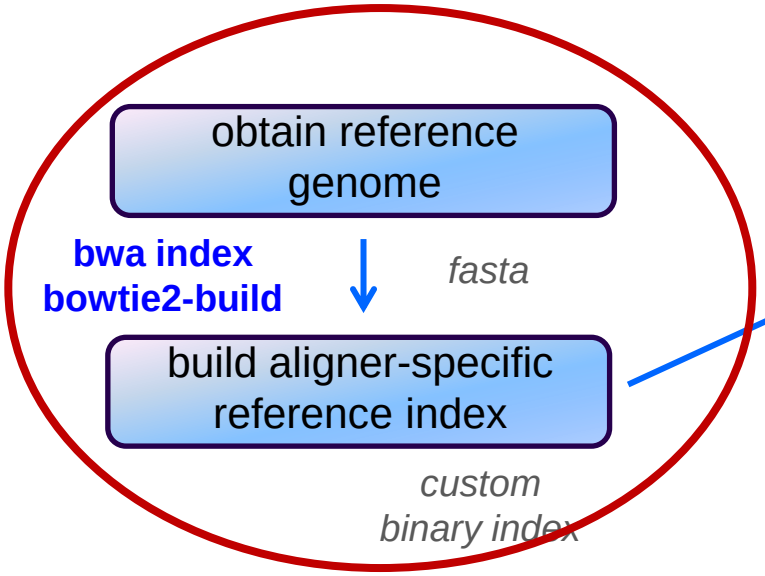

```
G _ A A T T C A G T T A
|   |   | |   |     |
G G _ A _ T C _ G _ _ A
```

- **Alignment score = Σ**
  - match reward
  - base mismatch penalty
  - gap open penalty
  - gap extension penalty
  - rewards and penalties may be adjusted for quality scores of bases involved

Reference sequence

```
ATTTGCGATCGGATGAAGACGAA
|||||||||||||||
ATTTGCGATCGGATGTTGACTTT

ATTTGCGATCGGATGAAGACG..AA
|||||||||||||||XX|||Xii||
ATTTGCGATCGGATGTTGACTTTAA
```



Alignment Workflow

fastq

QC & trim raw reads

fastq

FastQC, cutadapt

align reads to reference

SAM

bwa aln + bwa samse or **sampe**, **bwa mem**, or **bowtie2**

convert SAM to BAM

BAM

samtools view

sort BAM by position

BAM

samtools sort

handle duplicates *(optional)*

BAM

Picard MarkDuplicates
samtools rmdup

index BAM

BAM + .bai

samtools index

alignment metrics & QC

samtools flagstat
samtools idxstat

http://bio-bwa.sourceforge.net/bwa.shtml

http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

Obtaining/building a reference

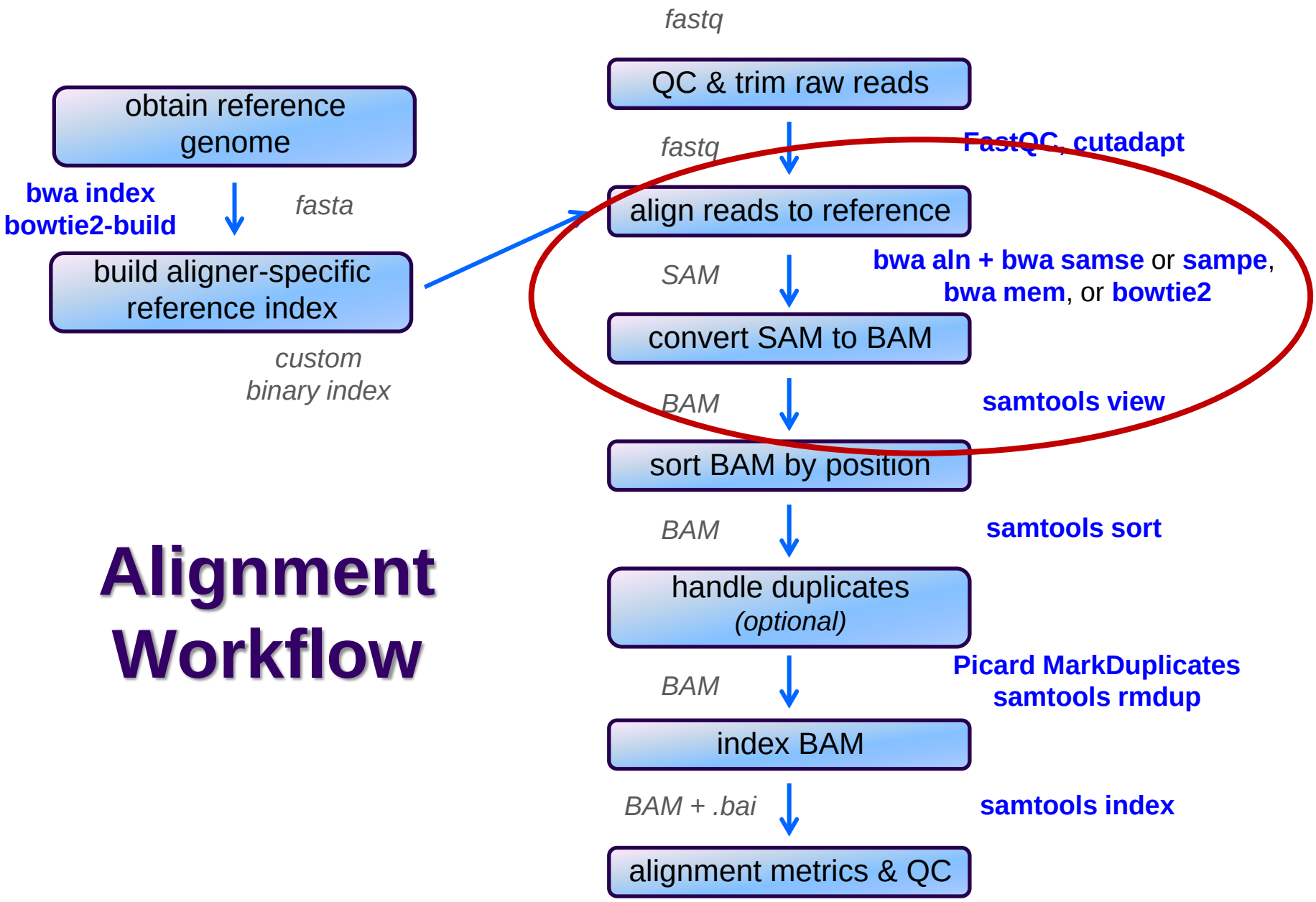
- What is a ***reference***?
  - assembled genomes or transcriptomes
    - Ensembl, UCSC, for eukaryotes
    - NCBI RefSeq or GenBank for prokaryotes/microbes *(prefer RefSeq)*
  - *any set of named DNA sequences*
    - names are chromosome or gene names (technically referred to as ***contigs***)
- Building a reference index (aligner-specific)
  - may take several hours to build
    - but you build each index once, use for multiple alignments
  - requires FASTA files (**.fa, .fasta**) containing DNA sequences
    - annotations (genome feature files, **.gtf**) may also be used to build the index, but will definitely be needed for downstream analysis

```
>chrM  Mitochondiral Chromosome
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCAT
TTGGTATTTTCGTCTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCTG
GAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATT
...
```

sequence name line
- ***always*** starts with **>**
- followed by a ***name*** and other (optional) descriptive information

one or more line(s) of ***sequence characters***
- ***never*** starts with **>**

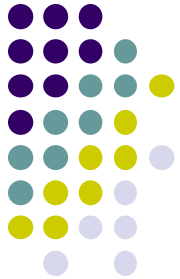
Alignment Workflow



http://bio-bwa.sourceforge.net/bwa.shtml

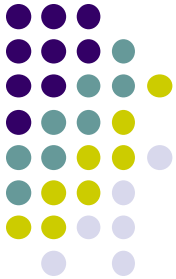
http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

SAM / BAM file format



- Aligners take **FASTQ** as input, output alignments in **S**equence **A**lignment **M**ap (**SAM**) format
  - plain-text file format that describes how reads align to a reference
    - http://samtools.github.io/hts-specs/SAMv1.pdf (the “Bible”)
    - and now https://github.com/samtools/hts-specs/blob/master/SAMtags.pdf
- **SAM** and **BAM** are two forms of the same data
  - **BAM** – **B**inary **A**lignment **M**ap
    - ***same data*** in a custom compressed (**gzip**’d) format
    - ***much*** smaller than **SAM** files
    - when sorted + indexed, support fast random access (**SAM** files do not)
- **SAM** file consists of
  - a ***header*** (includes reference sequence names and lengths)
  - ***alignment records***, one for each sequence read
    - alignments for R1 and R2 reads have *separate records*
    - records have 11 fixed fields + extensible-format **key:type:value** tuples

SAM file format
Fixed fields (tab-separated)



Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAGs
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{29}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next segment
8	PNEXT	Int	[0,2^{29}-1]	Position of the mate/next segment
9	TLEN	Int	[-2^{29}+1,2^{29}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

read name from fastq

contig + start = locus

use this to find end coordinate

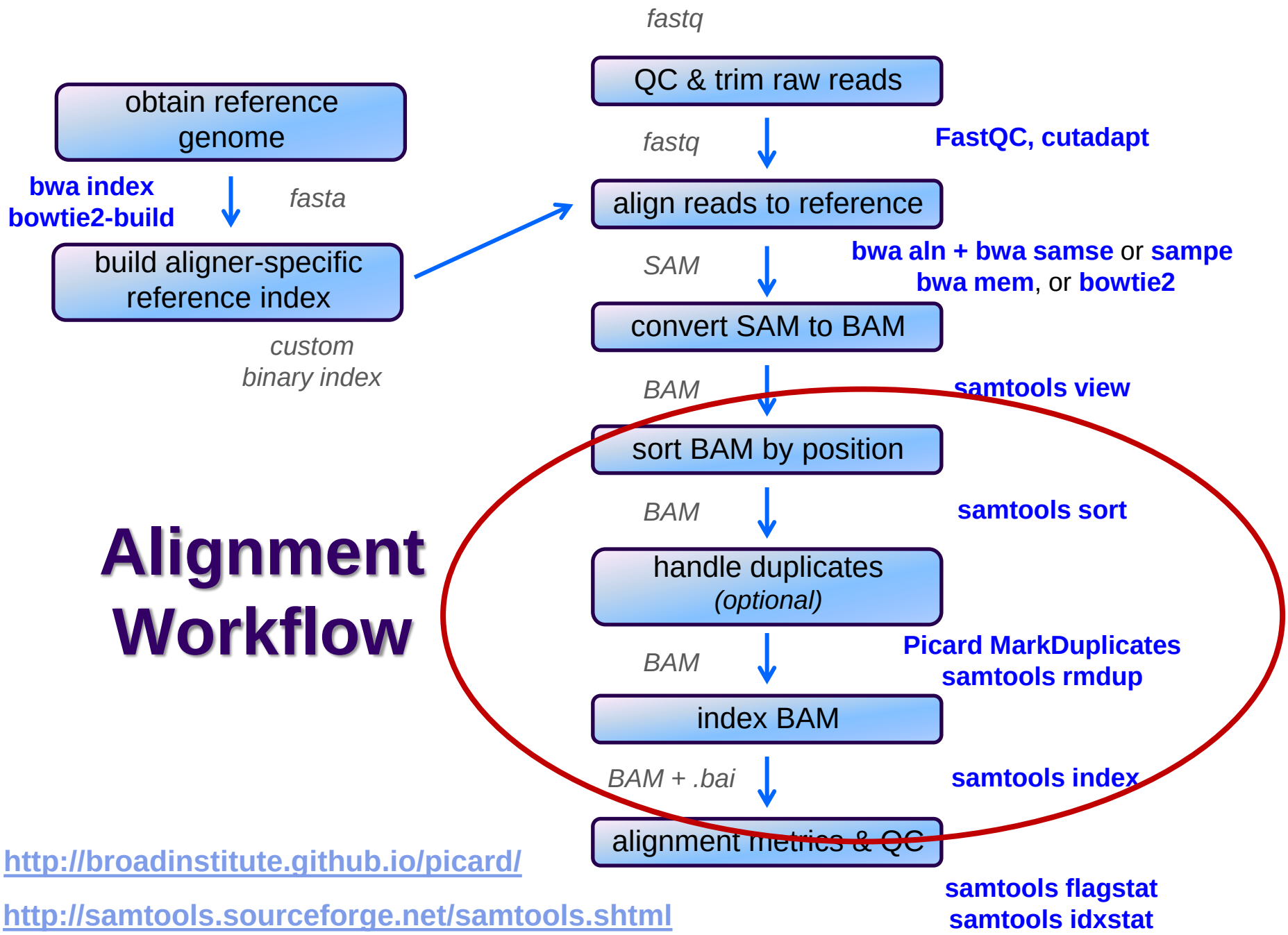
insert size, if paired

```
SRR030257.264529   99   NC_012967   1521   29   34M2S   =   1564   79
CTGGCCATTATCTCGGTGGTAGGACATGGCATGCCC
AAAAAA;AA;AAAAAA??A%.;?&'3735',()0*,
XT:A:M  NM:i:3  SM:i:29  AM:i:29  XM:i:3  XO:i:0  XG:i:0  MD:Z:23T0G4T4
```

positive for plus strand reads

```
SRR030257.2669090   147   NC_012967   1521   60   36M   =   1458   -99
CTGGCCATTATCTCGGTGGTAGGTGATGGTATGCGC
<<9:<<AAAAAAAAAAAAAAAAAAAAAAAAAAAA
XT:A:U  NM:i:0  SM:i:37  AM:i:37  X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:36
```

negative for minus strand reads

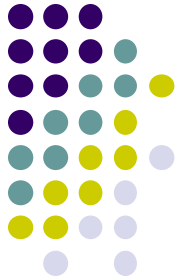


Alignment Workflow

http://broadinstitute.github.io/picard/

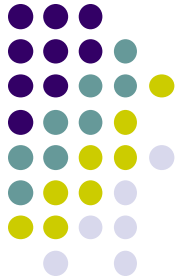
http://samtools.sourceforge.net/samtools.shtml

Sorting / indexing BAM files



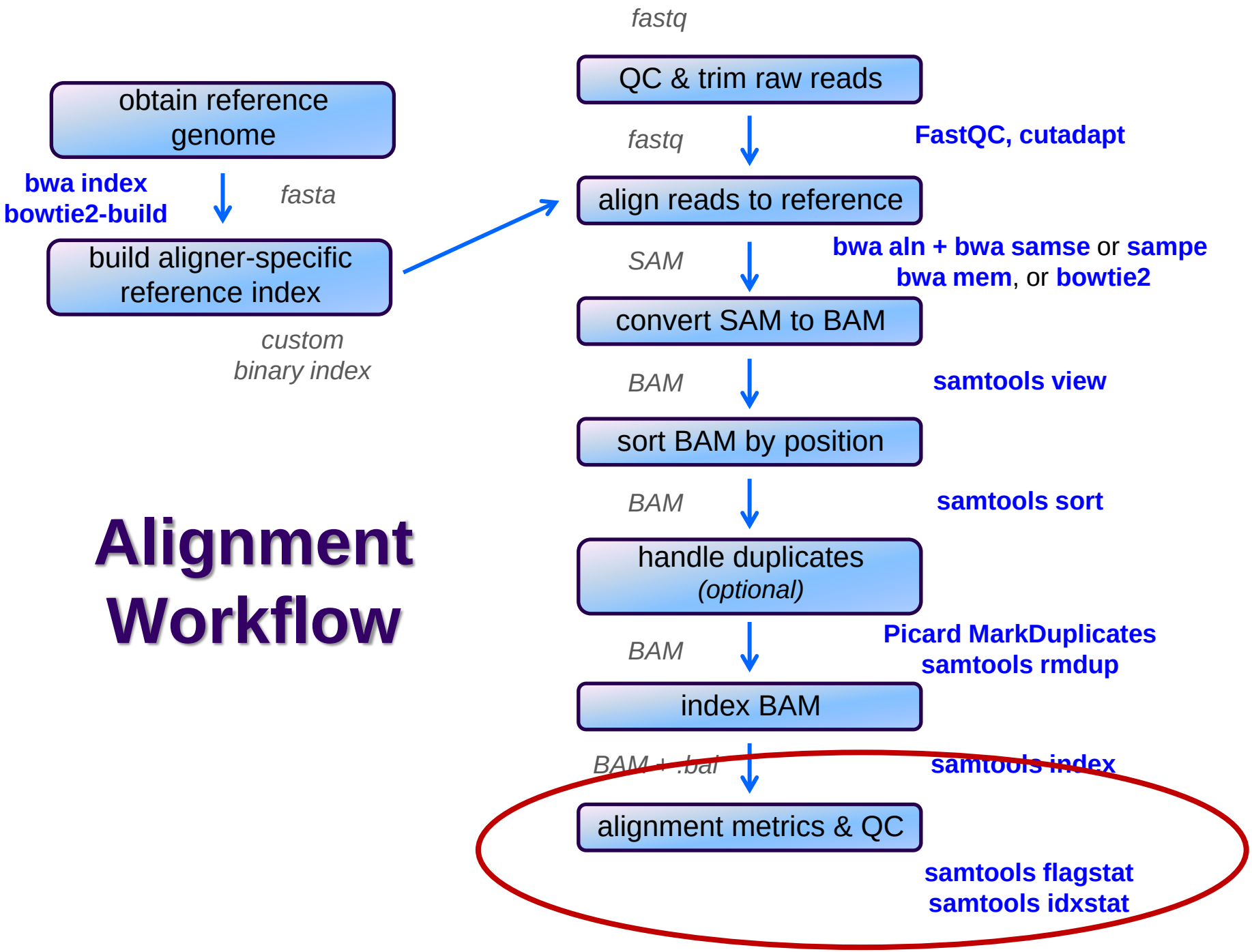
- **SAM** created by aligner contains read records in ***name order***
  - same order as read names in the input **FASTQ** file
    - R1, R2 have ***adjacent*** **SAM** records
  - **SAM** → **BAM** conversion does ***not*** change the name-sorted order
- Sorting **BAM** puts records in ***position (locus) order***
  - ***sorting is very compute, I/O and memory intensive!***
    - can take hours for large **BAM**
- Indexing a locus-sorted **BAM** allows fast random access
  - creates a small, binary alignment index file **(.bai)**
  - quite fast

Handling Duplicates

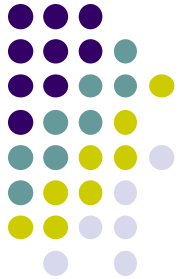


- Optional step, but very important for many protocols
- Definition of ***alignment duplicates***:
  - single-end reads or singleton/discordant PE alignment reads
    - alignments have the same ***start*** positions
  - paired-end reads
    - pairs have same ***external*** coordinates (5' + 3' coordinates of the ***insert***)
- Two choices for handling:
  - **samtools rmdup** – ***removes*** duplicates entirely
    - faster, but data is lost
  - **Picard MarkDuplicates** – ***flags*** duplicates only
    - slower, but all alignments are retained
  - both tools are quirky in their own ways

Alignment Workflow



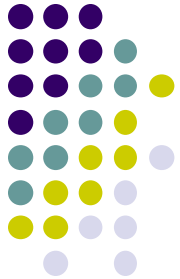
Alignment metrics



- **samtools flagstat**
  - simple statistics based on alignment record flag values
    - total sequences (R1+R2), total mapped
    - number properly paired
    - number of duplicates (0 if duplicates were not marked)

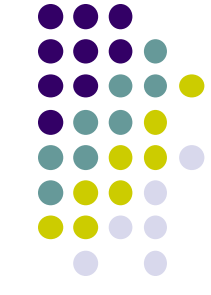
```
161490318 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
31602827 + 0 duplicates
158093331 + 0 mapped (97.90% : N/A)
161490318 + 0 paired in sequencing
80745159 + 0 read1
80745159 + 0 read2
153721151 + 0 properly paired (95.19% : N/A)
156184878 + 0 with itself and mate mapped
1908453 + 0 singletons (1.18% : N/A)
1061095 + 0 with mate mapped to a different chr
606632 + 0 with mate mapped to a different chr (mapQ>=5)
```

Alignment wrap up



- Many tools involved
  - choose one or two and learn their options well
- Many steps are involved in the full alignment workflow
  - important to go through manually a few times for learning
    - but gets tedious quickly!
  - best practice
    - automate series of complex steps by wrapping into a ***pipeline script***
      - e.g. **bash** or **python** script
  - the Bioinformatics team has a set of pipeline scripts available at TACC and BRCF "pods"
    - **align_bowtie2_illumina.sh**, **align_bwa_illumina.sh**, **trim_adapters.sh**, etc.
    - TACC: shared project directory **/work/projects/BioITeam/common/script/**
    - BRCF pods: read-only mount **/mnt/bioi/**

Part 4 summary



- Short read aligners determine placement of ***query sequences*** (your reads) against a known ***reference*** (e.g. a genome)
  - ***Mapping*** determines one or more ***positions*** (a.k.a. ***seeds*** or ***hits***) where a read shares a *short* sequence with the reference
  - ***Alignment*** starts with the seed and determines how read bases are best ***matched***, base-by-base, around the seed

- The alignment workflow:
  - Obtain a suitable reference (**FASTA**) + annotations (**gtf**)
  - Build an aligner-specific index (one-time)
  - Align QC'd reads to the reference → **SAM** file; convert **SAM** → **BAM** & sort by position
  - Handle duplicates (optional, but informative)
  - Index the **BAM** for fast random access
  - Gather and interpret alignment statistics

Other NGS Resources at UT

- CBRS training courses
  - Intro to NGS, RNAseq, many others
    https://research.utexas.edu/cbrs/cores/cbb/educational-resources/
  - "Summer School" (4 or 5 half-day sessions in June)
    - lots of hands-on, including w/TACC
  - Short courses (3-4 hour workshops) – starting next week!
    https://site.research.utexas.edu/cbrs/classes/short-courses/spring-2025-semester/
- Genome Sequencing & Analysis Facility (GSAF)
  - Jessica Podnar, Director, gsaf@utgsaf.org
- Bioinformatics consultants
  - Dennis Wylie, Dhivya Arasappan, Anna
  - BioITeam wiki – https://wikis.utexas.edu/display/bioiteam/
- Biomedical Research Support Facility (BRCF)
  - provides local compute and managed storage resources
    https://wikis.utexas.edu/display/RCTFUsers

Final thoughts

- Good judgement comes from experience

  unfortunately…
- Experience comes from bad judgement!
- So go get started making your 1st 1,000 mistakes….