Genomics and Big Data

A brief review of methods and practical utility

Amelia Weber Hall

January 2020

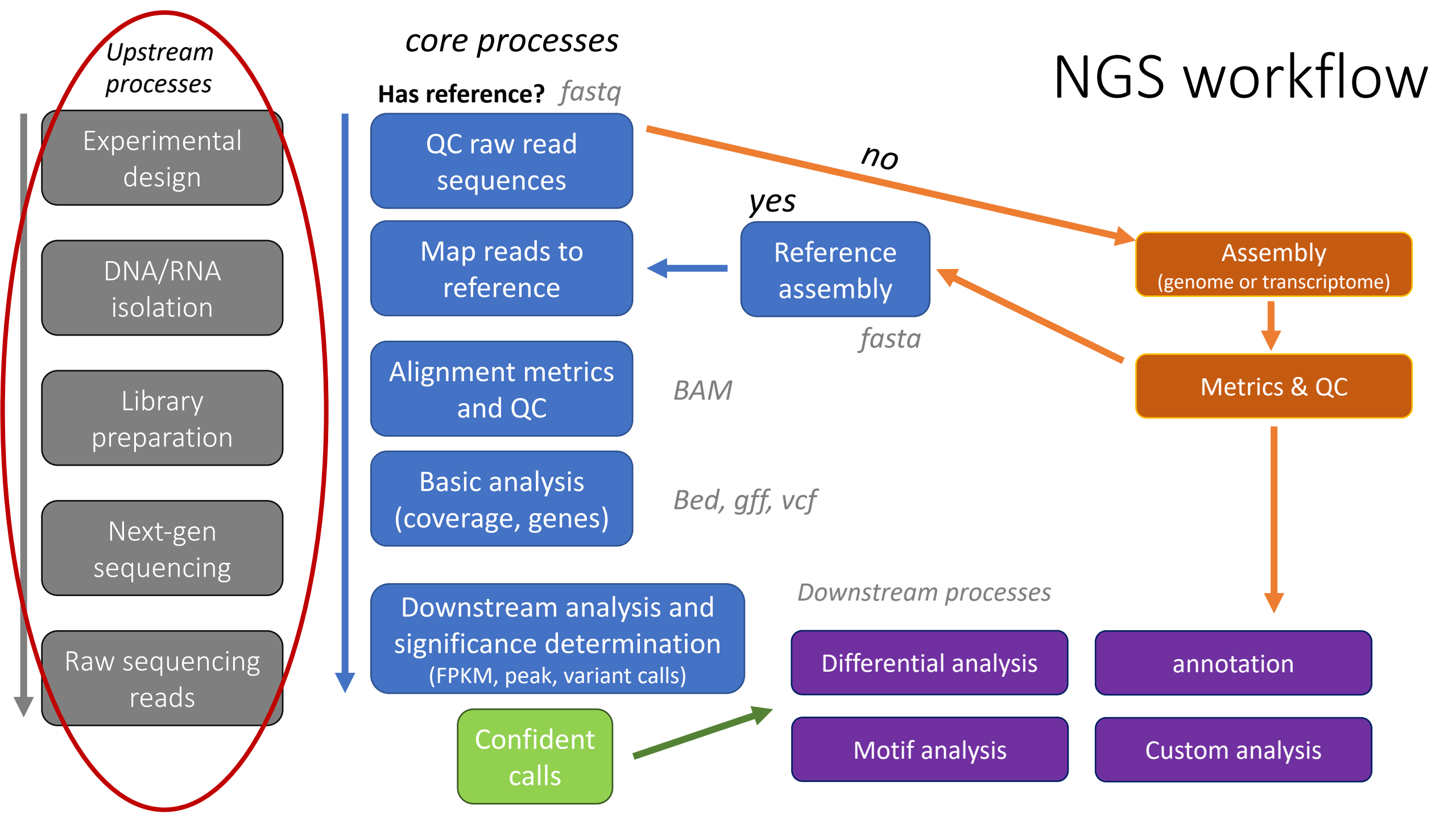
Computational Workgroup

With many thanks to Anna Battenhouse (UT Austin) who developed this original lecture back in ~2013

Outline

1. NGS workflow and experiment types
2. Read sequence terminology
3. The fastq format
4. Fastq QC methods
5. Alignment to a genome
6. Reference genomes: making and using
7. Alignment metrics and QC
8. UCSC genome browser time!

NGS workflow



Common Experiment Types: Genomic DNA

- Whole genome sequencing (WGS)
  - Library: all genomic DNA
  - Uses: Genome assembly, variant calling
    - Variants: methyl-seq
- Whole exome sequencing (WXS, or exome)
  - Library: coding regions (exons and sometimes adjacent regulatory regions)
  - Uses: Polymorphism/SNP detection, genotyping, de-novo variant discovery

Common Experiment Types: transcribed DNA (RNA)

- RNA-sequencing:
  - Bulk:
    - Library: extract all RNA from sample and convert to cDNA
      - All fragments → total RNA → polyDT selection → mRNA → ribozero (remove rRNA)

        → small fragments (miRNA)
    - Uses: differential gene expression, isoform discovery, exitrons, eQTLs (with WGS)
  - Single cell:
    - Library: unique to each cell (UMI), extract all RNA from sample and convert to cDNA
    - Uses: identify new cell types in complex samples such as tissue, many other analyses used for RNA-seq, but at cellular subtype resolution

Common Experiment Types: *NA-protein interactions

- ChIP-seq:
  - Library: isolated DNA bound by proteins (histones, transcription factors)
  - Uses: analysis of regulatory networks, annotating the non-coding genome,
    - Use targeted antibodies to pull down DNA:protein complexes after crosslinking
    - Being supplanted by methods such as ChIPmentation, CUT&RUN, ATAC-seq
- RIP-seq:
  - Library: isolated RNA bound by proteins (transcription factors, chromatin remodelers, RNA editing proteins)
  - Uses: protein-bound RNAs, RNA regulation and modification after transcription

Uncommon Experiment Types: a grab-bag!

- NET-seq
- GRO-seq
- ”C” methods: HiC, 5C, 4C-seq, 3C

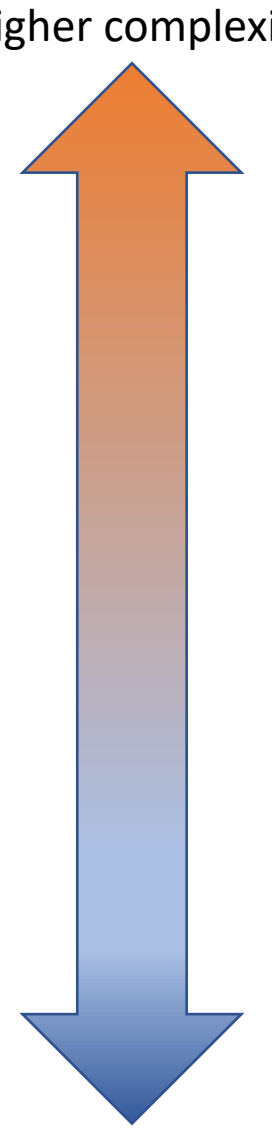
Library Complexity

- Is a measure of the number of diverse molecular species in a library

- Many different molecules → high complexity
- Few different molecules → low complexity

- Expected molecular diversity depends on the enrichment performed during library construction

Library complexity is primarily a function of experiment type



Less enrichment for specific sequences

Genomic (WGS)

Methyl-seq

RNA-seq (no selection)

Exome (WXS)

RNA-seq (polyA selected)

ChIP-seq

Amplicons/testing panels

More enrichment for specific sequences

Higher diversity of molecules
Less sequence duplication expected
More sequencing depth required

Other factors:
Genome size, sequencing depth
Method of library prep, luck

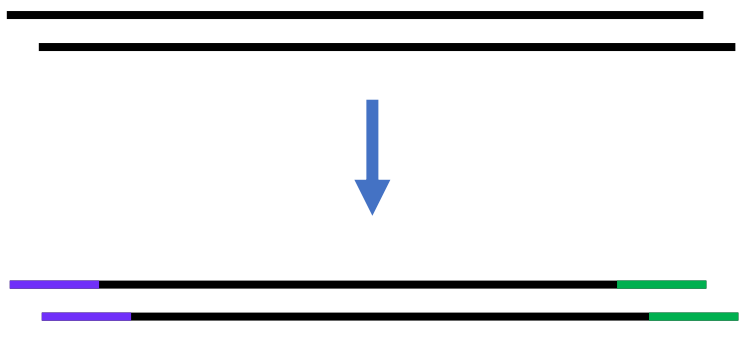
Lower diversity of molecules
Higher sequence duplication expected
Less sequencing depth required

Sequencing Technologies

- Illumina – short reads (can go up to 500bp)
  - Two PCR amplifications – library preparation, cluster generation
  - Amplification introduces bias!
- Single molecule sequencing
  - Sequencing of single molecules, not clusters
    - High error rate of reads, but much longer reads (multi-kilobase)
  - PACB – SMRT-seq system
    - Rolling circle replication, great for
  - Oxford Nanopore
    - Great for field applications, no refrigeration required, single use

Broad sequencing models and capabilities

Read Sequence Terminology



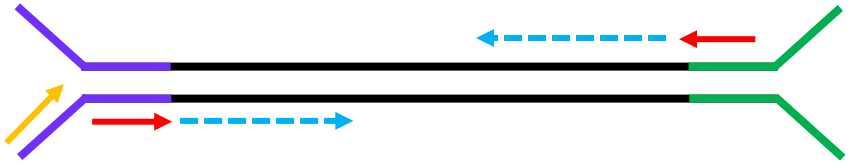
Original DNA sample (collection of fragmented DNA)

Library Prep

Sequencing Library
Double stranded or Y-adaptors added

Each side is different, which allows for paired-end sequencing

DNA sequencing



Primers: allow amplication and Sequence generation

Barcode – 6-12 bases
To allow for many samples to be run on a lane or flowcell
Known as **"multiplexing"**

Reads (50-1000+ bases)

Adapters include **primers** (P3/P7 for illumina) and a barcode: check supplier for details (Bioo, NEB, Illumina, others)
https://wikis.utexas.edu/display/GSAF/Illumina+-+all+flavors

Outline

1. NGS workflow and experiment types
2. Read sequence terminology
3. The fastq format
4. Fastq QC methods
5. Alignment to a genome
6. Reference genomes: making and using
7. Alignment metrics and QC
8. UCSC genome browser time!

Types of Illumina Sequencing

Single End

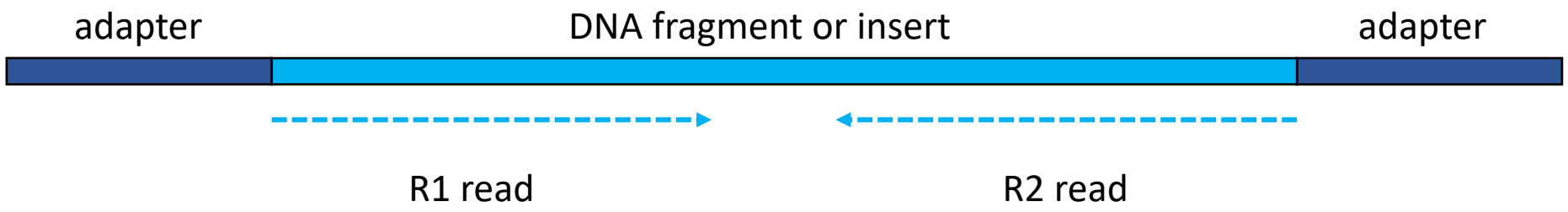
Independent reads from each molecule

Paired End

Two inwardly oriented reads, matched
And separated by ~200bp (or more)

Reads and Fragments

- When using paired end sequencing, keep in mind the distinction between:
  - The DNA ***fragment*** from your library that was sequenced
    - also called an ***insert***
  - The ***sequence reads*** you receive from the sequencing center
    - Called R1 and R2, older school lingo called them “tags”
  - An R1 and it’s associated R2 form a ***read pair***
    - This represents a readout of all or part of a DNA fragment/insert



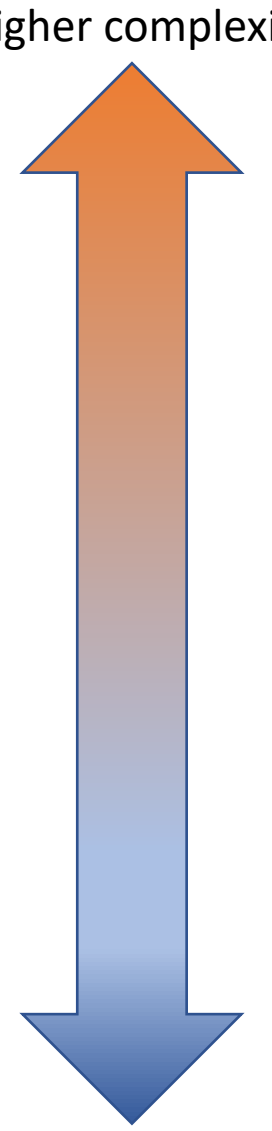
Paired end vs Single end read structures

- Single end is more rare due to the reduction in sequencing cost
- Paired end allows for more reliable mapping to a reference genome
  - Especially for lower complexity genomic regions
  - Able to determine actual fragment sizes
  - Allows better identification of duplicates
    - Better assessment of the true complexity of a library

Sequencing depth?

- Variable across sequencing experiments!
- Depends on:
  - Genome size
    - Prokaryotes: kilobases to 1-2 megabases
    - Lower eukaryotes (e.g. yeast) – megabases
    - Higher eukaryotes: gigabases
  - Library fragment enrichment
  - Theoretical library complexity
    - Less complex libraries don’t need as much depth
  - Desired sensitivity
    - Looking for rare mutations?

Library complexity is primarily a function of experiment type



Less enrichment for specific sequences

Genomic (WGS)

Methyl-seq

RNA-seq (no selection)

Exome (WXS)

RNA-seq (polyA selected)

ChIP-seq

Amplicons/testing panels

More enrichment for specific sequences

Higher diversity of molecules
Less sequence duplication expected
More sequencing depth required

More depth required for large genomes

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

Sequence Duplication

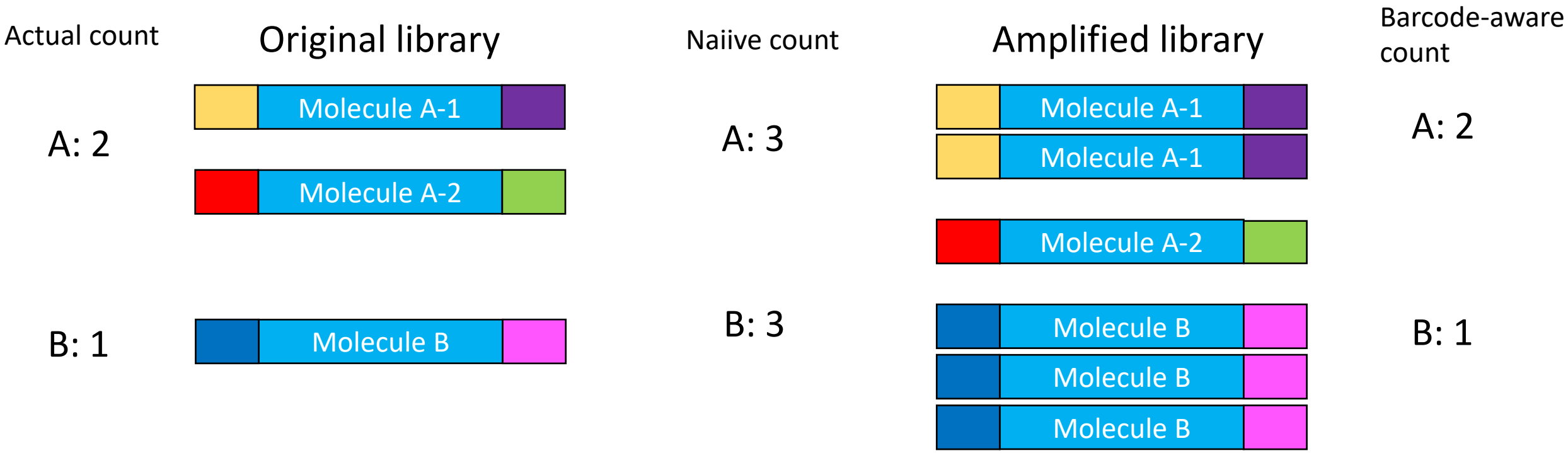
- Sequences from a library can contain exact duplicates
- Duplication can arise from
  - Sequencing of species enriched in your library (biological)
    - Each read comes from a different DNA cluster on the flowcell
  - Sequencing of PCR artifacts (technical)
    - Amplified PCR species (PCR duplicates)
    - Optical duplicates – two flowcell clusters overlap
- Current best practice is to ”mark duplicates” during the original processing of raw sequence reads
  - Can retain, discard, dose in duplicate reads
- Different experiment types have different expectation of duplication
  - Whole genome → high complexity and low duplication
  - Amplicon sequencing → low complexity and high duplication

Read vs Fragment Duplication

- 4 reads below: which are duplicates?
- Single end duplication: 50%
  - 2 unique and 2 duplicates
- Paired end duplication: 25%
  - 3 unique and 1 duplicate

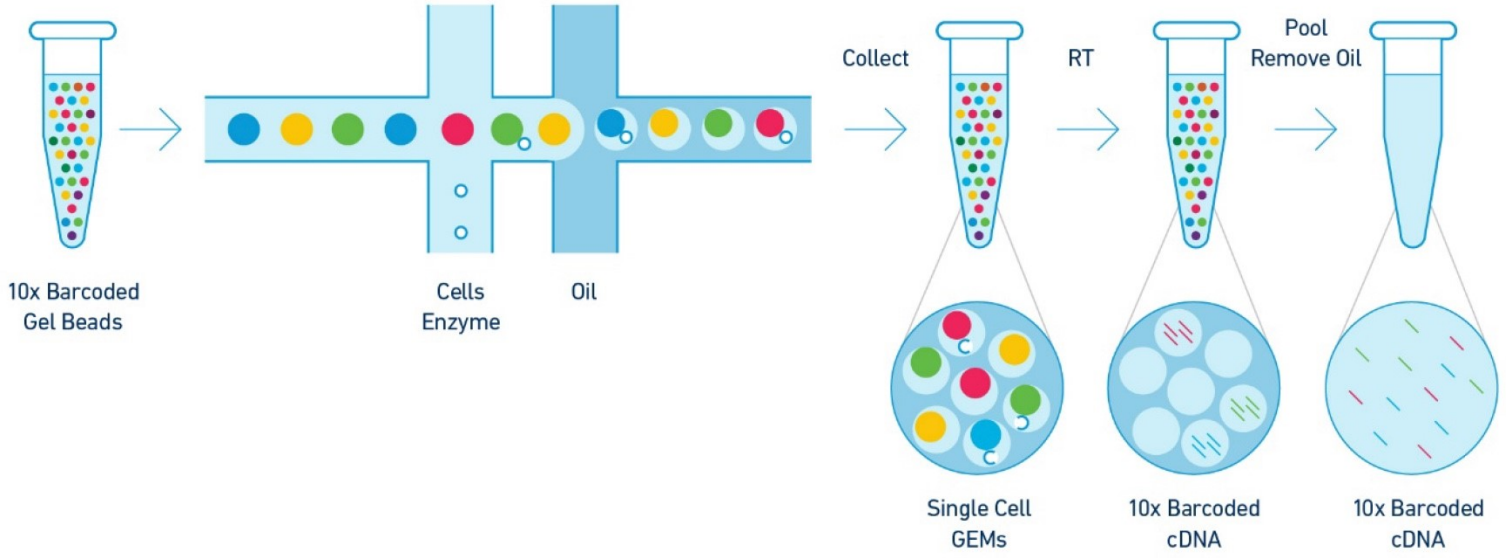
Molecular Barcoding (important for single cell methods!)

- Resolves ambiguity between biological and technical duplicates
  - Adds secondary barcodes to pre-PCR molecules
  - Barcodes + insert sequence can provide accurate quantification
  - Requires specialized pre and post processing



Single cell Sequencing

- A standard library takes DNA from many cells (thousands to millions)
  - If these cells are not all clones then “bulk” sequencing will not capture the true complexity of the RNA in each cell, and give an average signal
- Single cell sequencing aims to barcode each cell, so the reads from each cell are distinguishable
  - Allows for identification of new subtypes, subtype specific effects in cells



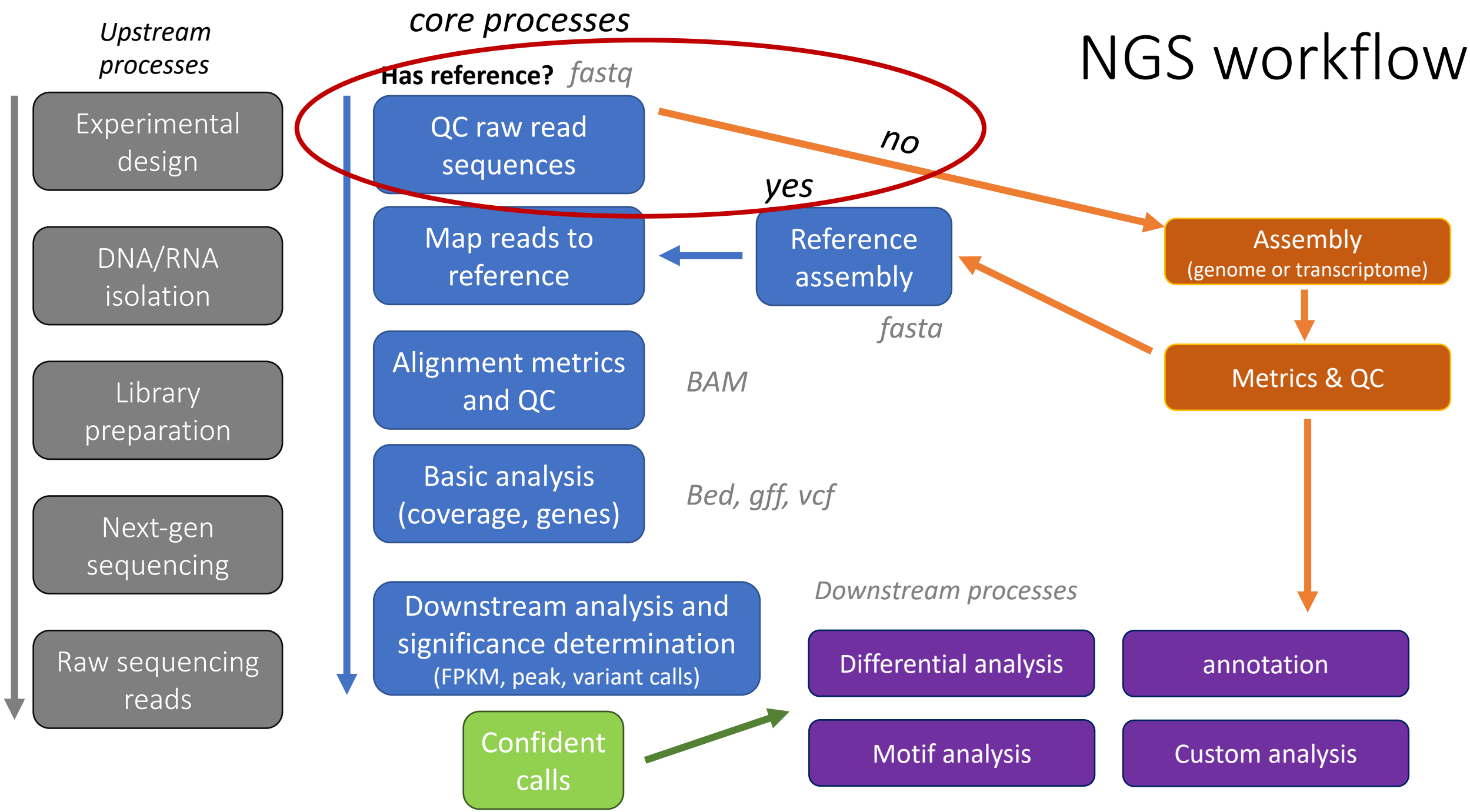
Some barcode/index types

- Library barcode
  - The same for all fragments in a library
  - About 96 available from Bioo, Illumina, NEB
- Molecular barcodes
  - Different small barcodes or pairs attached to DNA fragments before amplification
  - Diversity depends on barcode size and number
    - 4 well separated bases ~80
    - 2 x 4 well separated bases ~700
    - 2 x 8 well separated bases ~500K
    - finding well-separated sequencing compatible barcodes is not trivial
- Single cell barcodes or UMI
  - Unique barcodes generated using beads with 10^{12} possible unique molecular indices (UMI)

Outline

1. NGS workflow and experiment types
2. Read sequence terminology
3. The fastq format
4. Fastq QC methods
5. Alignment to a genome
6. Reference genomes: making and using
7. Alignment metrics and QC
8. UCSC genome browser time!

NGS workflow



FASTQ files

- Nearly all sequencing data are delivered as FASTQ files
  - FASTQ = FASTA sequences + Quality scores
  - Tend to have .fastq or .fq extensions
  - Generally compressed to save space (.gz extension)
    - Most tools will handle .fastq.gz files
- Paired end sequencing comes with 2 fastq files:
  - One for R1 and one for R2 – same number of rows

    ```
    1221-C_R1_001.fastq.gz
    1221-C_R2_001.fastq.gz
    ```

  - Order of reads is identical
    - Aligners rely on this identical ordering for paired end alignment

FASTQ format

- Text format for storing and manipulating sequence and quality data
  - https://en.wikipedia.org/wiki/FASTQ_format
- 4 lines per sequence:
  - **@**readname (generally specific to the sequencer)
  - Called base sequence (**ACTGN**)
    - Always 5’ → 3’
  - **+** optional read name
  - Base quality scores encoded as text characters

```
@GWNJ-0842:451:GW1902151877:2:1101:12753:1608 1:N:0:NTTACTCG+AGGATAGG
NACAGAACATAAAACATAAAAATATCAACCCTTTACAAAGATGATGAGAAATACAGCAAAGGCACCAGATCGGAAGA
+
#-A<-FFJFAJJJJJJJJJJFJJF-<-FFJJFFFFJJFF-FJJJJFFJJJJJJJFFAFJJ<A-7-A<7<FAFFA<J
```

FASTQ read names

- FASTQ readnames from Illumina data record information about the cluster location
  - Unique identifier (fragment name) begins with @
    - Sequencing machine name
    - Lane number
    - Flowcell grid coordinates
  - A space separates the name from extra read information
    - End number: 1 for R1, 2 for R2
    - Two quality fields (N = not QC failed)
    - Barcode sequence
      - This sample is dual indexed!
  - R1/R2 reads ***have the same fragment name***

```
@GWNJ-0842:451:GW1902151877:2:1101:12753:1608 1:N:0:NTTACTCG+AGGATAGG
@GWNJ-0842:451:GW1902151877:2:1101:12753:1608 2:N:0:NTTACTCG+AGGATAGG
```

FASTQ quality scores

- Base qualities expressed as ***Phred*** scores
  - Log scaled, higher = better
  - 20 = $1/10^2$ = 1/100 errors, 30 = $1/10^3$ = 1/1000 errors

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

- Integer Phred score converted to ACSII character (add 33)

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

```
#-A<-FFJFAJJJJJJJJJJFJJF-<-FFJJFFFFJJFF-FJJJJFFJJJJJJJJFFAFJJ<A-7-A<7<FAFFA<J
```

https://ascii.cl/

Handling sequencing data across multiple lanes

- For some sequencers (NextSeq) your sample will be split across all lanes
  - So you need to combine the lanes before processing the data for alignment
- Some argue that keeping data separate for as long as possible is best practices, but can also be hard to manage
  - Keep in mind that quality across all sequencing lanes is not necessarily identical
  - If you're having QC issues, try checking each lane separately
  - Keep all original FASTQ files until you've ascertained that alignment is unaffected by lane issues

Outline

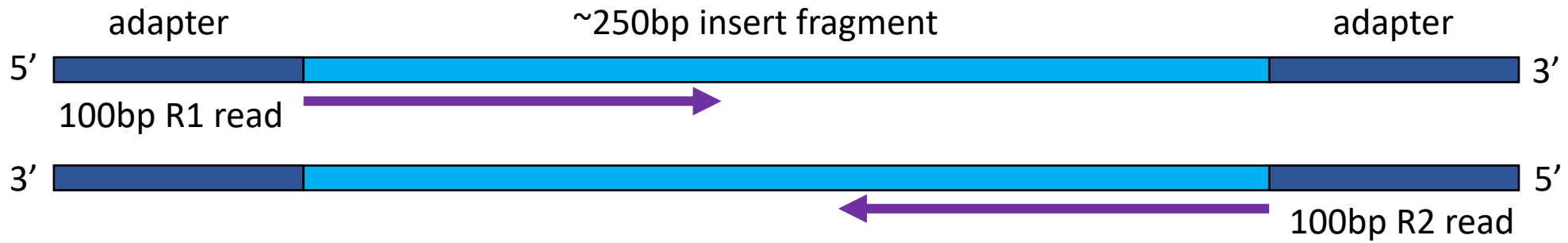
1. NGS workflow and experiment types
2. Read sequence terminology
3. The fastq format
4. Fastq QC methods
5. Alignment to a genome
6. Reference genomes: making and using
7. Alignment metrics and QC
8. UCSC genome browser time!

FASTQ QC and Raw sequence quality control

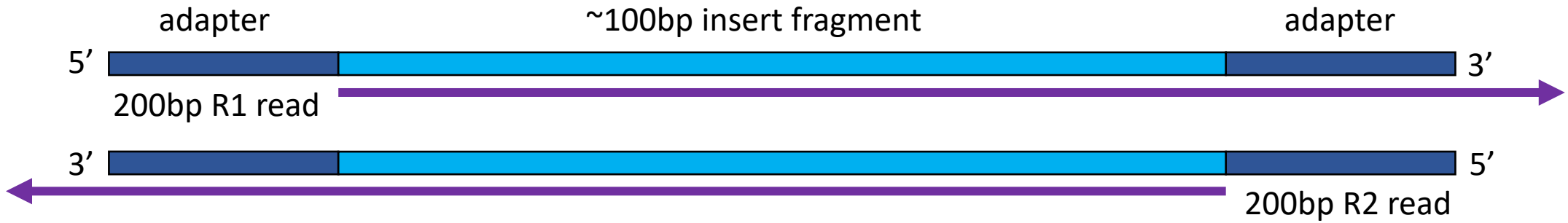
- Garbage in == Garbage out!
  - If your original sequences are junk, learn this as fast as possible, so you don't waste time on processing!
- General considerations:
  - Sequence quality distributions
  - Duplication rate
  - 3' adapter sequence trimming?
    - Can be important for RNA-seq (but mostly with shorter reads)
  - Contaminants?
    - Ribosomal RNA, cross-sample contamination (rare at Broad)
- Know your data:
  - Broad walkup processing pipelines (all hg19 oriented)

3’ adapter contamination

Condition A: reads shorter than insert length (no contamination)



Condition B: reads longer than insert length (contamination risk)



The presence of 3’ adapter information in the read is problematic because it can cause problems with genome alignment

FastQC: quality assurance tool for FASTQ

- Quality assurance tool for FASTQ sequences
- Accessible on prem as a dotkit
  - .**fastqc**-0.11.4 - .**fastqc**
- Input:
  - FASTQ files
  - Run on R1/R2 files
- Output
  - Directory with html and txt
    - Fastq_report.html
    - Fastqc_data.txt

Most useful FastQC reports

- Should I trim low-quality bases?
  - Per base sequence quality report
    - Based on all sequences
- Do I need to remove adapter sequences
  - Overrepresented sequences report
  - Based on first 100K sequences, trimmed to 75bp
- How complex is my library?
  - Sequence duplication levels report
    - Estimate based on first 100K sequences

FastQC resources

- FastQC website:
  - **http://www.bioinformatics.babraham.ac.uk**
- FastQC report documentation:
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/
- Good Illumina dataset:
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html
- Bad Illumina dataset:
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html
- Adapter contamination in RNA-seq example:
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/RNA-Seq_fastqc.html

Dealing with 3’ adapters

- Three main options:
  1. Hard trim all sequences by a specific amount
  2. Remove adapters specifically
  3. Perform a local alignment (vs global)

Hard trim all sequences by a specific amount

- E.g. trim 100 base reads to 50 bases
- ***Pro:***
  - Can eliminate vast majority of adapter contamination
  - Fast, easy to perform
  - Low quality 3’ bases also removed
- ***Con:***
  - Removes information you may want
    - e.g. splice junctions for RNAseq, coverage for mutation analysis
  - Not suitable for very short library fragments
    - e.g. miRNA libraries

Remove adapters specifically

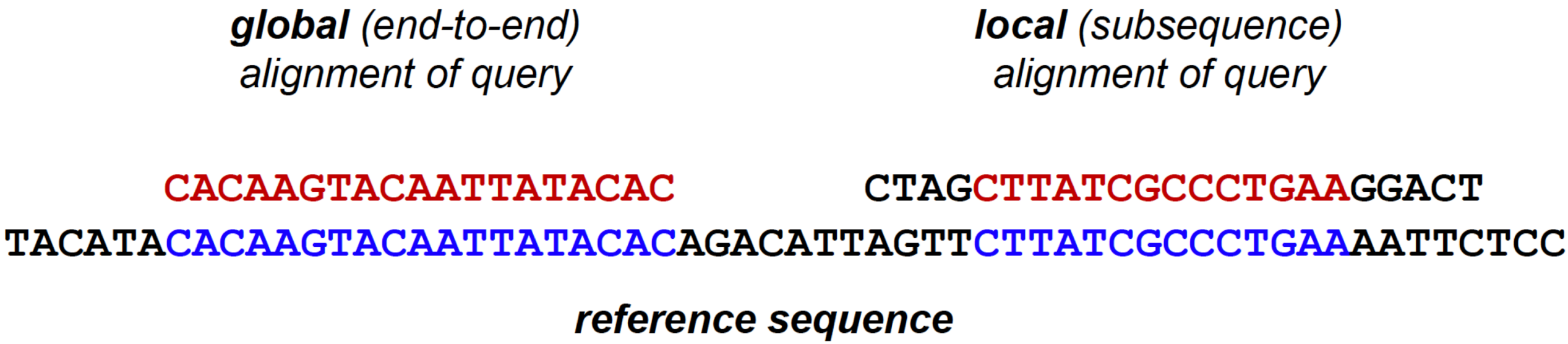
- ***Pro:***
  - Can eliminate vast majority of adapter contamination
  - Minimal loss of sequence information
    - still ambiguous: are 3’-most bases part of sequence or adapter?
- ***Con:***
  - Requires knowledge of insert fragment structure and adapters
  - Slower process; more complex to perform
  - Results in a heterogeneous pool of sequence lengths
    - can confuse some downstream tools (rare)
- Specific adapter trimming most common for RNA-seq
  - most transcriptome-aware aligners need adapter-trimmed reads

FASTQ trimming tools

- Tools: (none of these are on prem ☹)
  - **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)
    - typically, reads less than a specified length *after trimming* are discarded
    - leads to different sets of R1 and R2 reads unless care is taken
      - aligners do not like this!
- **cutadapt** has protocol for separating reads based on internal barcode

Perform a local alignment (vs global)

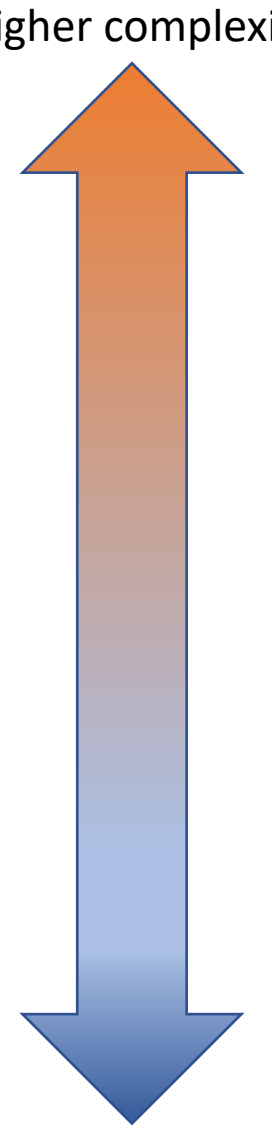
- ***Global*** alignment
  - requires query sequence to map ***fully*** (end-to-end) to reference
- ***Local*** alignment
  - allows a ***subset*** of the query sequence to map to reference
  - "untemplated" adapter sequences will be "soft clipped" (ignored)



Perform a local alignment

- ***Pro:***
  - mitigates adapter contamination while retaining full query sequence
  - minimal ambiguity
  - still ambiguous: are 3’-most bases part of sequence or adapter?
- ***Con:***
  - not supported by many aligners
  - e.g. not by the **tophat** or **hisat2** splice-aware aligners for RNAseq
  - slower alignment process
  - more complex post-alignment processing may be required
- Aligners with local alignment support:
  - **bwa mem**
  - **bowtie2 --local**

Library duplication is primarily a function of experiment type



Less enrichment for specific sequences

Genomic (WGS)

Methyl-seq

RNA-seq (no selection)

Exome (WXS)

RNA-seq (polyA selected)

ChIP-seq

Amplicons/testing panels

More enrichment for specific sequences

Higher diversity of molecules
Less sequence duplication expected
More sequencing depth required

More depth required for large genomes

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

Outline

1. NGS workflow and experiment types
2. Read sequence terminology
3. The fastq format
4. Fastq QC methods
5. Alignment to a genome
6. Reference genomes: making and using
7. Alignment metrics and QC
8. UCSC genome browser time!