

Introduction to NGS Analysis

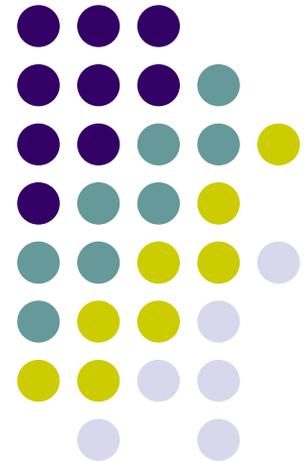
Anna Battenhouse

abattenhouse@utexas.edu

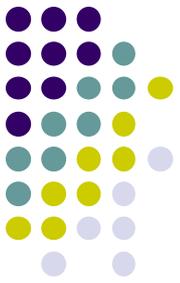
May, 2018

Associate Research Scientist
Center for Systems and Synthetic Biology (CSSB)
Ed Marcotte & Vishwanath Iyer labs

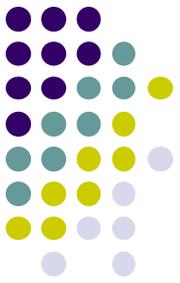
Center for Biomedical Research Support (CBRS)
Biomedical Research Computing Facility (BRCF)



Outline



1. NGS overview & terminology
2. The FASTQ format
3. Raw data QC and preparation
4. Alignment to a reference

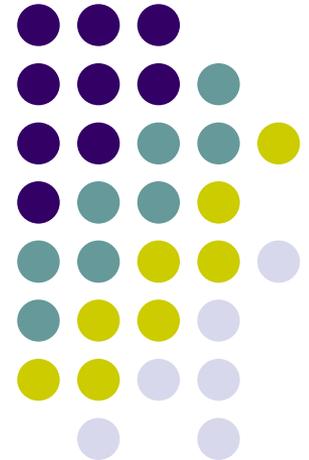


Other NGS Resources at UT

- CCBB short courses
 - 3-hour sessions
 - Intro to NGS, RNAseq, many others
- Genome Sequencing & Analysis Facility (GSAF)
 - Jessica Podnar, Associate Director, gsaf@utgsaf.org
- Bioinformatics consultants
 - Dennis Wylie, Dhivya Arasappan, Benni Goetz, Anna Battenhouse
- Byte Club (contact darasappan@gmail.com)
 - monthly presentations from bioinformatics community members
- BioTeam
 - <https://wikis.utexas.edu/display/bioiteam/>

Part 1: NGS Overview and Terminology

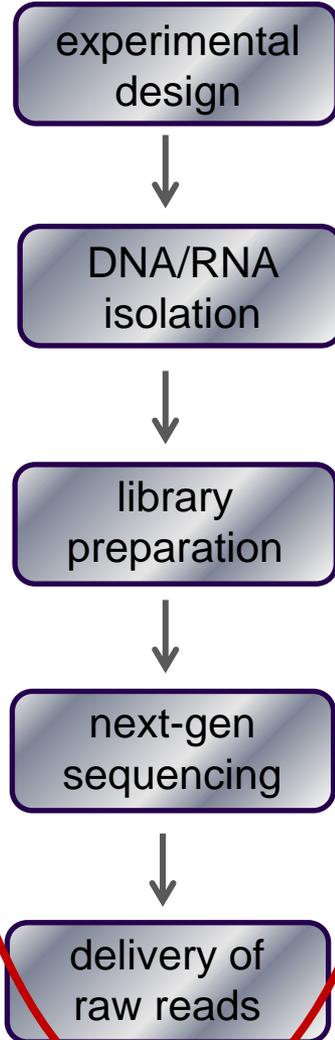
- NGS workflow overview
- Experiment types & library complexity
- Sequencing technology & terminology
- Sequence duplication issues



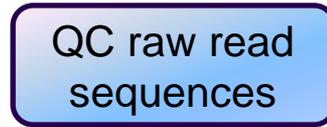
NGS Workflow

core processes

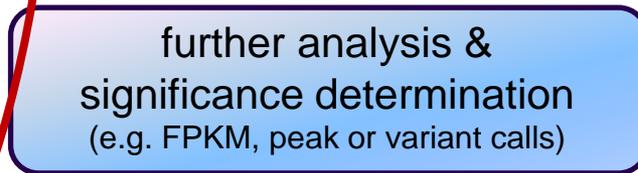
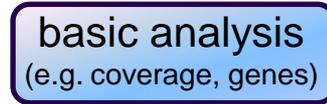
upstream processes



fastq



yes



has reference?

reference assembly

fasta

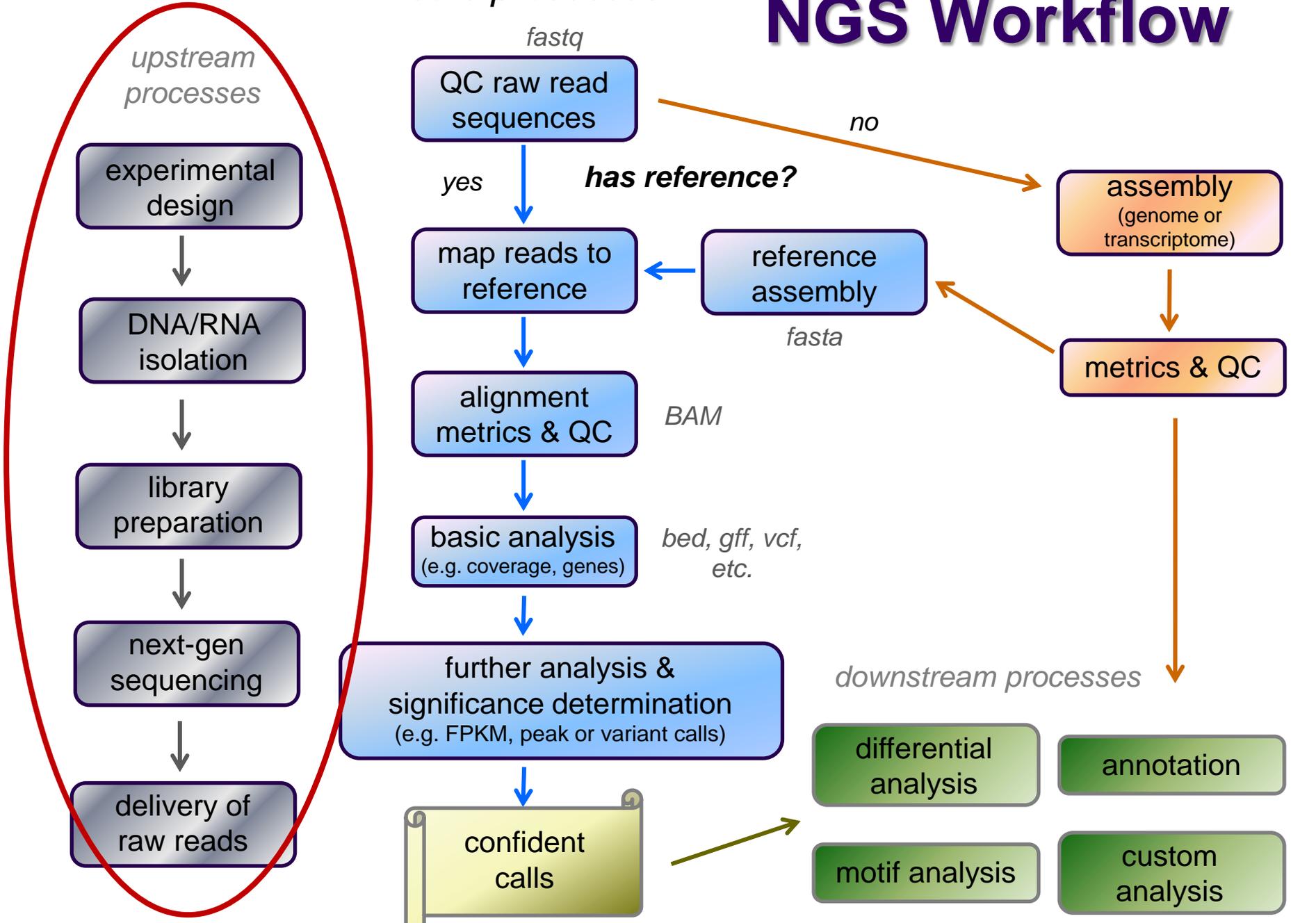
BAM

bed, gff, vcf, etc.

no



downstream processes



Popular Experiment Types



- **Whole Genome sequencing (WGS)**
 - **library:** all genomic DNA
 - **applications:** genome assembly, variant analysis
- **Exome sequencing (WXS)**
 - **library:** DNA from exonic regions (uses special kits)
 - **applications:** polymorphism/SNP detection; genotyping
- **RNA-seq**
 - **library:** extracted RNA converted to cDNA
 - all fragments → total or mRNA; small fragments → miRNA
 - **applications:** differential gene expression; isoform discovery
- **ChIP-seq**
 - **library:** DNA bound by transcription factor or histone proteins
 - or RNA bound by specific proteins (**RIP-seq**)
 - use target-specific antibodies to pull down after formaldehyde cross-linking
 - **applications:** analysis of regulatory networks, protein-bound RNAs



Library Complexity

Library complexity (diversity)

is a measure of the number of ***distinct molecular species*** in the library.

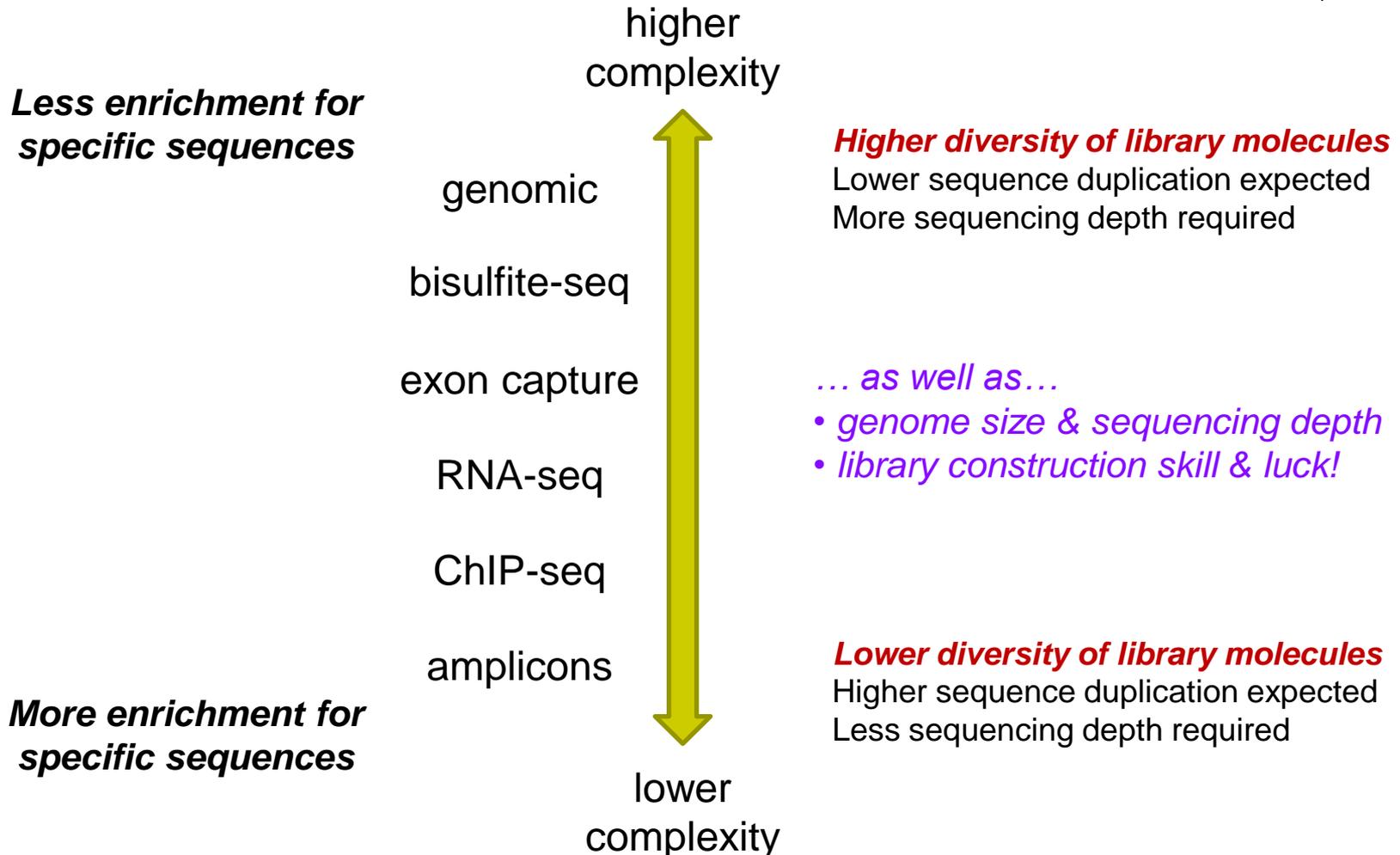
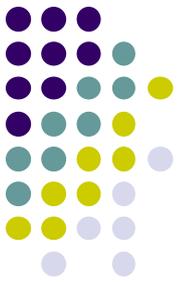
Many different molecules → high complexity

Few different molecules → low complexity

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

Type	Library prep	Applications	Complexity
Whole genome (WGS)	<ul style="list-style-type: none"> extract genomic DNA & fragment 	<ul style="list-style-type: none"> Genome assembly Variant detection, genotyping 	high
Bisulfite sequencing	<ul style="list-style-type: none"> bisulfite treatment converts C → U but not 5meC 	<ul style="list-style-type: none"> Methylation profiling (CpG) 	high
RAD-seq, ddRAD	<ul style="list-style-type: none"> restriction-enzyme digest DNA & fragment 	<ul style="list-style-type: none"> Variant detection (SNPs) Population genetics, QTL mapping 	high
Exome (WXS)	<ul style="list-style-type: none"> capture DNA from exons only (manufacturer kits) 	<ul style="list-style-type: none"> Variant detection, genotyping 	high-medium
ATAC-seq	<ul style="list-style-type: none"> high-activity transposase cuts DNA & ligates adapters 	<ul style="list-style-type: none"> Profile nucleosome-free regions (“open chromatin”) 	medium-high
RNA-seq, Tag-seq	<ul style="list-style-type: none"> extract RNA & fragment convert to cDNA 	<ul style="list-style-type: none"> Differential gene or isoform expression Transcriptome assembly 	medium, medium-low for Tag-seq
Transposon seq (Tn-seq)	<ul style="list-style-type: none"> create library of transposon-mutated genomic DNA amplify mutants via Tn-PCR 	<ul style="list-style-type: none"> Characterize genotype/phenotype relationships w/high sensitivity 	medium
ChIP-seq	<ul style="list-style-type: none"> cross-link proteins to DNA pull-down proteins of interest w/ specific antibody, reverse cross-links 	<ul style="list-style-type: none"> Genome-wide binding profiles of transcription factors, epigenetic marks & other proteins 	medium (but variable)
GRO-seq	<ul style="list-style-type: none"> isolate actively-transcribed RNA 	<ul style="list-style-type: none"> Characterize transcriptional dynamics 	medium-low
RIP-seq	<ul style="list-style-type: none"> like ChIP-seq, but with RNA 	<ul style="list-style-type: none"> Characterize protein-bound RNAs 	low
miRNA-seq	<ul style="list-style-type: none"> isolate 15-25bp RNA band 	<ul style="list-style-type: none"> miRNA profiling 	low
Amplicons	<ul style="list-style-type: none"> amplify 1-1000+ genes/regions 	<ul style="list-style-type: none"> genotyping, metagenomics, mutagenesis 	low

Library complexity is primarily a function of experiment type



Sequencing technologies



- Illumina (Solexa) now dominant for “short” (<300 bp) reads
 - [Short Illumina video](https://tinyurl.com/hvnmwjb) (<https://tinyurl.com/hvnmwjb>)
 - 2 PCR amplifications performed:
 1. during library preparation
 2. during “cluster generation”
 - ***amplification always introduces bias!***
- Newer “single molecule” sequencing
 - sequences single molecules, not clusters
 - one amplification (during library prep) usually still required
 - allows for longer reads (multi-kilobase!)
 - no signal wash-out due to lack of synchronization among cluster molecules
 - ***but:*** reads have high error rate (~10%)
 - [PacBio SMRT system](http://www.pacb.com/smrt-science/smrt-sequencing/) (<http://www.pacb.com/smrt-science/smrt-sequencing/>)
 - [PCR-free protocol](#) (limited applications)
 - [Oxford Nanopore](https://nanoporetech.com/) (<https://nanoporetech.com/>)

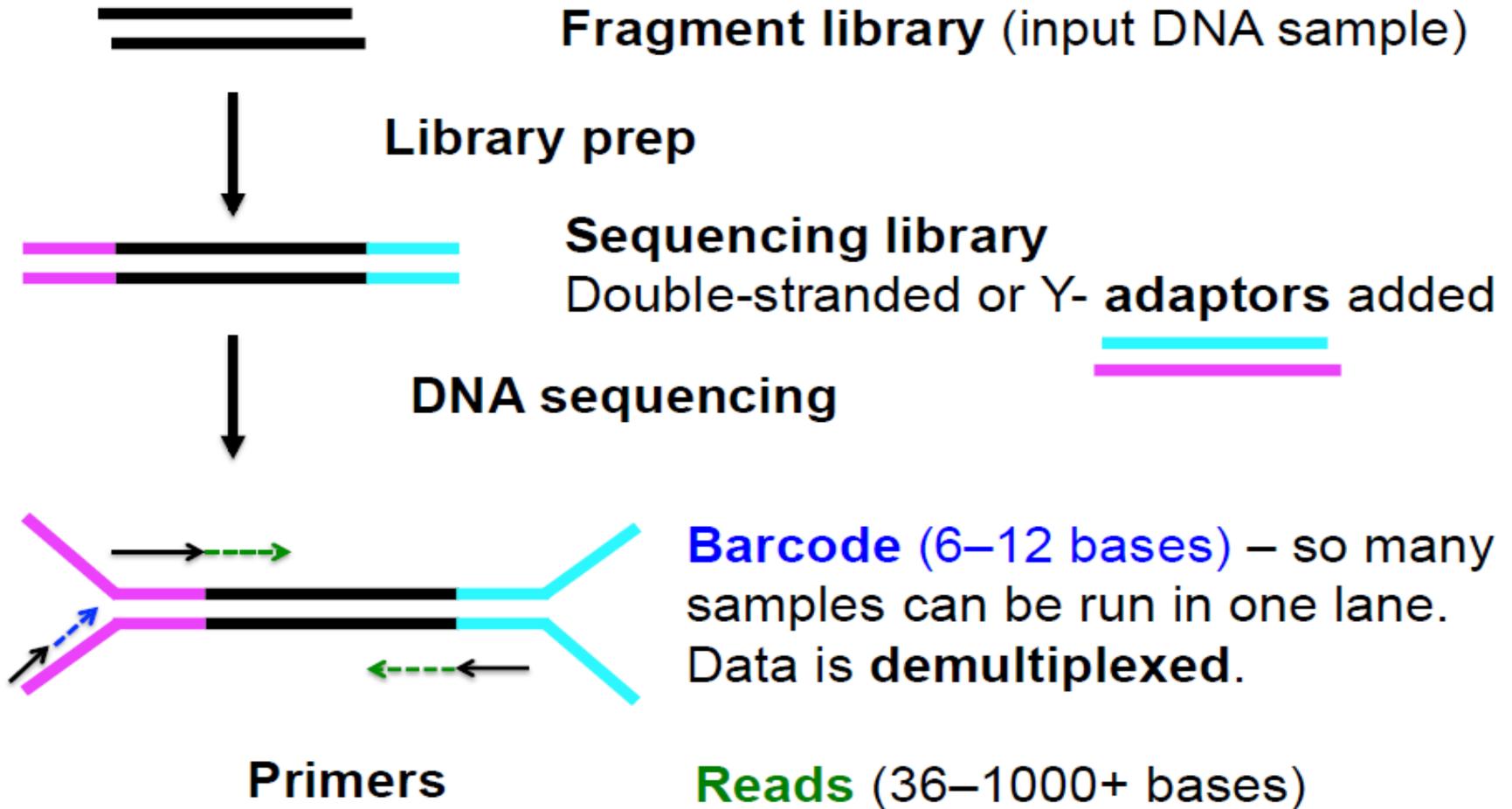
Illumina sequencer models (GSAF)



Model	Lanes	Typical reads per lane	Read lengths	Recommended applications
HiSeq 4000	8	240 M	50, 75, 150	WGS, WXS. RNA-seq, <i>GBS (Genotyping by Sequencing)</i>
HiSeq 2500	8	200 M	36, 50, 75, 100, 125 (150, 250 rapid run)	WGS, WXS, RNA-seq, GBS, targeted sequencing
NextSeq	4 (but all 4 get same DNA)	330 M	75, 150	WGS, WXS, RNA-seq, GBS, targeted sequencing
MiSeq	1	12 – 22 M (v2 vs v3 chemistry)	v2: 25, 36, 150, 250 v3: 75, 300	Amplicons, metagenomics, WGS for tiny genomes, RNA-seq for small transcriptomes

- When less than a full lane is needed, multiple samples with different *barcodes* (a.k.a. *indexes*) can be run on the same lane.
- Most current library prep kits on the market are limited to 96 barcodes.
- Custom indexing can allow for a mix of 1000's of samples (e.g. over 600+ indexes for 16S metagenomic preps).

Read sequence terminology



- Adapter areas include primers, barcode
 - sequencing facility will have more information

<https://wikis.utexas.edu/display/GSAF/Illumina+--+all+flavors>

Types of Illumina sequencing



single-end



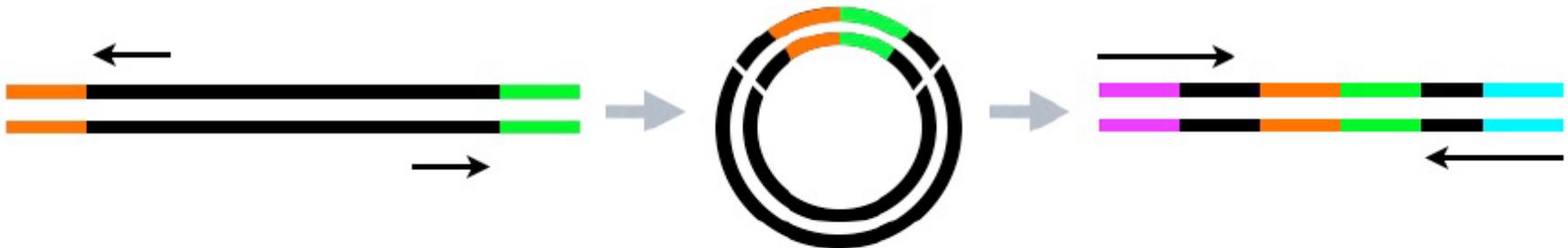
independent reads

paired-end



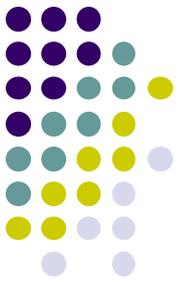
two inwardly oriented reads separated by ~200 nt

mate-paired

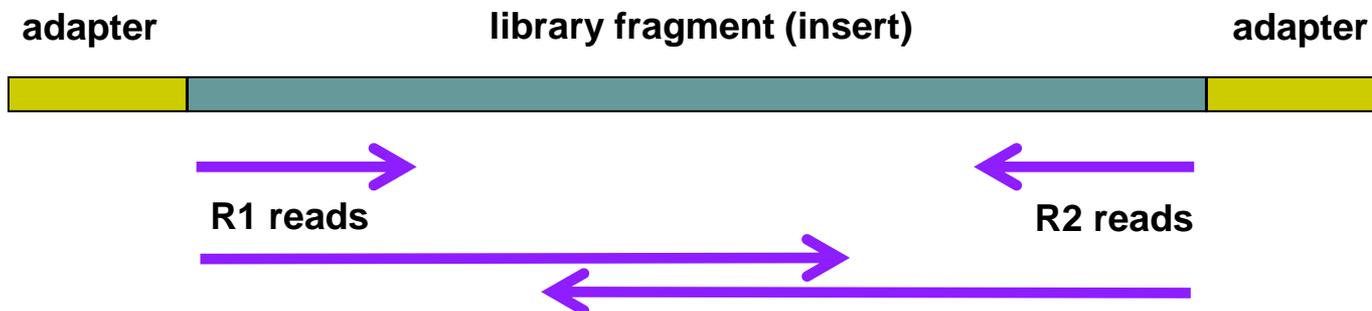


two outwardly oriented reads separated by ~3000 nt

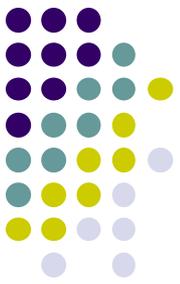
Reads and Fragments



- 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
 - also called **tags**
 - an R1 and its associated R2 form a **read pair**
 - a readout of part (or all) of the fragment molecule
- There is considerable confusion of terminology in this area!
 - Be sure you request depth in **read pairs** for paired-end sequencing



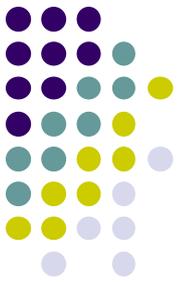
Single end vs Paired end



- **single end** (SE) reads are less expensive
 - **but** SE runs also less frequent
- **paired end** (PE) reads can be mapped more reliably
 - especially against lower complexity genomic regions
 - an unmapped read can be “rescued” if its mate maps well
 - they provide more bases around a locus
 - e.g. for analysis of polymorphisms
 - actual fragment sizes can be easily determined
 - from the alignment records for each dual-mapping “proper pair”
 - also help distinguish the true complexity of a library
 - by clarifying which *fragments* are duplicates (vs *read* 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 low-complexity, higher duplication experiments

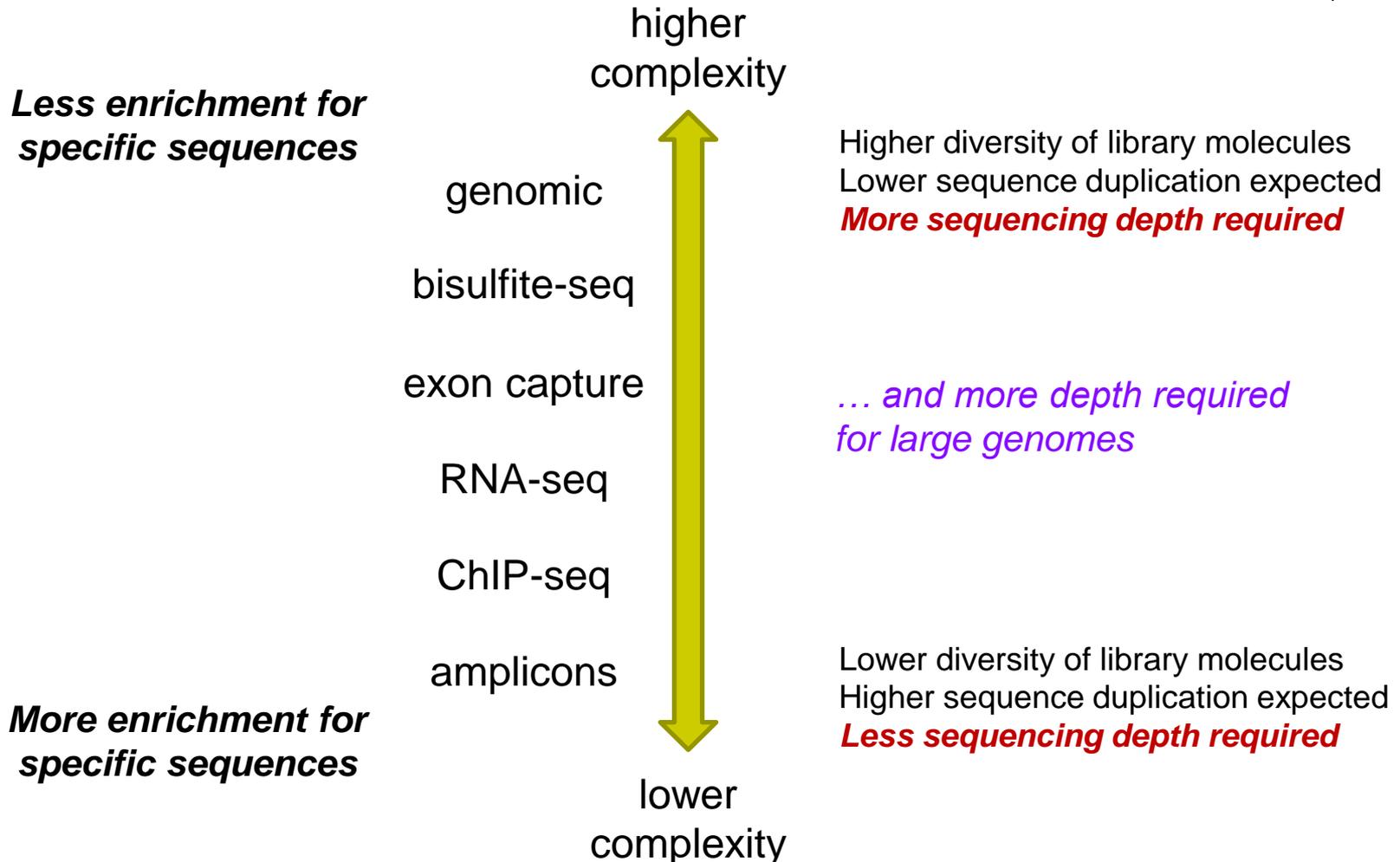


Sequencing depth



- How much sequencing depth is needed?
 - No single answer!
- Depends on:
 - genome size
 - prokaryotes – a few Kilobases
 - lower eukaryotes – some number of Megabases
 - higher eukaryotes – Gigabases
 - library fragment enrichment
 - e.g. ChIP-seq or RIP-seq
 - theoretical library complexity
 - genomic re-sequencing **vs** amplicon sequencing
 - desired sensitivity
 - e.g. looking for rare mutations

Sequencing depth required is a function of experiment type & genome size



Sequence Duplication

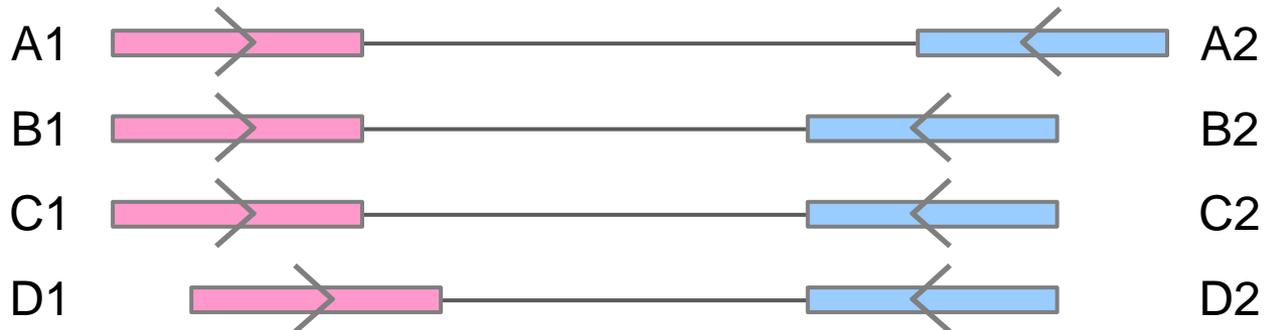


- The set of sequences you receive can contain exact duplicates
- Duplication can arise from:
 1. sequencing of species enriched in your library (**biological – good!**)
 - each read comes from a different cluster (DNA molecule)
 2. sequencing of artifacts (**technical – bad!**)
 - differentially amplified PCR species (PCR duplicates)
 - recall that 2 PCR amplifications are performed w/Illumina sequencing
 - optical duplicates, when two flowcell clusters overlap
- **cannot tell which using “standard” sequencing methods!**
- Current best practice is to “mark duplicates” during initial processing
 - then decide what to do with them later...
 - e.g. retain (all), remove (none), dose (some)
- Different experiment types have different expected duplication
 - whole genome/exome → high complexity & low duplication
 - amplicon sequencing → low complexity & high duplication

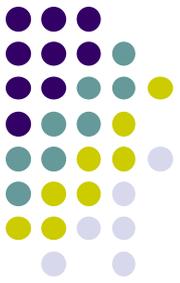


Read vs Fragment duplication

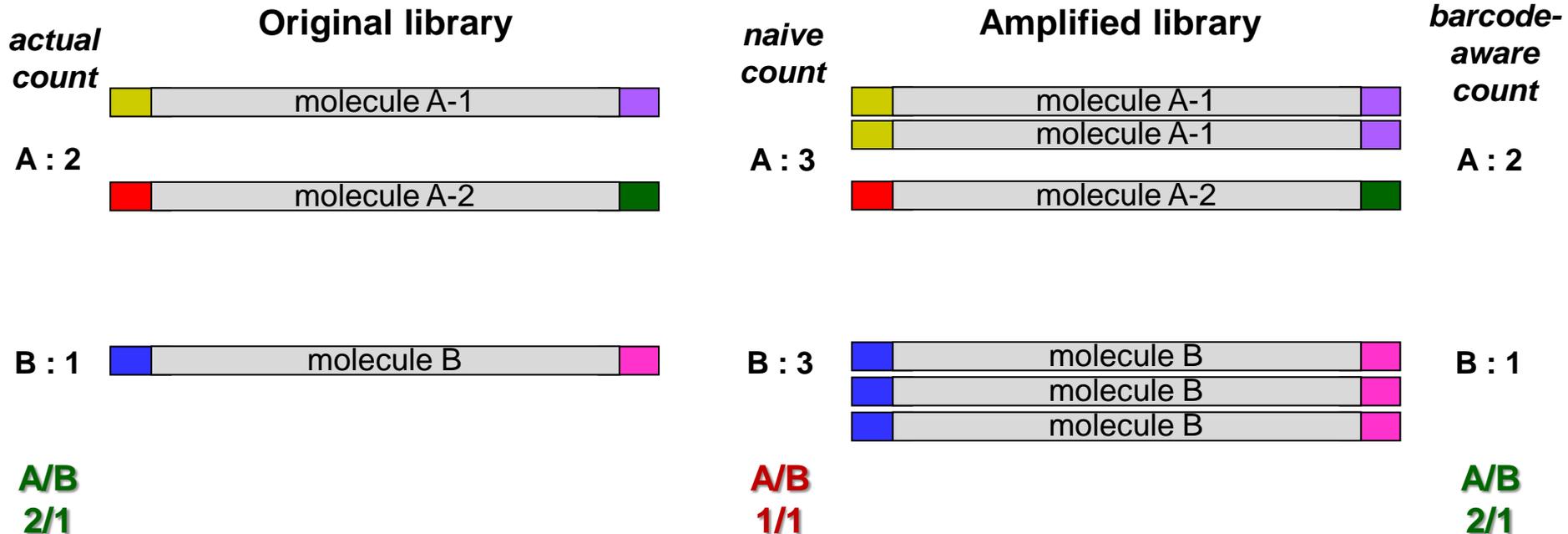
- Consider the 4 fragments below
 - 4 R1 reads (pink), 4 R2 reads (blue)
- Duplication when only 1 end considered
 - A1, B1, C1 have identical sequences, D1 different
 - 2 unique + 2 duplicates = 50% duplication rate
 - B2, C2, D2 have identical sequences, A2 different
 - 2 unique + 2 duplicates = 50%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



Molecular Barcoding



- Resolves ambiguity between biological and technical (PCR amplification) duplicates
 - adds secondary barcodes to pre-PCR molecules
 - combination of barcodes + insert sequence can provide accurate quantification
 - but requires specialized pre- and post-processing



Single Cell sequencing



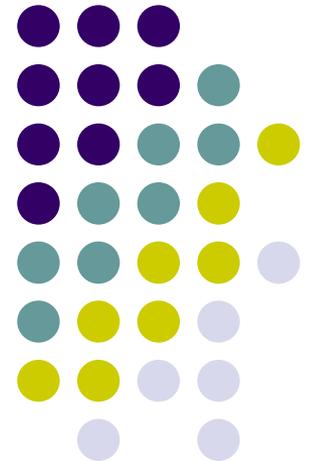
- Standard sequencing library starts with millions of cells
 - will be in different states unless synchronized
 - a heterogenous “ensemble” with (possibly) high cell-to-cell variability
- Single cell sequencing technologies aim to capture this variability
 - examples:
 - cells in different layers/regions of somatic tissue
 - cells in 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 sequencing
 5. custom downstream analysis links results to their cell (barcode) of origin

Some barcode (index) types



- **Library barcode**
 - the same for all fragments in a library
 - multiple barcoded samples can be pooled on one sequencer lane
 - on the order of dozens available (part of standard library prep kits)
- **Molecular barcodes**
 - different, small barcodes (or pairs) attached to library fragments **before amplification**
 - available diversity depends on barcode size and number, e.g.:
 - 4 well-separated bases → ~80
 - 2 x 4 well-separated bases → ~700
 - 2 x 8 well-separated bases → ~500,000
 - finding well-separated, sequencing-compatible barcodes is not trivial!
 - see also the Common Birthday Probability problem
 - 50% shared birthday probability with only 23 people (https://en.wikipedia.org/wiki/Birthday_problem)
- **Single cell barcode**
 - unique barcode attached to all cDNA molecules in each single cell
 - number of barcodes needed depends on # of single cells desired

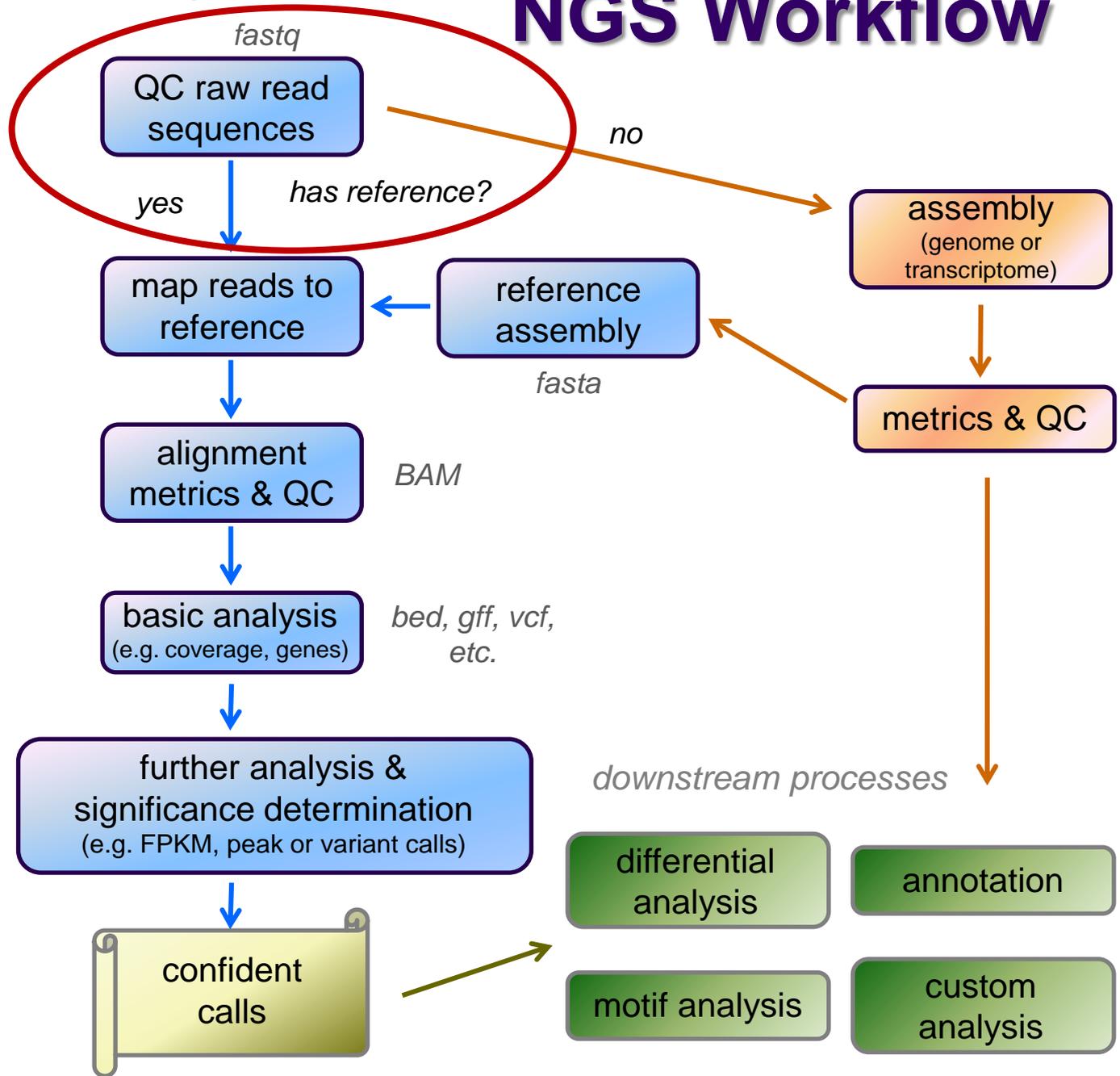
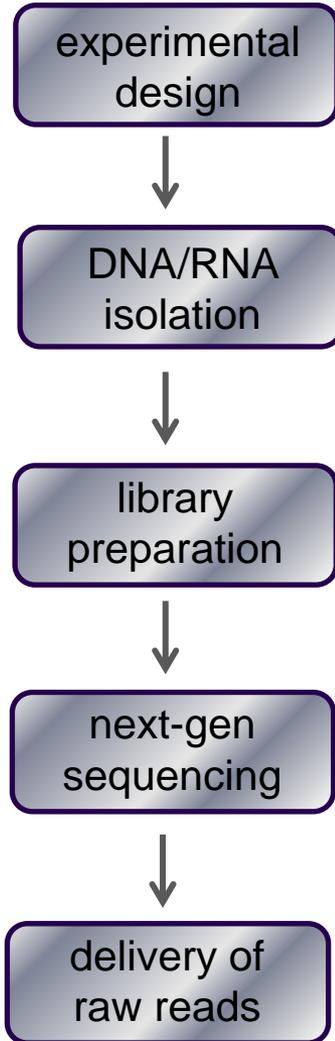
Part 2: The FASTQ format



NGS Workflow

core processes

upstream processes



FASTQ files



- Nearly all sequencing data delivered as **FASTQ** files
 - FASTQ = FASTA sequences + Quality scores
 - file names have **.fastq** or **.fq** extensions
 - usually compressed to save space
 - (**gzip**'d, with **.gz** file extension)
 - best practice: leave them that way!
 - 3x to 6x space saving
 - most tools handle **gzip**'d FASTQ
- Paired-end sequencing data comes in 2 FASTQs
 - one each for R1 and R2 reads, same number of rows
`Sample_MyTubeID_L008_R1.fastq.gz`
`Sample_MyTubeID_L008_R2.fastq.gz`
 - ***order of reads is identical***
 - aligners rely on this “name ordering” for PE

FASTQ format



- Text format for storing sequence and quality data
 - http://en.wikipedia.org/wiki/FASTQ_format
- 4 lines per sequence:
 1. **@read name**
 2. **called base sequence (ACGTN)**
always 5' to 3'; usually excludes 5' adapter
 3. **+optional read name**
 4. **base quality scores encoded as text characters**
- FASTQ representation of a single, 50 base R1 sequence

```
@HWI-ST1097:97:D0WW0ACXX:8:1101:2007:2085 1:N:0:ACTTGA  
ATTCTCCAAGATTTGGCAAATGATGAGTACAATTATATGCCCAATTTACA  
+  
?@@?DD;?;FF?HHBB+:ABECGHDHDCF4?FGIGACFDFH;FHEIIB9?
```

FASTQ read names



- Illumina fastq read names encode information about the source cluster
 - unique identifier (“fragment name”) begins with @, then:
 - 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
 - R1, R2 reads *have the same fragment name*

@HWI-ST1097:97:D0WW0ACXX:8:1101:2007:2085 1:N:0:ACTTGA

@HWI-ST1097:97:D0WW0ACXX:8:1101:2007:2085 2:N:0:ACTTGA

Multiple lanes

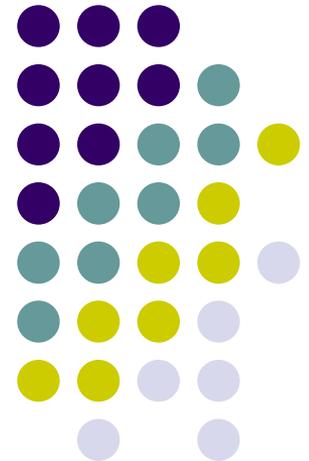


- Sometimes the sequencing facility splits your sample across lanes
 - one submitted sample may be delivered as multiple FASTQ files
 - Lane1: `Sample_MyTubeID_L001_R1.fastq.gz`, `Sample_MyTubeID_L001_R2.fastq.gz`
 - Lane8: `Sample_MyTubeID_L008_R1.fastq.gz`, `Sample_MyTubeID_L008_R2.fastq.gz`
- Your sample may be re-run to “top off” requested read depth
 - be careful with the file names!
 - if run in the same lane, the FASTQ file names will be the same
 - 1st run: `Sample_MyTubeID_L003_R1.fastq.gz`
 - 2nd run : `Sample_MyTubeID_L003_R1.fastq.gz`
- Best practice
 - keep original data in separate directories by date & project
 - process data from multiple lanes separately for as long as possible
 - e.g., through alignment, then merge the sorted, duplicate-marked BAMs
 - identical sequences from different lanes can be considered unique (non-duplicates)

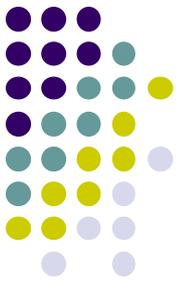
Part 3:

Data QC & preparation

- QC of raw sequences with **FastQC** tool
- Dealing with adapters



Raw sequence quality control



- Critical step! Garbage in = Garbage out
 - general sequence quality
 - base quality distributions
 - sequence duplication rate
 - trim 3' adapter sequences?
 - important for RNAseq
 - trim 3' bases with poor quality?
 - important for *de novo* assembly
 - other contaminants?
 - biological – rRNA in RNAseq
 - technical – samples sequenced w/other barcodes
- Know your data
 - sequencing center pre-processing
 - 5' adapter removed? QC-failed reads filtered?
 - PE reads? relative orientations?
 - technology specific issues?
 - e.g. PAR clip should produce C→T transitions





3' Adapter contamination

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

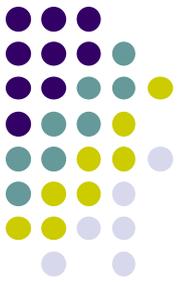


B. Reads long compared to library fragment (3' adapter contamination)



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

FastQC



- Quality Assurance tool for FASTQ sequences
- Can run as interactive tool or command line
- Input:
 - FASTQ file(s)
 - run on both R1, R2 files
- Output:
 - directory with html & text reports
 - `fastqc_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 *1st 100,000* sequences, trimmed to 75bp
- How complex is my library?
 - *Sequence duplication levels Report*
 - estimate based on *1st 100,000* sequences



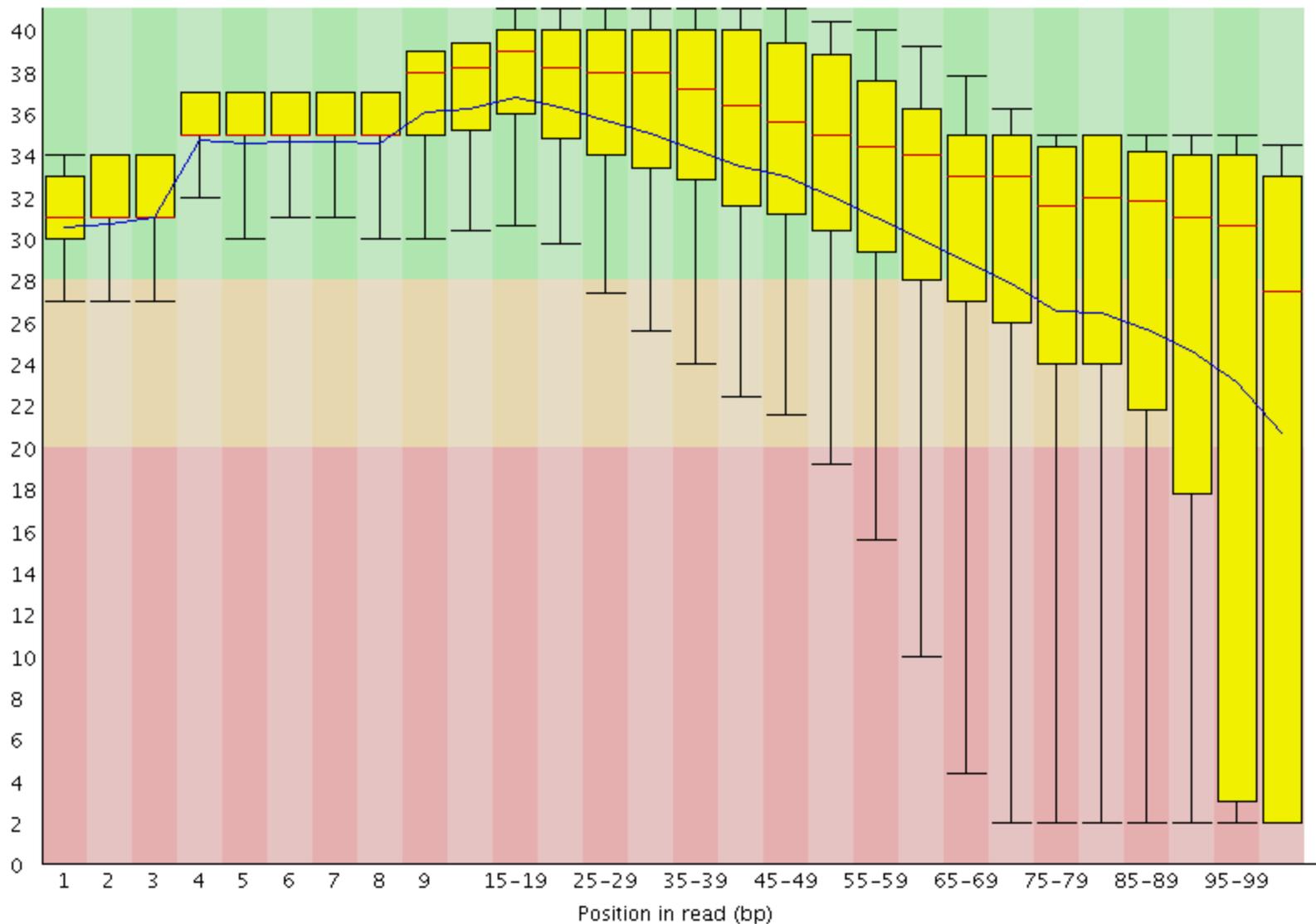
FastQC resources

- 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/fastqc_report.html
- Bad Illumina dataset:
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc/fastqc_report.html
- Real Yeast ChIP-seq dataset:
http://web.corral.tacc.utexas.edu/BiolTeam/yeast_stuff/Sample_Yeast_L005_R1.cat_fastqc/fastqc_report.html

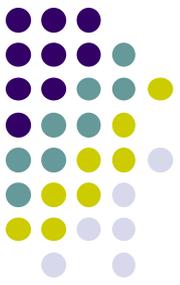
FastQC Per-base sequence quality report



Quality scores across all bases (Sanger / Illumina 1.9 encoding)



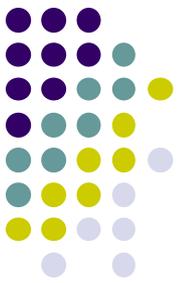
FastQC Overrepresented sequences report



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

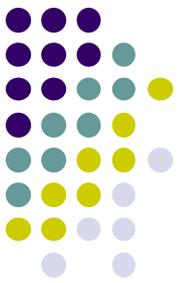
Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATG	60030	5.01369306977828	TruSeq Adapter, Index 1 (97% over 37bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATGC	42955	3.5875926338884896	TruSeq Adapter, Index 1 (97% over 37bp)
CACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATGCCGTCTTCTGCT	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)

Overrepresented sequences



- Here < 1% of sequences contain adapters
 - trimming optional

Sequence	Count	Percentage	Possible Source
AACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAAGAACGTA	102020	1.0707851766890004	No Hit
AATTCTAGAGCTAATACGTGCAACAAACCCCGACTTATGGAAGGGACGCA	89437	0.9387160737848865	No Hit
AAAGGATTGGCTCTGAGGGCTGGGCTCGGGGGTCCCAGTTCGAAACCCGT	89427	0.9386111154260659	No Hit
TACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCC	87604	0.9194772066130483	No Hit
ATTGGCTCTGAGGGCTGGGCTCGGGGGTCCCAGTTCGAAACCCGTCTGGCT	65829	0.6909303802809273	No Hit
TCTAGAGCTAATACGTGCAACAAACCCCGACTTATGGAAGGGACGCATT	65212	0.6844544495416888	No Hit
TAAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAAGAAC	61582	0.646354565289767	No Hit
CTCGGATAACCGTAGTAATTCTAGAGCTAATACGTGCAACAAACCCCGAC	59180	0.6211435675010296	No Hit
ATGGATCCGTAACCTTCGGGAAAAGGATTGGCTCTGAGGGCTGGGCTCGGG	56982	0.598073720232235	No Hit
AAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAAGAACG	54813	0.5753082522040206	No Hit
CTAGAGCTAATACGTGCAACAAACCCCGACTTATGGAAGGGACGCATT	40019	0.4200328561646452	No Hit
AGAACTCCGCAGTTAAGCGTGCTTGGGCGAGAGTAGTACTAGGATGGGTG	39753	0.4172409638200141	No Hit
ACTCGGATAACCGTAGTAATTCTAGAGCTAATACGTGCAACAAACCCCGA	38867	0.4079416532284981	No Hit
ACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAAGAACGTAG	38438	0.40343893963508914	No Hit
ACTTCGGGAAAAGGATTGGCTCTGAGGGCTGGGCTCGGGGGTCCCAGTTC	37406	0.3926072370047907	No Hit
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATG	34199	0.35894709133098535	TruSeq Adapter, Index 4 (100% over 49bp)
GAACCTTGGGATGGGTCTGGCCGGTCCGCCTTTGGTGTGCATTGGTCTGGCT	34099	0.3578975077427782	No Hit



Overrepresented sequences

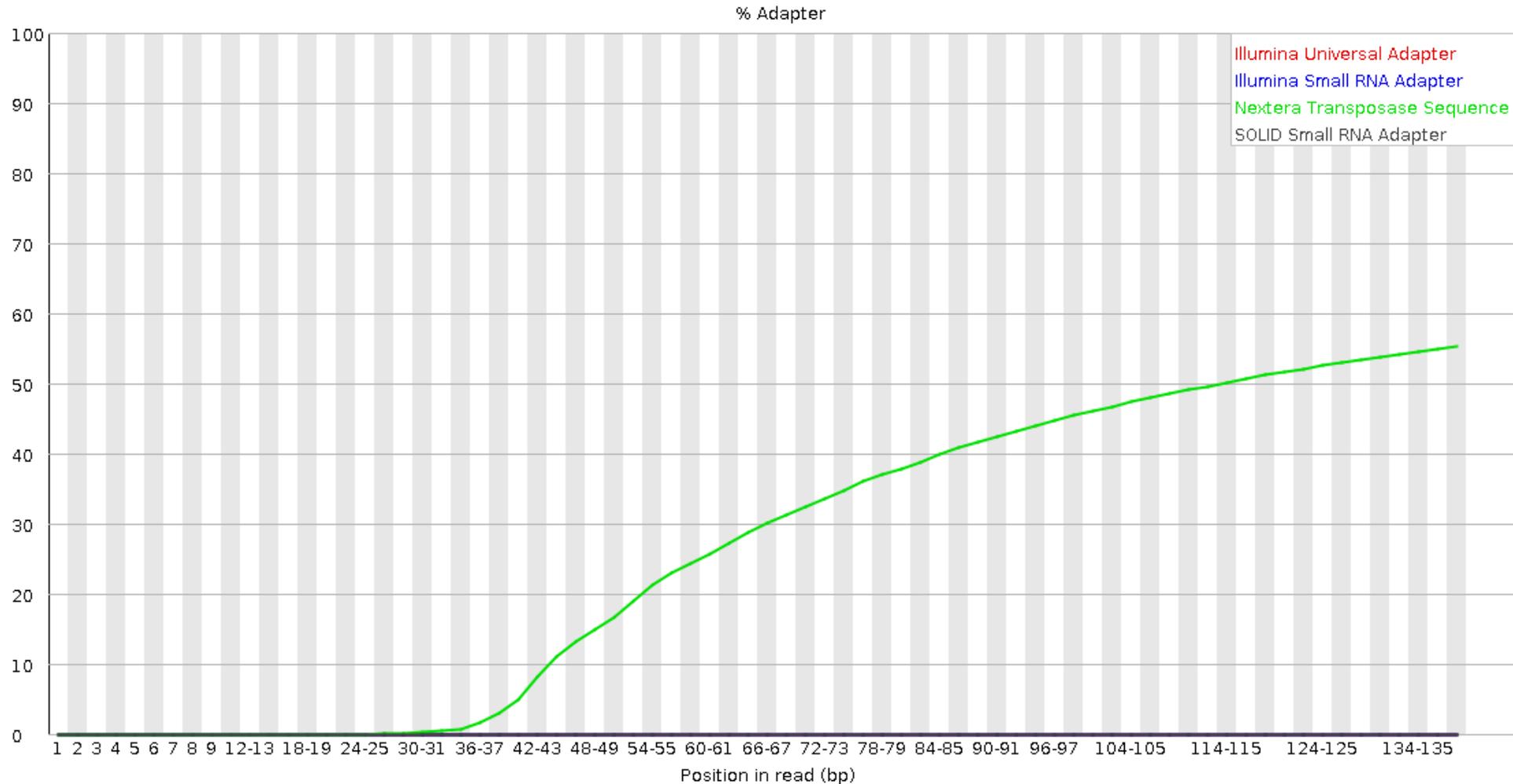
- Here nearly 1/3 of sequences some type of non-adapter contamination
 - **BLAST** the sequence to identify it

Sequence	Count	Percentage	Possible Source
GAAGGTCACGGCGAGACGAGCCGTTTATCATTACGATAGGTGTCAAGTGG	5632816	32.03026785752871	No Hit
TATTCTGGTGTCTTAGGCGTAGAGGAACAACACCAATCCATCCCGAACTT	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
AGAACATGAAACCCTAAGCTCCCAAGCAGTGGGAGGAGCCCTGGGCTCTG	111584	0.6345077504066322	No Hit
AAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAAGAACG	111255	0.6326369351474214	No Hit
ATTACGATAGGTGTCAAGTGGAAAGTGCAGTGATGTATGCAGCTGAGGCAT	73682	0.41898300890326096	No Hit
GAAGGTCACGGCGAGACGAGCCGTTTATCATTACGATAGGTGTCAAGGGG	71661	0.4074908580252516	No Hit
GGATGCGATCATACCAGCACTAATGCACCGGATCCCATCAGAACTCCGCA	69548	0.3954755612388914	No Hit
ATATTCTGGTGTCTTAGGCGTAGAGGAACAACACCAATCCATCCCGAACT	54017	0.30716057099328803	No Hit

Adapter Content report



- Newer versions of FastQC have a separate Adapter Content report
 - provides a per-base % adapter trace (Transposon-seq below)

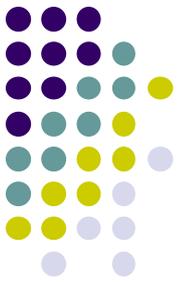


Dealing with 3' adapters



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

Hard trim by specific length



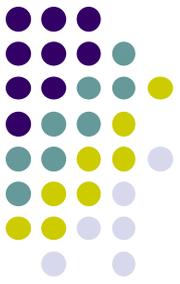
- E.g. trim 100 base reads to 50 bases
- **Pro:**
 - Can eliminate vast majority of adapter contamination
 - Fast, easy to perform
 - Low quality 3' bases also removed
- **Con:**
 - Removes information you may want
 - e.g. splice junctions for RNAseq, coverage for mutation analysis
 - Not suitable for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

Trim adapters specifically



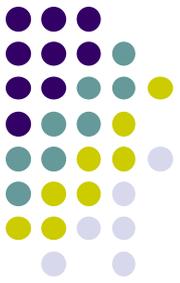
- **Pro:**
 - Can eliminate vast majority of adapter contamination
 - Minimal loss of sequence information
 - still ambiguous: are 3'-most bases part of sequence or adapter?
- **Con:**
 - Requires knowledge of insert fragment structure and adapters
 - Slower process; more complex to perform
 - Results 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:
 - **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

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” adapter sequences will be “soft clipped” (ignored)

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

local (*subsequence*)
alignment of query

CACAAGTACAATTATACAC

CTAGCTTATCGCCCTGAAGGACT

TACATA**CACAAGTACAATTATACAC**AGACATTAGTT**CTTATCGCCCTGAA**AATTCTCC

reference sequence



Perform 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**

FastQC Sequence duplication report

Yeast ChIP-seq

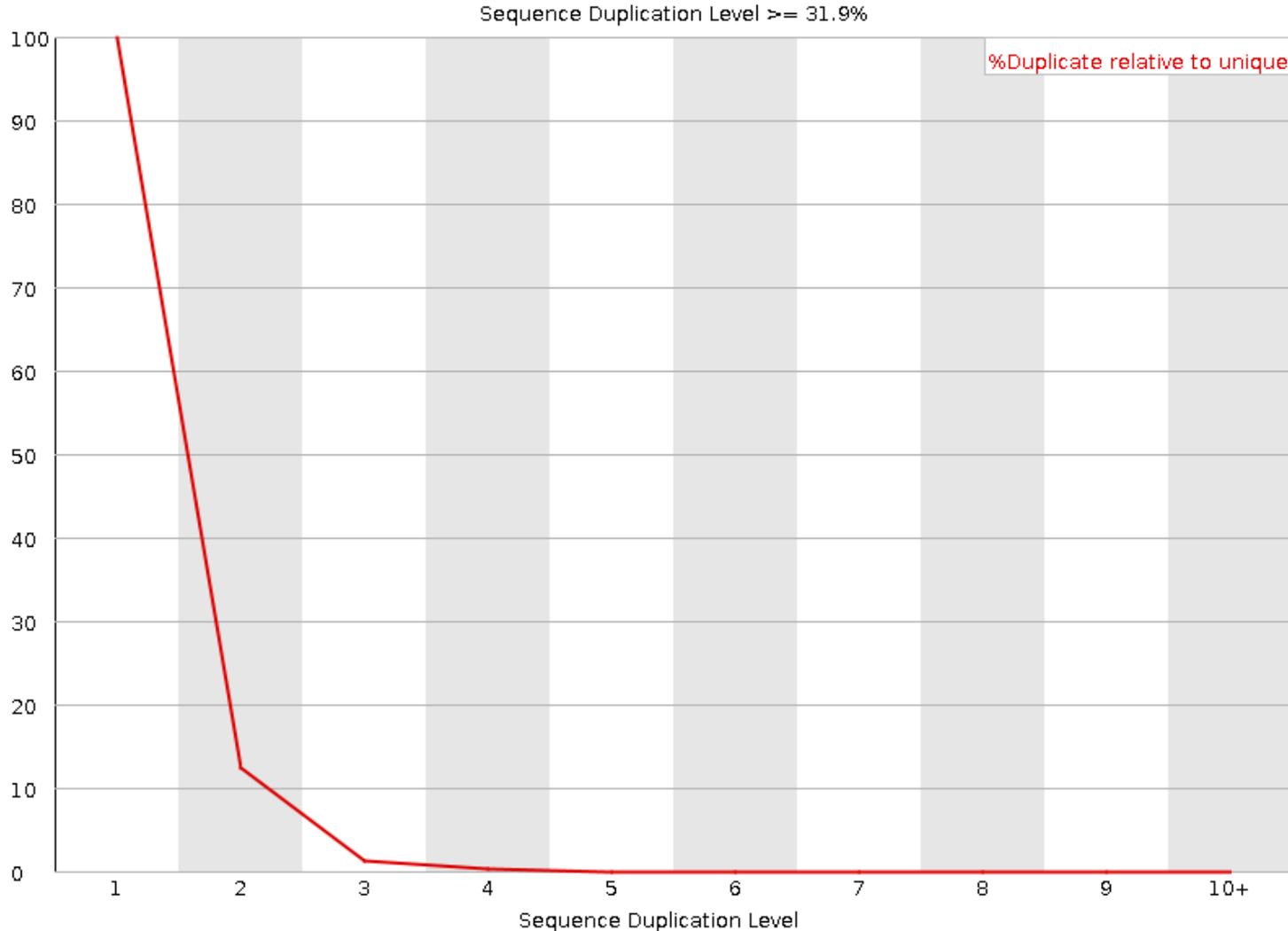


For every 100 unique sequences there are:

~12 sequences w/2 copies

~1-2 with 3 copies

Ok – Some duplication expected due to IP enrichment



Sequence duplication report

Yeast ChIP-exo

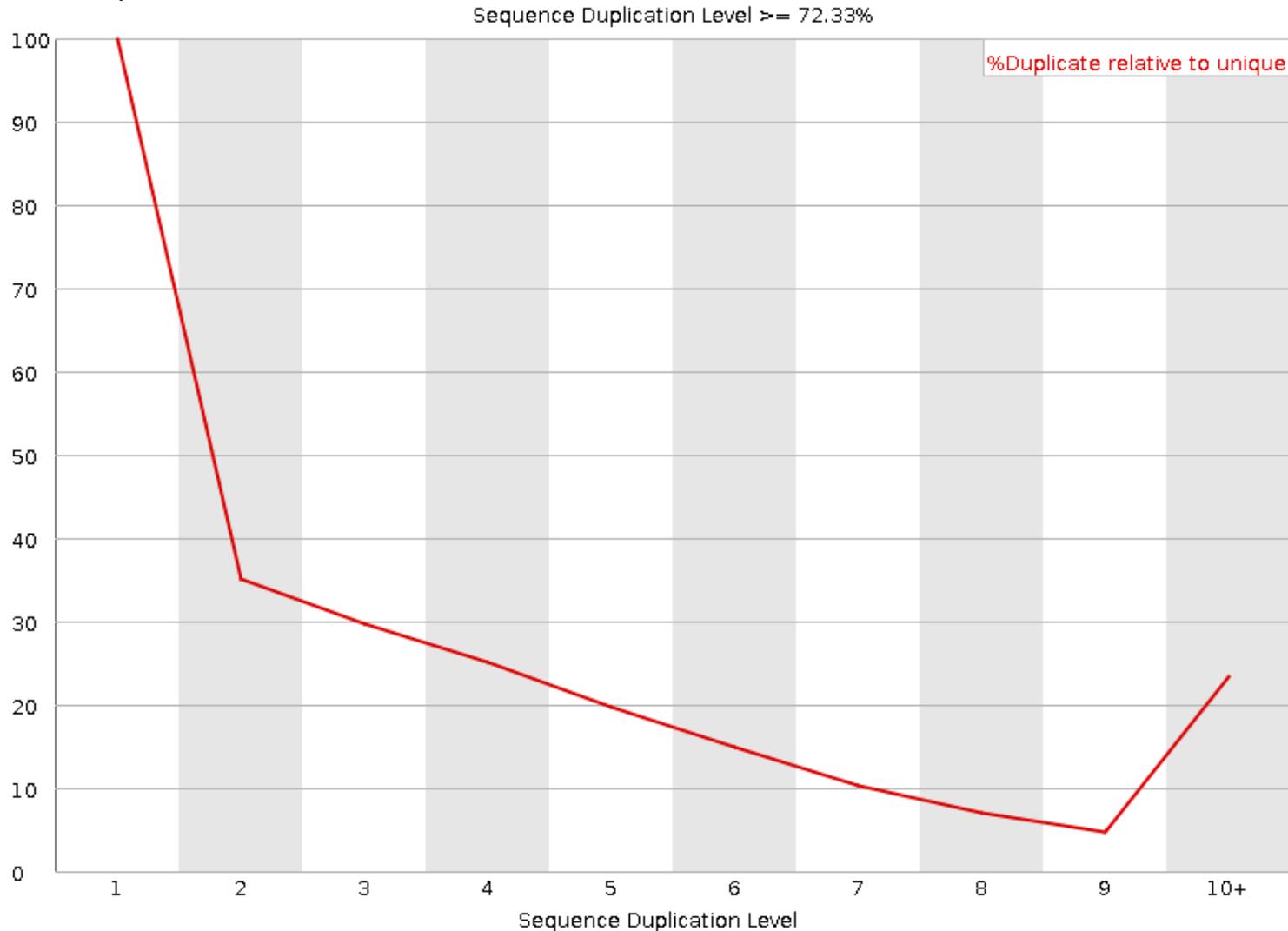


For every 100 unique sequences there are:

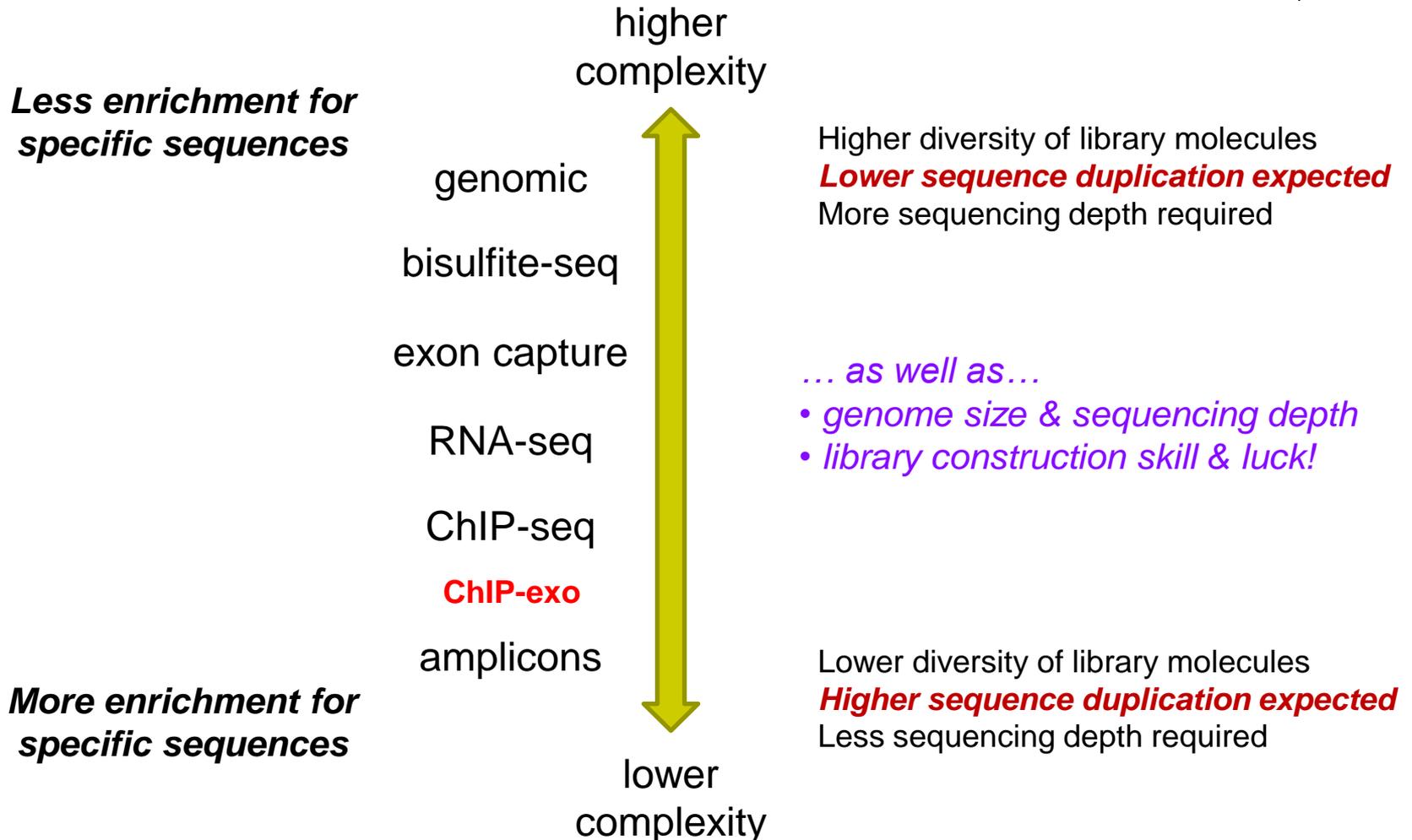
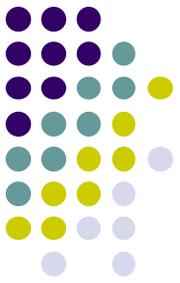
~35 sequences w/2 copies

~22 with 10+ copies

Success! Protocol expected to have high duplication

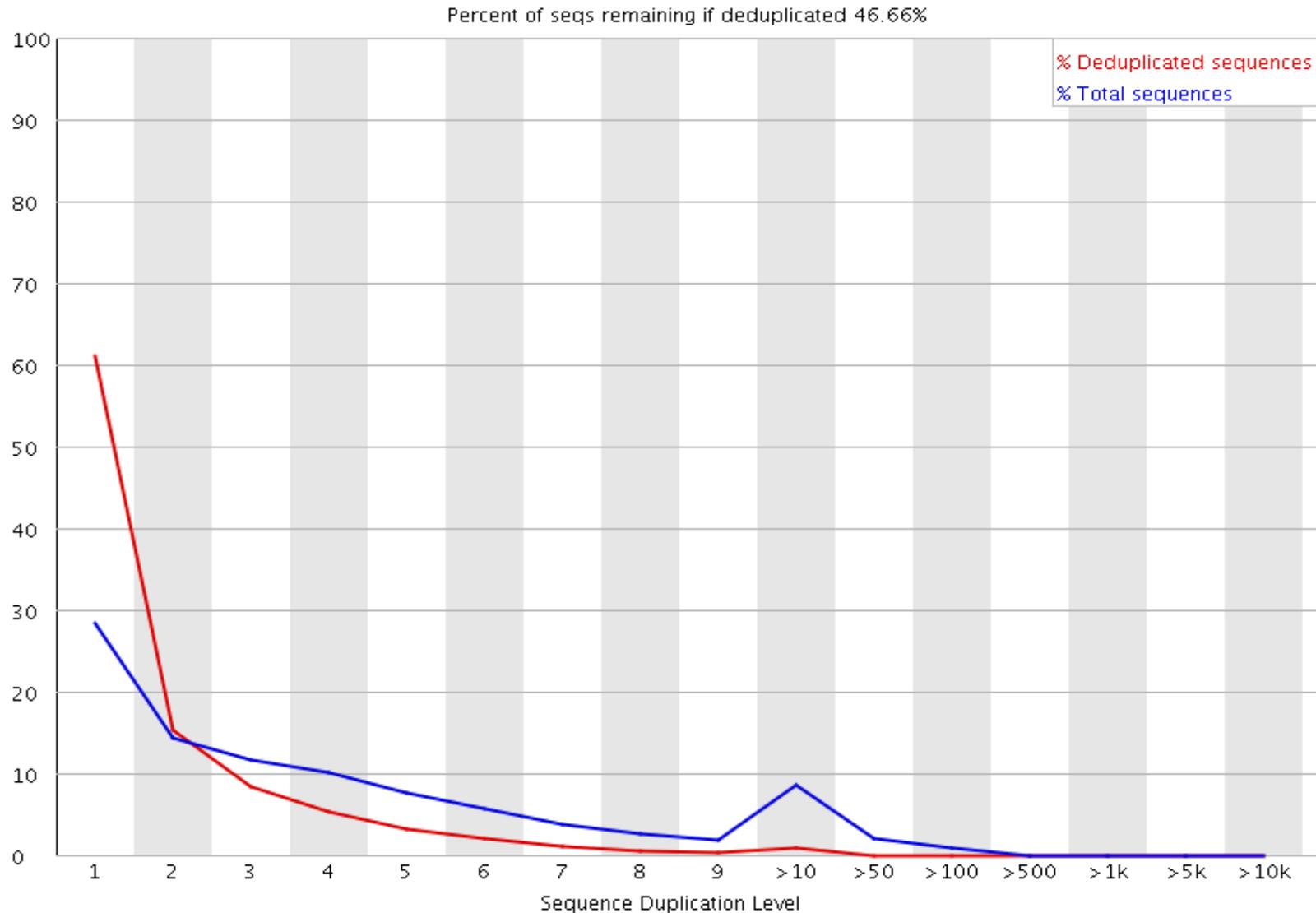
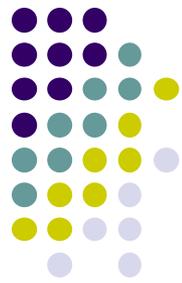


Expected sequence duplication is primarily a function of experiment type



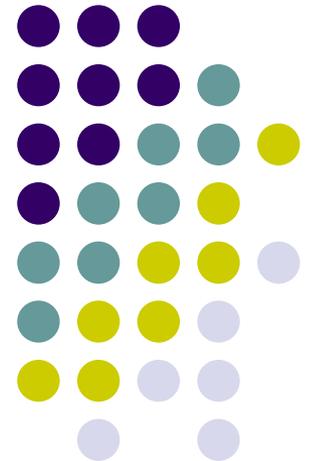
Newer FastQC versions have a slightly different Sequence Duplication report

- Red ***“deduplicated”*** line as previously described
- Blue ***“total”*** line is percentage histogram



Part 4: Alignment to a reference assembly

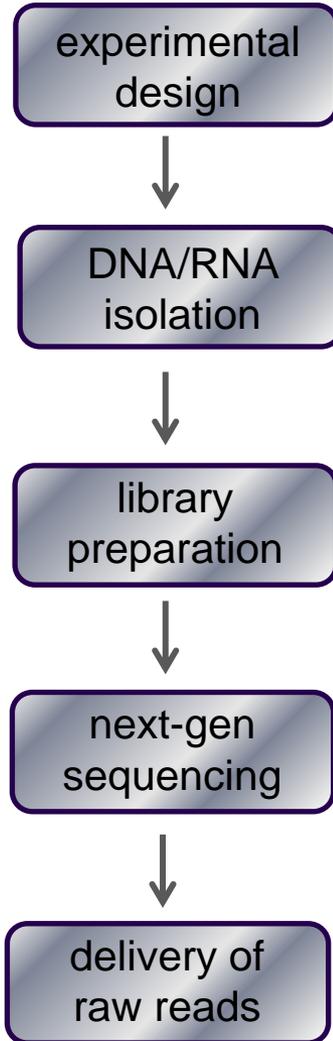
- Alignment overview & concepts
- Preparing a reference genome
- Alignment steps



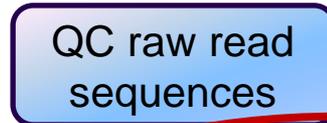
NGS Workflow

core processes

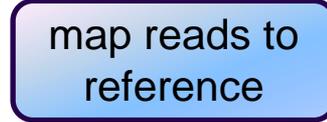
upstream processes



fastq



yes



has reference?



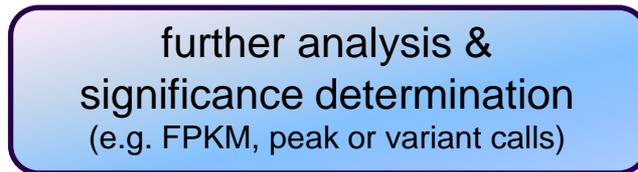
fasta



BAM



bed, gff, vcf, etc.



no



downstream processes

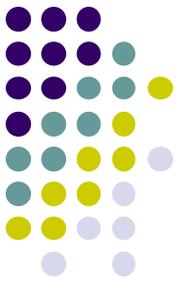


Short Read Aligners



- 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)
- 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

Aligner criteria



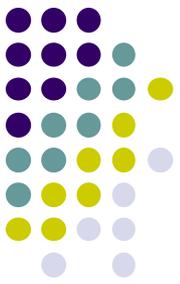
- Adoption and currency
 - widespread use by bioinformatics community
 - still being actively developed
- Features
 - well understood algorithm(s)
 - support for a variety of input formats and read lengths
 - detection of indels and gaps
 - makes use of base qualities
 - handling of multiple matches
- Usability
 - configurability and transparency of options
 - ease of installation and use
- Resource requirements
 - speed (“fast enough”)
 - scalability (takes advantage of multiple processors)
 - reasonable memory footprint

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 \neq High alignment score
 - **mapping quality** describes **positioning**
 - reflects the probability that the read is *incorrectly* mapped to the reported location
 - is a Phred score: $P(\text{incorrectly mapped}) = 10^{-\text{mappingQuality}/10}$
 - reflects the **complexity** or information content of the sequence (**mappability**)
 - **alignment score** describes **fit**
 - reflects the correspondence between the read and the reference sequences

	Read 1		Read 2	
• Maps to one location <i>high mapping quality</i>	GCGTAGTCTGCC		ATCGGGAGATCC	• Maps to 2 locations <i>low mapping quality</i>
• Has 2 mismatches <i>low alignment score</i>	TAGCCTAGTGTGCCGC	TAATCGGGAGATCCGC	TTATCGGGAGATCCGC	• Matches perfectly <i>high alignment score</i>
		<i>reference sequence</i>		

Mapping algorithms

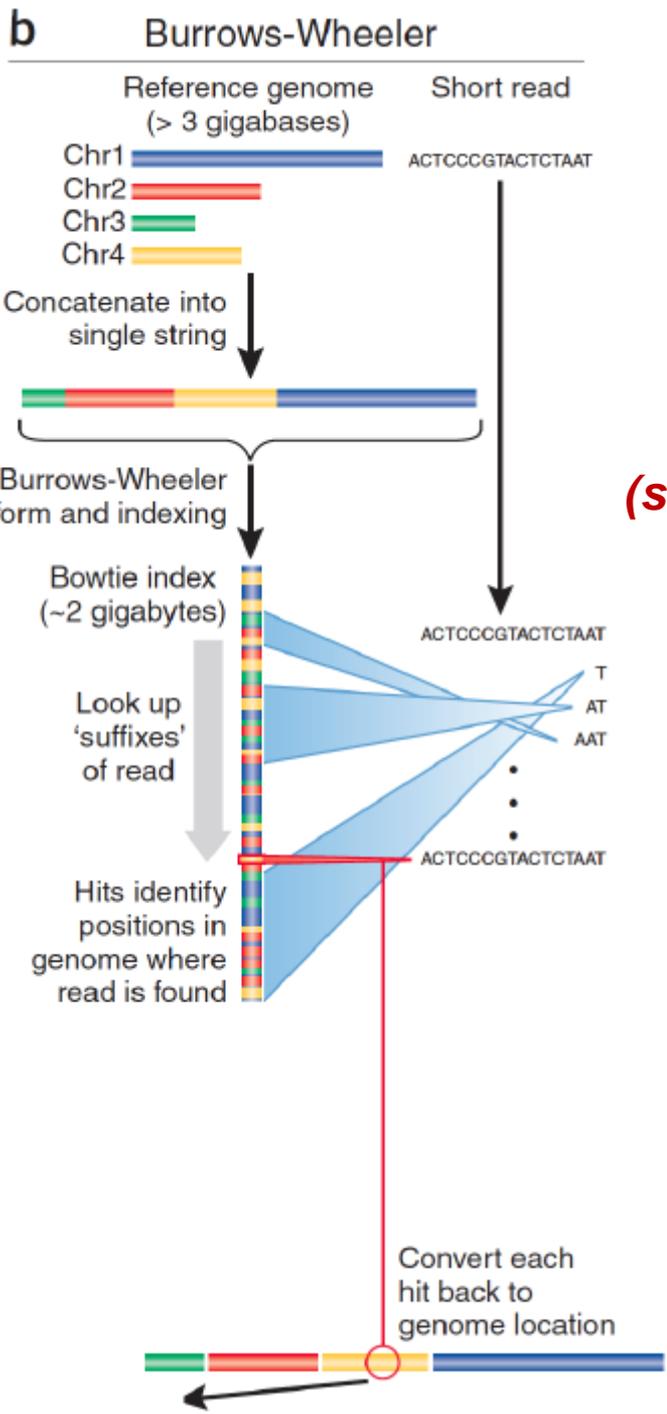


Two main mapping algorithms: *spaced seeds*, *suffix-array tries*

	Algorithm	Gapped	Quality-aware	Colorspace aware
BLAST	Hash table	Y	N	N
BLAT/SSHA2	Hash table	N	N	N
MAQ	Spaced seed	N	N	N
RMAP	Spaced seed	N	Y	N
ZOOM	Spaced seed	N	-	N
SOAP	Spaced seed	N	N	N
Eland	Spaced seed	N	N	N
SHRIMP	Q-gram/multi-seed	Y	Y	Y
BFAST	Q-gram/multi-seed	Y	Y	Y
Novoalign	Multi-seed + Vectorized SW	Y	Y	Y
clcBio	Multi-seed + Vectorized SW	Y	Y	Y
MUMmer	Tries	Y	N	N
OASIS	Tries	Y	-	-
VMATCH	Tries	Y	-	-
BWA/BWA-SW	Tries	Y	Y	Y
BOWTIE	Tries	Y	Y	Y
SOAP2	Tries	Y	N	N
Saruman	Exact (GPU)	Y	-	N

courtesy of Matt Vaughn, TACC

trie = tree structure for fast text retrieval.

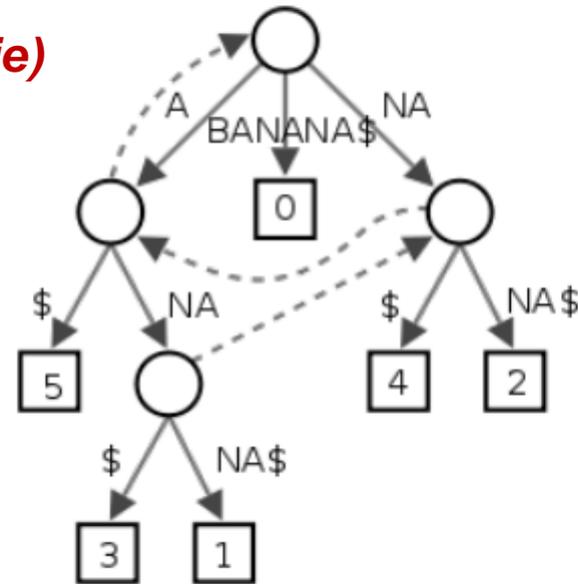


Burrows-Wheeler transform compresses sequence.

Input	SIX.MIXED.PIXIES.SIFT.SIXTY.PIXIE.DUST.BOXES
Output	TEXYDST.E.IXIXIXSSMPPS.B..E.S.EUSFXDIIIOIIT

Suffix tree enables fast lookup of subsequences.

(suffix array trie)

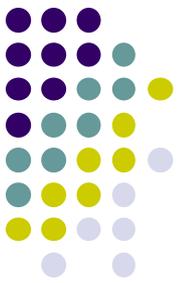


http://en.wikipedia.org/wiki/Suffix_tree

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

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

```

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

- **Alignment score** = \sum

- 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

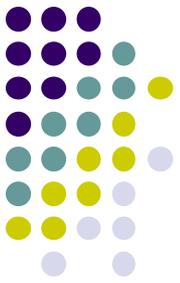
|||||

ATTTGCGATCGGATGTTGACTTT

ATTTGCGATCGGATGAAGACG..AA

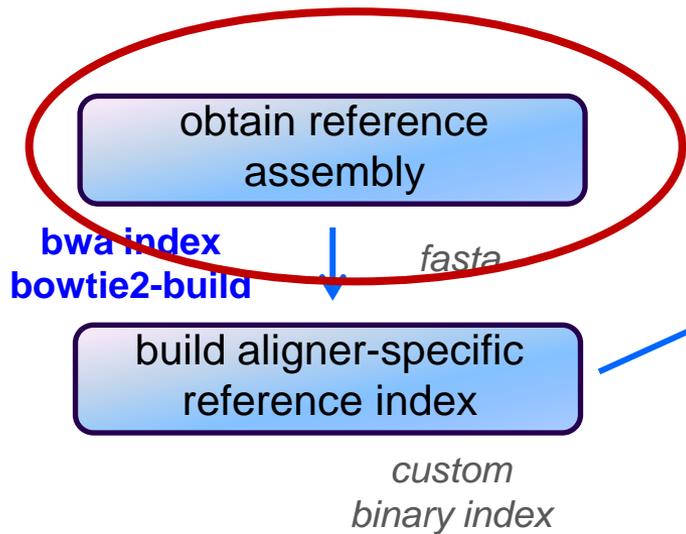
|||||XX|||Xi||

ATTTGCGATCGGATGTTGACTTAA

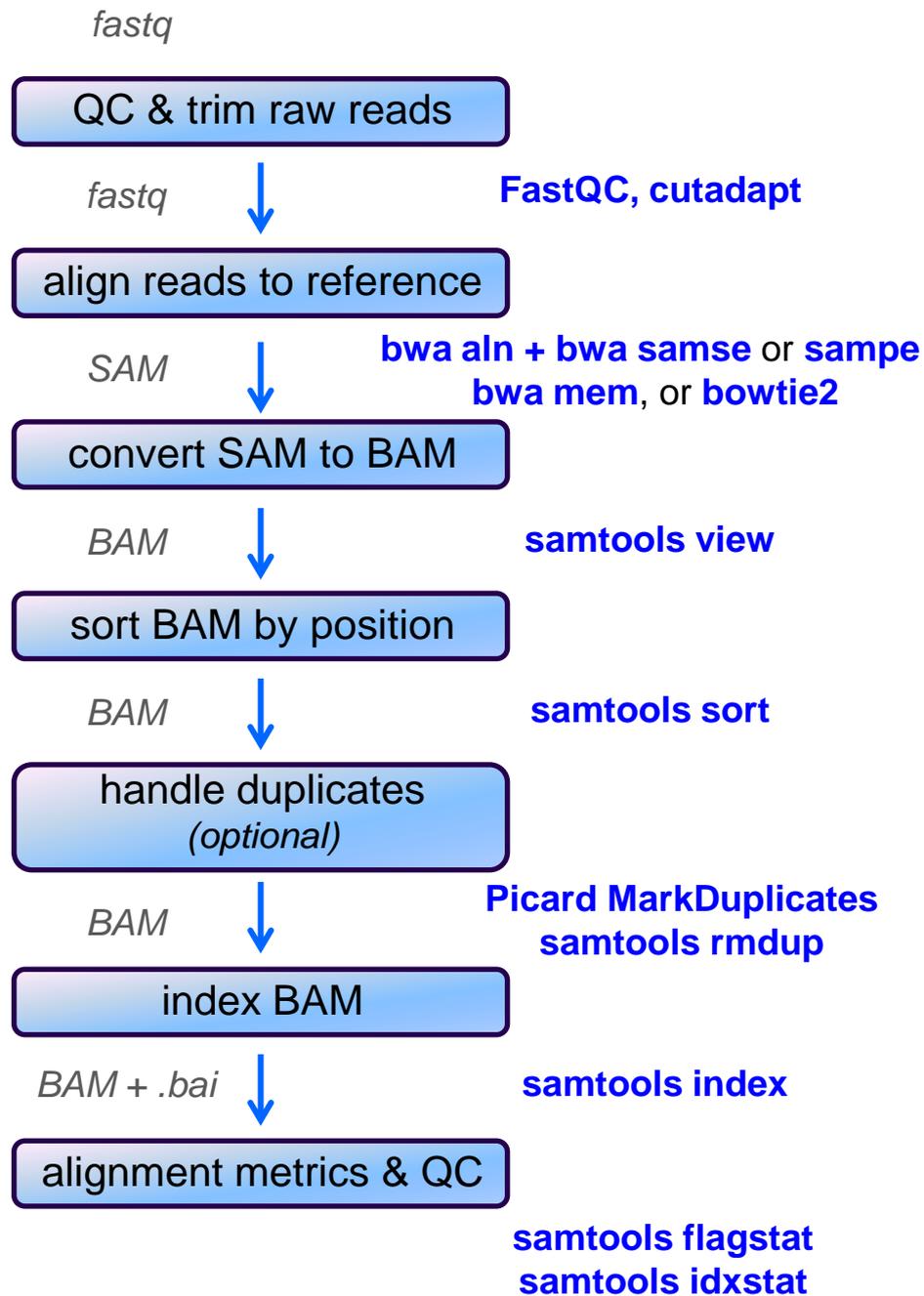


Paired End mapping

- Having paired-end reads improves mapping
 - mapping one read with high confidence anchors the pair
 - even when its mate read by itself maps several places equally
- Three possible outcomes of mapping an R1/R2 pair
 1. only one of a pair might map (*singleton/orphan*)
 2. both reads can map within the most likely distance range (*proper pair*)
 3. both reads can map but with an unexpected insert size or orientation, or to different contigs (*discordant pair*)
- Insert size is reported in the alignment record
 - for both proper and discordant pairs



Alignment Workflow

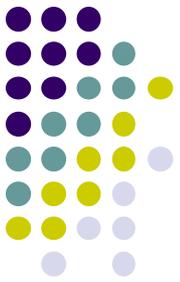




Obtaining a reference

- What is a reference?
 - *any set of named DNA sequences*
 - e.g. names are chromosome names
 - technically referred to as **contigs**
- Assembled genomes
 - Ensembl, UCSC, Gencode for eukaryotes
 - FASTA files (**.fa**, **.fasta**), + annotations (genome feature files, **.gff**)
 - GenBank, NCBI for prokaryotes/microbes
 - Records contain both fasta sequences and annotations
- Any set of sequences of interest, e.g:
 - transcriptome (set of transcribed gene sequences)
 - miRNA hairpin sequences from miRBase
 - rRNA/tRNA genes (e.g. for filtering)

FASTA format



- FASTA files contain a set of sequence records
 - sequence name line
 - always starts with >
 - followed by a **name** and other (optional) descriptive information
 - one or more sequence line(s)
 - never starts with >

- Mitochondrial chromosome sequence, human hg19

```
>chrM
GATCACAGGTCTATCACCCCTATTAACCACTCACGGGAGCTCTCCATGCAT
TTGGTATTTTCGTCTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCTG
GAGCCGGAGCACCCCTATGTTCGAGTATCTGTCTTTGATTCTGCCTCATT ...
```

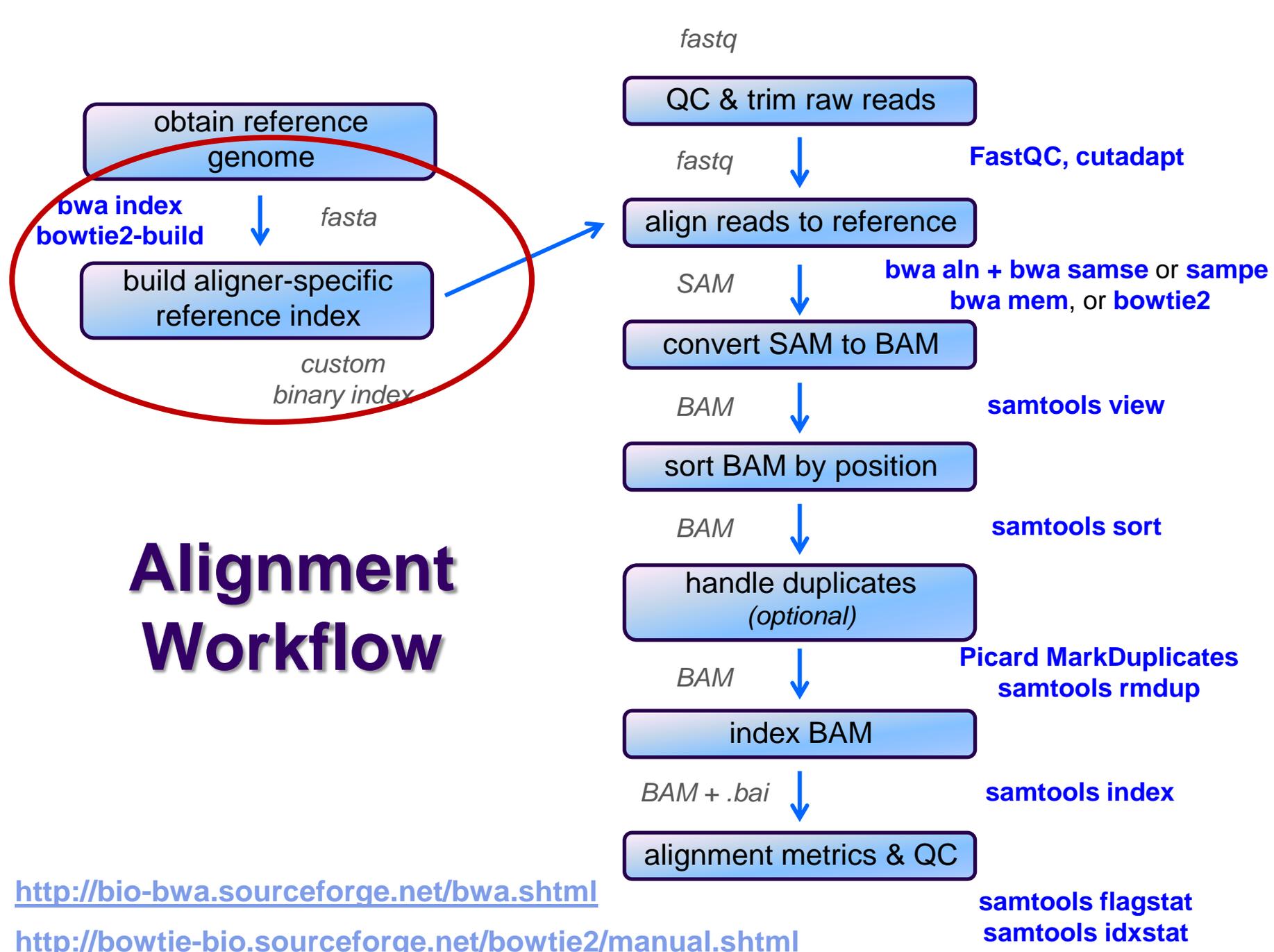
- Let-7e miRNA, human miRBase v21

```
>hsa-let-7e MI0000066 Homo sapiens let-7e stem-loop
CCCGGGCTGAGGTAGGAGGTTGTATAGTTGAGGAGGACACCCAAGGAGATCACTATACGG
CCTCCTAGCTTTCCCCAGG
```



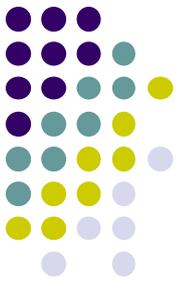
Reference considerations

- Is it appropriate to your study?
 - close enough to your species? complete?
- Does it contain repeats? What kinds?
 - know this up front or you will be confused
- From which source? And which version?
 - UCSC hg19 vs Ensembl GRCh37
- What annotations exist?
 - references lacking feature annotations are much more challenging
- Watch out for sequence name issues!
 - sequence names may be different between UCSC/Ensembl
 - e.g. “chr12” vs “12”
 - *annotation sequence names must match names in your reference!*
 - long sequence names can cause problems
 - rename: `>hsa-let-7e_MI0000066_Homo_sapiens_let-7e_stem-loop`
 - to: `>hsa-let-7e`



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

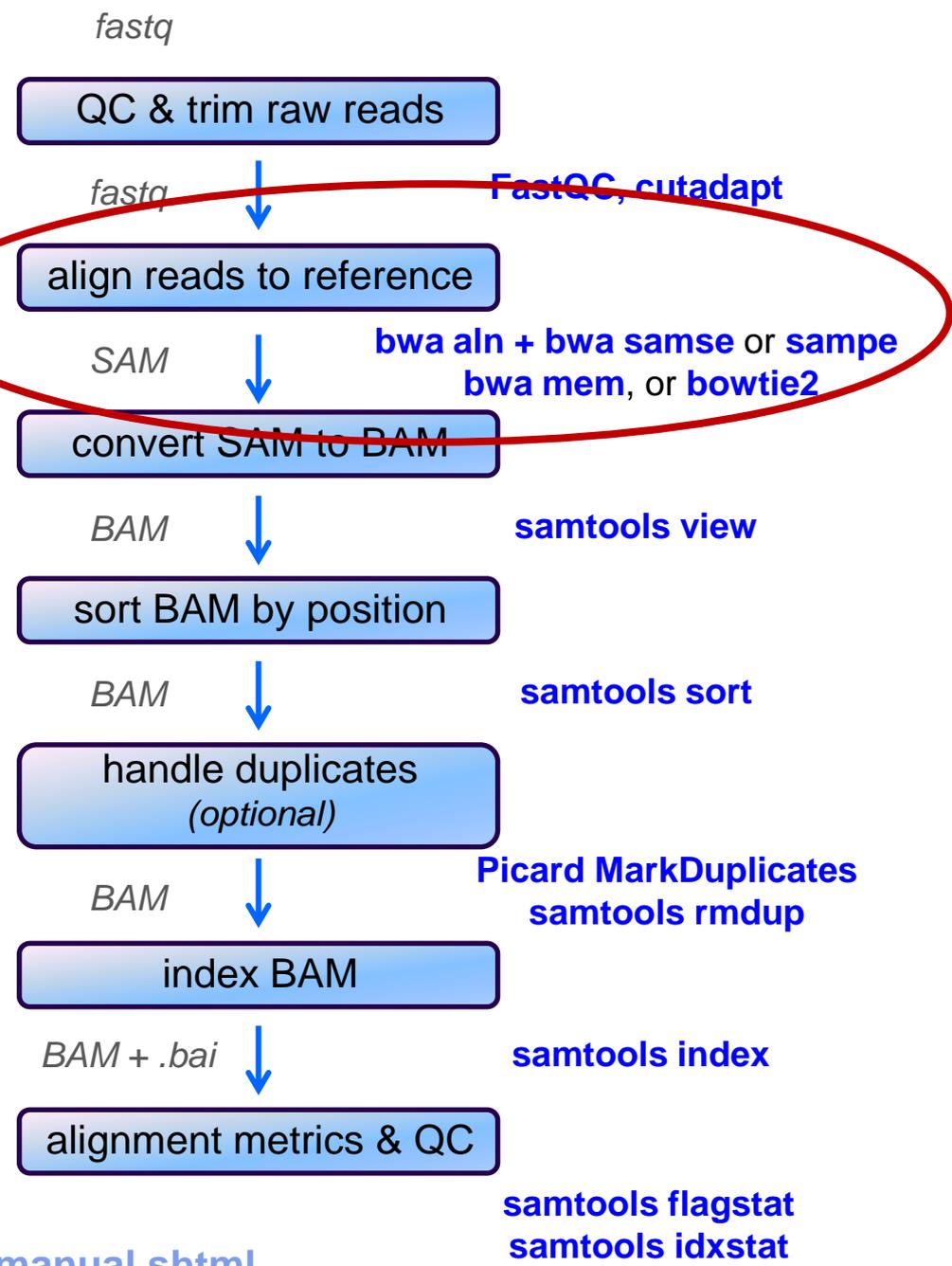
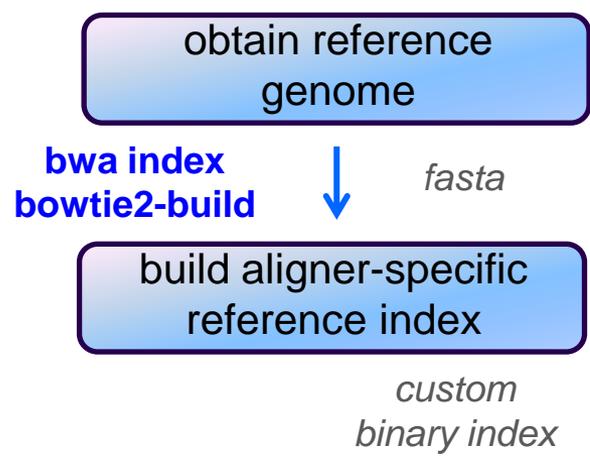
<http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml>



Building a reference index

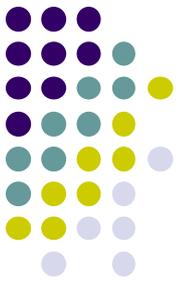
- Index format is specific to each aligner
 - may take several hours to build
 - but you build each index once, use for multiple alignments
 - Input:
 - one or more FASTA files
 - Output:
 - a number of binary files the aligner will use
- Best practice:
 - build each index in its own appropriately named directory, e.g.
 - `refs/bowtie2/UCSC/hg19`
 - `refs/bwa/Ensembl/GRCh37`

Alignment Workflow



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

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

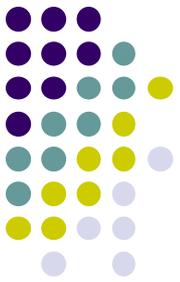


SAM file format

- Aligners take FASTQ as input, output alignments in **S**equence **A**lignment **M**ap (SAM) format
 - community file format that describes how reads align to a reference
 - the Bible: <http://samtools.github.io/hts-specs/SAMv1.pdf>
- SAM file consists of
 - a **header**
 - includes reference sequence names and lengths
 - **alignment records**, one for each sequence read
 - can include both mapped and unmapped reads
 - alignments for R1 and R2 reads have *separate records*
 - with fields that refer to the mate
 - 11 fixed fields + extensible-format **key:type:value** tuples

SAM file format

Fixed fields



Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME <i>read name from fastq</i>
2	FLAG	Int	[0,2 ¹⁶ -1]	bitwise <u>FLAGS</u>
3	RNAME	String	* [!-()+-<>-~] [!-~]*	Reference sequence NAME <i>contig + start</i>
4	POS	Int	[0,2 ²⁹ -1]	<u>1-based leftmost mapping POSITION</u> <i>= locus</i>
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 <i>insert size, if paired</i>
10	SEQ	String	* [A-Za-z=.]+	segment SEQUENCE
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

```

SRR030257.264529 99 NC_012967 1521 29 34M2S = 1564 79
CTGGCCATTATCTCGGTGGTAGGACATGGCATGCC
AAAAAA;AA;AAAAA??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
CTGGCCATTATCTCGGTGGTAGGTGATGGTATGCCG
<<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

SAM format – Bitwise flags

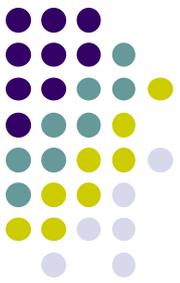


Bit

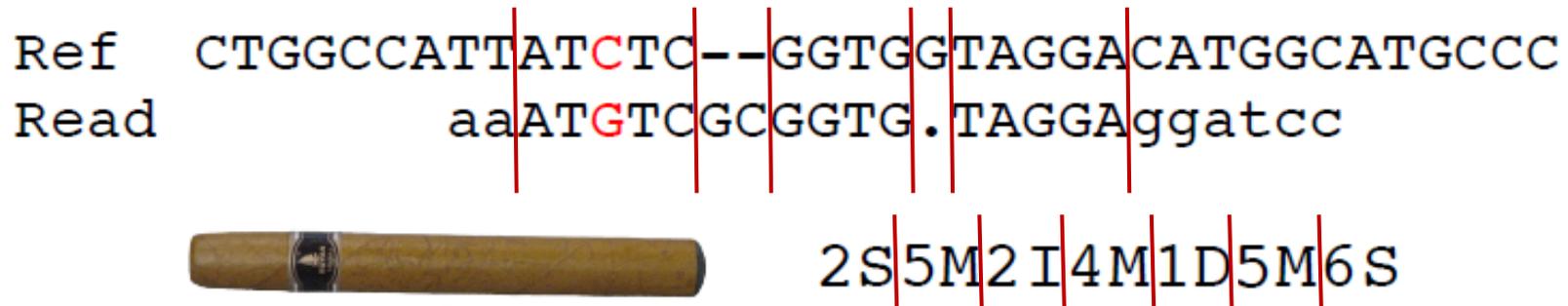
Decimal	Hex	Description	
1	0x1	template having multiple segments in sequencing	1 = part of a read pair
2	0x2	each segment properly aligned according to the aligner	1 = "properly" paired
4	0x4	segment unmapped	read <u>did</u> map = 0 1 = read <u>did not</u> map
8	0x8	next segment in the template unmapped	1 = mate <u>did not</u> map
16	0x10	SEQ being reverse complemented	<u>plus</u> strand read = 0 1 = minus strand read
32	0x20	SEQ of the next segment in the template being reverse complemented	1 = mate on minus strand
64	0x40	the first segment in the template	1 = R1 read
128	0x80	the last segment in the template	1 = R2 read
256	0x100	secondary alignment	1 = secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls	
1024	0x400	PCR or optical duplicate	1 = marked as duplicate
2048	0x800	supplementary alignment	1 = maps to ALT contig

	Decimal	Hex
SRR030257.264529 99 NC_012967 1521 29 34M2S = 1564 79	99	= 0x63
CTGGCCATTATCTCGGTGGTAGGACATGGCATGCC	= 64	= 0x40
AAAAAA;AA;AAAAA??A%.;?&'3735',()0*,	+ 32	+ 0x20
XT:A:M NM:i:3 SM:i:29 AM:i:29 XM:i:3 XO:i:0 XG:i:0 MD:Z:23T0G4T4	+ 2	+ 0x02
	+ 1	+ 0x01
SRR030257.2669090 147 NC_012967 1521 60 36M = 1458 -99	147	= 0x93
CTGGCCATTATCTCGGTGGTAGGTGATGGTATGCGC	= 128	= 0x80
<<9:<<AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	+ 16	+ 0x10
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	+ 2	+ 0x02
	+ 1	+ 0x01

<http://broadinstitute.github.io/picard/explain-flags.html>



Sometimes a CIGAR is just a way of describing how a read is aligned...

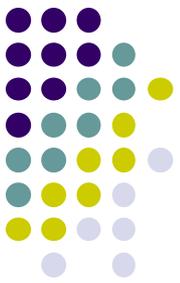


Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference <i>"N" indicates splicing event in RNAseq BAMs</i>
S	4	soft clipping (clipped sequences present in SEQ)
* H	5	hard clipping (clipped sequences NOT present in SEQ)
* P	6	padding (silent deletion from padded reference)
* =	7	sequence match
* X	8	sequence mismatch

CIGAR = "Concise Idiosyncratic Gapped Alignment Report"

SAM file format

key:type:value tuples



Tag ¹	Type	Description
X?	?	Reserved fields for end users (together with Y? and Z?)
...		
MD	Z	String for mismatching positions. <i>Regex</i> : [0-9]+((([A-Z] \^ [A-Z] +) [0-9] +)) * ²
MQ	i	Mapping quality of the mate/next segment
NH	i	Number of reported alignments that contains the query in the current record
NM	i	Edit distance to the reference, including ambiguous bases but excluding clipping
...		

details alignment of query to reference

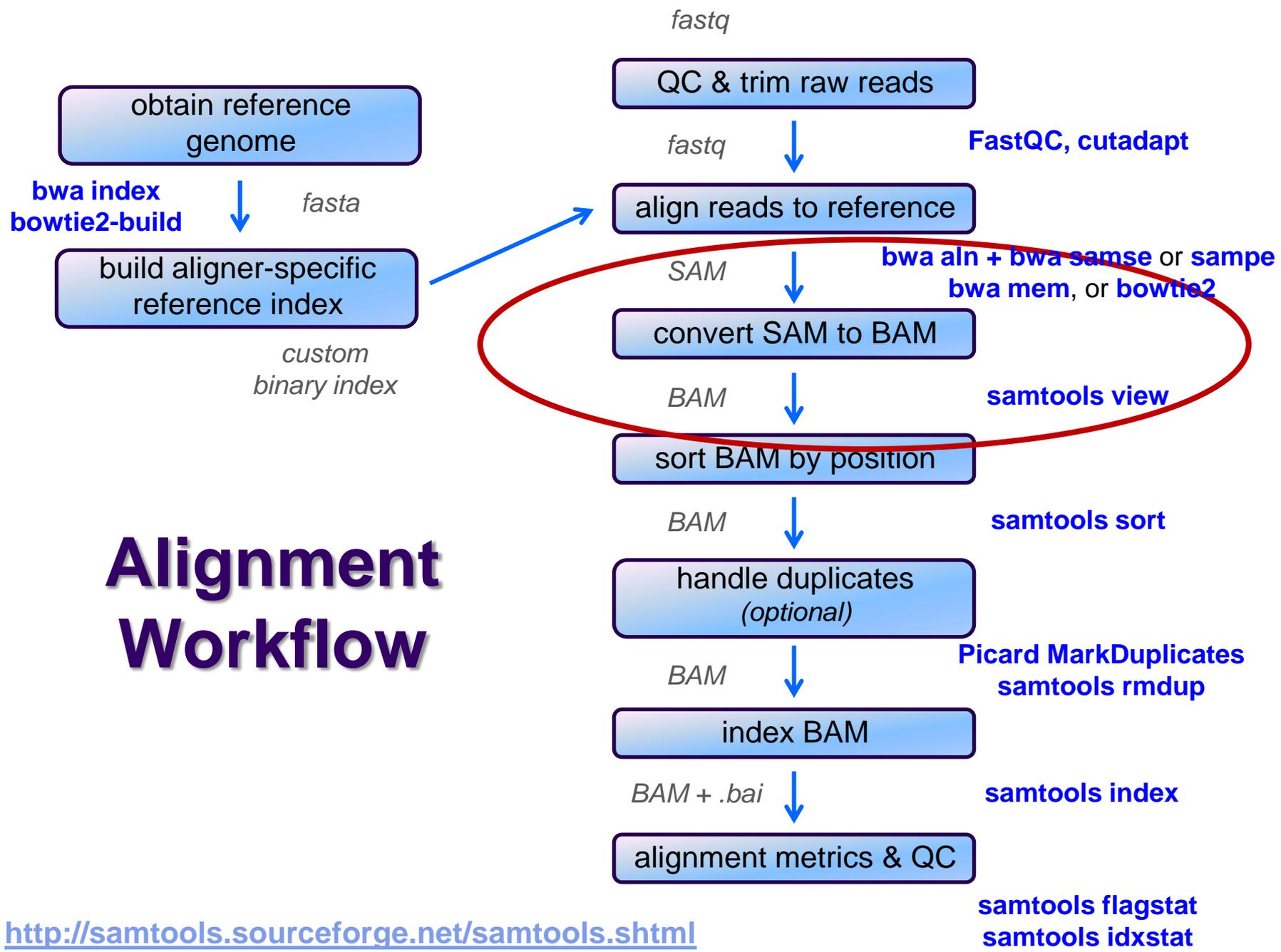
mismatches + insertions + deletions

²The MD field aims to achieve SNP/indel calling without looking at the reference. For example, a string '10A5^AC6' means from the leftmost reference base in the alignment, there are 10 matches followed by an A on the reference which is different from the aligned read base; the next 5 reference bases are matches followed by a 2bp deletion from the reference; the deleted sequence is AC; the last 6 bases are matches. The MD field ought to match the CIGAR string.

```

SRR030257.264529    99  NC_012967    1521    29  34M2S    =    1564    79
CTGGCCATTATCTCGGTGGTAGGACATGGCATGCC
AAAAAAAA:AA;AAAAA??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
    
```

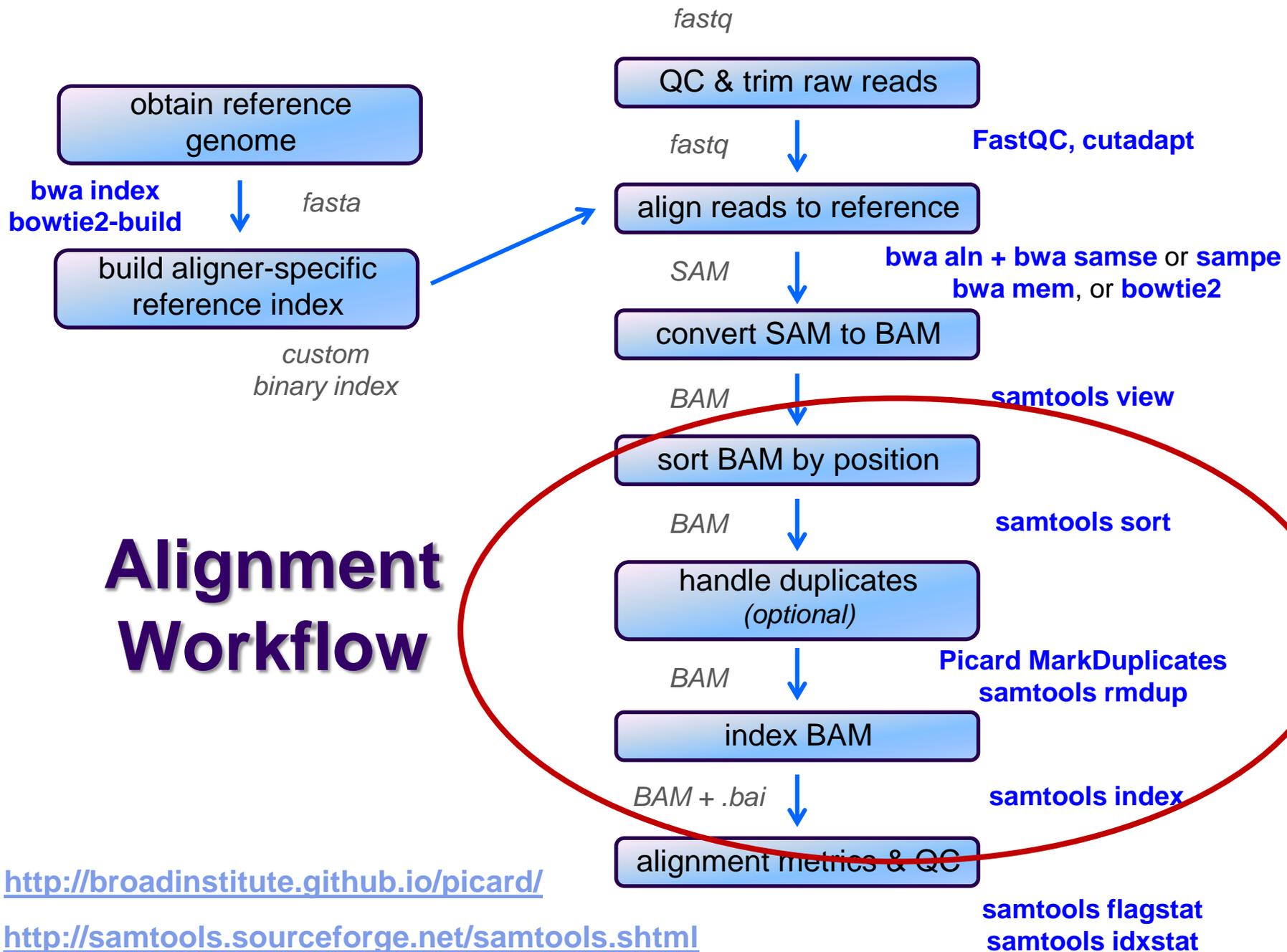
Alignment Workflow

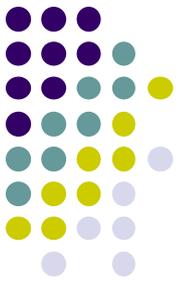


SAM / BAM files



- SAM and BAM are two forms of the same data
 - SAM – **S**equence **A**lignment **M**ap
 - plain text format
 - BAM – **B**inary **A**lignment **M**ap
 - **same data** in a custom compressed (**gzip**'d) format
- Differences
 - BAMs are **much** smaller than SAM files due to compression
 - BAM files support fast random access; SAM files do not
 - requires the BAM file to be *indexed*
 - most tools support BAM format and may require indexing
- Best practices
 - remove intermediate SAM and BAM files created during alignment and only save the final sorted, indexed BAM
 - keep your alignment artifacts (BAM, statistics files, log files) separate from the original FASTQ files
 - alignments can be re-generated – raw sequences cannot

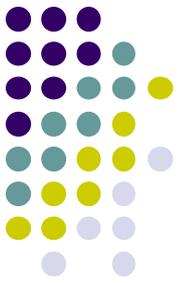




Sorting / indexing BAM files

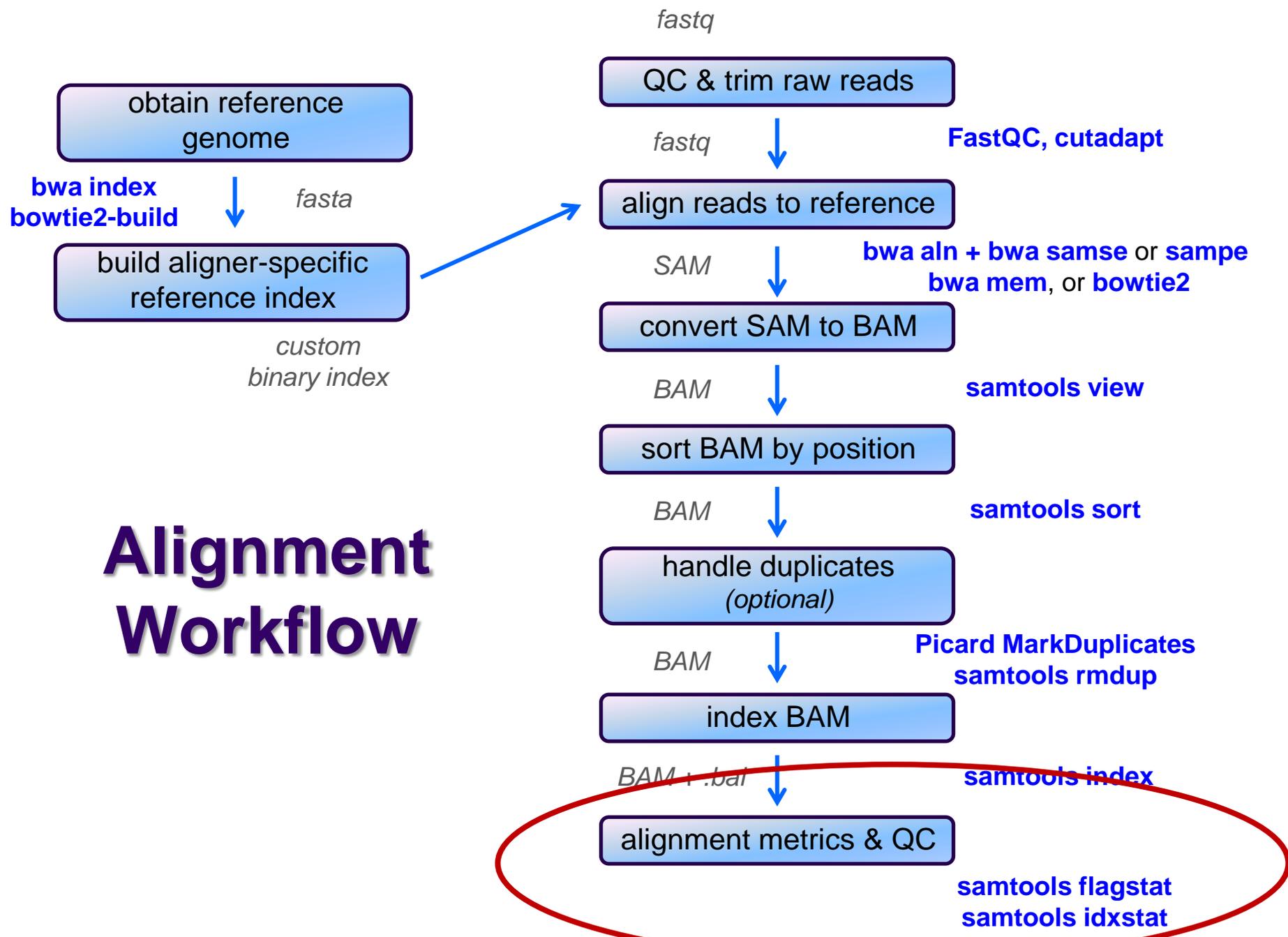
- SAM created by aligner contains read records in *name order*
 - same order as read names in the input FASTQ file
 - R1, R2 have adjacent SAM records
 - SAM → BAM conversion does not change the name-sorted order
- Sorting BAM puts records in *position (locus) order*
 - by contig name then start position (leftmost)
 - contig name order given in SAM/BAM header
 - based on order of sequences in FASTA used to build reference
 - **sorting is very compute and I/O intensive**
 - can take hours for large BAM
- Indexing a locus-sorted BAM allows fast random access
 - creates a 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
 - properly paired 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
 - does not intelligently handle data from multiple lanes
 - **Picard MarkDuplicates** – **flags** duplicates only (0x400 bam flag)
 - slower, but all alignments are retained
 - alignments from different lanes/replicates are considered separately
 - also newer **MarkDuplicatesWithMateCigar** tool
 - takes CIGAR string(s) into account; slower than plain **MarkDuplicates**
 - 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 metrics



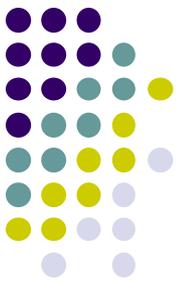
- **samtools idxstats**
 - reads aligning to each contig

<i>contig</i>	<i>length</i>	<i># map</i>	<i># not mapped</i>
chrI	230218	553609	2183
chrII	813184	1942996	5605
chrIII	316620	764449	2246
chrIV	1531933	3630237	10049
chrV	576874	1432940	4149
chrVI	270161	658338	1859
chrVII	1090940	2628838	7283
chrVIII	562643	1347702	4064
chrIX	439888	1079444	3057
chrX	745751	1861421	5576
chrXI	666816	1595615	4026
chrXII	1078177	4595061	23201
chrXIII	924431	2253102	6260
chrXIV	784333	1861773	5367
chrXV	1091291	2625205	7080
chrXVI	948066	2266237	6233
chrM	85779	210993	956
*	0	0	2291804

samtools notes



- There are 2 main “eras” of the **samtools** program
 - “old” **samtools**
 - v 0.1.19 last stable version
 - “new” **samtools**
 - v 1.0, 1.1, 1.2 – avoid these (very buggy!)
 - v 1.3+ stable
 - some functions have different arguments!
- **samtools** v 1.3+ has many new features
 - **samtools stats**
 - produces *many* different statistical reports
 - faster sorting
 - can use multiple threads



Computing average insert size

- Needed for some downstream analysis
 - e.g. RNAseq alignment using **tophat**
- Simple **awk** script that computes average insert size for a BAM
 - **-F 0x4** filter to **samtools view** says only consider mapped reads
 - technically “not unmapped”
 - the **-f 0x2** filter says consider only properly paired reads
 - they have reliable “insert size” values in column 9
 - insert size values are negative for minus strand reads
 - can ignore because each proper pair will have one plus and one minus strand alignment, with same insert size

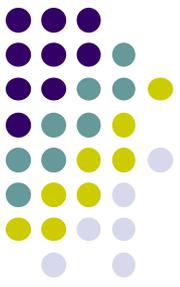
```
samtools view -F 0x4 -f 0x2 my_pe_data.bam | awk \  
  'BEGIN{ FS="\t"; sum=0; nrec=0; }\  
  { if ($9 > 0) {sum += $9; nrec++;} }\  
  END{ print sum/nrec; }'
```



Interpreting alignment metrics

- Table below is taken from a spreadsheet I keep on all our alignments
 - all are yeast paired-end read datasets from ChIP-seq experiments
- Alignment rates
 - samples 1-3 have excellent alignment rates & good rates of proper pairing
 - sample 4
 - has an unusually low alignment rate for a ChIP-seq dataset
 - has a median insert size of only 109, and these were un-trimmed 50 bp reads
 - could 3' adapter contamination be affecting the alignment rate?
 - try re-aligning the sequences after trimming, say to 35 bases
 - see if the alignment rate improves

#	totSeq	totAlign	% align	numPair	pePrAln	% prPr	nDup	% dup	multiHit	% multi	iszMed
1	149,644,822	145,228,810	97.0%	74,822,411	72,221,545	96.5%	49,745,225	34%	16,216,807	11%	181
2	981,186	860,940	87.7%	490,593	424,915	86.6%	609,378	71%	127,987	15%	148
3	22,573,348	21,928,789	97.1%	11,286,674	10,783,971	95.5%	9,408,725	43%	3,711,004	17%	132
4	7,200,628	3,460,992	48.1%	3,600,314	1,626,121	45.2%	1,234,524	36%	649,690	19%	109

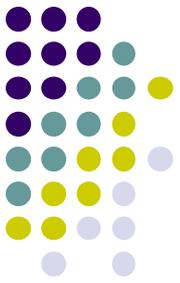


Interpreting alignment metrics

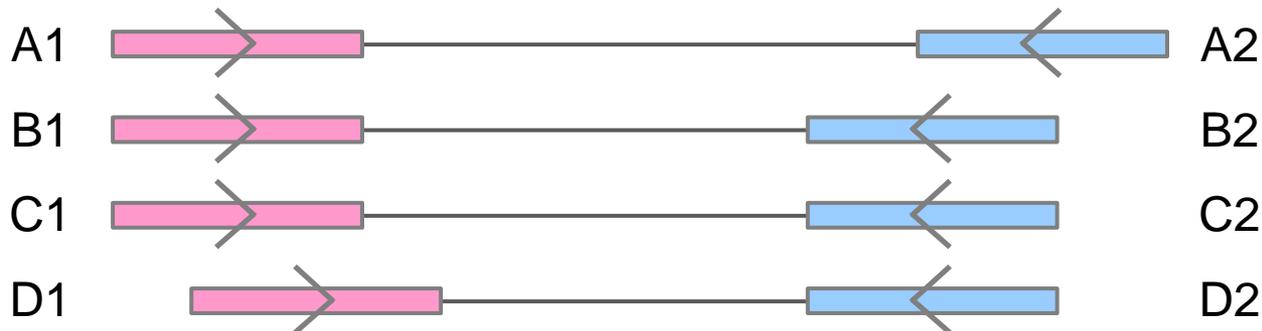
- Duplication rates
 - sample 1 is incredibly deeply sequenced (yeast genome only ~12 Mbase)
 - has a *very* low duplication rate considering! (34%)
 - turns out this is a control dataset (Mock ChIP), so is a great control to use (wonderfully complex!)
 - sample 2 is not very deeply sequenced but has a high duplication rate (71%)
 - subtracting duplicates from total aligned leaves only ~250,000 non-dup reads
 - not enough for further analysis (prefer 500,000+)
 - sample 3 has reasonable sequencing depth with substantial duplication (43%)
 - still leaves plenty of non-duplicate reads (> 12 million)

#	totSeq	totAlign	% align	numPair	pePrAln	% prPr	nDup	% dup	multiHit	% multi	iszMed
1	149,644,822	145,228,810	97.0%	74,822,411	72,221,545	96.5%	49,745,225	34%	16,216,807	11%	181
2	981,186	860,940	87.7%	490,593	424,915	86.6%	609,378	71%	127,987	15%	148
3	22,573,348	21,928,789	97.1%	11,286,674	10,783,971	95.5%	9,408,725	43%	3,711,004	17%	132
4	7,200,628	3,460,992	48.1%	3,600,314	1,626,121	45.2%	1,234,524	36%	649,690	19%	109

Read vs fragment duplication



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





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
- For UT folks with TACC accounts
 - The BioITeam has a set of TACC-aware alignment pipeline scripts
 - plus a set of pre-built reference indexes

Final thoughts

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

