NOTES

Lecture 1: Welcome to Bioinformatics

Explicit Controls in Biology Required

- S. aureus requires iron
- if you knock out IsdA gene, S. aureus grows poorly on heme
- if you rescue S. aureus with an IsdA plasmid, growth is restored

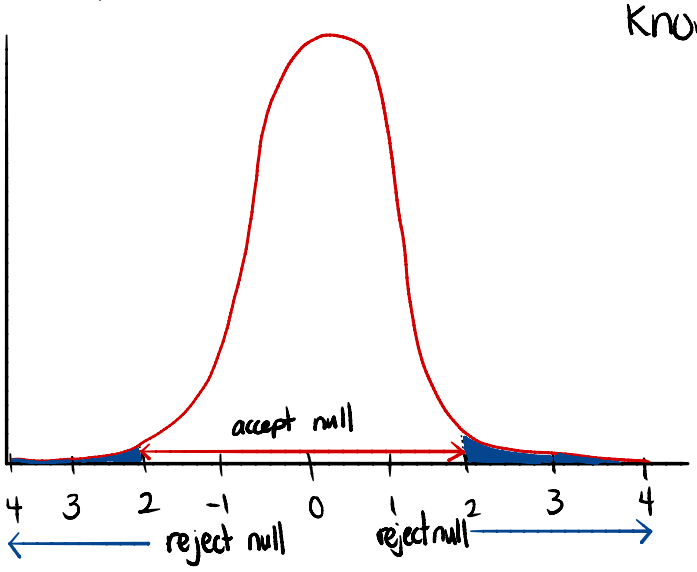
Explicit Controls in Dry Lab Absent

- Biophysical Methods
  - DNA sequencing
  - X-ray crystallography
- experimental results are assessed by
  1. statistical scores
  2. reproducibility
  3. expected outcomes

→ DNA sequence gives one of four bases at each position?
→ Dendrogram shows expected biological relationships

Critical Assessment in Science

- same critical assessment applies in vitro, in vivo, in silico
  - → what assumptions am I making?
  - → How do I assess that an experiment is working correctly?
    - → Am I using appropriate controls
    - → what statistical (or other) indicators are given to show significance of the program output (results)?
  - → Do the results meet the expectations (hypothesis)?
  - → what conclusions can I make?
  - → Do the results complement results derived from other methods (computational, or at the bench)? AND pre-existing biological knowledge?



Standards and References

- standardized data structures
  - FASTA/Q, sam/bam files
- curated resources
  - human reference genome builds
  - annotations associated with genome builds
  - resources constantly updated (e.g. hg38)

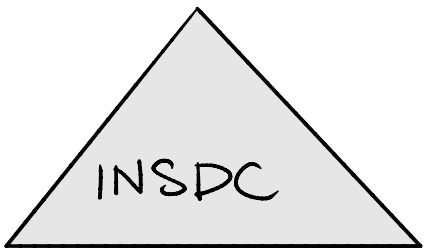
} foundational for sequence based bioinformatic analysis

Databases

- NCBI Entrez database
  - nucleic acid and protein sequences can be organized by relationships
  - an integrated database retrieval system that provides access to a diverse set of 39 databases that together contain over 1.7 billion records
    → includes

PubMed | Nucleotide and Protein Sequences | Protein Structures | Complete Genomes | Taxonomy

① Primary Nucleotide Sequence Databases



International Nucleotide Sequence Database Collaboration (INSDC)

NCBI
North America → USA

DDBJ
Asia → Japan

② NCBI Entrez Database Retrieval System

- each entry uses a unique identifier
- sometimes the identifier itself contains information
- multiple identifiers can exist for a single sequence

③ Entrez HardLinks vs Neighbours

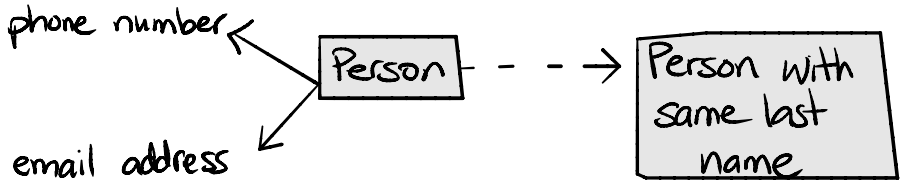
→ HardLinks

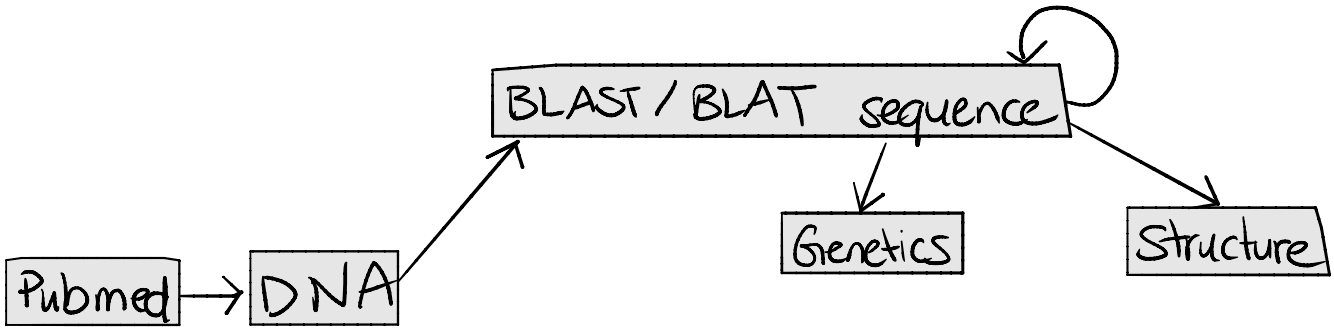
- DIRECT connections between entries in different databases
  - e.g. paper → nucleotide sequence
    taxonomy → protein sequence
    nucleotide sequence → protein CDS
    protein sequence → 3D structure
- not all possible links are present (depends on the source)

→ Neighbours

- another way to make connections between entries in different databases
  - e.g. similar sequences
    related papers
    similarity in 3D structure
- different definition of similarity for each database (subjective)
- e.g. related sequences
  - similar sequences identified using BLAST
    - precomputed BLAST results for all sequences in Genbank
    - sequence similarity meets a statistical criteria (cut off)
    - different list of neighbours for protein sequence vs. nucleotide sequence
    - two sequences that have a high level of sequence similarity often have related biological functions
- e.g. 3D structures
  - similar structures
    - proteins with the same fold or arrangement of secondary structure elements
- e.g. similar papers in Pubmed
  - measured by the number of "words" that the two papers have in common

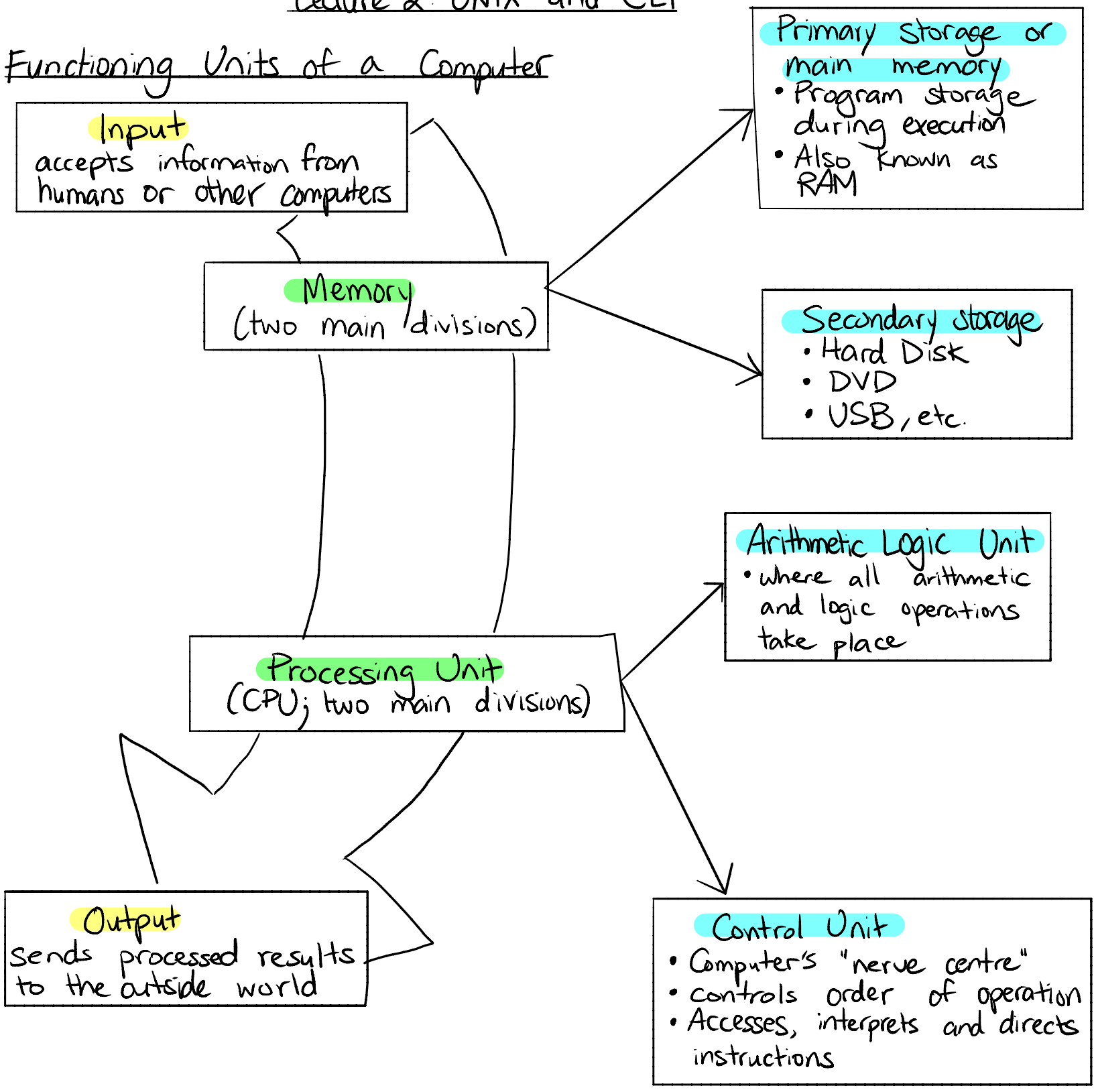
RECAP




BLAST / BLAT sequence
Pubmed
DNA
Genetics
Structure

Lecture 2: UNIX and CLI

Functioning Units of a Computer



① The Operating System
- the operating system is the suite of programs that make a computer work
  - e.g. Microsoft Windows 8, Vista, MacOS X, Linux or AIX

- the UNIX operating system is made of 3 parts
  - the kernel
  - the shell
  - the programs

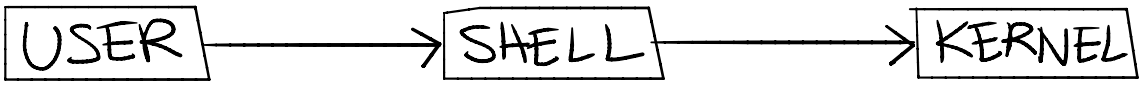
The Kernel

- the "hub" of the OS
  - allocates time and memory to programs
  - handles the filestore (files and directories)
  - handles the communications in response to system calls
- e.g. **rm *myfile***
  1. the shell searches the filestore for the file containing the program **rm**
  2. the shell requests the kernel (through system calls) to execute the program **rm** on ***myfile***
  3. when the process **rm *myfile*** has finished running, the shell then returns the UNIX prompt $ to you (indicating that it is waiting further commands)

SHELL → KERNEL → PROGRAM

The Shell

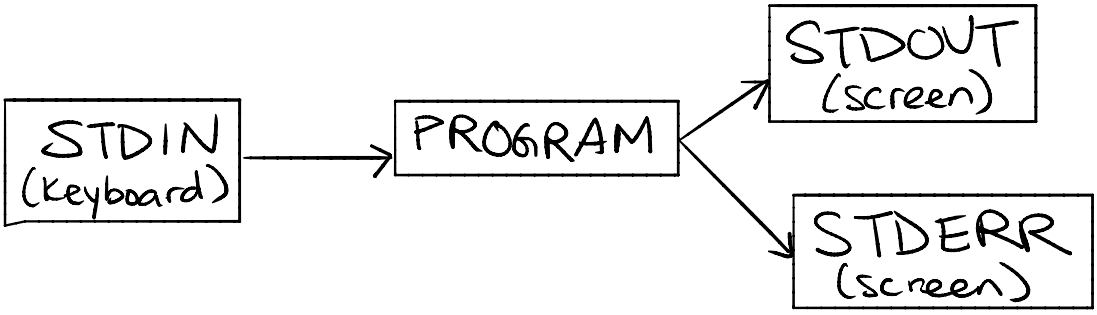
- interface between the user and the kernel
  - when you login, the login program checks the username and password, then starts another program called the shell
- a command line interpreter
  - interprets the commands and arranges for them to carried out
  - the commands themselves are programs and when they finish, the shell gives you another prompt
- common shells
  - Bourne shell (sh)
  - C shell (csh)
  - TC shell (tcsh)
  - Korn shell (ksh)
  - Bourne Again SHell (bash)



Files and Processes

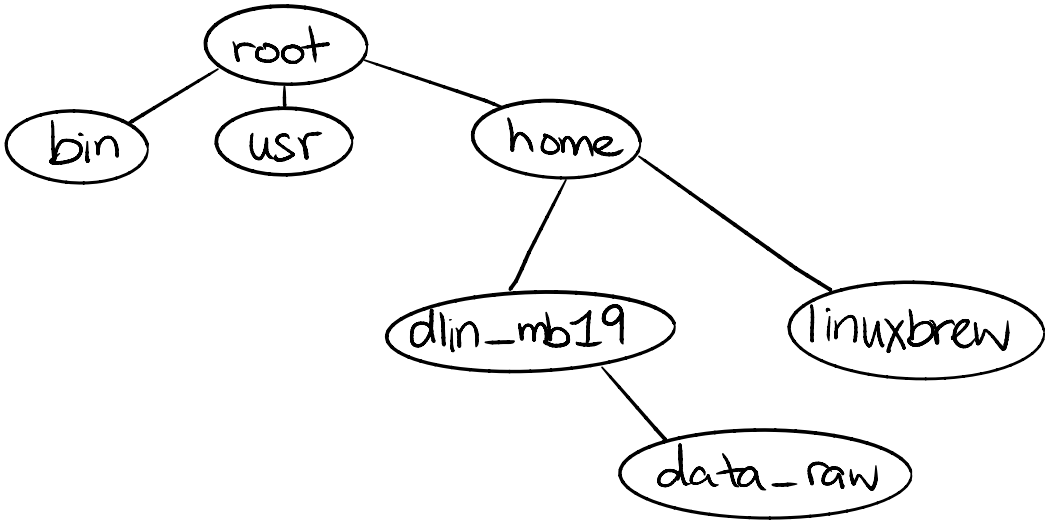
- everything in UNIX is either a file or a process
  - a process is an executing program identified by a unique process identifier (PID)
  - a file is a collection of data
    - they are created by users using text editors, running compilers, etc.
      - a document (report, essay etc.)
      - the text of a program (e.g. bwa) written in a programming language
      - a machine readable file (e.g. .bam file)
      - a directory, containing information about its contents

Standard in/out/error



The Directory Structure

- all files are grouped together in the directory structure
- file system arranged in hierarchical structure (ie. inverted tree)
- top of the hierarchy is traditionally called root.



② Navigating the file-system

- **ls**
- **cd**
- **mkdir**
- **pwd**

③ Copying, moving and viewing files

- **cp**
- **mv**
- **rm**
- **rmdir**
- **cat**
- **less**
- **head**
- **tail**
- **grep [-vnci]**
- **wc [-l]**

④ Wildcards, naming conventions and help

- * : match zero or more characters
- ? : match exactly one character

⑤ Redirection

- most processes initiated by UNIX commands write to standard output (the terminal screen), and many take their input from standard input (i.e. keyboard)

- >
- >>
- <
- |
- cat
- sort

⑥ File system security

- chmod
- chown

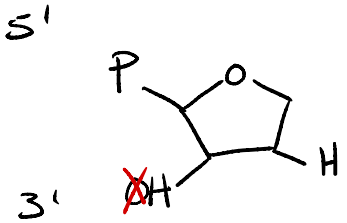
⑦ Other useful UNIX commands

- df
- gzip / gunzip
- history
- top
- !! (last command)
- !-2
- who
- !$

Lecture 3 : DNA Sequencing

Sanger Sequencing

- using labelled primer, labelled dNTP and labelled terminator (ddNTP)
- ddNTP : dideoxynucleotide triphosphates as DNA chain terminators
- labelled with P32 radioactive tags
- remove oxygen from hydroxyl group so no elongation



- Sanger won 2 Nobel prizes
- 250 bp off each sequence

Phred Score

↳ Phil Green

- PHil's Revised Editor (Phred) - the base quality standard

$$Q = -10 \log_{10} P$$

P = probability that the base is incorrect
Q = phrod quality score

- For Sanger sequencing (more complex for NGS), P is derivd from:
  1. Peak spacing
     - The ratio of the largest peak-to-peak spacing, in a window of seven peaks centred around the current one, to the smallest peak-to-peak spacing
  2. Uncalled/called ratio in a window of seven peaks around the current one.
  3. Uncalled/called ratio in a window of three peaks around the current one.
  4. Peak resolution
     - the number of bases between the current base and the nearest unresolved (N) base

→ a sequence of known bases is sequenced repeatedly and performance is analyzed

- an artificial cap of Phred is $Q = 50$

$$50 = -10 \log_{10} P$$

$$-5 = \log_{10} P$$

$$P = 10^{-5}$$

↳ 1 in 100,000 probability of being incorrect

The Human Reference Genome

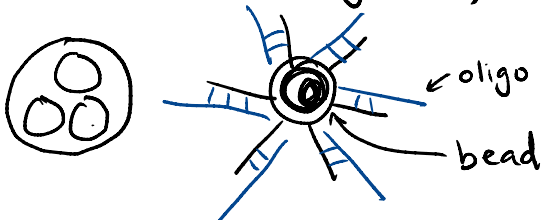
- $Q = 20$ with 6X coverage of NGS ("Bermuda Standard")
- currently: GRCh 38p12
- enabled studies of genetic variation and genomic function
- provided a framework for designing new tools for genome analysis
  - BAC arrays (CNV)
  - Oligonucleotide arrays
  - "Re-sequencing" strategies
  - Functional genomics

Second Generation Sequencers

- move from analog to digital sequencing
- individual sequence fragments are clonally amplified (replicates of ONE sequence strand)
- sequence by synthesis from single strands (typically ~ 1000)

Clonal Amplification

1. Oil / aqueous emulsion (e.g. 454, ion torrent)

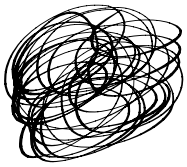


2. Solid Surface – microfluidic slide (e.g. Illumina)
3. Rolling circle amplification (e.g. Complete Genomics)

→ Sanger requires many many templates, but NGS makes replicates of ONE DNA strand

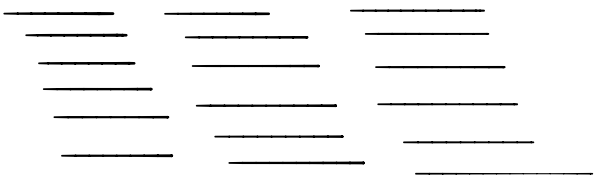
NGS Library Construction

DNA/cDNA



↓ Shear

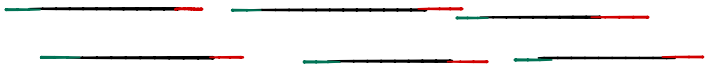
Pool of random fragments



Limitation: fragment size cap for clonal amplification

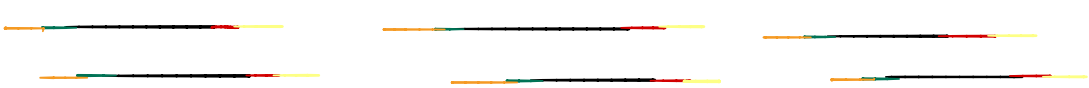
↓ end repair and ligate

ligate adapters



↓ PCR amplify

Amplified Library ready for sequencing



GS-FLX 454 Sequencer

- Clonal amplification via oil/aqueous emulsion
- Sequence by synthesis using light to detect nucleotide incorporation
- by Jonathan Rothberg

ADP

dNTP (NOT ddNTP)

pyrosequencing → light per base incorporated → flush and scan, one base at a time → flowed over plate, drop into wells

- Drawbacks: repetitive bases → how much more light is incorporated?

Ion Torrent Sequencer

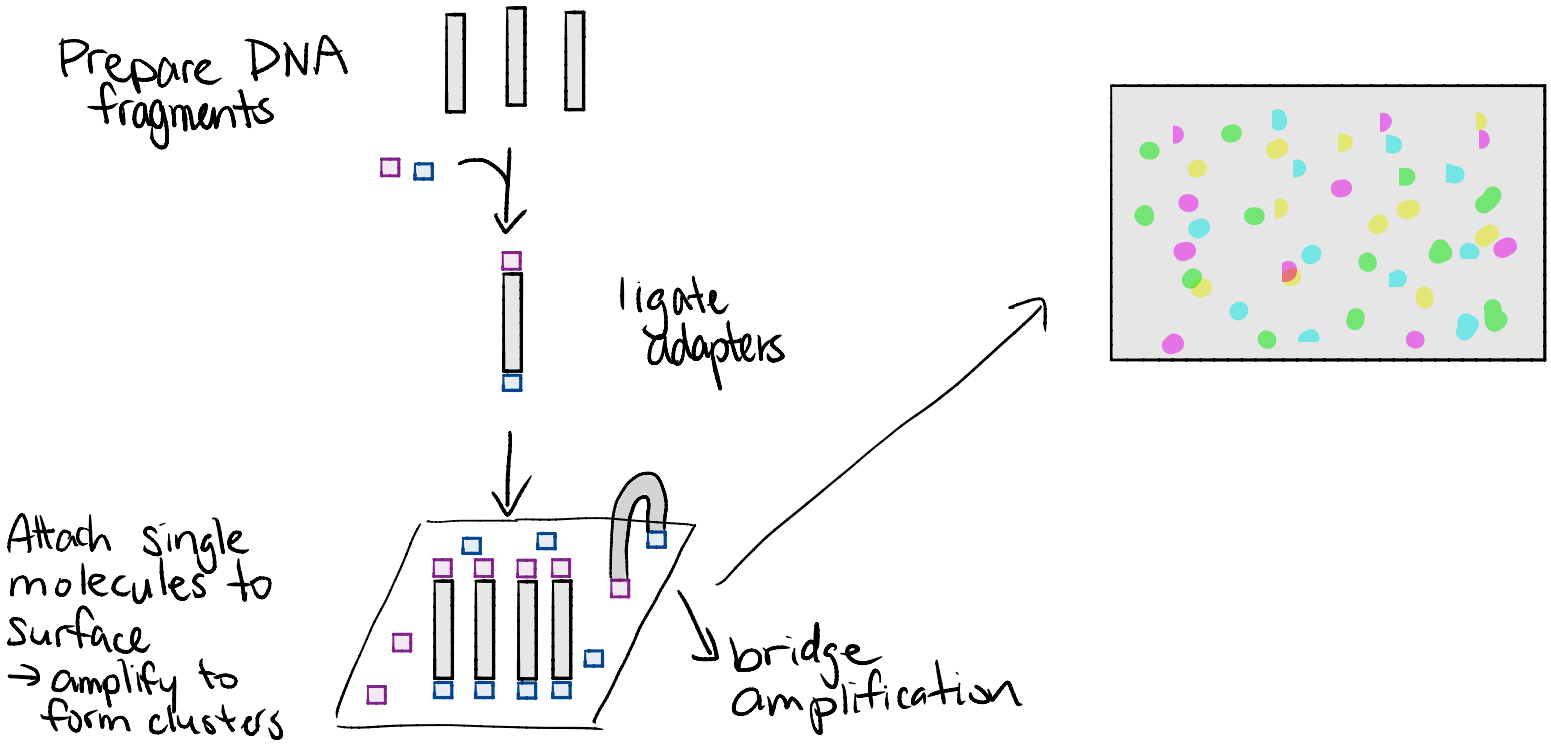
- Sequencing in microwells
- H^+ released during nucleotide incorporation
- no H^+ released if mismatched
- H^+ should scale with number of incorporations
- Drawback: ion detected via current across chip
  - ↳ # of ions given off is also hard to differ

Limitations of Emulsion-based Platforms

- relies on Poisson distribution to achieve 1 bead: 1 sequence
  - some beads will have more than one sequence, some will have none
    - hard to scale up
    - "large paint mixer"
- Single detection (e.g. light or H^+) strategy significantly increases insertion / deletion error types
  - ↳ single nucleotide at a time flowed over wells

Illumina Sequencing

- clonal arrays generated on a solid surface (cluster on flow cell surface - no emulsion)

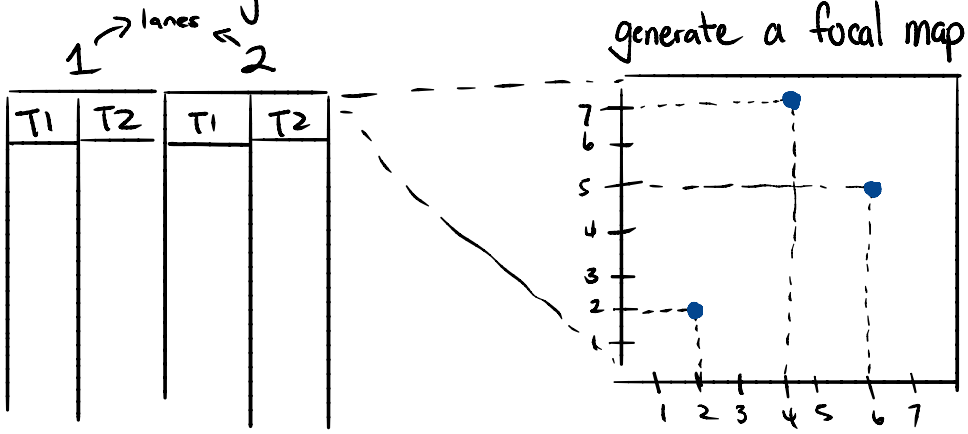


Sequencing By Synthesis (SBS)

- Cycle 1
  1. Add sequencing reagents
  2. First base incorporated → reversible terminators with fluores
  3. Remove unincorporated bases ↳ all bases present (not one kind of base flown at a time)
  4. Detect signal
- Cycle 2-n
  1. Add sequencing reagents and repeat

Base Calling From TIFF Images

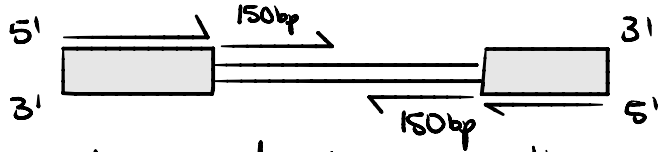
- the identity of each base of a cluster is deduced by analysis of sequential images



- Limitations: if all sequences have the same first nucleotide, every cluster is emitting the same fluore
  ↳ introduce diversity in library to overcome this
- Advantage: sequences homopolymers with high accuracy

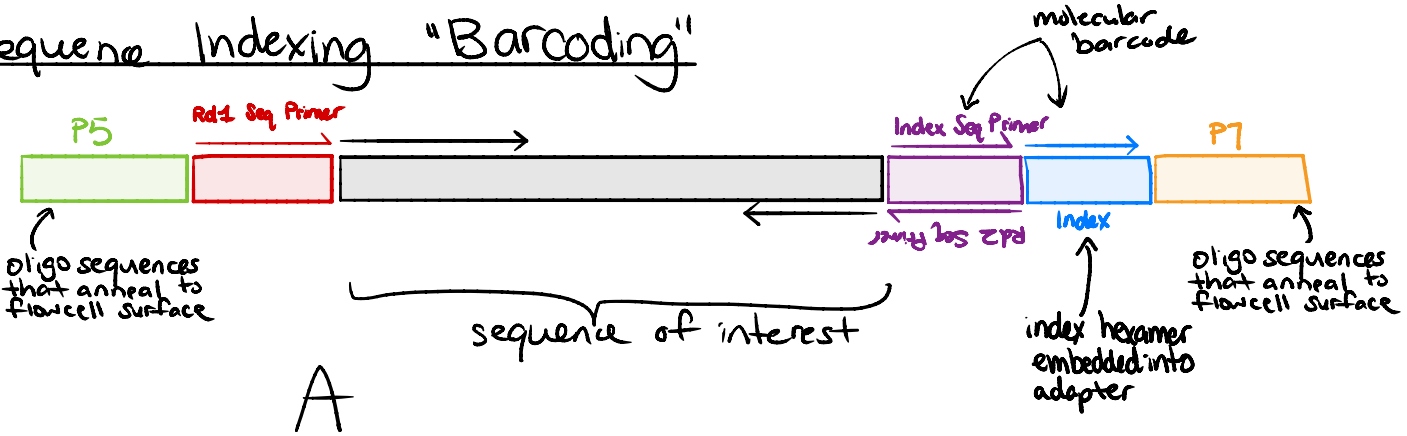
Paired End Read Chemistry

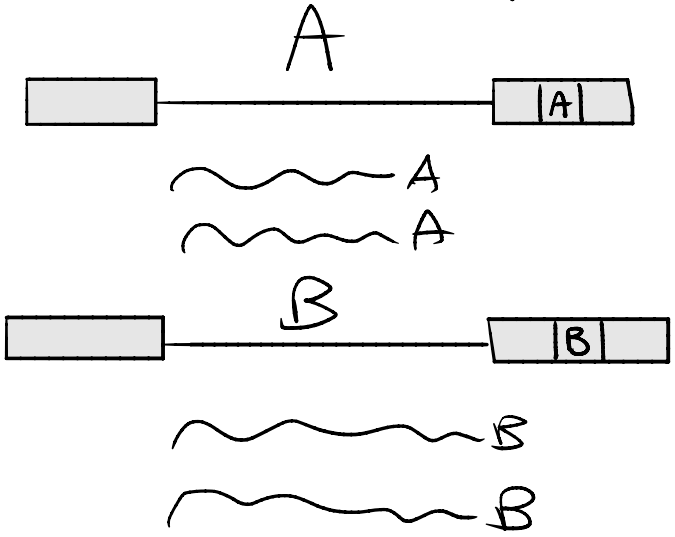
- 300 bp long fragments

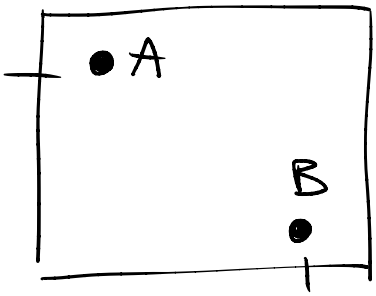


- second strand also uses the same focal map

Sequence Indexing "Barcoding"







- run 2 different samples on one lane
- associate sequenced reads with index barcode

Illumina HiSeq Base Calling

- Quality Considerations
  - phasing/prephasing
  - Chastity

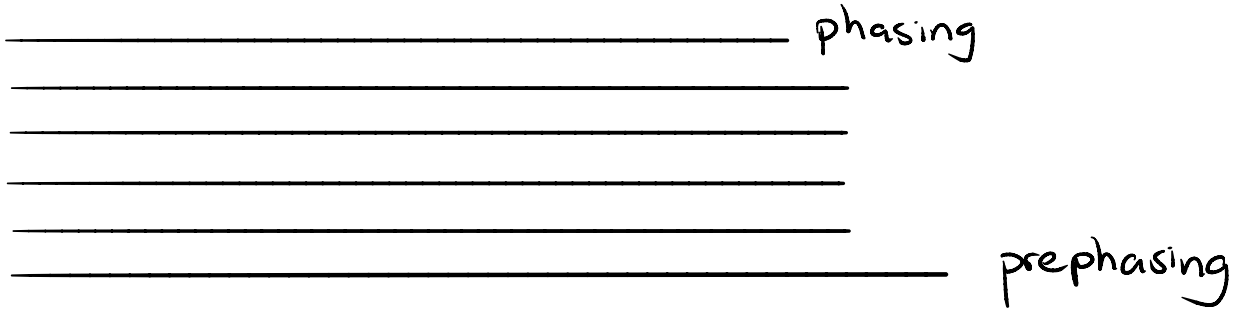
TB of Images ⟶ Intensity Files ⟶ 3.6 TB of Basecalls and qualities (.qseq files)

Base-calling - phasing/prephasing

- prephasing: molecules in each cluster running ahead } the current
- phasing: molecules in each cluster falling behind } incorporation cycle

→ mitigate effect by applying corrections in base calling step

- use statistical averaging over many clusters and sequences to estimate the correlation of signal between different cycles



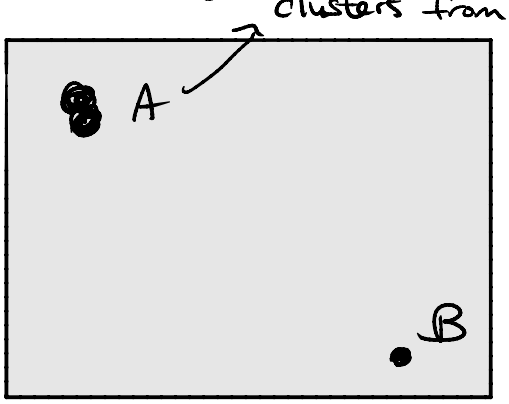
- some nucleotides extra or missed by polymerase
- becomes so significant in difference
  ↳ why the fragment length is STILL a limitation

Base-calling - Chastity

$$\frac{\text{Brightest Intensity}}{\text{Brightest Intensity} + \text{Second Brightest Intensity}} >= 0.6$$

- over the first 25 bases, 1 failure is allowed

→ Flags polyclonal clusters



ASCII Encoding

Base 33

If @ , @ → DEC = 64 , 64 - 33 = Q31

→ QSEQ file contains base of reads to pass onto aligner if using quality scores

FASTA file

K → G/T (keto)
M → A/C (amino)
B → G/T/C
V → G/C/A
S → G/C (strong)
W → A/T (weak)
D → G/A/T
Y → T/C (pyrimidine)
R → G/A (purine)
H → A/C/T
- → gap of indeterminate length

* less than 120 chars/line (80 for NCBI)

FASTQ file

Four lines:

1. Begins with @ followed by a sequence identifier (must be unique) and optional description
2. Raw sequence letters
3. Begins with '+' character followed by an optional identifier (identical to 1)
4. Encodes the ASCII base 33 quality values for the sequence in Line 2 and must contain the same number of symbols as letters in the sequence

→ R1 and R2 are generated as independent FASTQ files
- header

@INSTRUMENT:RUN:FLOWCELL:LANE:TILE:X:Y:READ:FILTERED Y/N:CTRL:IDX

FASTQC

- provides a standardized set of analysis tools for examining the quality of a massively parallel dataset

```
$ fastqc R1.fq -o &
```

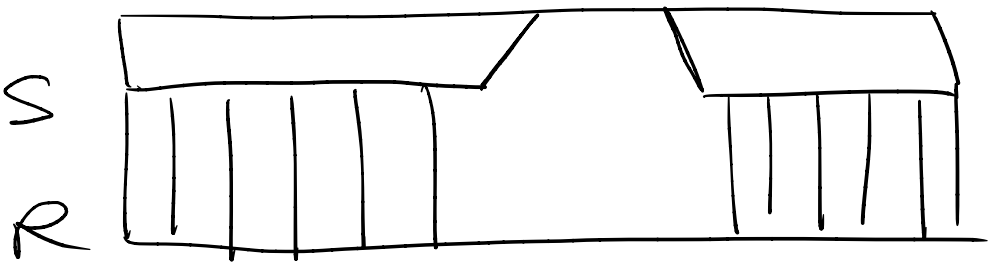
Third Generation Sequencing

- True single molecule sequencing - no clonal amplification (limitation of NGS)
- Very long read lengths possible (longest: 2Mb)
- high error rates (10%)

Lecture 4: Sequence Alignment

Similarity Searching

- alignments can be global or local (algorithm specific)
- Global alignment:
  - an optimal alignment that includes ALL characters from each sequence (e.g. Clustal, MSA)
  - place all characters of a string to the reference



- Local alignment
  - an optimal alignment that includes only the most similar local region or regions (e.g. BLAST)

- if aligned along the whole strand, global and local alignment will give the same solution

Broad Classification

- Sequence similarity searching with ranked solutions
  ↳ one against many (e.g. BLAST)
- sequence similarity searching → OPTIMAL SOLUTION
  ↳ many against ONE (e.g. BWA)

Sequence Similarity Searching Statistics

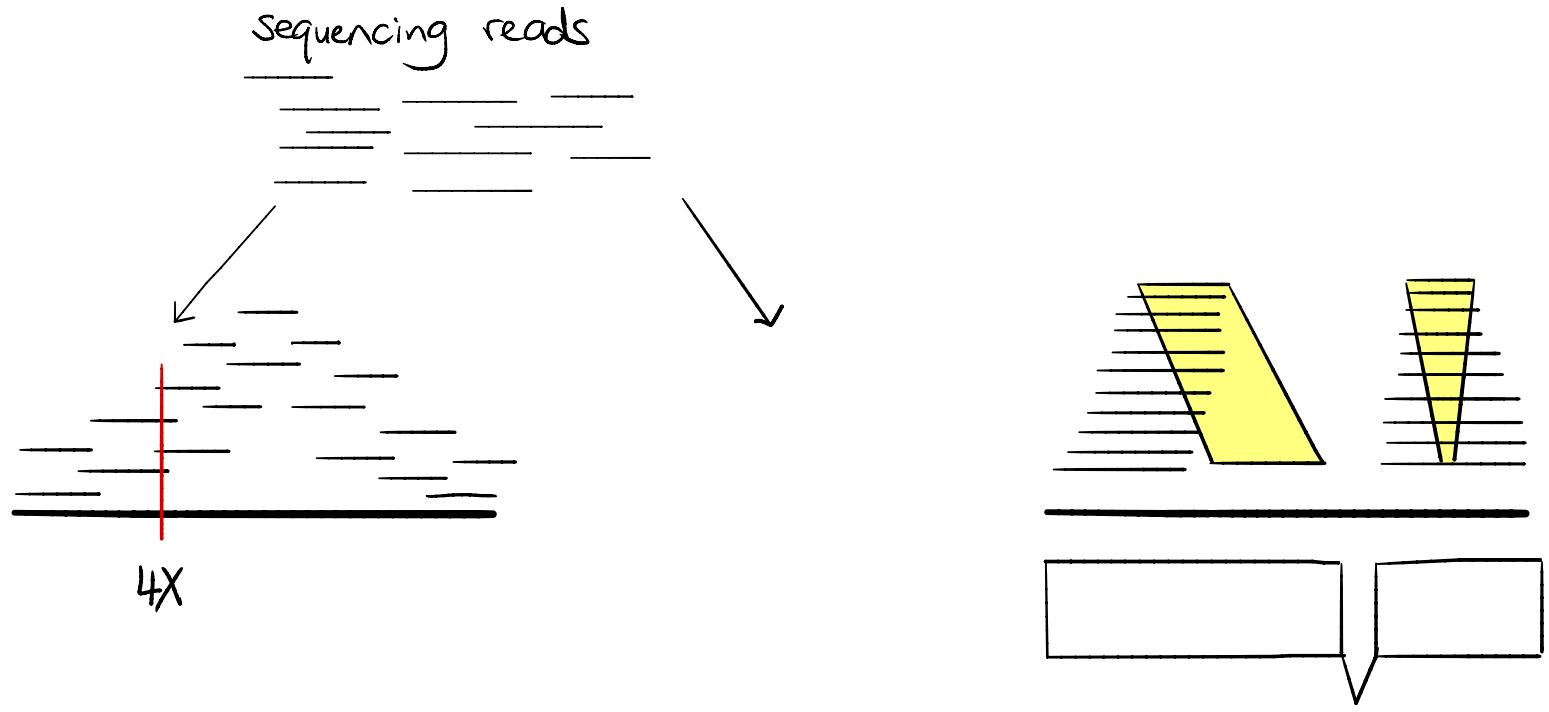
- Discriminating between real and artifactual matches is done using an estimate of probability that the match might occur by chance

- S (sum of matches and mismatches/gaps generates the expect score) and E-value are associated with BLAST hits
  - how well a sequence aligns to a similar sequence compared to aligning to a random sequence
- BWA reports a phred-like mappability probability

Know Your Reagents

- changing your reference (i.e. database in BLAST) is changing your search space
  - scores from different references are NOT comparable
- References impact alignment statistics
  - record...
    - BLAST parameters
    - database choice
    - database size

Sequence Alignment



Heuristic In Computational Biology

- heuristic: produce a "good enough" solution in a reasonable time frame
- solution may not be the best of all the actual solutions, but valuable since it does not require a long time

→ compromise time for optimality

Heuristic Balancing Act

- trading optimality, completeness, accuracy, or precision for compute time and resources

1. Optimality
   - if a read maps to multiple regions in the genome with 1 base mismatch do we need to know which is correct?
2. Completeness
   - if a sequence aligns to HUNDREDS of positions in the genome do we need to know them all?
3. Accuracy and precision
   - provide a way to rank quality of an alignment against other possible solutions
4. Compute time and resources
   - a heuristic may only marginally improve compute time with significant costs

→ GMAP (completeness) vs BWA (speed)

BWA

① to significantly increase the speed of alignment we convert the genome (and/or reads into an indexed table of short "words"

- "words" taken from sequence
  - always start "word" on 5' read (BLAST scans whole sequence)
  - because starts at 5' end, only needs to extend one way (blast extends both ways)

② Converting the reference into 4mer "words"

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

5' ATTAGCATTATGTCCTGA 3'

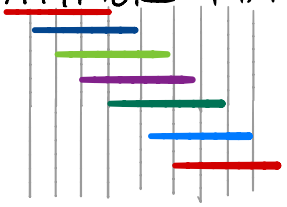
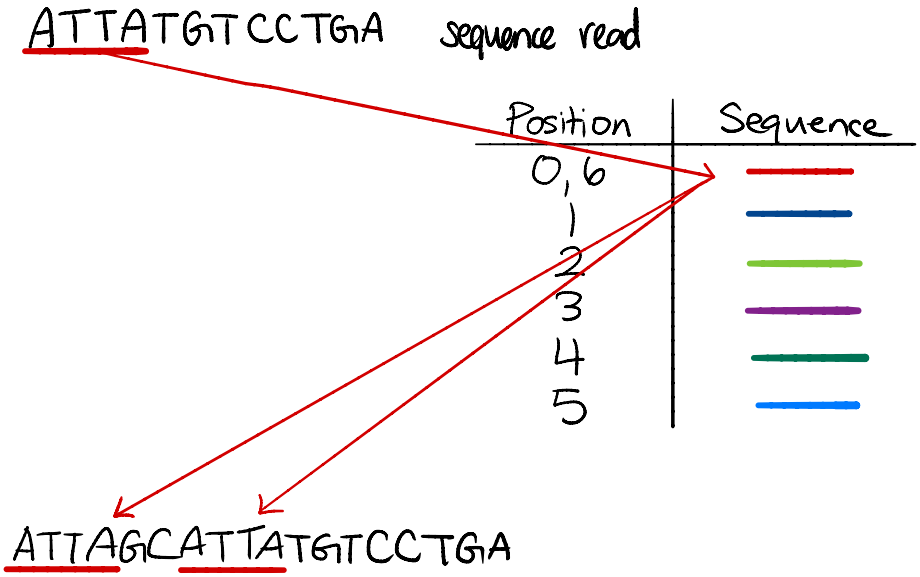


table of sequence words and their location in reference

Position	Sequence
0, 6	
1	
2	
3	
4	
5	

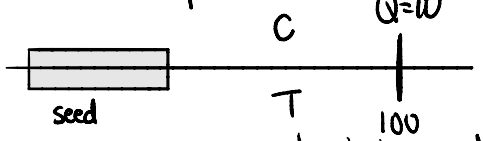
② Extract a 4mer "seed" starting from the 5' end of the sequence read and look it up in the table



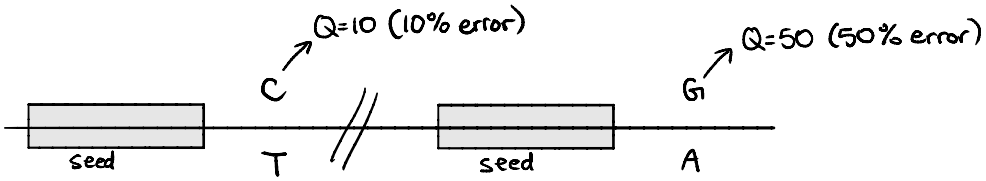
③ Attempt to resolve alignment by seed extension

BWA ALN

1. Reference is converted into an indexed 32mer table
2. The first (5') 32 bases of the sequence read (seed) is extracted and matched to the table
   - up to 2 mismatches are allowed in the seed region
3. Read is extended from seed using a set of adjustable parameters:
   - mismatches in extension allowed up to a set threshold (e.g. sum of base mismatched base qualities)
   - read can be truncated after threshold reached
4. Read is assigned a "mapping quality"
   - hard clip: cut read at that position and report read up to that point



   - soft clip: report whole read but report incident



Mapping Qualities

- A mapping quality is assigned to the read to indicate how confident the aligner is with the read mapping
- mapping qualities are NOT the same as BLAST expect values
  - they quantify the probability that a read is misplaced (i.e. $\frac{1}{1000}$)
- Mapping qualities are derived from base qualities and the number and frequency of mismatches for the best alignment vs. all other possible alignments
- like base qualities, mapping qualities are reported on a Phred scale
- base qualities are used to calculate mapping quality and where they're placed in the genome
- what happens to reads that align to more than 1 place in the reference?
  - assigned a mapping quality of ZERO
  - typically ignored in downstream applications, but in certain cases is randomly assigned one position on genome (with a MQ = 0)
- why does read length matter?
  - it allows us to place reads uniquely

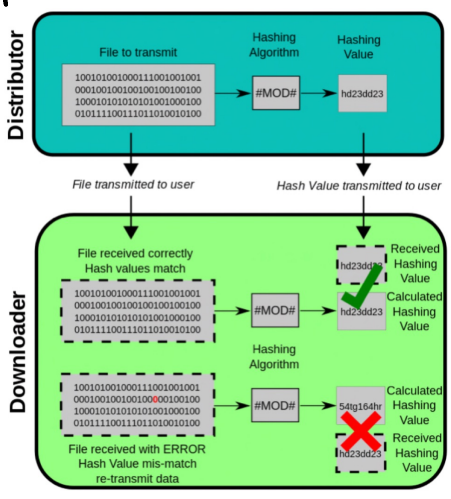
BWA Modules

1. aln (reads <= 75nt)
2. mem (reads >= 75nt)

→ difference in seed length (shorter seed length) and how gaps are handled in the alignment (mem uses S-W)

Digital Fingerprint - md5sum

- md5sum calculates and verifies 128-bit MD5 hashes of a file
- used to confirm the identity and completeness of any file

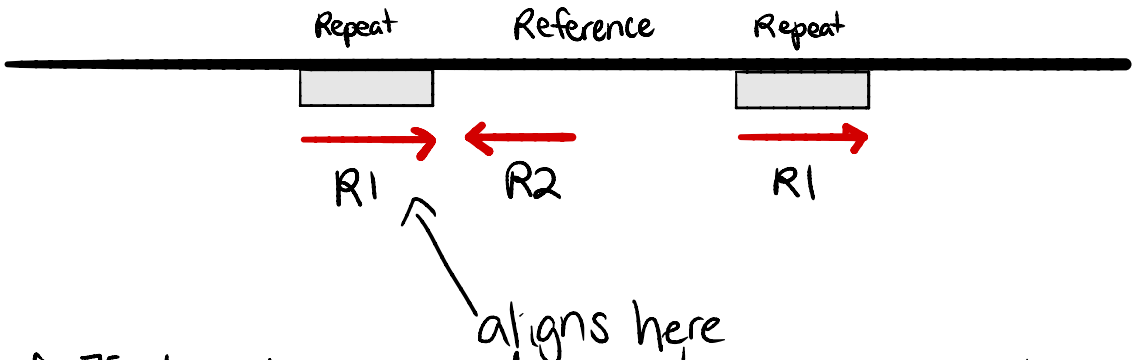


BWA Workflow

1. BWA – index reference
2. Align your reads to the indexed reference
3. Generate alignments from paired-end reads in SAM format

Read Pairs

- read pairs used to 'rescue' sequences aligned into repetitive genomic regions



- 2% of 75nt reads are rescued in a human genome shotgun sequence
- NGS has max fragment size of 600 bp

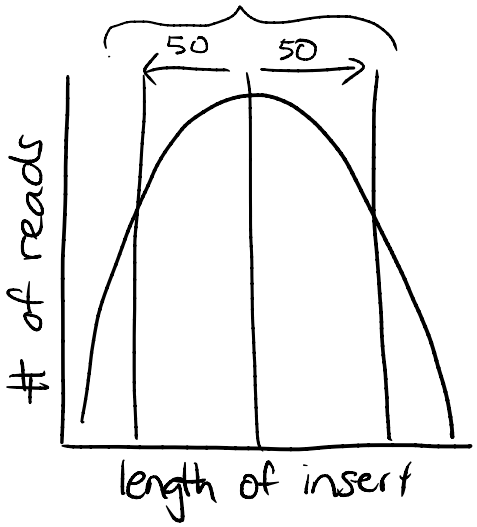
SAM – Mapping File Standard

- generic format for storing large nucleotide alignments
- flexible enough to store all alignment information generated by various alignment programs
- simple enough to be easily generated by alignment programs or converted from existing alignment formats
- allows most of operations on the alignment to work on a stream without loading the whole alignment into memory
- allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus

SAM/BAM – Header

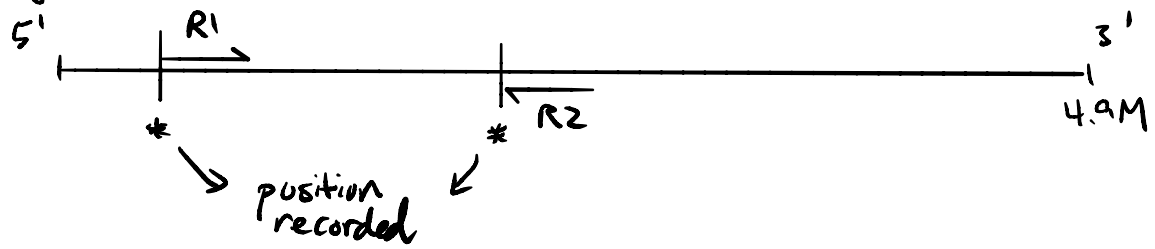
@SQ	SN: reference sequence name	LN: reference sequence length
	AS: genome assembly identifier	M5: MD5 checksum
	SP: species	UR: URI of sequence
@PG	ID: program record identifier	PN: Program name
	CL: Command line	

properly paired → read pairs aligned within insert size distribution



determine mean of insert size and standard deviation

with random fragmentation of genome, distribution of insert sizes should be normal distribution



position recorded

Lecture 5 - ChIP-seq

Histone Modification

- one type of epigenetic modification that acts to reinforce open or closed chromatin conformations

Modification	Effect
H3K4me3	Active Promoters
H3K4me1	Enhancers
H3K27ac	Active Enhancers
H3K36me3	Elongating Transcription
H3K9me3	Repressive / heterochromatin
H3K27me3	Repressive

ChIP-seq

- sequencing based approach used to measure histone modification patterns in genome
- histone modifications have very different patterns in the genome
  - studying these patterns can help to decode the regulatory state of a genome
- data is typically visualized using a genome browser
  - IGV can be used to visualize ChIP-seq datasets locally as BED files
    - bigBEDs → binary, indexed BEDs → load directly into UCSC

ChIP-seq Considerations

- antibody specificity and sensitivity (antibody recognizes histone modification)
- which marks (i.e. histone markers) should I profile
- sequencing depth (IHEC recommendations)
  - 50M read pairs for puncate marks (e.g. H3K4me3)
  - 100M read pairs for broad marks (e.g. H3K9me3)
- many potential biases (even more than genome analysis) in ChIP-seq analysis
  - garbage in is garbage out

ChIP-seq QC

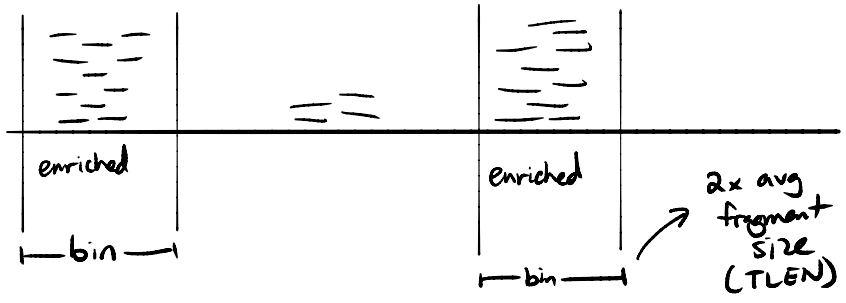
1. FASTQC
   - insert size distribution
   - diversity of fragments in library
2. Library Diversity
   - PCR duplicates occur with library amplification (NOT clonal)
     - reads that have identical post+ pos- coordinates
   - more PCRdups = poor quality library
3. IP quality

ChIP-seq Workflow

Operation	H3K27ac and Input	File Type
	↓	.fastq
Align and pair sequence reads	BWA mem	.sam
	↓	
Convert to binary format, sort and mark duplicates	SAMBAMBA/ SAMTOOLS	.mkdup.sorted.bam
	↓	
Generate density tracks and call enriched regions	MACS2	.wig .bed

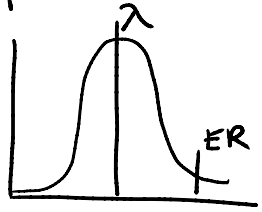
MACS2

- identify regions of enrichment comparing a treatment (IP) to a control (input)
- find significantly enriched bins (2 x average fragment size) with counts 'mfold' higher than random genome average



- for 1000 randomly chosen enriched bins, calculate the difference between max of the distribution read starts on (+) and (-) strands = d
- shift all (+) read starts by + d/2 and (-) reads by -d/2
- scale control experiment to the same number of reads as ChIPseq
- read start distributions within the 1000 randomly selected enriched bins are used to shift reads towards the centre
- the idea here being the nucleosome inhibits DNA shearing at the true peak centre
- assume Poisson distribution for read count distribution and scan genome with 2d bins
- for each bin, calculate the mean number of sequence reads from your IP (ie. treatment) that align within it

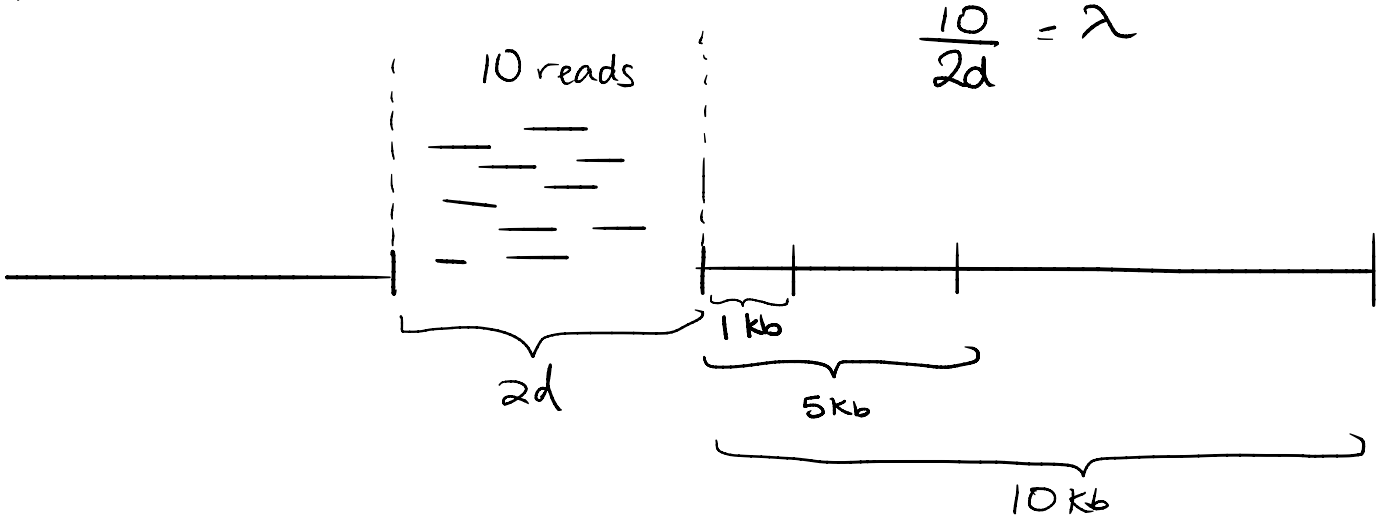
$$\frac{\#\text{ reads}}{\text{length of bin in nucleotides}} = \lambda$$



- calculate mean number of mean number of sequence reads from INPUT (ie control) for the next 1Kb, 5Kb, 10Kb bin and genome-wide
  - maximum value is your Poisson lambda
- Using Poisson distribution or Lambda, calculate P-value using IP mean (average # of reads that align in the bin)
- Estimate empirical False Discovery Rate (FDR) for each bin

d = space between read starts

$$\frac{10}{2d} = \lambda$$



- inputs:
  - used ChIP-seq data (e.g. Treatment) alone, or a Control (INPUT) sample to increase specificity of peak calls
  - REQUIRED: treatment file (only requirement) — BAM/BED (format detected using first treatment file provided)
  - Effective Genome Size — mappable genome size
    - genome build and read length dependent (larger fraction for longer reads)
- outputs:
  - a BED file of peaks (default)
  - a tabular file of peaks (default)
  - two bedGraph files of scores, for treatment and control lambda