SUMMER RCN BIOINFORMATICS MENTORING 2021

SEQUENCING DATA AND QUALITY CONTROL

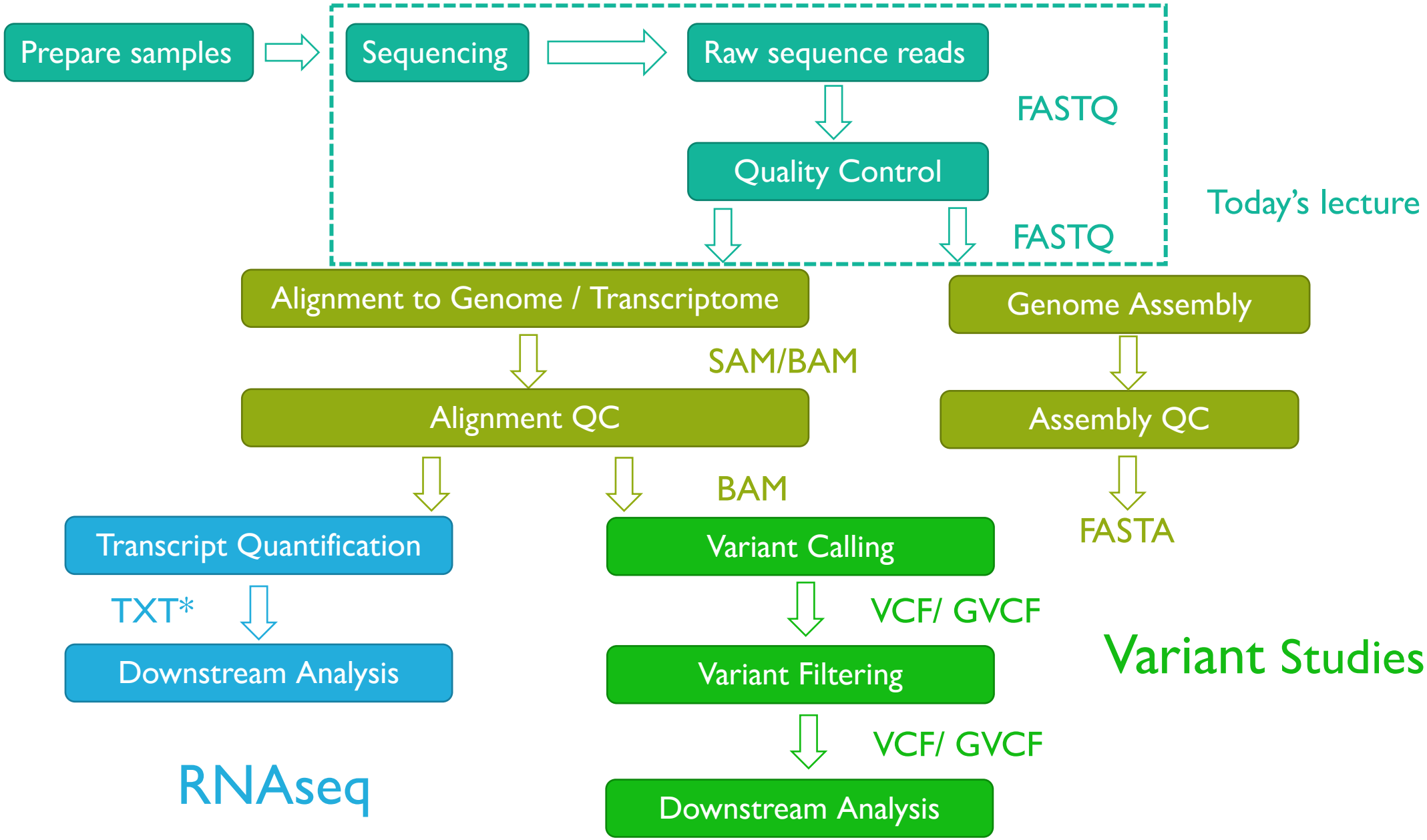
OUTLINE

Lecture

- Sequencing basics
- Sequence Formats
- Quality Metrics
- Quality Control with FastQC

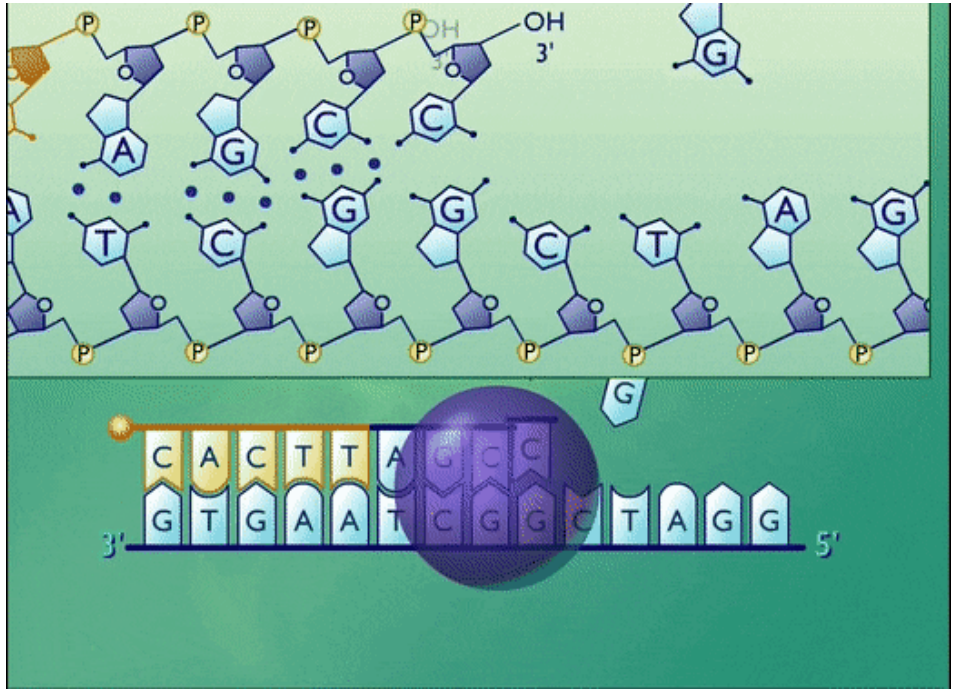
Tutorial

- FastQC
- Adapter Trimming


Prepare samples
Sequencing
Raw sequence reads
FASTQ
Quality Control
FASTQ
Today's lecture
Alignment to Genome / Transcriptome
Genome Assembly
SAM/BAM
Alignment QC
Assembly QC
BAM
FASTA
Transcript Quantification
Variant Calling
TXT*
VCF/ GVCF
Downstream Analysis
Variant Filtering
Variant Studies
VCF/ GVCF
Downstream Analysis
RNAseq

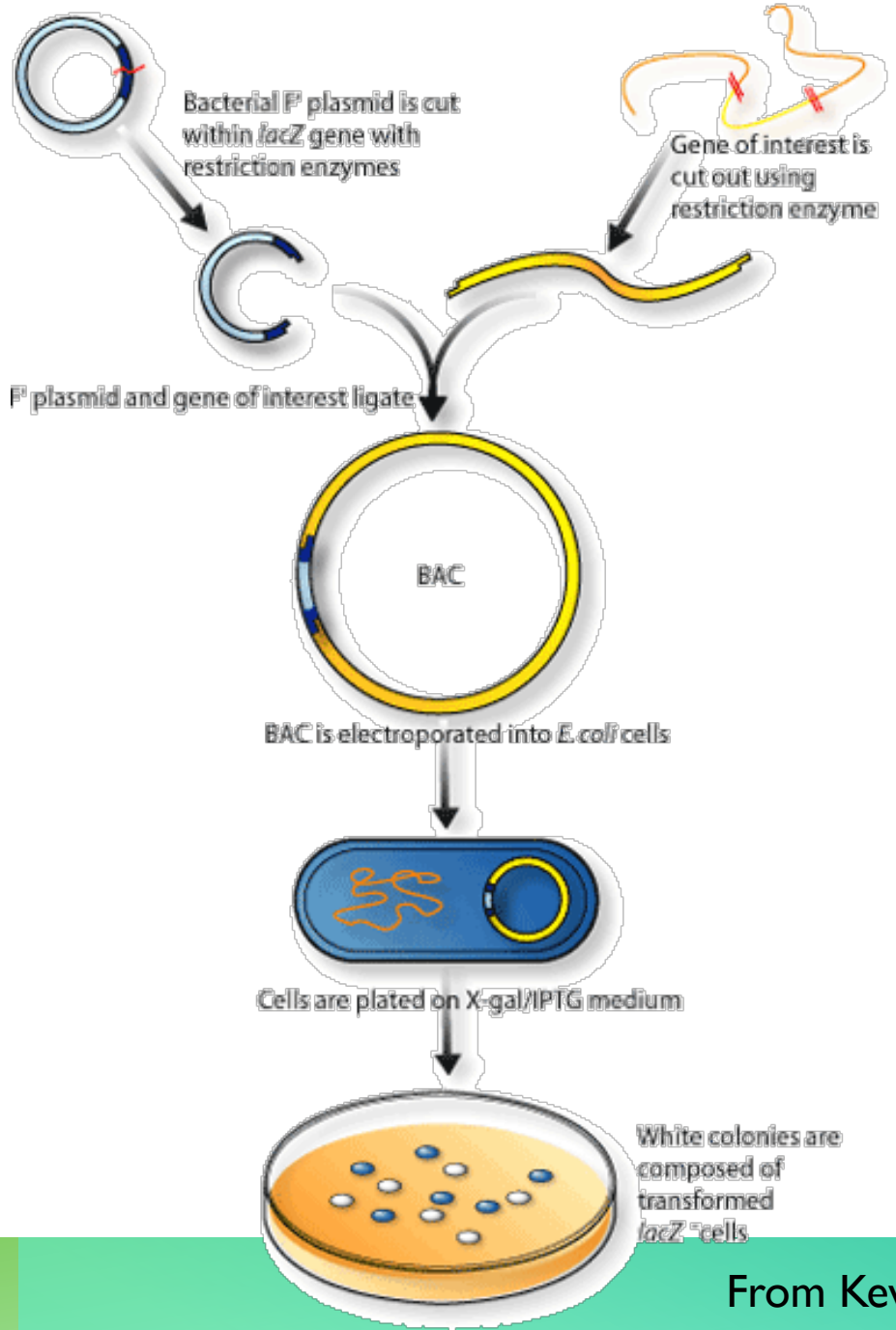
SEQUENCING

- Enzymatic reaction on adding complementary nucleotide
- Needs a “priming sequence” to start



HOW DO YOU GET PRIMED DNA?

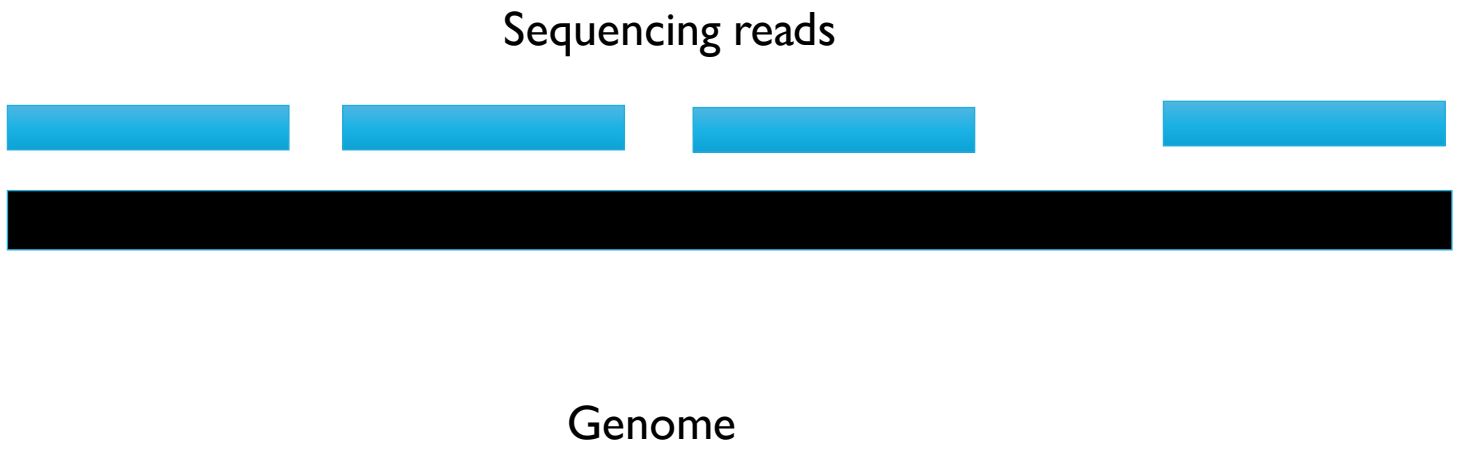
- *A priori* sequence knowledge: use Polymerase Chain Reaction(**PCR**) with Primers to select a DNA region
- In the old days if sequences weren't known then would use **cloning**



From Kevinshe (2004) Science Creative Quarterly

SHOTGUN SEQUENCING

- Shearing the whole genome
- Put adapters & barcodes on either end (synthetic primers) of sequence

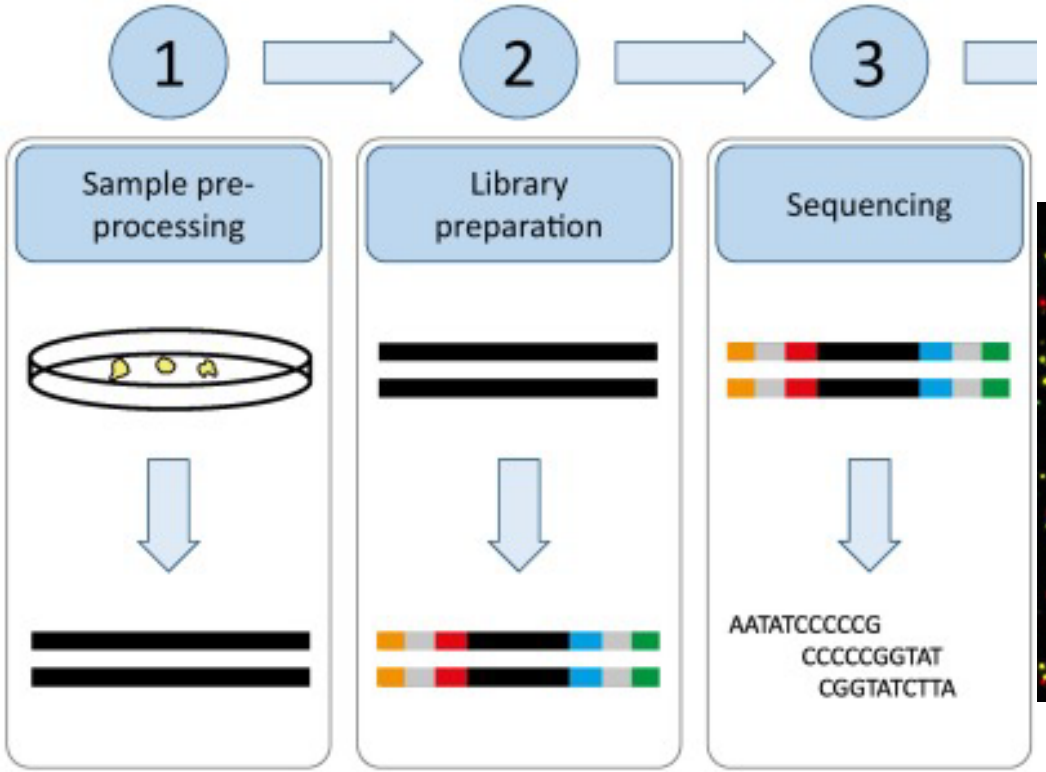


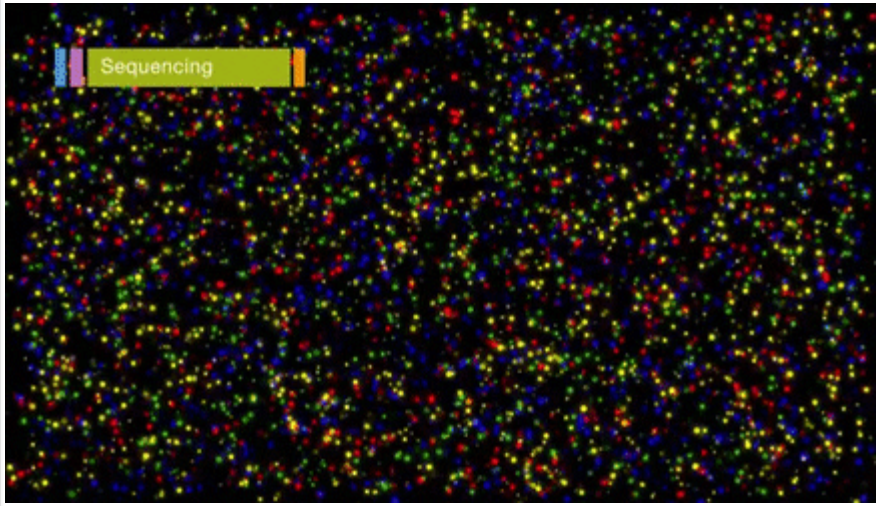
HIGH COVERAGE FILLS THE GAPS

Sequencing reads

Genome

TYPICAL NGS WORKFLOW





Workflow figure from Hess et al 2020 Biotech. Adv. 41: 107537

TYPICAL NGS WORKFLOW

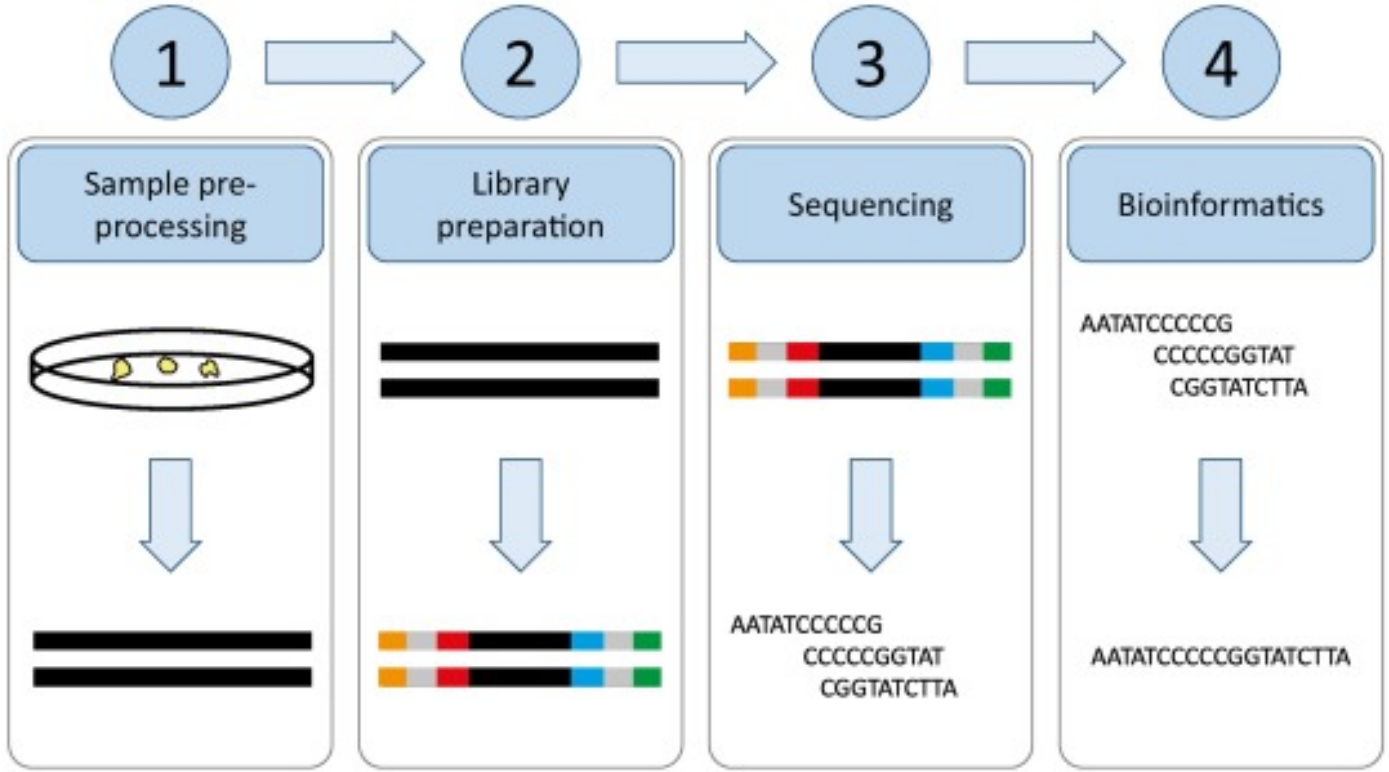
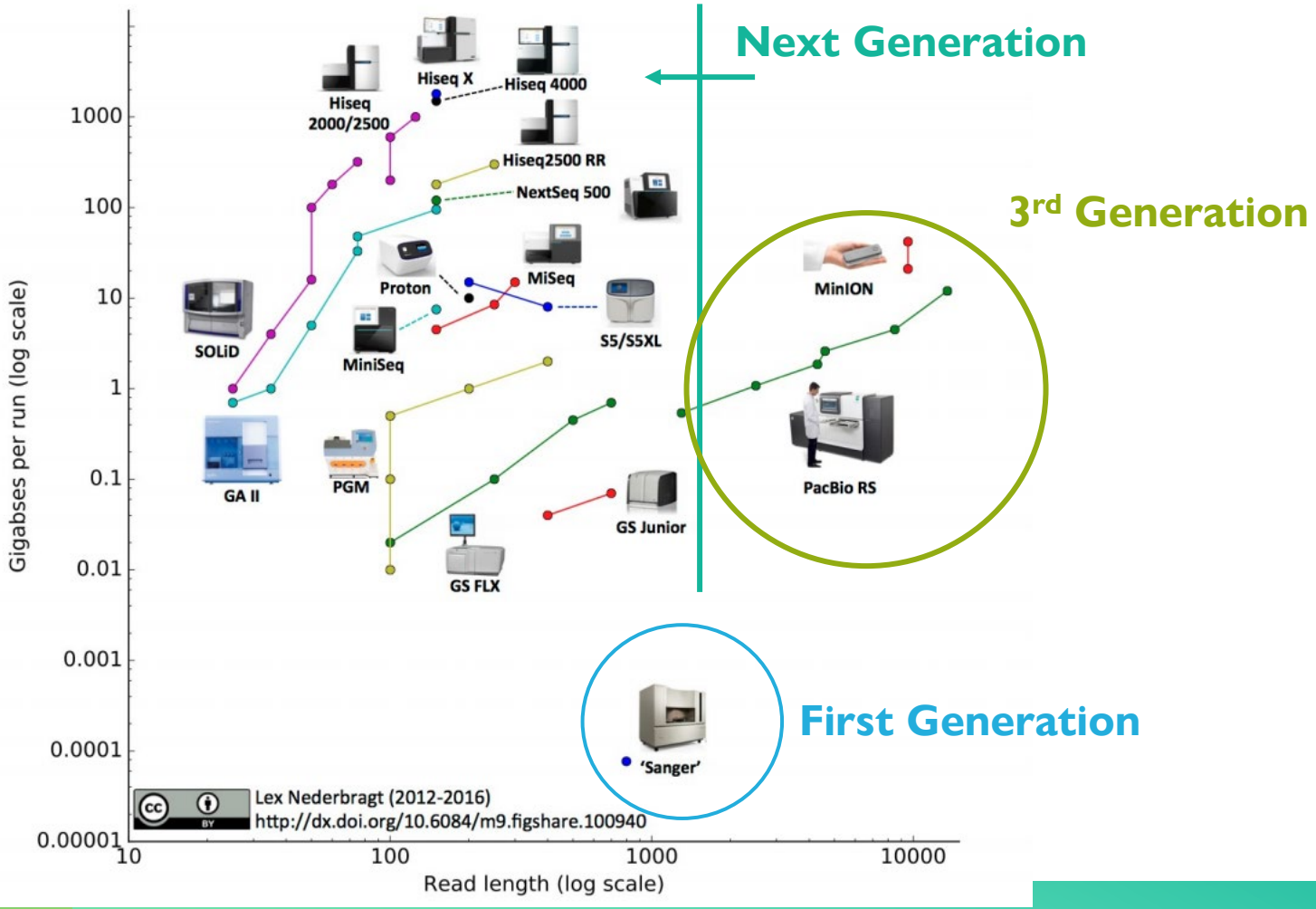


Figure from Hess et al 2020 Biotech. Adv. 41: 107537

"NEXT GENERATION" SEQUENCING

	Sanger Sequencing	"NGS"
Sequences Per Rxn	1 clone	Millions of molecules
Rxn per run	384	Millions
Sequence Quality	High	Low
Sequence Length	600-800 bp	35-2000
Cost per bp	High	Low (and decreasing)

SEQUENCING


Next Generation
3rd Generation
First Generation
Hiseq X
Hiseq 4000
Hiseq 2000/2500
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
SOLiD
MiniSeq
S5/S5XL
MinION
GA II
PGM
PacBio RS
GS Junior
GS FLX
'Sanger'
Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1000
10000
Read length (log scale)
Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940

LEADERS IN 2020

	Illumina (2nd gen)	PacBio (3rd Gen)
Sequence Length	150 bp	Up to 25kb
Error Rate	"Low"	"High" – but much better now
Price per Gbp	$7-$93	$100-$200
Different types of questions		

SEQUENCE ASSEMBLY

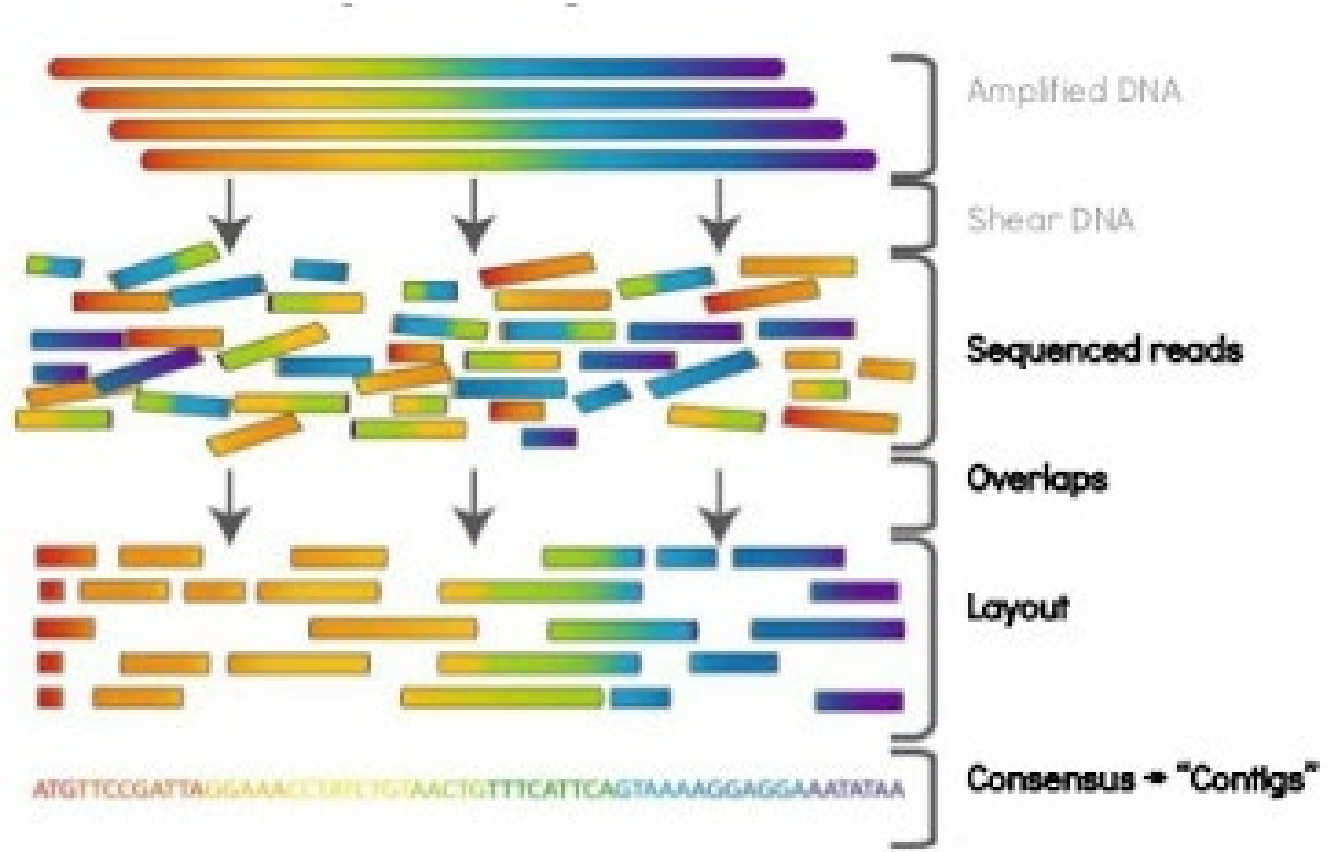
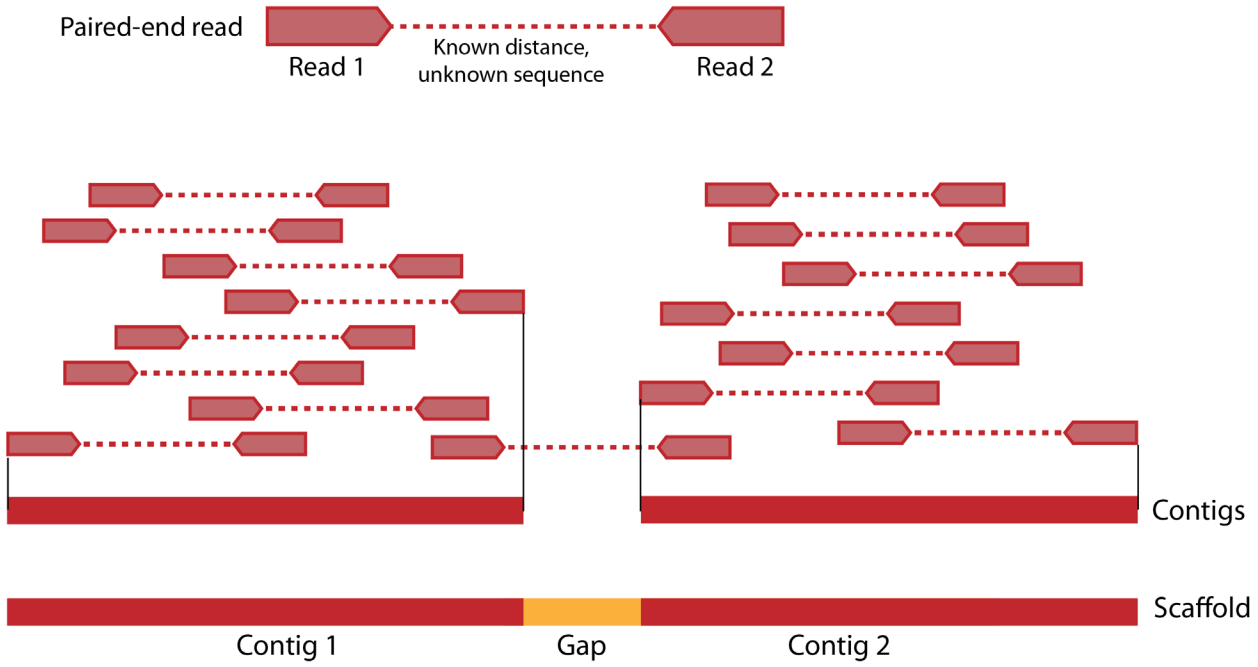


Figure from cromnet.net

GENOME ASSEMBLY

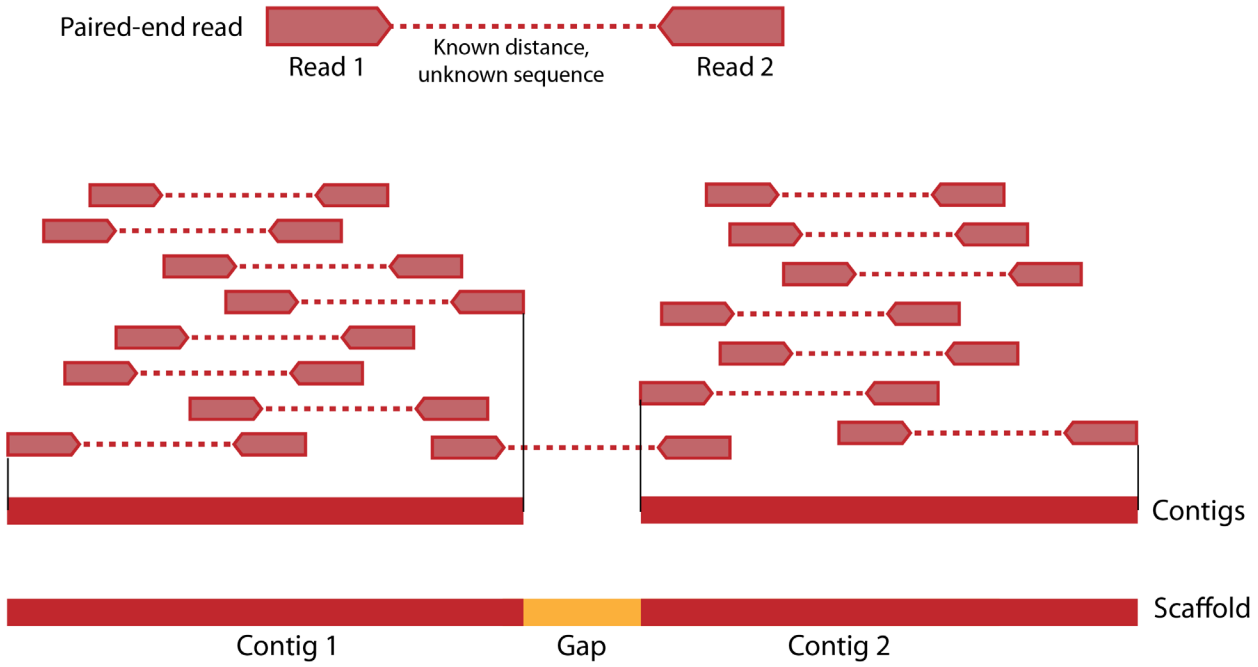
- N50 – Genome Contiguity – half of the genome is covered by contigs >= the N50 contig size.



Check out this explanation: https://www.molecularecologist.com/2017/03/whats-n50/

SCAFFOLDING

- Contig – high confidence of sequences
- Can then be arranged into Scaffolds e.g. using linkage information from the sequencing data itself



WHERE IN THE GENOME?

- **Genetic Mapping** – relative positions (cM) from pedigrees etc
- **Physical Mapping:**
  - Restriction Mapping
  - Florescent *in situ* Hybridisation (FISH)
  - Sequence Tagged Site (STS) Mapping
- **Sequencing information**
  - Subcloning
  - Long-read data
  - Linkage from PE short-read
- Synteny mapping from other species

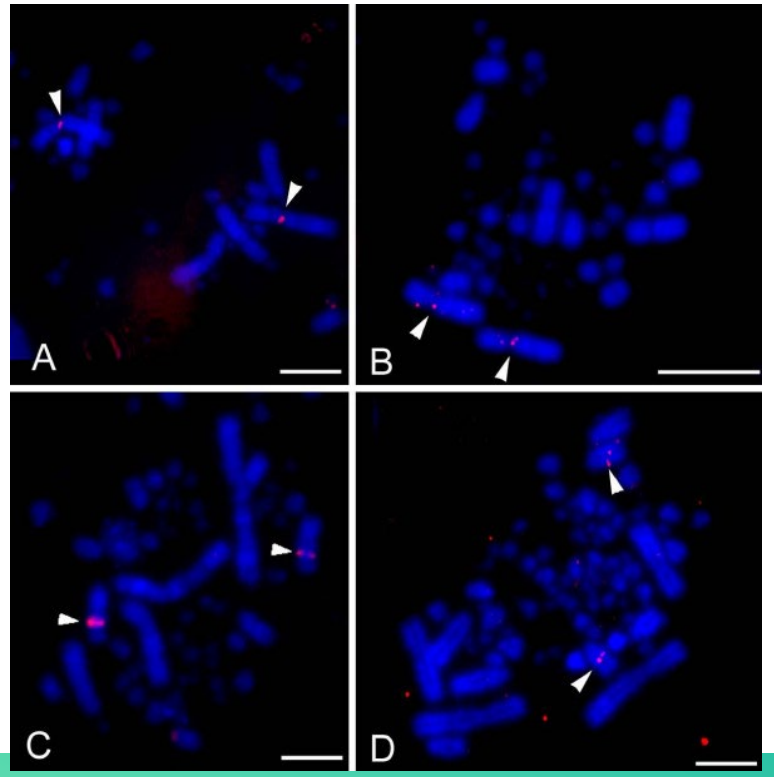


Figure Skinner et al. (2009) BMC Genomics

DIFFICULT SEQUENCES

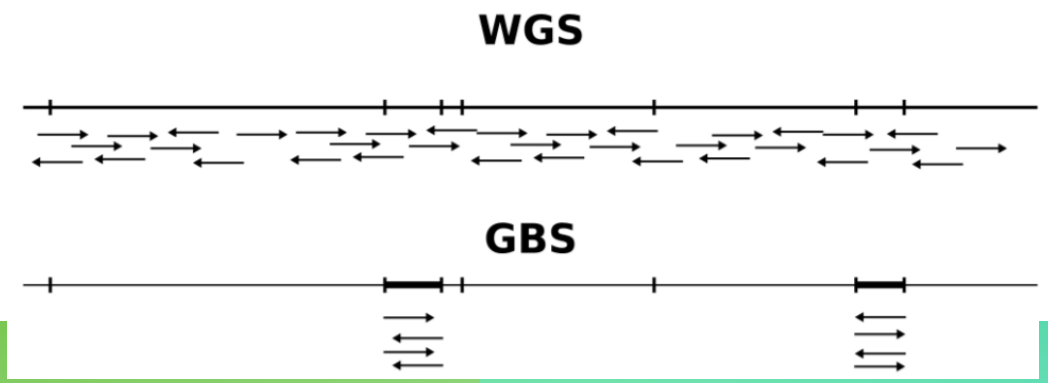
- High GC content (Library Prep & Sequencing)
  - >45% often difficult to obtain sequence for
- Repeated sequences (Bioinformatic)
  - E.g. short repeats (microsatellites)
  - Segment or gene duplications

REDUCING GENOME COMPLEXITY

- Trade off between depth and coverage
- Only Interested in Some Bits
- Cost

RADSEQ AND GENOTYPING-BY-SEQUENCING

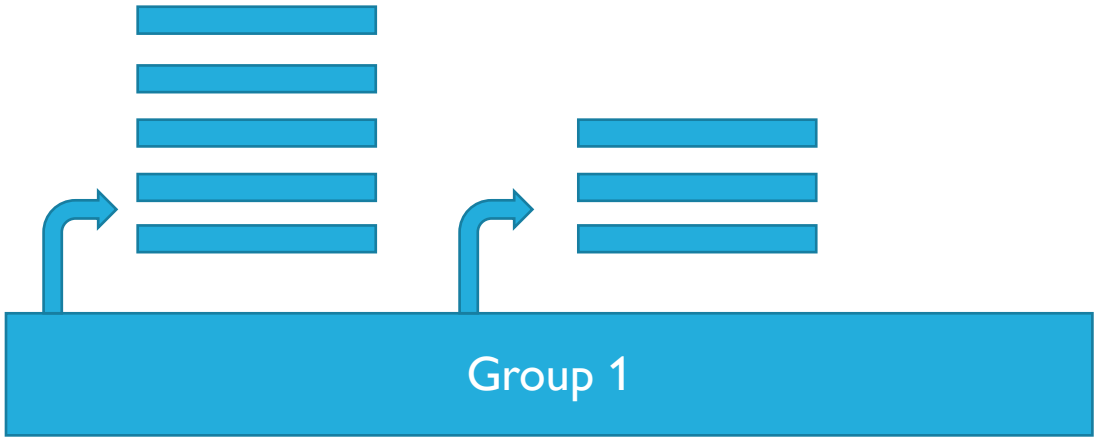
- RADseq, ddRAD, GBS, ezRAD, 2bRAD
- Cheap
- No Genome Required
- Usually only for SNPs
- Biases: dropout, PCR duplicates, coverage variance

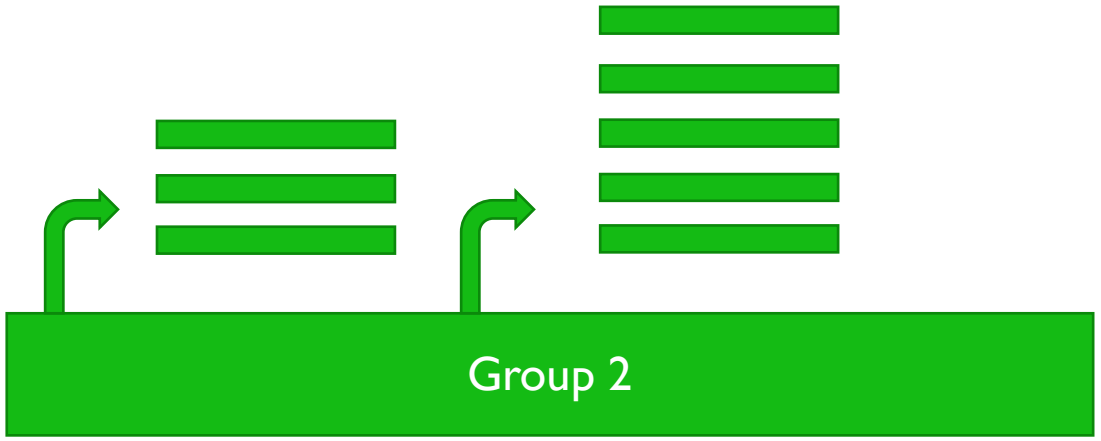


Andrews et al (2016) *Nature Reviews Genetics:* 17: 81-92

RNA-SEQ

- Only sequences gene exons
- Well understood methodologies and bioinformatics
- Can be used to understand gene expression

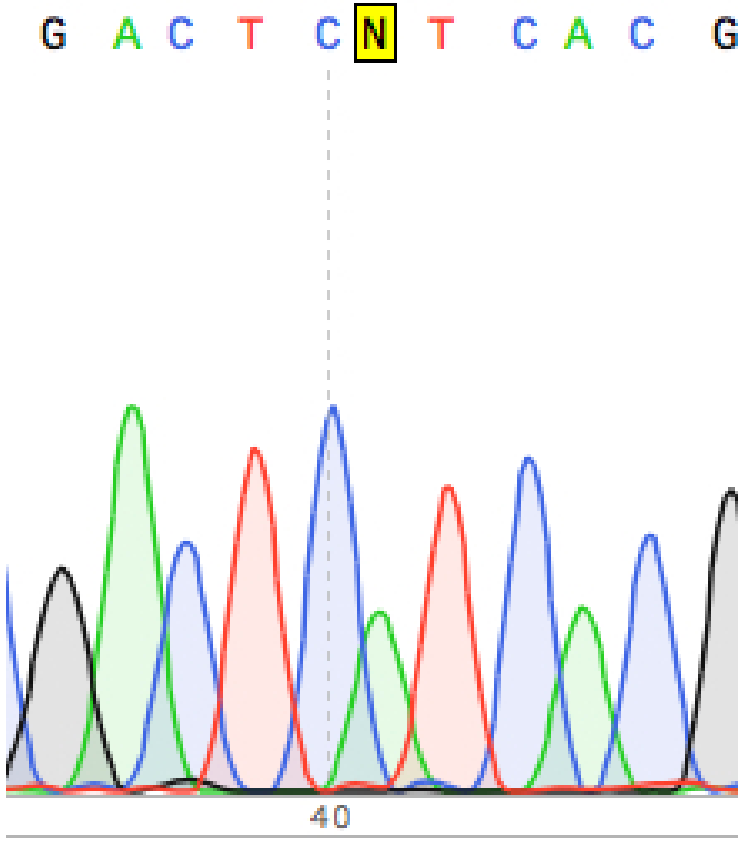




SO NOW YOU HAVE SOME DATA…

HOW DO COMPUTERS REPRESENT DNA SEQUENCES?

- Base information (ATG or C)
- Base Position
- Signal strength/confidence in base
- Where did the sequence come from?



FASTA FORMAT

- Nucleic acid or amino acid sequences
- File extension .fasta, .fas, .fna, .fa
- Header line “>” + information
- Interleaved or sequential sequences.
- Single vs multi-fasta

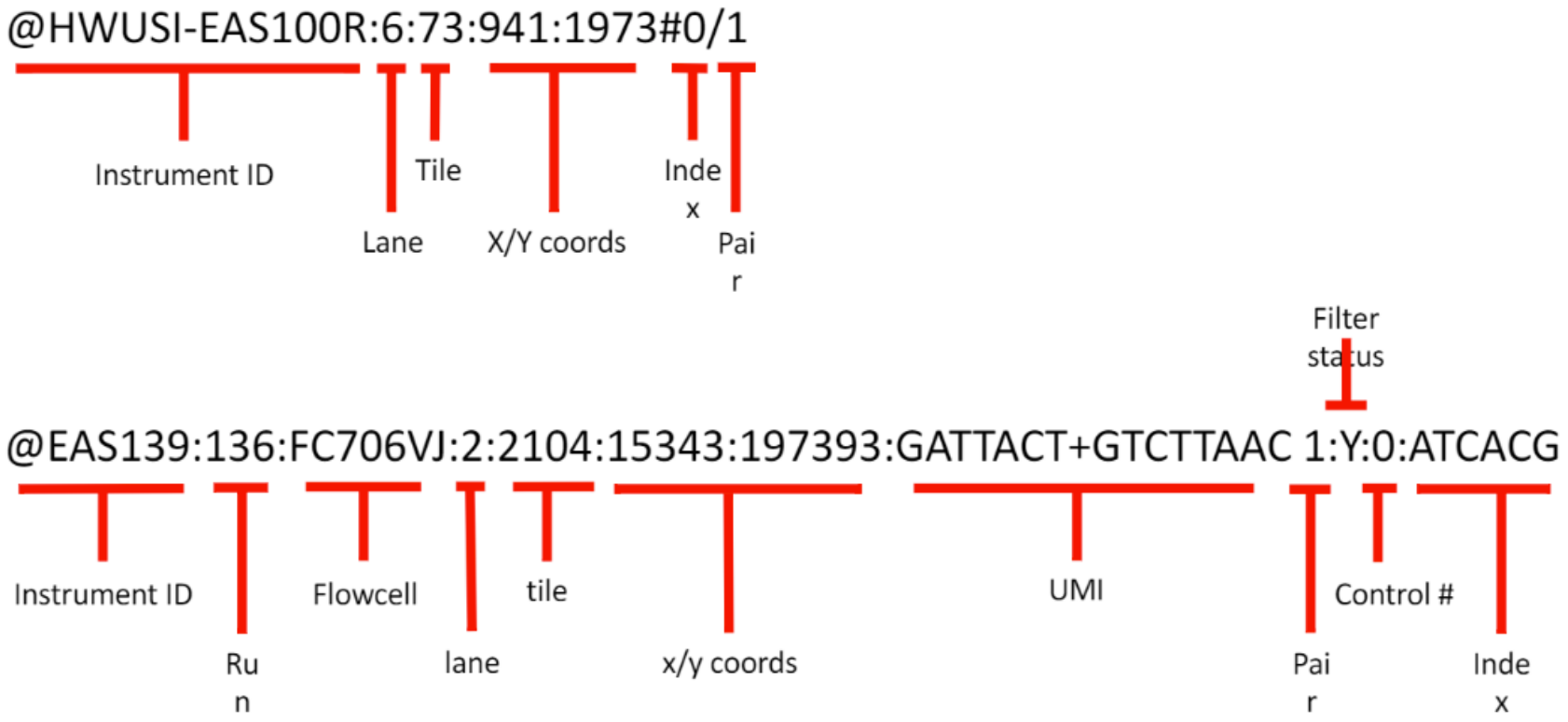
```
>ENSP00000354687 pep:known chromosome:GRCh37:MT:3307:4262:1 gene:ENS
MPMANLLLLIVPILIAMAFLMLTERKILGYMQLRKGPNVVGPYGLLQPFADAMKLFTKEP
LKPATSTITLYITAPTLALTIALLLWTPLPMPNPLVNLNLGLLFILATSSLAVYSILWSG
WASNSNYALIGALRAVAQTISYEVTLAIILLSTLLMSGSFNLSTLITTQEHLWLLLPSWP
LAMMWFISTLAETNRTPFDLAEGESELVSGFNIEYAAGPFALFFMAEYTNIIMMNTLTTT
IFLGTTYDALSPELYTTYFVTKTLLLTSLFLWIRTAYPRFRYDQLMHLLWKNFLPLTLAL
LMWYVSMPITISSIPPQT
>ENSP00000355046 pep:known chromosome:GRCh37:MT:4470:5511:1 gene:ENS
MNPLAQPVIYSTIFAGTLITALSSHWFFTWVGLEMNMLAFIPVLTKKMNPRSTEAAIKYF
LTQATASMILLMAILFNNMLSGQWTMTNTTNQYSSLMIMMAMAMKLGMAPFHFWVPEVTQ
GTPLTSGLLLLTWQKLAPISIMYQISPSLNVSLLLTLSILSIMAGSWGGLNQTQLRKILA
YSSITHMGWMMAVLPYNPNMTILNLTIYIILTTTAFLLLNLNSSTTTLLLSRTWNKLTWL
TPLIPSTLLSLGGLPPLTGFLPKWAIIEEFTKNNSLIIPTIMATITLLNLYFYLRLIYST
SITLLPMSNNVKMKWQFEHTKPTPFLPTLIALTTLLLPISPFMLMIL
>ENSP00000354499 pep:known chromosome:GRCh37:MT:5904:7445:1 gene:ENS
MFADRWLFSTNHKDIGTLYLLFGAWAGVLGTALSLLIRAELGQPGNLLGNDHIYNVIVTA
HAFVMIFFMVMPIMIGGFGNWLVPLMIGAPDMAFPRMNNMSFWLLPPSLLLLLASAMVEA
GAGTGWTVYPPLAGNYSHPGASVDLTIFSLHLAGVSSILGAINFITTIINMKPPAMTQYQ
TPLFVWSVLITAVLLLLSLPVLAAGITMLLTDRNLNTTFFDPAGGGDPILYQHLFWFFGH
PEVYILILPGFGMISHIVTYYSGKKEPFGYMGMVWAMMSIGFLGFIVWAHHMFTVGMDVD
TRAYFTSATMIIAIPTGVKVFSWLATLHGSNMKWSAAVLWALGFIFLFTVGGLTGIVLAN
SSLDIVLHDTYYVVAHFHYVLSMGAVFAIMGGFIHWFPLFSGYTLDQTYAKIHFTIMFIG
VNLTFFPQHFLGLSGMPRRYSDYPDAYTTWNILSSVGSFISLTAVMLMIFMIWEAFASKR
KVLMVEEPSMNLEWLYGCPPPYHTFEEPVYMKS
```

FASTQ FORMAT

- Results from next generation sequencing experiments
- Filename usually follows a standard: SampleCode_AdapterSequences.fq
- @header

```
@K00317:53:HGCNKBBXX:7:1101:12581:1525 1:N:0:TCTCGCGC+AGGCGAAG
CGGGGAAAAAAAACCAACAAAAACATTTTGGGCAATATAGGCGGCATTTCGGACCACGACAATGAGCGATTTAAAATGGACATGGGCACTTTCCAGAAA
+
AAFAFFJJJAJFJJJJJJJJJJFJJ7F<<FF<7FJFF<F<<<<AAJ7<<<-<F7-FJ77FA<<<FFF-7<<<<FJJ-<AFFA<<777AA-77<7J-<FF
@K00317:53:HGCNKBBXX:7:1101:16802:1525 1:N:0:TCTCGCGC+AGGCGAAG
AAACCATCCTGAACTTCAAAGGAATTTCTCATTCACGGTCTGCCACAAGTACCGGCCTGGTCATTGCCACTACCTTTCATTGTTGTACTGCCAAGCAGC
+
AAFFFJ7FJAAFFFA7FJJJ-7AFF-AF7<F-7<FJ<--7--7AFJJJA-<FJ-<FJ-7-77<--<<FFF7FFJ---<A<-<77<A<JF-F<<A-7A-7
@K00317:53:HGCNKBBXX:7:1101:19065:1578 1:N:0:TCTCGCGC+AGGCGAAG
CTCTACTGGCTCTCCCACTTGGGTTGTTAGGTTGTTGACAAGATTGCAGACAGGGAGCAGGGTCAGGTTCCAGGAAGGAATGTTGGCGCGGTTAATCGGG
```

FASTQ FORMAT



QUALITY SCORES

- ASCII character code – 33 = Phred Quality Score
- Minimum acceptable Phred = 20 (default in many trimming programs)

```
@K00317:53:HGCNKBBXX:7:1101:12581:1525 1:N:0:TCTCGCGC+AGGCGAAG
CGGGGAAAAAAAACCAACAAAAACATTTTGGGCAATATAGGCGGCATTTCGGACCACGACAATGAGCGATTTAAAATGGACATGGGCACTTTCCAGAAA
+
AAFAFFJJJAJFJJJJJJJJJJFJJ7F<<FF<7FJFF<F<<<<AAJ7<<<-<F7-FJ77FA<<<FFF-7<<<<FJJ-<AFFA<<777AA-77<7J-<FF
```

ASCII	Code	Probability of Wrong Base	Phred Quality Score
!	33	1	0
“	34	0.794	1
+	43	0.1	10
5	53	0.01	20
?	63	0.001	30

For the full list go here:

https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm

Table from rnabio.org

FASTQC

- Program for evaluating sequence quality in all reads in a file
- https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

PER BASE SEQUENCE QUALITY

- Ideal Illumina Data

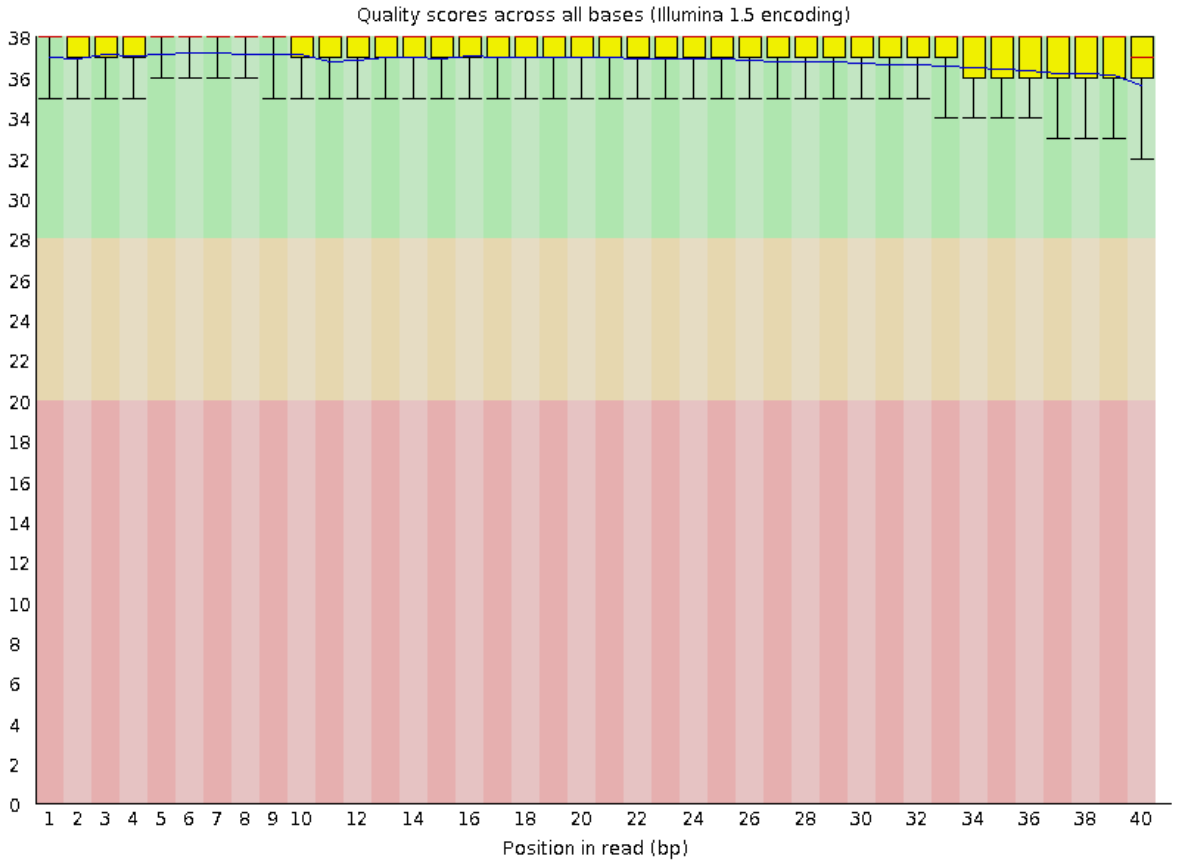


Figure From FastQC documentation

PER BASE SEQUENCE QUALITY

- Ideal Illumina Data … not universal
- Illumina sequence quality decreases along the read length
- Different sequences might have different typical error profiles
- Some can be corrected with quality trimming

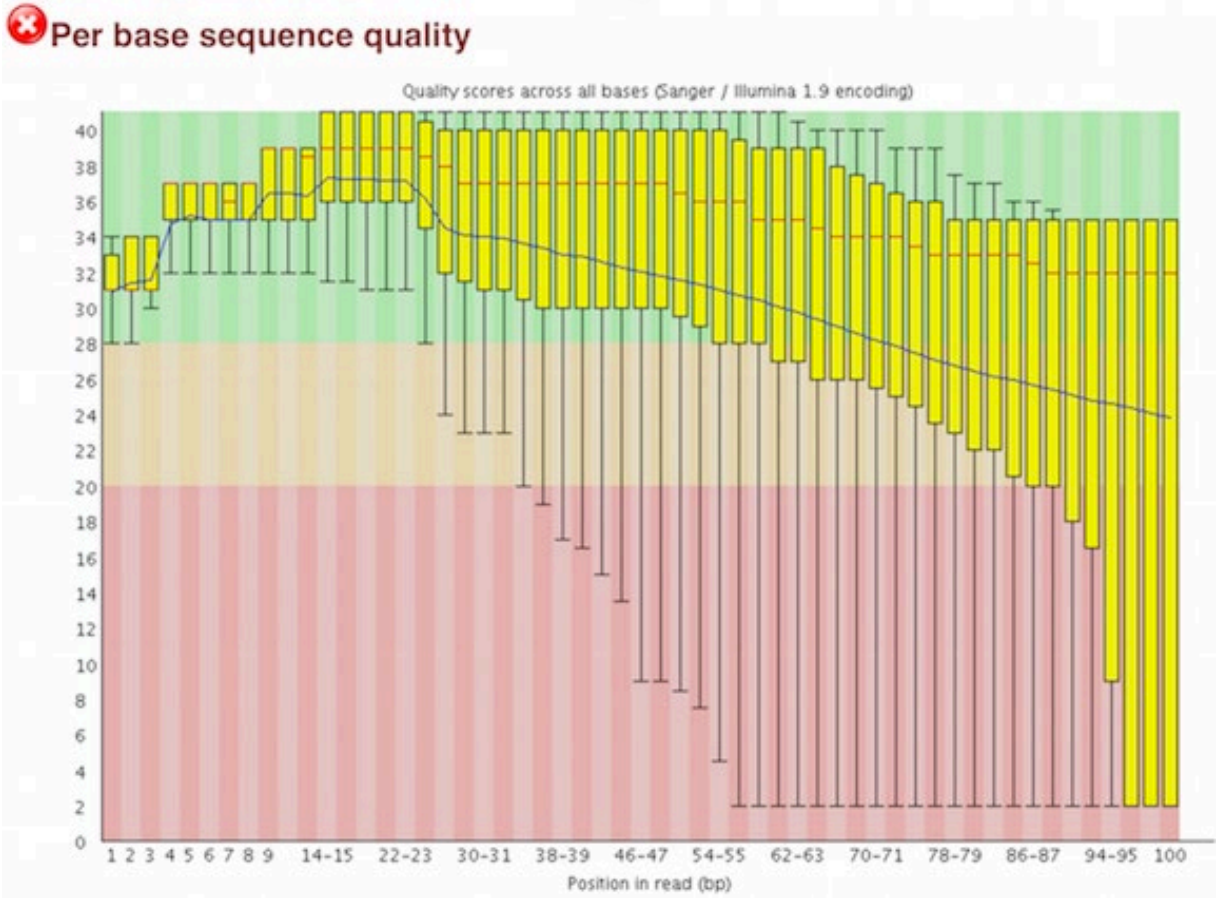
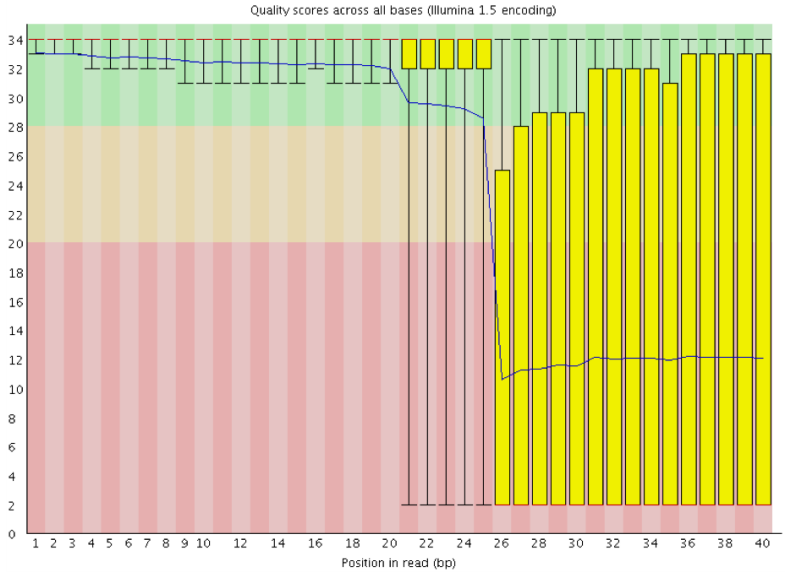
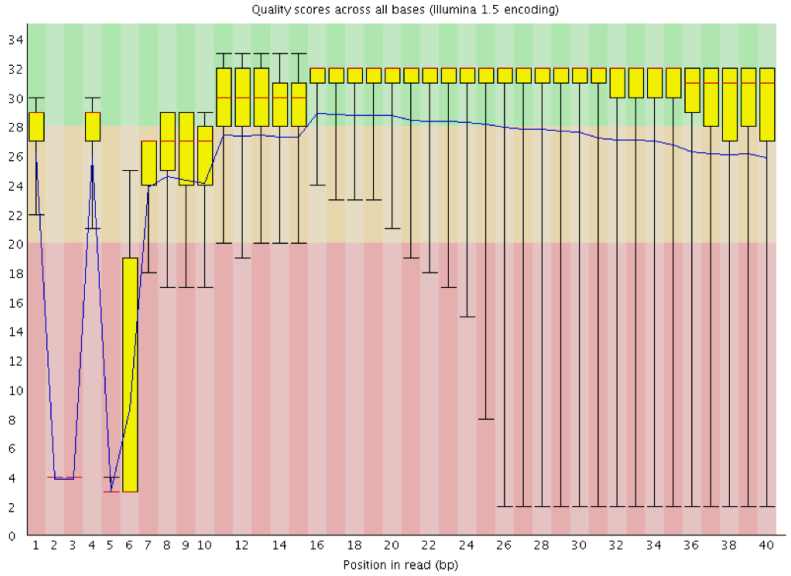


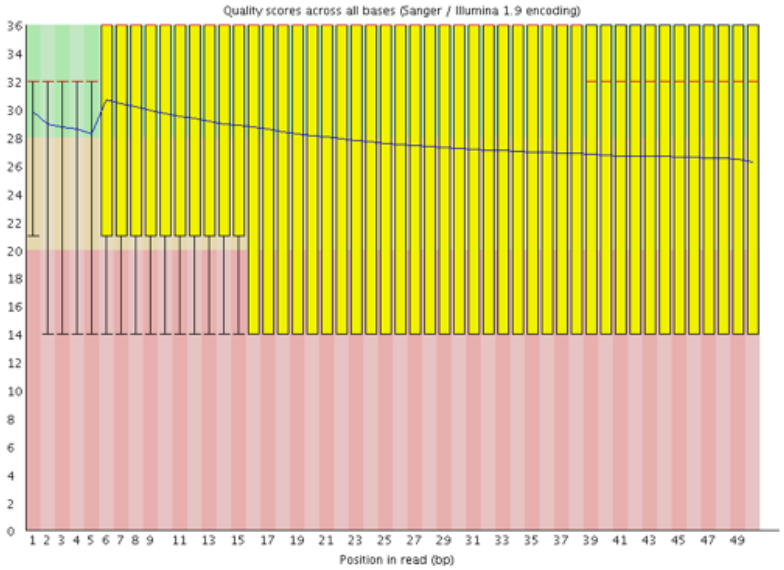
Figure From HBCTraining.github.io QC assessment

PER BASE SEQUENCE QUALITY

- Contact your sequencing facility

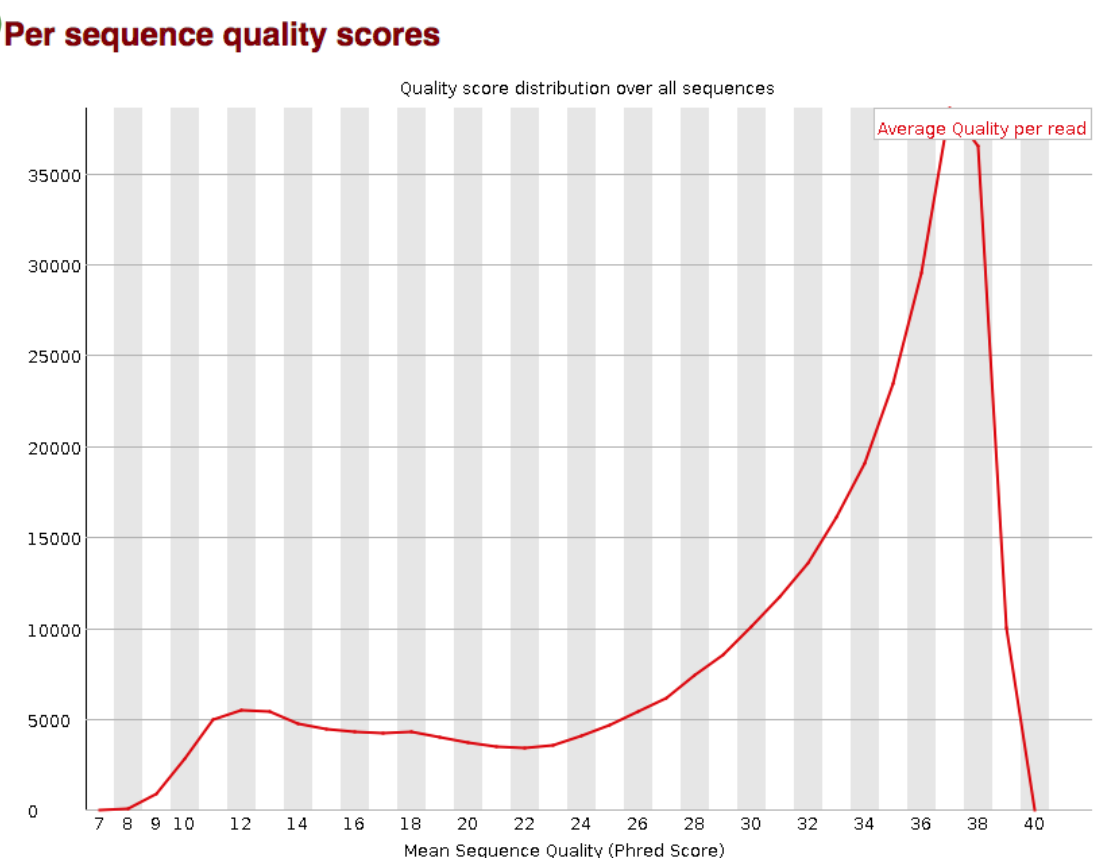






Figures From HBCTraining.github.io QC assessment

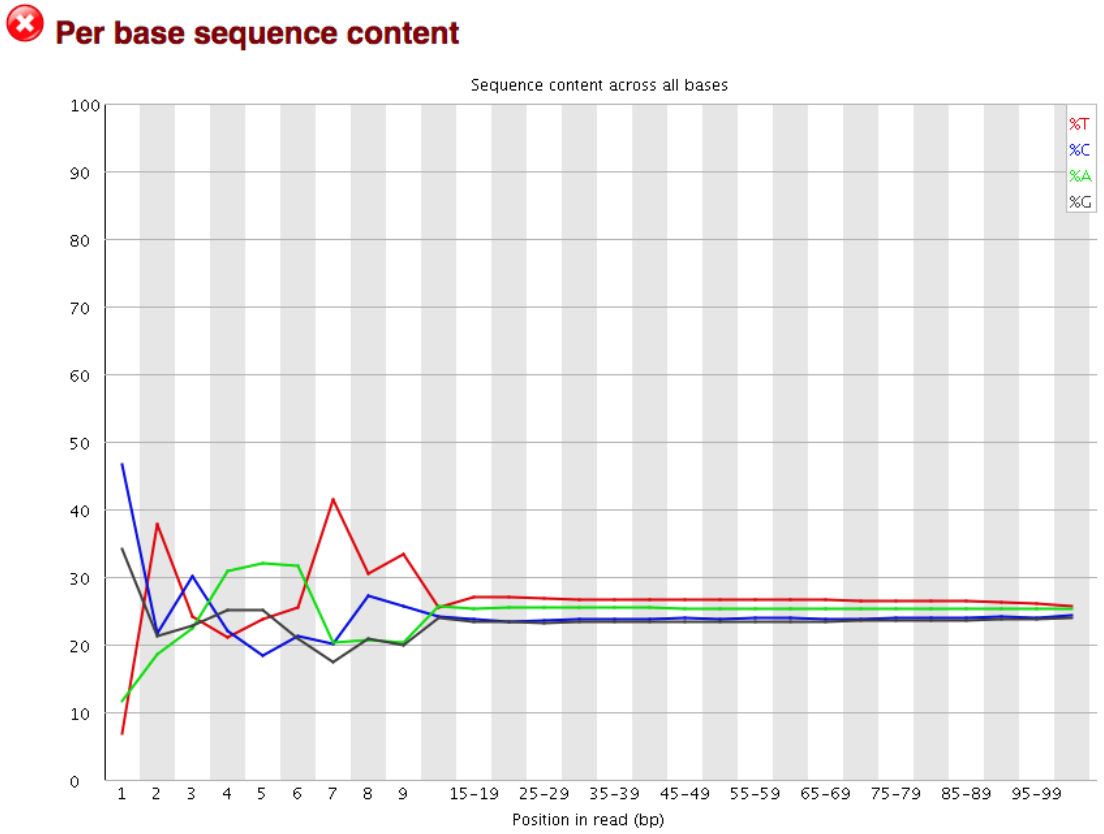
PER SEQUENCE QUALITY



Figures From HBCTraining.github.io QC assessment

PER BASE SEQUENCE CONTENT

- Always Fails for RNAseq data.



PER SEQUENCE GC CONTENT

- Ideal situation – Normal distribution of GC content
- Theoretical distribution derived from the dataset.

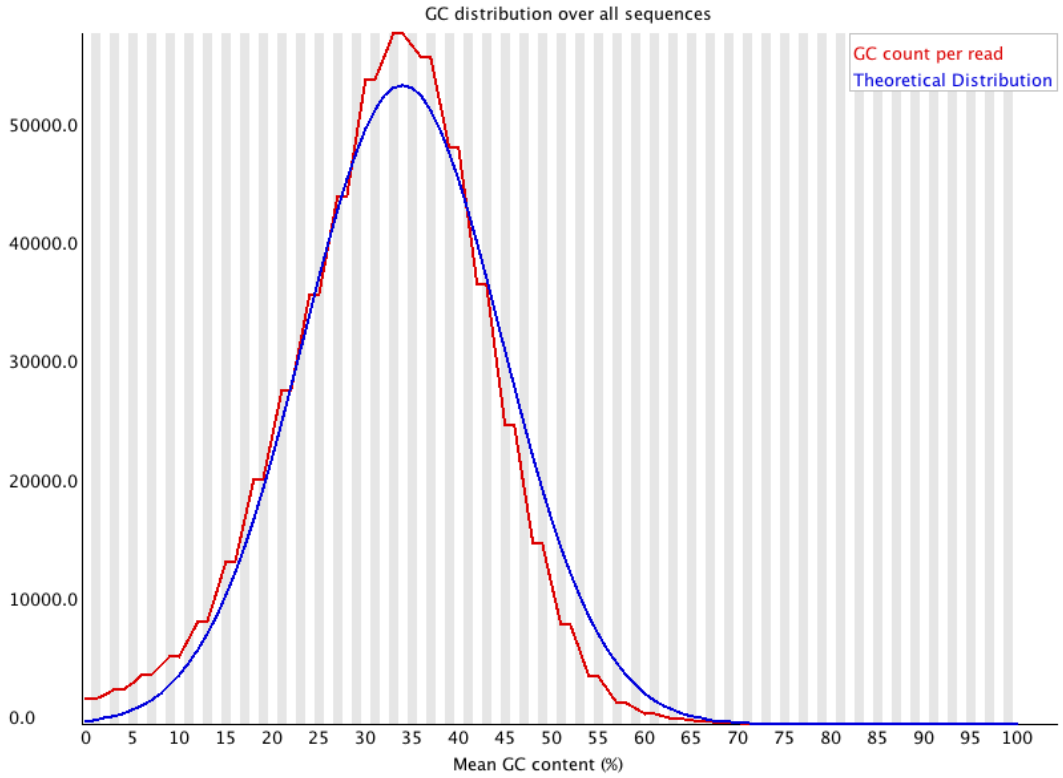
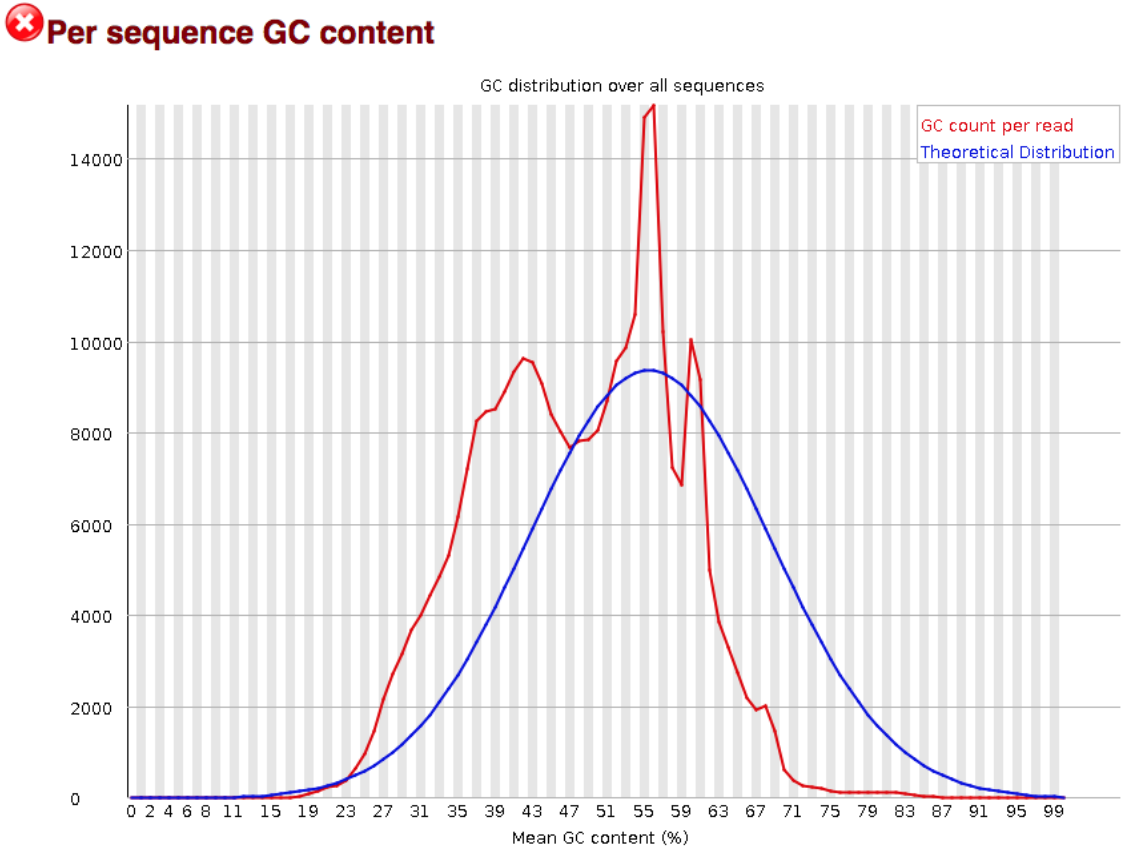


Figure from FastQC documentation

PER SEQUENCE GC CONTENT

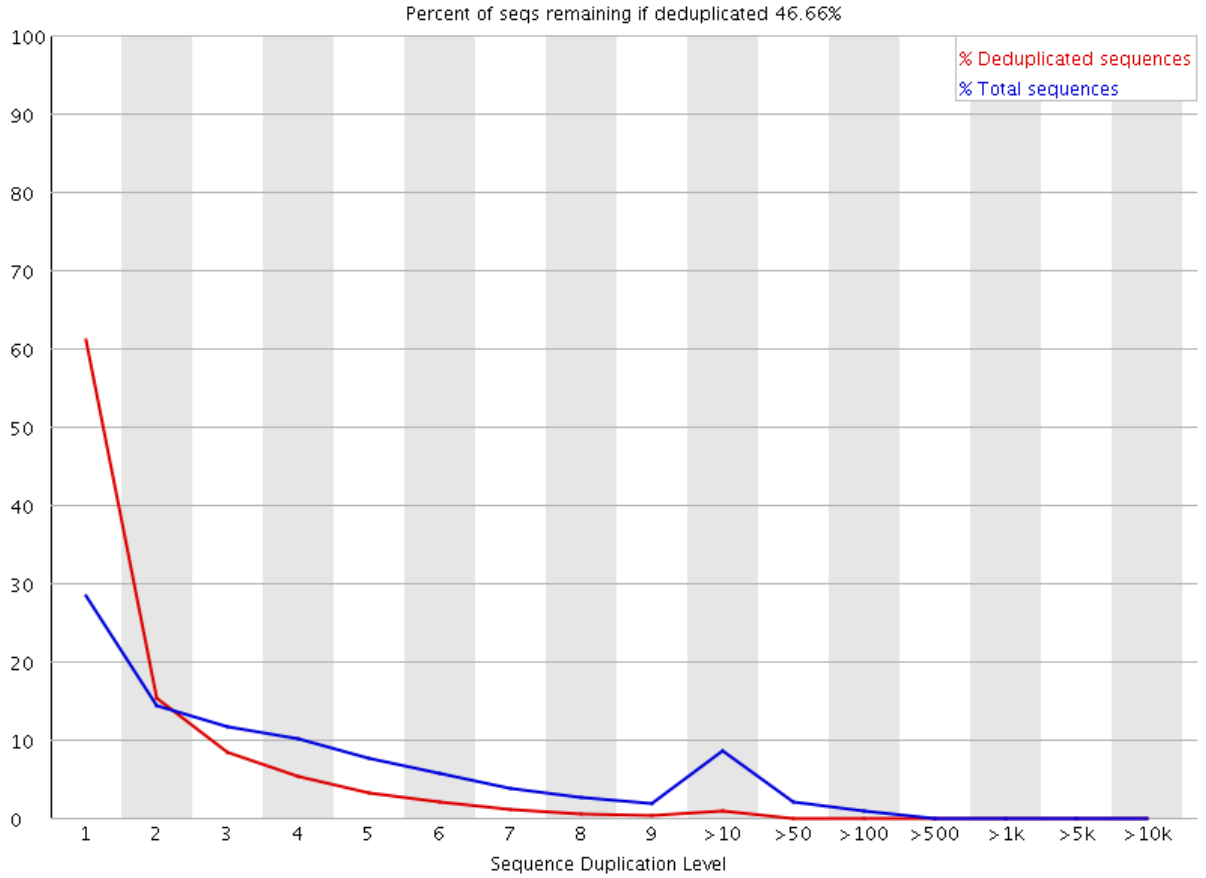
- Sharp peaks might be the result of a specific sequence contaminant such as:
  - Adapter-dimers
  - Highly expressed gene

- Broad peaks might indicate contaminant from another species (with different GC content)
  - Can check who with tools like taxonomer.io



DUPLICATE SEQUENCES

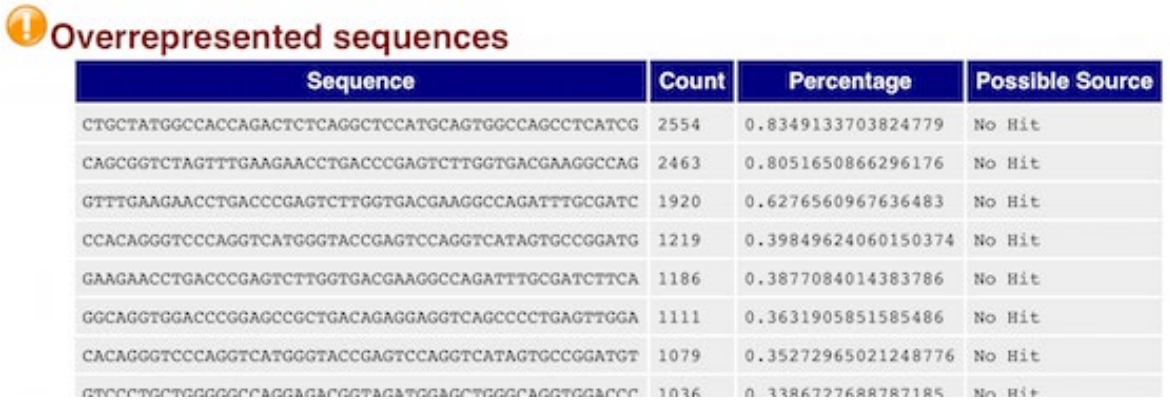
- Warnings/Failure if non-unique sequences >20%/50% total.
- Can arise from low complexity library – sequencing the same bit of DNA over and over again
- PCR duplicates from library prep – can be removed if have UMI
- Often warnings in RNAseq experiments – don't worry too much



Figures From FastQC Documentation

OVERREPRESENTED SEQUENCES

- Warnings/Failure if a given sequence represents >0.1/1% of total
- Sometimes have Adapter Sequences
- Can use BLAST Identify contaminant
- May also be flagged from RNAseq



Overrepresented sequences

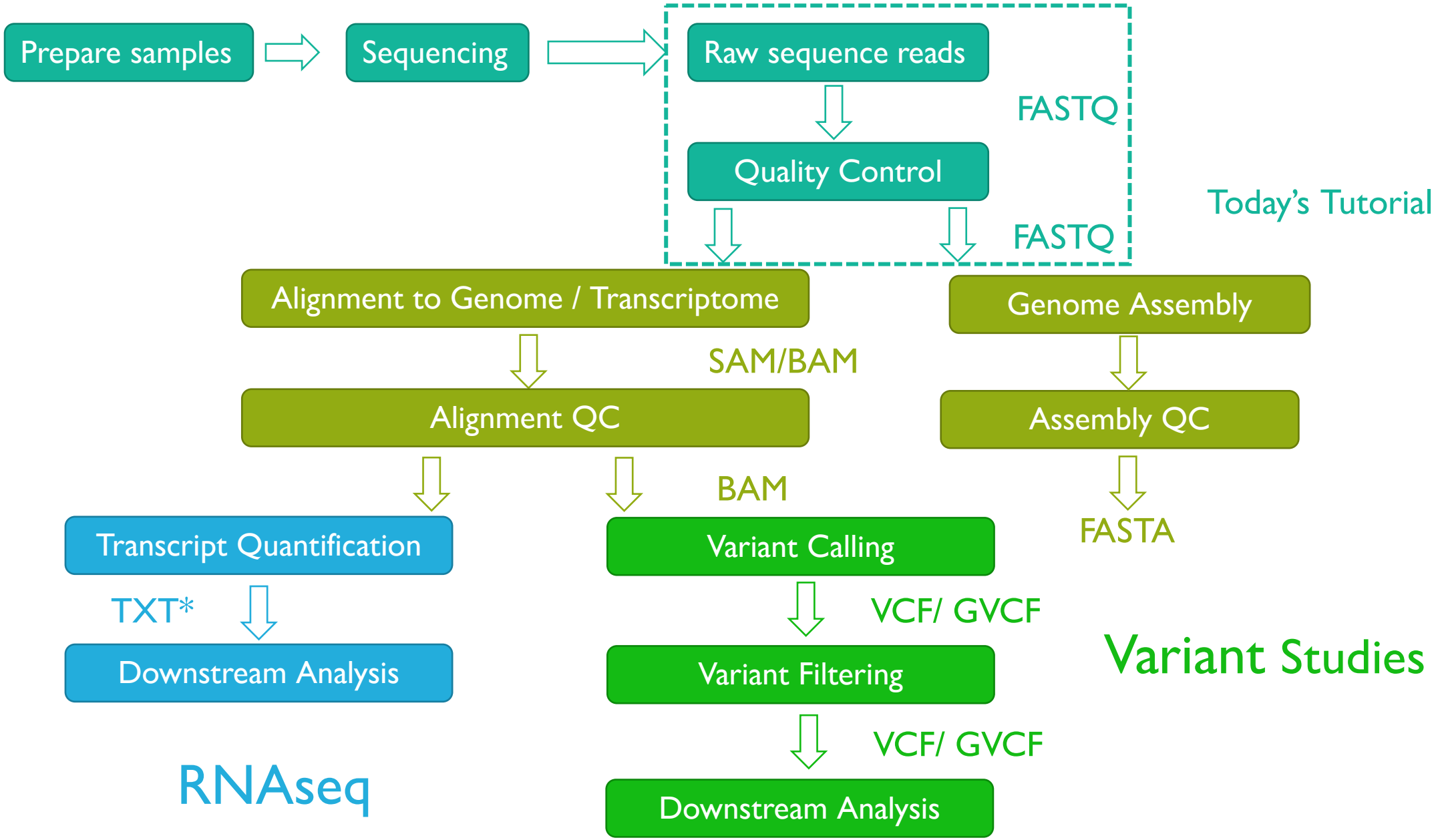
Sequence	Count	Percentage	Possible Source
CTGCTATGGCCACCAGACTCTCAGGCTCCATGCAGTGGCCAGCCTCATCG	2554	0.8349133703824779	No Hit
CAGCGGTCTAGTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAG	2463	0.8051650866296176	No Hit
GTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATC	1920	0.6276560967636483	No Hit
CCACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATG	1219	0.39849624060150374	No Hit
GAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATCTTCA	1186	0.3877084014383786	No Hit
GGCAGGTGGACCCGGAGCCGCTGACAGAGGAGGTCAGCCCCTGAGTTGGA	1111	0.3631905851585486	No Hit
CACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATGT	1079	0.35272965021248776	No Hit
GTCCCTGCTGGGGGCCAGGAGACGGTAGATGGAGCTGGGCAGGTGGACCC	1036	0.3386727688787185	No Hit

FASTQC TAKE HOME

- Warnings and Failure for modules very strict, a lot of data will have an issue.
- What is expected from your experimental design (sample source, library type etc)
- Can use quality trimming programs like trim_galore, Trimmomatic, Adapter Removal…
- RNAseq mapping tools are designed to account for adapter contamination and low sequence quality
- More important for questions that require variant calling
- Downstream ways to account for data quality too

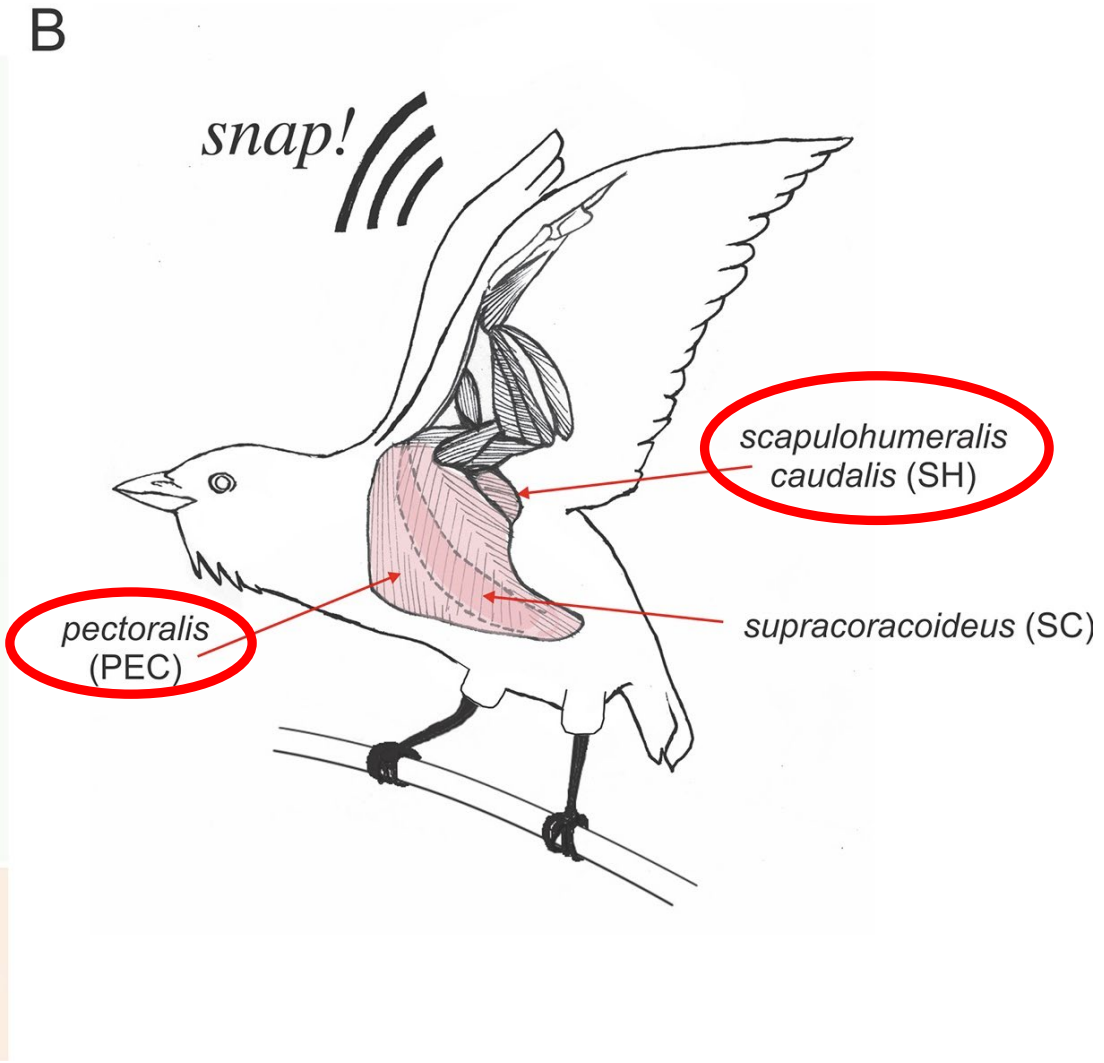
FURTHER READING & REFERENCES

- Ghurye & Pop (2019). Modern technologies and algorithms for scaffolding assembled genomes. PLoS computational biology. https://doi.org/10.1371/journal.pcbi.1006994
- MacManes et al (2014) On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*
- Logsdon et al (2020) Long-read human genome sequencing and its applications. *Nature Reviews Genetics* 21: 597-614
- Fabbro et al (2013) An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One* https://doi.org/10.1371/journal.pone.0085024
- Bush et al (2020) Read trimming has minimal effect on bacterial SNP-calling accuracy. *Microbial Genomics* 6: https://doi.org/10.1099/mgen.0.000434
- PacBio Contigs and Scaffolds (https://www.pacb.com/blog/genomes-vs-gennnnes-difference-contigs-scaffolds-genome-assemblies/)
- Griffith Lab RNA-seq Bioinformatics Course Lecture (https://rnabio.org/course/)
- HBC training Tutorial (https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html)
- FastQC documentation


Prepare samples
Sequencing
Raw sequence reads
FASTQ
Quality Control
FASTQ
Today's Tutorial
Alignment to Genome / Transcriptome
Genome Assembly
SAM/BAM
Alignment QC
Assembly QC
BAM
FASTA
Transcript Quantification
Variant Calling
TXT*
VCF/ GVCF
Downstream Analysis
Variant Filtering
Variant Studies
VCF/ GVCF
RNAseq
Downstream Analysis

YOUR DATA…

- *Manacus vitellinus*
- Lek display has ultra-fast wingsnap
- Normal (PEC) and Fast-twitch (SH)
- Your data from larger study on gene expression in muscles across manakins (Driver, Balakrishnan, Fuxjager et al Unpubl.)



YOUR DATA…

- *Taeniopygia guttata* (wild)
- Wild Zebra Finches from Singhal et al. (2015). Stable recombination hotspots in birds. *Science*
- Whole Genome Resequencing
- Illumina HiSeq 2000 Paired End.