Session 3: DNA Sequencing Strategies and Quality Control

Emerging Approaches For Tumor Analyses
in Epidemiological Studies

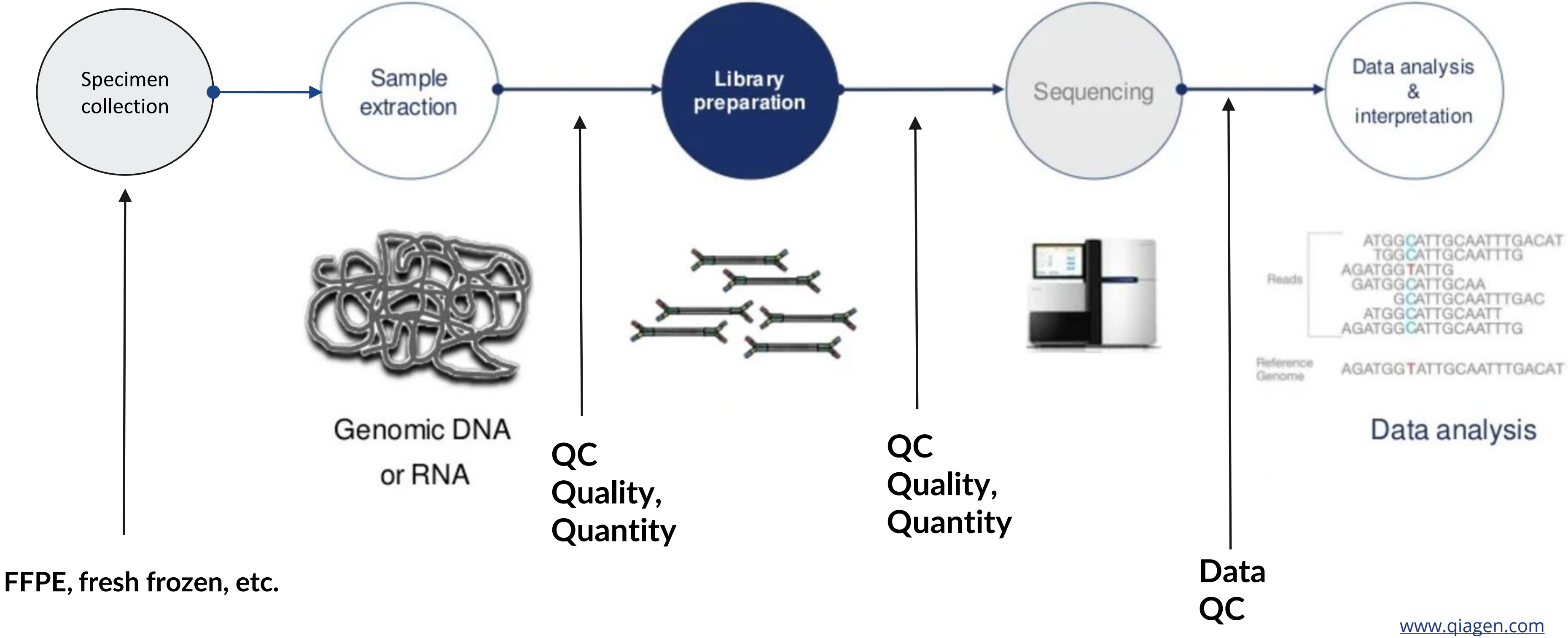
November 16, 2022
9:30 AM- 12:00 PM

Session Overview

- Introduction to Next Generation Sequencing (NGS)-DNA
- Sequencing Strategies and Study Design
- Quality Control: DNA Sequencing Data

Introduction to NGS-DNA

Next Generation Sequencing (NGS) Workflow



Specimen collection

Sample extraction

Library preparation

Sequencing

Data analysis & interpretation

Genomic DNA or RNA

Data analysis

FFPE, fresh frozen, etc.

QC
Quality,
Quantity

QC
Quality,
Quantity

Data
QC

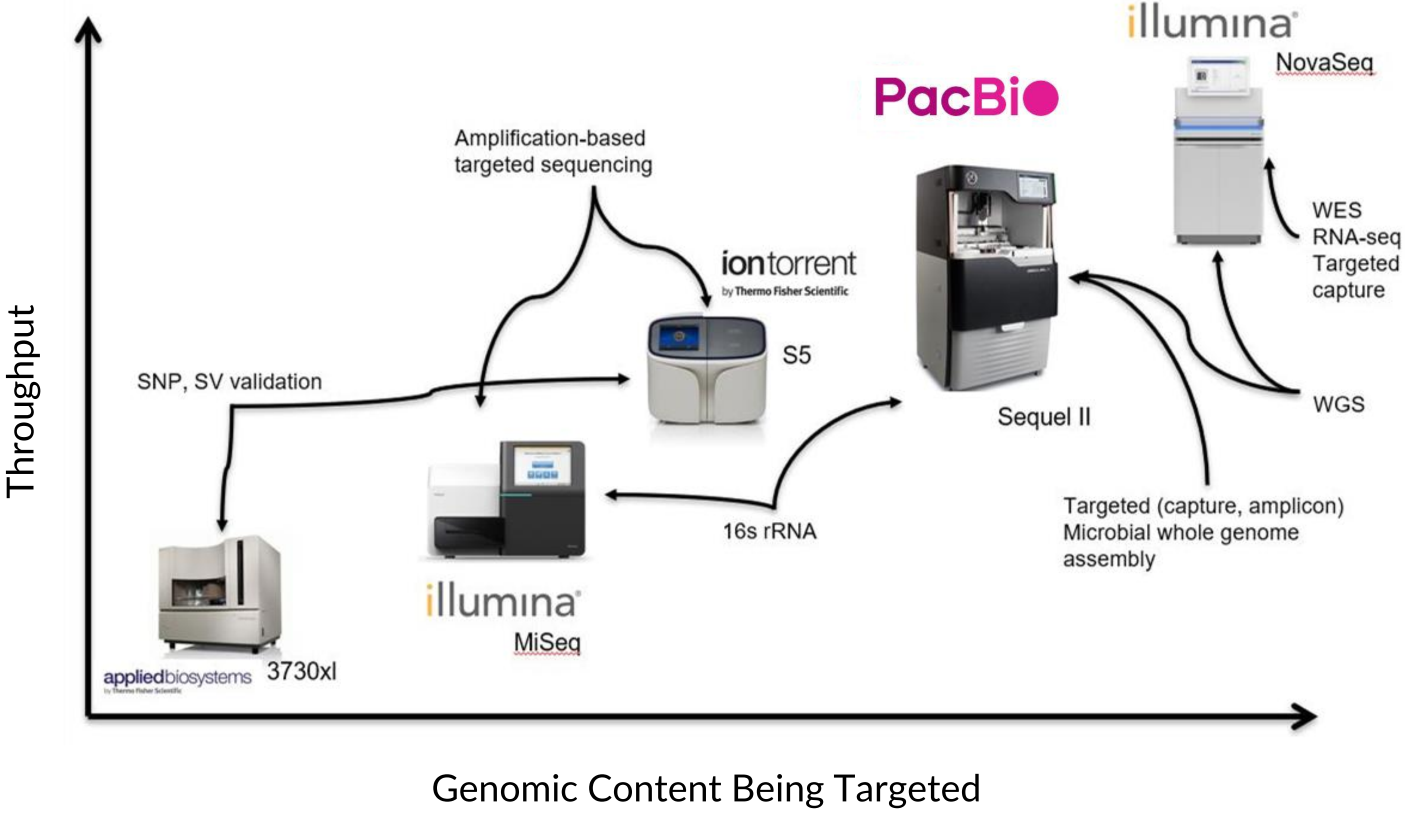
www.qiagen.com

Tailoring Sequencing Strategies for Study Design and Purposes

Sequencing-Major Platforms

	illumina	ion torrent	PacBio	Oxford NANOPORE Technologies
Read Length	**Short Read** (commonly 75bp-250bp; up to 600bp)	**Short Read** (commonly 100bp-250bp; up to 600bp)	**Long Read:** Up to 20kb or more	**Long Read:** Up to 30kb or more
Common applications	WGS, WES, targeted capture	Targeted amplification panels	WGS, targeted (difficult regions), base mod detection	WGS, targeted (difficult regions), base mod detection
Instruments	High Throughput: NovaSeq Low Throughput: NextSeq, MiSeq, iSeq	S5	Sequel II	Hight Throughput: PromethION Low Throughput: Flongle, MinION, GridION
What is DNA sequenced on?	Flow Cell	Chip	SMRT Cell	Flow Cell

Platform and Instrument Use Cases @ CGR



Throughput

Genomic Content Being Targeted

appliedbiosystems 3730xl

illumina MiSeq

iontorrent by Thermo Fisher Scientific S5

PacBio Sequel II

illumina NovaSeq

SNP, SV validation

Amplification-based targeted sequencing

16s rRNA

Targeted (capture, amplicon)
Microbial whole genome assembly

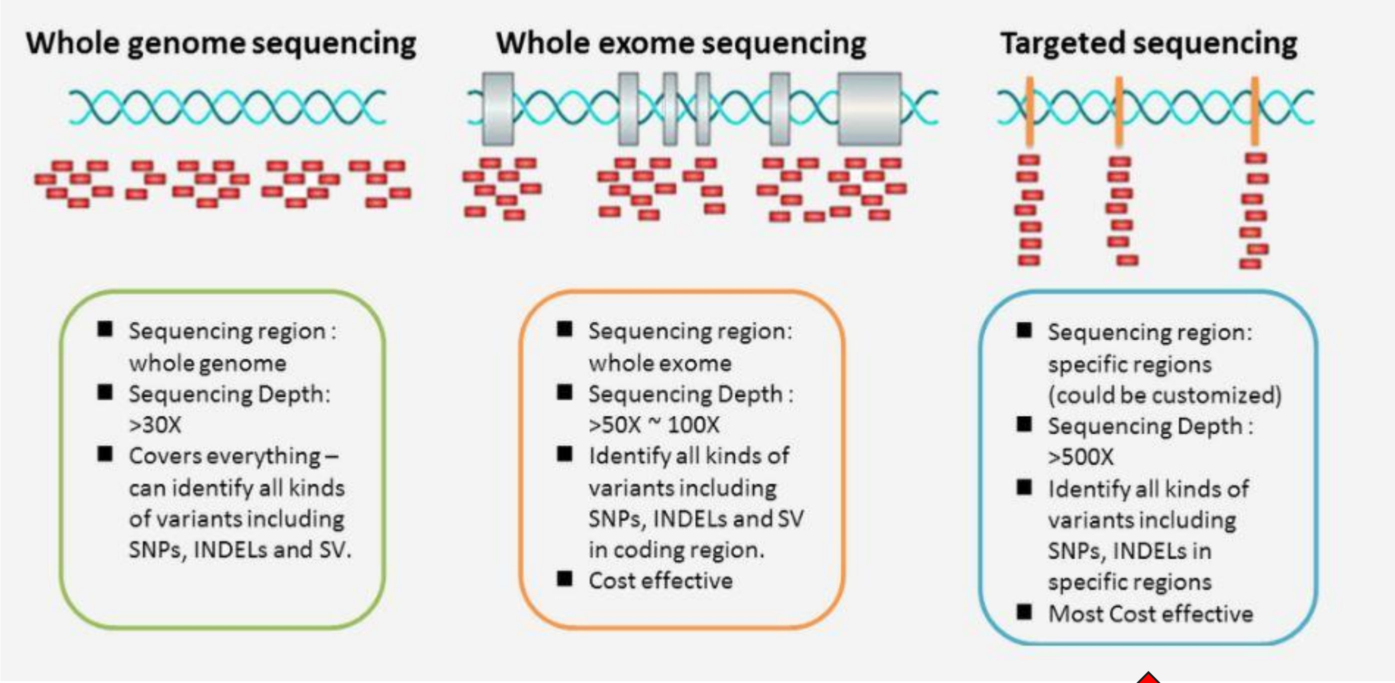
WGS

WES
RNA-seq
Targeted capture

Major Sequencing Platforms-Specs

Sequencing Platform	Instrument	Data Type	Read Length N50 (bp)	Read Accuracy (%)	Throughput per Run (Gb)	Cost per Gb ($)	Instrument Throughput per year (Gb)
PacBio	Sequel II	HiFi	10,000-20,0000	>99	15-30	50-100	10,000
Oxford Nanopore	MinION/GridION	Long	10,000-60,000	97-99	2-20	50-500	20,000-100,000
		Ultra-long	100,000-200,000		0.5-2	500-2,000	1,000-5,000
	PromethION	Long	10,000-60,000		50-100	20-40	3,000,000
Ion Torrent	S5	Single-end	100-600	98-99	0.3-50	30-300	10,000
Illumina	MiSeq	Paired-end	36-600	99.9	0.5-15	100-600	1,500
	NovaSeq	Paired-end	35-500		65-3,000	4-30	1,200,000

Different DNA Sequencing Strategies



Whole genome sequencing

- Sequencing region : whole genome
- Sequencing Depth: >30X
- Covers everything – can identify all kinds of variants including SNPs, INDELs and SV.

Whole exome sequencing

- Sequencing region: whole exome
- Sequencing Depth : >50X ~ 100X
- Identify all kinds of variants including SNPs, INDELs and SV in coding region.
- Cost effective

Targeted sequencing

- Sequencing region: specific regions (could be customized)
- Sequencing Depth : >500X
- Identify all kinds of variants including SNPs, INDELs in specific regions
- Most Cost effective

Hess et al. 2020

Targeted sequencing: Hybridization Capture vs. Amplicon Sequencing

https://www.thermofisher.com/ https://www.illumina.com/

	Hybridization Capture	Amplicon Sequencing
Principle	Capture by hybridization to biotinylated probes & isolated by magnetic pulldown	Amplified and purified using pools of carefully designed oligo probes
Size	20kb–62Mb regions. Typically >50 genes	A few to hundreds of genes in a single run. Typically <50 genes
Sample input	Higher input required (1-250ng for library prep)	Lower sample input required (needle biopsy aspirate or cDNA) (10-100ng)
Variant types	More comprehensive for all variant types	Ideal for SNVs and indels
Homologous regions (e.g. pseudogenes) **Hypervariable regions (e.g. TCR)** **Di/Tri nucleotide repeat regions (e.g. MSI)**	Difficulty distinguishing between the regions, resulting in non-specific enrichment	Better enrichment with specifically designed PCR primers
Overall	More comprehensive method, but more expensive with longer hands-on time and turnaround time	Less comprehensive, more affordable, and easier workflow

Targeted Sequencing - Common Panels

Gene Panel	Gene Count	Sample Type	Variants	Notes
Oncomine Comprehensive Plus	500+	DNA, RNA	SNVs, indels, CNVs, fusions, splice variants	Include TMB and MSI assays for potential immunotherapy applications. Also assess 46 genes in HRR pathway
TruSight RNA Pan-Cancer Panel	1385	RNA	SNVs, indels, fusions, novel transcripts, expression	Enables quantitative measurement of gene expression as well as the detection of gene fusions with both known and novel gene fusion partners.
MSK-IMPACT	505	DNA	SNVs, Indels, CNVs, fusions	Includes genes important in development and behaviour of tumors nominated by researchers and experts from across MSK. All actionable targets are also included.

- Also panels specific to certain cancers and diseases
- One can customize specific genes or regions of interests too

https://www.mygenomics.com/cancer-panels-gene-list/

WES - Common Capture Platforms

Bioinformatics pipeline for WES is somewhat similar to WGS, with the need to specify WES capture platforms BED files

Platform	Target Capture Region Length	Required input quantity	BED file links
Agilent SureSelect Human All Exon v8	35.1 Mb	10-400ng of DNA	https://kb.10xgenomics.com/hc/en-us/articles/115004150923-Where-can-I-find-the-Agilent-Target-BED-files-
Roche KAPA HyperExome	43 Mb - targeting hg38 genome assembly	100ng DNA	Download HG38 Design Files for the KAPA HyperExome Probes Download hg19 Design Files for the KAPA HyperExome Probes
Illumina TruSeq	45 Mb	100ng of DNA	https://emea.support.illumina.com/downloads/truseq-exome-product-files.html

Sequencing Strategies and Study Design

	Genomic Strategies		
Research area	**Targeted Panel**	**WES**	**WGS**
Coding driver genes	Pre-defined genes only	Yes - *De novo* discovery possible	Yes - *De novo* discovery possible
Non-coding	Pre-defined regions only	No	Best
Structural variant	Limited	Limited	Best
Tumor evolution	Limited	Limited	Best
Copy number analysis	Limited	Limited	Best
Mutational signatures	Being developed	Limited and potentially biased	Best
Gene fusion	Pre-defined regions only	Limited	Best - *De novo* discovery possible

Benefits of Long Read Sequencing with Respect to Study Design

Study Design	Short Read WGS (SRS, e.g. Illumina)	Long Read Sequencing (LRS, e.g. PacBio)
Copy Number Estimation	Relies heavily on PCR -> GC content bias (dependence between read coverage and GC content) -> influence copy number estimation	Does not rely on PCR -> Unbiased by GC content
Structural Variants (SVs)	Limited mostly to small SVs	Increased sensitivity and accuracy, better detection of large SVs (>50bp)
Genome Assembly and Resolution of Repeat-Heavy Regions	Limited resolution for repetitive regions	“Close the gaps” in the genome assembly Higher resolution for repetitive regions
Haplotype Phasing	Direct phasing is limited as SNVs are required to be on the same reads	Increased sensitivity and accuracy
Pseudogenes	Homologous pseudogenes might impact mapping rates and variants calling in functional counterparts	Better distinguishment of pseudogenes vs functional counterparts. Better identification of pseudogenes

Sequencing Strategies Comparison - Illumina

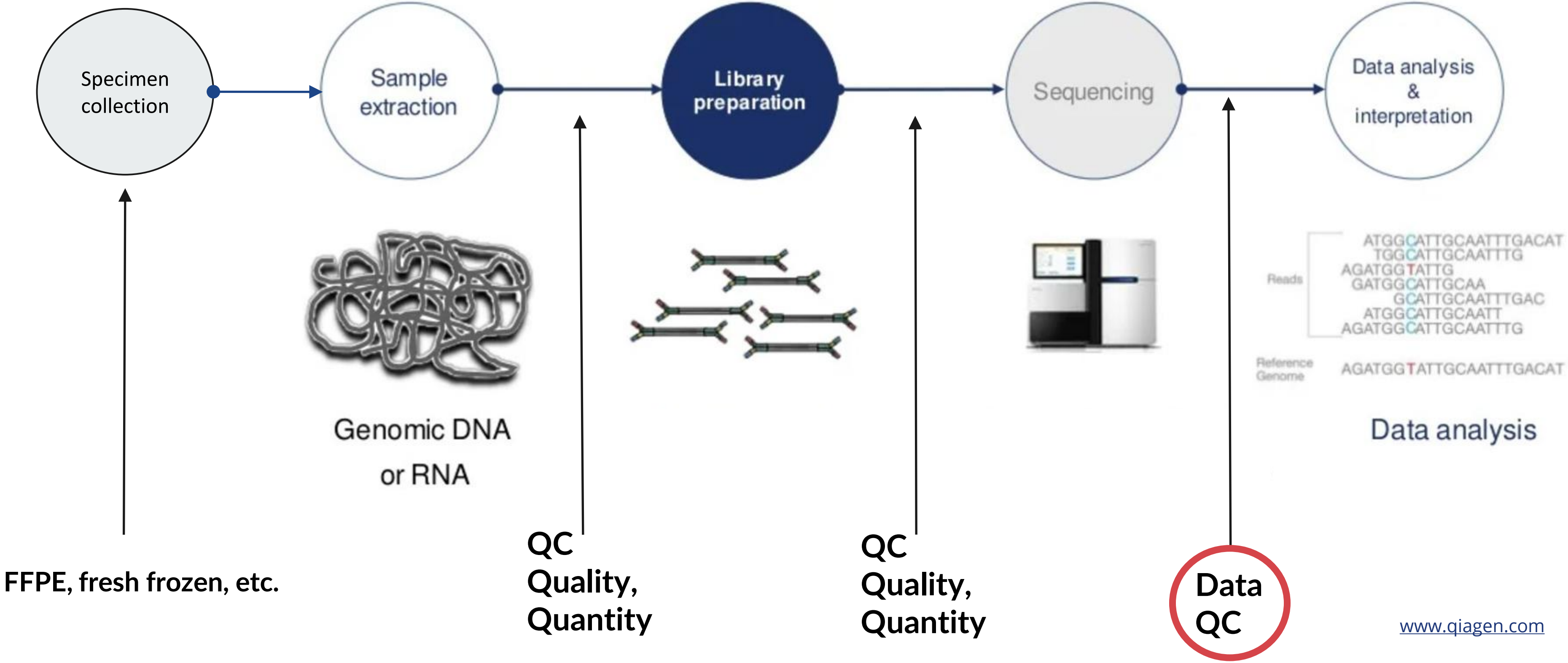
	Targeted Sequencing	WES	WGS
Cost (per sample)	$50-200	$90-200	$600-1500
DNA Quantity Required	50-200ng	50-200ng	**200-1000ng (PCR free)**
DNA Quality Required	Amenable FFPE	Amenable to FFPE	**FFPE not generally used**
Standard Coverage Depth	>100x	>40x (germline) >100x (somatic)	>30x (germline) >80x (somatic)
Samples (Per Run)	Up to 384	Up to 384	Up to 48

Sequencing Strategies Comparison - PacBio

	Targeted Sequencing	WES	WGS
Cost (per sample)	$15-$200	N/A	~$1500-5000
DNA Quantity Required	50-500 ng		**>3-5 ug***
DNA Quality Required	High quality		**Very high quality**
Standard Coverage Depth	>100x		10-30x
Samples (Per Run)	12-384		1

Quality Control: DNA Sequencing Data

Next generation DNA-Sequencing (NGS) Workflow



Specimen collection

Sample extraction

Library preparation

Sequencing

Data analysis & interpretation

Genomic DNA or RNA

Data analysis

FFPE, fresh frozen, etc.

QC
Quality,
Quantity

QC
Quality,
Quantity

Data
QC

www.qiagen.com

Common NGS Data QC

- Sequencing data quality control
    - QC *before* read mapping
        - Flowcell and Sample-level metrics
        - Tools: FastQC, FASTX toolkit
    - QC *after* read mapping
        - Sample-level metrics
        - Tools: samtools, picard, verifybamid, FASTQ screen, somalier

Data QC before read mapping

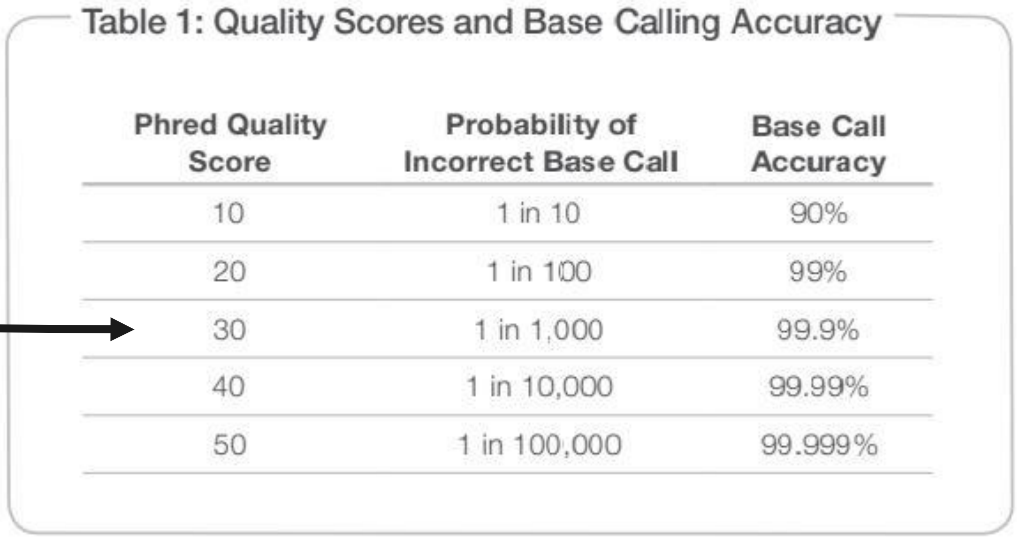


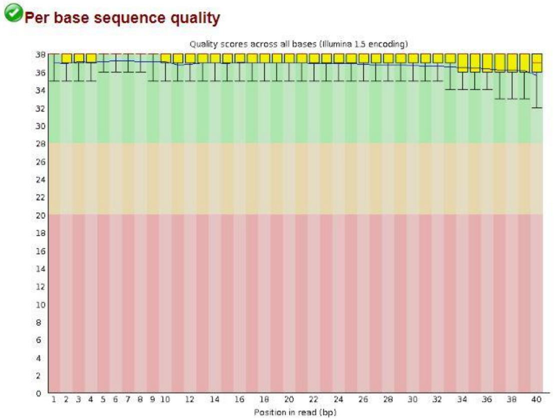
Table 1: Quality Scores and Base Calling Accuracy

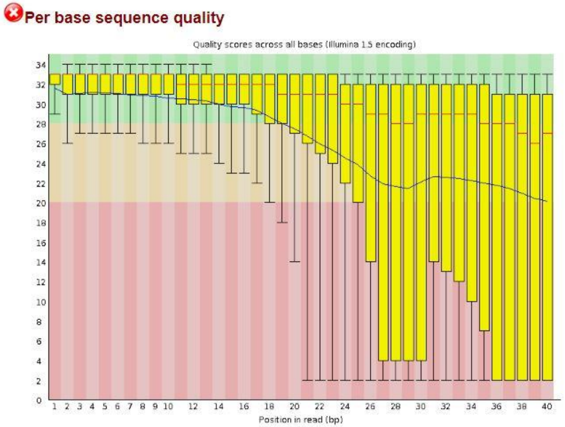
Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Q30 →

- Lab generating the sequence data should be able to provide a table of **Flow Cell level metrics**
- May include Total Yield (Gb) for the Flow Cell, sequence error rates, Q30%, # passed filter reads, etc.
- These should match to Illumina’s published specifications for the Flow Cell
- **Impact: Low Yield leads to lower coverage; low Q30s or high error rates may impact variant calls**

Data QC before read mapping



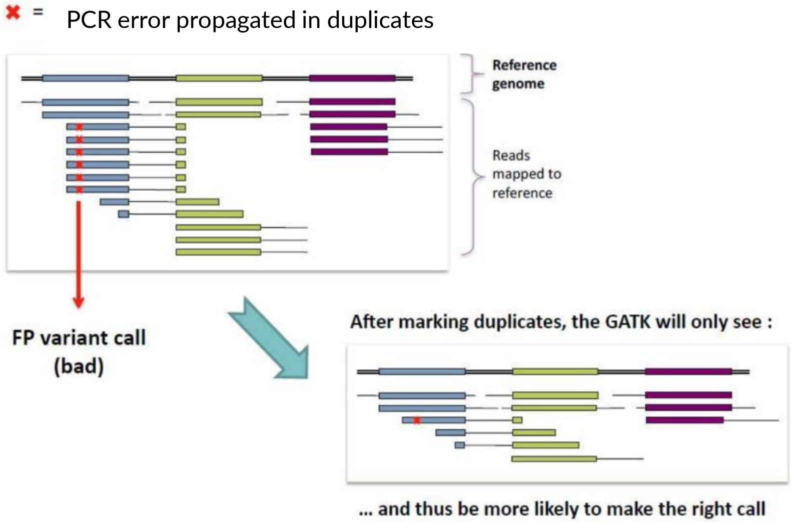


- **Sample level metrics** (read length, read quality)
- Many of these can be assessed with FastQC-tool to help visualize your current fastq files
- FASTX toolkit-also has tools to help address QC issues that are discovered
- **Can flag additional quality issues with the data before mapping is performed.**

Data QC after read mapping

- After fastq files are QC’ed, and reads are aligned to the reference genome, **additional QC should be done on the aligned reads/BAM files**
    - Duplicate rates
    - Insert size
    - Coverage depth
    - Contamination
    - Sex concordance, relatedness
- MultiQC-a reporting tool that parses output of many tools to help visualize QC checks of the BAM files

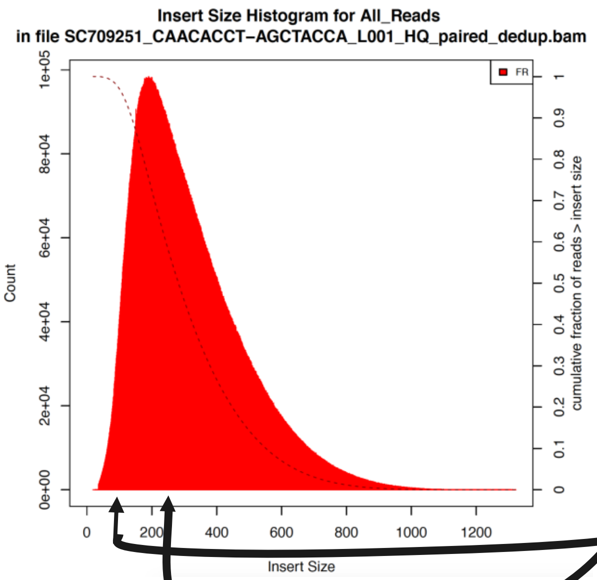
Duplicates


= PCR error propagated in duplicates
Reference genome
Reads mapped to reference
FP variant call (bad)
After marking duplicates, the GATK will only see :
... and thus be more likely to make the right call

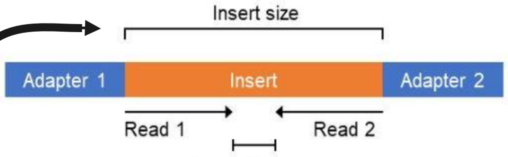
- **PCR amplification** (during library prep or sequencing) **creates copies of library molecules, called "duplicates".**
- For most DNA sequencing applications, duplicate copies of the same molecule need to be removed from the data, keeping only one copy.
- Your duplicate rate tells you what percentage of your reads are being removed.
- Tools: picard, samtools
- General rule: expect lower duplicate rates for WGS, WES with high quality DNA, higher duplicate rates for Targeted panels or low quality/low quantity DNA.
- **Impact of neglecting to remove duplicates: can bias or cause false variant calls**
- **Impact of high duplicate rates: can reduce overall coverage**

Duplicate Rates	WGS	WES	Targeted
Good DNA Quality/Quantity	low	low	mid
Low DNA Quality/Quantity	mid	mid	high

Insert size of the library



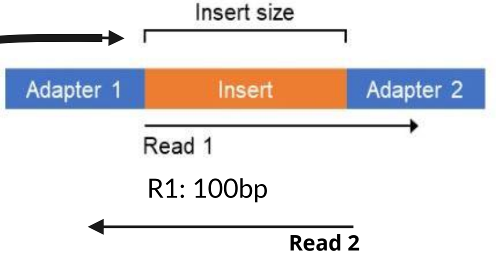
Insert Size: 250bp



R1: 100bp R2: 100bp

Insert size for some molecules/read pairs will be larger than the distance that read 1 and read 2 span-**ideal**

Insert Size: 85bp



R1: 100bp

R2: 100bp

Insert size for some molecules will mean that each read spans the entire insert (and may sequence into adapter); paired reads may overlap

Tools: picard, samtools

Impact of shorter insert size: reduction in coverage, possible small increase in false variant calls

Coverage Depth

- After read mapping and deduplication, coverage depth is calculated at each location being targeted.
- Coverage is generally reported as the mean depth across all target regions-some regions might have deeper or shallower coverage than the mean.
- Tools: picard, samtools
- **Impact: With lower coverage depth, some variant calls will be missed.**

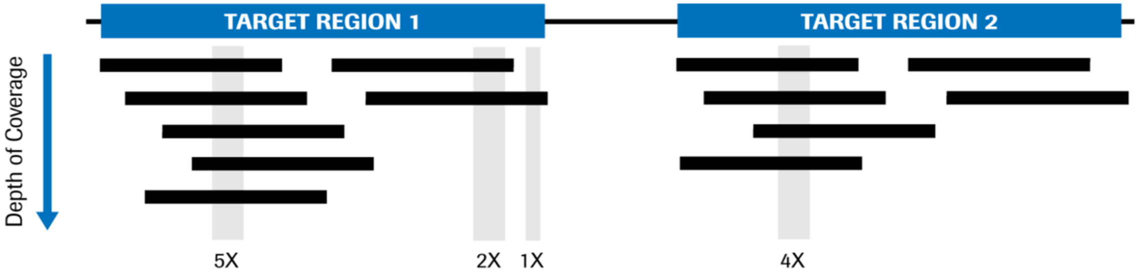
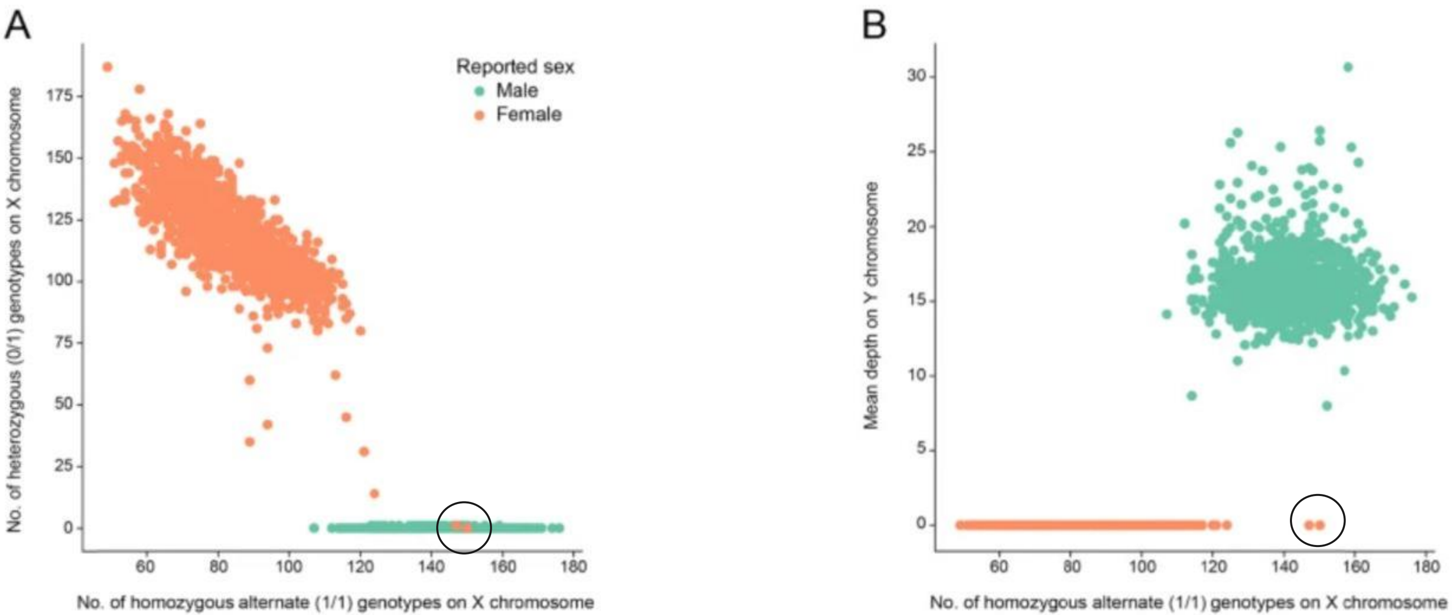


Figure 1. Illustration of coverage depth. Blue bars represent target genomic regions, black bars indicate unique mapped reads, and shaded boxes show various coverage depths across the target region. In this example, coverage of target regions ranges from 1X to 5X; required coverage depth varies widely across applications.

Contamination

- Inter-Sample Contamination:
  - one human sample contaminating another human sample
  - Tool: VerifyBamID
  - **Impact: High levels can lead to false variant calls**
- Inter-Species Contamination
  - A non-human sample contaminates a human sample
  - Tool: FASTQ screen
  - **Impact: High levels can reduce coverage, lead to false variant calls if reads map**

Sex Concordance

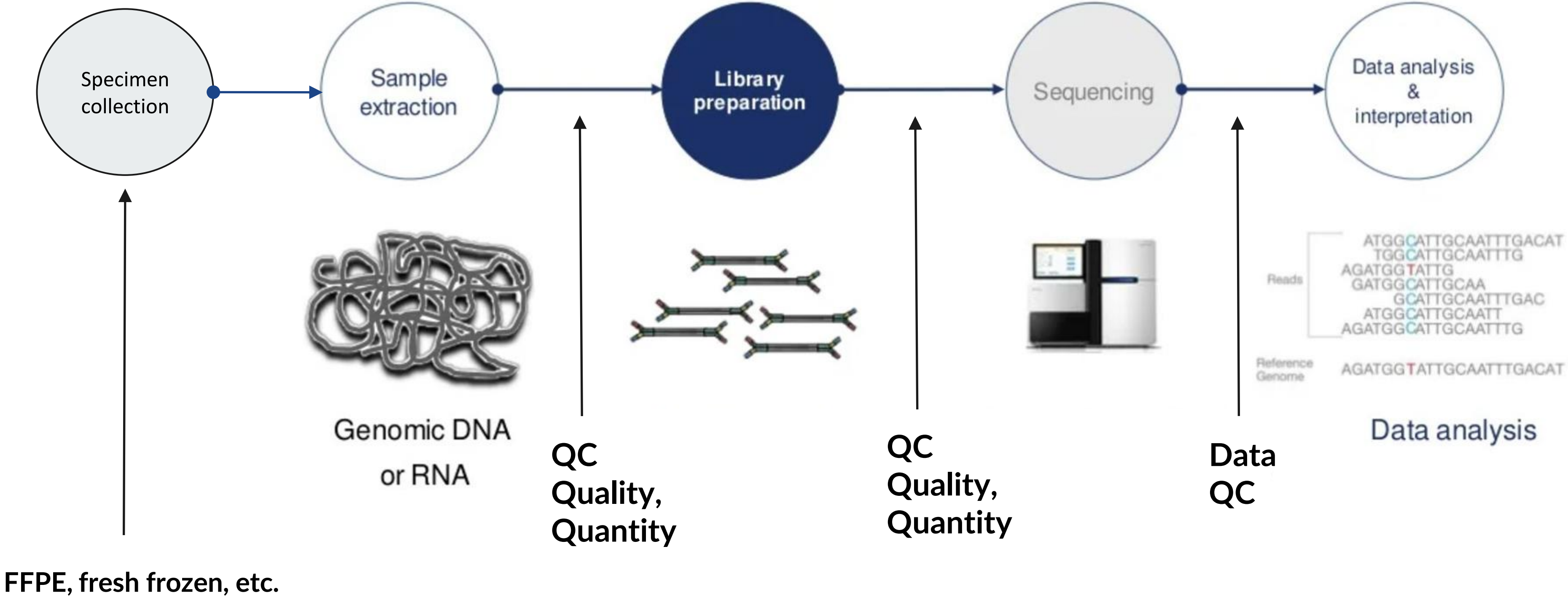


Data shown here is from 1000 Genomes data. Tool: Somalier

Impact: Analyzing data with incorrect subject-level information can lead to false conclusions

https://brentp.github.io/somalier/ex.html

Next Generation Sequencing (NGS) Workflow



Specimen collection

Sample extraction

Library preparation

Sequencing

Data analysis & interpretation

Genomic DNA or RNA

QC Quality, Quantity

QC Quality, Quantity

Data QC

Data analysis

FFPE, fresh frozen, etc.

www.qiagen.com