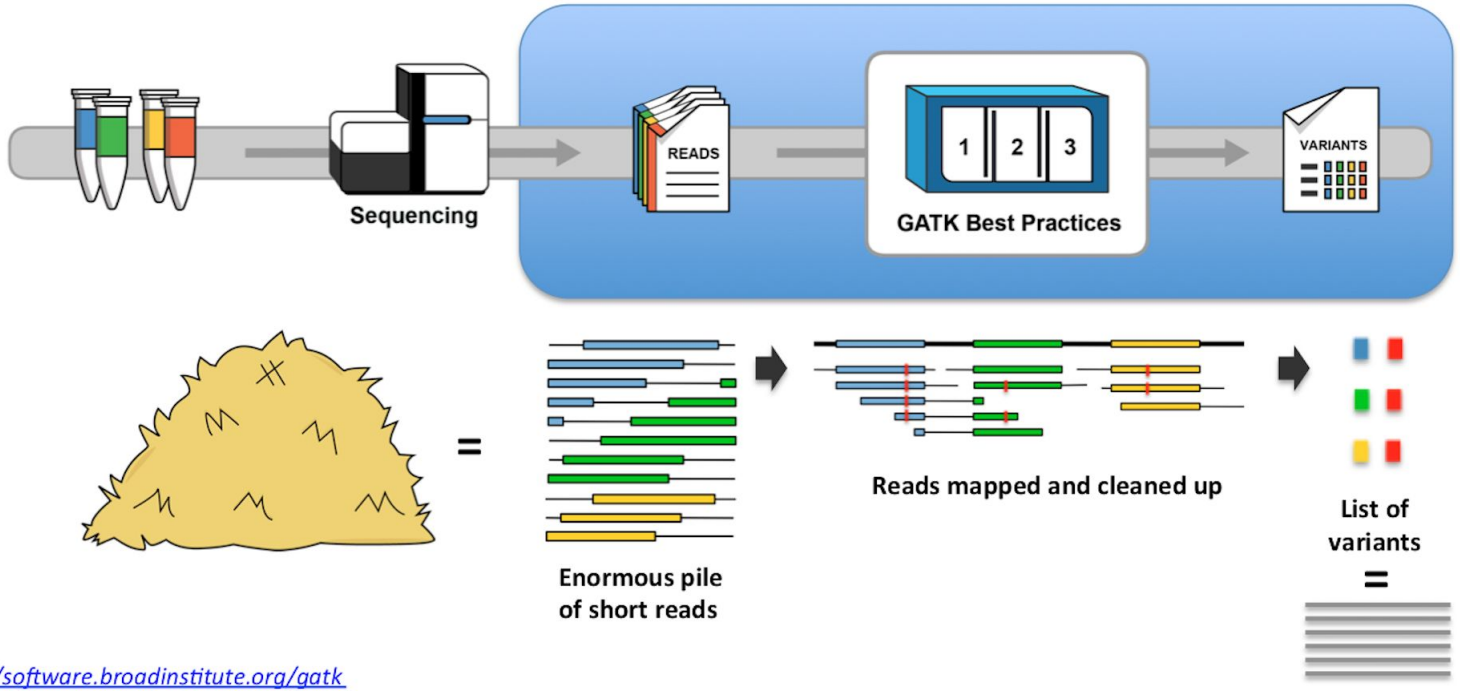
Quality Assurance/Quality Control (QA/QC) for Next-Generation Sequencing

Nicole Ruiz-Schultz
Utah Newborn Screening Program

Outline

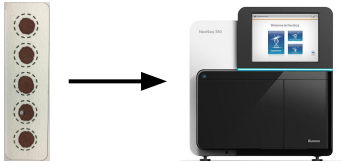
- Brief Overview of Next-Generation Sequencing (NGS)
- Overview of Utah NGS Pipeline
- Quality Control (QC) Checks in Wet-Lab Pipeline
- Quality Control (QC) Checks in Bioinformatics Pipeline
- Relevance to Clinical Laboratory Regulations
- Summary

Overview of NGS pipeline



https://software.broadinstitute.org/gatk

Utah NGS Pipeline



FASTQ File Generation
bcl2fastq

Pre-Alignment Data Processing
Seqyclean
Fastp
FastQC

Sequence Alignment
BWA

Post-Alignment Data Processing
Picard

Variant Calling
GATK HaplotypeCaller

Variant Filtering
Hard filtering recommendations from GATK

Variant Annotation
SnpEff
Variant Effect Predictor

Variant Normalization
vt normalize

Variant Interpretation
Local variant database
- Public variant data
- Utah NBS variant data

Pipeline output:
VCF file of annotated variants*

*Variant reinterpretation every 6 months

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma. NIH Public Access; 2013;43(1110):11.10.1-33.
Adrian Tan, Gonçalo R. Abecasis and Hyun Min Kang. Unified Representation of Genetic Variants. Bioinformatics (2015) 31(13): 2202-2204

Sources and Types of Sequencing Errors

Where in the pipeline can errors occur?

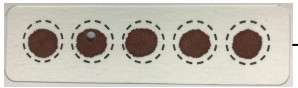
- DNA extraction
- Library prep
- Sequencing
- Data Analysis (error sources at every step)

What can cause problems with your NGS experiment?

- GC bias
- PCR amplification bias
- Low complexity samples

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121-32.
Robasky K, Lewis NE, Church GM. The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet. 15. England2014. p. 56-62.

Where do QC checks occur in the wet-lab pipeline?



DNA extraction

Pre-enriched library prep

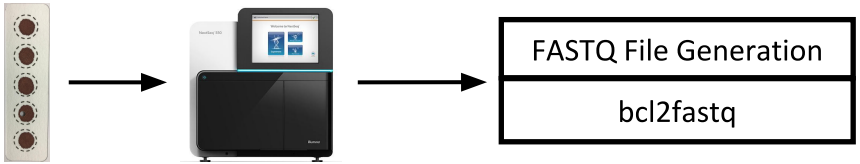
QC1 - Quantify pre-enriched library

Exome library prep

QC2 - Quantify post-enriched library

Sequence

Where do QC checks occur in the bioinformatics pipeline?



Sequencing Run Metrics

Metric	Definition	Example
PhiX Aligned (%)	Percentage of reads from clusters in each tile that aligned to the PhiX genome	6.15
Cluster Density (K/mm^2)	Density of clusters for each tile	189.5
Clusters Passing Filter (%)	Percentage of clusters passing filter	93.20
Q30 (%)	Percentage of bases with a quality score of ≥ 30	92.75

https://help.basespace.illumina.com/articles/descriptive/runs-charts/
https://support.illumina.com/bulletins/2017/02/how-much-phix-spike-in-is-recommended-when-sequencing-low-divers.html

What is a phred quality (Q) score?

- Measures the probability that a base is called incorrectly
- $Q = -10\log_{10}P$

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%

https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf
https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html

Where do QC checks occur in the bioinformatics pipeline?

FASTQ File Generation

bcl2fastq

Pre-Alignment Data Processing

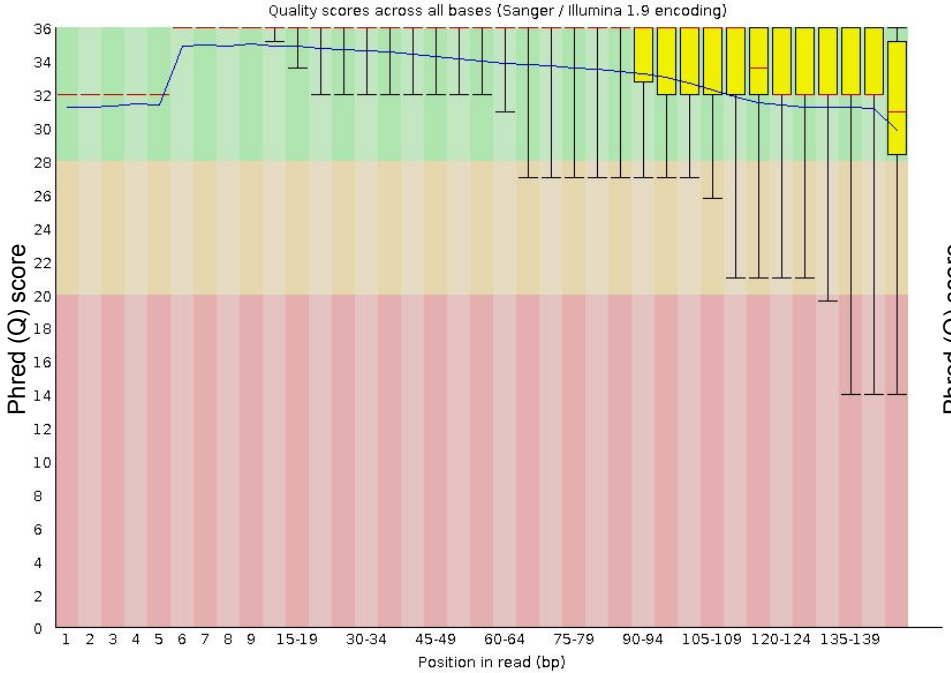
Seqyclean
Fastp
FastQC

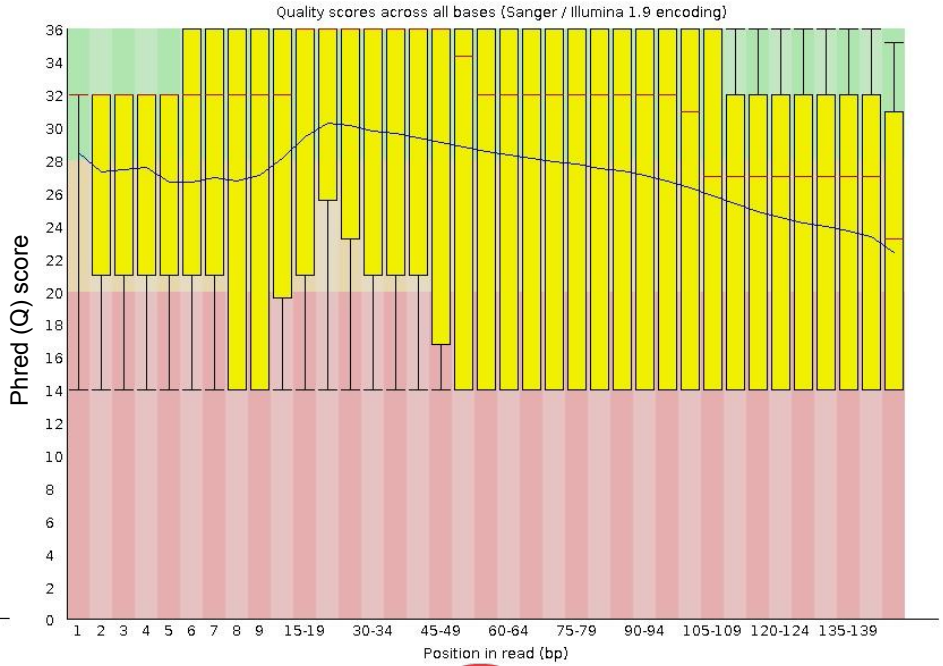
Pre-Alignment QC Check

- FastQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- Bioinformatics tool for visualizing quality of raw read sequencing data
- Generates a summary report
  - Basic statistics
  - Per base sequence quality
  - Per tile sequence quality
  - Per sequence quality scores
  - Per base sequence content
  - Per sequence GC content
  - Per base N content
  - Sequence length distribution
  - Sequence duplication levels
  - Overrepresented sequences
  - Adapter content

Andrews S. FastQC: A quality control tool for high throughput sequence data. 2010.

FastQC - Per base sequence quality - with increasing read length quality declines



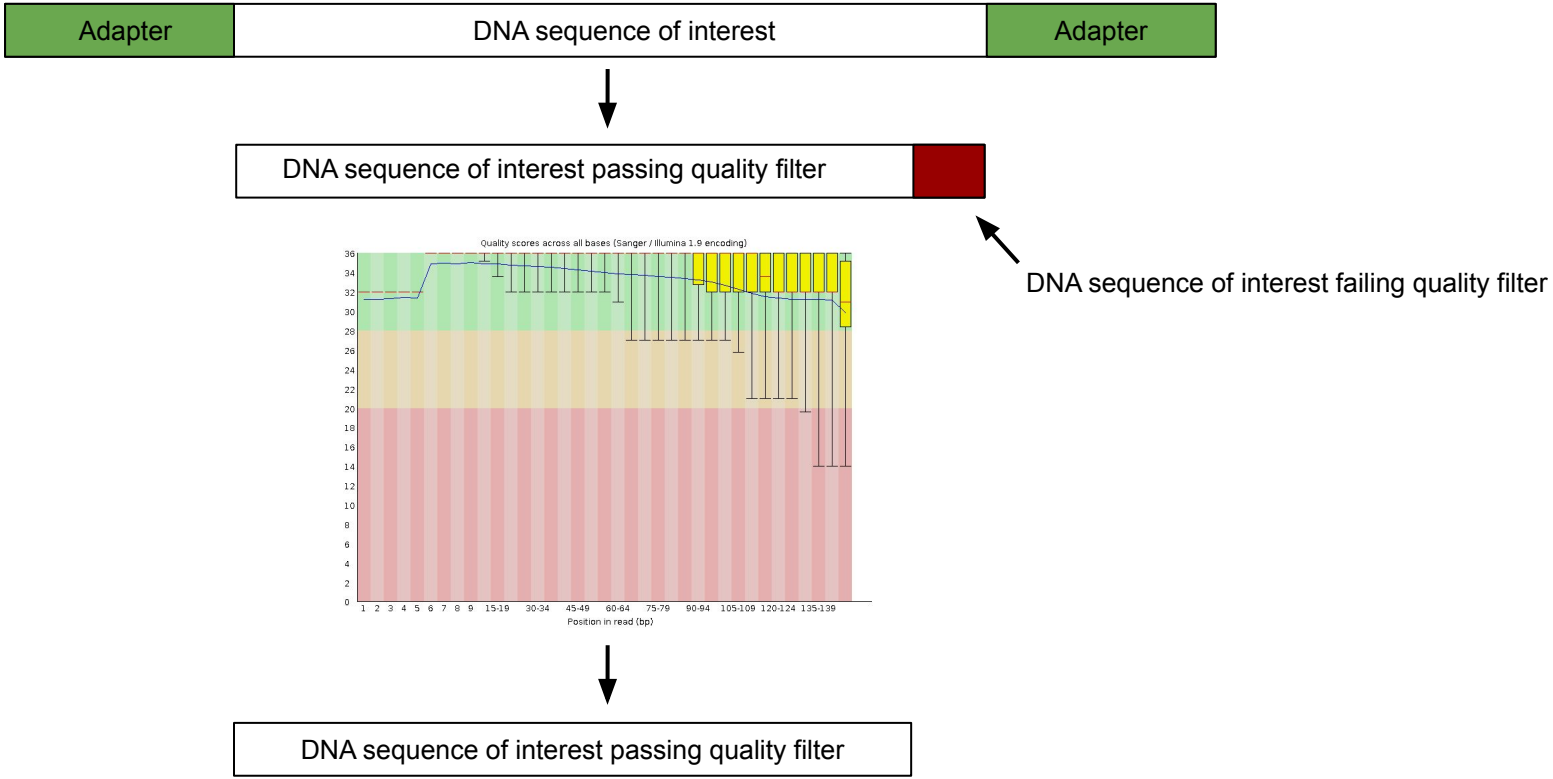


Read Trimming

- Adapter trimming
  - Removal of adapter sequences from reads
- Quality read trimming
  - Removal of low quality bases from reads
- Bioinformatics tools
  - Seqyclean https://github.com/ibest/seqyclean
    - `seqyclean -1 R1.fastq.gz -2 R2.fastq.gz -qual -o seqyclean_output/`
  - Fastp https://github.com/OpenGene/fastp
    - `fastp -i R1.fastq.gz -I R2.fastq.gz -o fastp_R1.fastq.gz -O fastp_R2.fastq.gz -h fastp.html -j fastp.json`

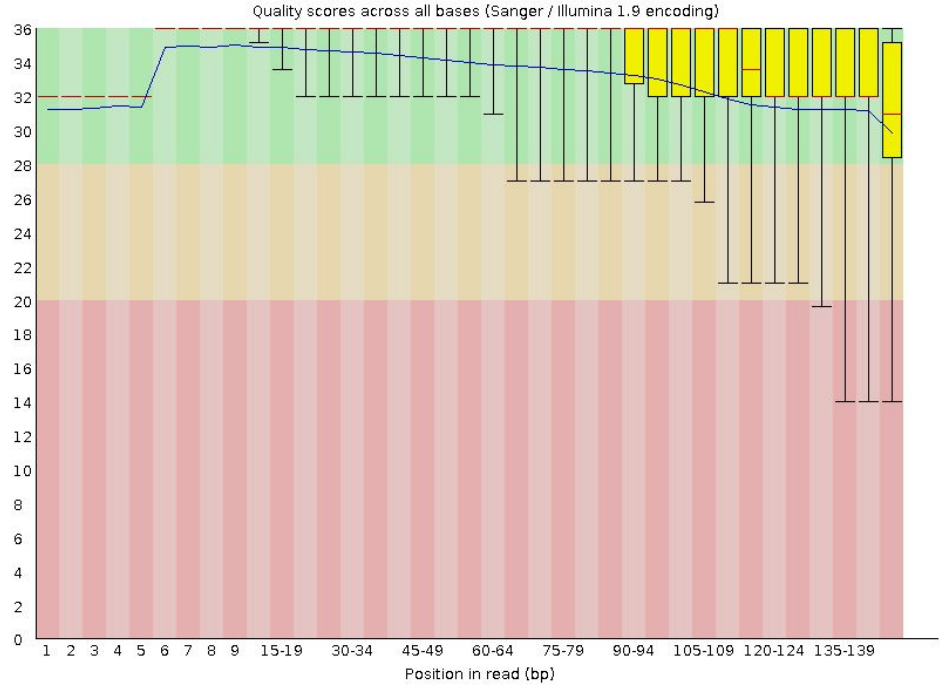
Zhbannikov IY, Hunter SS, Foster JA, Settles ML. SeqyClean: A Pipeline for High-throughput Sequence Data Preprocessing. Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics; 2017. New York, NY, USA: ACM; 2017. p. 407-16.
Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i90.

Adapter and Quality Trimming

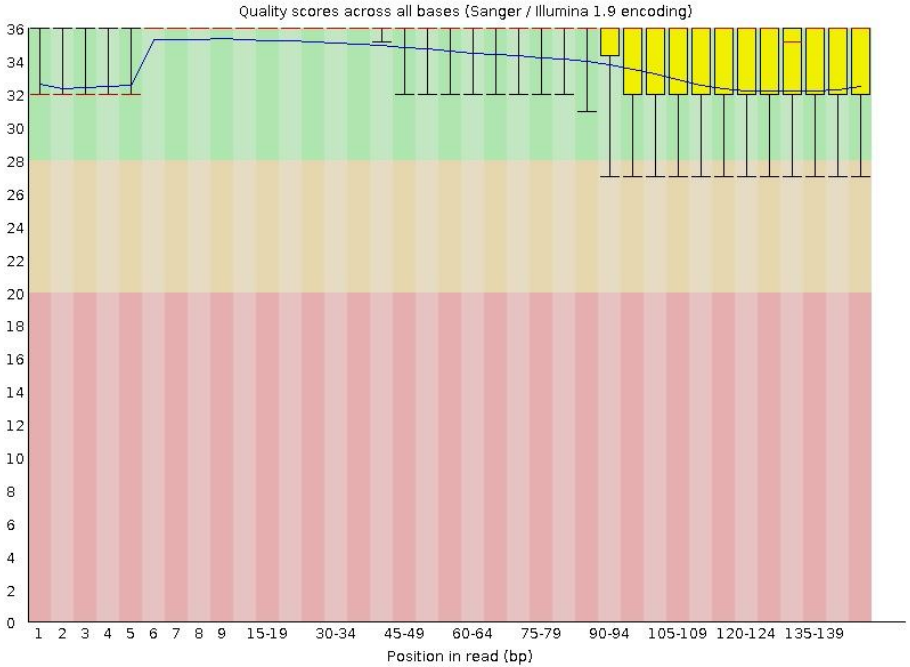


Aggressive Trimming Can Lead to Shorter Reads & Ambiguous Alignments

Result of Read Trimming

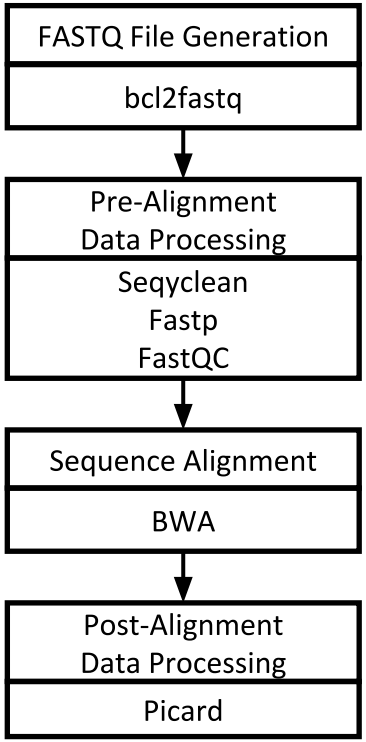


Before Seqyclean



After Seqyclean

Where do QC checks occur in the bioinformatics pipeline?

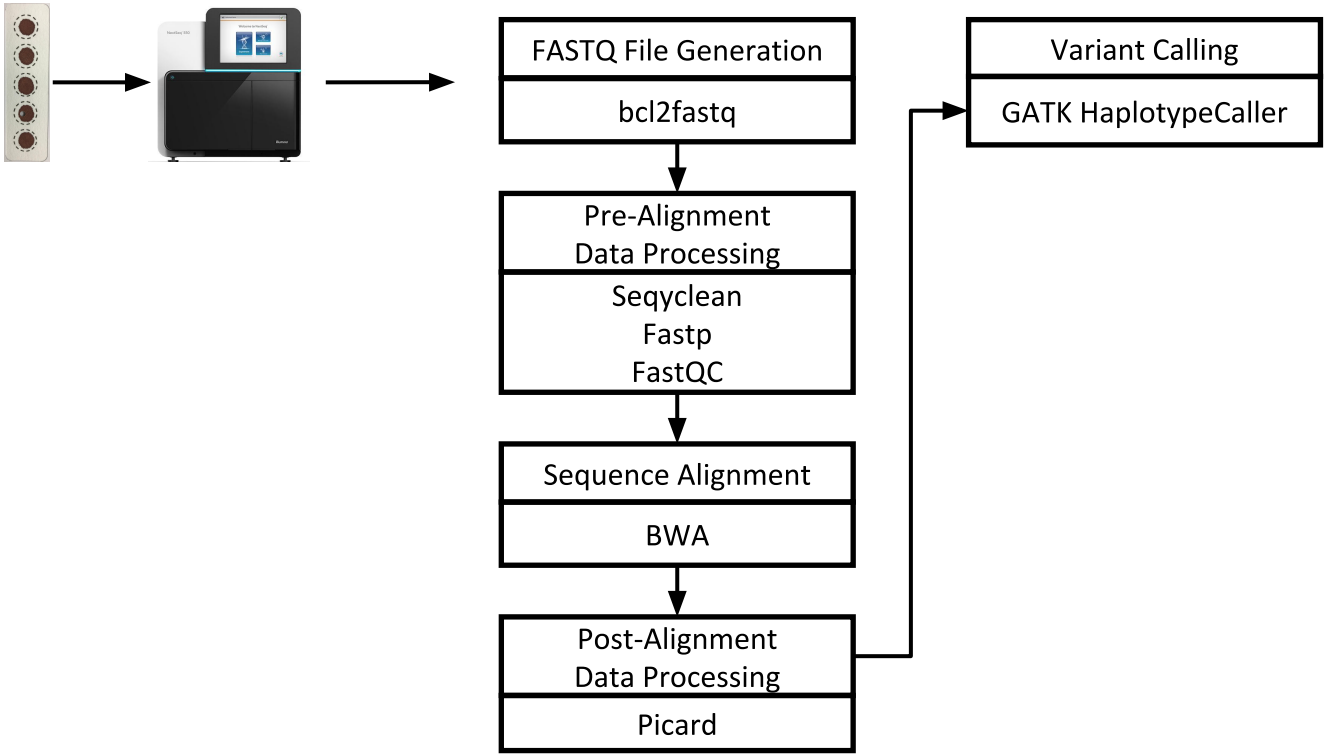


Post-Alignment QC Check

- Coverage (depth of coverage) - number of reads that align to a region of the reference sequence
  - Mean target coverage
  - % of target bases with ≥ 20x coverage
  - % of target bases with ≥ 30x coverage
- Percent duplication - Duplicates are sets of read pairs with the same alignment start and end
- Tools used to generate these statistics
  - Picard CollectHsMetrics
  - Picard MarkDuplicates
  - GATK DepthOfCoverage

http://broadinstitute.github.io/picard/
https://qcb.ucla.edu/wp-content/uploads/sites/14/2016/03/GATKwr12-2-Marking_duplicates.pdf

Where do QC checks occur in the bioinformatics pipeline?



Post-Variant Calling QC Check

- Coverage of variant
- Allelic fraction
  - Percentage of reads representing each allele in sample
- Additional filtering parameters

Parameter	**Definition**
QualByDepth (QD)	Variant call confidence normalized by depth of sample reads supporting a variant
FisherStrand (FS)	Strand bias estimated using Fisher's Exact Test
RMSMappingQuality (MQ)	Root mean square of the mapping quality of reads across all samples
MappingQualityRankSumTest (MQRankSum)	Rank sum test for mapping qualities of ref versus alt reads
ReadPosRankSumTest (ReadPosRankSum)	Rank sum test for relative positioning of ref versus alt alleles within reads

CLIA certification and the transition into clinical testing

- Performance characteristics used by CLIA for evaluation are not directly usable to evaluate NGS testing methodologies
- CLIA does not have guidelines for validation of bioinformatics pipelines
- Some expert groups have generated recommendations:
  - Use of biological reference materials (NIST standard NA12878)
  - Use of reference data (NA12878 high confidence variant call dataset)
  - Use of synthetic data modeling real-world variants
  - Proficiency testing

How do CLIA performance characteristics compare between methodologies?

Traditional quantitative biochemical assay - IRT

- Calibrators (quantitative assay)
- Standard curve
- Multilevel controls (kit component; low, mid high)
- Multilevel CDC controls
- Background control
- Limit of detection
- Accuracy
- Precision

Next-generation sequencing: 3 billion/30M bases

- Quantitative?
- Controls?
- In process controls? What is a positive control?
- What is the reference?
- NTC library prep only

Standards for NGS in Clinical Laboratories

- CLIA
- CAP
  - Aziz et al (2015) College of American Pathologists' laboratory standards for next-generation sequencing clinical test
  - Roy et al (2018) Standards and guidelines for validating next-generation sequencing bioinformatics pipelines
- FDA
  - Luh and Yen (2018) FDA guidance for next generation sequencing-based testing: balancing regulation and innovation in precision medicine
- ACMG
  - **Gargis et al (2012) Assuring the quality of next-generation sequencing in clinical laboratory practice**
  - Rehm et al (2013) ACMG clinical laboratory standards for next-generation sequencing
  - Gargis et al (2015) Good laboratory practice for clinical next-generation sequencing informatics pipelines
- NY Department of Health

  https://www.wadsworth.org/sites/default/files/WebDoc/2080900015/Germline_NextGen_Validation_Guidelines.pdf

How are we addressing CLIA compliance?

- Validation report will include
  - Concordance
  - Precision
  - Proficiency testing
    - Real samples (problem they are “common” & one dimensional)
    - In silico resources
- Variant report will include
  - Variant(s) detected
  - Pipeline version
  - Brief description of NGS method (wet lab and bioinformatics methods)
  - Bioinformatics tools used and versions
  - Variant coverage
- Revalidation of pipeline when changes are made
  - Use standard samples (NIST reference sample NA12878, archived DBS samples with confirmatory result information)
  - Synthetic data

Summary

- Quality control checks occur at multiple points throughout the NGS pipeline
- Sequencing errors can also occur at multiple points throughout the pipeline so QC is essential to attempt to mitigate these errors
- QC parameters must be defined prior to validation and implementation of NGS assay
- Regulatory agencies lack guidance for assessment and validation NGS assays, especially bioinformatics pipelines
- Expert groups have provided recommendations and guidelines which provide a solid starting point for validating NGS assays

Acknowledgements

Utah Department of Health Newborn Screening Program

- Andy Rohrwasser
- Kim Hart
- Bryce Asay

Utah Department of Health Infectious Disease

- Kelly Oakeson
- Erin Young

University of Utah Department of Biomedical Informatics

- Karen Eilbeck
- David Sant
- Jordan Little
- Krystal Chung