NGS Bioinformatics and File Types

Author Information and Affiliations:

W. Bailey Glen Jr., Ph.D

Medical University of South Carolina

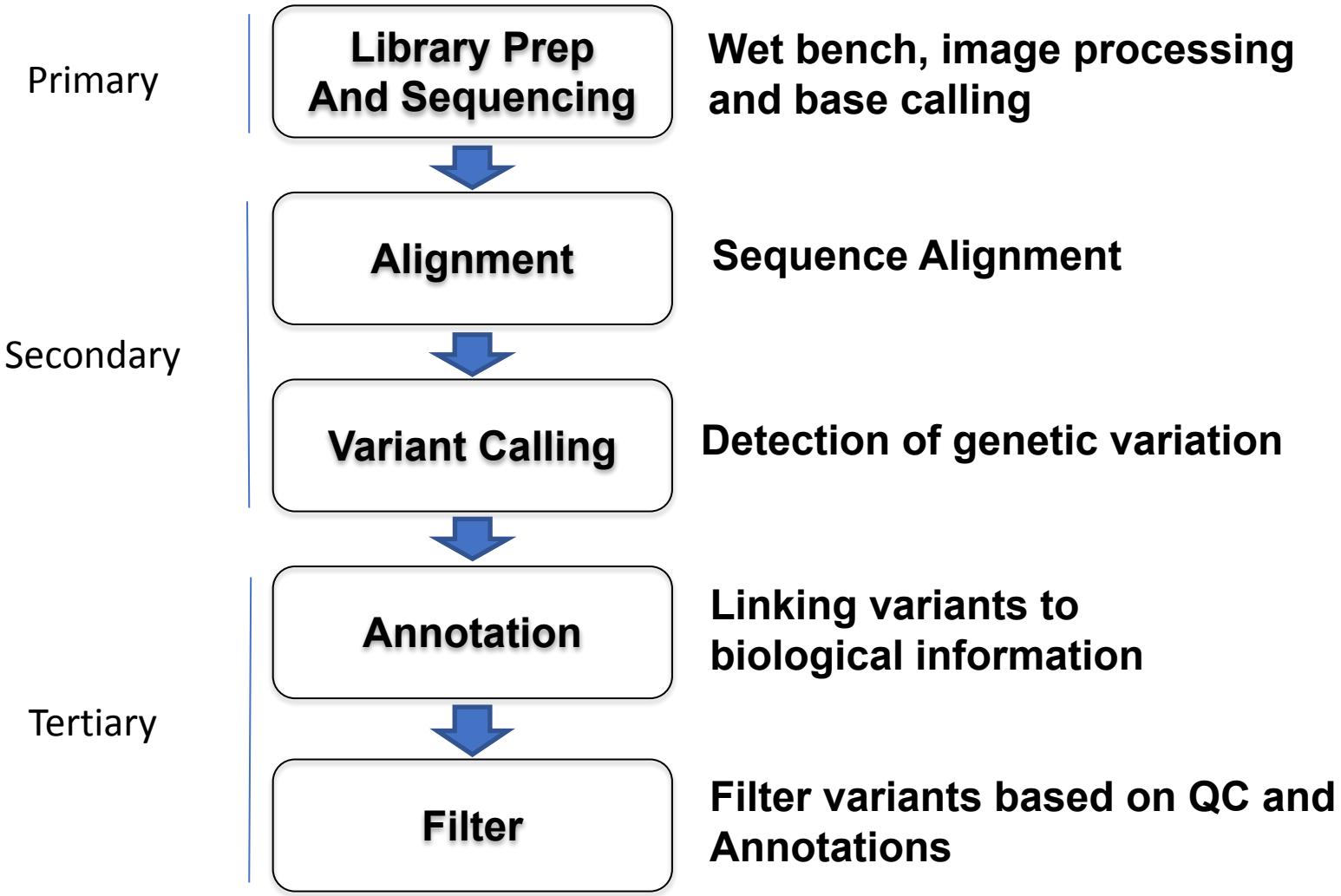
Pathology and Laboratory Medicine Department

Pathology Informatics Division

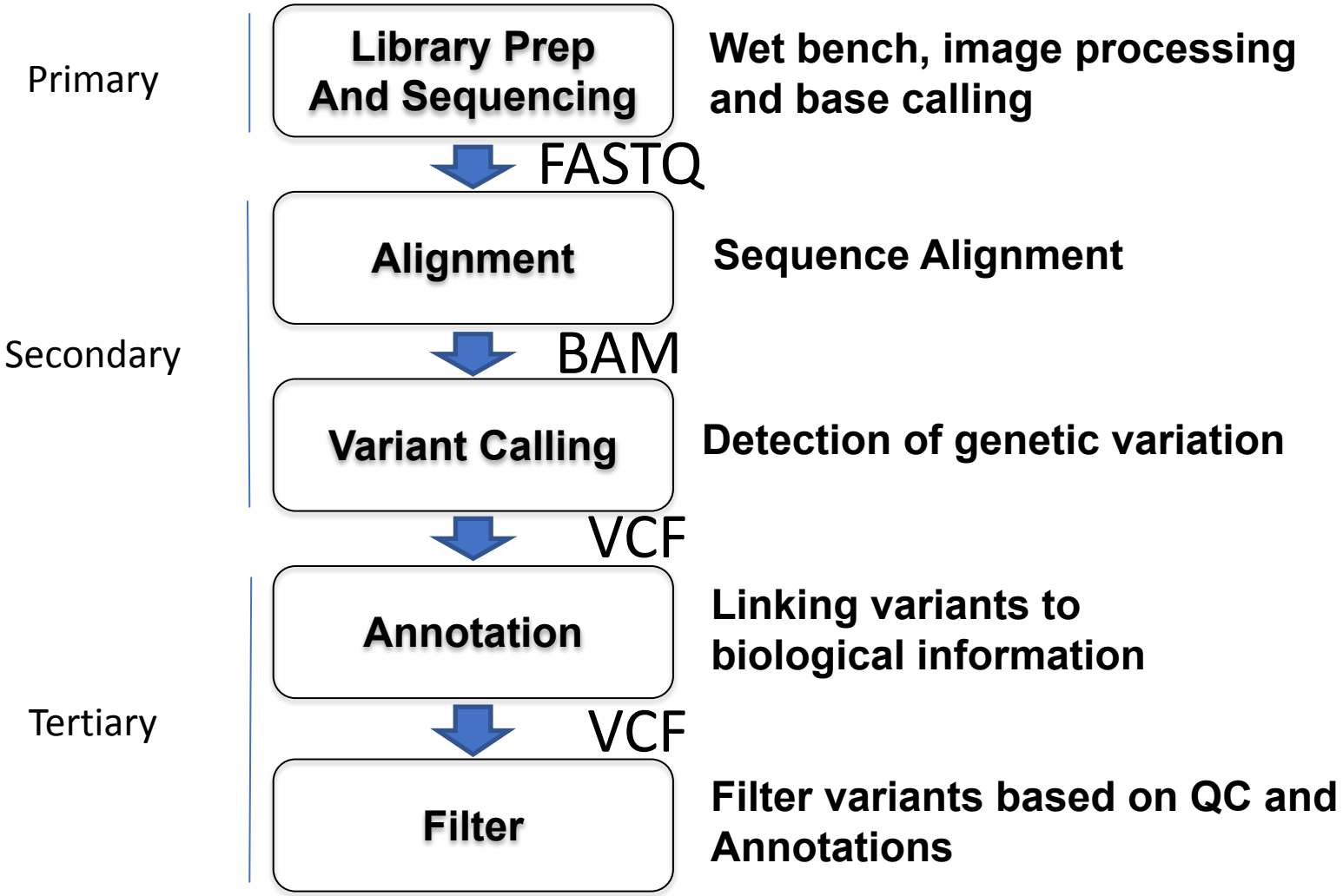
Date Created: 05/01/2024

CANCER GENOMICS CONSORTIUM

NGS Analysis Overview



NGS Analysis Overview: Common File Types



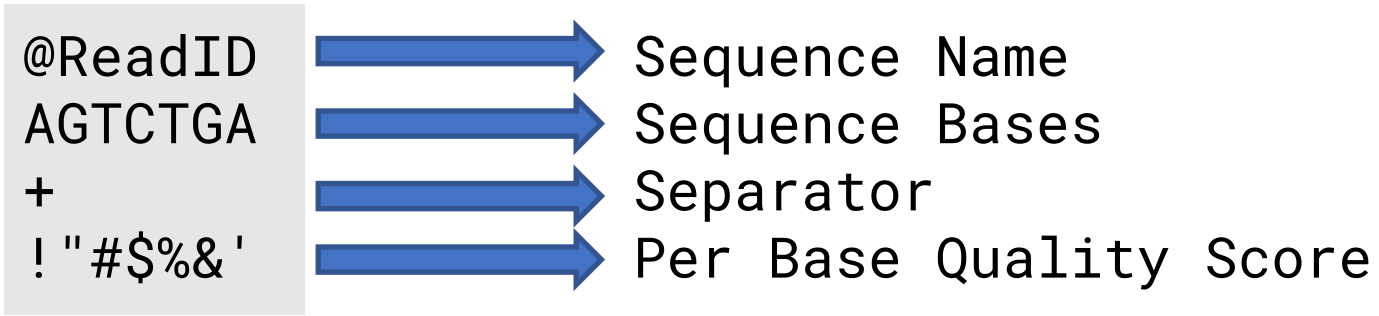
Primary Analysis Files

- FASTA
  - Text-based file format containing nucleotide (or amino acid) sequence(s)
  - Primarily used for reference or consensus sequences

- **FASTQ**
  - FASTA file with per-base quality metrics
  - Primary output file of most modern NGS Sequencing
  - Frequently stored in a compressed format (fastq.gz)
  - Frequently stored as a pair of files per sample (paired-end sequencing)

Example FASTQ File

- Four lines per sequence



Sequence Alignment Files

- Sequence Alignment Map (SAM)
  - Text-based file for storing aligned sequences
- **Binary Sequence Alignment Map (BAM)**
  - Binary version of a SAM file, reducing storage limits and improving read performance
  - Requires specific software to open
  - Frequently paired with BAM Index file (BAI) for rapid file access.

Sequence Alignment Files

- Rich and Complex file type
- Contains the full contents of the aligned FASTQ sequencing data
- Provides extra information including:
    - Alignment Positions
    - Alignment Quality Scores
    - Alternative Alignments
    - Differences from the Reference

Variant Call Files

- Variant Call File (VCF)
  - Text-based file format for storing genetic variation results.
  - Unlike FASTQ and SAM which are organized by sequencing read, VCF files are organized by consecutive genomic position.
  - Aggregates calculated results from many read alignments into a single call.
  - Can contain calls from multiple samples
  - Current specifications can contain multiple types of events (SNP, CNV, Fusion)
  - Flexible file format allowing for the representation of many custom annotations on individual calls
  - Frequently stored in a compressed format (vcf.gz)

Browser Extensible Data (BED) File

- Common text-based format for storing region based genetic information
- Three required fields (and many other optional):
  - Chromosome
  - Start
  - Stop
- Relatively human-readable file type
- Must be careful with indexing (0 vs 1)

Resources

- Detailed explanations of many common file types: https://genome.ucsc.edu/FAQ/FAQformat.html
- The Global Alliance for Genomics and Health (GA4GH) Large Scale Genomics (LSG) Work Stream maintains many of these file types: https://www.ga4gh.org/our-products/#

  https://github.com/samtools/hts-specs