Genome Variant Analysis Review

2016/05/26

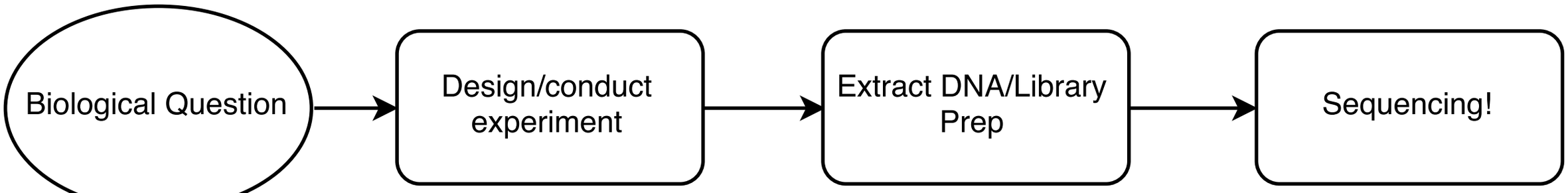
Sean P Leonard and Daniel Deatherage

Goals

- Student Goals:
    - Learn exactly how to analyze my data and have it done by the end of the course
- Our Goals:
    - Introduce you to overall methods, techniques, concerns for identifying genomic variants using NGS. Provide cookie-cutter tutorials for *some* use cases to get you started.

Overview

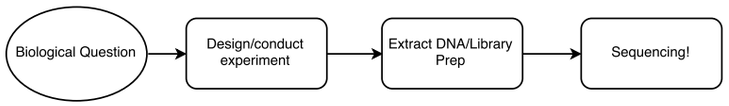
Extraction, shearing, enrichment



Technologies
(illumina vs nanopore, etc)

Read length, machine (miseq vs hiseq vs nextseq)
Sequencing facility

Steps for GVA



unix

Reads from Sequencing Center

fastQC

Check quality and trim adapters

bowtie2

Map to Reference

samtools

Call Variants

Unix, R

Data analysis

Variant filtering

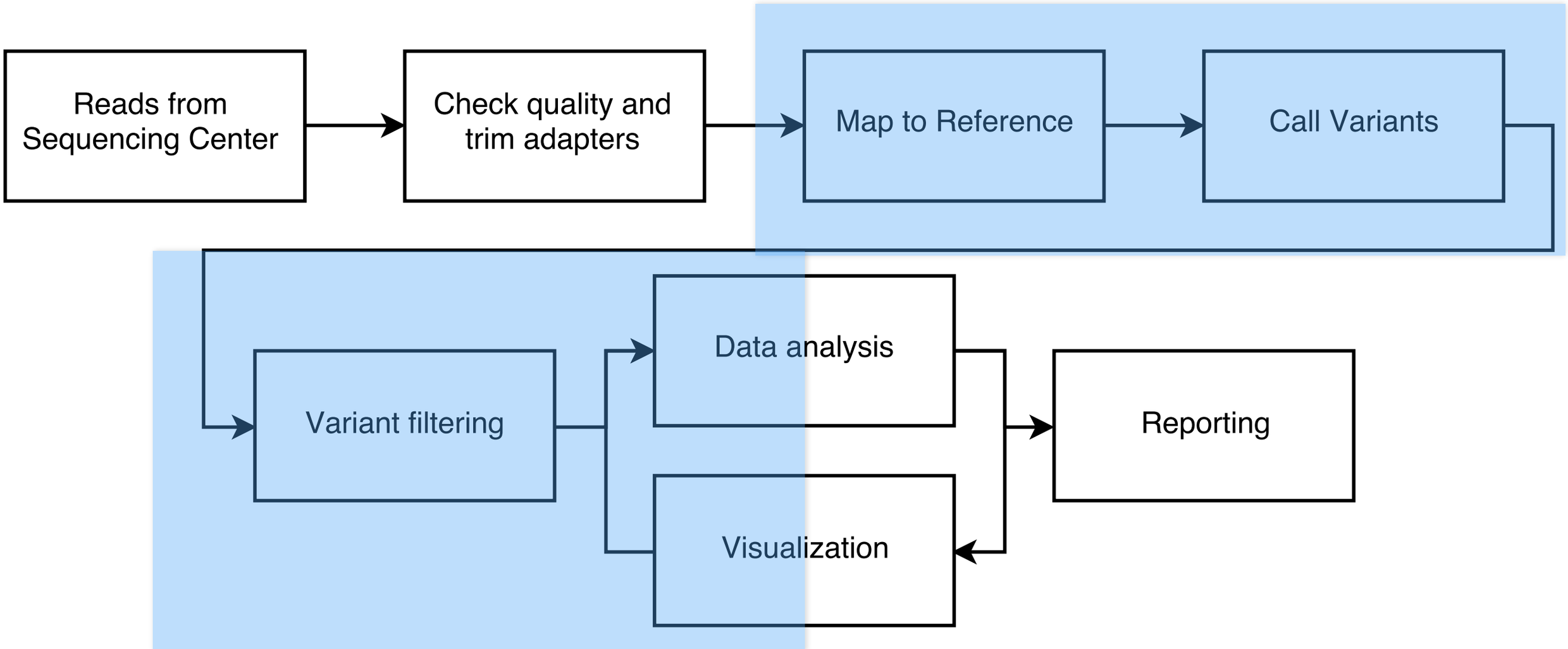
Reporting

samtools
annovar

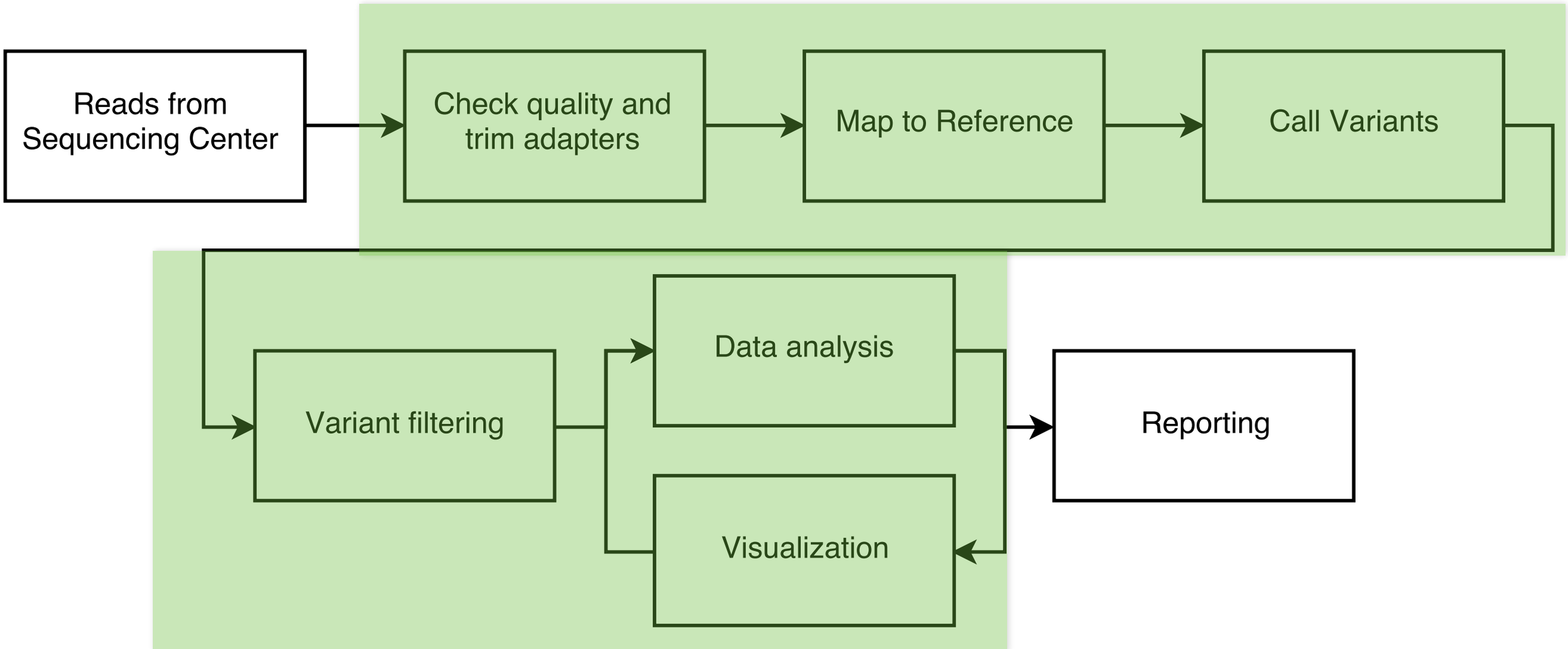
Visualization

IGV

microbial all-in-one: breseq



eukaryotic all-in-one: GATK



Common File Formats

.fasta -> sequences, usually obtained from database (not reads, no quality score)

.fastq -> unmapped reads from sequencer, 4 line records, contain quality information

.sam/.bam -> human readable (and binary) format for mapped reads. results of mapping programs (bowtie2) which align the .fastq reads to a .fasta reference

.vcf/.bcf -> variant call format. Shows variants using .sam files. Can be further analyzed/filtered

Further Resources (online)

- Galaxy : https://usegalaxy.org
- BioIteam website (more tutorials, info from other classes): https://wikis.utexas.edu/display/bioiteam/Home
- Coursera: Genomic data science : https://www.coursera.org/specializations/genomic-data-science

Further Resources (on campus)

- CCBB Short Courses (returning Fall 2016)
- Peer-led working groups (Fall and Spring)
- CCBB Open Coding Hour: Wednesdays 2-3pm, GDC 7th floor seminar room (GDC 7.514)