BIOINFORMATICS 101: GENOME ANALYSIS TOOLKIT (GATK) 4

W. Bailey Glen Jr.

02/26/2017

Broad Institute

- Independent research center partnered with Havard and MIT
- Cambridge, MA
- Broad Genomics
  - Large Scale Sequencing Core
  - One 30X human whole genome every 12 minutes
  - Computation Model and Software Development

Software Tools

- GATK
- Integrated Genome Viewer
- Hail
- Tumor Portal

GATK

- Analyzing sequencing analysis for genetic variation
  - Primarily Focused on Small Variants
    - Copy number changes recently added
  - Primarily Focused on DNA
    - RNA best practices have been previously defined
  - Primarily Constitutional
    - Tumor/Normal developed
    - Tumor only workflow now described
- Not
  - RNA expression analysis
  - Linkage disequilibrium or association testing

GATK 4 – Refactoring and more

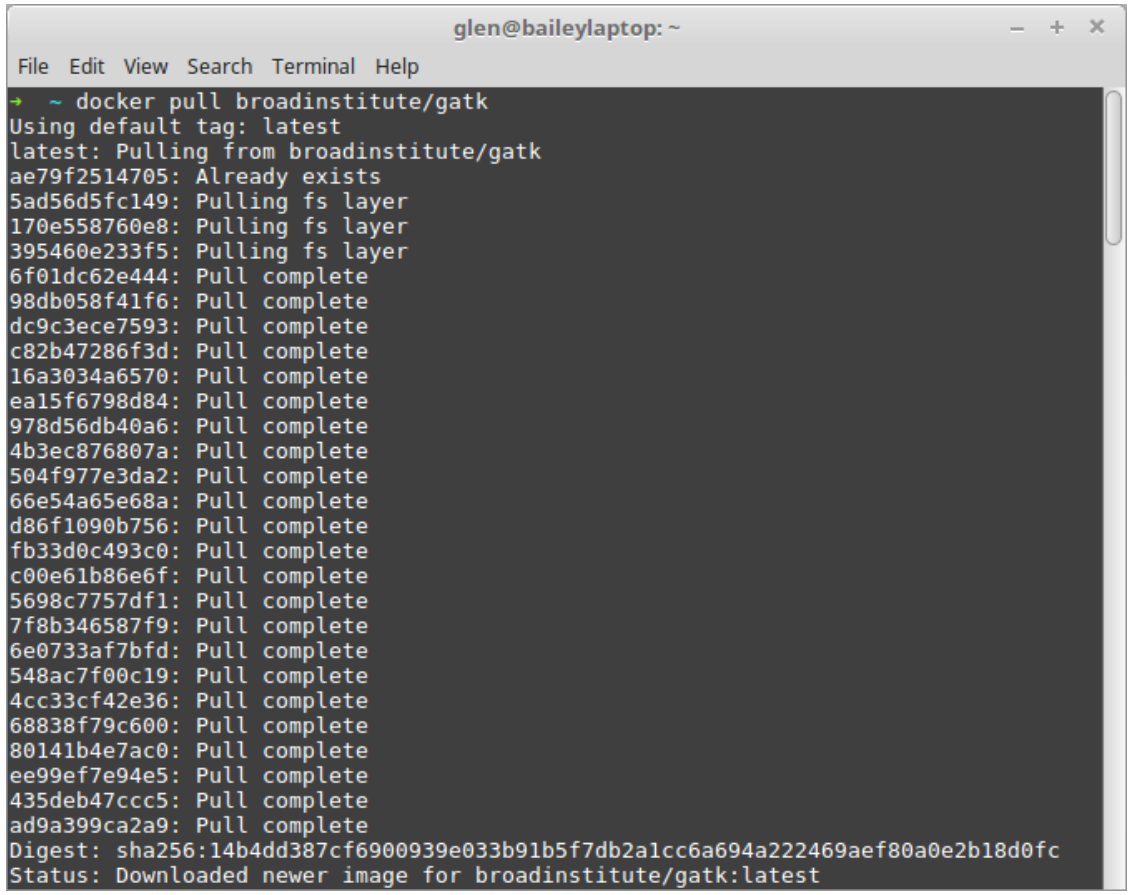
- In development for years
- Refactored GATK 3
  - In many cases, same commands producing equivalent results
- Partnered with Intel
- Large focus on computer science and software implementation
  - Performance
  - Deployment
  - Scalability
- Opensource

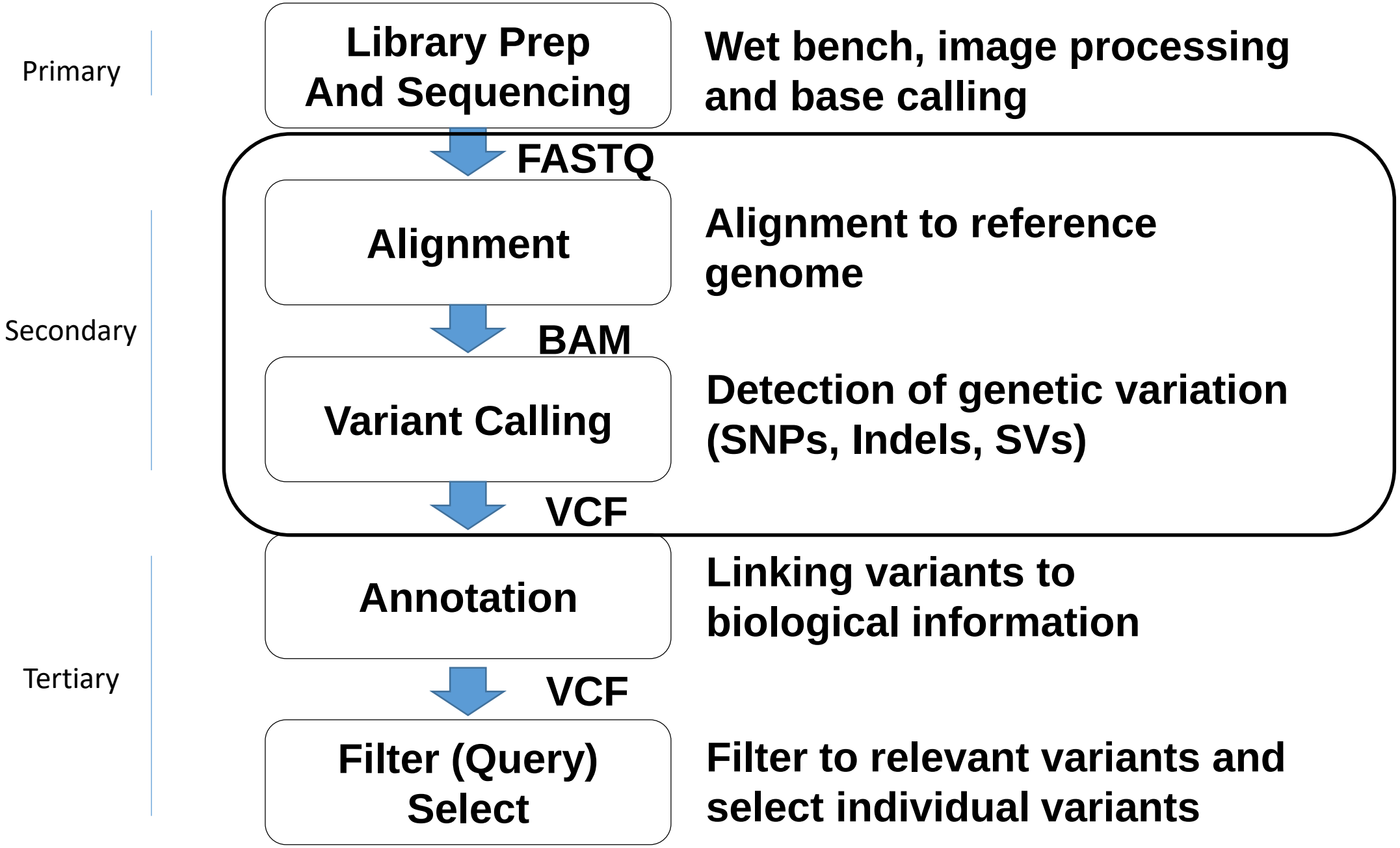
Component Software

- GATK
  - Java
- Dependencies
  - JVM
  - BWA
  - Picard tools
- Available as Docker or JAR

Docker

- Container Virtualization
  - Runs on linux, mac, windows
  - Creates an isolated controlled virtual computer on your computer
  - Comes with packages installed and version controlled
  - Dockers are "easily" deployed
- Emerging standard for sharing bioinformatic tools
- One new tool to learn, improve using many tools




Primary
Library Prep And Sequencing
Wet bench, image processing and base calling
FASTQ
Secondary
Alignment
Alignment to reference genome
BAM
Variant Calling
Detection of genetic variation (SNPs, Indels, SVs)
VCF
Tertiary
Annotation
Linking variants to biological information
VCF
Filter (Query) Select
Filter to relevant variants and select individual variants

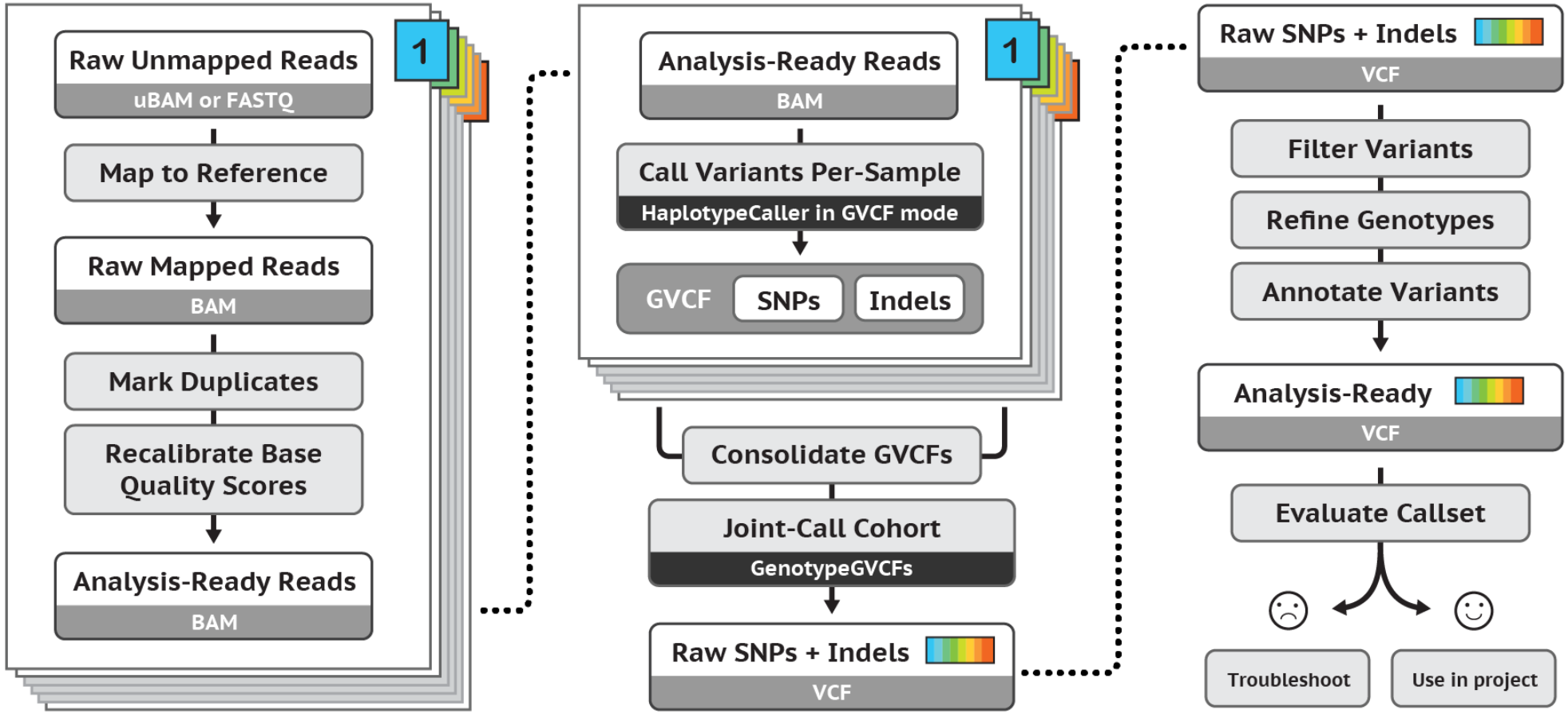
GATK Best Practices

- https://software.broadinstitute.org/gatk/best-practices/

- Small Nucleotide Polymorphisms
  - Germline SNPs + Indels
  - Somatic SNVs + Indels
  - RNAseq SNPs + Indels

- Copy Number Variations
  - Germline CNVs
  - Somatic CNVs

Example Genome Analysis Toolkit Workflow



Major File Types

- Unaligned Reads
  - FASTQ or uBAM

```
@HWI-ST1001:97:D0E5LABXX:7:1101:1180:2041 1:Y:0:
AATCNNTAGTTGGTGGCGTAAGGTCGCAAAGTAAGAGCTTCTCGGGCTGCGTCAGGATAGGTCGTATTTGCTCATTTTTCCCCTTCAGTCCACTTCGATTT
+
(079##2(3@@9)@(@##########################################################################################
```

- Aligned Reads
  - BAM

```
HWI-ST1001:97:D0E5LABXX:1:1102:11138:107030
163
chrM
1
41
28S70M3S
=
1
70
CGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGCTA
;39<====>=<<<<<<>=<==<=<6======<=<=>======<=====>>=>><<=<=<=<6=>=>=>=>=>>??=??><?><<>????8<=??>>=;9:8
```

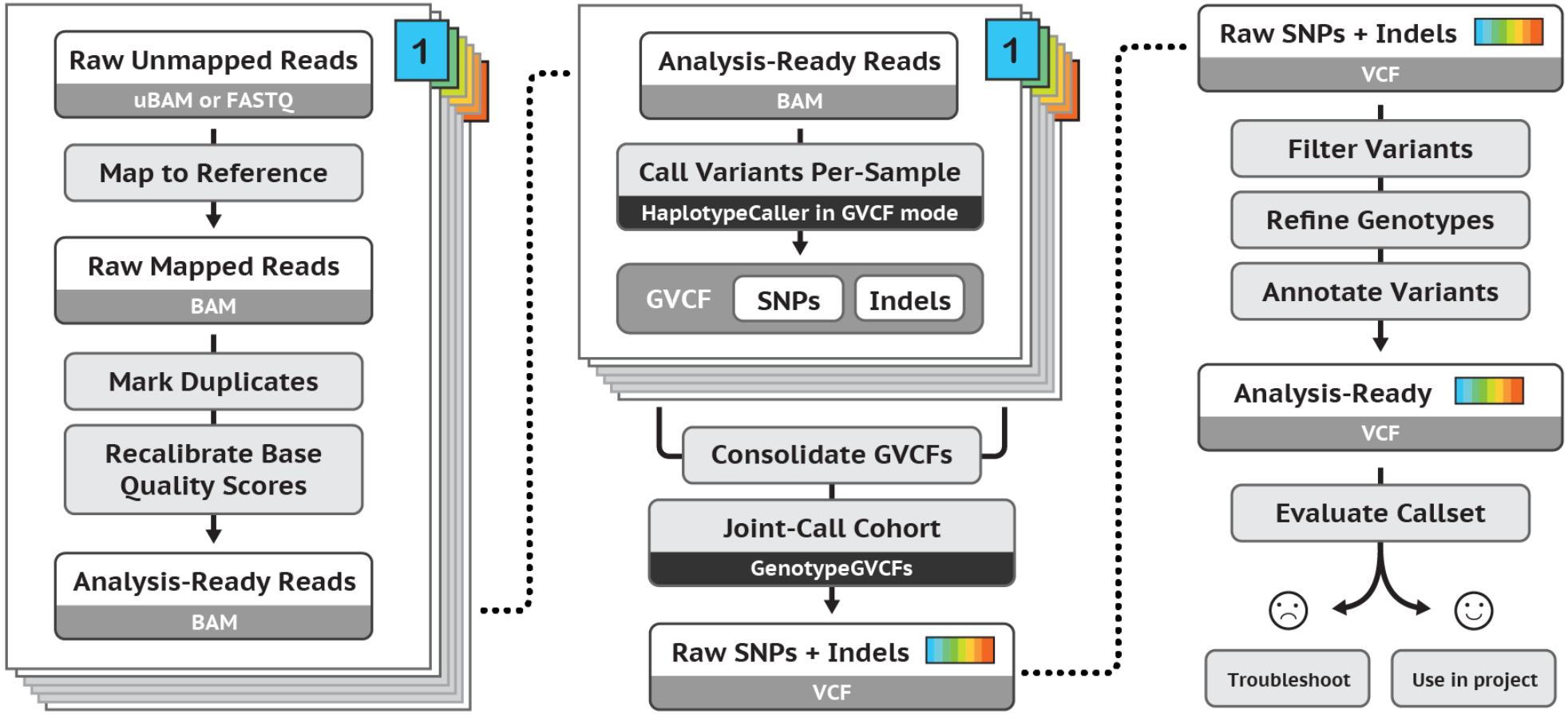
- Variant Call Files
  - GVCF (intermediate)
  - VCF

Alignment Software

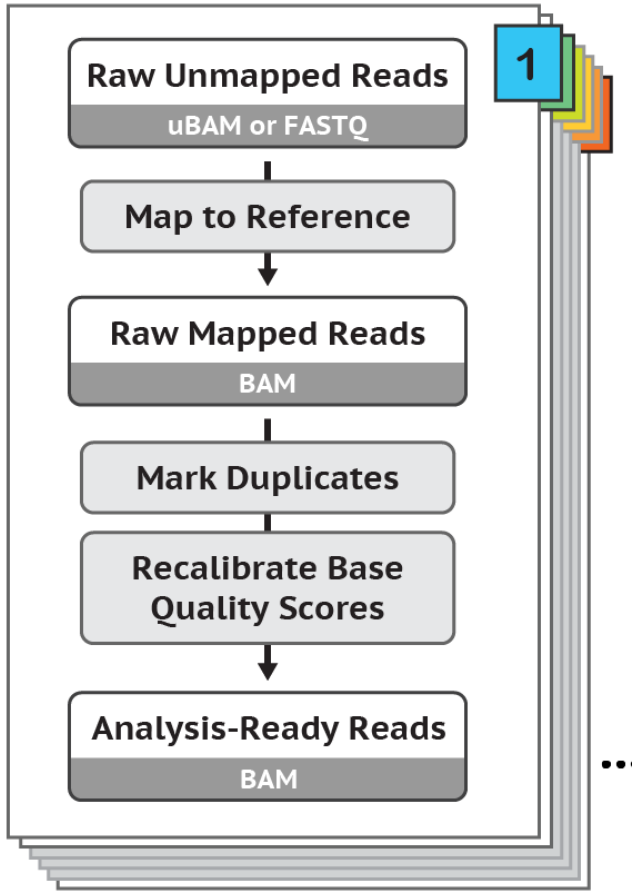
- Aligns reads to Reference Genome
  - Database Search Alignment
    - BLAST
  - Short Read Sequence Alignment
    - BWA-MEM
    - BOWTIE

- Reads can also be aligned to themselves if a reference alignment is missing (*de novo* assembly)
  - Long reads

Constitutional DNA Best Practices



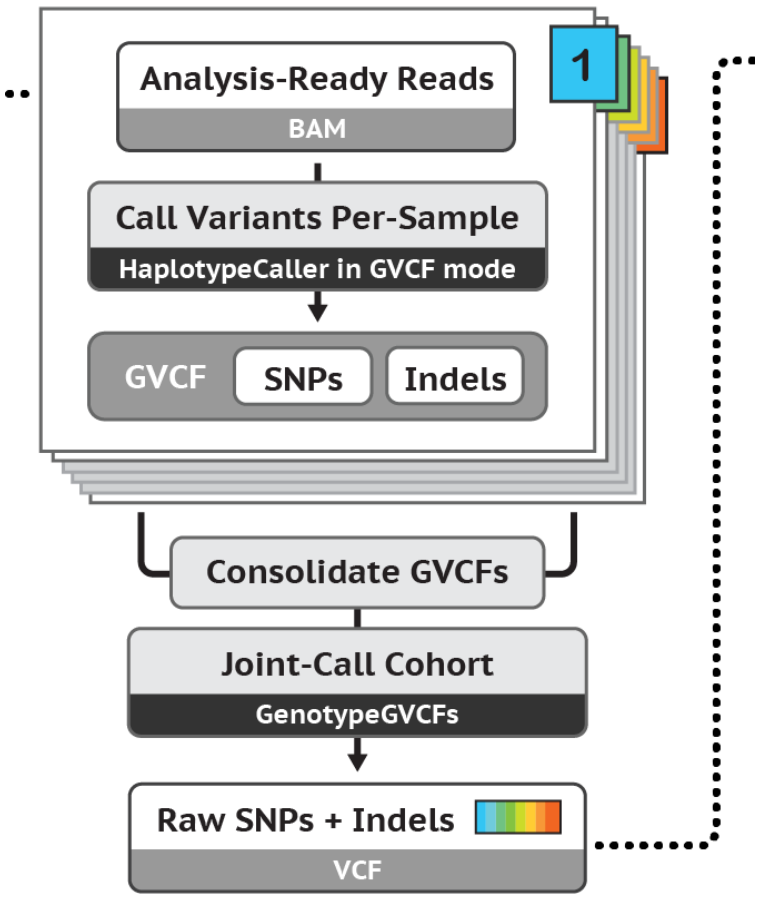
Preprocessing (Alignment)



Starting with unaligned reads

- Align to Reference Gene
  - BWA-MEM for DNA
  - STAR/BOWTIE for RNA
- Remove Duplicate Reads
  - Unless using an amplicon, most duplicate reads are sequencing errors
- Recalibrate Base Quality Scores
  - Statistical approach to improve base calls after completing a run

Alignment



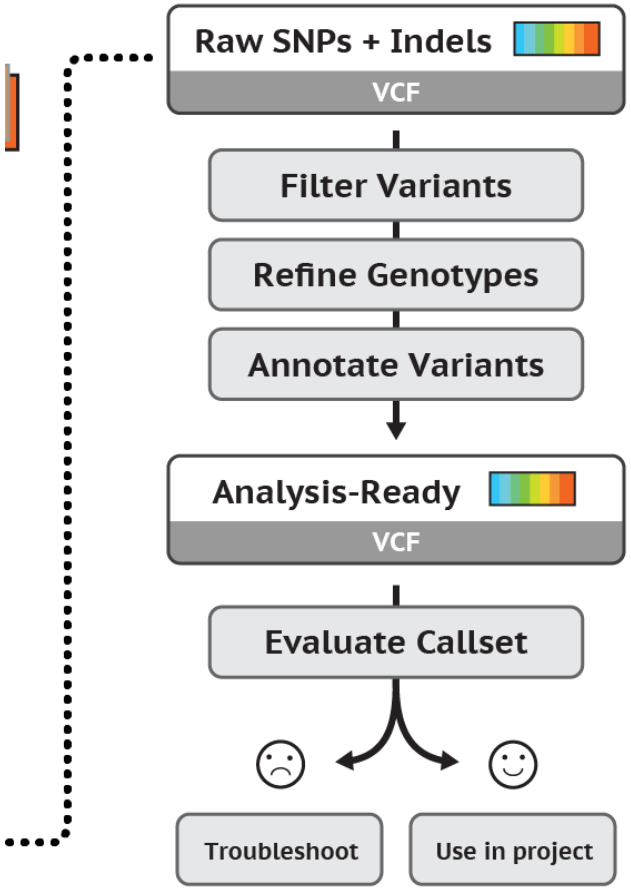
Starting with preprocessed bams

- Generate gVCF for each sample
- Join call gVCFs

OR

- Joint call bams to make VCF

Variant Calling



Starting with VCF

- Filter by Quality
- Refine Variants
  - Phasing (Trio, local within sample)
- Annotate Variants
  - Provide genomic context
- Perform Tertiary