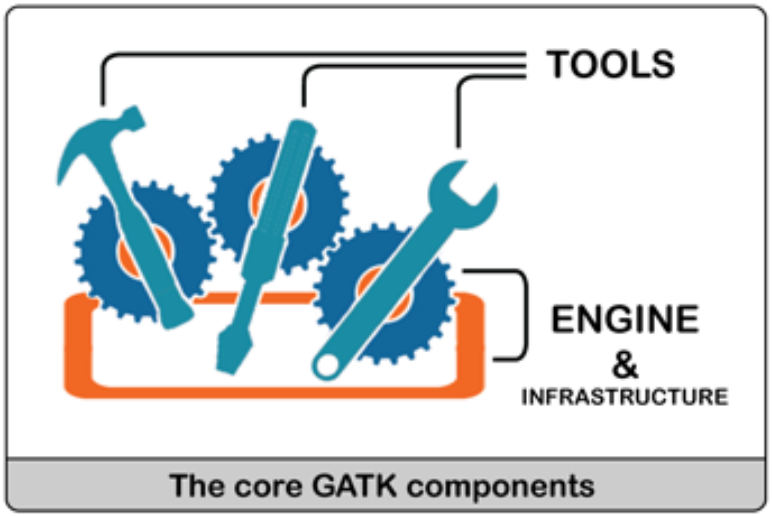
Introduction to the GATK Best Practices and the Broad production pipelines

Overview of the tools, methods
and pipelines for variant discovery

GATK BEST PRACTICES

GATK = Genome Analysis Toolkit

- **Toolkit** focused on variant discovery (SNP & indel)
- **Components:**
  - Engine and infrastructure
  - Tools (walkers)



The core GATK components

- Also a **programming framework** for developing genome analysis software

GATK command syntax

```
java –jar GenomeAnalysisTK.jar –T ToolName \
        –R reference.fasta \
        –I inputBAM.bam \
        –V inputVCF.vcf \
        –o outputs.someformat \
        –L 20:1000000-2000000
```

- Java-based command line tool (see running requirements in FAQs)
- Consult online documentation for details about each tool!
  - Argument names and default values can change
  - Exact arguments depend on the given tool

Picard tools: a companion package to GATK

- broadinstitute.github.io/picard
- Many useful utilities
- Java-based command line interface, much like GATK

- Example 1: sort by genomic coordinate

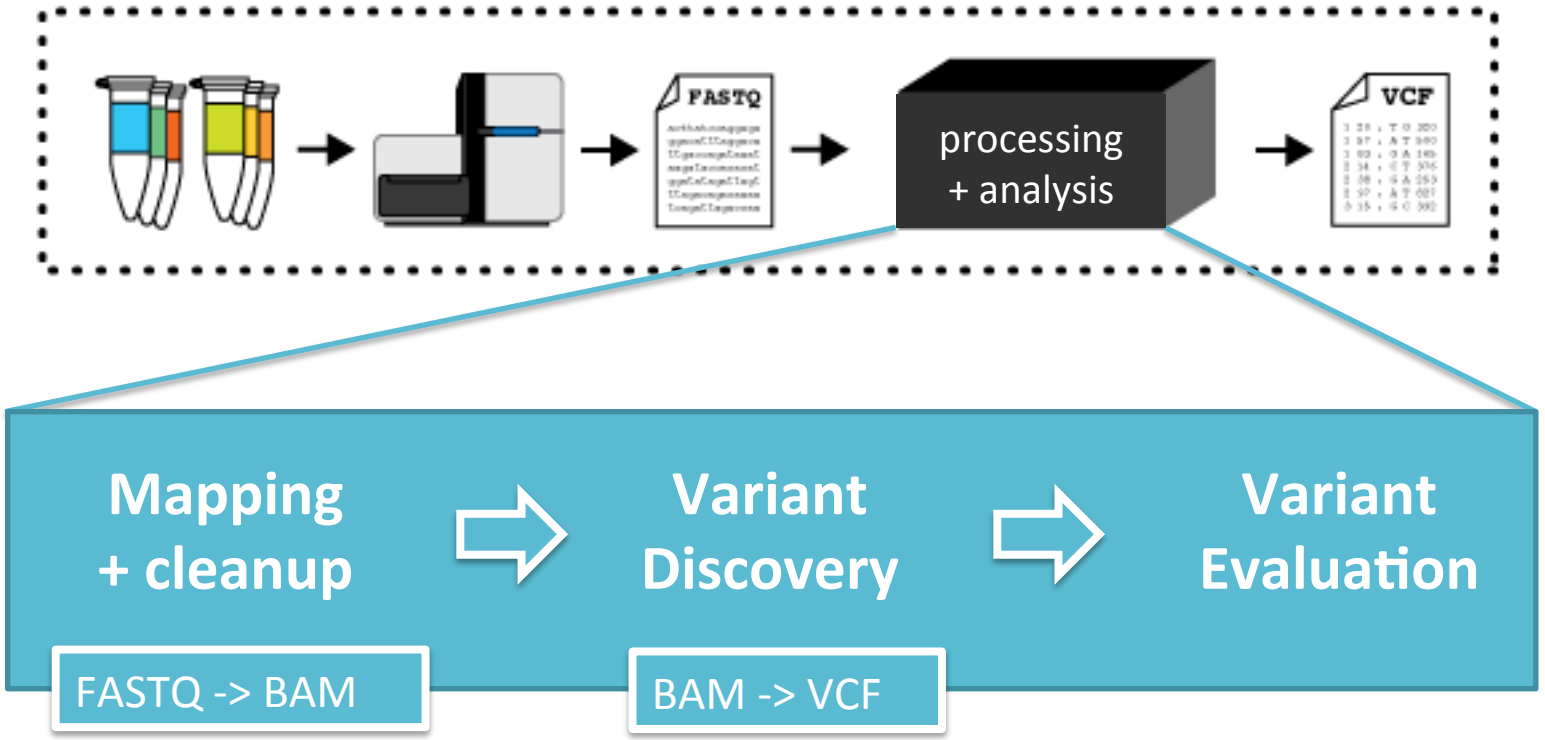
```
java –jar picard.jar SortSam INPUT=unsorted.sam OUTPUT=sorted.sam \
      SORT_ORDER=coordinate
```

- Example 2: mark duplicates

```
java –jar picard.jar MarkDuplicatesWithMateCigar \
      INPUT=unmarked.sam OUTPUT=marked.sam
```

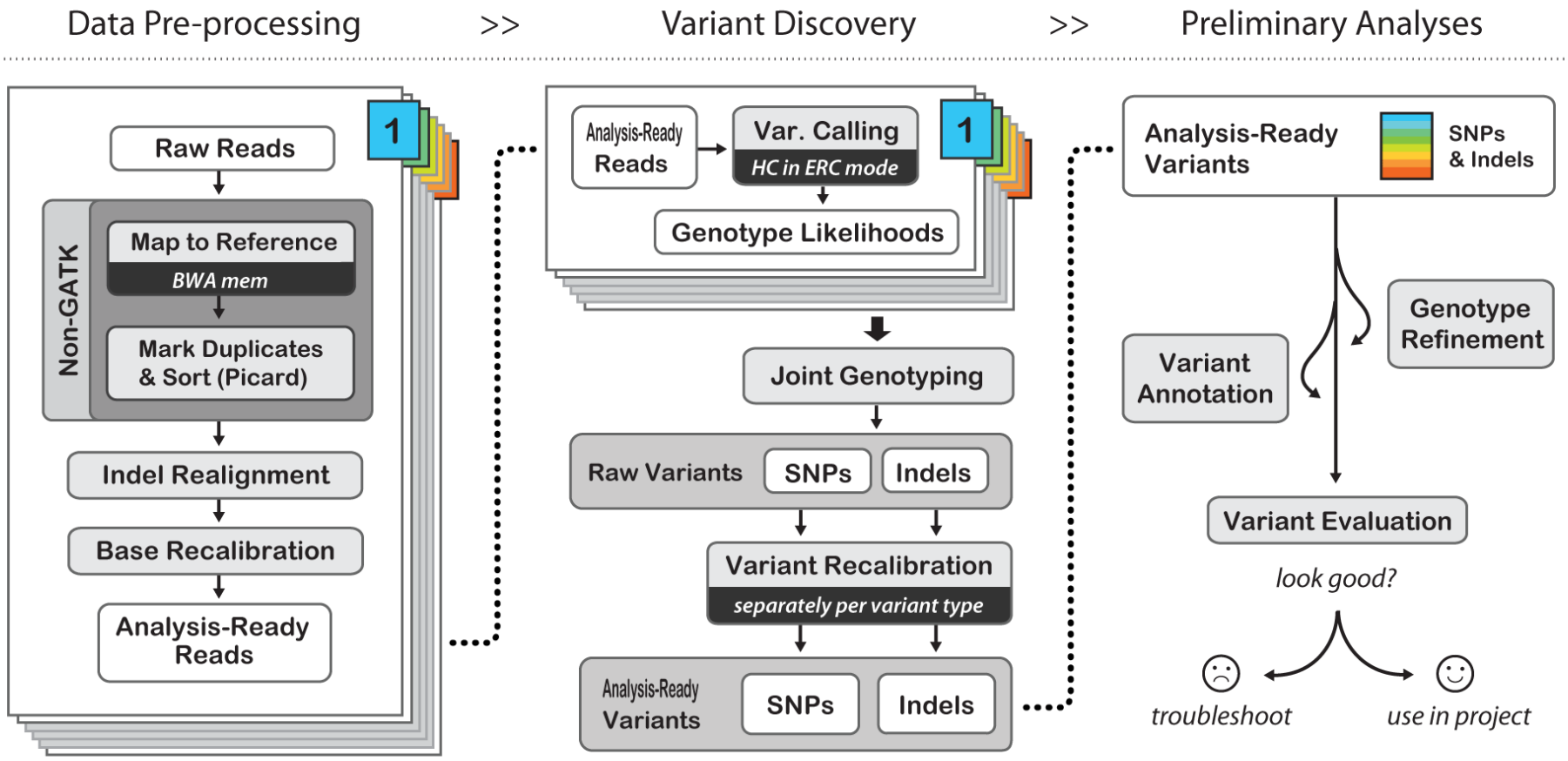
Picard tools are now supported on the GATK forum!

GATK Best Practices cover complete reads-to-variants workflow

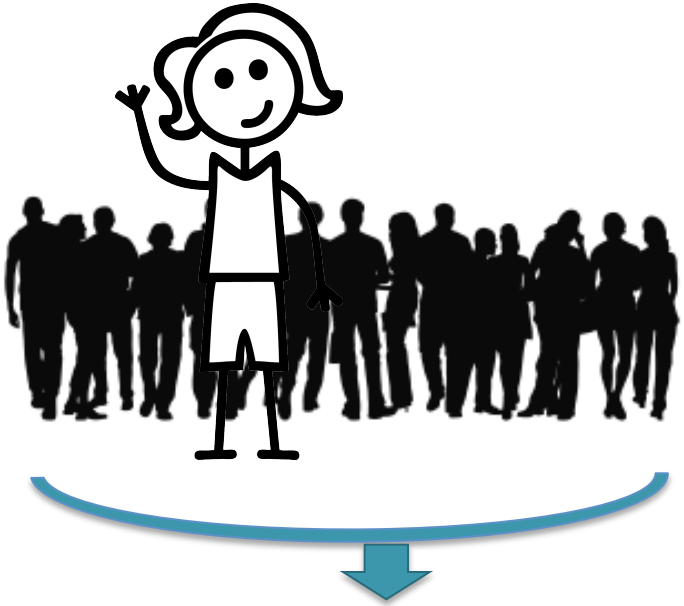


+ other software (BWA, STAR, Picard, Samtools)

Core competency:
Germline Variant Discovery in DNAseq

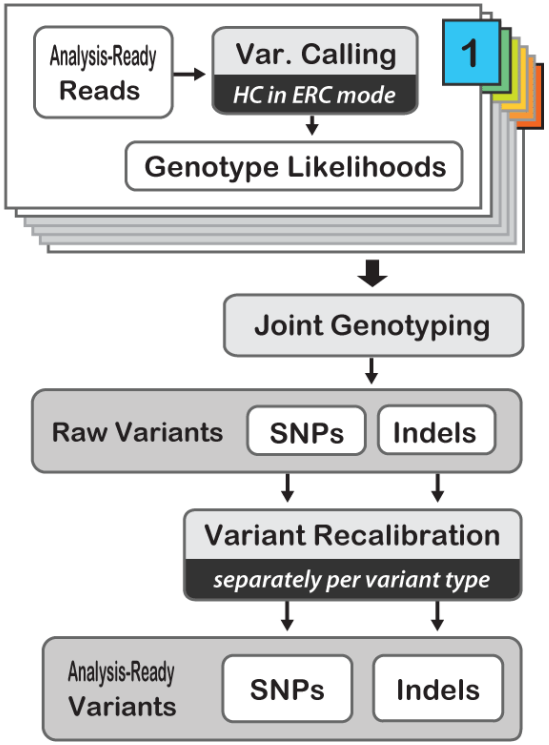


Emphasis on scalable solutions for cohort analysis

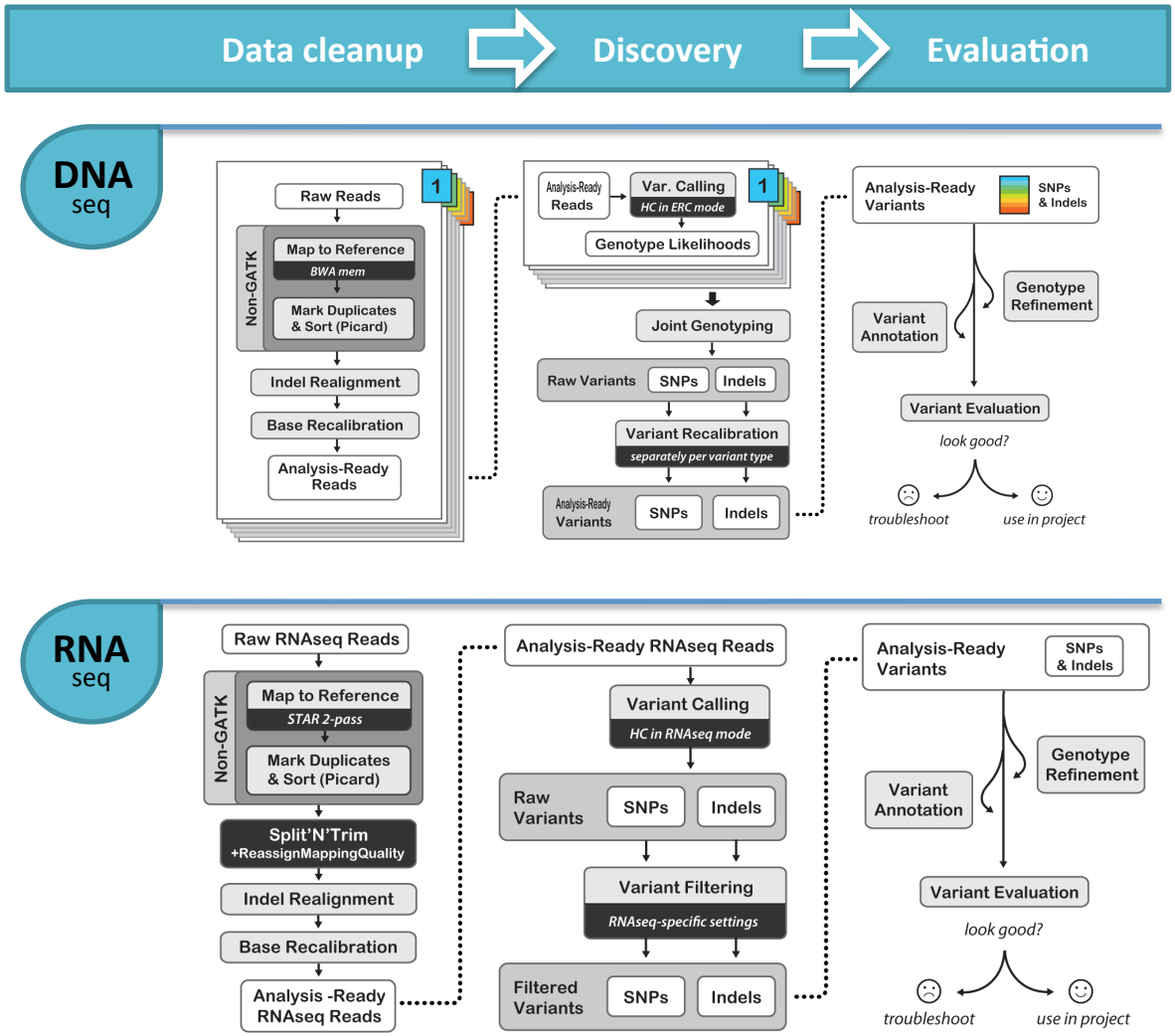


Joint callset

Empowered analysis



Branching out from DNAseq (WGS/Wex) to RNAseq



KEY HIGHLIGHTS

DNAseq workflow

Reference Confidence (ERC –GVCF mode)

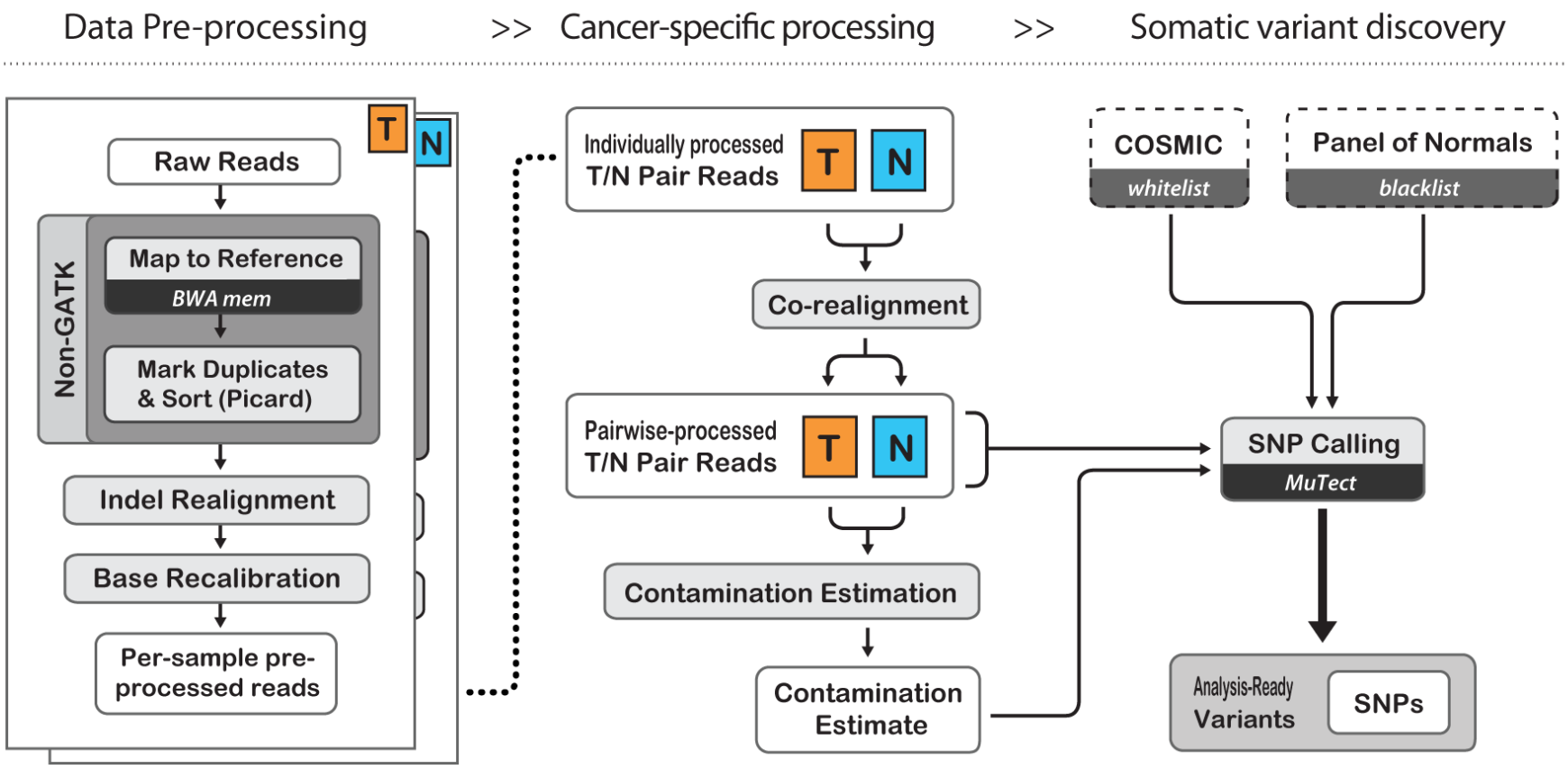
- Scalable & incremental
- Joint analysis of cohorts
- Sophisticated filtering

RNAseq workflow

Handling of splice junctions

- Mapping with STAR
- "Split N Trim"
- Specific filtering

Latest scope creep: Somatic Variant Discovery (via CGA)



Caveat Emptor

GATK Best Practices

= generic workflow recommendations

≠ concrete pipeline implementations

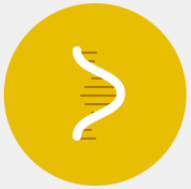
BROAD PIPELINES

Broad Genomic Services

Whole Genome Sequencing

Whole Exome Sequencing

Whole Transcriptome Sequencing



Customized Genomic Solutions

Clinical Research Sequencing

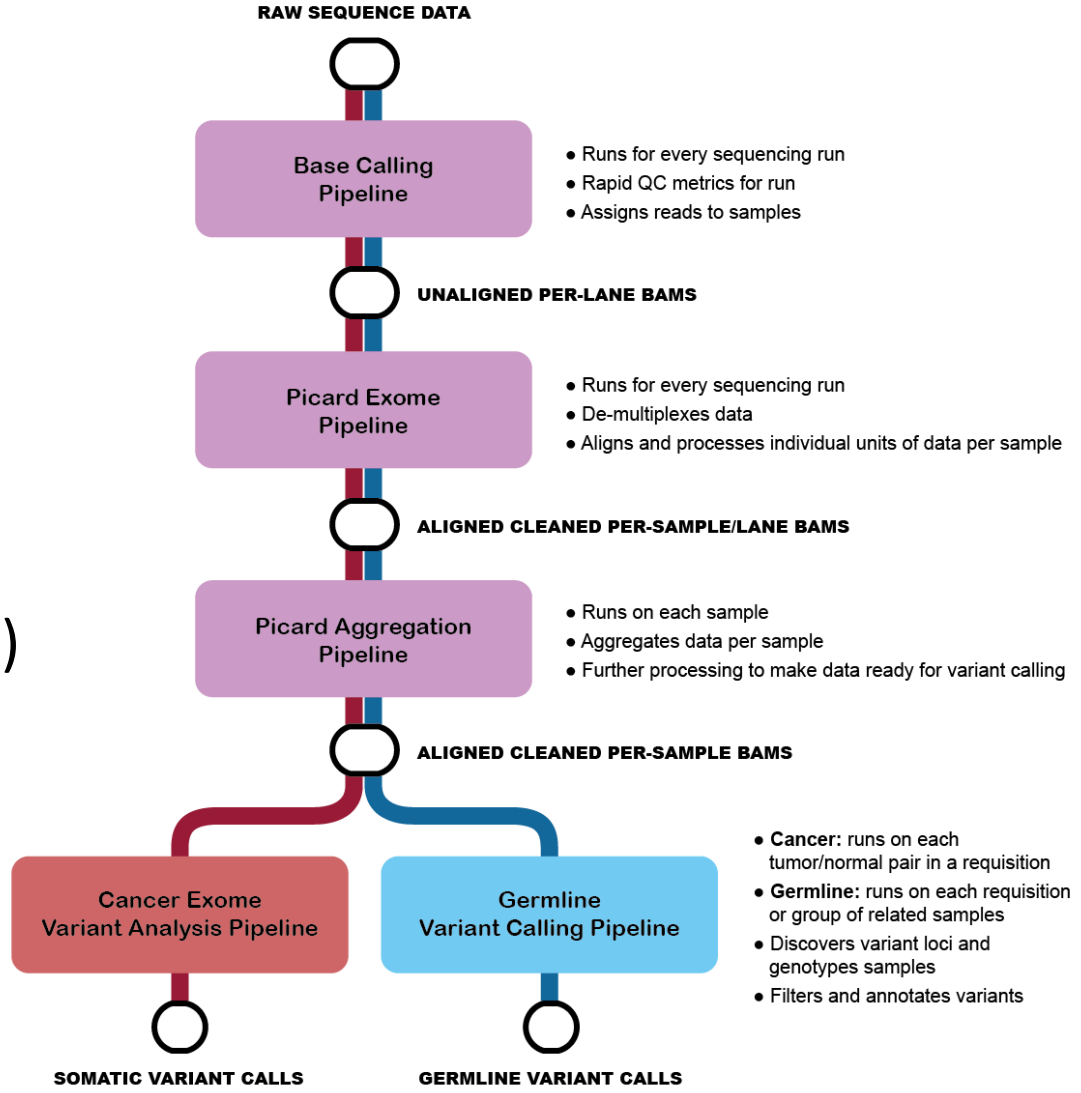
Data Analysis

http://genomics.broadinstitute.org/

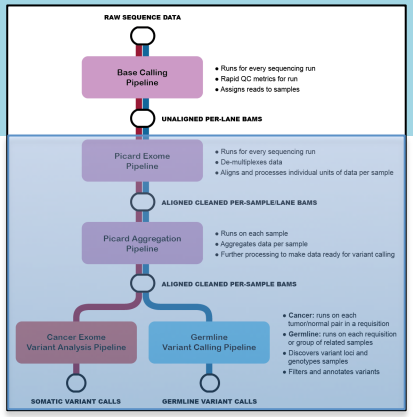
Research Exome Analysis Pipelines

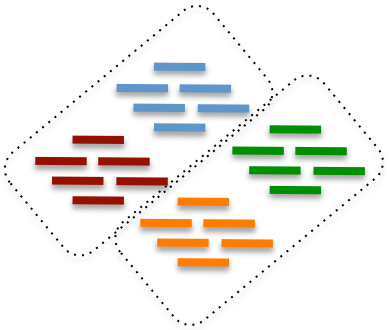
- Complete reads-to-variants pipeline implementations

- Two main versions:
  - Germline
  - Cancer (somatic)



Base Calling Pipeline

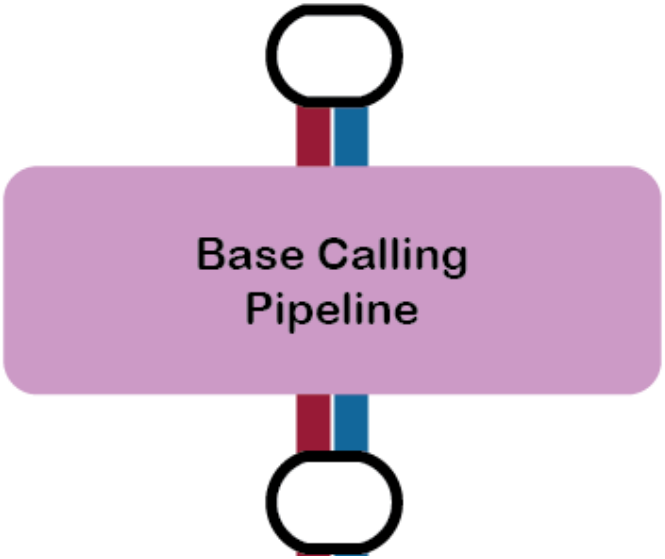


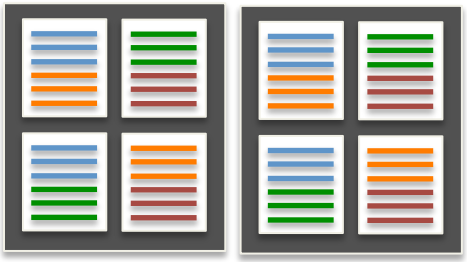


Prepared libraries (multiple per sample)

multiplex

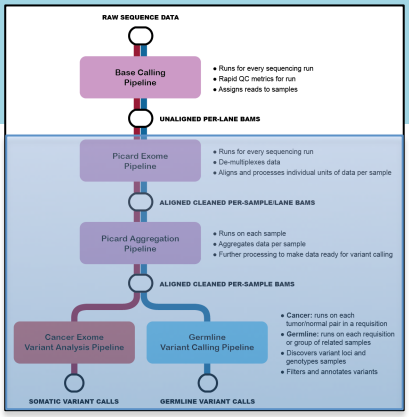
RAW SEQUENCE DATA **(BCL)**





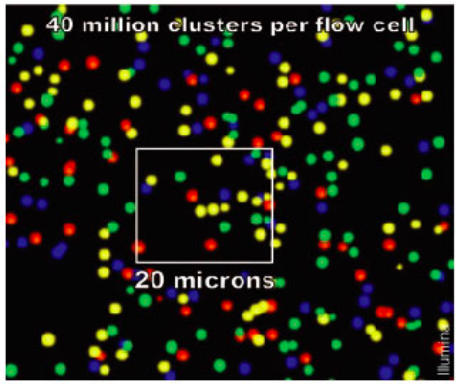
UNALIGNED PER-LANE BAMS (uBAMs)

Base Calling Pipeline

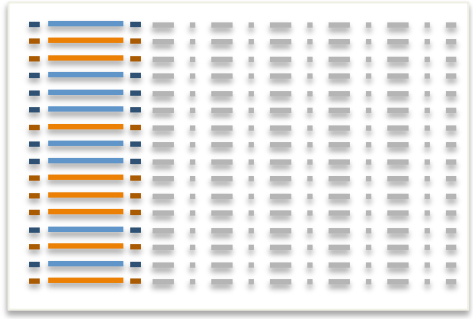


- Validate basecalls with **CheckIlluminaDirectory**
  - Sequencing run creates a directory with BCL data
  - Organized by lane, cycle, etc.
- Create **unmapped SAM file** with **IlluminaBasecallsToSam**
  - Unlike FASTQ, SAM can store file-level metadata (RG, LB, etc.)
- Low-level QC metrics to assess quality of run

Sequencer



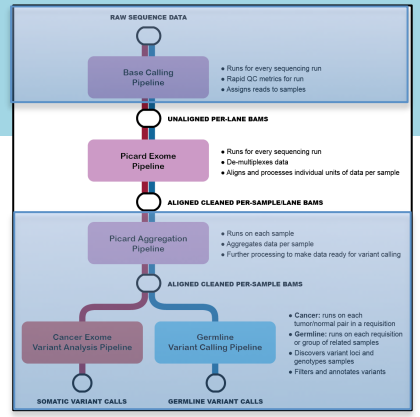
Basecalling (BCL) files

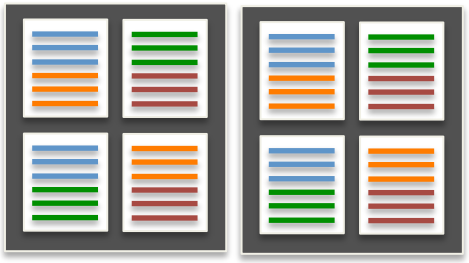


Unmapped SAM files

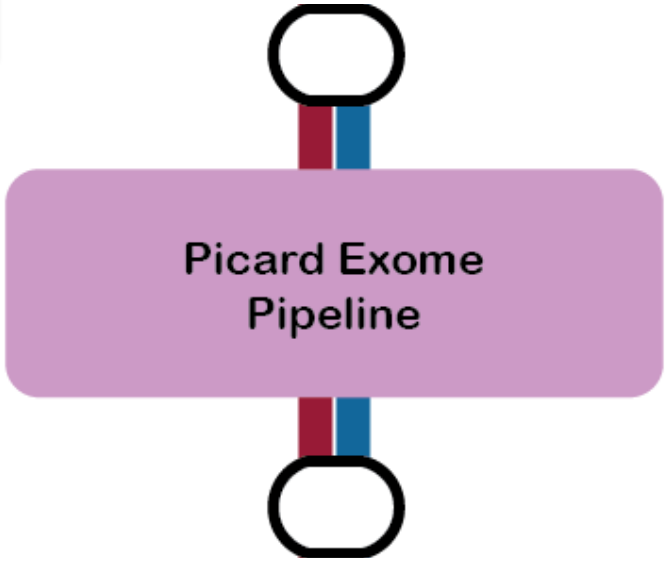
Note that this produces a SAM file rather than FASTQ as was presented in the intro to high-throughput sequencing. Here, we are using an unmapped SAM as a substitute for FASTQ because it can store metadata, which we want to attach to the data as soon as possible.

Picard Exome Pipeline



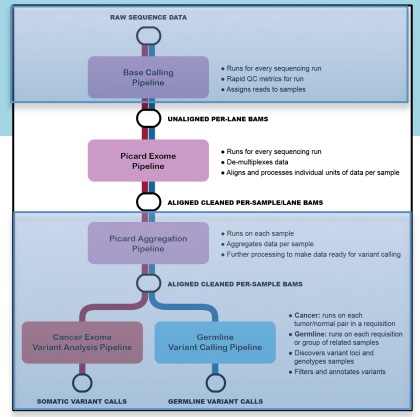


UNALIGNED PER-LANE BAMS (uBAMs)

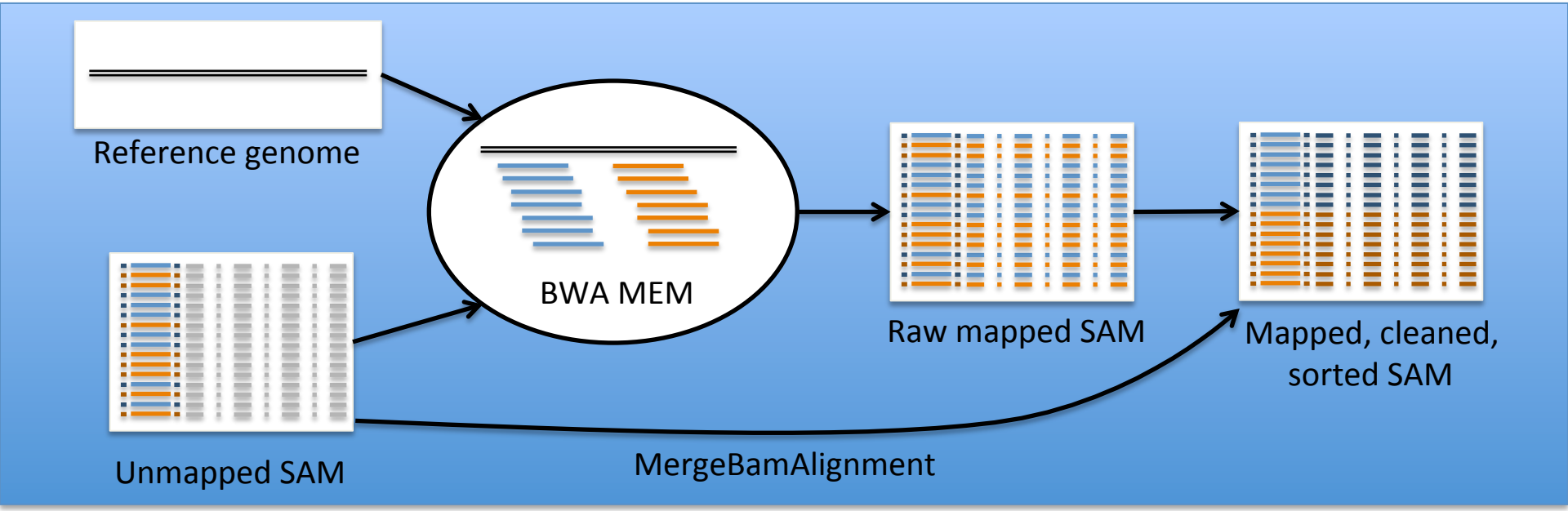


ALIGNED, CLEANED, DEDUPPED BAMS PER-SAMPLE PER-LANE

Picard Exome Pipeline

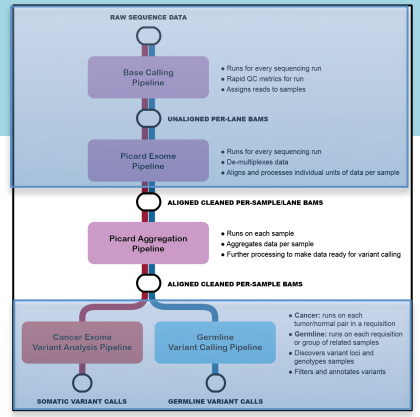


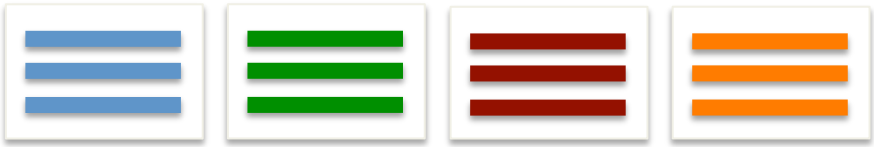
- **Demultiplex** data into per-sample per-lane files
- Map reads to reference using **BWA MEM**
- Combine output with original unmapped SAM using **MergeBamAlignment**
  - Cleans up mapping issues, sorts reads, adds read group information



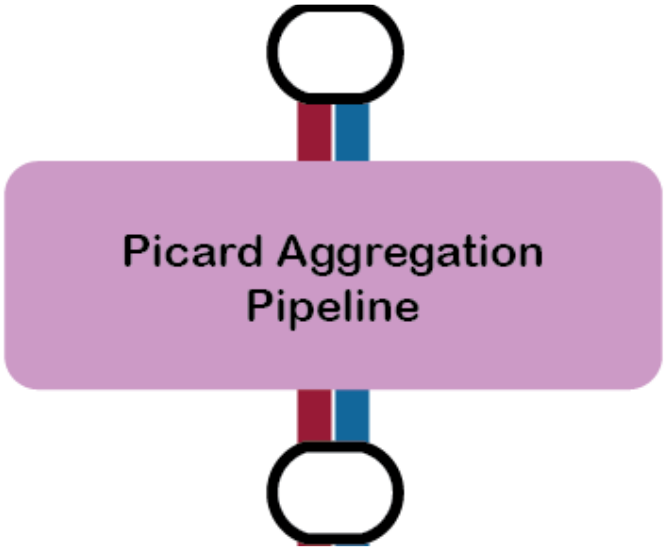
- Mark duplicates with **MarkDuplicates**

Picard Aggregation Pipeline



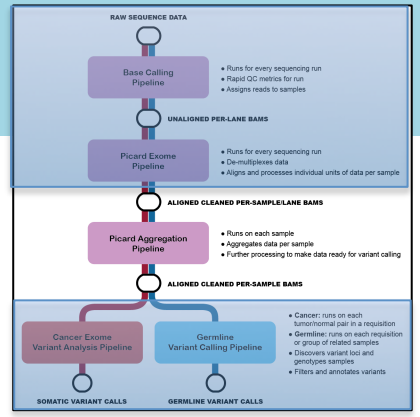


ALIGNED, CLEANED, DEDUPPED BAMS PER-SAMPLE PER-LANE



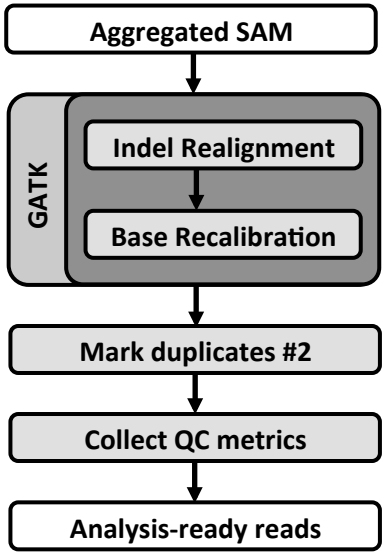
ALIGNED, CLEANED, DEDUPPED BAMS PER-SAMPLE

Picard Aggregation Pipeline



- Aggregate data per sample
- Perform GATK processing steps
  - Indel Realignment
  - Base Recalibration
- Mark duplicates with **MarkDuplicates** (2nd pass)
- Collect QC metrics

ALIGNED, CLEANED, DEDUPPED BAMS PER-SAMPLE



Summary of the pre-processing

LAB WORK

Library Prep

- Sample may have multiple libraries
- Libraries might be split across multiple flowcells

BASECALLING + PICARD EXOME

Flowcell-level processing

- Convert basecalls to SAM
- *Map with BWA*
- Sorting + cleaning
- Mark duplicates
- Collect QC metrics

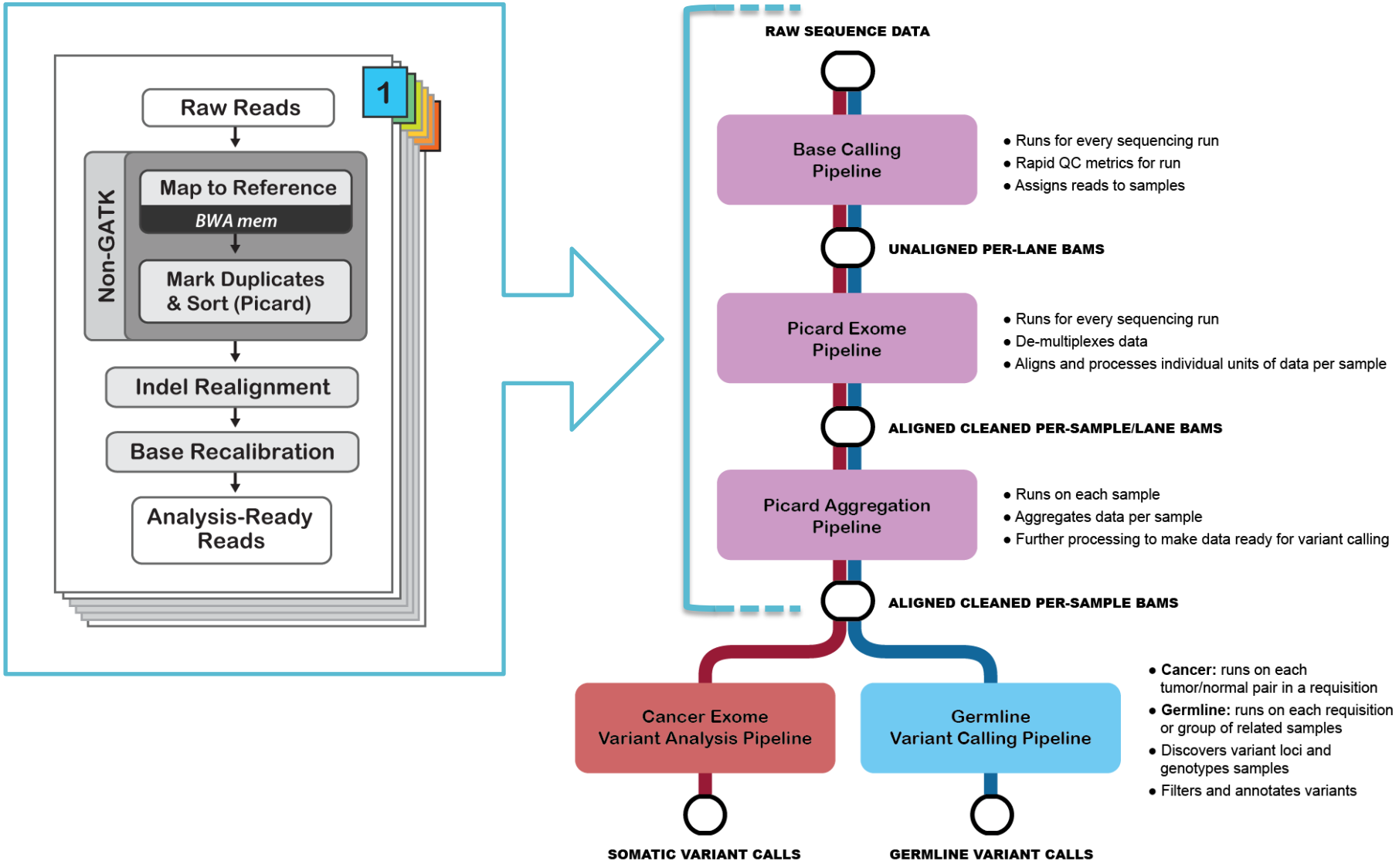
PICARD AGGREGATION

Aggregation-level processing

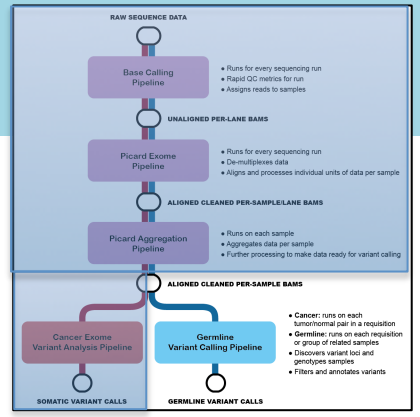
- Merge sample across flowcells
- Mark duplicates (again)
- *Indel realignment*
- *BQSR*
- Collect QC metrics

Prepared libraries | **Flowcells / lanes** | **Aggregated samples**

Generic Best Practices vs. production implementation

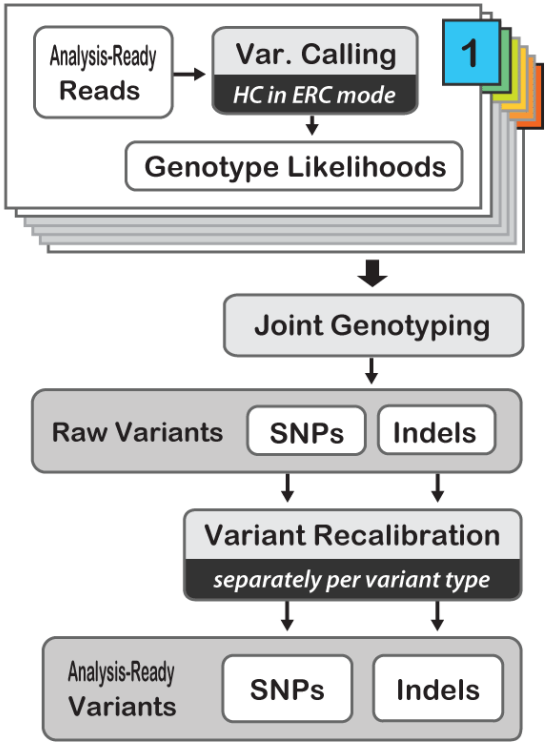


Germline Variant Discovery

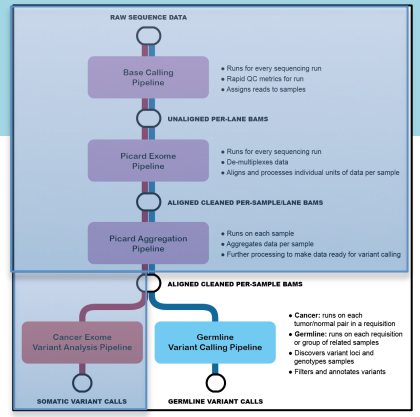


- Call variants per-sample
- Joint genotyping per cohort
- Variant recalibration

ALIGNED, CLEANED, DEDUPPED BAMS PER-SAMPLE

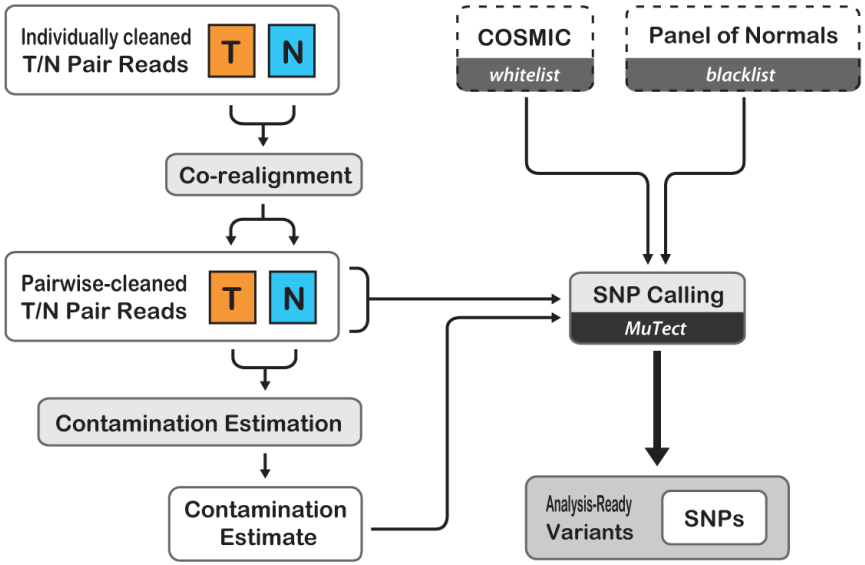


Somatic Variant Discovery



- Co-cleaning of Tumor/Normal pair
- Call variants per T/N pair
- Variant filtering and processing

ALIGNED, CLEANED, DEDUPPED BAMS PER T/N PAIR



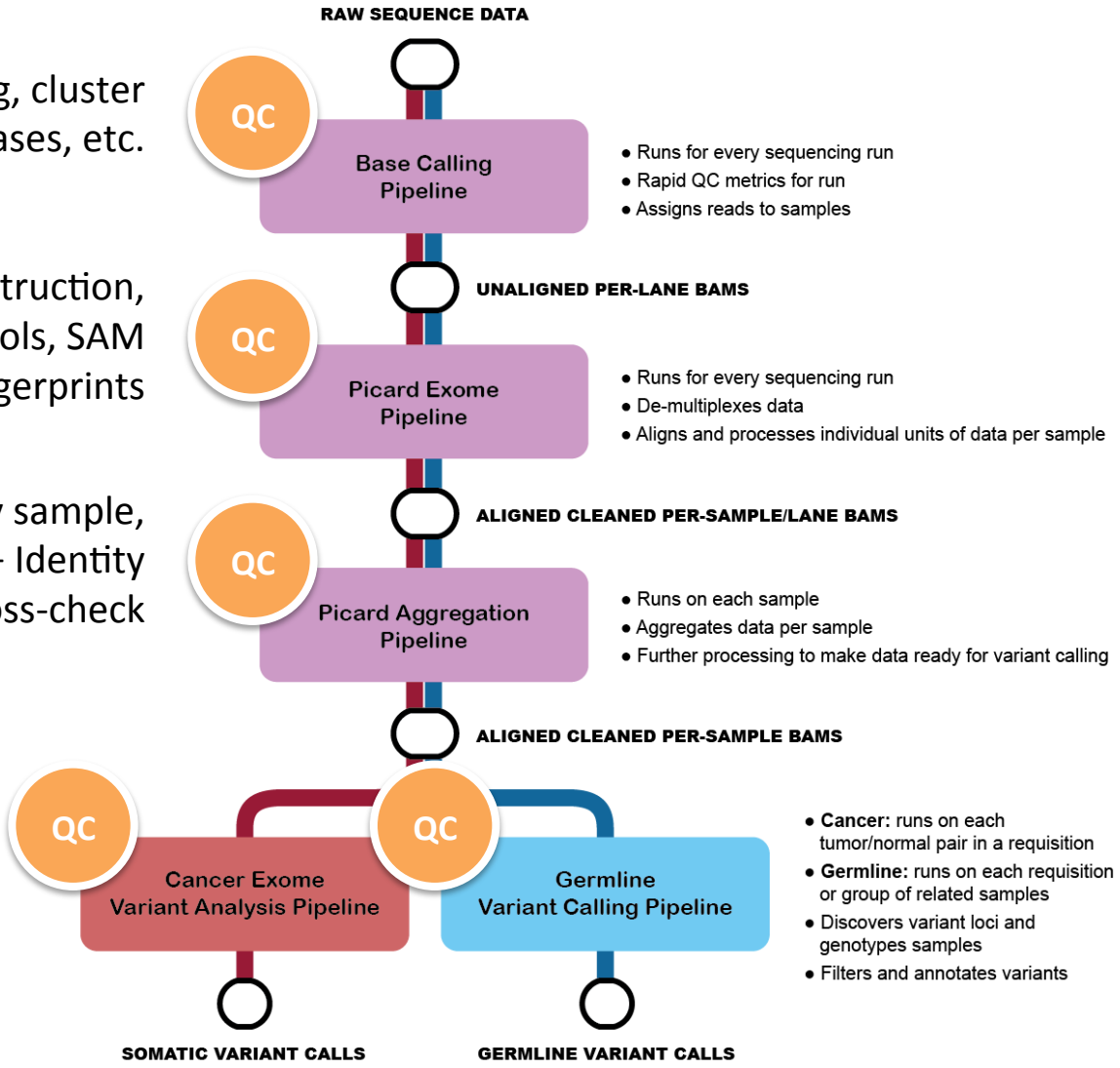
Up next: Quality Control (QC)

(1) Quality of barcode matching, cluster density, number of reads, bases, etc.

(2) Quality of alignment, library construction, coverage, base quality, internal controls, SAM format validation + Identity through fingerprints

(3) Cumulative quality from (2) by sample, cross-sample contamination + Identity fingerprints, read groups cross-check

(4) VCF format validation, genotype concordance on control samples, variant calling quality metrics



Further reading

http://www.broadinstitute.org/gatk/guide/best-practices