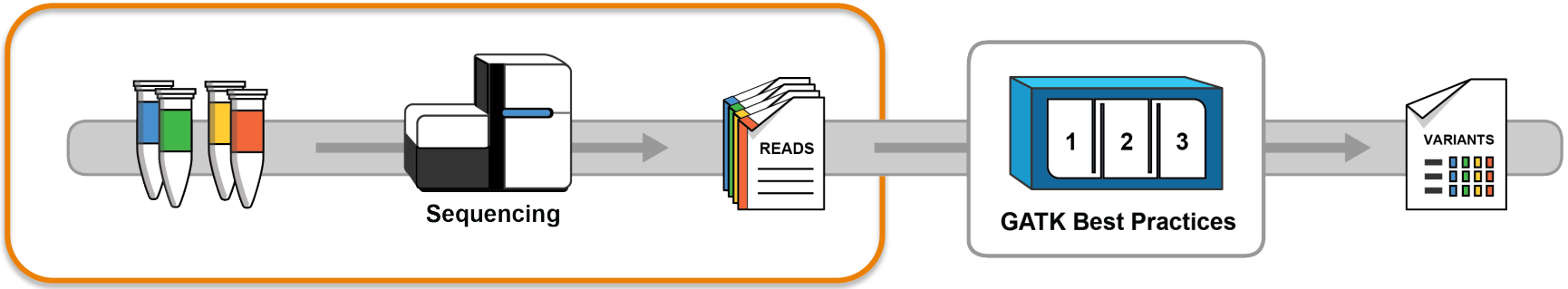
Introduction to High-Throughput Sequence Data

How it is generated and how we process it

BROAD INSTITUTE

http://software.broadinstitute.org/gatk/

gatk

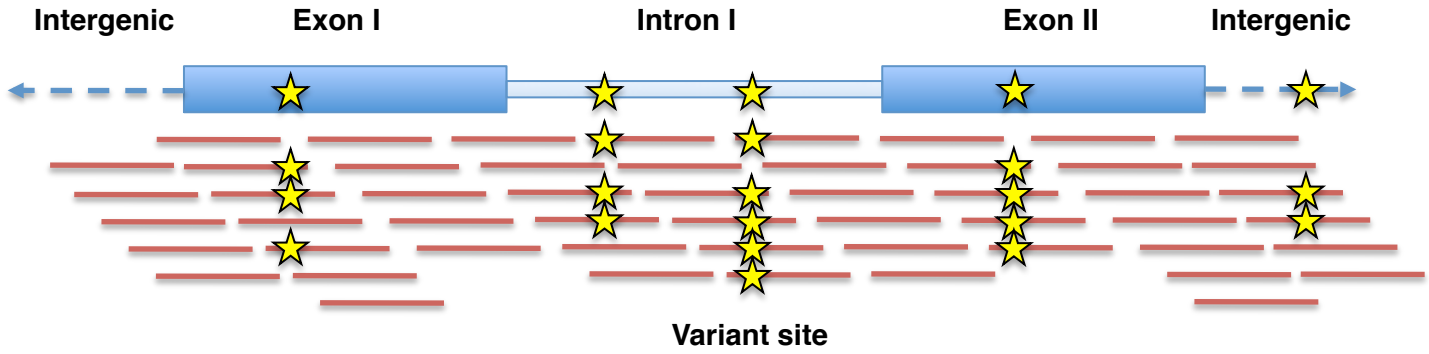


Part 1

DATA GENERATION

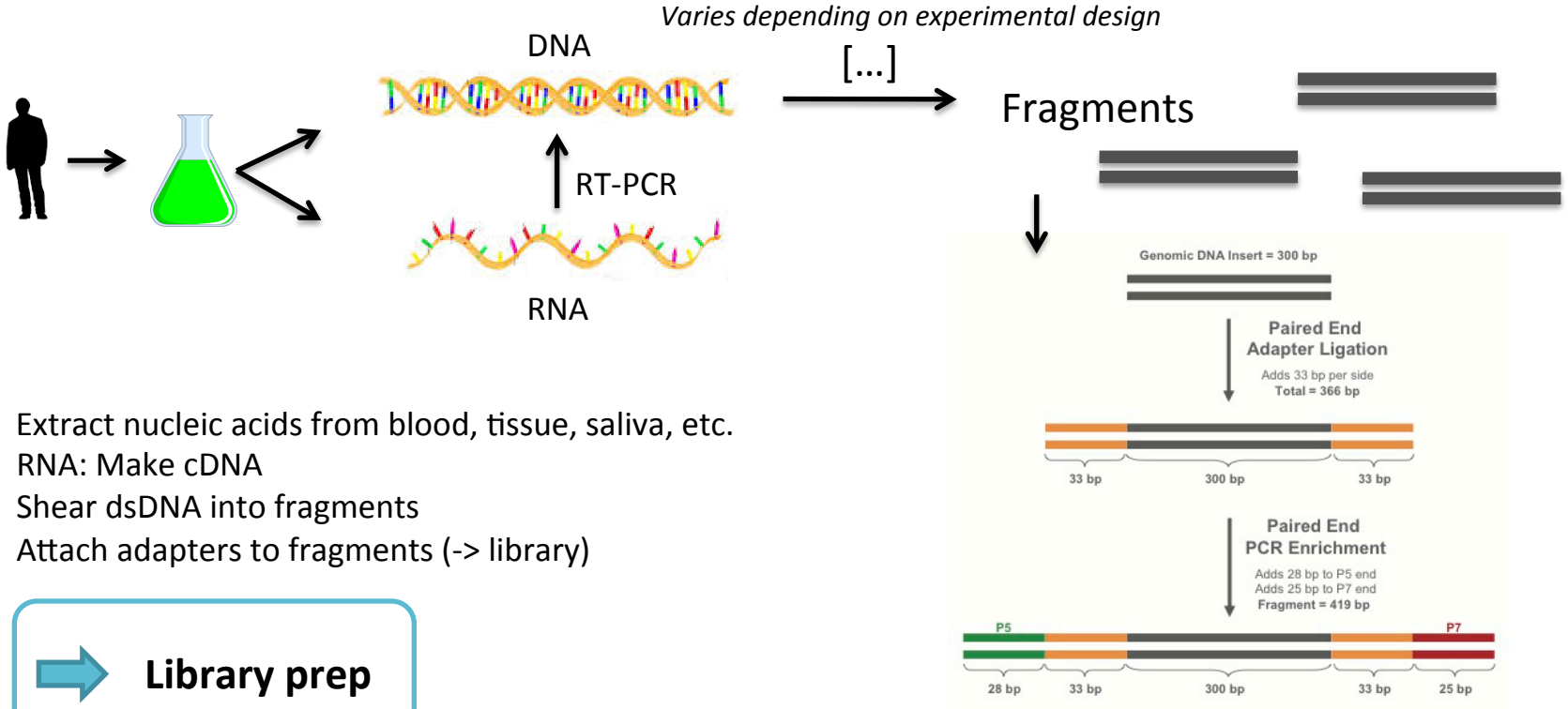
Different types of experimental design

Whole Genome



Exome (+ gene panels)

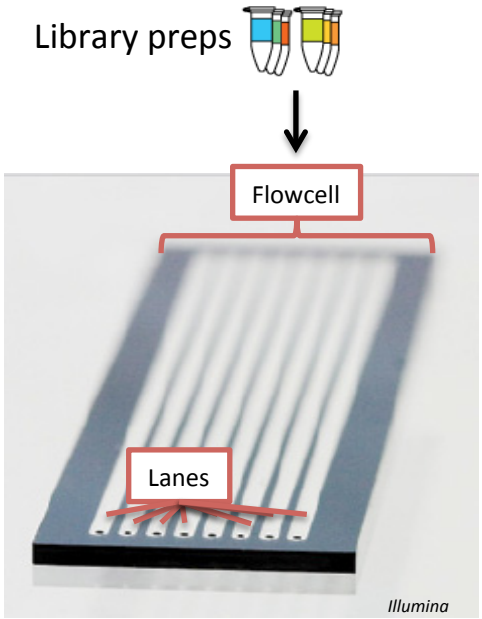
Library preparation



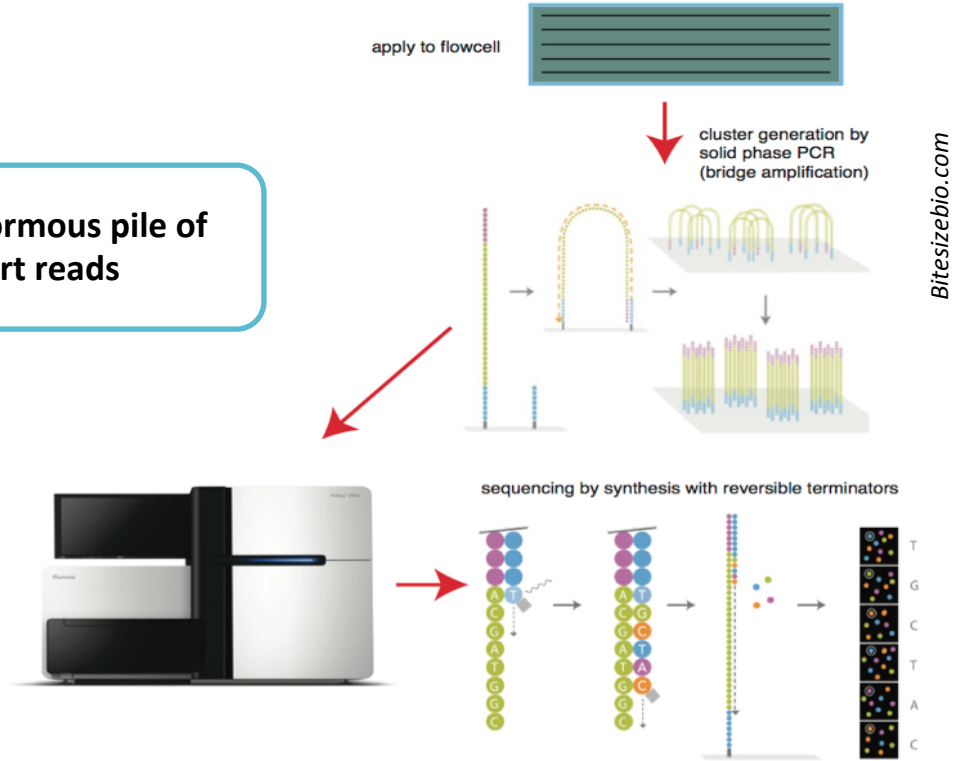
1) Extract nucleic acids from blood, tissue, saliva, etc.
2) RNA: Make cDNA
3) Shear dsDNA into fragments
4) Attach adapters to fragments (-> library)

Library prep

Sequence the library



Enormous pile of short reads



HTS machine processes a **flowcell** containing **lanes**; each lane may contain multiple sample (indexed With a DNA barcode), each of there is a **read group (RG)**

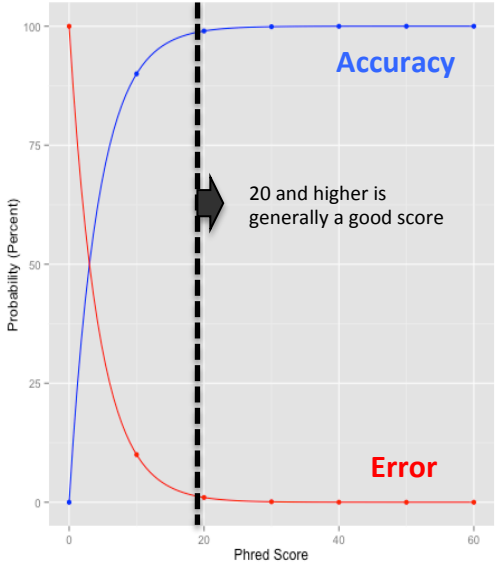
Raw sequence: typically in FASTQ format

- Sequence Name (read name, group, etc.)
- Sequence
- + (optional: Sequence name again)
- Associated quality score

Example record

- `@EAS54_6_R1_2_1_413_324`
- `CCCTTCTTGTCTTCAGCGTTTCTCC`
- `+`
- `;;3;;;;;;;;;;7;;;;;;;88`

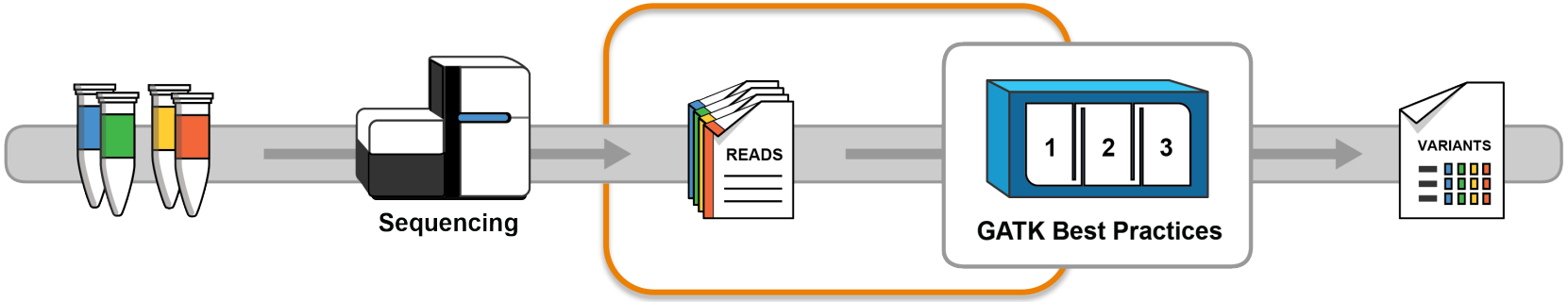
-> ASCII code translates to Phred-scale Q scores



Phred value = $-10 * log_{10}(\varepsilon)$

90% confidence (10% error rate) = Q10
99% confidence (1% error rate) = Q20
99.9% confidence (0.1% error rate) = Q30

Format specification http://maq.sourceforge.net/fastq.shtml



Part 2

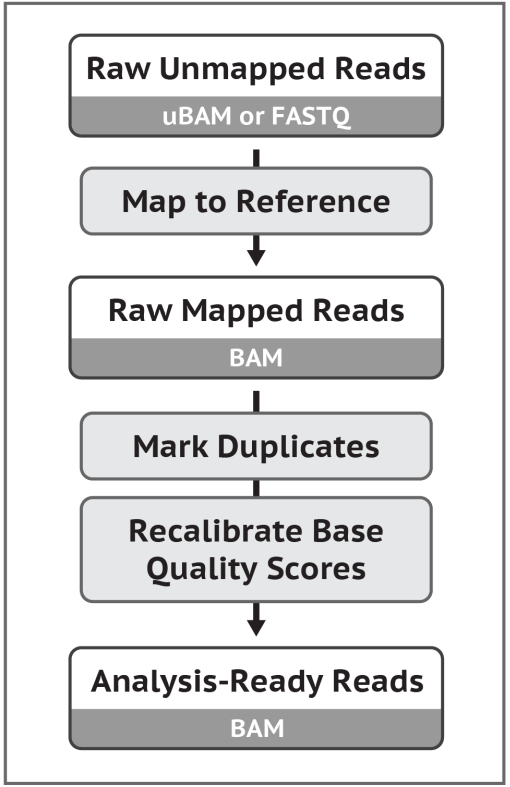
DATA PRE-PROCESSING

What is data pre-processing?

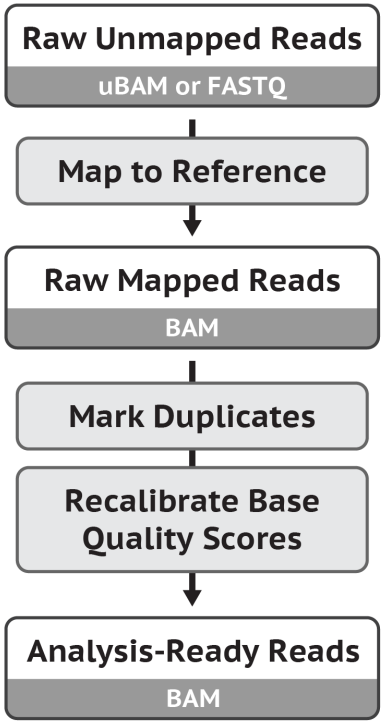
Data produced by the sequencers:

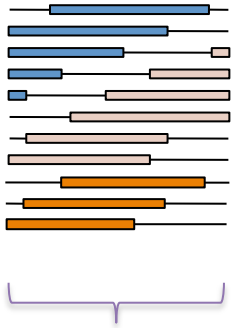
- A huge pile of *paired* reads without mapping information
- Afflicted by various technical biases and artifacts
- May include artificial duplicates reads from the same original molecule

➢ **Need to map, sort and clean up**



Step 1: Map the reads produced by the sequencer to the reference

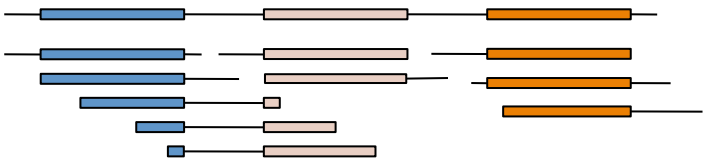




Enormous pile of short reads from HTS

Mapping and alignment algorithms

- **BWA for DNA**
- **STAR for RNAseq**

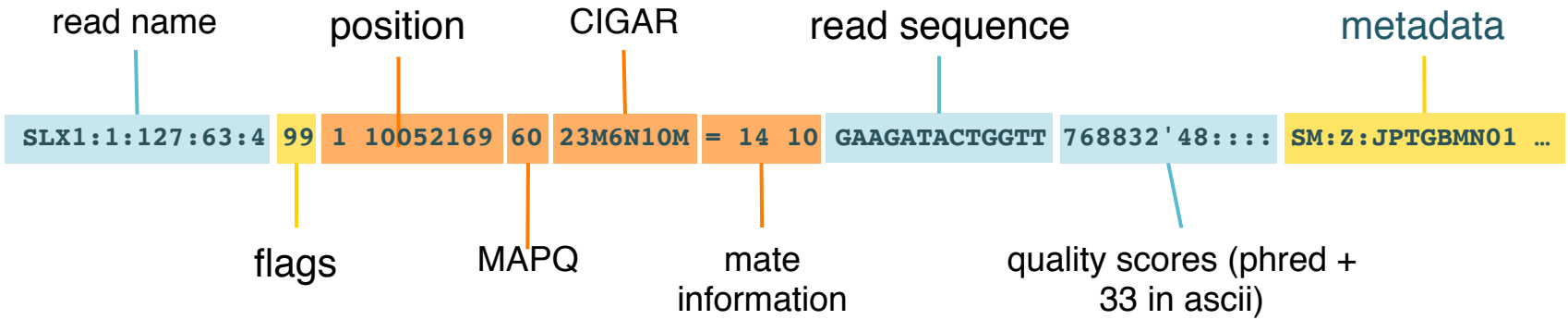


Reference genome

Reads mapped to reference

Output format: Sequence/Binary Alignment Map (SAM/BAM)

HEADER containing metadata (sequence dictionary, read group definitions, etc.)
RECORDS containing structured read information (1 line per read record)



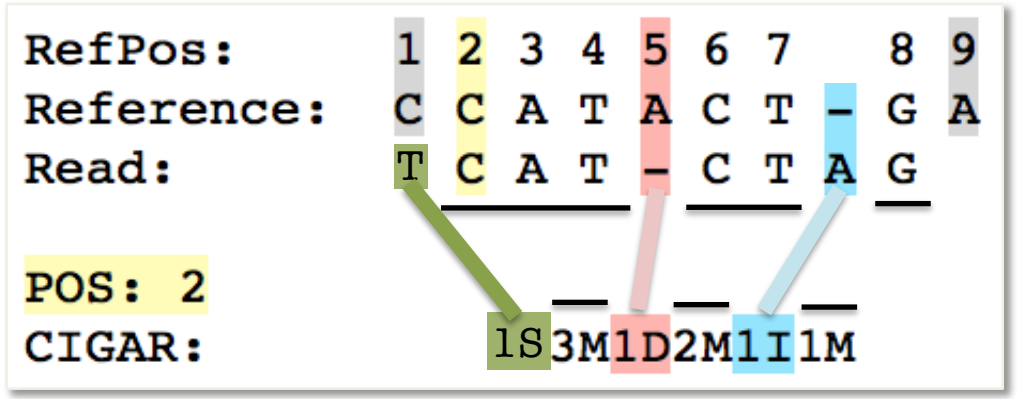
- Added mapping info summarizes **position**, **quality**, and **structure** for each **read**
- Mate information points to the read from the other end of the molecule

http://samtools.github.io/hts-specs/SAMv1.pdf

CIGAR summarizes **alignment structure**

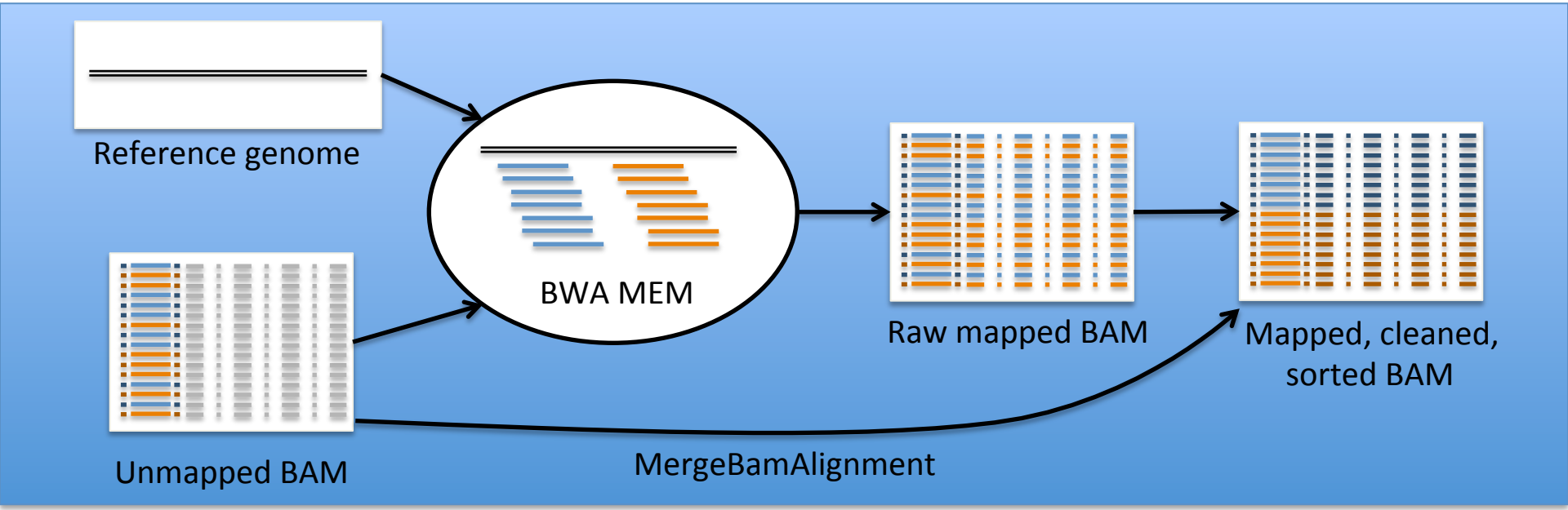
CIGAR = Concise Idiosyncratic Gapped Alignment Report

```
read1   99   ref   2   30   1S3M1D2M1I1M   =   14   20   CATCTAG   …
```

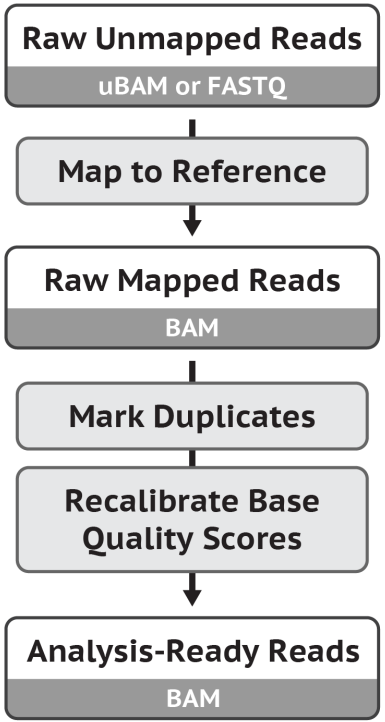


At Broad: Unmapped BAM instead of FASTQ

Special workflow using Picard tools for improved data management

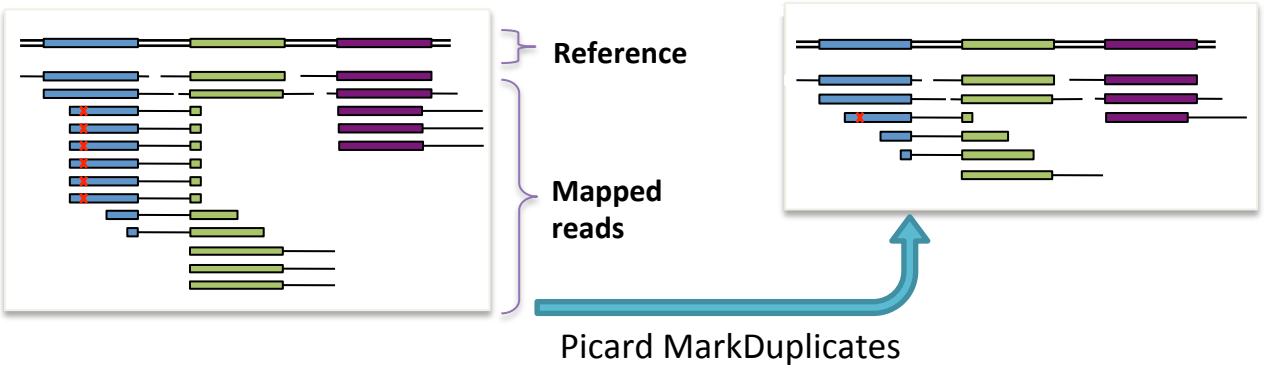


Step 2: Mark duplicates to mitigate duplication artifacts



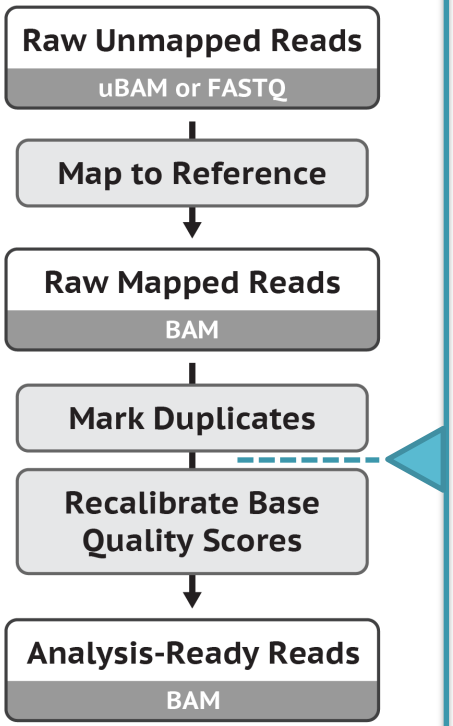
Duplicates = **non-independent measurements** of a sequence fragment

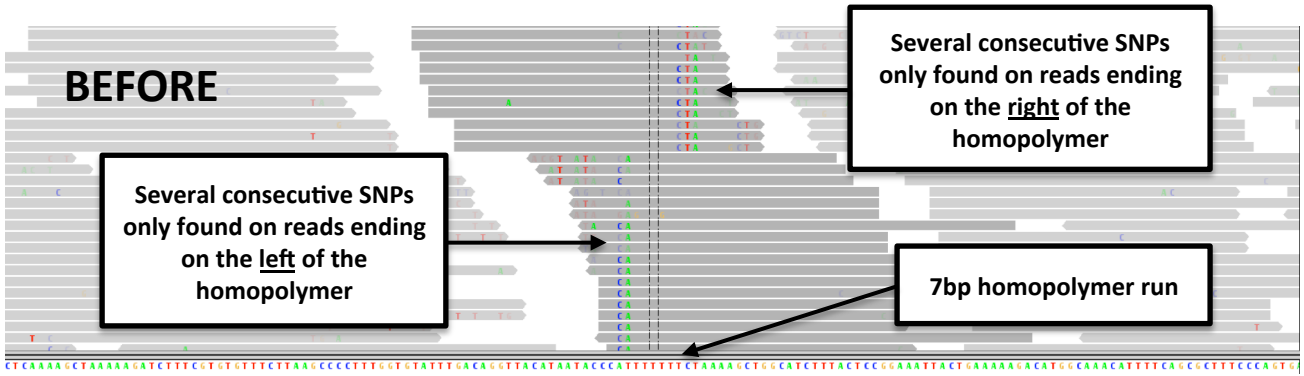
-> Must be removed to assess support for alleles correctly

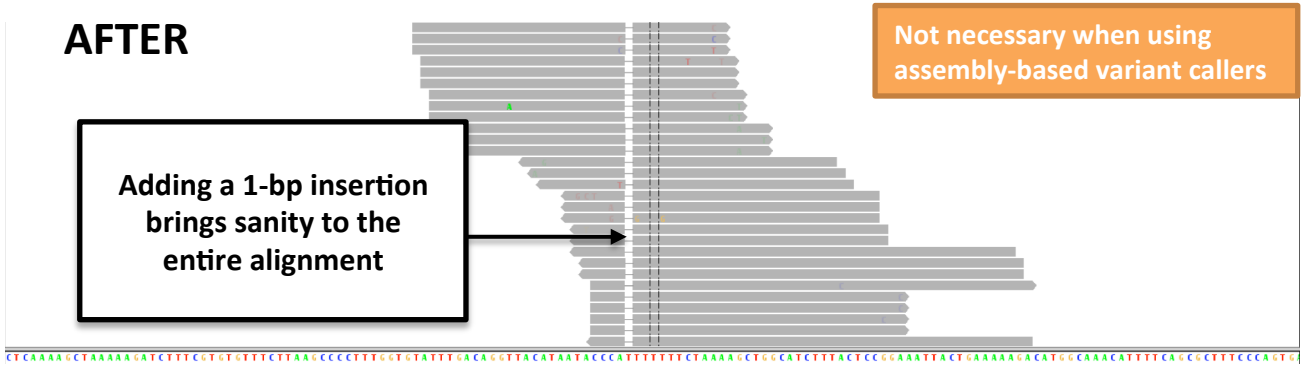


✖ = sequencing error propagated in duplicates

Step 3: DEPRECATED: Local realignment around indels

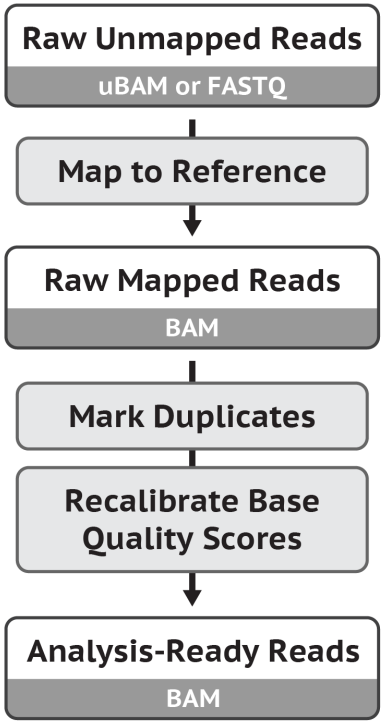






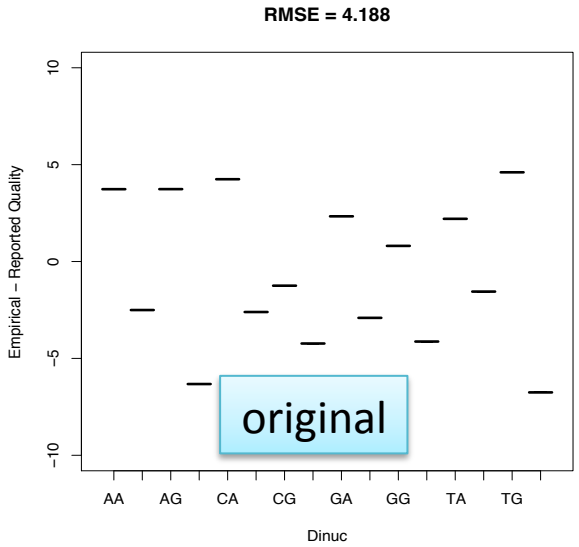
** For implications, see <https://gatkforums.broadinstitute.org/gatk/discussion/7847/changing-workflows-around-calling-snps-and-indels>*

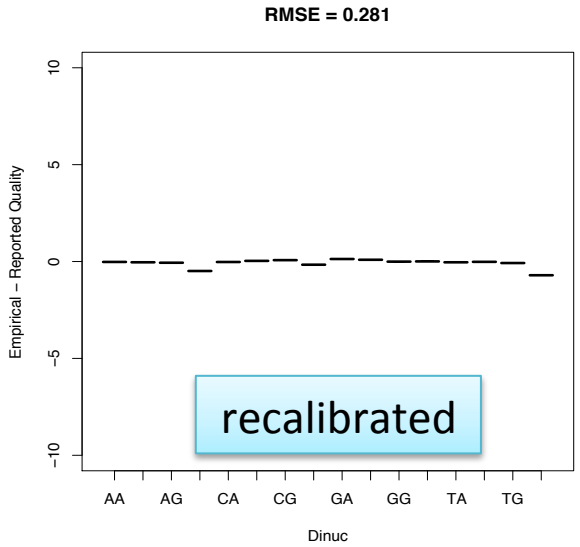
Step 4: Base Recalibration (BQSR) corrects for machine errors



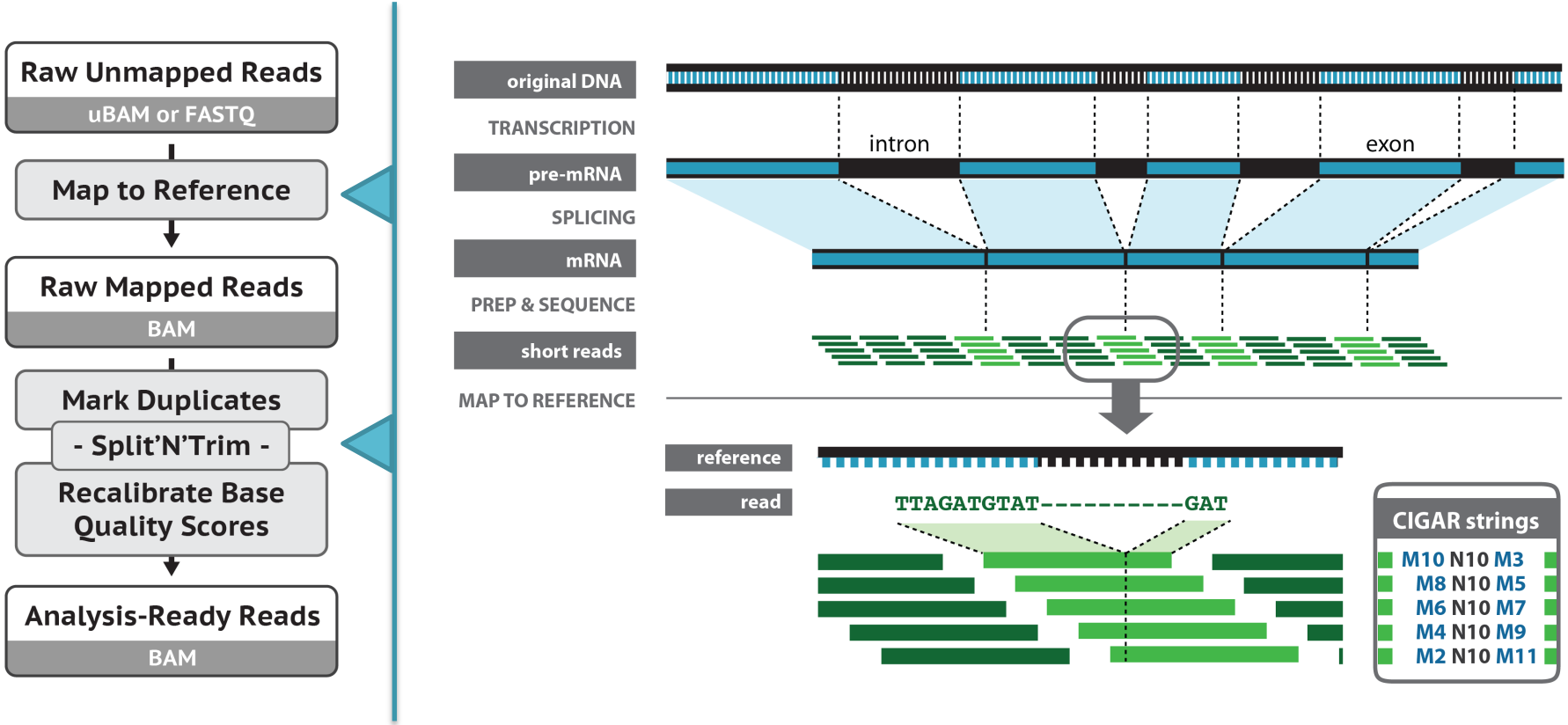
- Sequencers make systematic errors in base quality scores
- Sequencer quality cannot include PCR-based errors
- BQSR corrects the quality scores (not the bases)

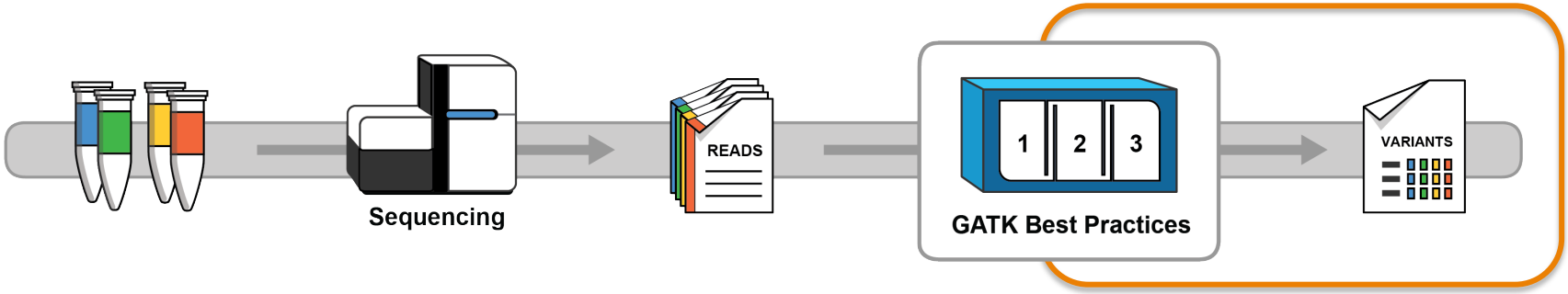
Example of bias: qualities reported depending on nucleotide context





Special handling for RNAseq splice junctions





Next step:

VARIANT DISCOVERY