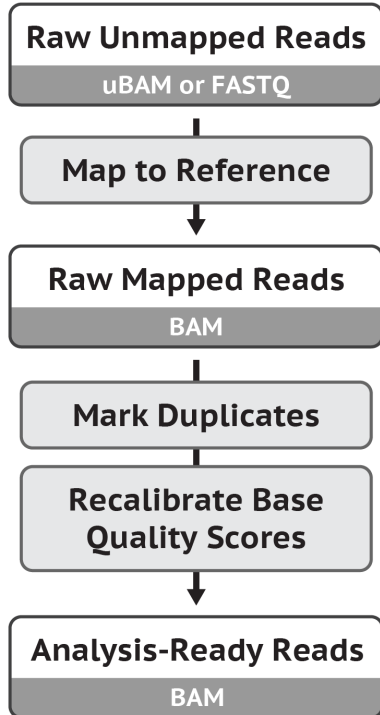




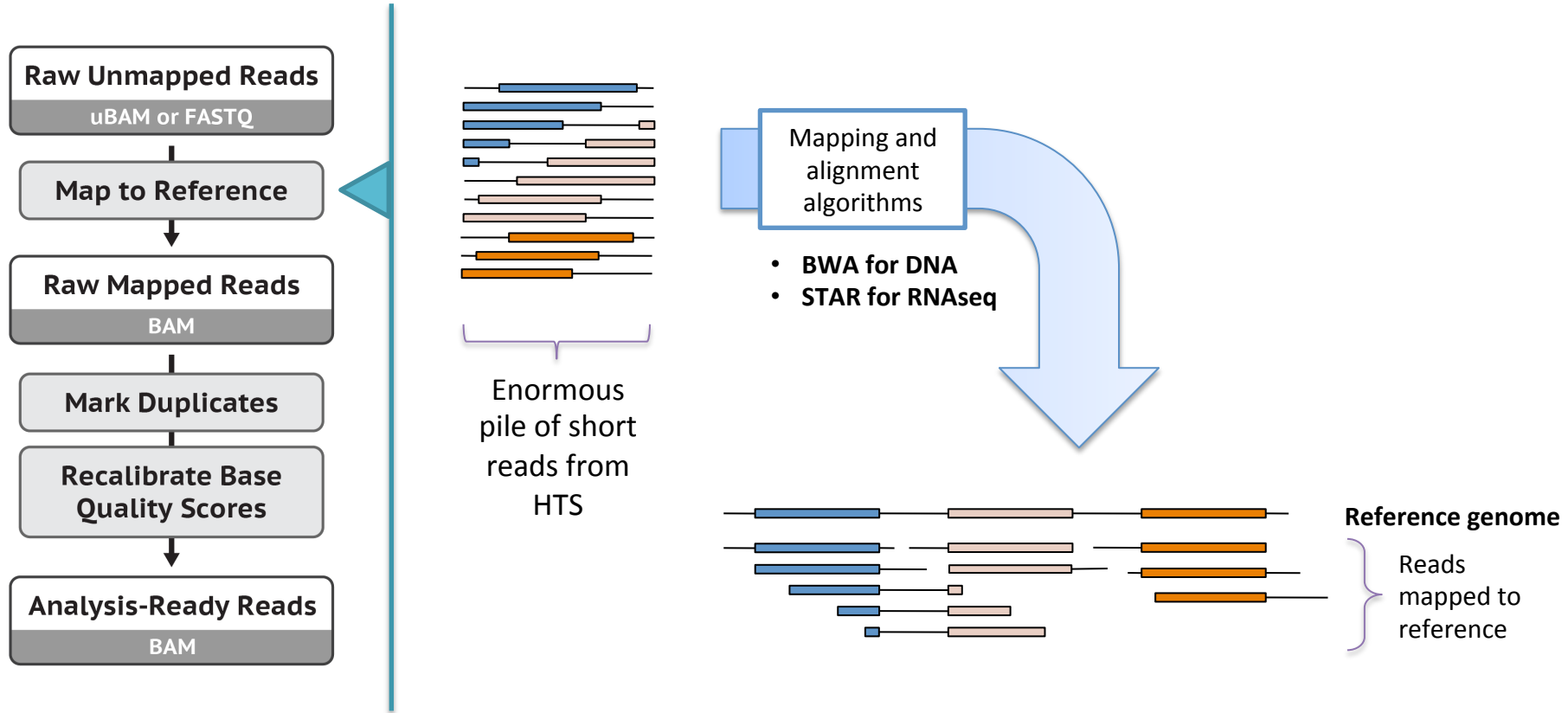
Mapping

Finding where reads belong in the genome

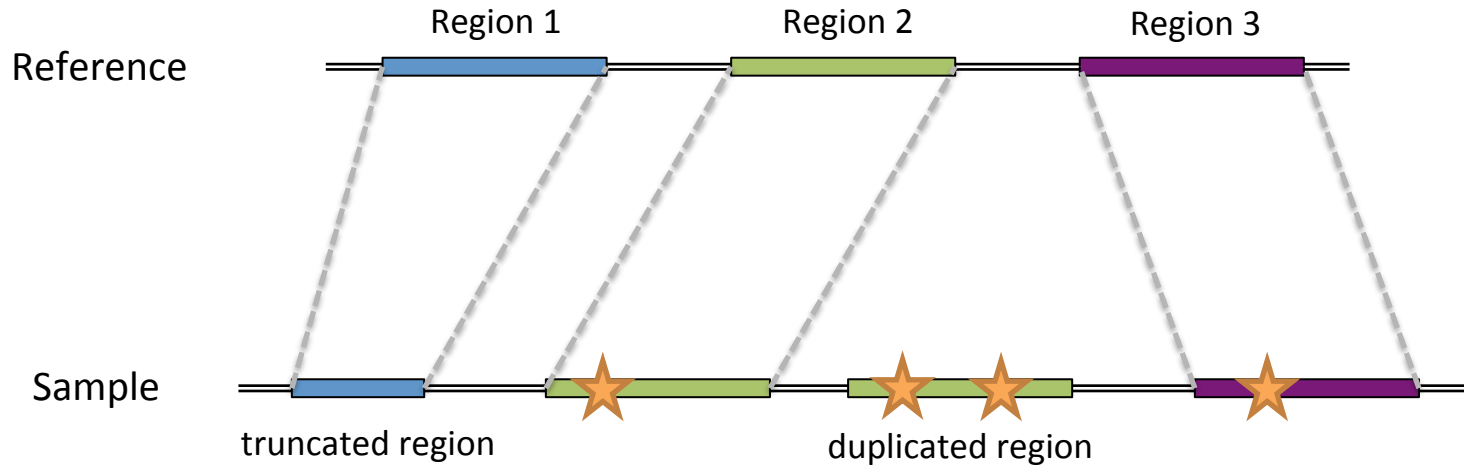
Data Pre-processing for Variant Discovery



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



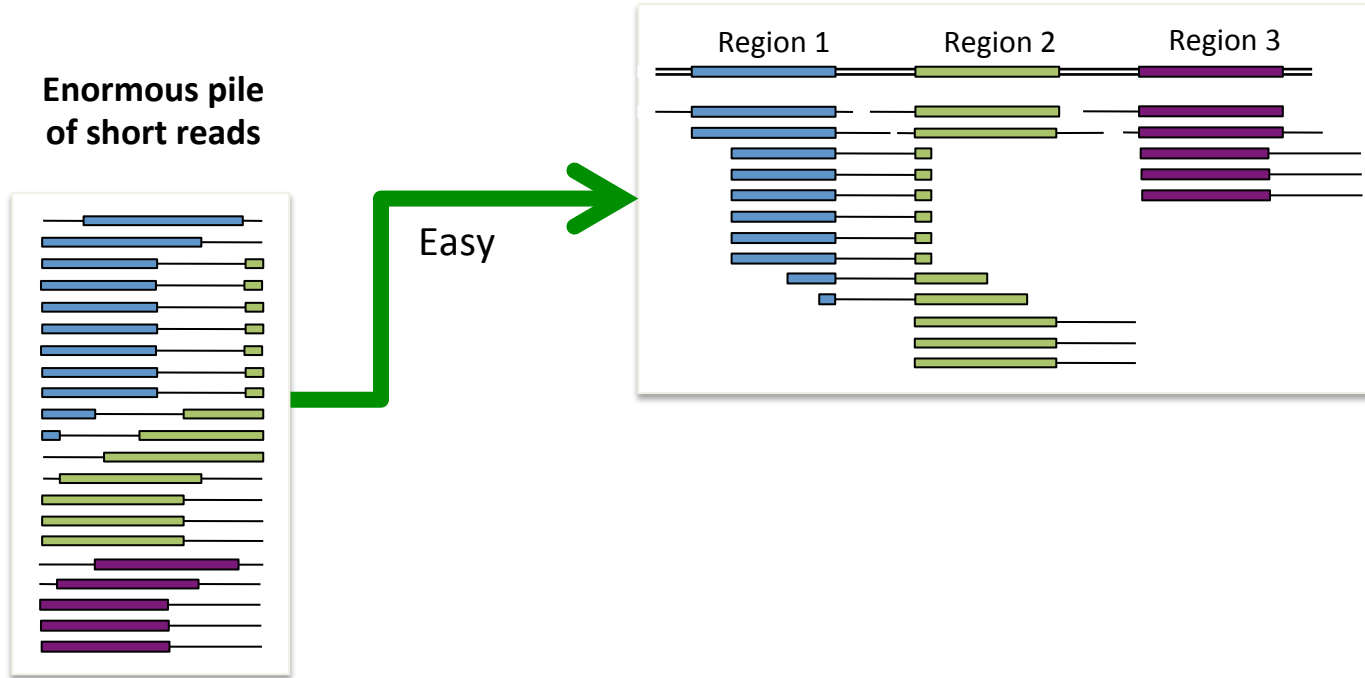
Goal: align the sample genome to the reference genome



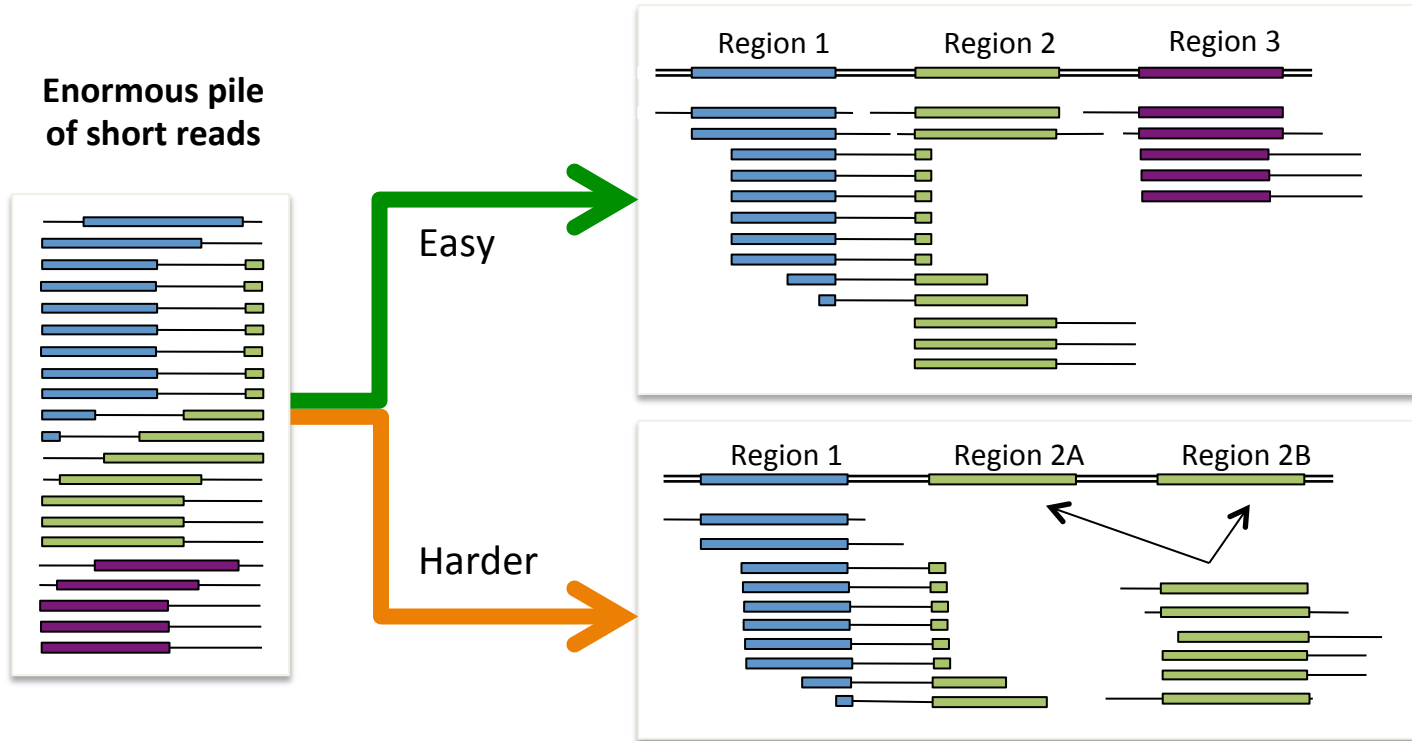
★ = local variant (SNP/indel)

...But we don't have the whole sample in one piece.

So we have to map each little bit one by one

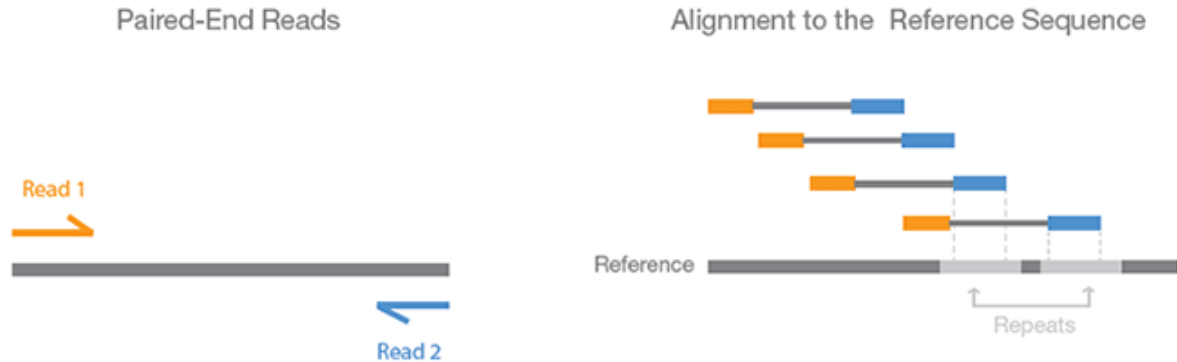


Complication: mismatches, indels, duplicated regions...



Paired-end sequencing helps a lot

Figure 4. Paired-End Sequencing and Alignment

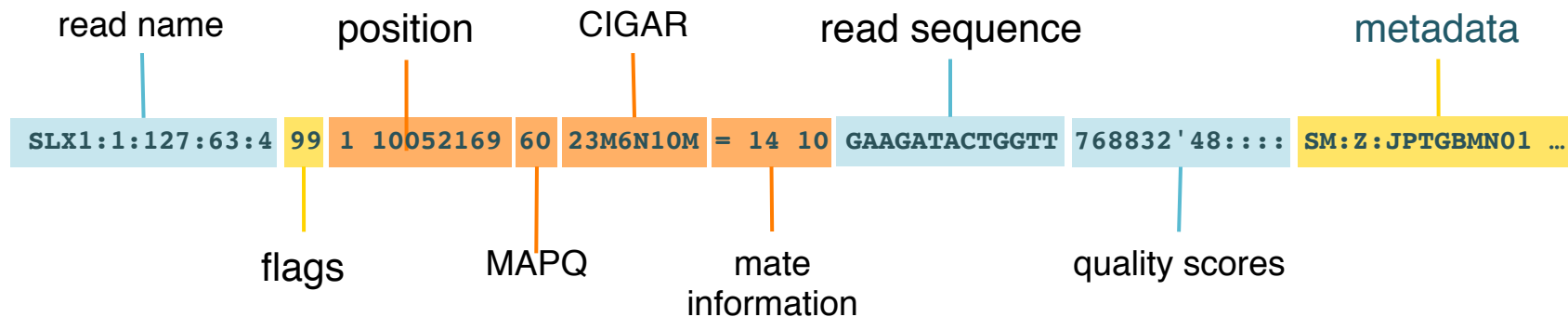


Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

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**

Special Note #1

ALT CONTIGS IN HG38



How BWA handles ALT contigs

Read: ATCAGCATC

```
ALT ctg 1:      TGAAA---CGAATGCAAATGGTCAATCAGCATCGAACTAGTCACAT
                ||||| (high div) ||||| (novel ins) |||||
Chromosome: GCGTACATGATACGAATCgGCATCATGGTC-----CTAGTCACATCGTAATC
                ||||| ||||| (novel ins) |||||
ALT ctg 2:      TGATACGAATCgcCATCATGGTCAATCgcCAgCGAACTAGTCACAT
```

4 potential hits: ATCAGCATC > ATCgGCATC > ATCgcCATC > ATCgcCAgC

2 hit groups: {ATCAGCATC, ATCgcCAgC} and {ATCgGCATC, ATCgcCATC}

Hits considered in mapQ: ATCAGCATC and ATCgGCATC (best from each group)

In the output SAM: ATCgGCATC as the primary SAM line with mapQ=0

ATCAGCATC as a supplementary line with mapQ>0

ATCgcCAgC as a supplementary line with mapQ>0

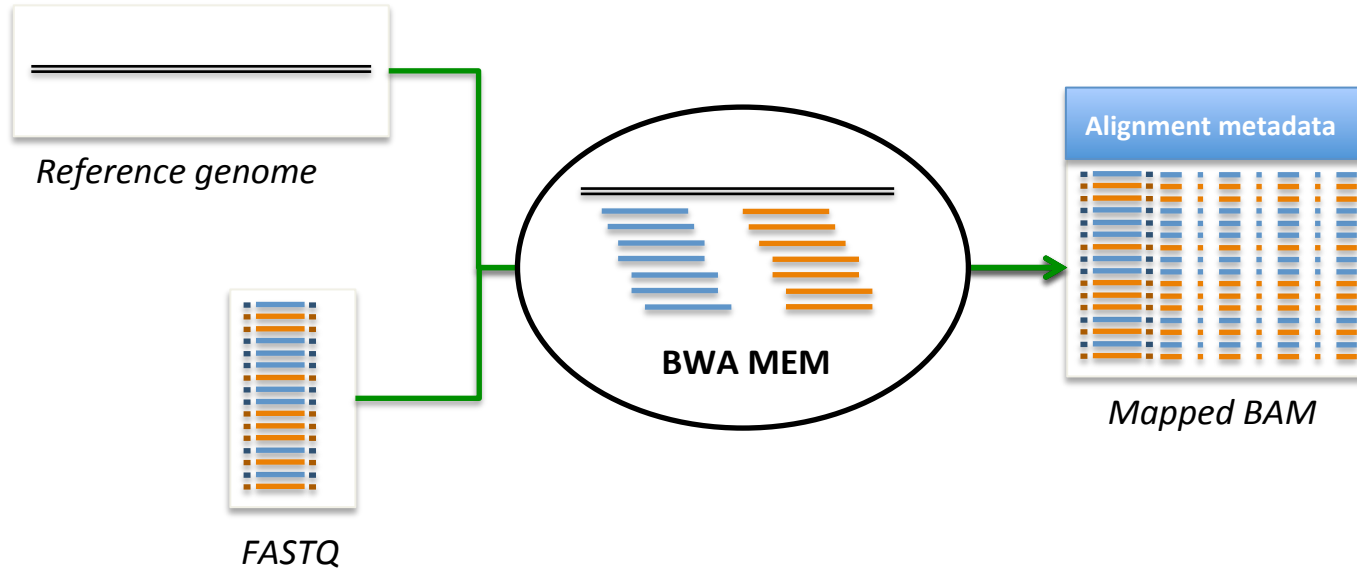
ATCgcCATC in an XA tag, not as a separate line

Special Note #2

THE UNMAPPED BAM WORKFLOW

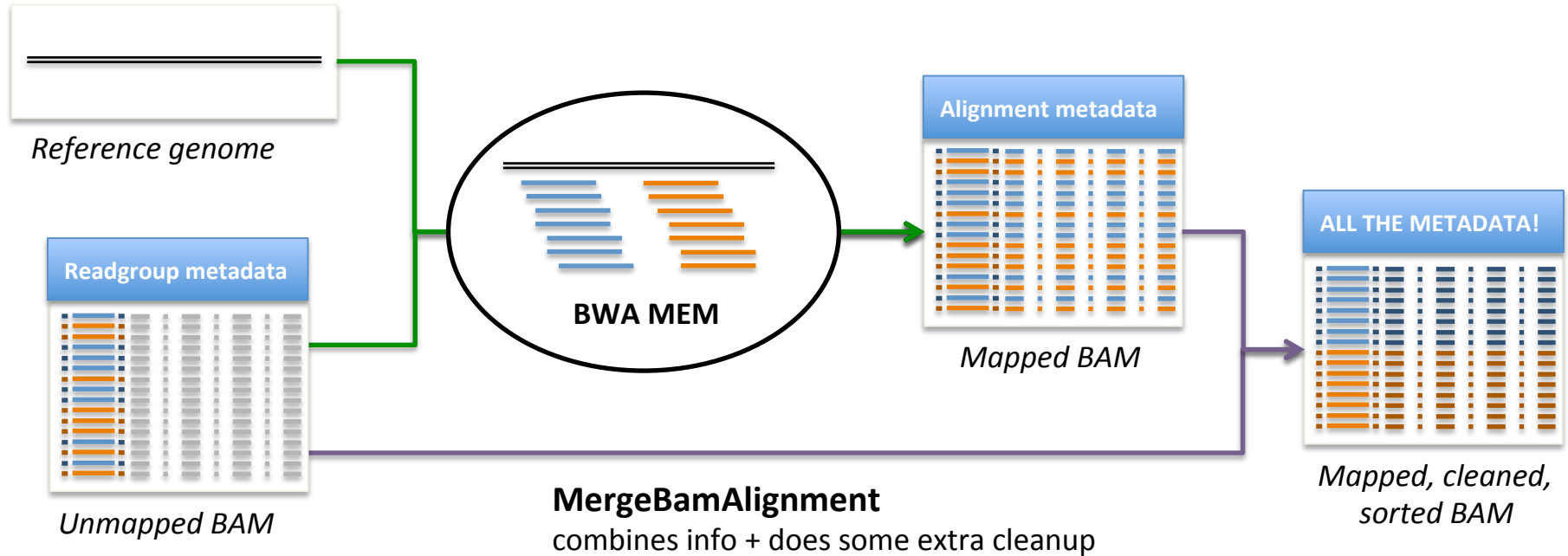


Regular FASTQ -> BAM workflow



! Adding Readgroup metadata requires additional step
or injection of metadata into BWA command

Unmapped BAM -> BAM workflow

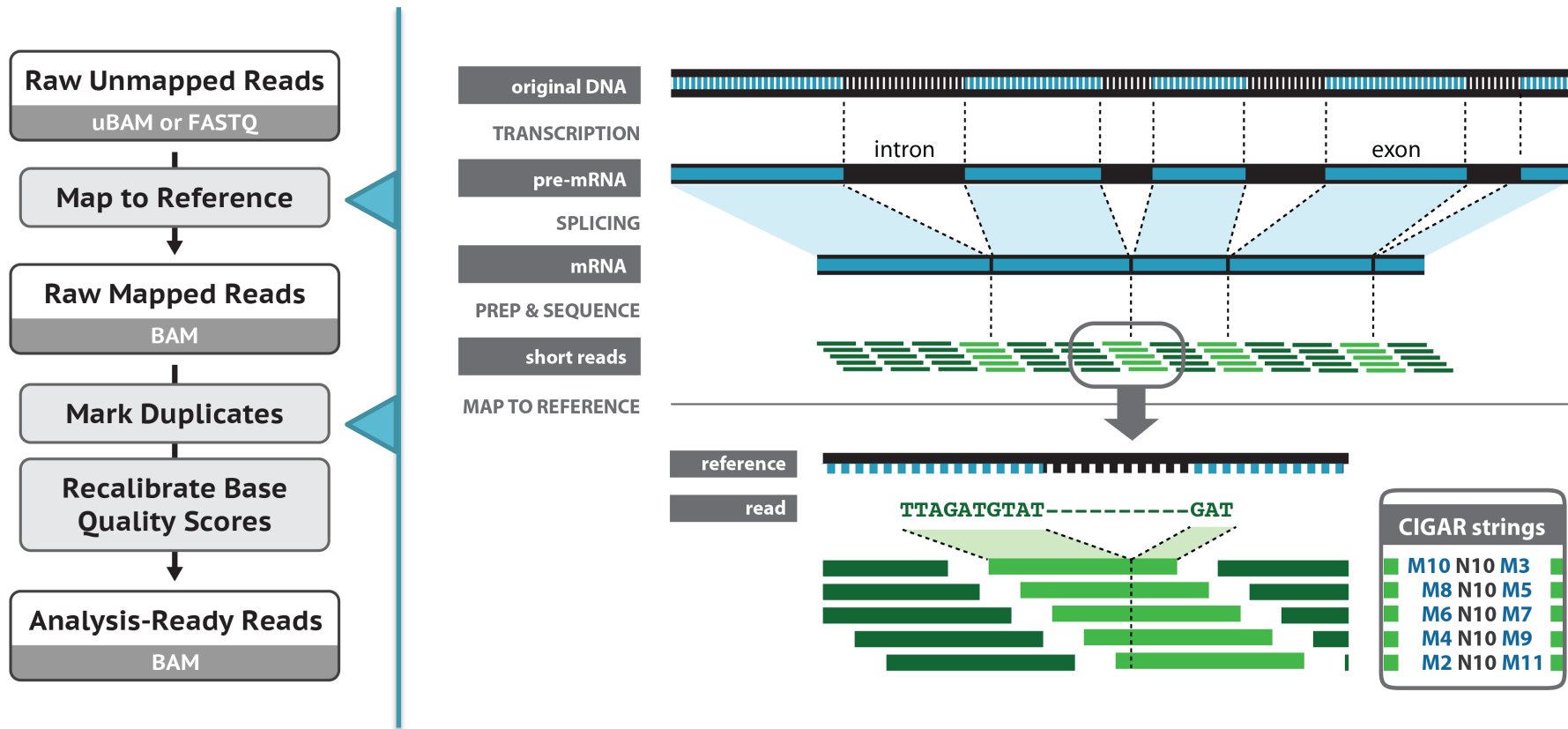


Special Note #3

RNASEQ MAPPING

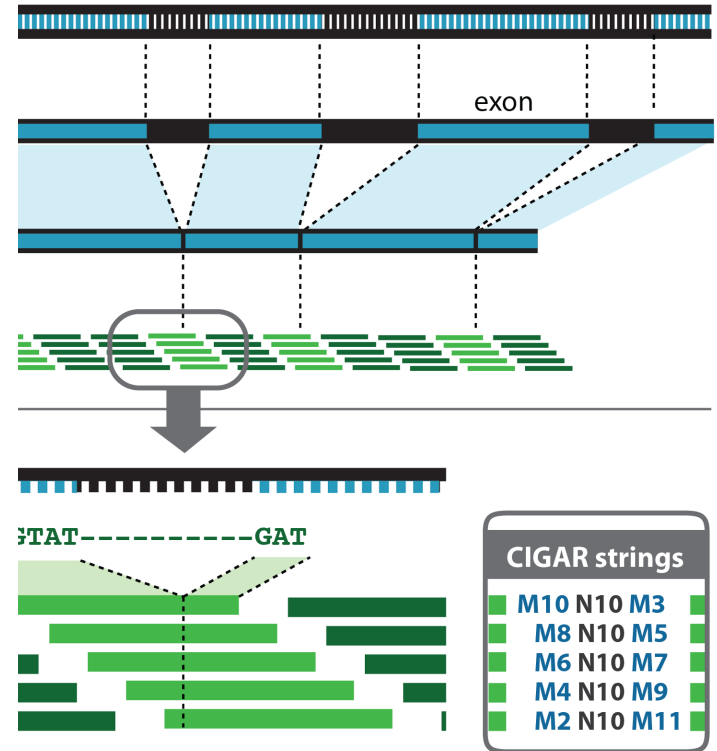


Special handling for RNAseq splice junctions



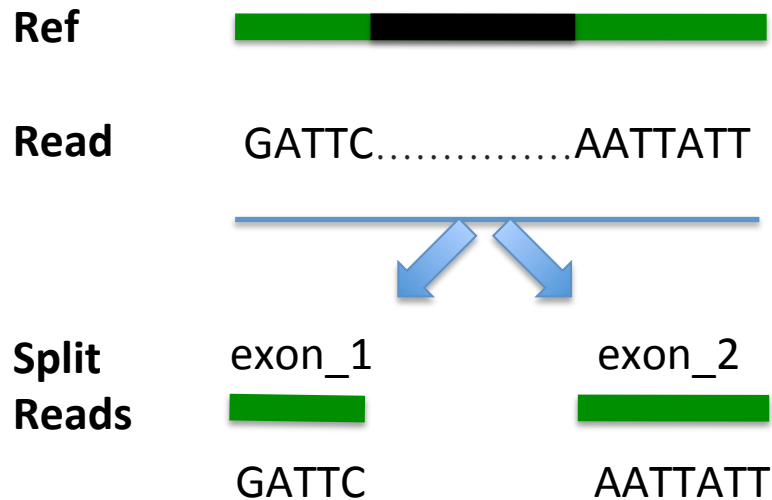
Mapping RNAseq data with STAR

- Highest sensitivity for **both SNPs and indels** among all programs tested
- 2-pass approach described in
Pär G Engström et al. “Systematic evaluation of spliced alignment programs for RNA-seq data”. *Nature Methods*, 2013
(see Suppl.I text p. 43 for detailed protocol)
 - First pass identifies splice junctions (SJ)
 - Use the SJ to guide the second round of alignment



Split'N'Trim

1. Split reads with Ns in the CIGAR string



2. Trim overhangs



Data Pre-processing for Variant Discovery

