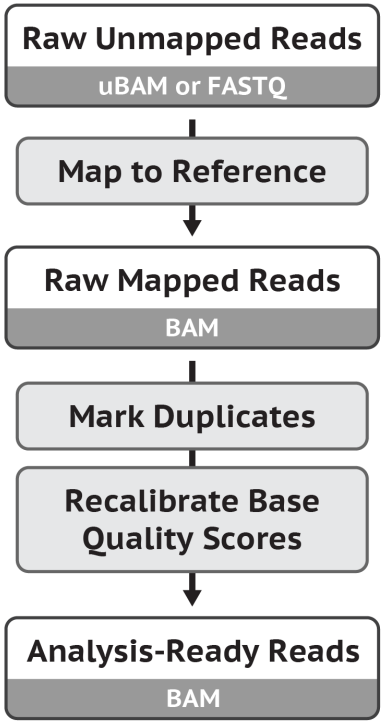
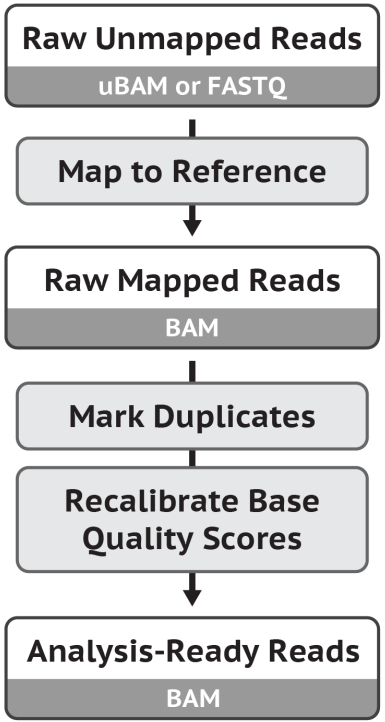
Mapping

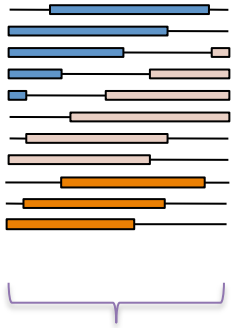
Finding where reads belong in the genome

Data Pre-processing for Variant Discovery



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

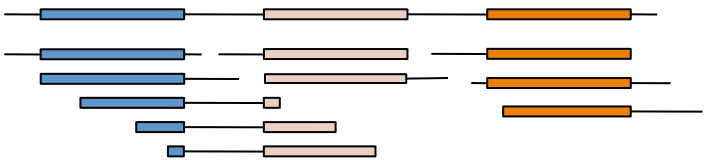




Enormous pile of short reads from HTS

Mapping and alignment algorithms

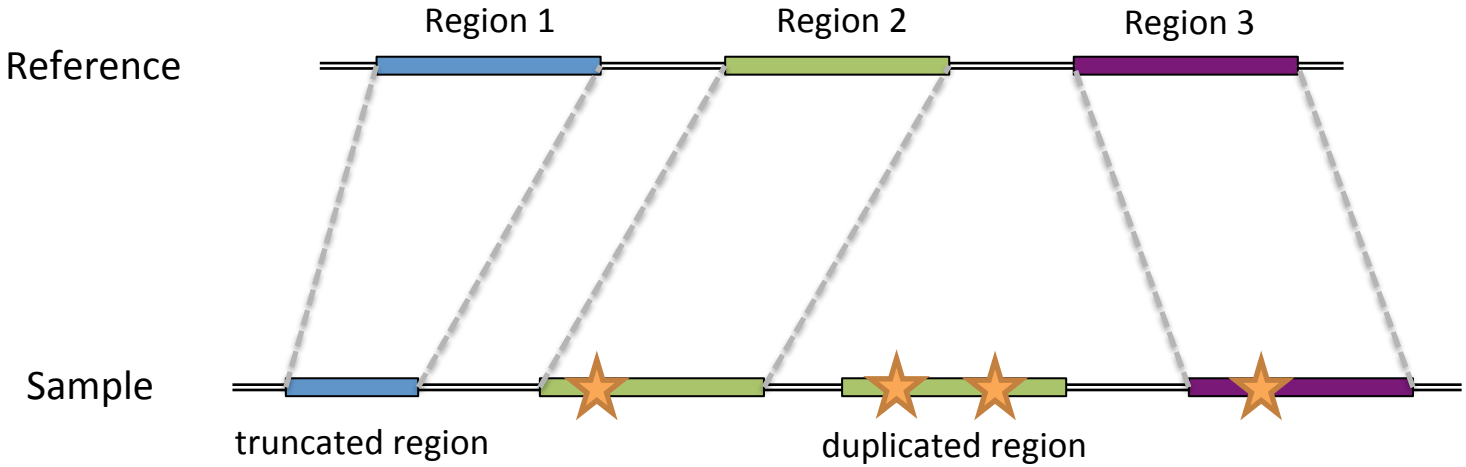
- **BWA for DNA**
- **STAR for RNAseq**



Reference genome

Reads mapped to reference

Goal: align the sample genome to the reference genome

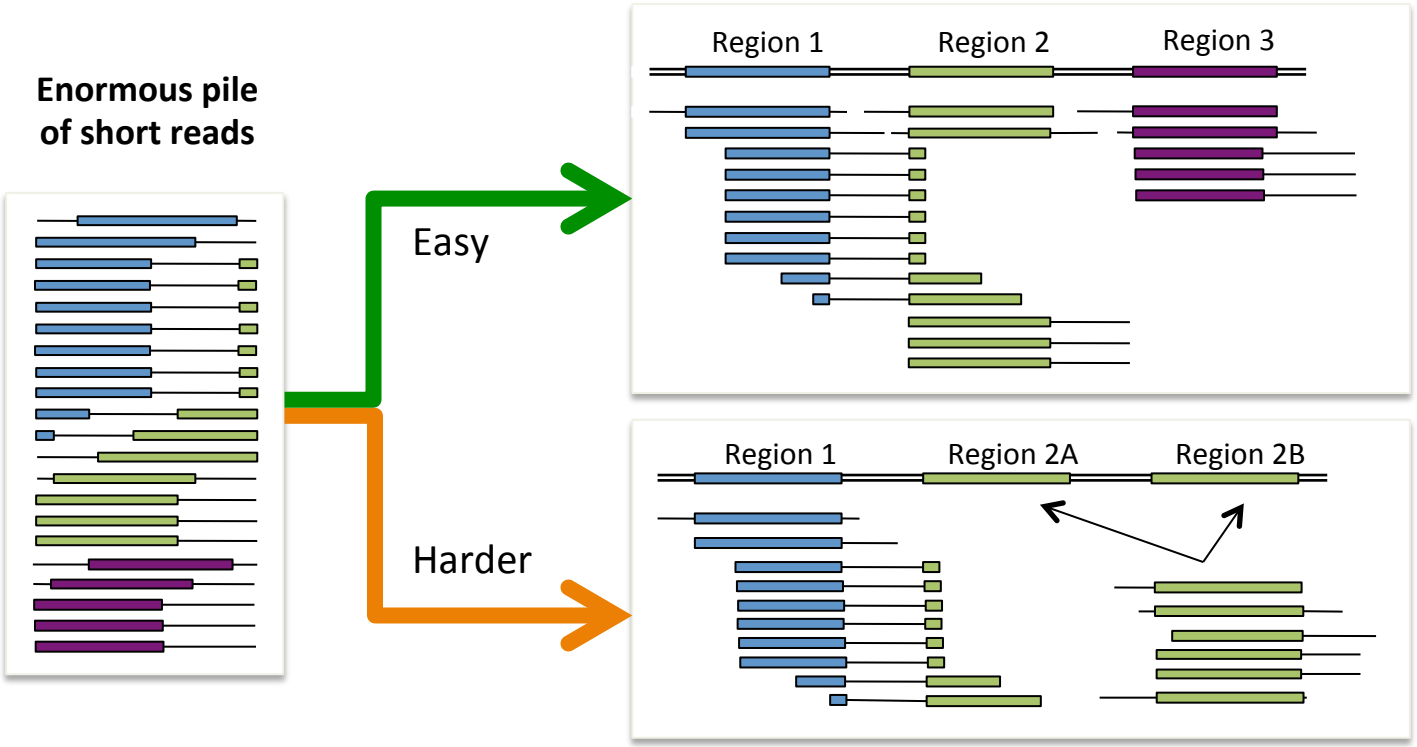


= local variant (SNP/indel)

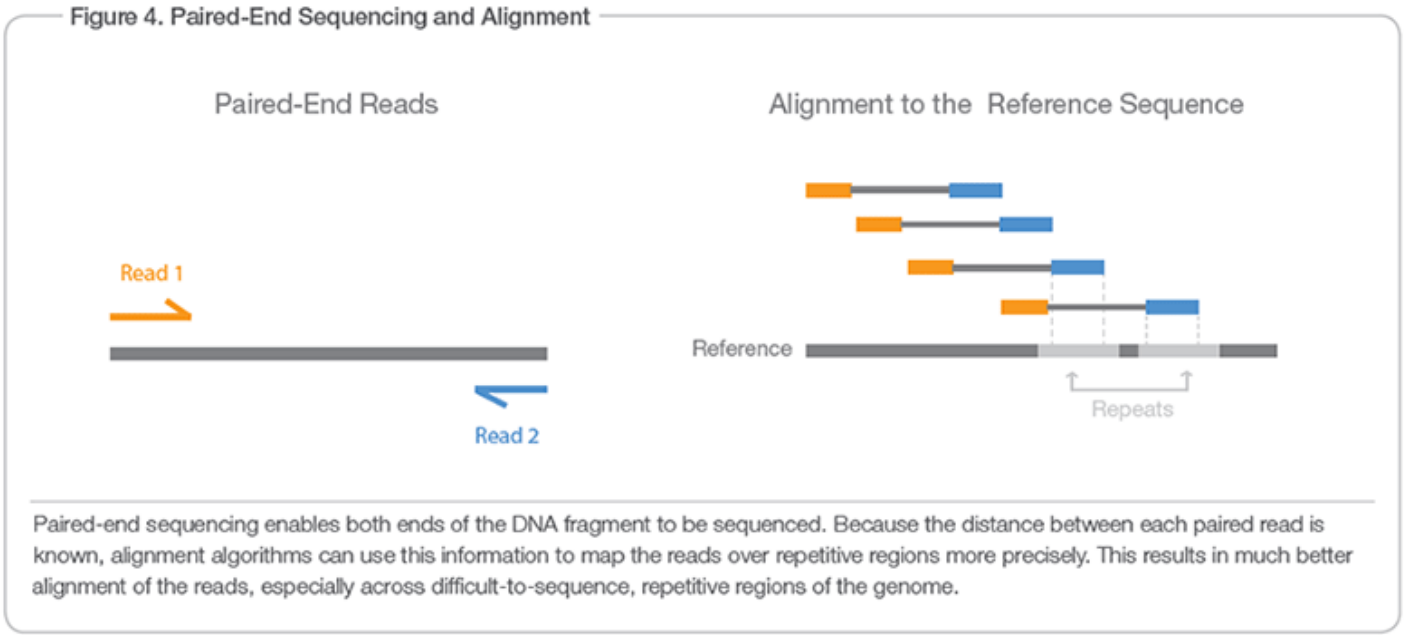
…But we don’t have the whole sample in one piece.

So we have to map each little bit one by one

Complication: mismatches, indels, duplicated regions…



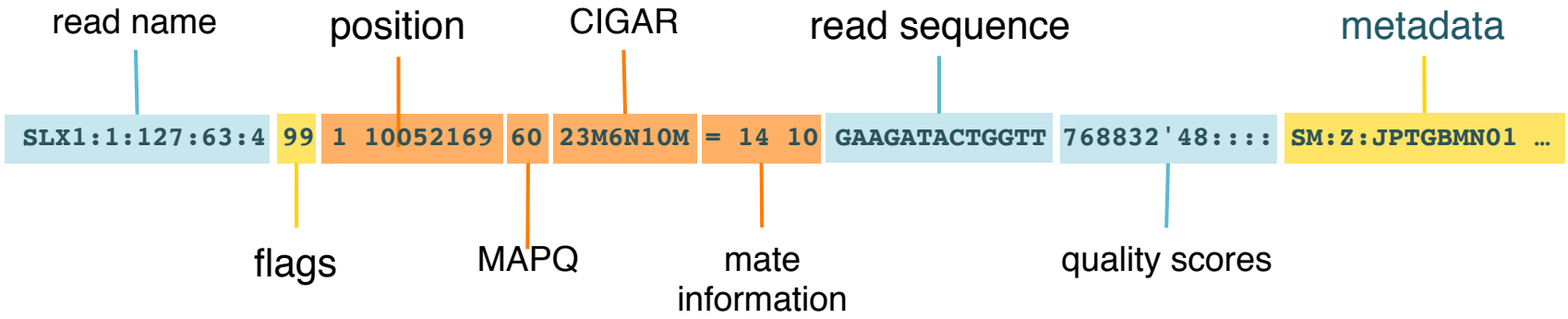
Paired-end sequencing helps a lot



https://www.illumina.com/science/technology/next-generation-sequencing/paired-end-vs-single-read-sequencing.html

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)



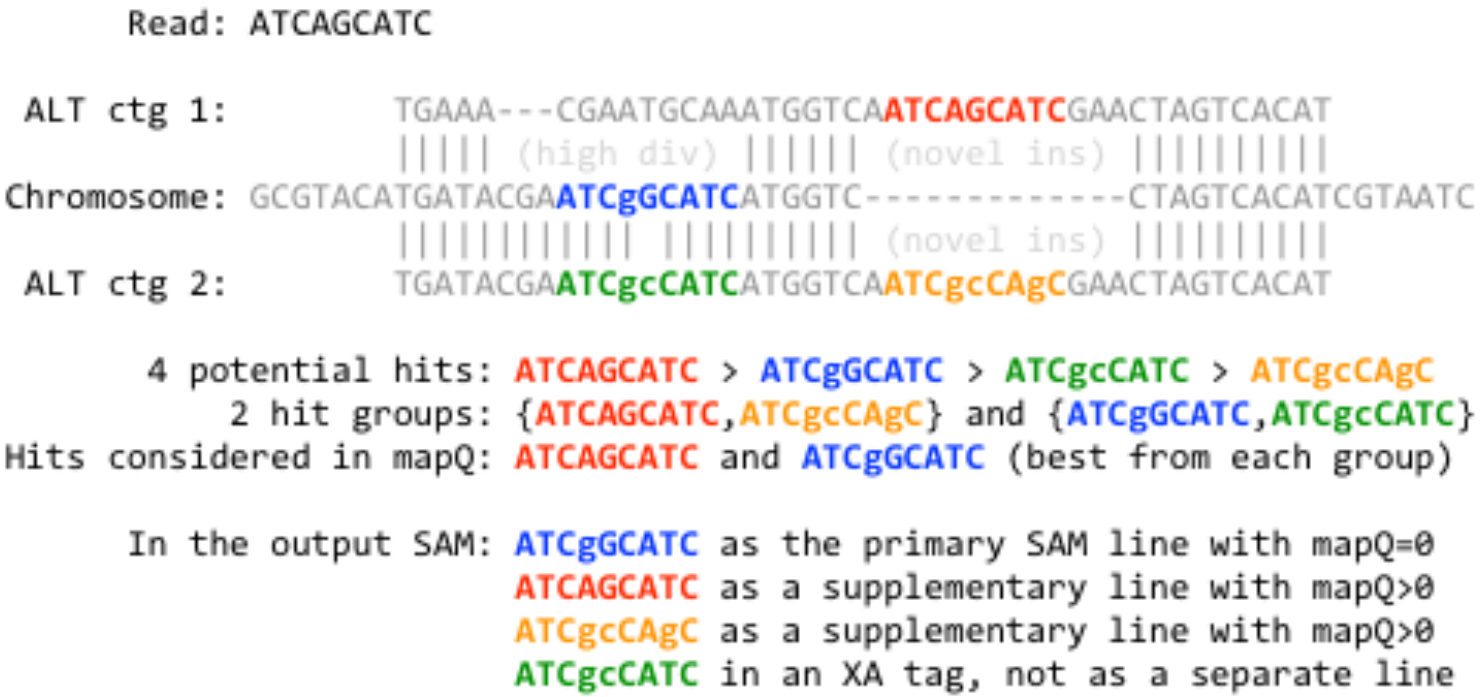
- Added mapping info summarizes **position**, **quality**, and **structure** for each **read**

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

Special Note #1

ALT CONTIGS IN HG38

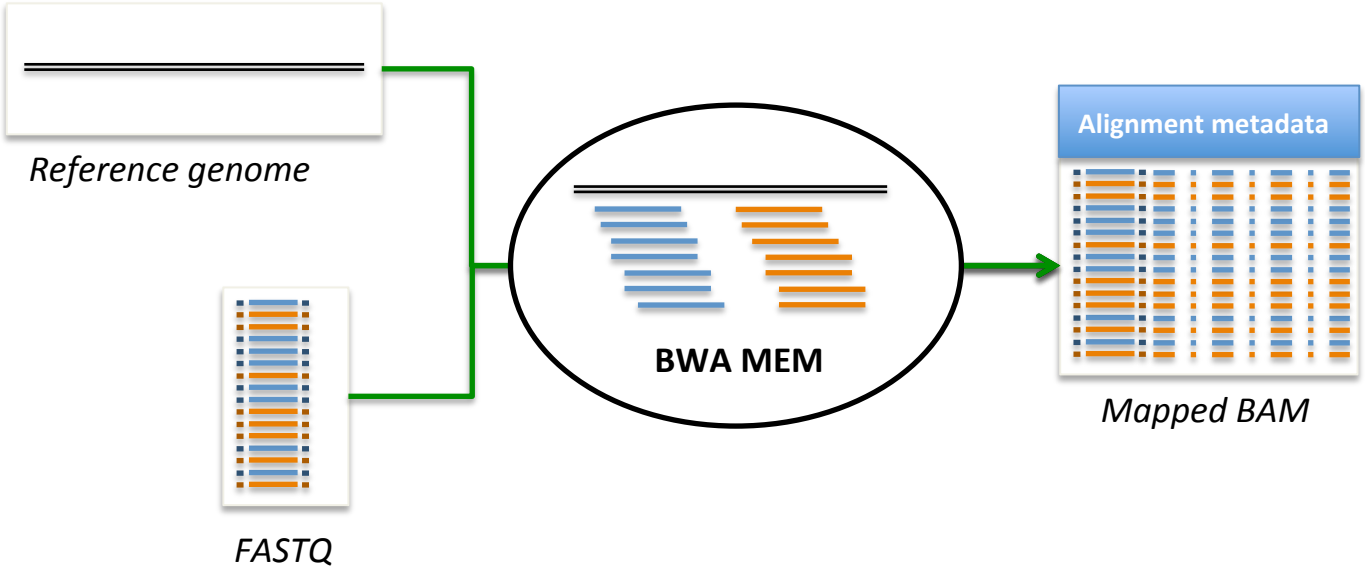
How BWA handles ALT contigs



Special Note #2

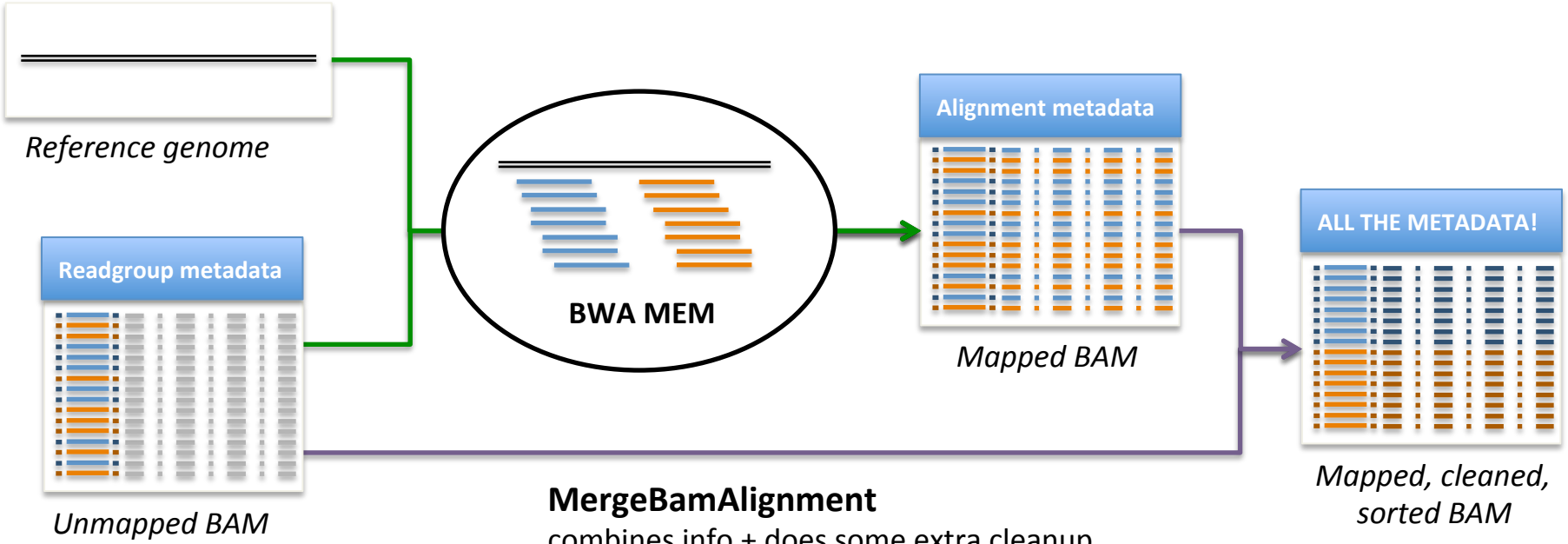
THE UNMAPPED BAM WORKFLOW

Regular FASTQ -> BAM workflow



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

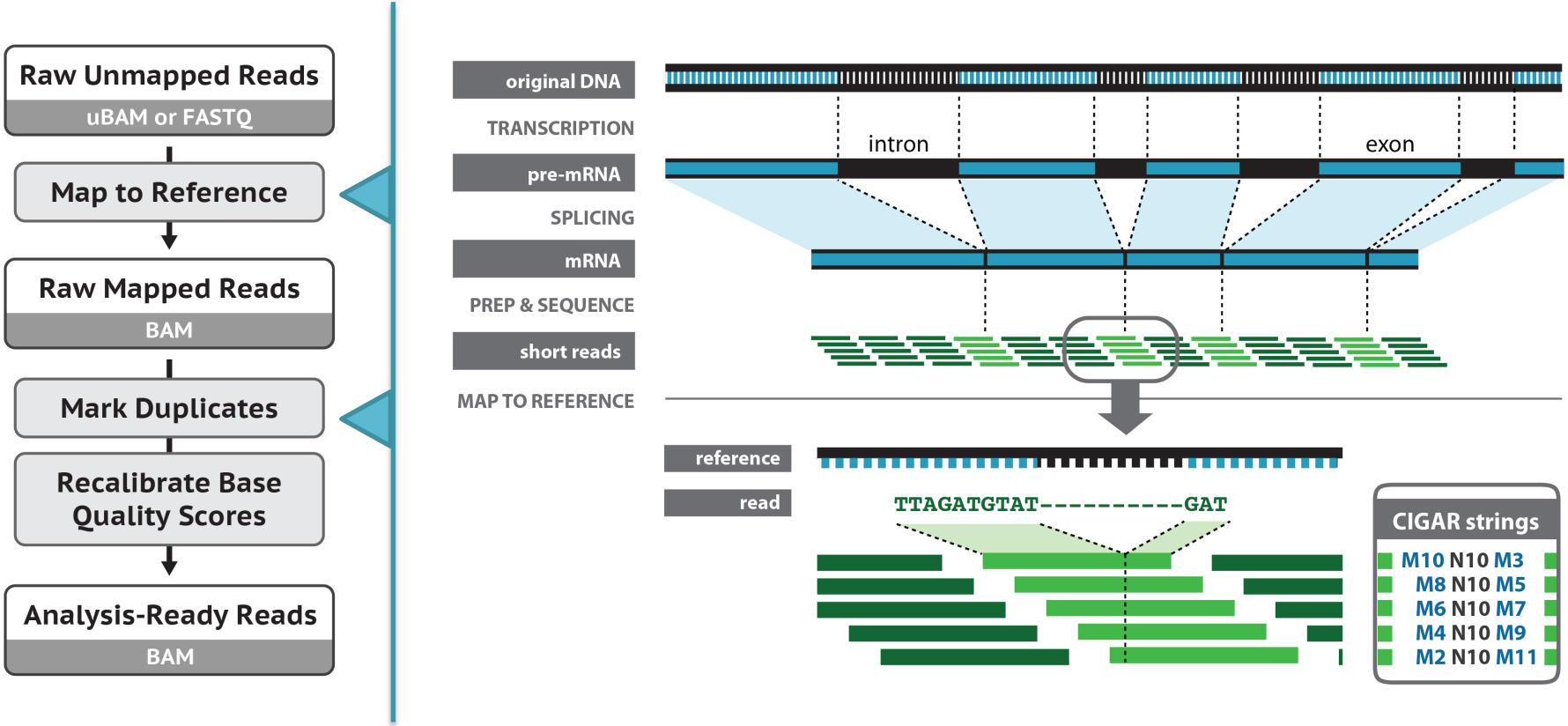
Unmapped BAM -> BAM workflow



Special Note #3

RNASEQ MAPPING

Special handling for RNAseq splice junctions



Mapping RNAseq data with STAR

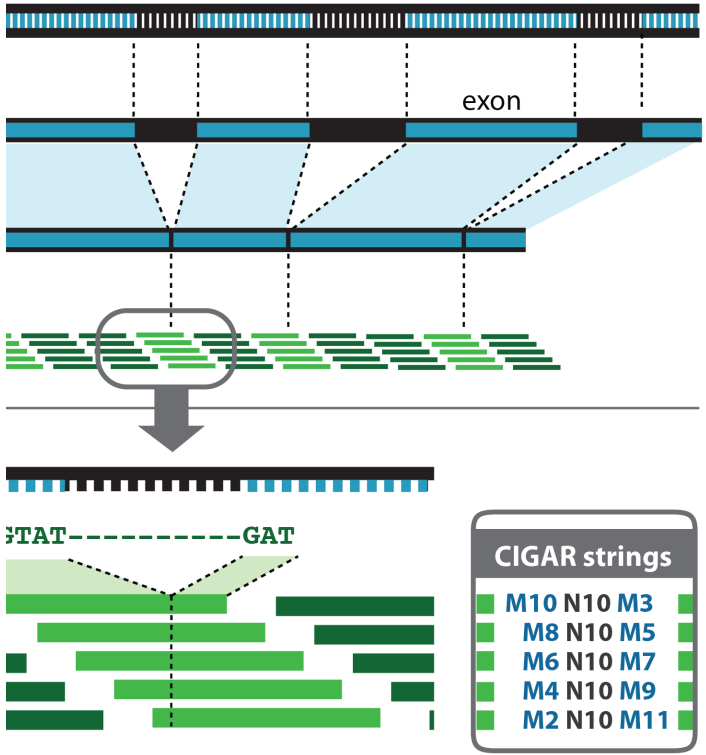
- Highest sensitivity for
  both SNPs and indels
  among all programs tested

- 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.l text p. 43 for detailed protocol)

  – First pass identifies
  splice junctions (SJ)

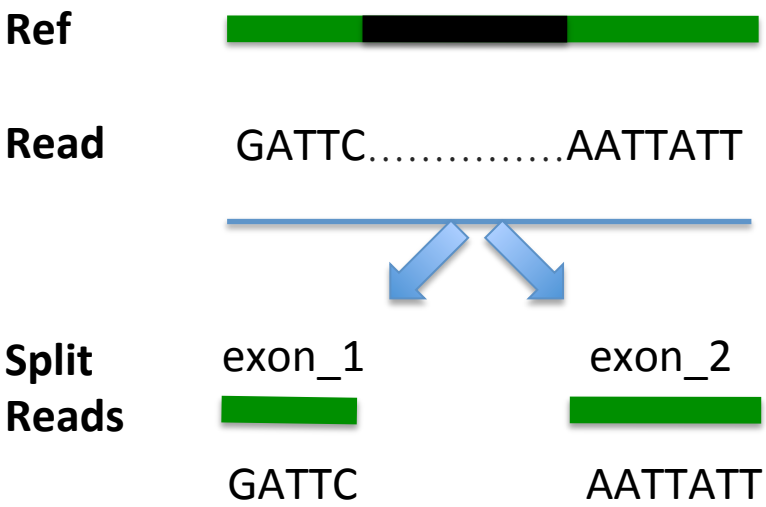
  – Use the SJ to guide
  the second round of alignment



STAR by Dobin *et al.*, 2012
http://bioinformatics.oxfordjournals.org/content/29/1/15

Split'N'Trim

1. Split reads with Ns in the CIGAR string



2. Trim overhangs

Data Pre-processing for Variant Discovery