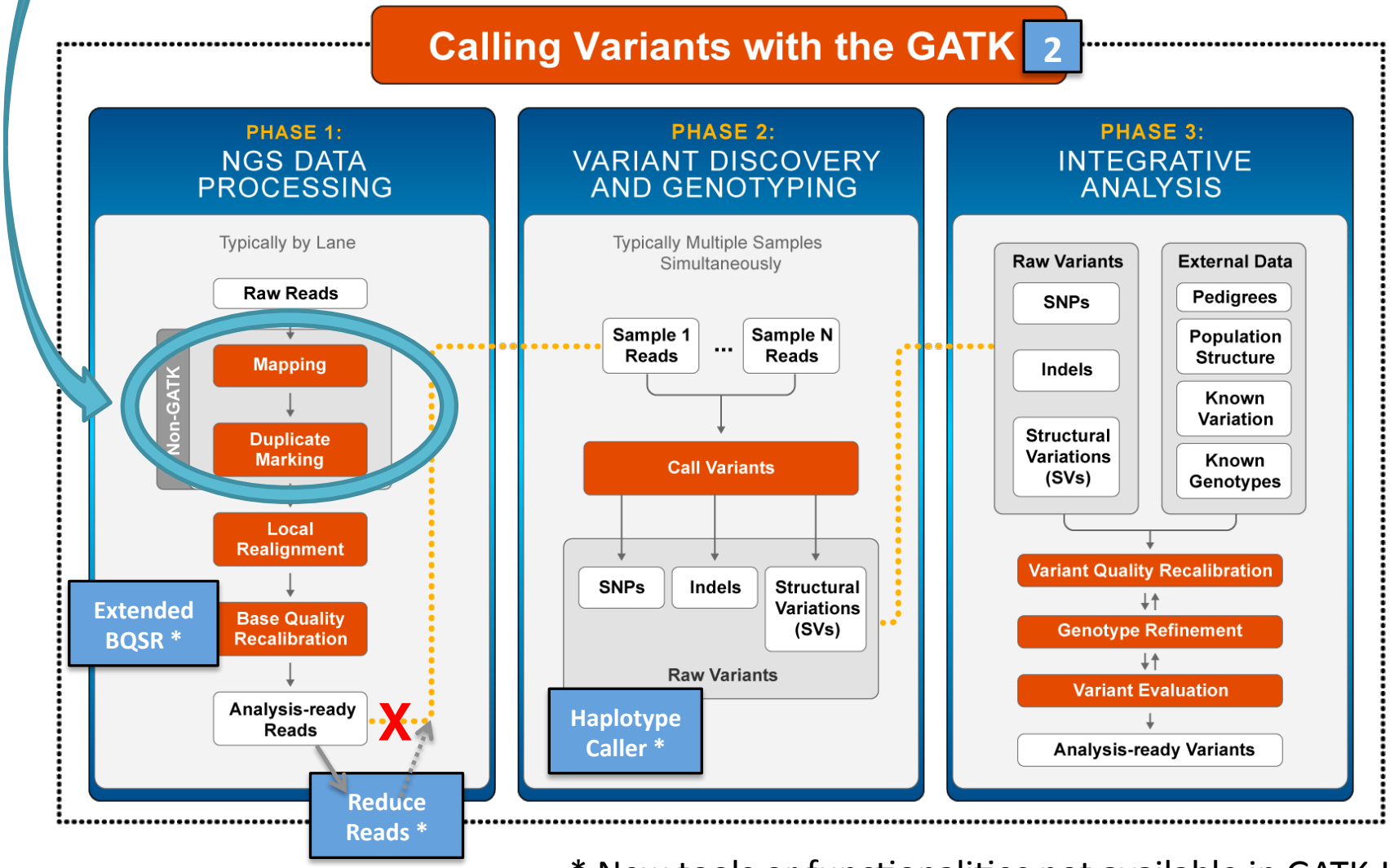
Mapping and duplicate marking

First step from machine to analysis:

place the reads in the right place on the reference and eliminate duplicates

We are here in the Best Practices workflow

MAPPING & DEDUPPING

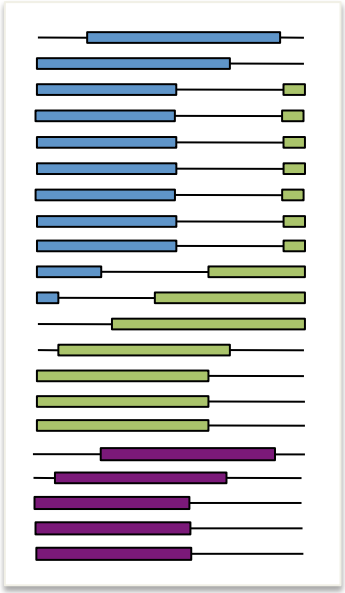


* New tools or functionalities not available in GATK-Lite

Overview of this step's goals

Reference genome

Enormous pile of short reads from NGS



➔ Map reads to reference with **BWA**

All later steps assume that reads are placed in the right location and represent that region of the genome.

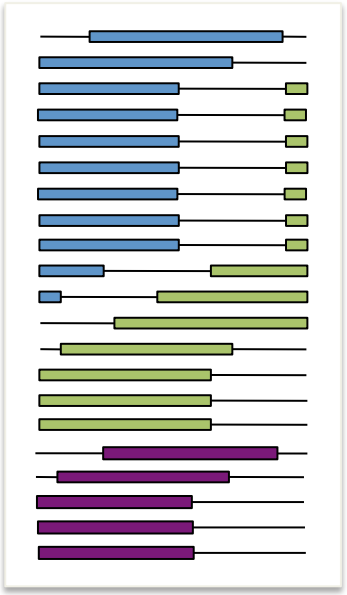
➔ Mark duplicates with **Picard tools**

Duplicates originate mostly from DNA prep methods and cause biases that skew variant calling results.

MAPPING

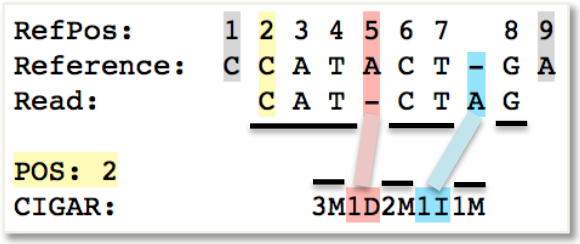
Mapping short reads to a reference is simple in principle

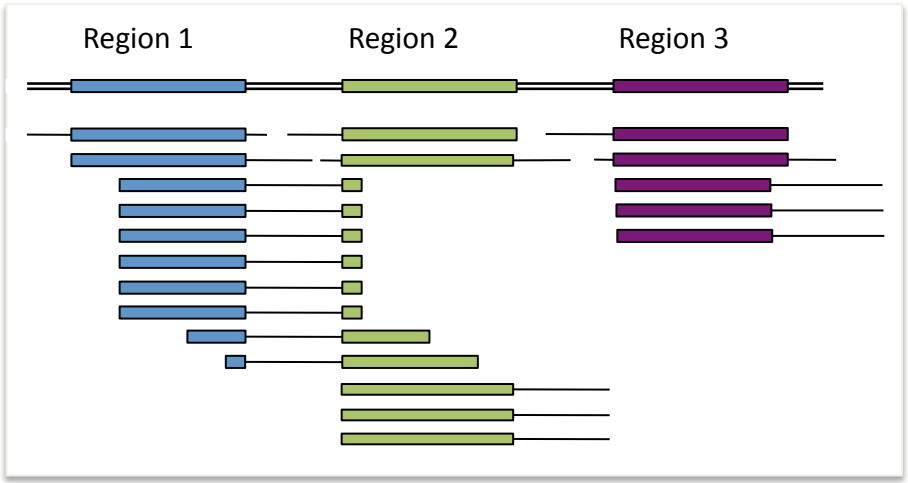
Enormous pile of short reads from NGS



Mapping and alignment algorithms

Identify where the read matches the reference sequence and record match details as CIGAR string



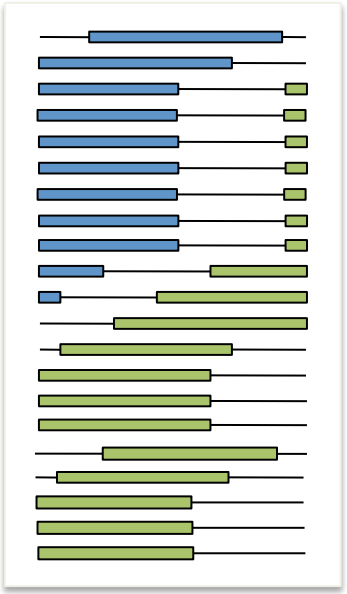


Reference genome

Reads mapped to reference

But mapping is actually very hard because of mismatches (true mutations or sequencing errors), duplicated regions etc.

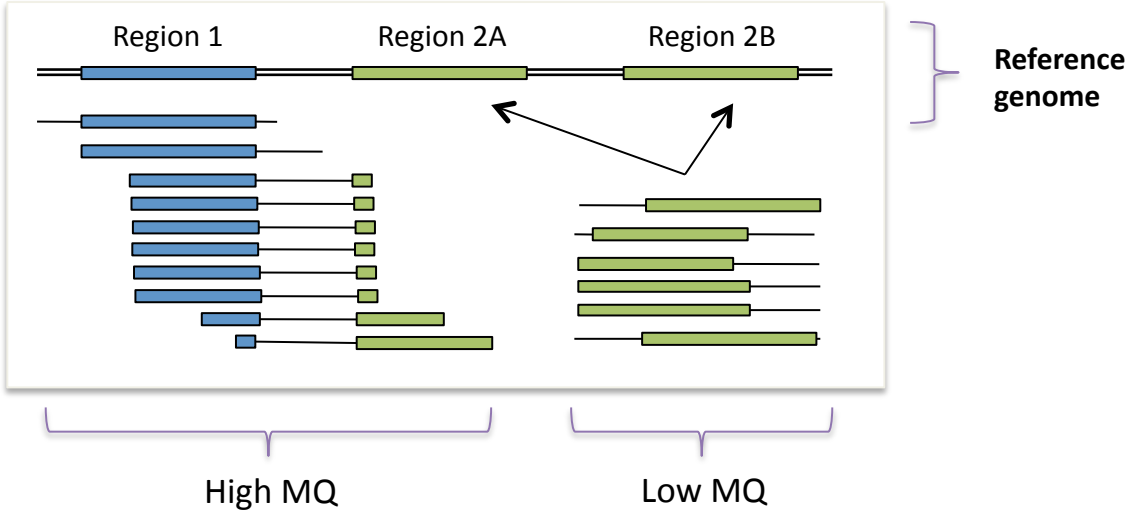
Enormous pile of short reads from NGS



Mapping and alignment algorithms

Mapping algorithms account for this by choosing the most likely placement

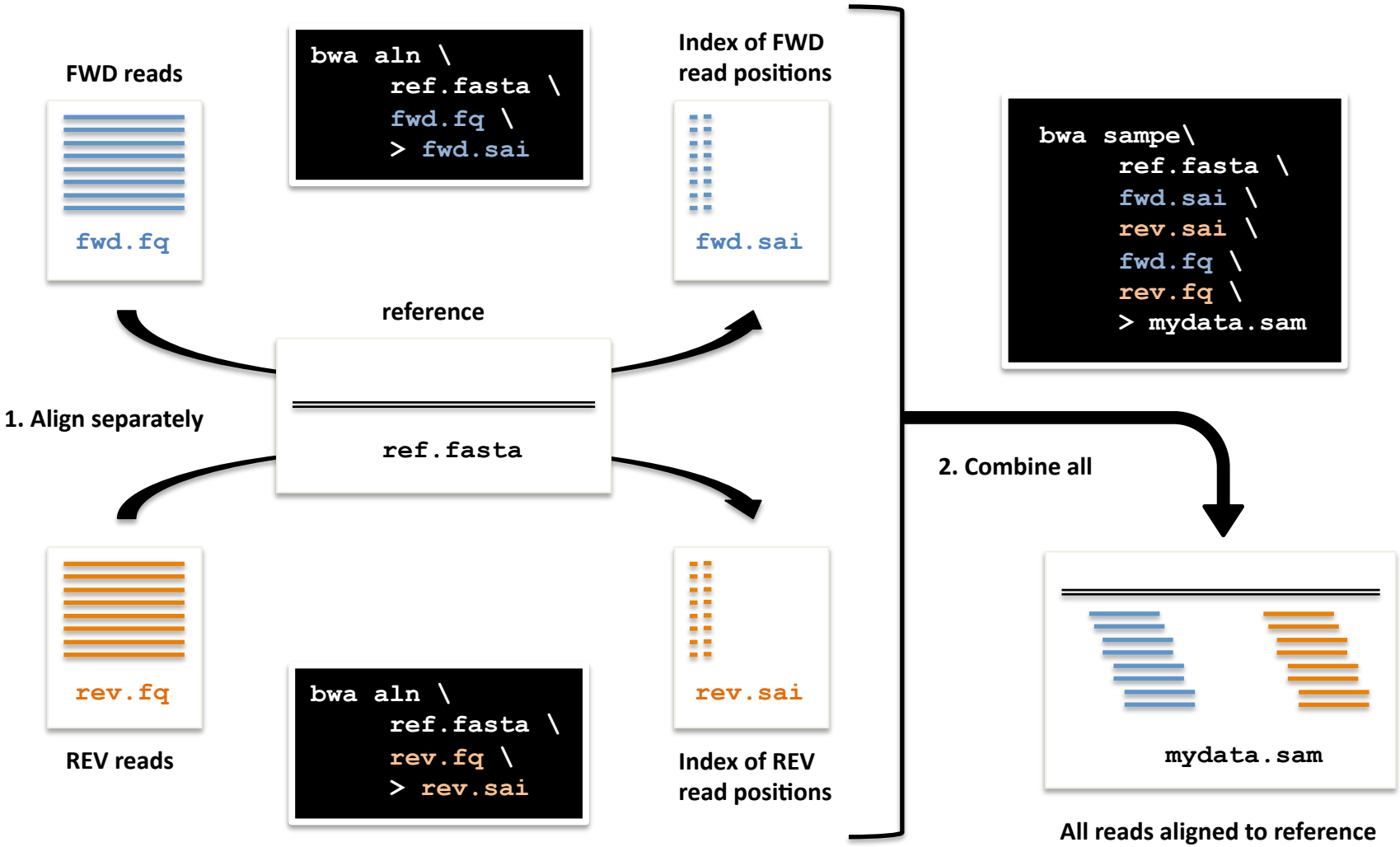
➔ **mapping quality (MQ)**



For more information see:

Li and Homer (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*.

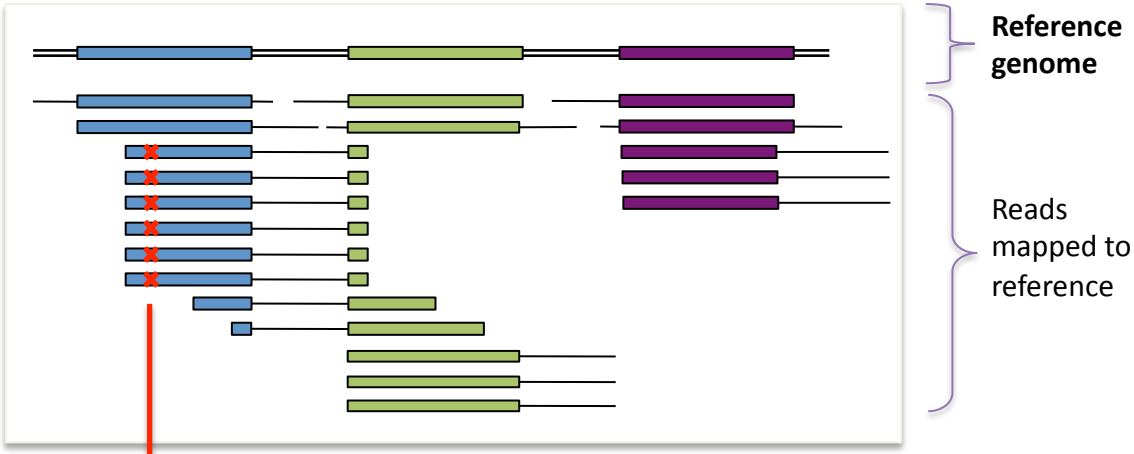
Typical workflow using BWA to map paired-end data



MARKING DUPLICATES

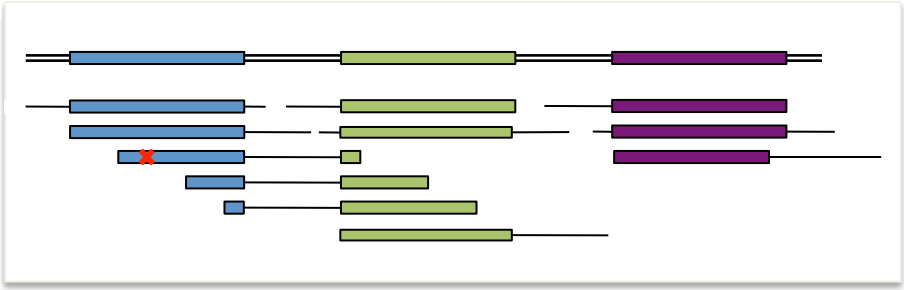
The reason why duplicates are bad

✖ = sequencing error propagated in duplicates



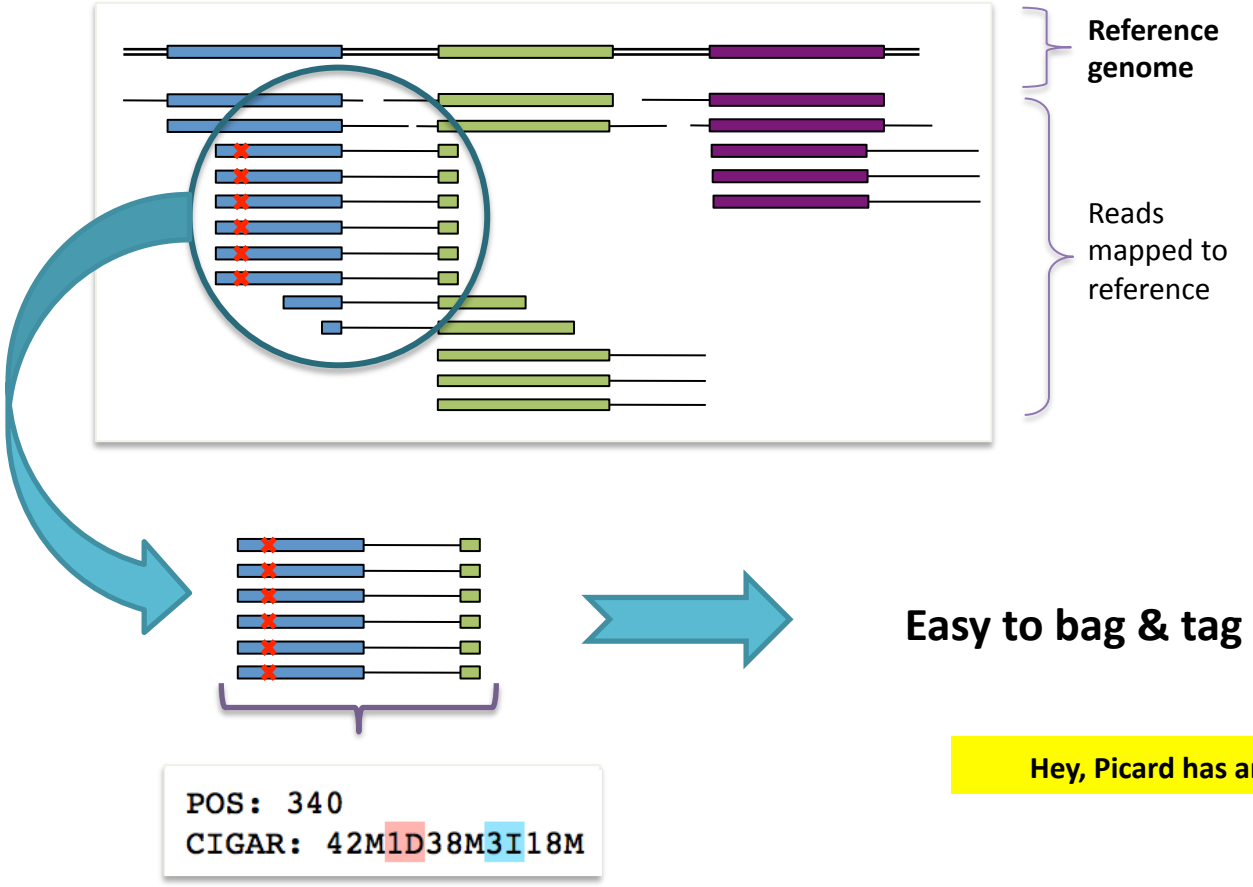
FP variant call (bad)

After marking duplicates, the GATK will only see :



… and thus be more likely to make the right call

Duplicates have the same starting position and the same CIGAR string



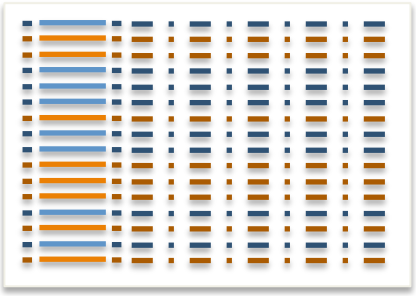
Easy to bag & tag

Hey, Picard has an app for that!

A quick diversion about sorting and read groups

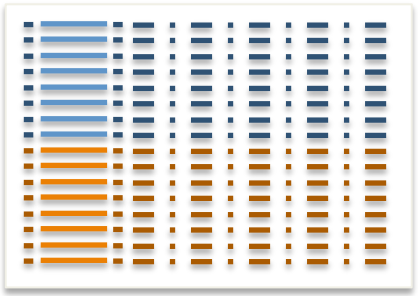
The information for this:

… is actually stored as a text file with one line per read which from far away looks like this:



The reads are in no particular order…

… but the GATK wants reads to be sorted by starting position like this:



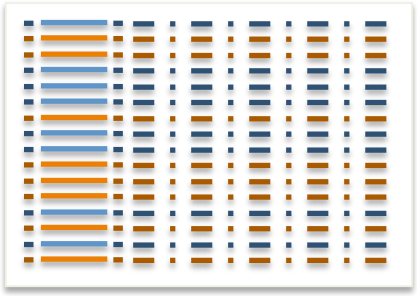
So we need to explicitly sort the SAM file…

… and Picard has an app for that!

And while we're at it, let's add **read group** information if it isn't already there, so **the GATK will know what read belongs to what sample** (that's kind of important).

Hey, Picard has an app for that too!

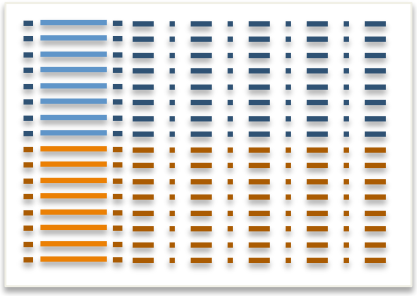
Typical workflow using Picard tools to mark duplicates *et al.*



Original SAM

```
java -jar SortSam.jar \
     I=input.sam \
     O=output.bam \
     SO=coordinate
```

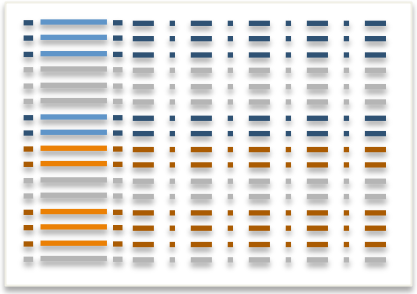
sort



Ordered BAM

```
java -jar MarkDuplicates.jar \
     I=input.sam \
     O=output.bam
```

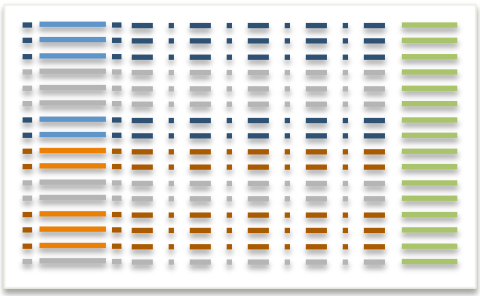
"dedup"



Dedupped BAM

```
java -jar AddOrReplaceReadGroups.jar \
     I=input.bam \
     O=output.bam \
     RGID=id RGLB=solexa-123 \
     RGPL=illumina RGPU=AXL2342 \
     RGSM=NA12878 RGCN=bi \
     RGDT=12/12/2011
```

add RG



Final BAM

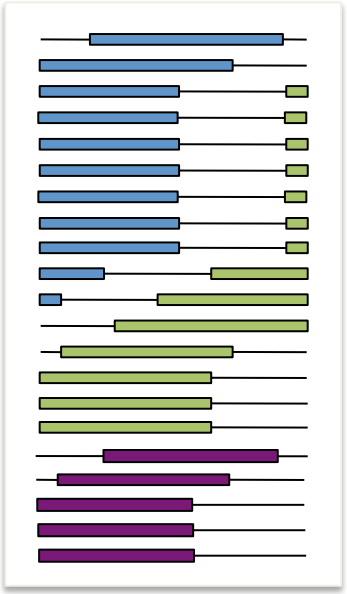
Wham, BAM, thank you Picard!

TO CONCLUDE

Recap of what was achieved

Reference genome

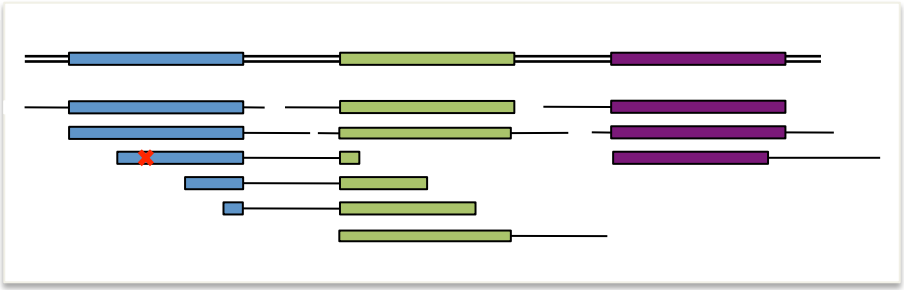
Enormous pile of short reads from NGS



➔ Mapped reads to reference with **BWA**

➔ Marked duplicates with **Picard tools**

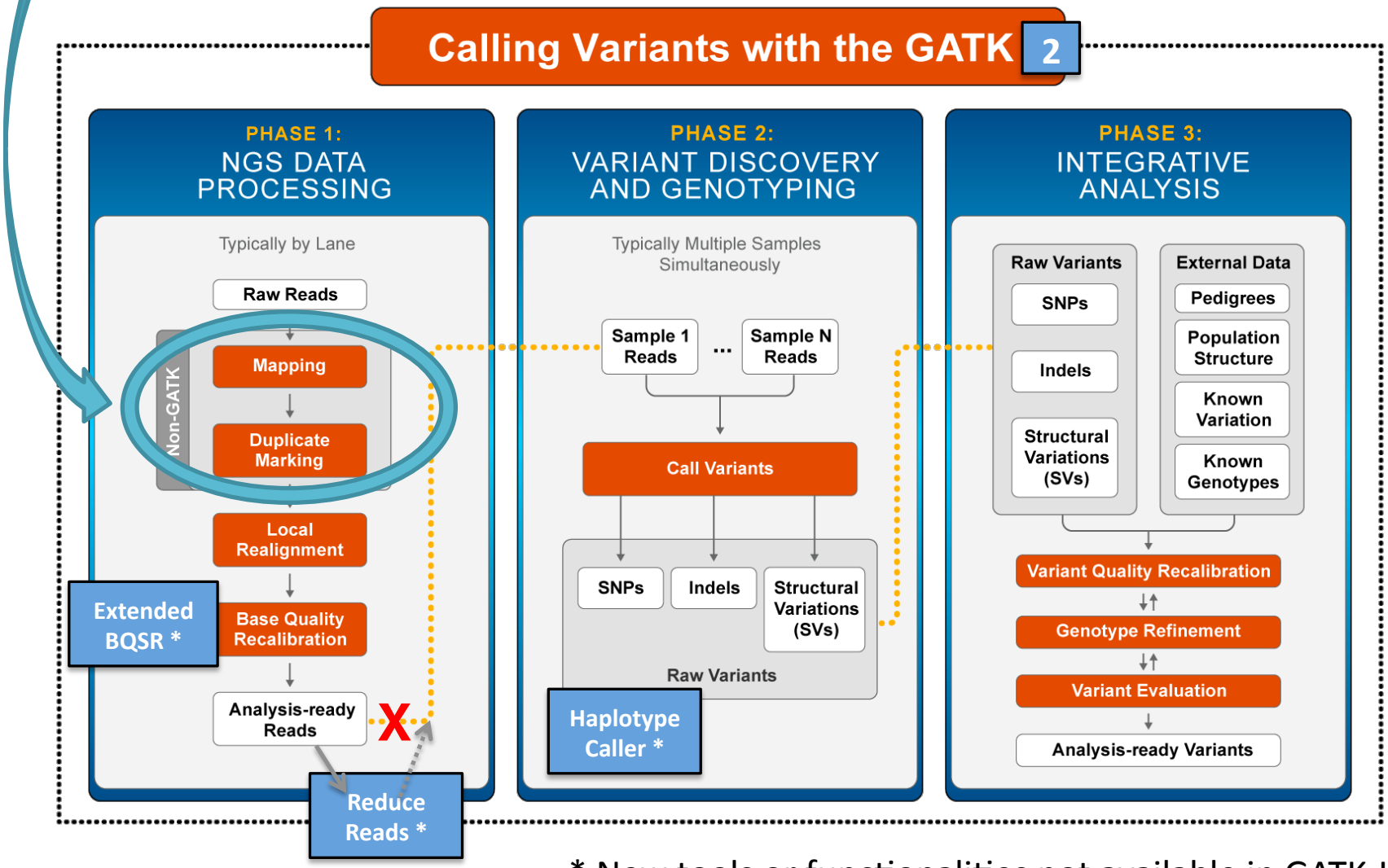
(and did some additional prep work)



Reads mapped, sorted, dedupped, +RGs

We were here in the Best Practices workflow

NEXT STEP: REALIGNMENT



* New tools or functionalities not available in GATK-Lite

Further reading

http://www.broadinstitute.org/gatk/guide/topic?name=intro

http://www.broadinstitute.org/gatk/guide/topic?name=best-practices