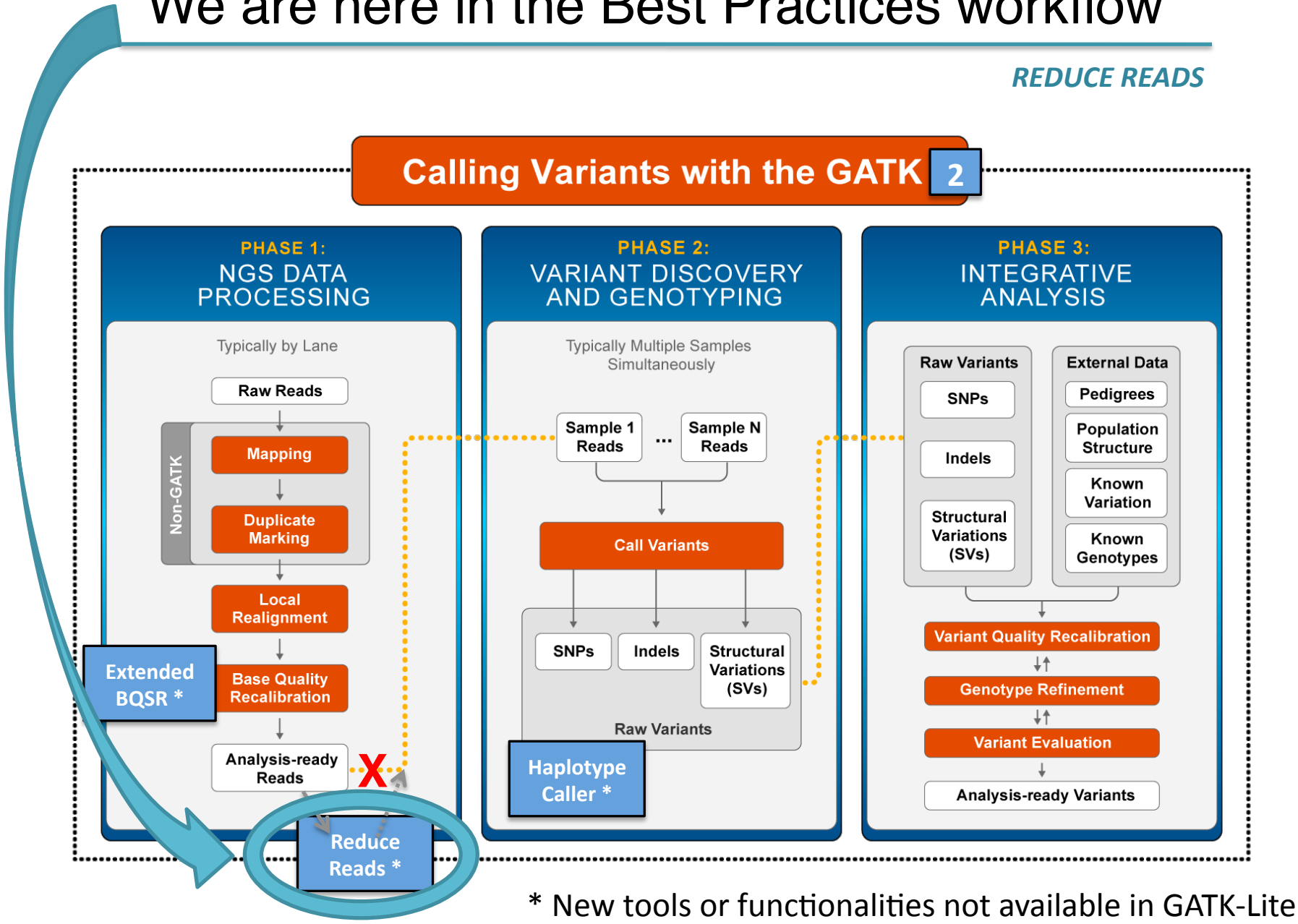
Data Compression with Reduce Reads

Reducing the BAM file to a manageable size that allows greater performance and scalability for the GATK analysis tools

We are here in the Best Practices workflow



* New tools or functionalities not available in GATK-Lite

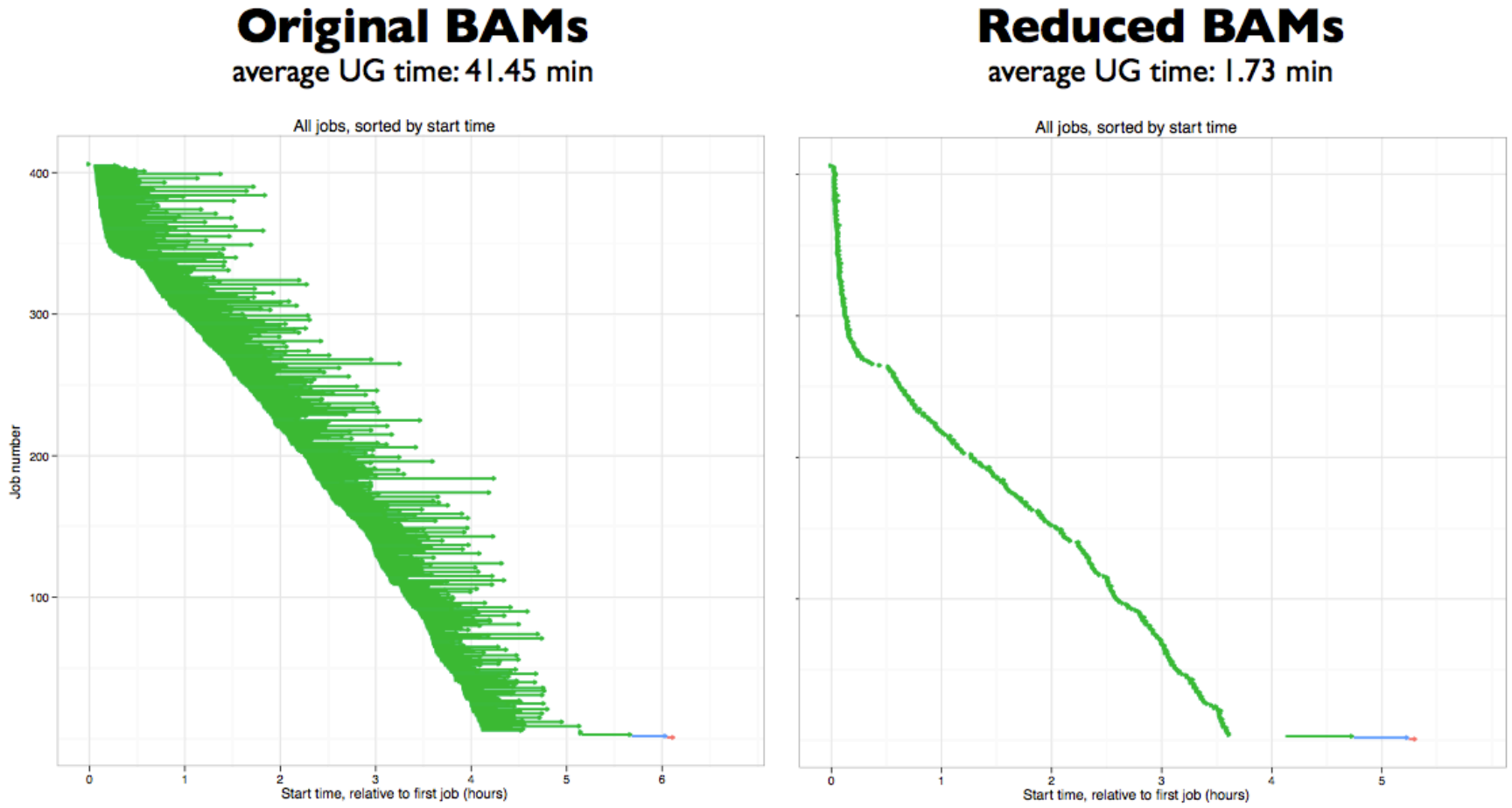
PURPOSE

Why compress NGS data?

- BAM file sizes are huge
- File transfer is impractical
- Simple analysis takes too long
- Complex or large scale analysis is non-viable
  - Batching also has issues

☑ **The size of the BAM file is a major roadblock for data analysis scalability**

Reducing the size of the BAM file allows greater performance in simple analysis and scaling to tens of thousands of BAMs in complex or large scale projects



PRINCIPLES

Compression = throw out redundant information

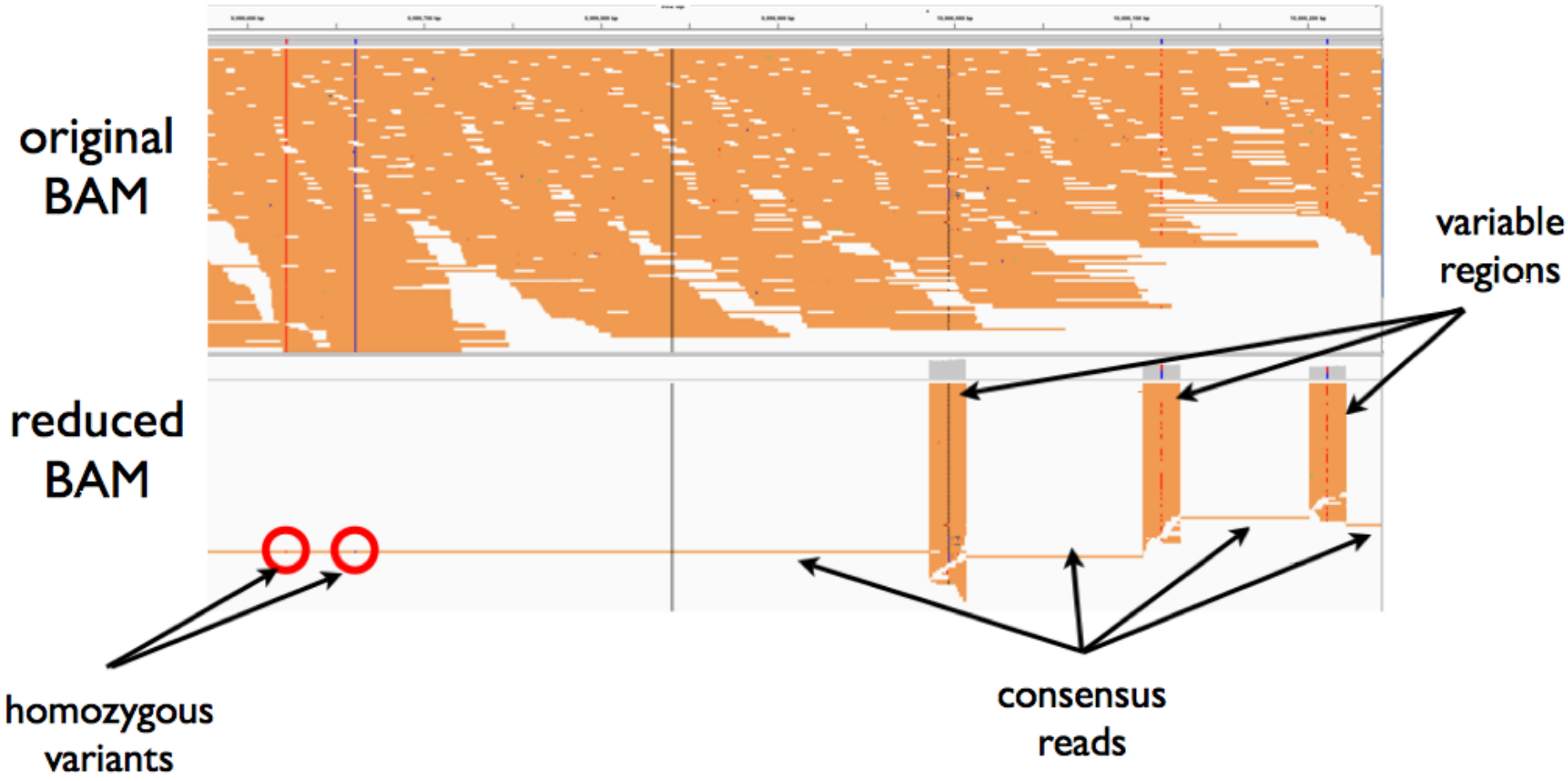
All reads agree

Almost all reads agree

Read-based compression keeps only essential information for variant calling

- Distinguishes variable and consensus regions
- Variable regions are windows around the disagreement between the reads with sufficient information for subsequent analysis.
- A disagreement can be triggered for any generic analysis goal with different thresholds (e.g. heterozygous sites, insertions, deletions).
- Original reads are downsampled to a “more than reasonable” coverage for analysis.
- Despite being clipped, original offsets and length information can still be inferred from the reads in the variable region for annotations.
- Tumor and Normal samples (or any set of samples) get co-reduced, meaning that every variable region triggered by one sample will be forced in every sample.

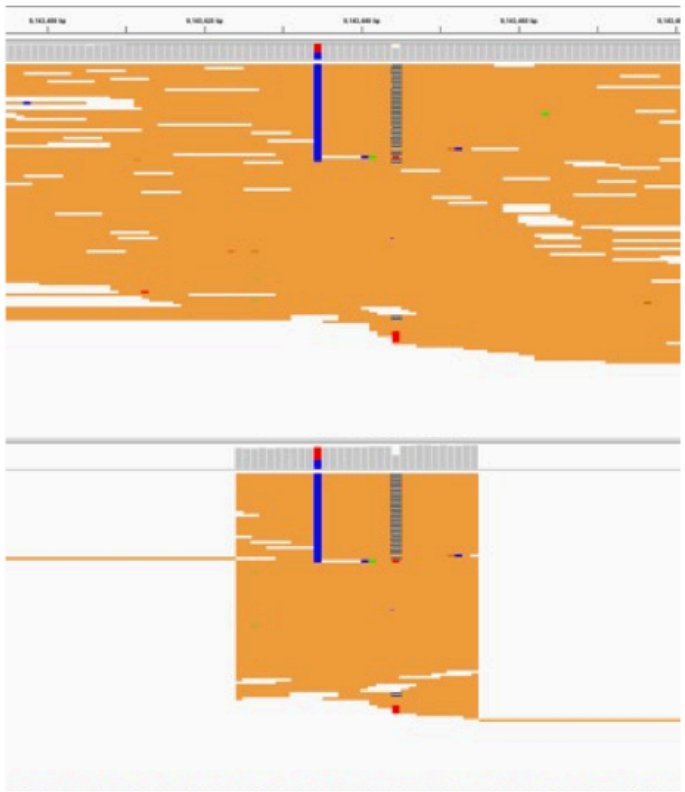
This is what a compressed BAM looks like



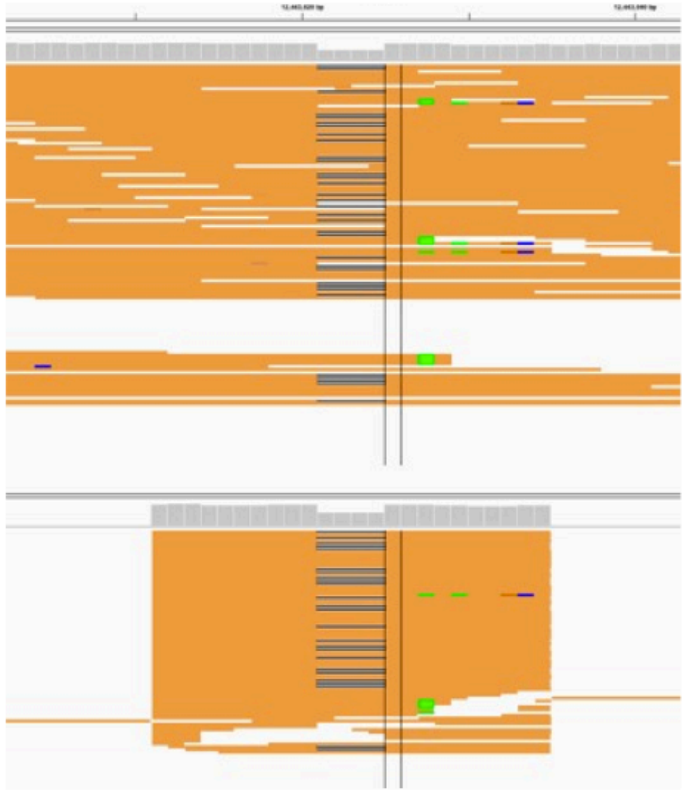
Important to handle complex cases properly

original BAM

reduced BAM



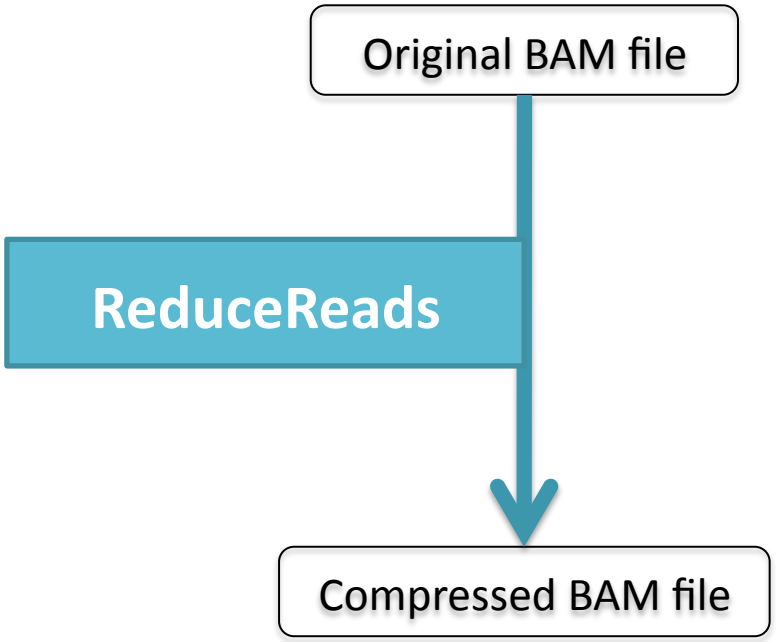
multiple variants merging variant region



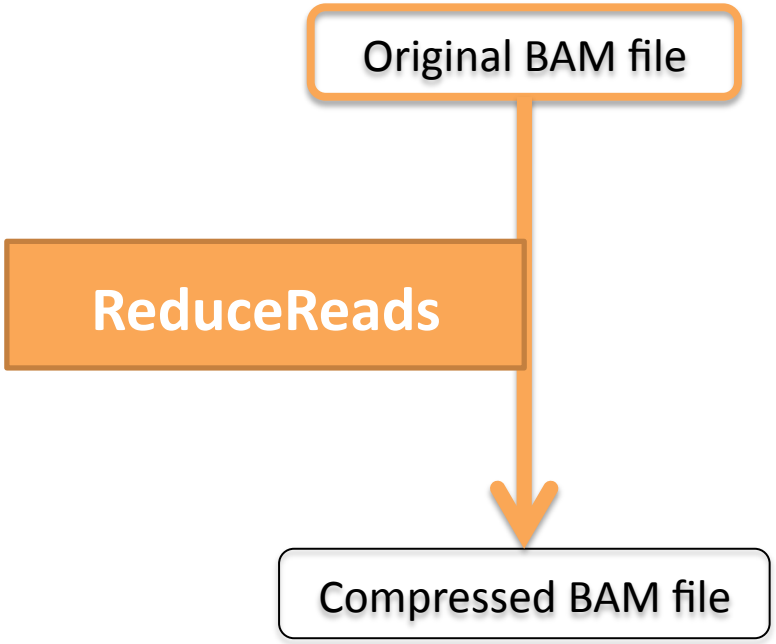
long deletion

PROTOCOL

Compression workflow



Compression workflow



TOOL TIPS

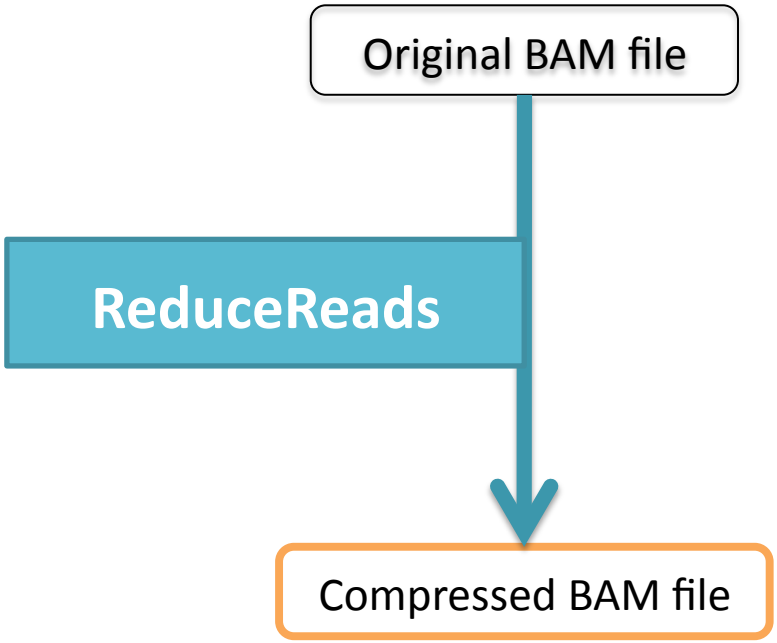
ReduceReads

New in 2

- Reduces the BAM file using read based compression that keeps only essential information for variant calling

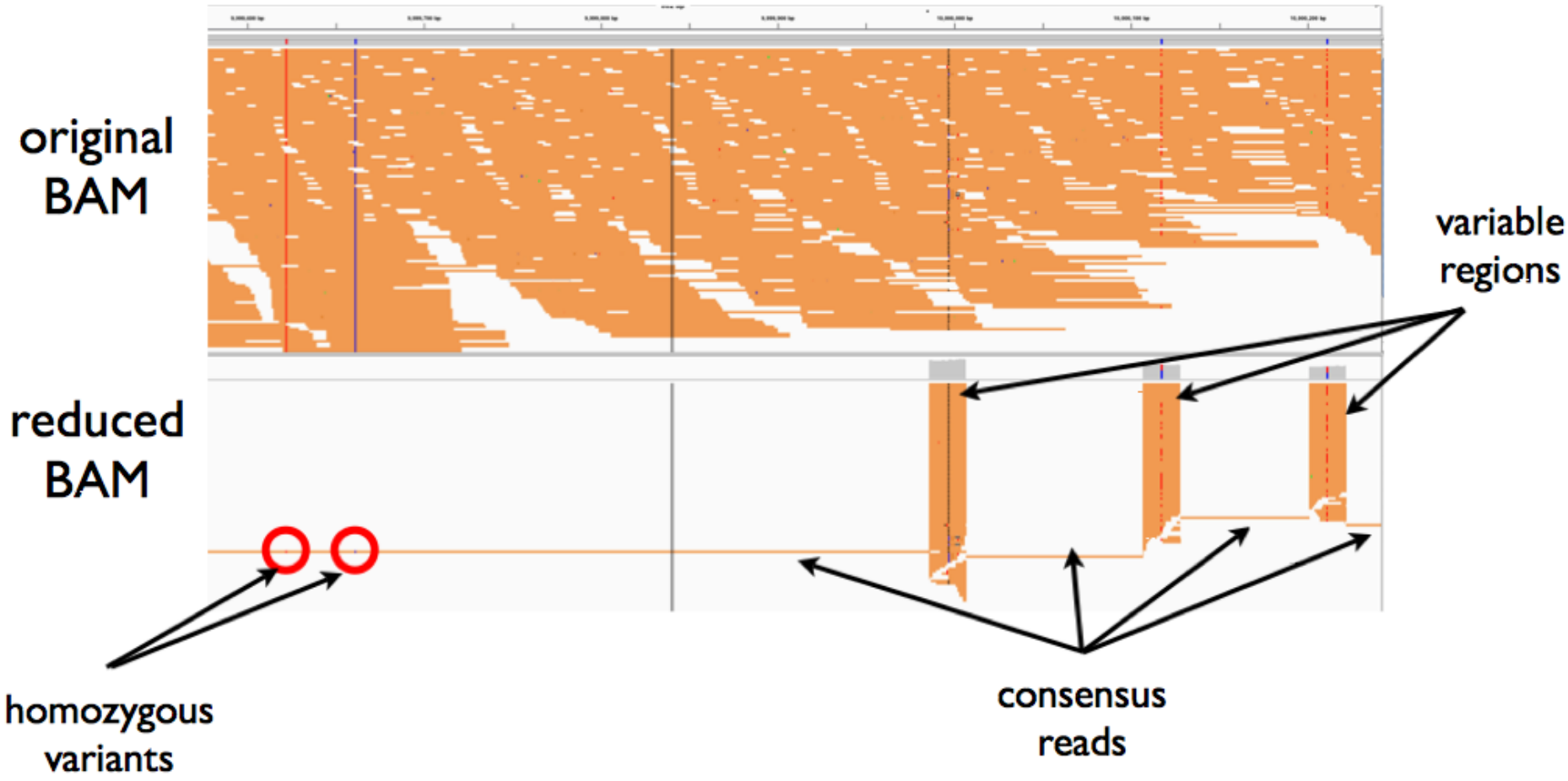
```
java –jar GenomeAnalysisTK.jar –T ReduceReads \
      –R human.fasta \
      –I recal.bam \
      –o reduced.bam
```

Compression workflow



RESULTS

This is what a compressed BAM looks like

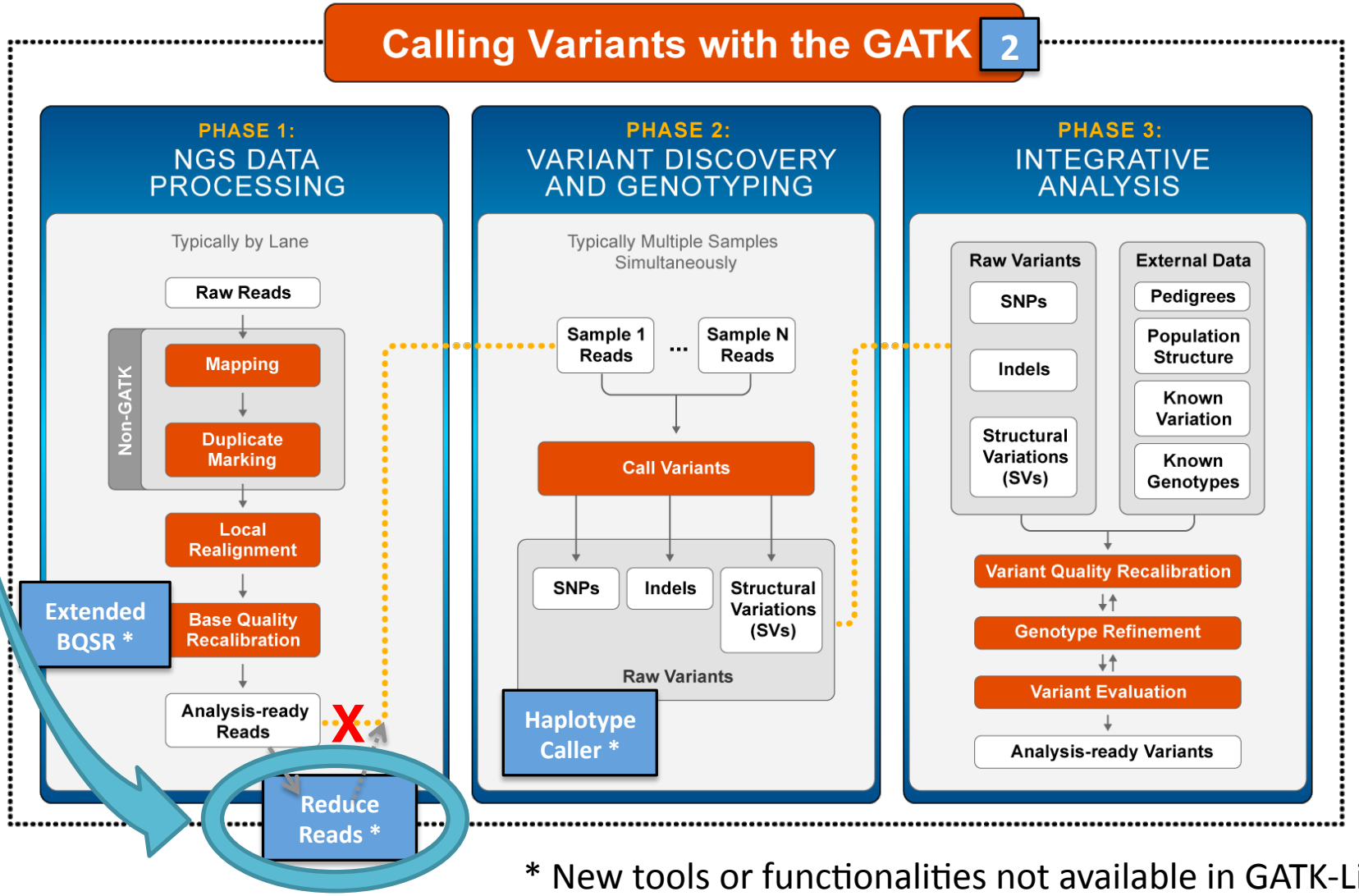


Did the compression work properly?

- Reads should be stripped out of all extra tags in the BAM file.
- A quick variant calling run on a small region of the genome (such as chr20:10,000,000-20,000,000) on both full and reduced BAM and look for highly similar variant calls. If numbers are too disparate (either compressed BAM is missing variants or is carrying many new variants) a more cursory look at the file is advised.
- Coverage test with DiagnoseTargets should yield similar results for variant regions and capped results for consensus regions.

We were here in the Best Practices workflow

NEXT STEP: CALLING VARIANTS



* 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

http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_compression_reducereads_ReduceReads.html