University of Zurich[UZH]

URPP Evolution

SIB Swiss Institute of Bioinformatics

URPP Tutorial NGS

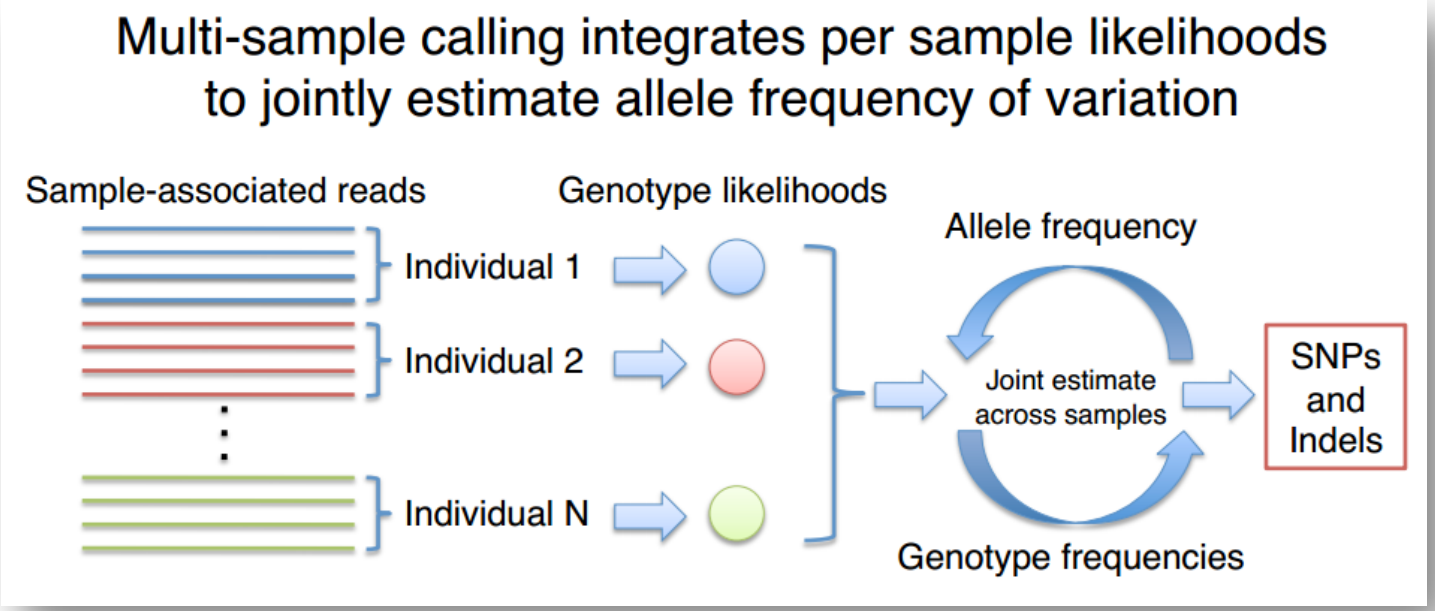
Variant Calling Part 3

Dr. Heidi E.L. Lischer
University of Zurich
Switzerland

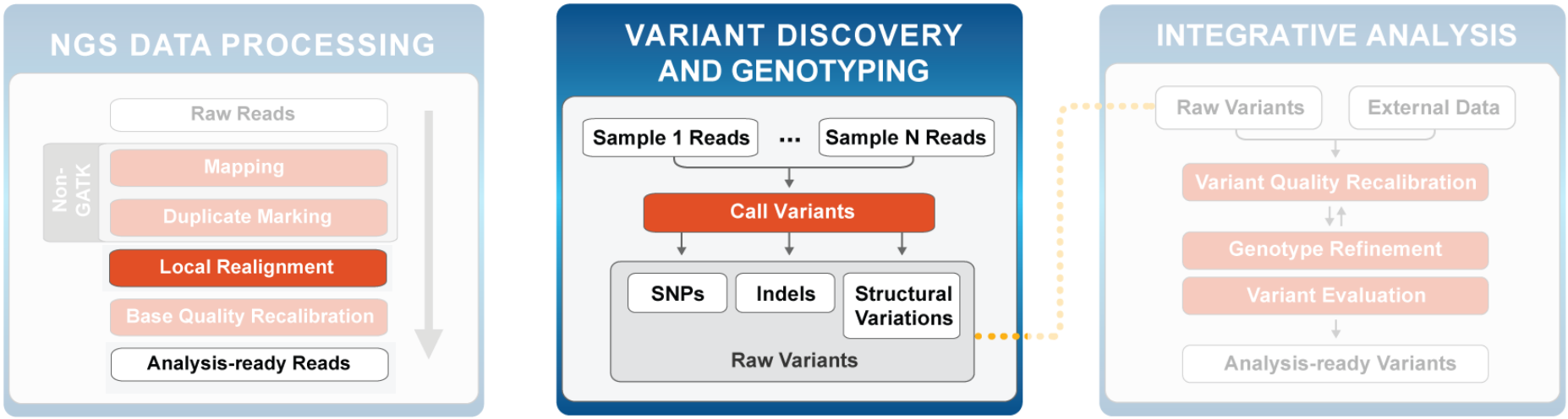
23 April 2014

Variant calling with GATK

- originally developed for human genetics
- handles genome data from any organism, with any level of ploidy
- Gain power to detect SNPs by multi-sample SNP calling



GATK pipeline

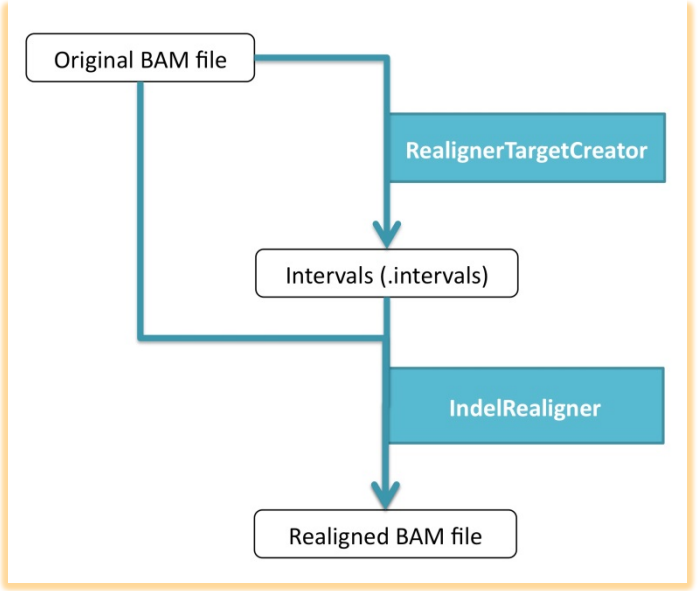


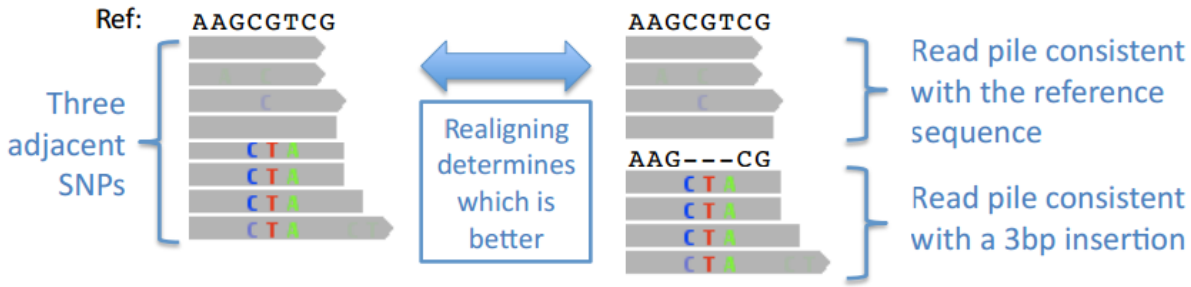
http://www.broadinstitute.org/gatk

Local realignment

- Mapping algorithms tend to produce various types of artifacts
  - mismatching bases around indels
- **Realignments:** identify the most consistent placement of the reads relative to the indels

GATK workflow:



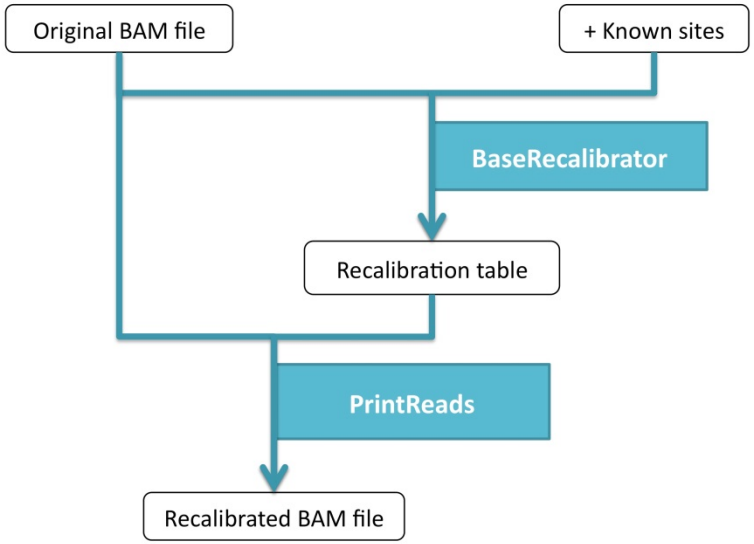


Base recalibration

- Base quality scores are per-base estimates of error emitted by the sequencing machines
  → various sources of systematic error
  → over- or under-estimated base quality

- **Base quality score recalibration:** apply machine learning to model errors empirically and adjust the quality scores

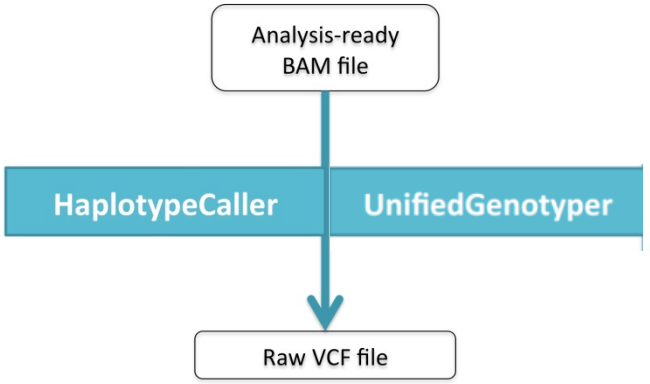
GATK workflow:



Variant calling

- Some observed variation are caused by mapping and sequencing artifacts
  → challenge: balance the need for sensitivity (to minimize false negatives) vs. specificity (to minimize false positives)

GATK workflow:



- GATK includes two variant calling tools:

UnifiedGenotyper

- Calls SNPs and indels separately
- Each variant locus independently

- Any ploidy
- Pooled calling
- High sample numbers

HaplotypeCaller → new (may replace UG)!

- Calls SNPs, indels and some SVs simultaneously
- Performing a local de-novo assembly

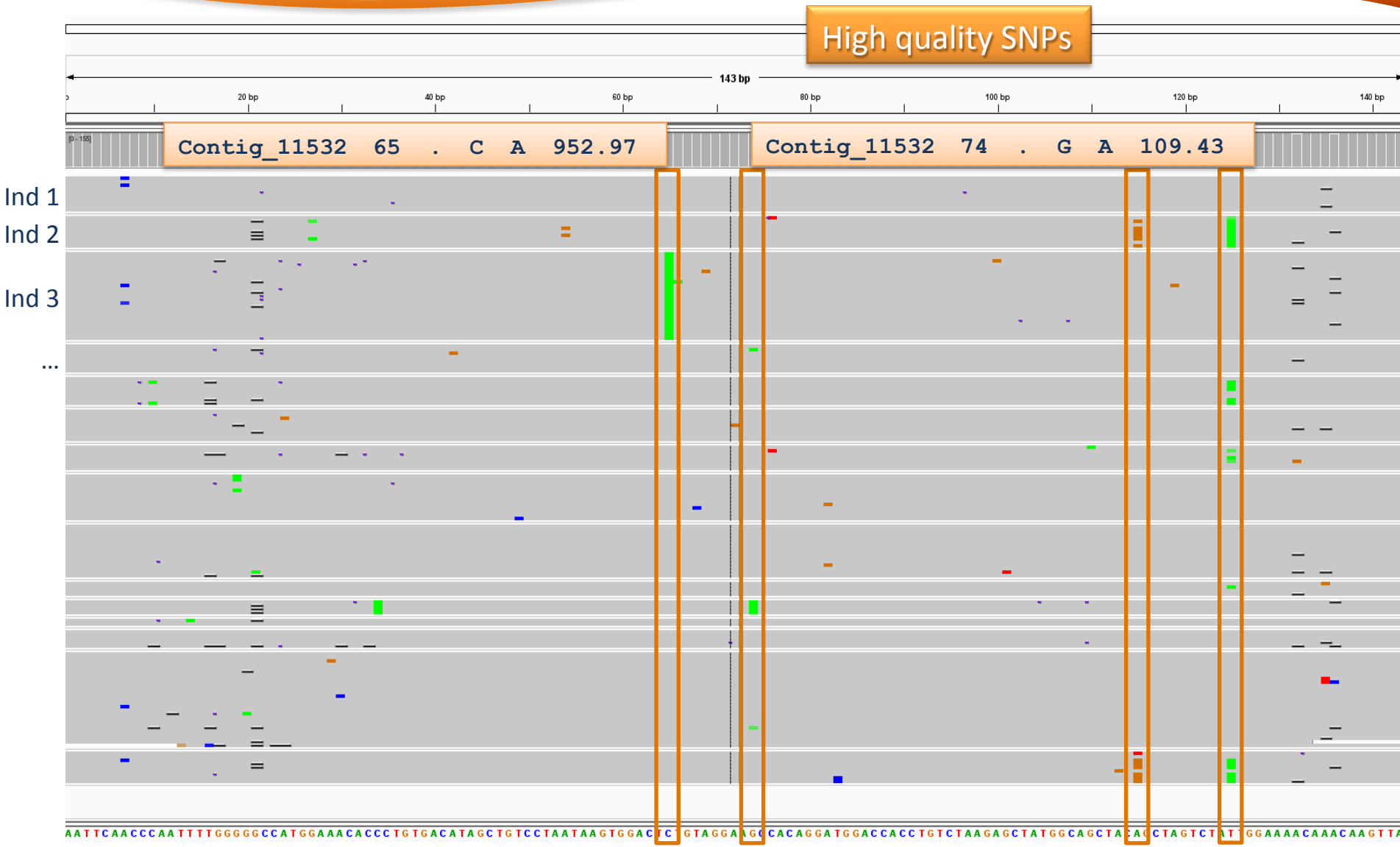
- More accurate (especially for indels)
- Computationally more intensive
- Handles only diploids

Variant filtering

- VariantFiltration:
  Filters variant calls using a number of user-selectable and parameterizable criteria
  - Read depth
  - SNP calling quality
  - Genotype quality
  - …

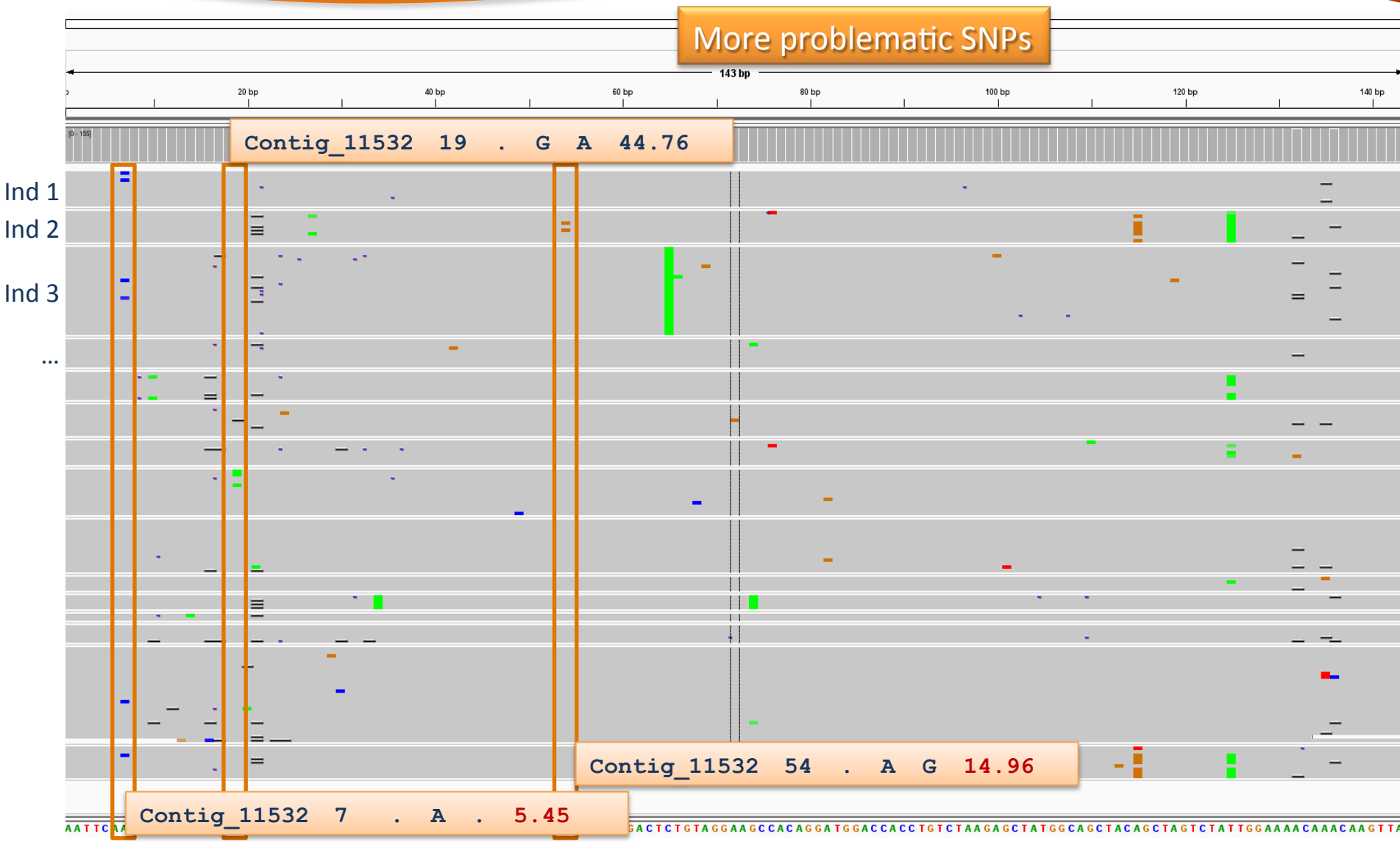
Example –

SNP calling across 15 ind at relative low coverage



Example –

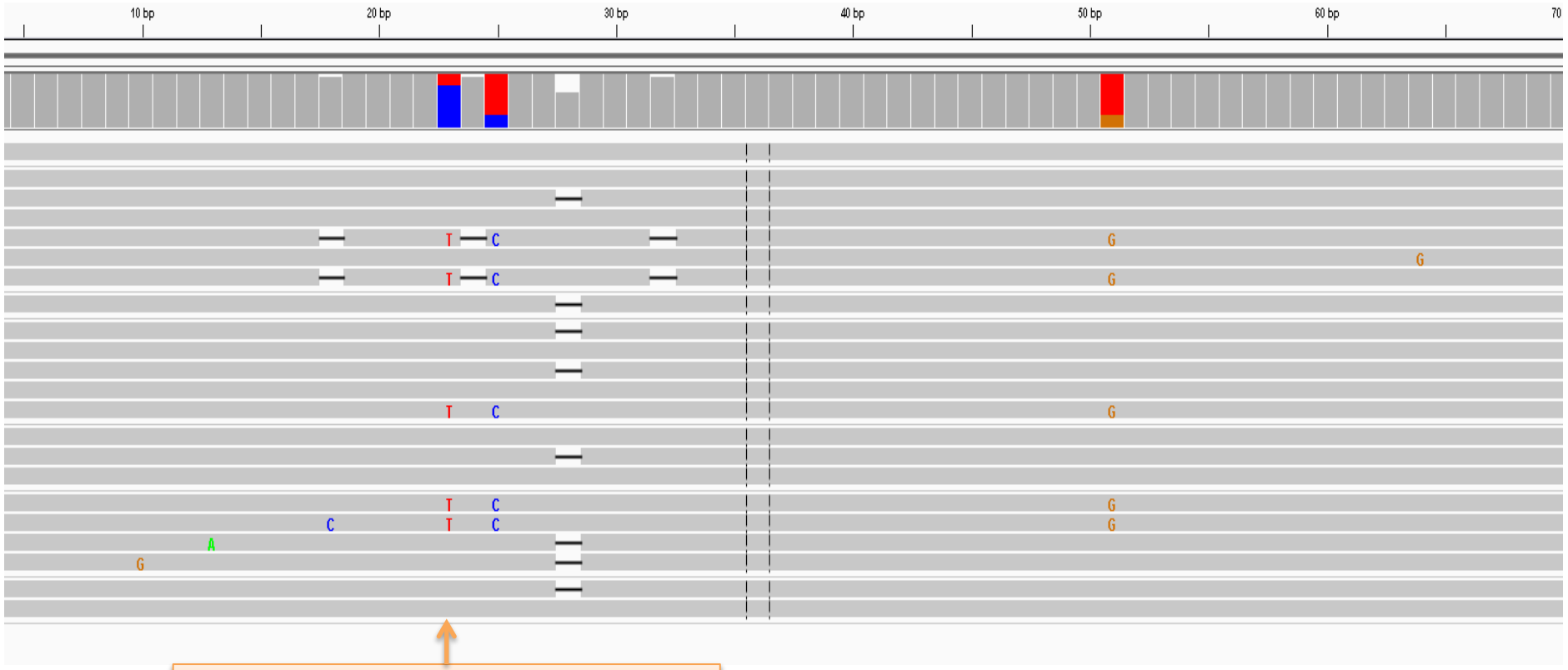
SNP calling across 15 ind at relative low coverage



Example –

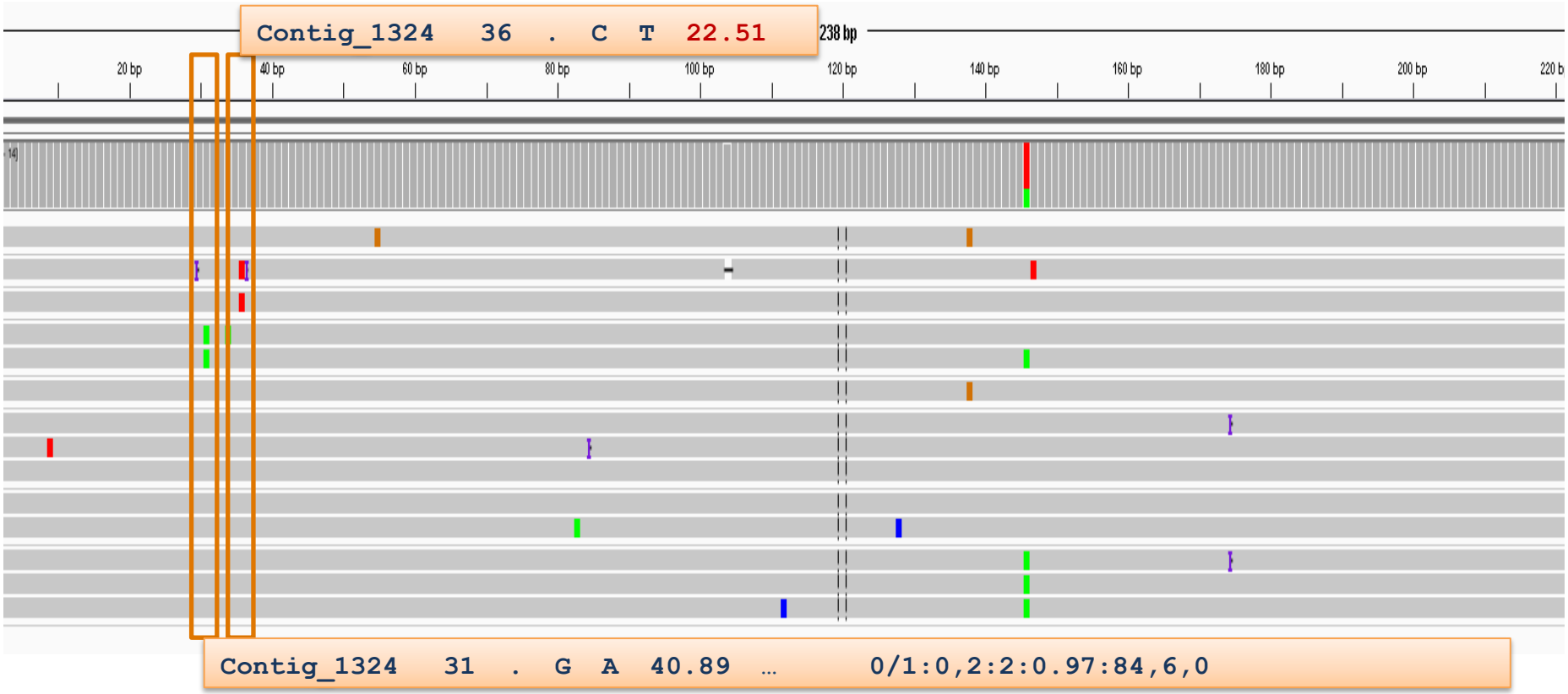
SNP calling across 15 ind at relative low coverage

- Information from other samples increase the power to call SNPs



Example –
SNP calling across 15 ind at relative low coverage

- Higher weight on SNPs within sample than between



→ Bias towards reference!
(real likelihood of 0/1 is only 10^{-6})