



Genomics in the Cloud

Book Club - Week 6

January 4, 2021

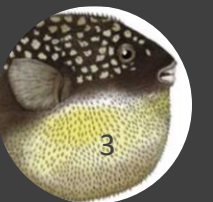
Agenda

- Chapter 5: First Steps with GATK
- Additional resources
- Open discussion



Our guest speaker

Dr. Mikhael
Manurung



Chapter 5: First Steps in GATK

Genomics in the Cloud by Geraldine A. Van der Auwera and Brian D. O'Connor (O'Reilly). Copyright 2020 The Broad Institute, Inc. and Brian O'Connor, 978-1-491-97519-0.



- Genome Analysis Toolkit
- Collection of tools with a primary focus on variant discovery
- <https://gatk.broadinstitute.org/hc/en-us>



Calling GATK

```
# gatk
```

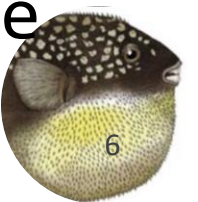
```
# gatk --list
```

Coverage Analysis:	Tools that count coverage, e.g. depth per allele
ASEReadCounter	Generates table of filtered base counts at het sites for allele specific expression
AnalyzeSaturationMutagenesis	(BETA Tool) (EXPERIMENTAL) Processes reads from a MITESeq or other saturation mutagenesis experiment.
CollectAllelicCounts	Collects reference and alternate allele counts at specified sites
CollectAllelicCountsSpark	Collects reference and alternate allele counts at specified sites
CollectF1R2Counts	Collect F1R2 read counts for the Mutect2 orientation bias mixture model filter
CollectReadCounts	Collects read counts at specified intervals
CountBases	Count bases in a SAM/BAM/CRAM file
CountBasesSpark	Counts bases in the input SAM/BAM
CountReads	Count reads in a SAM/BAM/CRAM file
CountReadsSpark	Counts reads in the input SAM/BAM
GetPileupSummaries	(BETA Tool) Tabulates pileup metrics for inferring contamination
Pileup	Prints read alignments in samtools pileup format
PileupSpark	(BETA Tool) Prints read alignments in samtools pileup format

```
# gatk ToolName [tool arguments]
```

```
# gatk ToolName --help
```

```
# gatk --java-options "insert-options-here" ToolName  
[tool arguments]
```



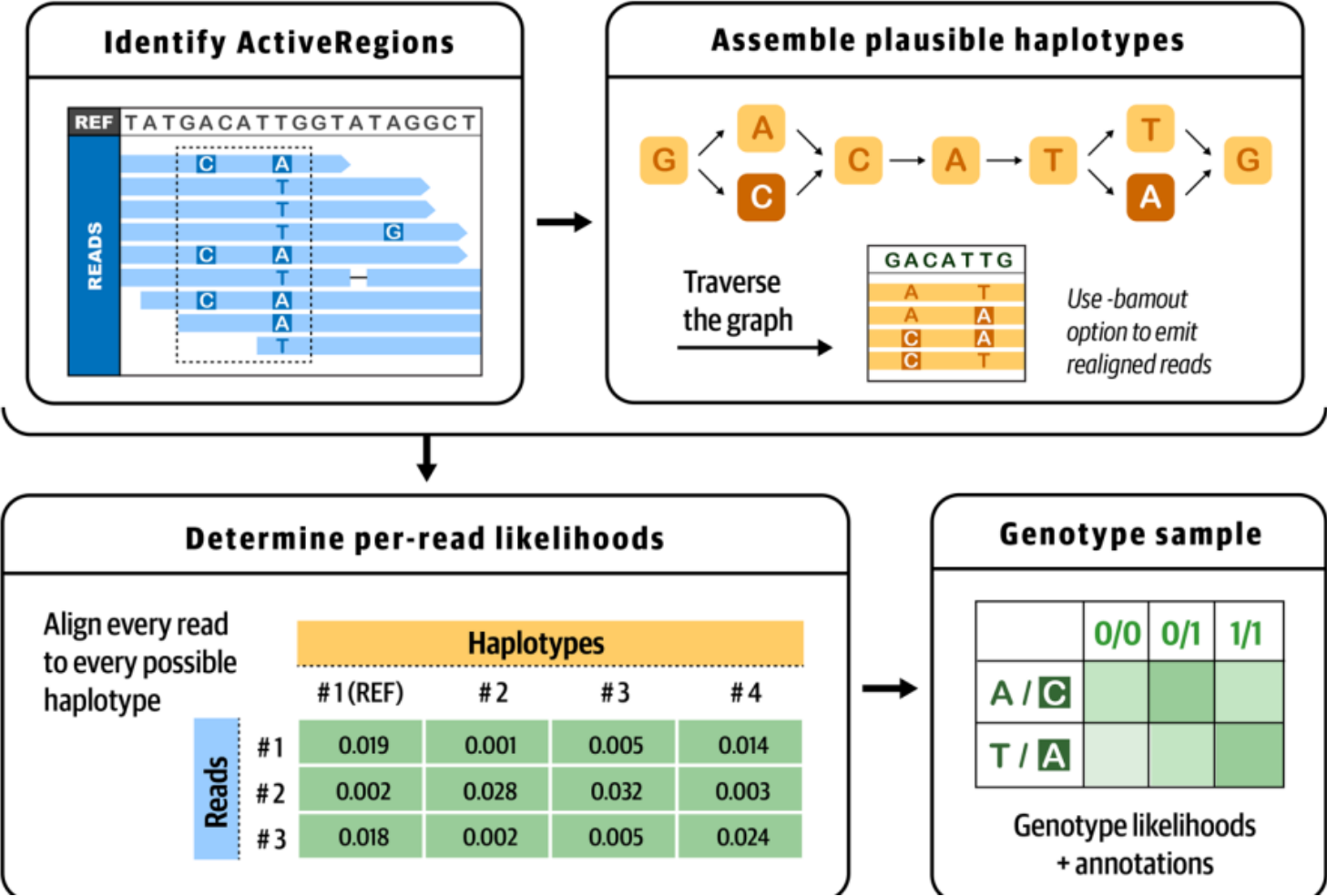
Running your first GATK command

```
# cd /home/book/data/germline
# mkdir sandbox
# gatk HaplotypeCaller \
    -R ref/ref.fasta \
    -I bams/mother.bam \
    -O sandbox/mother_variants.vcf
```

- Move to `home/book/data/germline` folder
- Create `sandbox` directory
- Run `HaplotypeCaller` with the following **required** arguments:
 - -R: reference
 - -I: input file
 - -O: output file

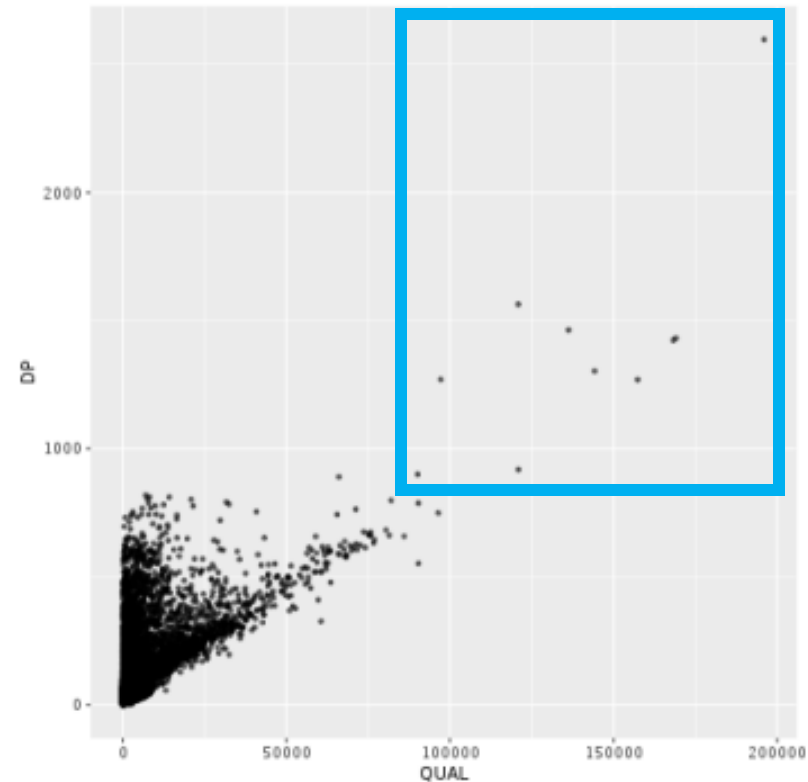
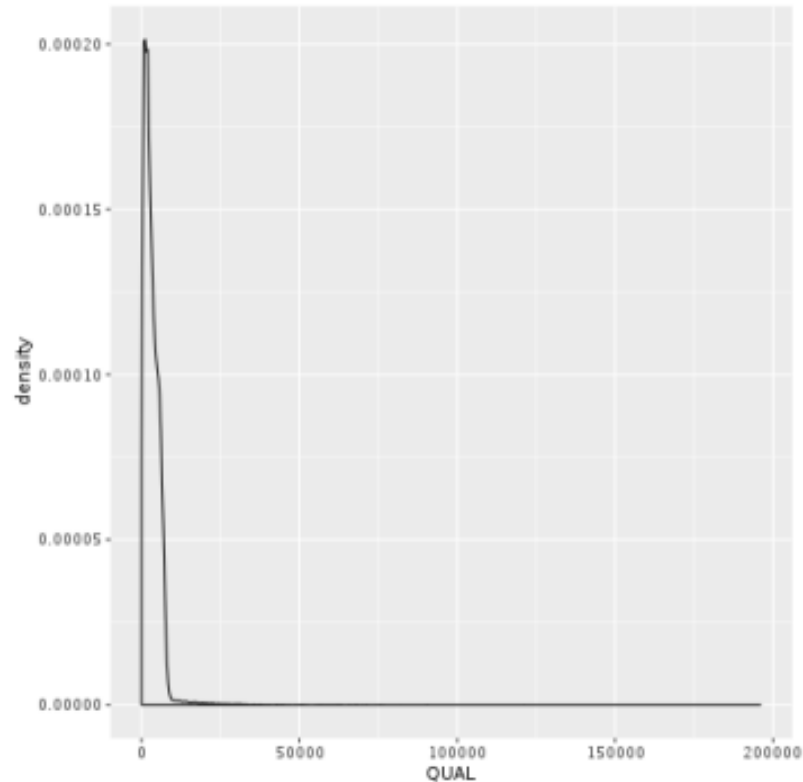


Calling variants with HaplotypeCaller



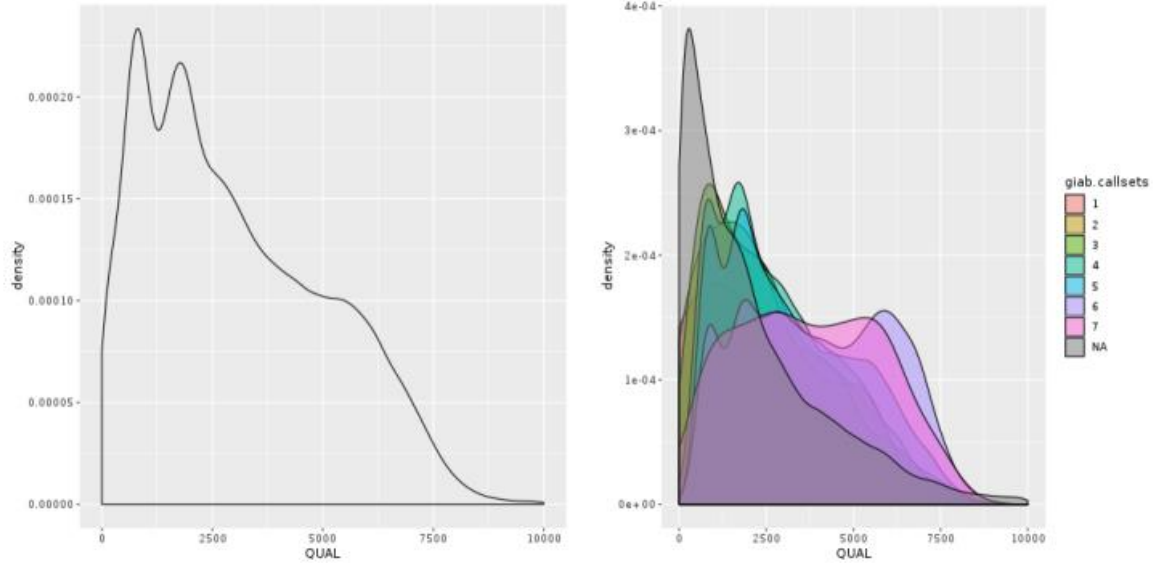
Quality Control

- Quality is confounded by high depth of coverage

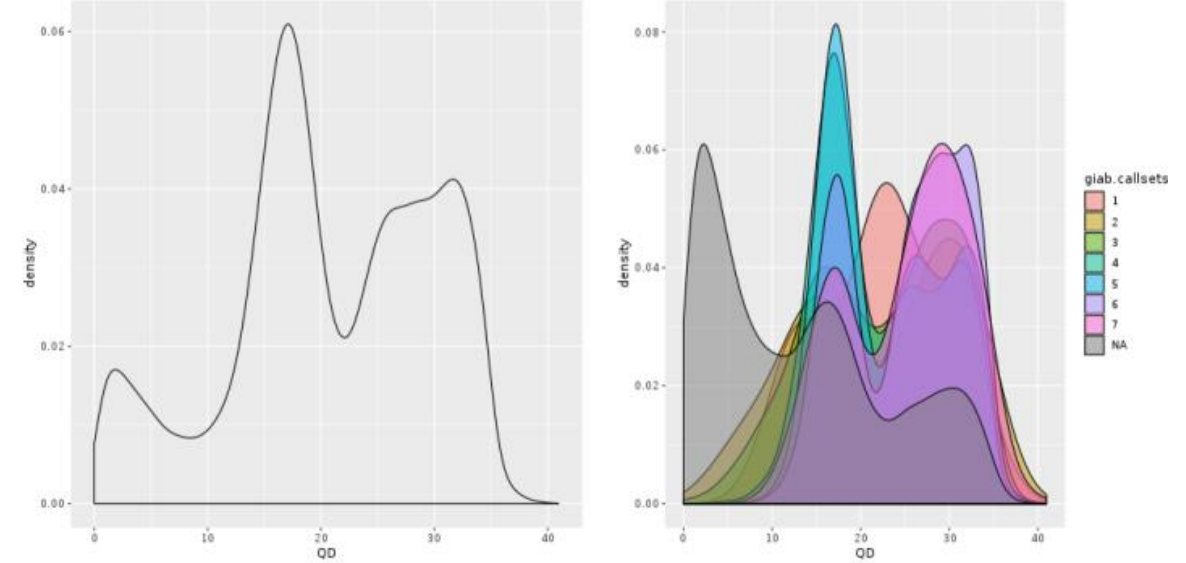


Looking at the truth set...

Quality



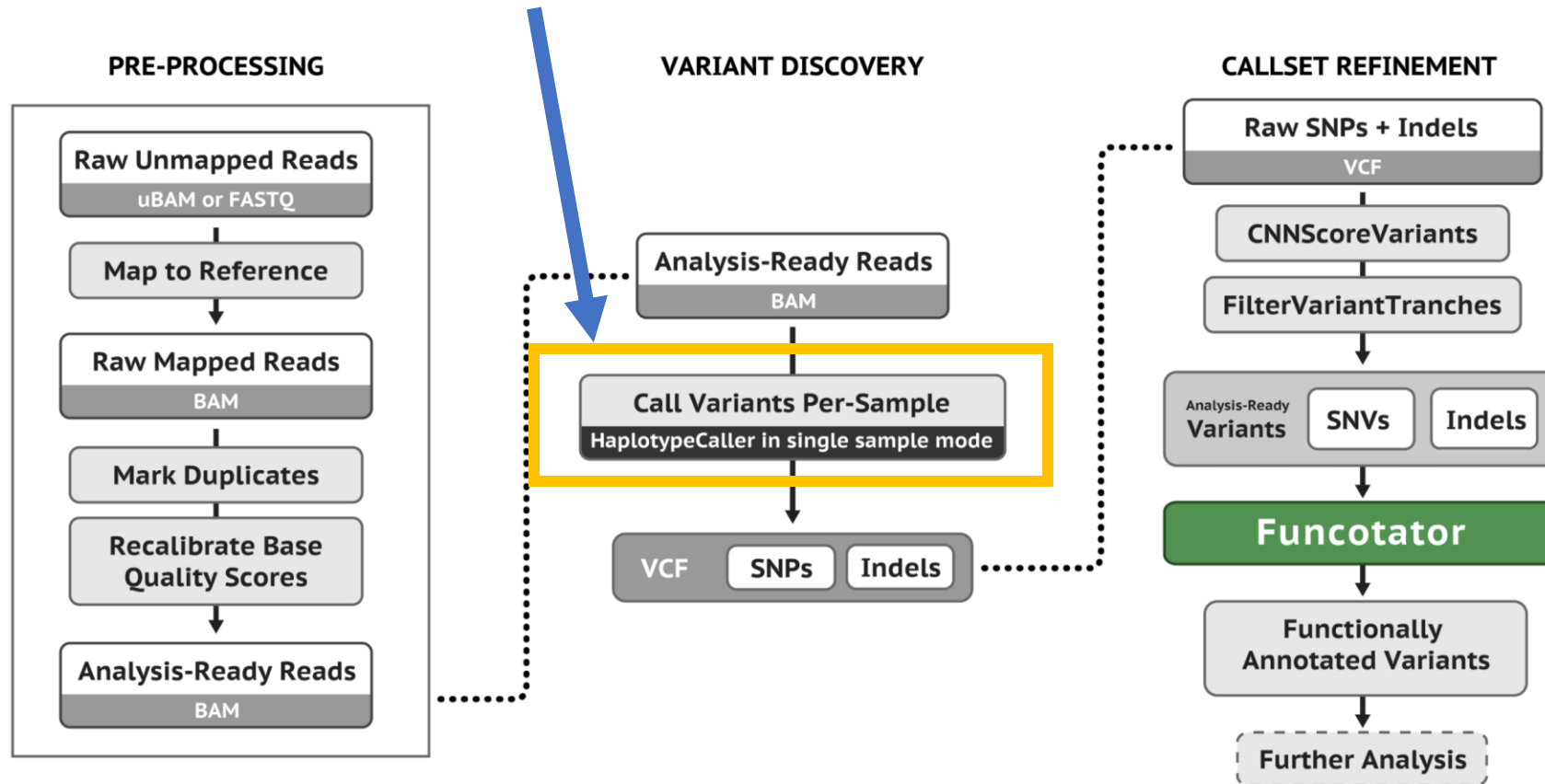
QualByDepth



Normalizing quality by amount of coverage available




Single-sample germline SNPs/indels workflow

We are here



GATK Best Practices



	Germline	Somatic
SNPs & Indels	 HaplotypeCaller / Joint Calling	 MuTect2 / Tumor-Normal
Copy number	GATK gCNV	 GATK CNV + aCNV
Structural Variation	GATK SVDDiscovery (beta)	<i>on the roadmap</i>



Additional Resources

- GATK Workshop @ University of Costa Rica
 - https://youtu.be/0DoS_m2iyKU (4-day workshop--look within the channel for videos from the other days)
- Artificial Haplotypes
 - https://www.youtube.com/watch?v=vocBk2MHP1A&index=7&ab_channel=BroadInstitute
- Genome in a Bottle (reference genomes)
 - <https://www.nist.gov/programs-projects/genome-bottle>
- GATK Best Practices Pipeline Index
 - <https://gatk.broadinstitute.org/hc/en-us/articles/360035894751-Pipeline-Index>
- WDL Analysis Research Pipelines (WARP)
 - <https://support.terra.bio/hc/en-us/articles/360050981492-Introducing-WARP-A-collection-of-cloud-optimized-workflows-for-biological-data-processing-and-analysis>





Thank you for joining us today!

Next week: Chapter 6

Next meeting: January 11, 2021