Variant Calling Using GATK (Genome Analysis Toolkit)

Architecture of large projects in bioinformatics

Konstanty Kraszewski

University of Warsaw

April 9, 2024

Presentation outline

1. Introduction
2. Definitions
3. Genome Analysis Toolkit
4. Variant Calling
   - HaplotypeCaller
5. Conclusion

Presentation outline

1. Introduction
2. Definitions
3. Genome Analysis Toolkit
4. Variant Calling
   - HaplotypeCaller
5. Conclusion

Introduction

- Massive data sets make developing analysis tools challenging.
- Many professionals are limited due to the complexity of accessing and manipulating NGS data.
- GATK optimizes for correctness, stability, CPU and memory efficiency, and supports distributed and shared memory parallelization.

Presentation outline

1. Introduction
2. Definitions
3. Genome Analysis Toolkit
4. Variant Calling
    - HaplotypeCaller
5. Conclusion

Definitions

- **Variant**: A variation in the DNA sequence compared to a reference genome.
- **Variant Calling**: The process of identifying and characterizing genetic variants from next-generation sequencing data.
- **SNP (Single Nucleotide Polymorphism)**: A variation at a single position in a DNA sequence among individuals, where each variation is present to some appreciable degree within a population.

Presentation outline

1. Introduction
2. Definitions
3. Genome Analysis Toolkit
4. Variant Calling
   - HaplotypeCaller
5. Conclusion

What is GATK?

- Open-source programming framework.
- Developed by the Broad Institute.
- Widely used in bioinformatics for variant discovery and genotyping.
- Provides a suite of tools for processing high-throughput sequencing data.

GATK technical details

GitHub repository at github.com/broadinstitute/gatk/.

GATK technical details

GitHub repository at github.com/broadinstitute/gatk/.

Requirements for running GATK:

- Java 17
- Python 2.6 or greater (frontend)
- Python 3.6.2, with additional Python packages (some tools and workflows)
- R 3.2.5 (plots in some tools)

All are found in prepared Docker images.

Licensing

License at project website and dockerhub repository:

- BSD 3-clause

Licensing

License at project website and dockerhub repository:

- BSD 3-clause

License at project repository:

- Apache 2.0

How is GATK Used?

- Identifying single nucleotide polymorphisms (SNPs)
- Detecting insertions and deletions (indels)
- Calling germline and somatic mutations
- Filtering variants based on quality metrics
- Somatic short variant calling
- Copy number (CNV) and structural variation (SV)
- Other tasks like processing and quality control

Presentation outline

1. Introduction
2. Definitions
3. Genome Analysis Toolkit
4. Variant Calling
   - HaplotypeCaller
5. Conclusion

Workflow

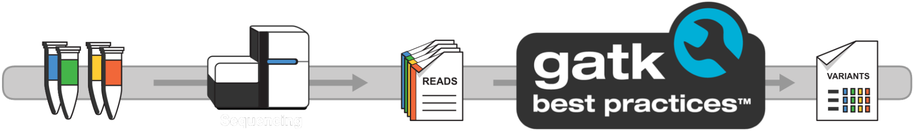


Figure: General workflow for finding variants.

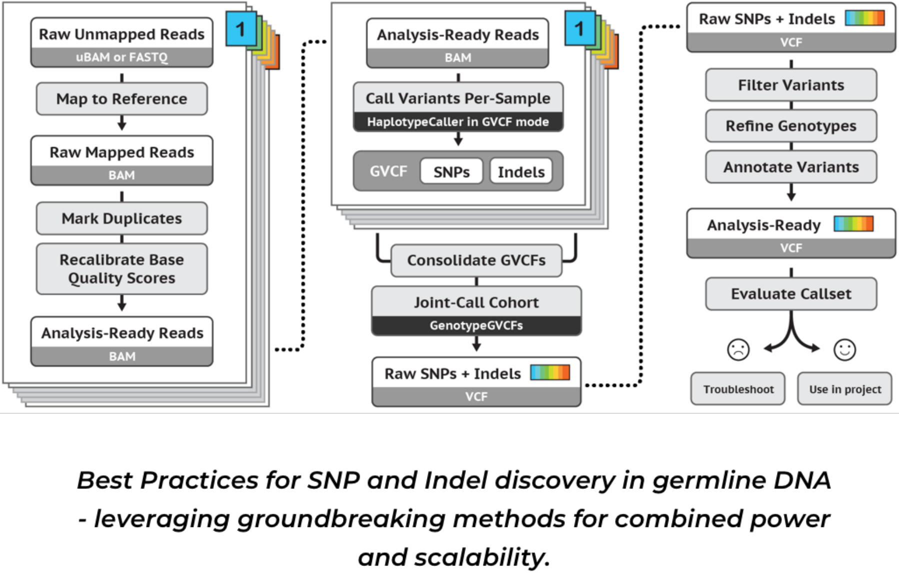


Figure: GATK Best Practices workflow.

HaplotypeCaller algorithm

1. **Define active regions.** Finds regions with the most variation.

HaplotypeCaller algorithm

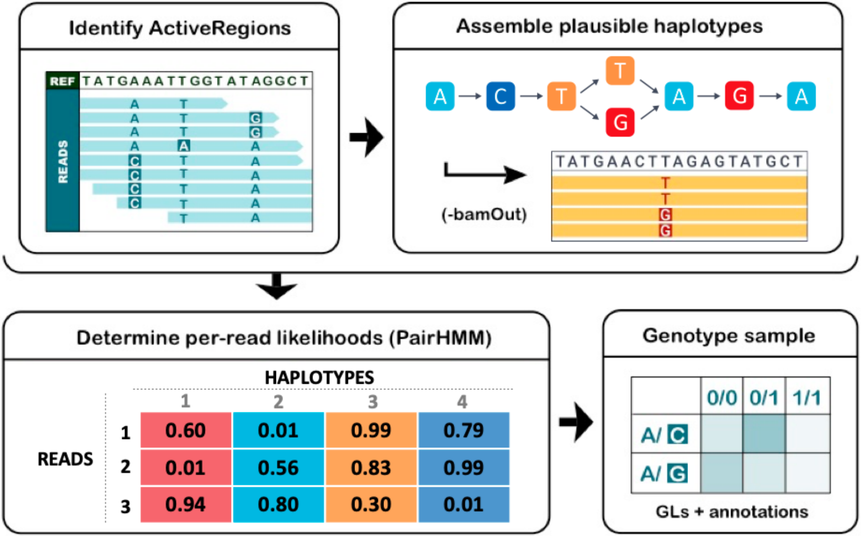
1. **Define active regions.** Finds regions with the most variation.
2. **Determine haplotypes by re-assembly of the active region.** Builds possible sequences *de novo* from the input reads for each active region.

HaplotypeCaller algorithm

1. **Define active regions.** Finds regions with the most variation.
2. **Determine haplotypes by re-assembly of the active region.** Builds possible sequences *de novo* from the input reads for each active region.
3. **Determine likelihoods of the haplotypes given the read data.** Computes how much evidence is there in the reads for each haplotype.

HaplotypeCaller algorithm

1. **Define active regions.** Finds regions with the most variation.
2. **Determine haplotypes by re-assembly of the active region.** Builds possible sequences *de novo* from the input reads for each active region.
3. **Determine likelihoods of the haplotypes given the read data.** Computes how much evidence is there in the reads for each haplotype.
4. **Assign sample genotypes.** Computes likelihoods for each genotype.



		HAPLOTYPES			
		1	2	3	4
READS	1	0.60	0.01	0.99	0.79
	2	0.01	0.56	0.83	0.99
	3	0.94	0.80	0.30	0.01

	0/0	0/1	1/1
A/ C			
A/ G			

Figure: HaplotypeCaller workflow.

Presentation outline

1. Introduction
2. Definitions
3. Genome Analysis Toolkit
4. Variant Calling
   - HaplotypeCaller
5. Conclusion

Conclusion

GATK plays a crucial role in variant calling and genomic analysis. Its robust algorithms and tools enable accurate identification of genetic variants from next-generation sequencing data, facilitating various research and clinical applications.

GATK has been incorporated into large-scale sequencing projects like the 1000 Genomes Project and The Cancer Genome Atlas, highlighting its importance.

Thank you for your attention.

Sources

- https://gatk.broadinstitute.org/
- McKenna et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. Epub 2010 Jul 19. PMID: 20644199; PMCID: PMC2928508
- https://github.com/broadinstitute/gatk/