Disclosure for:

ASHG Interactive Workshop:
Variant Discovery with GATK 4

We are collaborating with Google to set up GATK-as-a-service on their commercial cloud platform.
Variant Discovery with GATK 4

ASHG 2016 Interactive Workshop
Vancouver, CA 18 October, 2016

Geraldine Van der Auwera
Soo Hee Lee
GATK = Genome Analysis ToolKit

- Toolkit focused on **variant discovery**
- Developed at Broad Institute
- [http://www.broadinstitute.org/gatk/](http://www.broadinstitute.org/gatk/)

- Supported on GATK forum at [http://gatkforums.broadinstitute.org/gatk/categories/ask-the-team](http://gatkforums.broadinstitute.org/gatk/categories/ask-the-team)

- Freely available to academics for non-profit work
- Commercial license for for-profit applications
GATK Best Practices for Variant Discovery

1. Data Pre-Processing
   - FASTQ -> BAM

2. Variant Discovery
   - BAM -> VCF

3. Callset Refinement
1. GATK 4 - Features & development roadmap
2. GATK on the Cloud - Live demos
3. CNV analysis with GATK 4 - Hands-on tutorial
Part 1

GATK 4
GATK Development Roadmap

GATK 4 Alpha: Completely rewritten engine
Support for Apache Spark
+ Extended functionality: CNVs, SVs, Picard integration
Different types of variants – all relative to reference genome
Variant discovery workflows by variant type

**Germline SNPs & Indels**
- Analysis-Ready Reads
  - Var. Calling
    - HC in ERC mode
  - Genotype Likelihoods
  - Joint Genotyping
  - Raw Variants
    - SNPs
    - Indels
  - Variant Recalibration
    - separately per variant type
  - Analysis-Ready Variants
    - SNPs
    - Indels

**Somatic SNVs & Indels**
- T/N Pair Reads
  - T
  - N
  - Panel of Normals
  - ContEst
  - COSMIC
  - Cross-sample Contamination
  - dbSNP
  - SNV & Indel Calling
    - MuTect2
  - Filter Variants
  - Analysis-Ready Variants
    - SNVs
    - Indels

**Somatic CNVs**
- Reads
  - Targets
  - PoN
  - Collect proportional coverage
  - Proportional coverage profile
  - Normalize
  - Segment
  - Plots
  - Segment coverage
  - Called segments
**GATK Best Practices**

Recommended workflows for variant discovery analysis with GATK

**Reads-to-variants workflows used at the Broad Institute.**

The GATK Best Practices provide step-by-step recommendations for performing variant discovery analysis in high-throughput sequencing (HTS) data. There are several different Best Practices workflows tailored to particular applications depending on the type of variation of interest and the technology employed. The Best Practices documentation attempts to describe in detail the key principles of the processing and analysis steps required to go from raw reads coming off the sequencing machine, all the way to an appropriately filtered variant callset that can be used in downstream analyses. Wherever we can, we try to provide guidance regarding experimental design, quality control (QC) and pipeline implementation options, but please understand that these are dependent on many factors including sequencing technology and the hardware infrastructure that are at your disposal, so you may need to adapt our recommendations to your specific situation.

### GERMLINE

<table>
<thead>
<tr>
<th>SNPs &amp; Indels</th>
<th>COPY NUMBER</th>
<th>SNVS &amp; Indels</th>
<th>COPY NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXOME/PANEL + WGS</td>
<td>BWA + HaplotypeCaller GVCF</td>
<td>EXOME/PANEL + WGS</td>
<td>BWA + MuTect2 BETA</td>
</tr>
<tr>
<td>RNASeq</td>
<td>STAR + HaplotypeCaller</td>
<td>RNASeq</td>
<td>STAR + HaplotypeCaller</td>
</tr>
</tbody>
</table>

### SOMATIC

<table>
<thead>
<tr>
<th>SNVS &amp; Indels</th>
<th>COPY NUMBER</th>
<th>SNVS &amp; Indels</th>
<th>COPY NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXOME/PANEL + WGS</td>
<td>BWA + MuTect</td>
<td>EXOME/PANEL + WGS</td>
<td>BWA + CaliSegments</td>
</tr>
<tr>
<td>RNASeq</td>
<td>STAR + HaplotypeCaller</td>
<td>RNASeq</td>
<td>STAR + HaplotypeCaller</td>
</tr>
</tbody>
</table>

In development
GATK ON THE CLOUD

Part 2
What is the Cloud and how does it help?

- **Elasticity**
  scale throughput at will

- **Availability**
  no waiting in queues

- **Collaboration**
  data sharing + reproducibility

- **Transparency**
  of costs/billing
Our Hg38 bundle is on the Cloud!
Ways to run GATK on Cloud

- Directly from GATK4 with Spark
- Pipelining service: Google Genomics API (others to come)

COMMAND-LINE

WEB GUI / FRONT END

compute + storage

Broad’s Genome Analysis Workbench
(workshop Wed 10/19)
E.g. FireCloud platform
GATK 4 Spark demo

JUST 3 SIMPLE STEPS

• Put data in a **bucket**
• Create a **custom cluster**
• Run GATK 4 with **-sparkRunner**

Enjoy!

https://console.cloud.google.com
Google Genomics API demo

JUST 3 SIMPLE STEPS

• Put data in a **bucket**
• Write or get a **WDL script**
• Run **gcloud** command

Enjoy!

[link](https://cloud.google.com/genomics/v1alpha2/gatk)
Google Genomics API demo: The Pipeline

Single-sample workflow for WGS

**Input:** Raw reads (uBAM)

**Output:**
- (Analysis-ready BAM)
- CRAM for storage
- GVCF for joint genotyping

**Written in WDL**

Script shared at
https://github.com/broadinstitute/wdl/tree/develop/scripts/broad_pipelines
SOMATIC CNV (COPY NUMBER VARIANT) ANALYSIS WITH GATK 4
Somatic copy-number variation can be dramatic

- Spectral karyotyping paints each chromosome pair with a color
- Alterations can vary dramatically between cancers and within cancers
Copy number variants alter coverage

[Diagram showing copy number variants and control]
Coverage is variable across WES *targets* and *kits*

**WES bait-capture and library amplification add to variability.**

*WES: ICE*

*WES: NEXTERA*

*In comparison, WGS gives even coverage.*

*WGS*
Detect copy-number variation using proportional coverage

1. Count reads overlapping each target

2. Divide by total sample reads for proportional coverage

Copy ratio

Genomic loci
Remove noise* to reveal copy-number variation

- By tangent normalization (PCA)
- By target medians

**Raw copy ratios**

*Copy ratio*  

*Genomic loci*
Overview of the targeted somatic CNV case-sample workflow

**Inputs:**
- Tumor BAM
- Panel of normals (PoN) created from BAMs
- Genomic targets, e.g. exome

**Outputs:**
- Denoised copy-ratio profile
- Segmented somatic CNV calls
- Plots and QC metrics
What is different between the plots? Look closely.

The 4th CNV event is also called by GISTIC analysis of SNP6 array data.
Upcoming tool: *Allelic* Copy Number Variation (ACNV)

- Uses matched-normal and CNV results
- Uses heterozygous SNP sites identified in normal to estimate allele fraction in tumor
- Updates segmented copy-ratio estimates
- Detects the following:
  - minor-allele fraction, e.g. 1/3
  - copy-neutral loss of heterozygosity (LoH)
Join Broad Genomics at other ASHG events

Experts @ Booth 329

Jeff Gentry
Expertise: Cloud-based Analysis Execution
Wednesday, October 19th
11:00am - 12:00pm

Heidi Rehm, Ph.D.
Expertise: Clinical Research + Med & Pop Genetics
Wednesday, October 19th
3:00pm - 4:00pm

Geraldine Van der Auwera, Ph.D.
Expertise: Genome Analysis Toolkit (GATK)
Thursday, October 20th
10:00am - 11:00am

Daniel MacArthur, Ph.D.
Expertise: ExAc and Medical & Population Genetics
Thursday, October 20th
3:00pm - 4:00pm

Alexander Baumann
Expertise: Broad Institute Workbench Analysis Tool
Friday, October 21st
1:00pm - 2:00pm

Workshops

Introducing Broad Institute Workbench
A Cloud-Based Platform for Data Analysis, Management, and Sharing
Wednesday, October 19th
1:00pm - 2:30pm
Room 16, East Building
Vancouver Convention Centre

Camp ASHG

Wednesday, October 19th
6:30pm - 8:30pm
Exhibit Hall A, East Building
Vancouver Convention Centre

Posters

RNASeq 2.0 @ Broad Genomics: Advances in RNA Sequencing Analysis
Poster Number: 3226

Scaling Variant Calling to Hundreds of Thousands of Samples with GATK
Poster Number: 1953

Detection and Depletion of Bacteria Contamination in Saliva Derived DNA Samples for Human Whole Genome Sequencing
Poster Number: 3218

A Sequencing Deliverable: Theoretical Sensitivity to Heterozygous SNPs
Poster Number: 1819