Introduction to Variant Discovery with GATK Best Practices

All you need to know about high-throughput sequencing, data formats and analysis methods
OVERVIEW

SCOPE & TERMINOLOGY
Variant discovery = identify **variants** in sequencing data
What do we mean by variant?

- Genetic changes in individuals relative to a reference genome
  - Germline (inherited)
  - Somatic (cancer)

- Reference genome = a standardized genomic sequence

- Human genome reference sequence
  - Current standard: hg19 / b37
  - New standard (not yet widely adopted): hg38

- Other organisms
  - Some have a fully assembled reference available
  - Most do not
Different types of variants

- SNP/SNV
- Indel
- CNV/CNA
- SV
GATK Best Practices = complete reads-to-variants workflows

Data Pre-Processing

FASTQ -> BAM

Variant Discovery

BAM -> VCF

Callset Refinement

SAM/BAM

VCF

FASTQ
Informational content of the important file formats

- **FASTQ**
  - Enormous pile of short reads
  - Sequence + base qualities

- **SAM/BAM**
  - Reference genome
  - Mapped reads
  - Alignment map

- **VCF**
  - site 1 description + sample genotypes
  - site 2 description + sample genotypes
  - site 3 description + sample genotypes

Site-by-site genomic variation
EXPERIMENTAL DESIGNS
WGS VS. EXOME
Whole genome (WGS) vs. Exome (WEx)

**Whole genome**

Intergenic  Exon I  Intron I  Variant site  Exon II  Intergenic

**Exome**

Intergenic  Exon I  Intron I  Variant site  Exon II  Intergenic

Small targeted experiments, gene panels, RADseq

• Similar to exomes for most purposes
Evenness of coverage is very different

- To achieve 80% at 20X target for exomes, many bases are covered up to 100X
- WGS achieves mean coverage >20X with fewer high coverage bases
Exome capture relies on target regions -> kit-specific

- Involves **baits** (complementary sequences) tilled to capture segments of target regions

(Broad uses a custom-designed bait set; the target intervals list is available in our resource bundle)

- Resulting covered intervals are specific to capture kit manufacturer

(Clar et al, Nat. Biotech., 2011)
Exome analysis: restrict analyses to capture intervals!

- Obtain the appropriate interval list from the capture kit manufacturer or sequence provider

  - Use `--L` argument to restrict analyses at key steps*

```java
java -jar GenomeAnalysisTK.jar [tool and other arguments] \n   -L intervals.list \n   --interval_padding 50
```

* RealignerTargetCreator, BaseRecalibrator, HaplotypeCaller
What about gene panels, RADseq, ...?

- **Targeted gene panels**
  - Functionally similar to exome with very small interval list

- **RADseq**
  - Uses *Restriction Analysis Digest* i.e. chop up DNA at frequent restriction enzyme sites
  - Can be done without an assembled reference

  “stacks” of sequence

- **Same drawbacks as exomes PLUS some new ones**
  - Produces many independent fragments with identical start and end, so MarkDuplicates should not be run
  - Too few variants for VQSR, must use hard-filtering
GENERATING THE INPUT DATA
LABWORK & SEQUENCING
1) Extract nucleic acids from blood, tissue, saliva
2) RNA: Make cDNA
3) Shear dsDNA into fragments
4) Attach fragments to adapters (-> library)

Library prep
Sequencing the library

HTS machine processes a **flowcell** containing up to 8 **lanes**; each lane constitutes a **read group (RG)** (unless multiplexed).

Enormous pile of short reads
Most common format is FASTQ = FASTA + base quality scores

• Extension of FASTA format
• Four data elements:
  • Sequence Name (read name, group, etc.)
  • Sequence
  • + (optional: Sequence name again)
  • Associated quality score.
• Example record:
  • @EAS54_6_R1_2_1_413_324
  • CCCTTCTTGTCTTCAGCGTTTCTCC
  • +
  • ;;;3;;;;;;;;;;;;;;;;7;;;;;;;;;;;;;;;;88

A FASTQ file typically contains data from a single lane of sequence

-> ASCII code translates to Phred-scale Q scores

Official specification in http://maq.sourceforge.net/fastq.shtml
Phred scaling makes it easier to handle probability scores

- **Phred value** = \(-10 \times \log_{10}(\varepsilon)\)

- Examples:
  - 90% confidence (10% error rate) = Q10
  - 99% confidence (1% error rate) = Q20
  - 99.9% confidence (.1% error rate) = Q30

- SAM encoding adds 33 to the value (because ASCII 33 is the first visible character)

![Accuracy vs Error Graph](image)
DATA PRE-PROCESSING

MAP + CORRECT FOR TECHNICAL ERRORS
Step 1: Map the reads produced by the sequencer to the reference

Enormous pile of short reads from HTS

Mapping and alignment algorithms

• BWA for DNA
• STAR for RNAseq

Reference genome

• Reads mapped to reference

Data Pre-processing

1 Raw Reads

Non-GATK

• Map to Reference
  - BWA mem

• Mark Duplicates & Sort (Picard)

• Indel Realignment

• Base Recalibration

• Analysis-Ready Reads
A BAM file can contain data from a single or from several samples

Mapping produces a SAM alignment summarizing **position**, **quality**, and **structure** for each read

<table>
<thead>
<tr>
<th>POS (alignment start)</th>
<th>MAPQ (quality)</th>
<th>CIGAR (structure)</th>
<th>SEQ (sequence)</th>
</tr>
</thead>
<tbody>
<tr>
<td>read1 99 ref 2 30 3M1D2M1I1M = 14 20 CATCTAG *</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mate information

RefPos: 1 2 3 4 5 6 7 8 9
Reference: C C A T A C T - G A
Read: C A T - C T A G
POS: 2
CIGAR: 3M1D2M1I1M

See also:
How we visualize aligned HTS data (Integrated Genomics Viewer)

Non-reference bases are colored; reference bases are grey

Clean C/T heterozygote

Depth of coverage

Individual reads aligned to the genome

Reference sequence
SPECIAL HANDLING FOR RNASEQ
Key differences for mapping & processing RNAseq

DNAseq

1. Map to Ref (BWA) & Add Read Groups
2. Mark Duplicates & Sort (Picard)
3. Indel Realignment
4. Base Recalibration
5. Analysis-Ready Reads

RNAseq

1. Map to Ref (STAR) & Add Read Groups
2. Mark Duplicates & Sort (Picard)
3. Split’N’Trim + Reassign Mapping Quality
4. Indel Realignment
5. Base Recalibration
6. Analysis-Ready RNAseq Reads

Different mapper
Extra processing step
RNAseq reads mapped across splice junctions need special handling.
Map reads using STAR aligner

- Highest sensitivity for both SNPs and indels among all programs tested

- 2-pass approach described in
  (see Suppl.I text p. 43 for detailed protocol)
  - First pass identifies splice junctions (SJ)
  - Use the SJ to guide the second round of alignment

STAR by Dobin et al., 2012 http://bioinformatics.oxfordjournals.org/content/29/1/15
BACK TO PRE-PROCESSING
Step 2: Mark duplicates to mitigate duplication artifacts

- Duplicates = **non-independent measurements** of a sequence
  - Sampled from same template of DNA
  - Violates assumptions of variant calling

- Errors in sample/library prep will get propagated to *all* the duplicates

-> Just pick the “best” copy – mitigates the effects of errors

[Diagram showing reference, mapped reads, and Picard MarkDuplicates process]

- = sequencing error propagated in duplicates
RNAseq ONLY:
Split reads that overlap splice junctions + assign flat mapping quality

- **SplitNCigarReads**
  - splits reads with Ns in the CIGAR string
  - trims overhangs
- For now, need to use –U ALLOW_N_CIGAR_READS
- Use **ReassignOneMappingQuality** read filter to reassign mapping qualities from 255 (usable by GATK) to 60
Step 3: Local realignment around indels corrects mapping errors

BEFORE

Several consecutive “SNPs” only found on reads ending on the right of the homopolymer

7bp “T” homopolymer run

AFTER

Adding a 1-bp insertion brings sanity to the entire alignment
**Step 4**: Base Recalibration (BQSR) corrects for machine errors

- Sequencers make systematic errors in base quality scores
- BQSR corrects the quality scores (not the bases)

Example of bias: qualities reported depending on nucleotide context

Data Pre-processing

- Raw Reads
- Map to Reference
- Mark Duplicates & Sort (Picard)
- Indel Realignment
- Base Recalibration
- Analysis-Ready Reads

**Original**

**Recalibrated**

RMSE = 4.188

RMSE = 0.281
VARIANT DISCOVERY
CALLING AND FILTERING VARIANTS
Distinguishing real variants from noise is hard

How we deal with this depends on what type of variant we’re looking for
Germline vs. Somatic Variant Discovery

Different approaches depending on the type of variants we are looking for

Germline SNPs and 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 SNPs and Indels

- Individually processed T/N Pair Reads
  - T
  - N
  - SNV & Indel Calling
    - M2 (MuTect2)
    - COSMIC whitelist
    - dbSNP redlist
    - Panel of Normals blacklist
    - Contamination Estimate
    - Additional Filtering
    - Analysis-Ready Variants
      - SNVs & Indels
Germline Variant Discovery

1. Analysis-Ready Reads → Var. Calling (HC in ERC mode) → Genotype Likelihoods
2. Joint Genotyping
3. Raw Variants → SNPs, Indels
4. Variant Recalibration (separately per variant type)
5. Analysis-Ready Variants → SNPs, Indels
Germline variant discovery works best with large cohorts

- Single genome in isolation: almost never useful
- Family or population data add valuable information
  - rarity of variants
  - *de novo* mutations
  - ethnic background
  - ...
Joint analysis from early stages empowers discovery

Individual callsets

Underpowered analysis

Joint callset

Empowered analysis
Joint analysis empowers discovery at difficult sites

- Sample #1 or Sample #N alone:
  - weak evidence for variant
  - may miss calling the variant

- Both samples seen together:
  - unlikely to be artifact
  - call the variant more confidently
Joint analysis ensures complete information at every site of interest

• Analyzed individually:
  – No call for either sample
  – Very different reasons!

• In joint analysis with other samples:
  – Hom-ref call and no-call genotypes emitted
Traditional **multi-sample calling** approach is very inefficient

It gives us the right answers, but...

Want to add new samples?

Got to re-run pipeline from scratch!

Compute requirements scale exponentially with number of samples

(combinatorial problem)
New **GVCF workflow** solves both problems, yields same results

Want to add new samples? Just call it by itself then re-genotype the cohort at will!

GVCF stores all genotype likelihoods

Compute requirements scale linearly with number of samples

---

Project Timeline:

- **Samples**
  - #1
  - ...
  - #N

- **Single-sample Variant calling**
  - HC
  - HC
  - HC

- **Genomic VCF intermediate**
  - Genomic VCFs

- **Joint genotyping**
  - GGVCFs
  - GGVCFs
  - GGVCFs

- **Results**
  - #1
  - ...
  - #1 to #N
At this point we have raw calls that we still need to filter
Somatic Variant Discovery

Individually processed T/N Pair Reads → SNV & Indel Calling

M2 (MuTect2)

COSMIC whitelist

Panel of Normals blacklist

dbSNP redlist

Additional Filtering

Analysis-Ready Variants

SNVs & Indels

Contamination Estimate
Aim is to detect to detect **drivers of cancer development**, not inherited risk

**Cancer determinants**

- **Germline variation**
  - Normal
  - High penetrance genes < 5%
  - Metabolic & immunologic polymorphisms
    - Sex
    - Tissue
      - Age & aging Epigenetics
        - Lifestyle factors
        - Environment
        - Infectious agents

- **Somatic alterations**
  - Tumor

**Accumulated mutations drive cancer evolution, persistence and metastasis**
Matched Tumor/Normal samples from the same individual

<table>
<thead>
<tr>
<th>SOMATIC EVENT</th>
<th>ARTIFACT</th>
<th>GERMLINE EVENT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Diagram showing somatic and germline processes in tumor and normal samples.
Tumor heterogeneity, contamination by normal & other samples

Adapted from https://science.education.nih.gov/supplements/nih1/cancer/guide/understanding1.html
Overall challenge: amount of signal may be comparable to noise

Expectation for germline variants

- Signal
- Noise

+ AF expected to follow ploidy

Expectation for somatic variants

- Signal
- Noise
- Heterogeneity
- Contamination

+ No reliance on ploidy for AF
Workflow for somatic SNVs and indels

Individually processed T/N Pair Reads

SNV & Indel Calling

T N

M2 (MuTect2)

COSMIC

blacklist

dbSNP

redlist

Panel of Normals

Additional Filtering

Contamination Estimate

cross-sample

Analysis-Ready Variants

SNVs & Indels
Workflow for calling somatic CNVs

Underlying method (CapSeg) has been around for years and has been cited in many papers

**Start:**
- Genome reference
- Reads (BAM file)
- Targets (exome or gene panel)
- Panel of normals (PoN)

**End:** Amplification/deletion calls and copy ratio for each segment
VARIANT CALLSET REFINEMENT
OPTIONS FOR REFINEMENT & EVALUATION
Callset refinement -- only for germline variants for now
GATK BEST PRACTICES
RECAP OF KEY POINTS
GATK Best Practices = complete reads-to-variants workflows

Data Pre-Processing → Variant Discovery → Callset Refinement

FASTQ → BAM → BAM → VCF

SAM/BAM → VCF
Further reading

http://www.broadinstitute.org/gatk/best-practices