Best Practices GATK Variant Detection: Theory and Practice

MSI RISS
12/7/2011
Background on the GATK

• **Genome Analysis Tool Kit** is an open source java API for NGS data handling and variant detection

• Created to support the 1000 Genomes Project
A framework for variation discovery and genotyping using next-generation DNA sequencing data

Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, Mark J Daly
Best practices pipeline overview

Phase 1: nGS data processing
Typically by lane

Input
Raw reads

Mapping

Local realignment

Duplicate marking

Base quality recalibration

Output
Analysis-ready reads

Phase 2: variant discovery and genotyping
Typically multiple samples simultaneously but can be single sample alone

Input
Sample 1 reads

Sample N reads

SNPs

Indels

Structural variation (SV)

Raw variants

Phase 3: integrative analysis

Input
Raw indels

Raw SNPs

Raw SVs

External data

Pedigrees

Population structure

Known variation

Known genotypes

Variant quality recalibration

Genotype refinement

Output
Analysis-ready variants
NGS data processing

**Phase 1: NGS data processing**
Typically by lane

- **Input**
  - Raw reads
- **Mapping**
- **Local realignment**
- **Duplicate marking**
- **Base quality recalibration**
- **Output**
  - Analysis-ready reads

**Phase 2: Variant discovery and genotyping**
Typically multiple samples simultaneously but can be single sample alone

- Sample 1 reads
- Sample N reads
- **SNPs**
- **Indels**
- **Structural variation (SV)**
- **Raw variants**

**Phase 3: Integrative analysis**

- **External data**
  - Pedigrees
  - Population structure
  - Known variation
  - Known genotypes
- **Variant quality recalibration**
- **Genotype refinement**
- **Analysis-ready variants**
Mapping

• Map with Burroughs-Wheeler Aligner (BWA)
• Result is Sequence Alignment & Mapping (SAM)
• **NOTE**: Observe GATK Central Dogma:
  • *All must be sorted in order of one of the canonical references sequences*
    • hg19: chrM, chr1, chr2, ... chrX, chrY
    • b37: 1, 2, 3, ..., 22, X, Y, MT
• Specify read group and metadata via ‘sampe –r’
• You must run Picard tools SortSam on your raw SAM to set SORT_ORDER=coordinate flag
Indel realigner removes artifactual variants near indels
Indel realigner removes artifactual variants near indels

NOTE: inclusion of known indel sites (e.g., from 1000 Genomes Project) in addition to novel sites will improve performance!
PCR duplicates can throw off coverage depth and variant frequency.

Phase 1: NGS data processing

Typically by lane

Input
Raw reads

Mapping

Local realignment

Duplicate marking

Base quality recalibration

Output
Analysis-ready reads

Duplicate removal
Raw base call qualities are inaccurate
How do they compute “empirical quality”?

- Mapping disagreements with reference genome are dominated by errors!

 Ignored known dbSNP site
 Random errors
How do they compute “empirical quality”?  

- We can gather statistics on the mapping discrepancies and compare directly to nominal quality values. E.g.,  
  - Reads from a run with assigned Q=15 may have a mapping discrepancy of 1 in 100 suggesting Q’=20
We can break down empirical quality estimates across many covariates

- E.g., an Illumina run may systematically differentially under- or over-estimate base quality across the length of the read.
We can break down empirical quality estimates across many covariates

- E.g., some dinucleotides have reliable quality values while others have huge biases.
Base recalibration strategies

• Base recalibration is crucial when
  – Individual samples spread over many lanes
  – Combining data from different platforms
  – Performing multi-sample analyses

• Inaccurate base calls will lead to false-positive SNPs
The GATK makes empirical base quality adjustment across specified covariates.

Phase 1: nGS data processing

Typically by lane

Input  Raw reads
Mapping
Local realignment
Duplicate marking

Base quality recalibration

Analysis-ready reads
Pre-processing nuances required if alignment was done externally
Additional Recommendations

• Many GATK commands can be run at the lane level, but will give better results seeing all of the data for a single sample, or even all of the data for all samples.

• For contrastive calling projects -- such as cancer tumor/normals -- we recommend cleaning both the tumor and the normal together in general to avoid slight alignment differences between the two tissue types.
Variant discovery and genotyping

Phase 1: nGS data processing
- Typically by lane
- Input: Raw reads
- Mapping
- Local realignment
- Duplicate marking
- Base quality recalibration
- Output: Analysis-ready reads

Phase 2: variant discovery and genotyping
- Typically multiple samples simultaneously
- Raw read processing:
  - Sample 1 reads
  - Sample N reads
- Raw variants
- Structural variation (SV)
- Indels
- SNPs

Phase 3: integrative analysis
- Typically but can be single sample alone
- Raw indels
- Raw SNPs
- Raw SVs
- External data
  - Pedigrees
  - Known variation
  - Known genotypes
  - Variant quality recalibration
  - Genotype refinement
  - Analysis-ready variants
GATK single sample genotype likelihoods

Bayesian model

\[
L(G \mid D) = P(G) P(D \mid G) = \prod_{b \in \{\text{good bases}\}} P(b \mid G)
\]

- Priors applied during multi-sample calculation; \(P(G) = 1\)
- Likelihood of data computed using pileup of bases and associated quality scores at given locus
- Only “good bases” are included: those satisfying minimum base quality, mapping read quality, pair mapping quality, NQS
- \(P(b \mid G)\) uses a platform-specific confusion matrix
- \(L(G \mid D)\) computed for all 10 genotypes

http://www.broadinstitute.org/gsa/wiki/images/a/ab/UGgenotypeLikelihoods.jpg
The Broad Unified Genotyper SNP caller multiple-sample allele frequency and genotype estimates

- This approach allows us to combine weak single sample calls to discover variation among samples with high confidence

Genotyper Assumptions

• Biallelic
  – the population cannot have both AC and AG at a locus

• Diploid
  – No aneuploidy: An individual sample cannot be AAC at one locus and GT at another locus.
  – Pooled samples must be separately barcoded

• MuTect will soon be available to deal with these heterogeneous conditions
Genotyper Practical issues

- For performance, keep BAM files 10-300 Gb
  - Sample-level for deep datasets (e.g., WG, exome)
  - Multi-sample for targeted or WG low-pass
- You can send multiple BAMs to the genotyper
- Unified Genotyper calls SNPs only by default
  - Use \(-glm\ BOTH\) argument for indels too.
Genotyper Practical issues (2)

• Requires input of confidence score threshold
  – Q30 for deep data (>10X); Q4 for shallow data
• Unified Genotyper does not use reads with mapping Q255. Beware bowtie users!!
  – Use `-rf ReassignMappingQuality -DMQ 60`
• Good calls on the sex chromosomes (X,Y) require care
  – See http://db.tt/TIOWVvKU
Output: Variant Call Format (VCF)

### Types of variants

#### SNPs

<table>
<thead>
<tr>
<th>Alignment</th>
<th>VCF representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGT</td>
<td>POS  REF  ALT</td>
</tr>
<tr>
<td>ATGT</td>
<td>2      C      T</td>
</tr>
</tbody>
</table>

#### Insertions

<table>
<thead>
<tr>
<th>Alignment</th>
<th>VCF representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC - GT</td>
<td>POS  REF  ALT</td>
</tr>
<tr>
<td>ACTGT</td>
<td>2      C      CT</td>
</tr>
</tbody>
</table>

#### Deletions

<table>
<thead>
<tr>
<th>Alignment</th>
<th>VCF representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGT</td>
<td>POS  REF  ALT</td>
</tr>
<tr>
<td>A - TT</td>
<td>1      ACG    A</td>
</tr>
</tbody>
</table>

#### Complex events

<table>
<thead>
<tr>
<th>Alignment</th>
<th>VCF representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGT</td>
<td>POS  REF  ALT</td>
</tr>
<tr>
<td>A - TT</td>
<td>1      ACG    AT</td>
</tr>
</tbody>
</table>

#### Large structural variants

<table>
<thead>
<tr>
<th>VCF representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS  REF  ALT  INFO</td>
</tr>
<tr>
<td>100  T   &lt;DEL&gt; SVTYPE=DEL; END=300</td>
</tr>
</tbody>
</table>

http://vcftools.sourceforge.net/specs.html
# Example VCF

**Mandatory header lines**

```plaintext
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=Integer,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

**Optional header lines (meta-data about the annotations in the VCF body)**

```plaintext
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0/1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1/0:77 1/1:95
1 100 T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
```

**Reference alleles (GT=0)**

**Alternate alleles (GT>0 is an index to the ALT column)**

**Phased data (G and C above are on the same chromosome)**

**Large SV**
Integrative analysis
All SNP calls are not created equal: Variant Quality Recalibration

- **GOAL**: separate true genotype calls from machine artifacts
- **APPROACH**: use known characteristics of SNPs to ascribe a quality score to the genotype call itself.
  - Known dbSNP rates
    - As a result of the 1000 Genomes project, 99% of all variants will be cataloged for a European sample (dbSNP currently >90%).
    - SNPs at known sites are more likely to be real
    - Differences in known site rates between samples are indicative of quality
  - Transitions/Transversions are non-random
    - Random: 0.5
    - Whole genome: 2.0-2.1
    - Exome: 3.0-3.3
  - Population-specific heterozygosity should hold
    - E.g., observing all AT with no AA or TT is bad!
- Minimal read strand bias should exist
- Depth of coverage (ridiculous pileups)
- Can also check concordance with non-reference genotyping chips
Recalibrator options

- **VS**
- good
- bias indicative of false positive

**variant allele**

- largest homopolymer run of variant allele in either direction

- test hypothesis alt base (C) is consistently at beginning or end of read (bias could indicate reads are mis-mapped)

- test hypothesis reads with alt base (T) have a consistently lower MAPQ (white: MAPQ=0)

- confidence assigned by unified genotyper on variant site divided by number of reads (QD).
  - As depth increases, so should confidence.
  - \[ QD = \frac{((AB+BB)/AA\text{ from PLs})}{\text{unfiltered depth}} \]

\[ PL = \text{probability likelihood of genotypes (AA, AB, BB: 0/0, 0/1, 1/1)} \]
Variant Quality Recalibrator Implementation

• Uses a Gaussian Mixture Model trained on annotated variants

• Anything can be used as truth data.
  – Validation assays
  – several 1000G callsets, or
  – auto-generate your own by subsetting to the highest quality SNPs

• There is no reason to decide between high sensitivity or high specificity. Just use a probabilistic callset.

Ryan Poplin, http://db.tt/nR0c7My
Variant Quality Score Recalibration: training on highly confident known sites to determine the probability that other sites are true

DePristo et al. Nature Genetics. 2011
Removed SNPs are often near the centromere or telomere

T2D exome batch 005: delta hand filter vs. VQSR

Called with hand filters, removed by VQSR
Called by VQSR, removed with hand filters

Ryan Poplin, http://db.tt/nR0c7My
Practical Issues for Variant Recalibrator

- GATK Bundle includes training files for SNPs from HapMap v3.3 and OmniChip array from the 1000G plus Indels
- Certain annotations may not make sense for your project
  - InbreedingCoeff is a population level statistic that requires at least 10 samples in order to be calculated.
  - DP (depth) should not be used with hybridization capture datasets since there is extreme variation in the depth to which targets are captured.
# Evaluation on 3 different applications

## Table 1  Next-generation DNA sequencing datasets analyzed

<table>
<thead>
<tr>
<th></th>
<th>HiSeq</th>
<th>Exome</th>
<th>Low-pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples</td>
<td>NA12878</td>
<td>NA12878</td>
<td>NA12878 + 60 unrelated CEPH individuals</td>
</tr>
<tr>
<td>Sequencing</td>
<td>Whole genome shotgun;</td>
<td>Agilent exome hybrid</td>
<td>Whole genome shotgun; Illumina GenomeAnalyzer^{17}; Life/SOLiD^{33}; Roche/454 (ref. 19)</td>
</tr>
<tr>
<td>technologies</td>
<td>Illumina HiSequation</td>
<td>capture^{31,32}; Illumina</td>
<td>~4x</td>
</tr>
<tr>
<td></td>
<td>(2000)^{17}</td>
<td>GenomeAnalyzer^{17}</td>
<td></td>
</tr>
<tr>
<td>Coverage per sample</td>
<td>~60x</td>
<td>~150x; 93% of bases at &gt;20x coverage</td>
<td></td>
</tr>
<tr>
<td>Read architecture</td>
<td>101 bp paired end</td>
<td>76/101 bp paired end</td>
<td>25, 36, 51, 76, ~250 (454) bp single and paired ends</td>
</tr>
<tr>
<td>Targeted area</td>
<td>2.85 Gb of autosomes and chr. X</td>
<td>28 Mb</td>
<td>2.85 Gb of autosomes and chr. X</td>
</tr>
<tr>
<td>Data set source</td>
<td>New, generated for this article</td>
<td>New, generated for this article</td>
<td>1000 Genomes Project</td>
</tr>
<tr>
<td>Aligner(s)</td>
<td>BWA^{10}</td>
<td>MAQ^{9}</td>
<td>MAQ^{10}; Corona Lite; SSAHA^{12}</td>
</tr>
</tbody>
</table>

Chr., chromosome.
## Summary of SNPs from pipeline

<table>
<thead>
<tr>
<th>Call set</th>
<th>Site discovery</th>
<th>Comparison to NA12878 variants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of SNPs</td>
<td>Ti/Tv</td>
</tr>
<tr>
<td></td>
<td>All Known Novel dbSNP% Known Novel</td>
<td>NR sensitivity NRD rate NR sensitivity NRD rate</td>
</tr>
<tr>
<td>HiSeq</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recalibrated, MSA raw calls</td>
<td>4.23M 3.46M</td>
<td>770K 81.81 2.06 1.53</td>
</tr>
<tr>
<td>Hard filtered</td>
<td>3.56M 3.20M</td>
<td>359K 89.92 2.10 1.96</td>
</tr>
<tr>
<td>Variant recalibrated</td>
<td>3.62M 3.27M</td>
<td>348K 90.37 2.14 2.07</td>
</tr>
<tr>
<td>Low-pass*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recalibrated, MSA raw calls</td>
<td>20.27M 6.73M</td>
<td>13.54M 33.21 2.05 1.01</td>
</tr>
<tr>
<td>Variant recalibrated call set</td>
<td>8.19M 5.94M</td>
<td>2.25M 72.54 2.17 1.91</td>
</tr>
<tr>
<td>Exome capture</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recalibrated, MSA raw calls</td>
<td>18.7K 16.9K</td>
<td>1.8K 90.28 3.19 1.55</td>
</tr>
<tr>
<td>Hard filtered</td>
<td>16.0K 15.2K</td>
<td>828 94.85 3.37 2.70</td>
</tr>
<tr>
<td>Variant recalibrated</td>
<td>17.0K 16.2K</td>
<td>855 94.97 3.27 2.73</td>
</tr>
</tbody>
</table>
For more information

• GATK wiki

• Paper