The 1000 Genomes Project

Lessons From

Variant Calling and Genotyping

October 13th, 2011

Hyun Min Kang

University of Michigan, Ann Arbor
OVERVIEW OF PHASE 1 CALL SET
1000 Genomes integrated genotypes

- Deep Exomes
  - SNPs 38M
- Low-pass Genomes
  - INDELs 4.0M
  - SVs 15k

Integrated Genotypes ~42M
# Methods for integrated genotypes

<table>
<thead>
<tr>
<th>Components</th>
<th>SNPs</th>
<th>INDELs</th>
<th>SVs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low-Pass Genomes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Call Sets</td>
<td>BC, BCM, BI NCBI, SI, UM</td>
<td>BC, BI, DI OX, SI</td>
<td>BI, EBI, EMBL UW, Yale</td>
</tr>
<tr>
<td>Consensus</td>
<td>VQSR</td>
<td>VQSR</td>
<td>GenomeSTRiP</td>
</tr>
<tr>
<td><strong>Deep Exomes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Call Sets</td>
<td>BC, BCM, BI UM, WCMC</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Consensus</td>
<td>SVM</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Likelihood</td>
<td>BBMM</td>
<td>GATK</td>
<td>GenomeSTRiP</td>
</tr>
<tr>
<td>Site Models</td>
<td>Variants are linearly ordered as point mutations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haplotyper</td>
<td>MaCH/Thunder with BEAGLE’s initial haplotypes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
From PILOT to PHASE1

PILOT
• 14.8M SNPs
• Ts/Tv 2.01
• Includes 97.8% HapMap3

PHASE1
• 36.8M SNPs
• Ts/Tv 2.17
• Includes 98.9% HapMap3

Autosomal chromosomes only
From PILOT to PHASE1

- **PILOT \(\land\) PHASE1**
  - 13.1M SNPs
  - Ts/Tv 2.18
  - Includes 97.7% of HapMap3

- **PILOT-only**
  - 1.7M SNPs
  - Ts/Tv 1.11
  - Includes 0.15% HapMap3

- **PHASE1-only**
  - 23.8M SNPs
  - Ts/Tv 2.16
  - Includes 1.2% of HapMap3
From PILOT to PHASE1

PILOT-only
- 1.7M SNPs
- Ts/Tv 1.11
- Includes 0.15% HapMap3

PILOT ∩ PHASE1
- 13.1M SNPs
- Ts/Tv 2.18
- Includes 97.7% of HapMap3

PHASE1-only
- 23.8M SNPs
- Ts/Tv 2.16
- Includes 1.2% of HapMap3

100k monomorphic SNPs in 2.5M OMNI Array (>1,000 individuals)
From PILOT to PHASE1: Improved SNP calls

PILOT-only
- 1.7M SNPs
- Ts/Tv 1.11
- Includes 0.15% HapMap3

PHASE1-only
- 13.1M SNPs
- Ts/Tv 2.18
- Includes 97.7% of HapMap3

PILOT and PHASE1
- 23.8M SNPs
- Ts/Tv 2.16
- Includes 1.2% of HapMap3

100k monomorphic SNPs in 2.5M OMNI Array (>1,000 individuals)

59.6% of OMNI-MONO

OMNI-MONO information was not used in making phase1 variant calls
IMPROVEMENT IN METHODS SINCE PILOT
1000 Genomes’ engines for improved variant calls and genotypes

• INDEL realignment
• Per Base Alignment Quality (BAQ) adjustment
• Robust consensus SNP selection strategy
  – Variant Quality Score Recalibration (VQSR)
  – Support Vector Machine (SVM)
• improved Genotype Likelihood Calculation
  – BAM-specific Binomial Mixture Model (BBMM)
  – Leveraging off-target exome reads
INDEL Realignment: How it works...

• Given a list of potential indels...
• Check if reads consistent with SNP or indel
• Adjust alignment as needed
• Greatly reduces false-positive SNP calls

Ref: AAGCGTCGG
AAGCGT
AAGCGTC
AAGCGTCG
AAGCGC
AAGCGCG
AAGCGCGG

Neighboring SNPs?

With one read, hard to choose between alternatives

AAGCGTCG
AAGCGT
AAGCGTC
AAGCGTCG
AAGCGC
AAGCGCG
AAGCGCGG

Short Indel?

Read pile consistent with a 1bp deletion

Read pile consistent with the reference sequence

Eric Banks and Mark DePristo
Per Base Alignment Qualities

Short Read

GATAGCTAGCTAGCTGATGA GCCG
5’-AGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGGATC-3’

Reference Genome
Per Base Alignment Qualities

Should we insert a gap?

Short Read

GATAGCTAGCTAGCTGATGAGGCC-G

5'-AGCTGATAGCTAGCTAGCTAGCTAGCTGATGAGCCCCGATC-3'

Reference Genome

Heng Li
Per Base Alignment Qualities

Compensate for Alignment Uncertainty With Lower Base Quality

Short Read

GATAGCTAGCTAGCTAGCTGATGAGCCG
5’-AGCTGATAGCTAGCTAGCTAGCTGATGAGCCCCGATC-3’

Reference Genome

Heng Li
Per Base Alignment Qualities

Compensate for Alignment Uncertainty With Lower Base Quality

Short Read

GATAGCTAGCTAGCTGATGAGCCG

5′-AGCTGATAGCTAGCTAGCTAGCTAGCTGATGAGCCCGATC-3′

Reference Genome

Improves quality near new indels and sequencing artifacts

Heng Li
## Producing high-quality consensus call sets

<table>
<thead>
<tr>
<th>Center</th>
<th>Total # variants</th>
<th>dbSNP% (129)</th>
<th>Novel Ts/Tv</th>
<th>Omni poly sensitivity</th>
<th>Omni MONO false discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broad</td>
<td>36.6M</td>
<td>22.7</td>
<td>2.17</td>
<td>96.5%</td>
<td>5.45%</td>
</tr>
<tr>
<td>Sanger</td>
<td>34.8M</td>
<td>22.9</td>
<td>2.18</td>
<td>96.1%</td>
<td>4.94%</td>
</tr>
<tr>
<td>UMich</td>
<td>34.5M</td>
<td>24.4</td>
<td>2.16</td>
<td>98.0%</td>
<td>2.77%</td>
</tr>
<tr>
<td>Baylor</td>
<td>34.1M</td>
<td>21.8</td>
<td>2.13</td>
<td>93.8%</td>
<td>1.43%</td>
</tr>
<tr>
<td>BC</td>
<td>33.3M</td>
<td>23.9</td>
<td>2.10</td>
<td>94.9%</td>
<td>9.72%</td>
</tr>
<tr>
<td>NCBI</td>
<td>30.7M</td>
<td>25.7</td>
<td>2.33</td>
<td>94.6%</td>
<td>10.47%</td>
</tr>
<tr>
<td>VQSR Consensus</td>
<td>37.9M</td>
<td>21.7</td>
<td>2.16</td>
<td>98.4%</td>
<td>1.80%</td>
</tr>
<tr>
<td>2 of 6</td>
<td>39.1M</td>
<td>22.2</td>
<td>2.15</td>
<td>98.6%</td>
<td>11.23%</td>
</tr>
</tbody>
</table>

Ryan Poplin
Consensus SNP site selection under multidimensional feature space

Goo Jun – Joint variant calling and … - Platform 192, Friday 5:30 (Room 517A)
### Improved likelihood estimation produces more accurate genotypes

<table>
<thead>
<tr>
<th>Likelihood Model</th>
<th># SNPs Evaluated</th>
<th>HET (OMNI)</th>
<th>NONREF -EITHER</th>
<th>OVERALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAQ</td>
<td>51,002</td>
<td>1.86%</td>
<td>2.03%</td>
<td>0.65%</td>
</tr>
<tr>
<td>BBMM</td>
<td>51,002</td>
<td>1.49%</td>
<td>1.86%</td>
<td>0.60%</td>
</tr>
</tbody>
</table>

**Evaluation in April 2011**

**KEY IDEA in BBMM:**
Re-estimate the genotype likelihood by clustering the variants based on the read distribution.
Off-target exome reads improves genotype quality

<table>
<thead>
<tr>
<th>Sites</th>
<th>#chr20 Variants</th>
<th>#OMNI Overlaps</th>
<th>HET (OMNI)</th>
<th>NREF-EITHER</th>
<th>OVER-ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-coverage SNPs (May 2011)</td>
<td>824,876</td>
<td>52,329</td>
<td>1.10%</td>
<td>1.41%</td>
<td>0.46%</td>
</tr>
<tr>
<td>Integrated (Nov 2011) - LC+EX/ INDELs/ SVs</td>
<td>907,452</td>
<td>52,329</td>
<td>0.79%</td>
<td>1.07%</td>
<td>0.35%</td>
</tr>
</tbody>
</table>

Integrated on-target coding genotypes are also more accurate than low-coverage-only or exome-only platforms
## Genotype Qualities in SVs and INDELs

<table>
<thead>
<tr>
<th>SV genotypes</th>
<th>Sites</th>
<th>Call Rate</th>
<th>Evaluation Data</th>
<th># Sites Evaluated</th>
<th>HET (eval)</th>
<th>NONREF-EITHER</th>
<th>OVERALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEFORE Integration</td>
<td>13,973</td>
<td>95.2% (^2)</td>
<td>Conrad (80% RO)</td>
<td>1962</td>
<td>0.61%</td>
<td>1.60%</td>
<td>0.20%</td>
</tr>
<tr>
<td>AFTER integration</td>
<td>13,973</td>
<td>100%</td>
<td>Conrad (80% RO)</td>
<td>1962</td>
<td>0.62%</td>
<td>0.93%</td>
<td>0.11%</td>
</tr>
<tr>
<td>IMPUTED</td>
<td>13,973</td>
<td>100%</td>
<td>Conrad (80% RO)</td>
<td>1962</td>
<td>4.17%</td>
<td>5.75%</td>
<td>0.74%</td>
</tr>
</tbody>
</table>

### INDEL genotypes

<table>
<thead>
<tr>
<th>Evaluation Data</th>
<th>#Sites Evaluated</th>
<th>HOMREF</th>
<th>HET</th>
<th>HOMALT</th>
<th>NREF-EITHER</th>
<th>OVERALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000G CGI</td>
<td>1,029</td>
<td>0.65%</td>
<td>2.68%</td>
<td>1.24%</td>
<td>2.65%</td>
<td>1.35%</td>
</tr>
<tr>
<td>1000G Array (Mills et al)</td>
<td>1,029</td>
<td>2.21%</td>
<td>7.16%</td>
<td>3.77%</td>
<td>7.56%</td>
<td>3.97%</td>
</tr>
</tbody>
</table>
MORE IN-DEPTH VIEW OF PHASE 1 INTEGRATED GENOTYPES
Sensitivity at low-frequency SNPs

Sensitivity compared to OMNI-HapMap2 overlapping SNPs (chr1)

% SNPs detected in 1000G

Non-reference allele count among 1,092 individuals (OMNI)

- 0.1%
- 0.5%
- 1.0%
>96% SNPs are detected compared to deep genomes

![Graph showing the number of chr20 SNPs shared between 1000G and CGI, and those exclusive to CGI.](image)
Genotype discordance by frequency
Impact of sequencing depth on genotype accuracy
(interim integrated panel, chr20)
Highlights

- The quality of phase 1 call set is much more improved compared to pilot call set
- 1000G engines for phase 1 variant calls produced high-sensitivity, high-specificity variant calls
- >99% of genotypes are concordant with array-based genotypes
- Likelihood-based integrated improves off-target & on-target genotyping qualities
Acknowledgements

The 1000 Genomes Project
1000 Genomes Analysis Group