Implementation of Consensus Variant Calling using Globus Genomics

Vassily Trubetskoy
April 16, 2014
Road Map

Context:

Disease/medical genetics, next generation sequencing

Consensus Variant Calling:

motivation, high level description

Globus genomics experiments:

Implementation, testing in ACE autism dataset
Context
Disease and Medical Genetics

Population

Diseased

Normal

AAATTCCGC

AAATTCCGC

AAATTCCGC

AAATTCCGC

AAATTCCGC

AAATTCCGC

A TGCCTTCGGAAATATATAT C GGGCTTAGGCT

A TGCCTTCGGAAATATATAT C GGGCTTAGGCT

A TGCCTTCGGAAATATATAT C GGGCTTAGGCT

A TGCCTTCGGAAATATATAT C GGGCTTAGGCT

A TGCCTTCGGAAATATATAT C GGGCTTAGGCT

A TGCCTTCGGAAATATATAT C GGGCTTAGGCT
Published Genome-Wide Associations through 12/2012
Published GWA at $p \leq 5 \times 10^{-8}$ for 17 trait categories

NHGRI GWA Catalog
www.genome.gov/GWAStrudies
www.ebi.ac.uk/fgpt/gwas/
**Medical Genetics**

IMURAN® (azathioprine) 50-mg Scored Tablets Rx only

*TPMT Testing:* It is recommended that consideration be given to either genotype or phenotype patients for TPMT. Phenotyping and genotyping methods are commercially available. The most common non-functional alleles associated with reduced levels of TPMT activity are *TPMT*2, *TPMT*3A and *TPMT*3C. Patients with two non-functional alleles (homozygous) have low or absent TPMT activity and those with one non-functional allele (heterozygous) have intermediate activity. Accurate phenotyping (red blood cell TPMT activity) results are not

credit: PharmGKB
What is next generation sequencing?

Figure 1: Conceptual Overview of Whole-Genome Resequencing

A. Extracted gDNA.
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.
C. Individual sequence reads are reassembled by aligning to a reference genome.
D. The whole-genome sequence is derived from the consensus of aligned reads.

Credit: Illumina pub. 770-2012-008
Next gen sequencing

extract DNA

fragment and *sequence

TATATCGGGCTTAGGCTAAATT
GCTTGCTTCGGAATATATATCGGGGC
TTAGGCTAAATTCGCTTGCTTGCGGA
ATATATCGGGCTTAGGCTAAATTCCGC

ATCGGGCTTAGGCTAAA
TGCTTTCGGAATATATATCGGGCTTAG
Next gen sequencing

align to reference

fragment and *sequence
Next gen sequencing

align to reference

identify variants

SNP  SNP  SNP
Data scale

Type II Diabetes Genes Consortium

- 600 whole genomes with 250 TB in raw reads (~40 GB in genotype data)

The Cancer Genome Atlas Project

- large scale data collection in many patients, many cancers, and many tissues
- Sept 2013: 9k patients, 147k files, ~13Tb of genotype data
Cumulative Error

Robasky et. al Nature Reviews Genetics 2013
Consensus variant calling
Motivation

- Cumulative sequencing error
- Performance of different models of variation
- Existing ensemble methods
Overview

aligned reads

fit models

consensus sites

consensus genotypes

SNP1: A|T
SNP2: C|T
SNP3: G|C
## Motivation

<table>
<thead>
<tr>
<th>Accessibility</th>
<th>Reproducibility</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usable by non-programmers</td>
<td>Share <em>exact</em> methods</td>
<td>Scale storage</td>
</tr>
<tr>
<td>Single interface for similar tasks</td>
<td>Data provenance</td>
<td>Parallel execution</td>
</tr>
<tr>
<td>Open source</td>
<td></td>
<td>Iterate workflow construction</td>
</tr>
</tbody>
</table>
Accessibility -> Galaxy + Globus Online
Reproducibility -> Galaxy + Globus Online
Efficiency -> Amazon Elastic Compute
Consensus calling with Globus Genomics
Autism ACE

- Autism Centers of Excellence consortium
- 132 samples with 40 complete trios
- Illumina Whole Exome Capture product
- 1.8 TB of raw data
Estimating Genotype Error

- Mendelian transmission
- Sanger validated genotypes
- Variant rediscovery
Estimating genotyping error: mendelian transmission
Estimating genotyping error: Sanger validated variants
Estimating genotyping error: Prior information

Rediscovering variants from the 1000 Genomes and Exome Variant Server projects

NHLBI Exome Sequencing Project (ESP)
Exome Variant Server

1000 Genomes
A Deep Catalog of Human Genetic Variation
## Results

<table>
<thead>
<tr>
<th>Call Set</th>
<th># of Sites</th>
<th>Mendel Rate</th>
<th>EVS discover</th>
<th>1000G discover</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>AtlasSNP2</td>
<td>214,149</td>
<td>0.245%</td>
<td>72.7%</td>
<td>69.0%</td>
<td>92.3%</td>
</tr>
<tr>
<td>Freebayes</td>
<td>140,803</td>
<td>0.449%</td>
<td>78.9%</td>
<td>74.3%</td>
<td>81.1%</td>
</tr>
<tr>
<td>GATK</td>
<td>265,625</td>
<td>1.03%</td>
<td>61.3%</td>
<td>58.3%</td>
<td>84.4%</td>
</tr>
<tr>
<td>Consensus</td>
<td>129,706</td>
<td>0.0459%</td>
<td>82.9%</td>
<td>78.1%</td>
<td>93.5%</td>
</tr>
</tbody>
</table>
## Runtimes

<table>
<thead>
<tr>
<th>Model</th>
<th>Runtime (days)</th>
<th>CPU time (days)</th>
<th>Nodes used</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATK UG</td>
<td>0.875</td>
<td>10.1</td>
<td>23</td>
</tr>
<tr>
<td>Freebayes</td>
<td>1.31</td>
<td>30.1</td>
<td>23</td>
</tr>
<tr>
<td>ATLAS</td>
<td>4.6</td>
<td>x</td>
<td>135</td>
</tr>
</tbody>
</table>
Concurrent work

Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls

Justin M Zook, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide & Marc Salit

Affiliations | Contributions | Corresponding author

Received 14 December 2013 | Accepted 27 January 2014 | Published online 16 February 2014

Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing

Jason O’Rawe1,2, Tao Jiang3, Guangqing Sun3, Yiyang Wu1,2, Wei Wang4, Jingchu Hu1, Paul Bodily5, Lifeng Tian6, Hakon Hakonarson6, W Evan Johnson7, Zhi Wei6, Kai Wang8,9 and Gholson J Lyon1,2,12

Published online 10 December 2013

The role of replicates for error mitigation in next-generation sequencing

Kimberly Robasky, Nathan E. Lewis & George M. Church

Affiliations | Corresponding author

Published online 10 December 2013
Acknowledgements

ACE Consortium
• Ed Cook
• Jim Sutcliffe

Globus Group
• Ravi Madduri
• Ian Foster
• Alex Rodriguez
• Paul Dave

Cox Lab
• Nancy Cox
• Lea Davis

Funding
• Amazon EC2 Academic grants
Description:

This is an implementation of an ensemble variant calling method. Specifically, it takes VCF files generated by various calling algorithms and merges them according to specified thresholds on variant and genotype concordance. The resulting VCF can range from a strict consensus among inputs, to a union of all possible observations.