variant calling in high throughput data

Analysis of Biological Sequences 410.638
types of genetic variation

SNV (single nucleotide variant)

ACTTGC   ACGTGC

indel

ACTTGC   ACTCCTTGC
types of genetic variation

CNV (copy number variation)

chromosome-level changes: insertion, deletion, inversion, translocation
lots of genetic data!!!

several large consortia have managed to sequence altogether ~100,000 human exomes and 10,000 human genomes, and counting!

Smaller, disease-focused groups have contributed thousands more genomes and exomes.

Data are variably accessible.
exome sequencing
laboratory workflow
About IGSR and the 1000 Genomes Project

The 1000 Genomes Project ran between 2008 and 2015, creating the largest public catalogue of human variation and genotype data. As the project ended, the Data Coordination Centre at EMBL-EBI has received continued funding from the Wellcome Trust to maintain and expand the resource. The International Genome Sample Resource (IGSR) was set up to do this and has the following aims:

1. Ensure the future access to and usability of the 1000 Genomes reference data
2. Incorporate additional published genomic data on the 1000 Genomes samples
3. Expand the data collection to include new populations not represented in the 1000 Genomes Project

The 1000 Genomes Project

Overview of the 1000 Genomes Project

The goal of the 1000 Genomes Project was to find most genetic variants with frequencies of at least 1% in the populations studied.

The 1000 Genomes Project took advantage of developments in sequencing technology, which sharply reduced the cost of sequencing. It was the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation. Data from the 1000 Genomes Project was quickly made available to the worldwide scientific community through freely accessible public databases.
The goal of the NHLBI GO Exome Sequencing Project (ESP) is to discover novel genes and mechanisms contributing to heart, lung and blood disorders by pioneering the application of next-generation sequencing of the protein coding regions of the human genome across diverse, richly-phenotyped populations and to share these datasets and findings with the scientific community to extend and enrich the diagnosis, management and treatment of heart, lung and blood disorders. The groups participating and collaborating in the NHLBI GO ESP include:

- Seattle GO – University of Washington, Seattle, WA
- BroadGO – Broad Institute of MIT and Harvard, Cambridge, MA
- WHISP – Ohio State University Medical Center, Columbus, OH
- Lung GO – University of Washington, Seattle, WA
- WashU GO – Washington University, St. Louis, MO
- Heart GO – University of Virginia Health System, Charlottesville, VA
- ChargeS GO – University of Texas Health Sciences Center at Houston

The group includes some of the largest well-phenotyped populations in the United States, representing more than 200,000 individuals altogether from the:

- Women’s Health Initiative (WHI)
- Framingham Heart Study (FHS)
- Jackson Heart Study (JHS)
- Multi-Ethnic Study of Atherosclerosis (MESA)
- Atherosclerosis Risk in Communities (ARIC)
- Coronary Artery Risk Development in Young Adults (CARDIA)
ExAC Browser (Beta) | Exome Aggregation Consortium

About ExAC

The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

The data set provided on this website spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. The ExAC Principal Investigators and groups that have contributed data to the current release are listed here.

All data here are released under a Fort Lauderdale Agreement for the benefit of the wider biomedical community - see the terms of use here.

Recent News

August 8, 2016
- CNV calls are now available on the ExAC browser

March 14, 2016
- Version 0.3.1 ExAC data and browser (beta) is released! (Release notes)

January 13, 2015
- Version 0.3 ExAC data and browser (beta) is released! (Release notes)

October 23, 2014
The 100,000 Genomes Project

The project will sequence 100,000 genomes from around 70,000 people. Participants are NHS patients with a rare disease, plus their families, and patients with cancer.

The aim is to create a new genomic medicine service for the NHS – transforming the way people are cared for. Patients may be offered a diagnosis where there wasn't one before. In time, there is the potential of new and more effective treatments.

The project will also enable new medical research. Combining genomic sequence data with medical records is a ground-breaking resource. Researchers will study how best to use genomics in healthcare
humans are diploid
humans are diploid

mismatch in some of the reads
humans are diploid

mismatch in roughly half of the reads

scattered additional mismatches
typical heterozygous variant
somewhat unclear
very naïve variant caller

we could just assume that the genomic DNA gets sheared/cut randomly and that we sequence without bias, and correctly align, a random subset of those pieces.

The depth of coverage would follow a Poisson distribution, and we could use a t test to determine whether a site has enough alternate alleles to be called a homozygous or heterozygous variant.
why my very naïve variant caller doesn’t work

biased sampling
sequencing error
PCR duplicates
confusing behavior of aligners (index ordering)
multimapping reads
mappability
low complexity sequence
reference genome problems
short reads
why my very naïve variant caller doesn’t work

biased sampling: coverage varies wildly, so one size won’t fit all. We’ll need to approach low and high coverage regions differently. Many variant callers have both a minimum and maximum read cutoff.

sequencing error: 0.5% to 15% from currently used sequencers. Not entirely random.
heterozygous variant?

coverage: 14
G: 8
T: 5
A: 1

ACGGGCTAGCTAGCGGCGCGATATCGAT
GATGATCGTGTACGTGGGGCGATCGATCGGGC
ACGGGCTAGCTAGCGGCGCGATATCGAT
GATGATCGTGTACGTGGGGCGATCGATCGGGC
ACGGGCTAGCTAGCGGCGCGATATCGAT
GATGATCGTGTACGTGGGGCGATCGATCGGGC
ACGGGCTAGCTAGCGGCGCGATATCGAT
GATGATCGTGTACGTGGGGCGATCGATCGGGC
ACGGGCTAGCTAGCGGCGCGATATCGAT
GATGATCGTGTACGTGGGGCGATCGATCGGGC
ACGGGCTAGCTAGCGGCGCGATATCGAT
GATGATCGTGTACGTGGGGCGATCGATCGGGC
ACGGGCTAGCTAGCGGCGCGATATCGAT
GATGATCGTGTACGTGGGGCGATCGATCGGGC
ACGGGCTAGCTAGCGGCGCGATATCGAT
GATGATCGTGTACGTGGGGCGATCGATCGGGC
ACGGGCTAGCTAGCGGCGCGATATCGAT
GATGATCGTGTACGTGGGGCGATCGATCGGGC
ACGGGCTAGCTAGCGGCGCGATATCGAT
GATGATCGTGTACGTGGGGCGATCGATCGGGC
ACGGGCTAGCTAGCGGCGCGATATCGAT
GATGATCGTGTACGTGGGGCGATCGATCGGGC
short read sequencing errors

read quality scores do not accurately reflect this problem
why my very naïve variant caller doesn’t work

PCR duplicates: constructing a sequencing library involves quite a few rounds of PCR. Some molecules (particularly shorter ones) will get amplified just a little bit more in early rounds; this is known as a “jackpot event”
why my very naïve variant caller doesn’t work

classifying behavior of aligners (index ordering): if I change the order in which the chromosomes appear in a BWT index, I’ll get different variant calls!!!

this is probably due to multimapping reads.

If a read can align in several places equally well, an aligner will pick one of those locations “randomly.”
why my very naïve variant caller doesn’t work

mappability: the human genome is highly repetitive!

---

Reconsidering the significance of genomic word frequencies

Miklós Csűrös¹, Laurent Noé² and Gregory Kucherov²

¹ Department of Computer Science and Operations Research, Université de Montréal, CP 6128, succ. Centre-Ville, Montréal, Québec H3C 3J7, Canada
² Laboratoire d'Informatique Fondamentale de Lille, Bât. M3, 59655 Villeneuve d'Ascq Cédex, France
predicted 13mer frequencies if human genome were random

actual kmer frequencies
why my very naïve variant caller doesn’t work

low complexity sequence: large portions of the human genome are low-entropy sequence like

ACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACAC

this is difficult to sequence & is nearly unalignable
why my very naïve variant caller doesn’t work

reference genome problems: none of us has the reference genome. Aligning to it ensures that we will miss variation and will create false positive variants.

approaches: first-pass alignment to roughly establish proper reference
align to ethnicity-specific genome
why my very naïve variant caller doesn’t work

short reads: structural variants are hard to find!! reads have to span the edges of these variants in order to be informative

nanopore and PacBio sequencing have helped to resolve large structural variants.
menagerie of randomly selected variant callers

GATK
FreeBayes
Platypus
considerations (in addition to the previous list)

- proximity of variants to indels
- strand bias
- mapping quality
The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data

Aaron McKenna,¹ Matthew Hanna,¹ Eric Banks,¹ Andrey Sivachenko,¹ Kristian Cibulskis,¹ Andrew Kernytsky,¹ Kiran Garimella,¹ David Altshuler,¹,² Stacey Gabriel,¹ Mark Daly,¹,² and Mark A. DePristo¹,³

¹Program in Medical and Population Genetics, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA; ²Center for Human Genetic Research, Massachusetts General Hospital, Richard B. Simches Research Center, Boston, Massachusetts 02114, USA
GATK

- uses MapReduce algorithm (Google, Yahoo!) to divide computations into pieces to feed to map function, then reduce map results into a final product
- excellent parallelization
- SNP discovery, genotyping perform calculations at genomic loci independently
- peak calling integrates height of read pileups to reduce to intervals
- “sharding” creates kb-size pieces (size determined by data & OS) that will be handled independently
- can merge multiple input files
- uses bam format
GATK coverage estimation

**Figure 3.** MHC depth of coverage in JPT samples of the 1000 Genomes Project pilot 2, calculated using the GATK depth of coverage tool. Coverage is averaged over 2.5-kb regions, where lines represent a local polynomial regression of coverage. The track containing all known annotated genes from the UCSC Genome Browser is shown in gray, with HLA genes highlighted in red. Coverage drops near 32.1 M and 32.7 M correspond with increasing density of HLA genes.
GATK genotyping

at every position calculates prior probability of AA, AC, AT, AG, CC, CT, CG, TT, TG, GG using Bayes’ formula

\[ p(D|G) = \prod_{b \in \text{pileup}} p(b|G), \]

\[ p(b|A) = \begin{cases} 
\frac{e}{3}: b \neq A \\
1 - e: b = A 
\end{cases} \]

D=data
G=genotype
b=each base at the target locus
e=quality score for the base

very fast, very accurate, state-of-the-art
Haplotype-based variant detection from short-read sequencing

Erik Garrison and Gabor Marth

July 24, 2012

Abstract

The direct detection of haplotypes from short-read DNA sequencing data requires changes to existing small-variant detection methods. Here, we develop a Bayesian statistical framework which is capable of modeling multiallelic loci in sets of individuals with non-uniform copy number. We then describe our implementation of this framework in a haplotype-based variant detector, FreeBayes.

\[ P(\text{genotype}|\text{data}) \]
\[
P(G_1, \ldots, G_n | R_1, \ldots, R_n) = \frac{P(G_1, \ldots, G_n) P(R_1, \ldots, R_n | G_1, \ldots, G_n)}{P(R_1, \ldots, R_n)}
\]

\[
P(G_1, \ldots, G_n | R_1, \ldots, R_n) = \frac{P(G_1, \ldots, G_n) \prod_{i=1}^{n} P(R_i | G_i)}{\sum_{\forall G_1, \ldots, G_n} (P(G_1, \ldots, G_n) \prod_{i=1}^{n} P(R_i | G_i))}
\]

\[
P(R_i | G_i) = \sum_{\forall (R_i \in G_i)} \left( \frac{s_i!}{\prod_{j=1}^{k_i} o_j'} \prod_{j=1}^{k_i} \left( \frac{f_{ij}}{m_i} \right)^{o_j'} \prod_{l=1}^{s_i} P(b'_i | b_l) \right)
\]

probability of obtaining a set of reads given an underlying genotype = probability of sampling the set of observations from the underlying genotype * probability that reads are correct
Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications

Andy Rimmer\textsuperscript{1,5}, Hang Phan\textsuperscript{1,5}, Iain Mathieson\textsuperscript{1}, Zamin Iqbal\textsuperscript{1}, Stephen R F Twigg\textsuperscript{2}, WGS500 Consortium\textsuperscript{3}, Andrew O M Wilkie\textsuperscript{2}, Gil McVean\textsuperscript{1,4} & Gerton Lunter\textsuperscript{1}
Indel variant analysis of short-read sequencing data with Scalpel

Han Fang¹⁻³, Ewa A Bergmann⁴, Kanika Arora⁴, Vladimir Vacic⁴, Michael C Zody⁴, Ivan Iossifov¹, Jason A O’Rawe²,³, Yiyang Wu²,³, Laura T Jimenez Barron²,⁵, Julie Rosenbaum¹, Michael Ronemus¹, Yoon-ha Lee¹, Zihua Wang¹, Esra Dikoglu², Vaidehi Jobanputra²,⁶, Gholson J Lyon²,³, Michael Wigler¹, Michael C Schatz¹,⁷ & Giuseppe Narzisi¹,⁴
<table>
<thead>
<tr>
<th></th>
<th>REF</th>
<th></th>
<th>ALT1</th>
<th></th>
<th>ALT2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>...</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Extract reads

Construct de Bruijn graph and resolve repeats

Enumerate haplotype paths

Detect mutations

Deletion

Insertion

An example deletion

...A C T G T A G C T...
structural variants

DELLY: structural variant discovery by integrated paired-end and split-read analysis

Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes and Jan O. Korbel

1European Molecular Biology Laboratory (EMBL), Genome Biology, Meyerhofstr. 1, 69117 Heidelberg, Germany and
2EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
and 3European Molecular Biology Laboratory (EMBL), Core Facilities and Services, Meyerhofstr. 1, 69117 Heidelberg, Germany

main idea: first-pass alignment to find potential structural variants, then modification of the reference genome and realignment
Fig. 4. The build-up of the split-read alignment reference depends on the type of paired-end call. For tandem duplications, inversions and translocations, we modify the reference in such a way that a standard ‘deletion-type’ split-read alignment can be carried out.
main idea: intelligent heuristics for aligning, dumping reads that align perfectly, and looking for known structural variant signatures
**Figure 1** The LUMPY framework for integrating multiple structural variation signals. (A) A scenario in which LUMPY integrates three different sequence alignment signals (read-pair, split-read and read-depth) from a genome single sample. Additionally, sites of known variants are provided to LUMPY as prior knowledge in order to improve sensitivity. (B) A single signal type (in this case, read-pair) that is integrated from three different genome samples. We present these as example scenarios and emphasize that multi-signal and multi-sample workflows are not mutually exclusive. CNV, copy number variation.
assessing variant callers

sole conclusion that all of the assessments agree on: variant callers do not agree with each other.

In fact, a recent publication noted that taking the intersection of 7 variant callers on one dataset left “a modest 1% of variants originally called.”
assessing variant callers

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 Hu³, Paul Bodily⁵, Lifeng Tian⁶, Hakon Hakonarson⁶, W Evan Johnson⁷, Zhi Wei⁴, Kai Wang⁸,⁹* and Gholson J Lyon¹,²,⁹*
concordance of novel SNVs
indel concordance
assessing variant callers: gold standard datasets

pedigrees: sidestep artifacts by gathering family information

Genome In A Bottle: sidestep artifacts by sequencing incredibly deeply, from dozens of platforms and chemistries, and validating all variants
A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree

Michael A. Eberle,¹ Epameinondas Fritzilas,² Peter Krusche,² Morten Källberg,² Benjamin L. Moore,² Mitchell A. Bekritsky,² Zamin Iqbal,³ Han-Yu Chuang,¹ Sean J. Humphray,² Aaron L. Halpern,¹ Semyon Kruglyak,¹ Elliott H. Margulies,¹ Gil McVean,³,⁴ and David R. Bentley²

¹Illumina Incorporated, San Diego, California 92122, USA; ²Illumina Limited, Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex, CB10 1XL, United Kingdom; ³Wellcome Trust Centre for Human Genetics, Oxford, OX3 7BN, United Kingdom; ⁴Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, OX3 7BN, United Kingdom
Figure 1. Pedigree of the family sequenced for this study (CEPH pedigree 1463). The Coriell ID for each sample is defined by adding the prefix NA128 to each numbered individual; e.g., 77 = NA12877. Samples filled with dark orange are used in this analysis but the founder generations (blue) were also sequenced and used as further validation of the haplotypes generated in this study. The trio, 91-92-78, was also sequenced during Phase I of the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010).
assessing variant callers: genome in a bottle

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

Justin M Zook¹, Brad Chapman², Jason Wang³, David Mittelman³,⁴, Oliver Hofmann², Winston Hide² & Marc Salit¹
Table 1 Description of data sets and their processing to produce bam files for our integration methods

<table>
<thead>
<tr>
<th>Source</th>
<th>Platform</th>
<th>Mapping algorithm</th>
<th>Coverage</th>
<th>Read length</th>
<th>Genome/exome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Genomes</td>
<td>Illumina Gallx</td>
<td>BWA</td>
<td>39</td>
<td>44</td>
<td>Genome</td>
</tr>
<tr>
<td>1000 Genomes</td>
<td>Illumina Gallx</td>
<td>BWA</td>
<td>30</td>
<td>54</td>
<td>Exome</td>
</tr>
<tr>
<td>1000 Genomes</td>
<td>454</td>
<td>Ssaha2</td>
<td>16</td>
<td>239</td>
<td>Genome</td>
</tr>
<tr>
<td>X Prize</td>
<td>Illumina HiSeq</td>
<td>Novoalign</td>
<td>37</td>
<td>100</td>
<td>Genome</td>
</tr>
<tr>
<td>X Prize</td>
<td>SOLiD 4</td>
<td>Lifescope</td>
<td>24</td>
<td>40</td>
<td>Genome</td>
</tr>
<tr>
<td>Complete Genomics</td>
<td>Complete Genomics</td>
<td>CGTools 2.0</td>
<td>73</td>
<td>33</td>
<td>Genome</td>
</tr>
<tr>
<td>Broad</td>
<td>Illumina HiSeq</td>
<td>BWA</td>
<td>68</td>
<td>93</td>
<td>Genome</td>
</tr>
<tr>
<td>Broad</td>
<td>Illumina HiSeq</td>
<td>BWA</td>
<td>66</td>
<td>66</td>
<td>Exome</td>
</tr>
<tr>
<td>Illumina</td>
<td>Illumina HiSeq</td>
<td>CASAVA</td>
<td>80</td>
<td>100</td>
<td>Genome</td>
</tr>
<tr>
<td>Illumina</td>
<td>Illumina HiSeq – PCR-free</td>
<td>BWA</td>
<td>56</td>
<td>99</td>
<td>Genome</td>
</tr>
<tr>
<td>Illumina</td>
<td>Illumina HiSeq – PCR-free</td>
<td>BWA-MEM</td>
<td>190</td>
<td>99</td>
<td>Genome</td>
</tr>
<tr>
<td>Life Technologies</td>
<td>Ion Torrent</td>
<td>tmap</td>
<td>80</td>
<td>237</td>
<td>Exome</td>
</tr>
<tr>
<td>Illumina</td>
<td>Illumina HiSeq – PCR-free</td>
<td>BWA-MEM</td>
<td>60</td>
<td>250</td>
<td>Genome</td>
</tr>
<tr>
<td>Life Technologies</td>
<td>Ion Torrent</td>
<td>tmap</td>
<td>12</td>
<td>200</td>
<td>Genome</td>
</tr>
</tbody>
</table>

databases of variation

dbSNP
ClinVar
COSMIC
somatic variant calling

big problem: cancers are highly heterogeneous. If we take a tissue sample, extract DNA, and sequence it, we are looking at a bunch of subclones!

One small area of the tumor may have a variant not present in the rest of the tumor. This variant may then show up at 10% allele frequency, but it’s real!

Cancers also have aneuploidy, extensive copy number variation, and loss of heterozygosity.
In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data

Lei Cai¹,², Wei Yuan¹, Zhou Zhang¹,³, Lin He¹,⁴ & Kuo-Chen Chou²,⁵
somatic variant calling
Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data

Anne Bruun Krøigård¹,²*, Mads Thomassen¹,², Anne-Vibeke Lænkhølm³, Torben A. Kruse¹,², Martin Jakob Larsen¹,²

¹ Department of Clinical Genetics, Odense University Hospital, Sdr. Boulevard 29, 5000, Odense, Denmark,
² Human Genetics, Institute of Clinical Research, University of Southern Denmark, Winsløwparken 19, 5000,
Odense, Denmark,
³ Department of Pathology, Slagelse Hospital, Ingemannsvej 18, 4200, Slagelse, Denmark

* anne.kroeigaard@rsyd.dk
somatic variant calling

<table>
<thead>
<tr>
<th>SNVs</th>
<th>EBCall</th>
<th>Mutect</th>
<th>Seurat</th>
<th>Shimmer</th>
<th>Somatic Sniper</th>
<th>Strelka</th>
<th>Varscan 2</th>
<th>Virmid</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBCall</td>
<td>1.00</td>
<td>0.51</td>
<td>0.60</td>
<td>0.36</td>
<td>0.50</td>
<td>0.43</td>
<td>0.57</td>
<td>0.48</td>
</tr>
<tr>
<td>Mutect</td>
<td>0.20</td>
<td>1.00</td>
<td>0.47</td>
<td>0.18</td>
<td>0.25</td>
<td>0.26</td>
<td>0.30</td>
<td>0.33</td>
</tr>
<tr>
<td>Seurat</td>
<td>0.08</td>
<td>0.16</td>
<td>1.00</td>
<td>0.10</td>
<td>0.26</td>
<td>0.09</td>
<td>0.19</td>
<td>0.12</td>
</tr>
<tr>
<td>Shimmer</td>
<td>0.32</td>
<td>0.41</td>
<td>0.66</td>
<td>1.00</td>
<td>0.39</td>
<td>0.39</td>
<td>0.46</td>
<td>0.40</td>
</tr>
<tr>
<td>Somatic Sniper</td>
<td>0.08</td>
<td>0.10</td>
<td>0.31</td>
<td>0.07</td>
<td>1.00</td>
<td>0.06</td>
<td>0.20</td>
<td>0.09</td>
</tr>
<tr>
<td>Strelka</td>
<td>0.52</td>
<td>0.80</td>
<td>0.79</td>
<td>0.53</td>
<td>0.52</td>
<td>1.00</td>
<td>0.69</td>
<td>0.68</td>
</tr>
<tr>
<td>Varscan</td>
<td>0.21</td>
<td>0.28</td>
<td>0.52</td>
<td>0.20</td>
<td>0.50</td>
<td>0.21</td>
<td>1.00</td>
<td>0.24</td>
</tr>
<tr>
<td>Virmid</td>
<td>0.37</td>
<td>0.64</td>
<td>0.72</td>
<td>0.35</td>
<td>0.47</td>
<td>0.43</td>
<td>0.51</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indels</th>
<th>EBCall</th>
<th>Seurat</th>
<th>Indelocator</th>
<th>Strelka</th>
<th>Varscan 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBCall</td>
<td>1.00</td>
<td>0.14</td>
<td>0.14</td>
<td>0.06</td>
<td>0.20</td>
</tr>
<tr>
<td>Seurat</td>
<td>0.02</td>
<td>1.00</td>
<td>0.18</td>
<td>0.01</td>
<td>0.13</td>
</tr>
<tr>
<td>Indelocator</td>
<td>0.01</td>
<td>0.13</td>
<td>1.00</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td>Strelka</td>
<td>0.38</td>
<td>0.59</td>
<td>0.62</td>
<td>1.00</td>
<td>0.65</td>
</tr>
<tr>
<td>Varscan</td>
<td>0.08</td>
<td>0.39</td>
<td>0.27</td>
<td>0.04</td>
<td>1.00</td>
</tr>
</tbody>
</table>
somatic variant calling