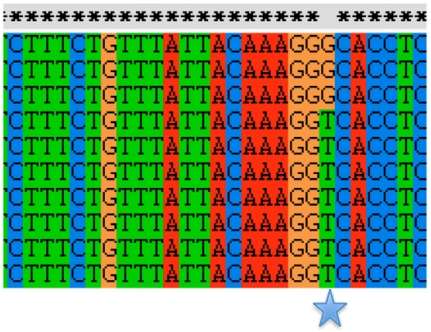
Variant Calling

- SNP = single nucleotide polymorphism
- SNV = single nucleotide variant
- Indel = insertion/deletion

- Examine the alignments of reads and look for differences between the reference and the individual(s) being sequenced

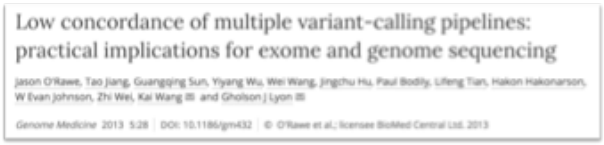


Feng et al 2014. Novel Single Nucleotide Polymorphisms of the Insulin-Like Growth Factor-I Gene and Their Associations with Growth Traits in Common Carp (Cyprinus carpio L.)

Objective: From reads to variants

Your mileage may vary

- Different decisions about how to align reads and identify variants can yield very different results



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

Jason O'Rawe, Tao Jiang, Guangqing Sun, Yiyang Wu, Wei Wang, Jingchu Hu, Paul Bodily, Lifeng Tian, Hakon Hakonarson, W Evan Johnson, Zhi Wei, Kai Wang and Gholson J Lyon

Genome Medicine 2013, 5:28 | DOI: 10.1186/gm432

- 5 pipelines
- "SNP concordance between five Illumina pipelines across all 15 exomes was 57.4%, while 0.5 to 5.1% of variants were called as unique to each pipeline. Indel concordance was only 26.8% between three indel-calling pipelines"

Variant Calling Difficulties

- Difficulties:
  - Cloning process (PCR) artifacts
  - Errors in the sequencing reads
  - Incorrect mapping
  - Errors in the reference genome
- Heng Li, developer of BWA, looked at major sources of errors in variant calls*:
  - erroneous realignment in low-complexity (tandem) regions
  - the incomplete reference genome with respect to the sample

* Li 2014 Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics.

Indel

```
coor      12345678901234    567890123456789012345678
ref       aggttttataaaac----aattaagtctacagagcaacta
sample    aggttttataaaacAAATaattaagtctacagagcaacta
read1     aggttttataaaac****aaAtaa
read2      ggttttataaaac****aaAtaaTt
read3          ttataaaacAAATaattaagtctaca
read4              CaaaT****aattaagtctacagagcaac
read5                aaT****aattaagtctacagagcaact
read6                  T****aattaagtctacagagcaacta
```

Can be difficult to decide where the best alignment actually is.

Indels are far more problematic to call than SNPs.

*https://bioinf.comav.upv.es/courses/sequence_analysis/snp_calling.html

Many tool options: review literature!

Article | OPEN

Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data

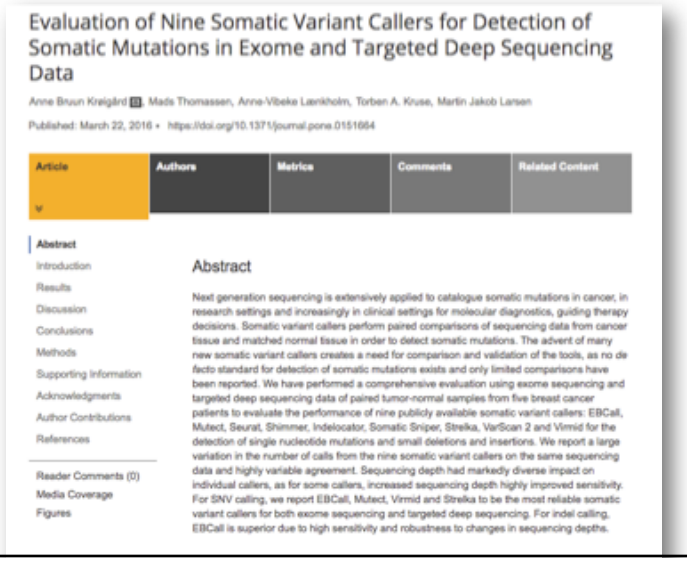
Sarah Sandmann, Aniek O. de Graaf, Mohsen Karimi, Bert A. van der Reijden, Eva Hellström-Lindberg, Joop H. Jansen & Martin Dugas

Scientific Reports **7**, Article number: 43169 (2017)
doi:10.1038/srep43169
Download Citation

Cancer genetics | Cancer genomics | DNA sequencing

Received: 27 October 2016
Accepted: 20 January 2017
Published online: 24 February 2017

Many tool options: review literature!



Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data

Anne Bruun Kreigård, Mads Thomassen, Anne-Vibeke Lænkholm, Torben A. Kruse, Martin Jakob Larsen

Published: March 22, 2016 • https://doi.org/10.1371/journal.pone.0151664

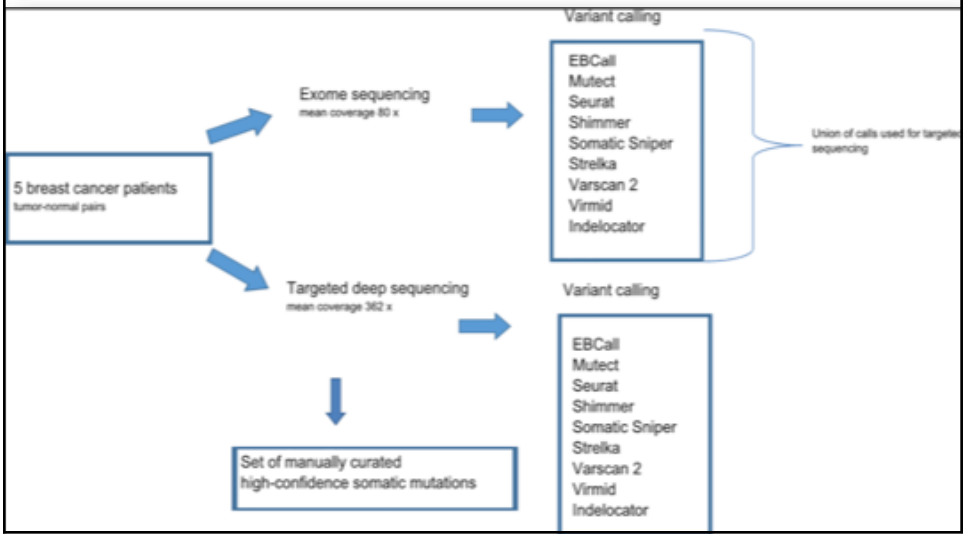
Article | Authors | Metrics | Comments | Related Content

Abstract
Introduction
Results
Discussion
Conclusions
Methods
Supporting Information
Acknowledgments
Author Contributions
References
Reader Comments (0)
Media Coverage
Figures

Abstract

Next generation sequencing is extensively applied to catalogue somatic mutations in cancer, in research settings and increasingly in clinical settings for molecular diagnostics, guiding therapy decisions. Somatic variant callers perform paired comparisons of sequencing data from cancer tissue and matched normal tissue in order to detect somatic mutations. The advent of many new somatic variant callers creates a need for comparison and validation of the tools, as no *de facto* standard for detection of somatic mutations exists and only limited comparisons have been reported. We have performed a comprehensive evaluation using exome sequencing and targeted deep sequencing data of paired tumor-normal samples from five breast cancer patients to evaluate the performance of nine publicly available somatic variant callers: EBCall, Mutect, Seurat, Shimmer, Indelocator, Somatic Sniper, Strelka, VarScan 2 and Virmid for the detection of single nucleotide mutations and small deletions and insertions. We report a large variation in the number of calls from the nine somatic variant callers on the same sequencing data and highly variable agreement. Sequencing depth had markedly diverse impact on individual callers, as for some callers, increased sequencing depth highly improved sensitivity. For SNV calling, we report EBCall, Mutect, Virmid and Strelka to be the most reliable somatic variant callers for both exome sequencing and targeted deep sequencing. For indel calling, EBCall is superior due to high sensitivity and robustness to changes in sequencing depths.

Many tool options: review literature!



Tool Recommendations

	Transcriptome	Exome Capture	GBS	RNA-Seq	Somatic
Aligner	Bowtie2/HISAT2	BWA	HISAT2	HISAT2	BWA
SNP detection	Freebayes	GATK	GATK (local re-alignment)	GATK (local re-alignment)	MuTect Platypus

	Rare alleles	RAD-Seq (mod) - model	Ploidy (Pooled)	RAD-Seq (low) – non-model
Aligner	HISAT2	BWA/HISAT2	HISAT2	HISAT2
SNP detection	VarDict FreeBayes	Stacks/PyRAD	Freebayes	dDocent

Filtering

- Genotype Likelihood calculation should render this unnecessary, but alas, real data sets often benefit from additional filtering
- Hard cut off on depth
  - How many reads do you need to sample to confidently call a SNP? (For a diploid?)
  - > 20X = very good
  - 5-20X = okay
  - < 5X = missing many heterozygous calls
- High coverage – can indicate a duplicated region in the genome
- Highly variable region – can also indicate a duplicated region (take into account HWE)
- Low complexity regions

Last step (?): Imputation

If one site has low coverage but is tightly linked to other sites with high coverage, the information can be "imputed"

- Utilize LD across loci (i.e. known haplotypes)
- Depends on haplotype estimation (phasing)
- Many sogware options
  - BEAGLE
  - Impute2
  - MaCH

Rescue missing data!

Phasing

Heterozygous genotypes at 3 sites

AC TG AT

The 4 possible consistent pairs of haplotypes

ATT	ATA	AGT	AGA
CGA	CGT	CTA	CTT

Novel SNPs/Indels

- What is the effect of this variant?
- Is the variant inside a gene?
  - Does it change an amino acid?
  - Does it create a stop codon?
  - Does it shift the open reading frame?
- Software:
  - **SnpEff/SnpSift**
  - Annovar
  - Variant Effect Predictor

SAMtools, BCFtools, HTSLib

- http://www.htslib.org/
- Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

1. Samtools
   Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format
2. BCFtools
   Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarizing SNP and short indel sequence variants
3. HTSlib
   A C library for reading/writing high-throughput sequencing data

- Example workflow:
- http://www.htslib.org/workflow/#mapping_to_variant

Mpileup format

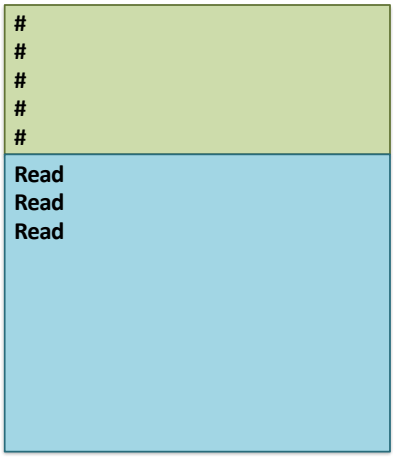
- Mpileup format
- For each base in the reference
  - reference base
  - the number of reads covering the site
  - read bases
  - base qualities
  - alignment mapping qualities
- You will rarely ever use this format, just need to generate it and pass it straight to the SNP caller

Overview

Samtools	• Works with SAM/BAM files • Produces mpileup	Alignment Data
Bcftools	• Call SNPs from mpileup • Works with VCF/BCF files	Variant Data

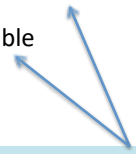
VCF

- Variant Call Format
- Official spec: http://samtools.github.io/hts-specs/VCFv4.2.pdf
- Header lines starting with # signs
- Lines with variants agerward



VCF (cont)

- Tab delimited fields
  - Chromosome
  - Location
  - ID (if this is a named variant)
  - Reference sequence
  - Alternate sequence
  - Quality score
  - Filter (true/false – whether or not it passed filtering)
  - Info – lots of additional info such as CIGAR string, depth across different samples, etc.
  - Columns follow for each genotype if available
- BCF is the compressed binary format
  - SAM <-> BAM
  - VCF <-> BCF



Quite variable depending on software used to call SNPs

VCF Example

#CHROM	20
POS	14370
ID	rs6054257
REF	G
ALT	A
QUAL	29
FILTER	PASS
INFO	NS=3;DP=14;AF=0.5;DB;H2
FORMAT	GT:GQ:DP:HQ
NA00001	0|0:48:1:51,51
NA00002	1|0:48:8:51,51
NA00003	1/1:43:5:.,.

Standard

VCF Example

#CHROM	20
POS	14370
ID	rs6054257
REF	G
ALT	A
QUAL	29
FILTER	PASS
INFO	NS=3;DP=14;AF=0.5;DB;H2
FORMAT	GT:GQ:DP:HQ
NA00001	0|0:48:1:51,51
NA00002	1|0:48:8:51,51
NA00003	1/1:43:5:.,.

Info field gives general information about this position across all samples. The codes are defined in the header of the file, can vary.

NS = Number of samples with data

VCF Example

#CHROM	20
POS	14370
ID	rs6054257
REF	G
ALT	A
QUAL	29
FILTER	PASS
INFO	NS=3;DP=14;AF=0.5;DB;H2
FORMAT	GT:GQ:DP:HQ
NA00001	0|0:48:1:51,51
NA00002	1|0:48:8:51,51
NA00003	1/1:43:5:.,.

DP = combined depth across samples

VCF Example

#CHROM	20
POS	14370
ID	rs6054257
REF	G
ALT	A
QUAL	29
FILTER	PASS
INFO	NS=3;DP=14;AF=0.5;DB;H2
FORMAT	GT:GQ:DP:HQ
NA00001	0|0:48:1:51,51
NA00002	1|0:48:8:51,51
NA00003	1/1:43:5:.,.

AF = allele frequence for alternate allele

VCF Example

#CHROM	20
POS	14370
ID	rs6054257
REF	G
ALT	A
QUAL	29
FILTER	PASS
INFO	NS=3;DP=14;AF=0.5;DB;H2
FORMAT	GT:GQ:DP:HQ
NA00001	0|0:48:1:51,51
NA00002	1|0:48:8:51,51
NA00003	1/1:43:5:.,.

DB = dbSNP membership

H2 = HapMap2 membership

VCF Example

#CHROM	20
POS	14370
ID	rs6054257
REF	G
ALT	A
QUAL	29
FILTER	PASS
INFO	NS=3;DP=14;AF=0.5;DB;H2
FORMAT	GT:GQ:DP:HQ
NA00001	0|0:48:1:51,51
NA00002	1|0:48:8:51,51
NA00003	1/1:43:5:.,.

Format field

Explains the format used for information about each sample.

Variable by SNP caller.

VCF Example

#CHROM	20
POS	14370
ID	rs6054257
REF	G
ALT	A
QUAL	29
FILTER	PASS
INFO	NS=3;DP=14;AF=0.5;DB;H2
FORMAT	GT:GQ:DP:HQ
NA00001	0|0:48:1:51,51
NA00002	1|0:48:8:51,51
NA00003	1/1:43:5:.,.

GT = genotype
0/0 0/1 1/1 1/2

The / is replaced with a | if the alleles are phased

0|0 0|1 1|1

VCF Example

#CHROM	20
POS	14370
ID	rs6054257
REF	G
ALT	A
QUAL	29
FILTER	PASS
INFO	NS=3;DP=14;AF=0.5;DB;H2
FORMAT	GT:GQ:DP:HQ
NA00001	0|0:48:1:51,51
NA00002	1|0:48:8:51,51
NA00003	1/1:43:5:.,.

GQ = Genotype Quality

Phred-scaled confidence in genotype call

VCF Example

#CHROM	20
POS	14370
ID	rs6054257
REF	G
ALT	A
QUAL	29
FILTER	PASS
INFO	NS=3;DP=14;AF=0.5;DB;H2
FORMAT	GT:GQ:DP HQ
NA00001	0|0:48:1 51,51
NA00002	1|0:48:8 51,51
NA00003	1/1:43:5 .,.

DP = Read Depth

# of reads from this location for this individual

VCF Example

#CHROM	20
POS	14370
ID	rs6054257
REF	G
ALT	A
QUAL	29
FILTER	PASS
INFO	NS=3;DP=14;AF=0.5;DB;H2
FORMAT	GT:GQ:DP:HQ
NA00001	0|0:48:1:51,51
NA00002	1|0:48:8:51,51
NA00003	1/1:43:5:.,.

HQ = Haplotype Quality

Only for phased loci, added by phasing software

Flexible info fields

- **SNPEff** has standardized the addition of variant effect information
- Additional tag ANN in the info field

Chromosome 1411926 . G C 228.0 PASS DP=97;VDB=1.42407e-36;SGB=-0.693147;MQSB=1 ;MQ0F=0;AC=2;AN=2;DP4=0,0,45,41;MQ=60;ANN= C|missense_variant|MODERATE|ttcA|b1344| transcript|AAC74426.1...