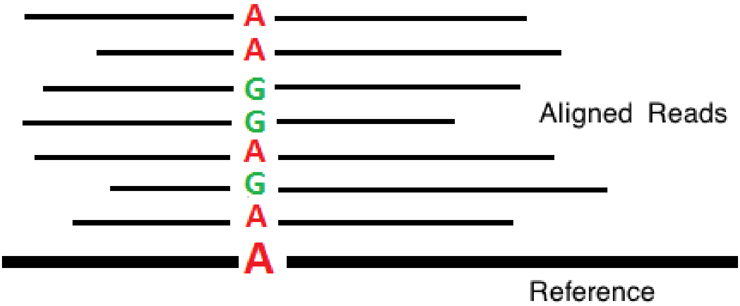
Validated variant calling: structural, joint, somatic

Brad Chapman
Bioinformatics Core, Harvard School of Public Health
https://github.com/chapmanb/bcbio-nextgen
http://j.mp/bcbiolinks

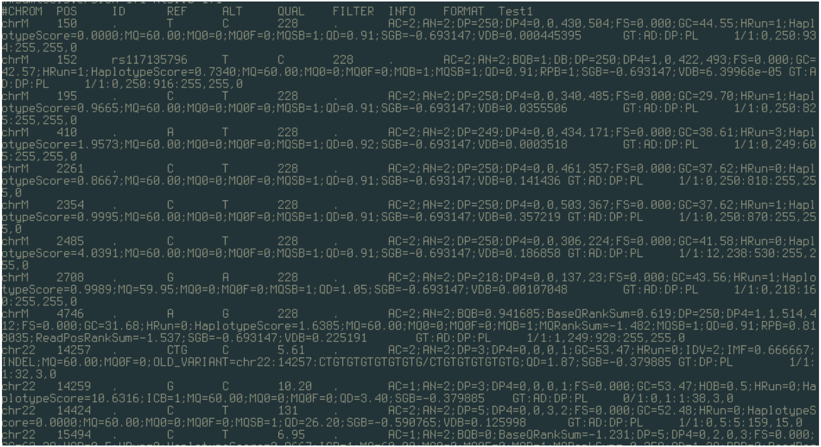
15 October 2014

Variant calling



http://en.wikipedia.org/wiki/SNV_calling_from_NGS_data

Ok, VCFs



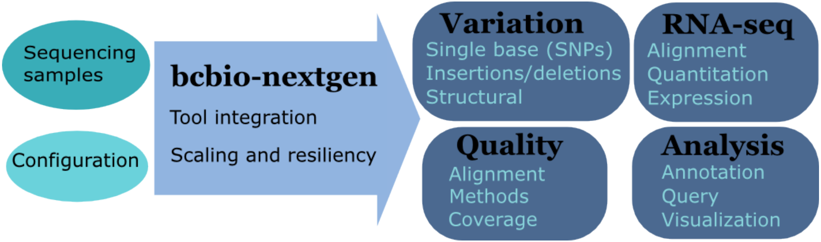
```
#CHROM  POS     ID        REF       ALT       QUAL      FILTER  INFO      FORMAT  Test1
chrM    150     .         T         C         228       .       AC=2;AN=2;DP=250;DP4=0,0,430,504;FS=0.000;GC=44.55;HRun=1;Hapl
otypeScore=0.0000;MQ=60.00;MQ0=0;MQ0F=0;MQSB=1;QD=0.91;SGB=-0.693147;VDB=0.000445395      GT:AD:DP:PL     1/1:0,250:93
4:255,255,0
chrM    152     rs117135796     T         C         228     .       AC=2;AN=2;BQB=1;DB;DP=250;DP4=1,0,422,493;FS=0.000;GC=
42.57;HRun=1;HaplotypeScore=0.7340;MQ=60.00;MQ0=0;MQ0F=0;MQB=1;MQSB=1;QD=0.91;RPB=1;SGB=-0.693147;VDB=6.39968e-05 GT:A
D:DP:PL     1/1:0,250:916:255,255,0
chrM    195     .         C         T         228       .       AC=2;AN=2;DP=250;DP4=0,0,340,485;FS=0.000;GC=29.70;HRun=1;Hapl
otypeScore=0.9665;MQ=60.00;MQ0=0;MQ0F=0;MQSB=1;QD=0.91;SGB=-0.693147;VDB=0.0355506        GT:AD:DP:PL     1/1:0,250:82
5:255,255,0
chrM    410     .         A         T         228       .       AC=2;AN=2;DP=249;DP4=0,0,434,171;FS=0.000;GC=38.61;HRun=3;Hapl
otypeScore=1.9573;MQ=60.00;MQ0=0;MQ0F=0;MQSB=1;QD=0.92;SGB=-0.693147;VDB=0.0003518        GT:AD:DP:PL     1/1:0,249:60
5:255,255,0
chrM    2261    .         C         T         228       .       AC=2;AN=2;DP=250;DP4=0,0,461,357;FS=0.000;GC=37.62;HRun=0;Hapl
otypeScore=0.8667;MQ=60.00;MQ0=0;MQ0F=0;MQSB=1;QD=0.91;SGB=-0.693147;VDB=0.141436 GT:AD:DP:PL     1/1:0,250:818:255,25
5,0
chrM    2354    .         C         T         228       .       AC=2;AN=2;DP=250;DP4=0,0,503,367;FS=0.000;GC=37.62;HRun=1;Hapl
otypeScore=0.9995;MQ=60.00;MQ0=0;MQ0F=0;MQSB=1;QD=0.91;SGB=-0.693147;VDB=0.357219 GT:AD:DP:PL     1/1:0,250:870:255,25
5,0
chrM    2485    .         C         T         228       .       AC=2;AN=2;DP=250;DP4=0,0,306,224;FS=0.000;GC=41.58;HRun=0;Hapl
otypeScore=4.0391;MQ=60.00;MQ0=0;MQ0F=0;MQSB=1;QD=0.91;SGB=-0.693147;VDB=0.186858 GT:AD:DP:PL     1/1:12,238:530:255,2
55,0
chrM    2708    .         G         A         228       .       AC=2;AN=2;DP=218;DP4=0,0,137,23;FS=0.000;GC=43.56;HRun=1;Haplo
typeScore=0.9989;MQ=59.95;MQ0=0;MQ0F=0;MQSB=1;QD=1.05;SGB=-0.693147;VDB=0.00107048        GT:AD:DP:PL     1/1:0,218:16
0:255,255,0
chrM    4746    .         A         G         228       .       AC=2;AN=2;BQB=0.941685;BaseQRankSum=0.619;DP=250;DP4=1,1,514,4
12;FS=0.000;GC=31.68;HRun=0;HaplotypeScore=1.6385;MQ=60.00;MQ0=0;MQ0F=0;MQB=1;MQRankSum=-1.482;MQSB=1;QD=0.91;RPB=0.81
8035;ReadPosRankSum=-1.537;SGB=-0.693147;VDB=0.225191       GT:AD:DP:PL     1/1:1,249:928:255,255,0
chr22   14257   .         CTG       C         5.61      .       AC=2;AN=2;DP=3;DP4=0,0,0,1;GC=53.47;HRun=0;IDV=2;IMF=0.666667;
INDEL;MQ=60.00;MQ0F=0;OLD_VARIANT=chr22:14257:CTGTGTGTGTGTGTG/CTGTGTGTGTGTG;QD=1.87;SGB=-0.379885 GT:DP:PL        1/1:
1:32,3,0
chr22   14259   .         G         C         10.20     .       AC=1;AN=2;DP=3;DP4=0,0,0,1;FS=0.000;GC=53.47;HOB=0.5;HRun=0;Ha
plotypeScore=10.6316;ICB=1;MQ=60.00;MQ0=0;MQ0F=0;QD=3.40;SGB=-0.379885    GT:AD:DP:PL     0/1:0,1:1:38,3,0
chr22   14424   .         C         T         131       .       AC=2;AN=2;DP=5;DP4=0,0,3,2;FS=0.000;GC=52.48;HRun=0;HaplotypeS
core=0.0000;MQ=60.00;MQ0=0;MQ0F=0;MQSB=1;QD=26.20;SGB=-0.590765;VDB=0.125998      GT:AD:DP:PL     1/1:0,5:5:159,15,0
chr22   15494   .         C         T         6.95      .       AC=1;AN=2;BQB=0;BaseQRankSum=-1.231;DP=5;DP4=0,2,0,3;FS=0.000;
```

Summary

- What is bcbio?
- Community software development
- Variation validation
- Software support

White box software

Overview



https://github.com/chapmanb/bcbio-nextgen

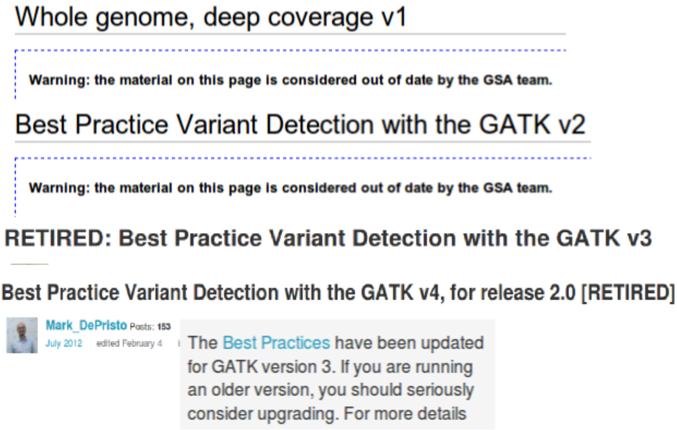
Uses

- Aligners: bwa-mem, novoalign, bowtie2
- Variantion: FreeBayes, GATK, Platypus, MuTecT, scalpel, SnpEff, VEP, GEMINI, Lumpy, Delly
- RNA-seq: Tophat, STAR, cufflinks, HTSeq
- Quality control: fastqc, bamtools, RNA-SeQC
- Manipulation: bedtools, bcftools, biobambam, sambamba, samblaster, samtools, vcflib

Provides

- Community – collected set of expertise
- Tool integration
- Validation – outputs + automated evaluation
- Scaling
- Installation of tools and data

Complex, rapidly changing pipelines



Whole genome, deep coverage v1

Warning: the material on this page is considered out of date by the GSA team.

Best Practice Variant Detection with the GATK v2

Warning: the material on this page is considered out of date by the GSA team.

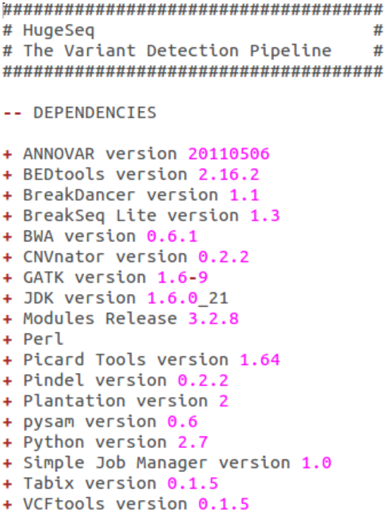
RETIRED: Best Practice Variant Detection with the GATK v3

Best Practice Variant Detection with the GATK v4, for release 2.0 [RETIRED]

Mark_DePristo Posts: 153
July 2012 edited February 4

The Best Practices have been updated for GATK version 3. If you are running an older version, you should seriously consider upgrading. For more details

Large number of specialized dependencies



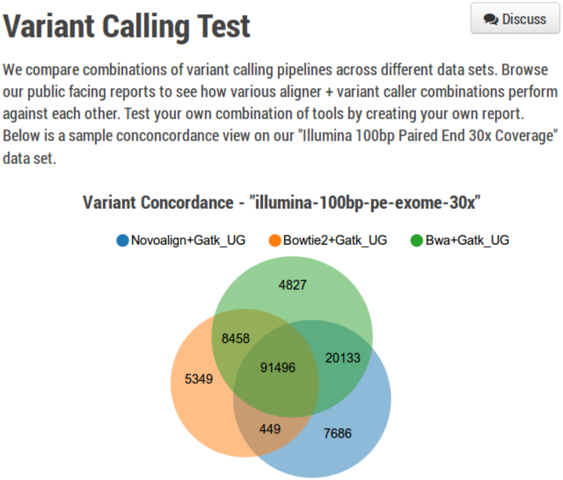
```
#####################################
# HugeSeq                           #
# The Variant Detection Pipeline    #
#####################################

-- DEPENDENCIES

+ ANNOVAR version 20110506
+ BEDtools version 2.16.2
+ BreakDancer version 1.1
+ BreakSeq Lite version 1.3
+ BWA version 0.6.1
+ CNVnator version 0.2.2
+ GATK version 1.6-9
+ JDK version 1.6.0_21
+ Modules Release 3.2.8
+ Perl
+ Picard Tools version 1.64
+ Pindel version 0.2.2
+ Plantation version 2
+ pysam version 0.6
+ Python version 2.7
+ Simple Job Manager version 1.0
+ Tabix version 0.1.5
+ VCFtools version 0.1.5
```

https://github.com/StanfordBioinformatics/HugeSeq

Quality differences between methods



Variant Calling Test

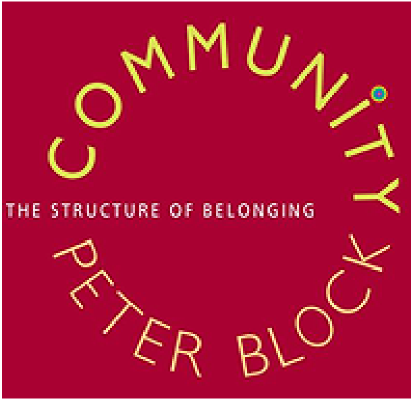
Discuss

We compare combinations of variant calling pipelines across different data sets. Browse our public facing reports to see how various aligner + variant caller combinations perform against each other. Test your own combination of tools by creating your own report. Below is a sample conconcordance view on our "Illumina 100bp Paired End 30x Coverage" data set.

Variant Concordance - "illumina-100bp-pe-exome-30x"

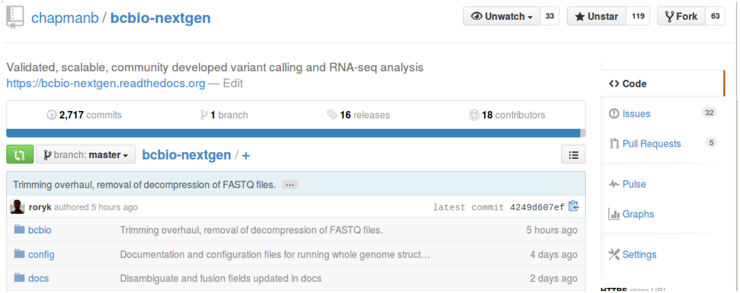
http://www.bioplanet.com/gcat

Solution



http://www.amazon.com/Community-Structure-Belonging-Peter-Block/dp/1605092770

Community: contribution



chapmanb / bcbio-nextgen

Validated, scalable, community developed variant calling and RNA-seq analysis
https://bcbio-nextgen.readthedocs.org — Edit

2,717 commits · 1 branch · 16 releases · 18 contributors

bcbio-nextgen / +

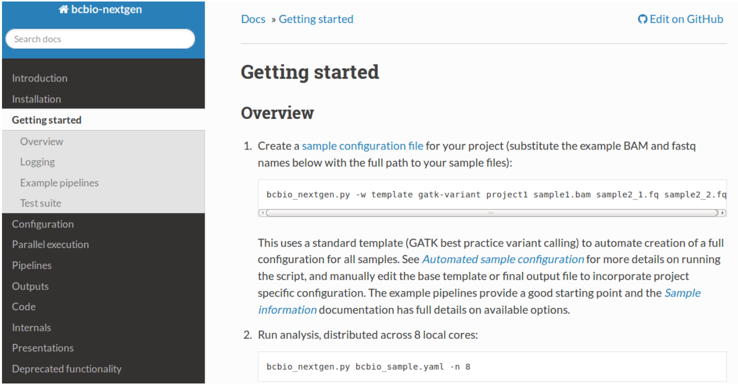
Trimming overhaul, removal of decompression of FASTQ files. ···

roryk authored 5 hours ago — latest commit 4249d607ef

bcbio	Trimming overhaul, removal of decompression of FASTQ files.	5 hours ago
config	Documentation and configuration files for running whole genome struct…	4 days ago
docs	Disambiguate and fusion fields updated in docs	2 days ago

https://github.com/chapmanb/bcbio-nextgen

Community: documentation



Getting started

Overview

1. Create a sample configuration file for your project (substitute the example BAM and fastq names below with the full path to your sample files):

```
bcbio_nextgen.py -w template gatk-variant project1 sample1.bam sample2_1.fq sample2_2.fq
```

   This uses a standard template (GATK best practice variant calling) to automate creation of a full configuration for all samples. See *Automated sample configuration* for more details on running the script, and manually edit the base template or final output file to incorporate project specific configuration. The example pipelines provide a good starting point and the *Sample information* documentation has full details on available options.

2. Run analysis, distributed across 8 local cores:

```
bcbio_nextgen.py bcbio_sample.yaml -n 8
```

https://bcbio-nextgen.readthedocs.org

Validation

Tests for implementation and methods

- Family/population calling
- Structural variations
- Cancer tumor/normal

Reference materials

Genome in a Bottle
Consortium

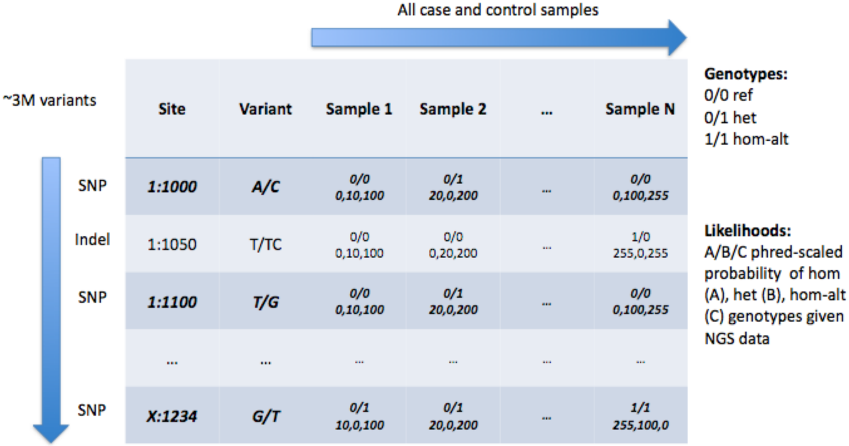
http://www.genomeinabottle.org/

Joint variant calling definitions

- Joint calling
- Squaring off/backfilling
- Pooled calling
- Single sample calling

http://j.mp/bcbiojoint

Squared off VCF



All case and control samples

~3M variants	Site	Variant	Sample 1	Sample 2	...	Sample N
SNP	***1:1000***	***A/C***	***0/0*** ***0,10,100***	***0/1*** ***20,0,200***	...	***0/0*** ***0,100,255***
Indel	1:1050	T/TC	0/0 0,10,100	0/0 0,20,200	...	1/0 255,0,255
SNP	***1:1100***	***T/G***	***0/0*** ***0,10,100***	***0/1*** ***20,0,200***	...	***0/0*** ***0,100,255***
	...	...	...	...	...	...
SNP	***X:1234***	***G/T***	***0/1*** ***10,0,100***	***0/1*** ***20,0,200***	...	***1/1*** ***255,100,0***

Genotypes:
0/0 ref
0/1 het
1/1 hom-alt

Likelihoods:
A/B/C phred-scaled probability of hom (A), het (B), hom-alt (C) genotypes given NGS data

http://gatkforums.broadinstitute.org/discussion/4150/should-i-analyze-my-samples-alone-or-together

Implementation

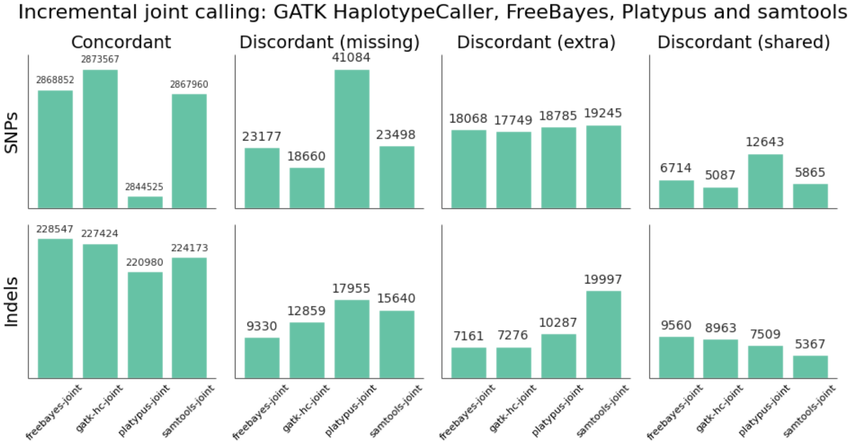
- GATK HaplotypeCaller – gVCFs
- FreeBayes – recalling
- Platypus – recalling
- samtools 1.x – recalling

https://github.com/chapmanb/bcbio.variation.recall

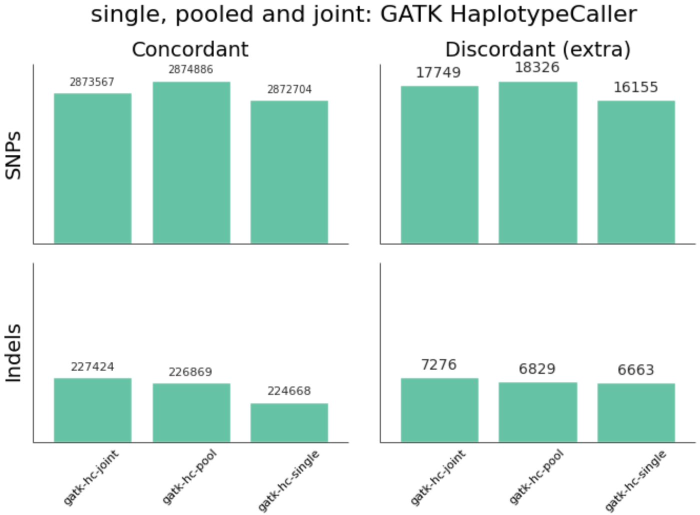
Scaling and analysis flexibility

- Parallelize: call samples individually
- Add single new sample to analysis
- Combine existing populations

Multiple approaches



Joint vs batch vs single



Structural variations

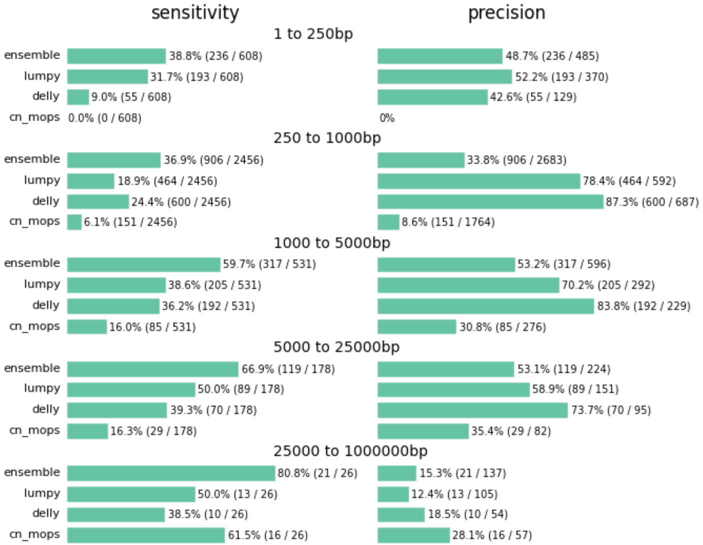
- Goal: identify regions with potential issues
- Rough boundaries for additional analysis
- Ensemble: union of all calls
- Understand sensitivity and precision

`http://j.mp/bcbiosv`

Structural variant callers

- LUMPY https://github.com/arq5x/lumpy-sv
- Delly https://github.com/tobiasrausch/delly
- cn.mops http://www.bioconductor.org/packages/release/bioc/html/cn.mops.html
- CNVkit http://cnvkit.readthedocs.org/
- WHAM https://github.com/jewmanchue/wham

Structural variant evaluation

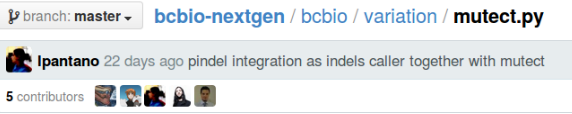


Cancer tumor/normal

- Truth calls: synthetic data from DREAM challenge
- Mixed population of subclones
- Need additional complexity: mixed cellularity

http://j.mp/dreamsyn3

Community built



branch: master

bcbio-nextgen / bcbio / variation / mutect.py

lpantano 22 days ago pindel integration as indels caller together with mutect

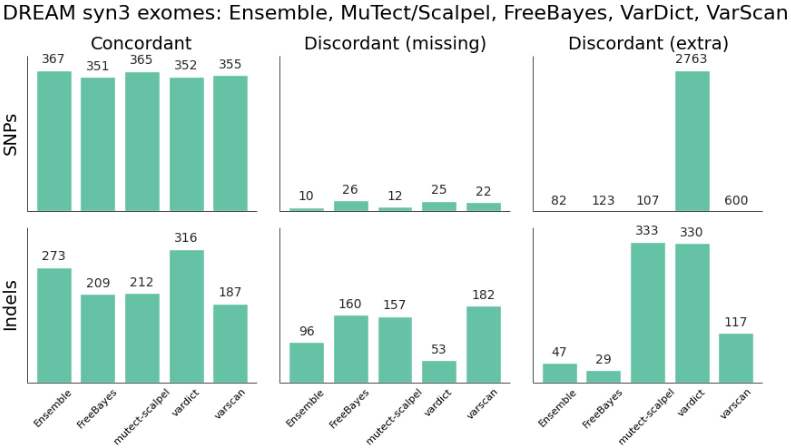
5 contributors

- Luca Beltrame – IRCCS, Italy
- Miika Ahdesmaki – AstraZeneca
- Mario Giovacchini – SciLifeLab, Sweden
- Lorena Pantano – HSPH

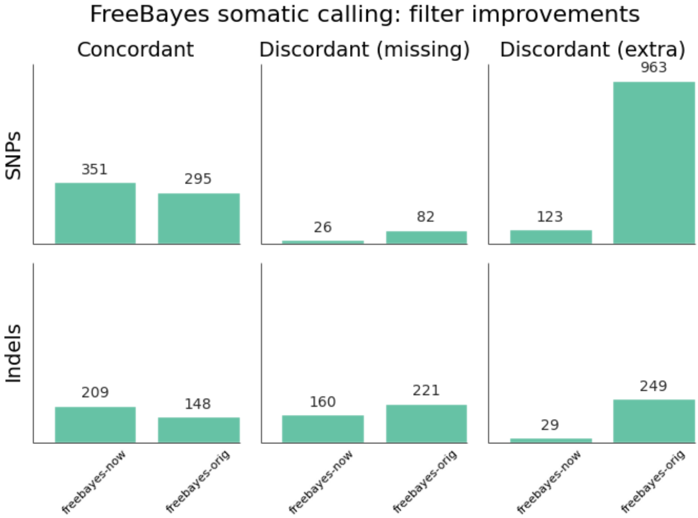
Callers available

- MuTect https://www.broadinstitute.org/cancer/cga/mutect
- FreeBayes https://github.com/ekg/freebayes
- VarScan http://varscan.sourceforge.net/
- VarDict https://github.com/AstraZeneca-NGS/VarDict
- Ensemble

Somatic evaluation



Benefits of improved filtering

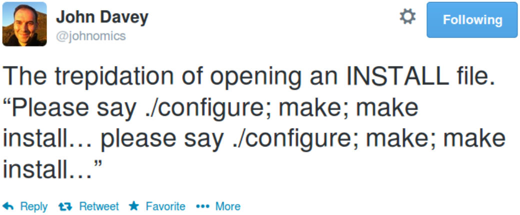


Validation enables scaling

- Little value in realignment when using haplotype aware caller
- Little value in recalibration when using high quality reads
- Streaming de-duplication approaches provide same quality without disk IO

`http://j.mp/bcbioeval2`

Make installation easy



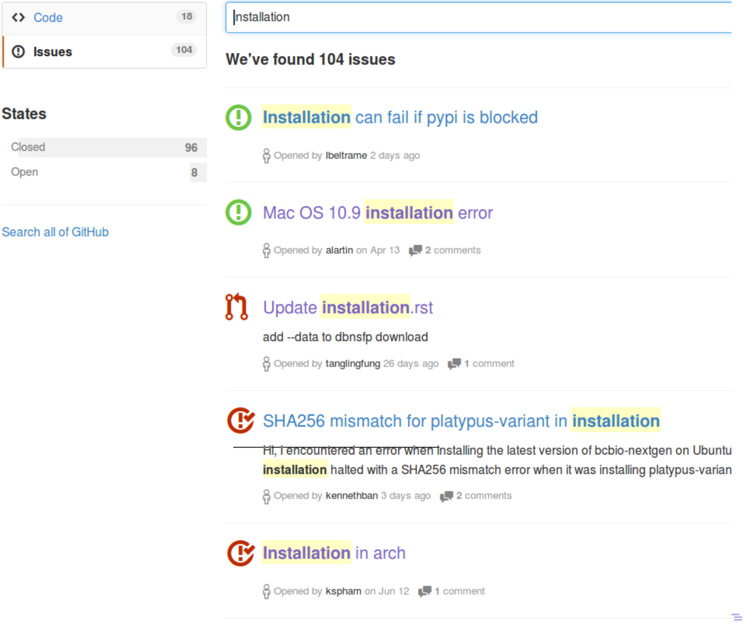
John Davey
@johnomics

The trepidation of opening an INSTALL file. “Please say ./configure; make; make install… please say ./configure; make; make install…”

Automated Install

We made it easy to install a large number of biological tools. Good or bad idea?

Need a consistent support environment



Installation

We've found 104 issues

States

Closed	96
Open	8

Installation can fail if pypi is blocked

Opened by lbeltrame 2 days ago

Mac OS 10.9 **installation** error

Opened by alartin on Apr 13 2 comments

Update **installation**.rst

add --data to dbnsfp download

Opened by tanglingfung 26 days ago 1 comment

SHA256 mismatch for platypus-variant in **installation**

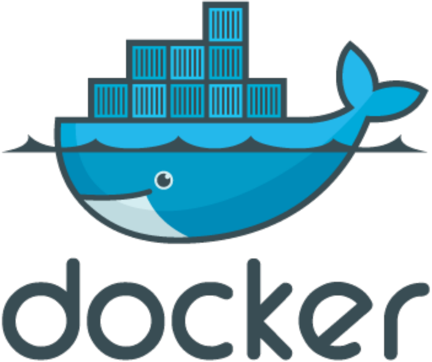
~~Hi, I encountered an error when~~ installing the latest version of bcbio-nextgen on Ubuntu
installation halted with a SHA256 mismatch error when it was installing platypus-varian

Opened by kennethban 3 days ago 2 comments

Installation in arch

Opened by kspham on Jun 12 1 comment

Docker lightweight containers



http://docker.io

Docker benefits

- Fully isolated
- Reproducible – store full environment with analysis (1Gb)
- Improved installation – single download + data

bcbio with Docker

- External Python wrapper
  - Installation
  - Start and run containers
  - Mount external data into containers
  - Parallelize
- All analysis tools inside Docker

https://github.com/chapmanb/bcbio-nextgen-vm

http://j.mp/bcbiodocker

Sustainability

A piece of software is being sustained if people are using it, fixing it, and improving it rather than replacing it.

http://software-carpentry.org/blog/2014/08/sustainability.html

Summary

- What is bcbio?
- Community software development
- Variation validation
- Software support

https://github.com/chapmanb/bcbio-nextgen