

Validated variant calling, clouds and containers

Brad Chapman

Bioinformatics Core, Harvard Chan School

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

<http://bcb.io>

<http://j.mp/bcbiolinks>

23 January 2015

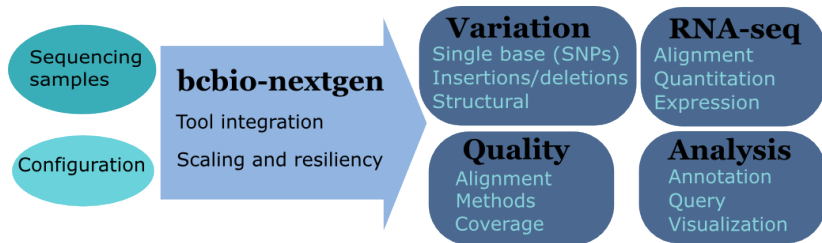
Summary

- What is bcbio?
- Validation
- Variant management
- Docker and AWS

White box software



Overview



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

- Aligners: bwa-mem, novoalign, bowtie2
- Variation: 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

- Community – collected set of expertise
- Validation – outputs + automated evaluation
- Scaling
- Ready to run parallel processing on AWS
- Local 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Warning: the material on this page is considered out of date by the GSA team.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Warning: the material on this page is considered out of date by the GSA team.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>

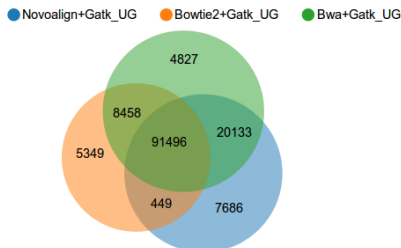
Quality differences between methods

Variant Calling Test

Discuss

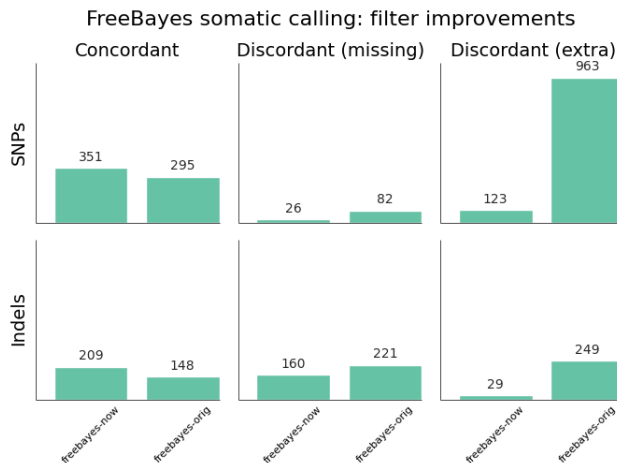
We compare combinations of variant calling pipelines across different data sets. Browse our public facing reports to see how various aligner + variant caller combinations perform against each other. Test your own combination of tools by creating your own report. Below is a sample concordance view on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

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



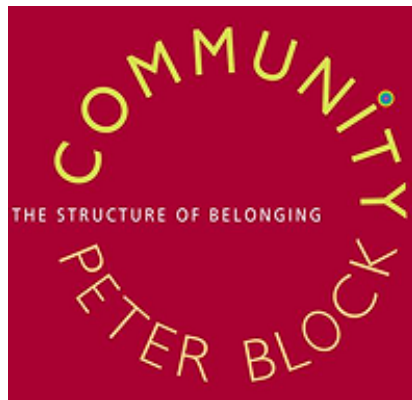
<http://www.bioplanet.com/gcat>

Benefits of improved filtering



<http://j.mp/cancervalpre>

Solution



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

Community: contribution

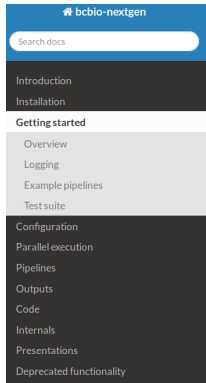
The screenshot shows the GitHub repository page for **chapmanb / bcbio-nextgen**. At the top, there are buttons for **Unwatch** (33), **Unstar** (119), and **Fork** (63). The repository description is "Validated, scalable, community developed variant calling and RNA-seq analysis" with a link to <https://bcbio-nextgen.readthedocs.org> and an **Edit** button. Below this, statistics show **2,717 commits**, **1 branch**, **16 releases**, and **18 contributors**. A green button indicates the current branch is **master**. The main content area shows a commit titled "Trimming overhaul, removal of decompression of FASTQ files." by user **roryk**, authored 5 hours ago. Below the commit message is a table of files changed in the commit:

File	Change	Time
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

On the right sidebar, there are links for **Code**, **Issues** (32), **Pull Requests** (5), **Pulse**, **Graphs**, and **Settings**.

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

Community: documentation



Docs » Getting started

[Edit on GitHub](#)

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>

Tests for implementation and methods

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



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

~3M variants

All case and control samples

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

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

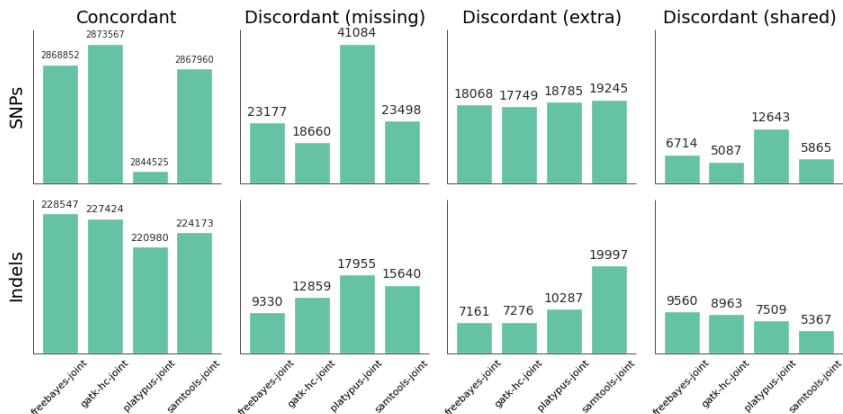
<https://github.com/chapmanb/bcbio.variation.recall>

Scaling and analysis flexibility

- Parallelize: call samples individually
- Add single new sample to analysis
- Combine existing populations
- Inform calls based on previously known variants

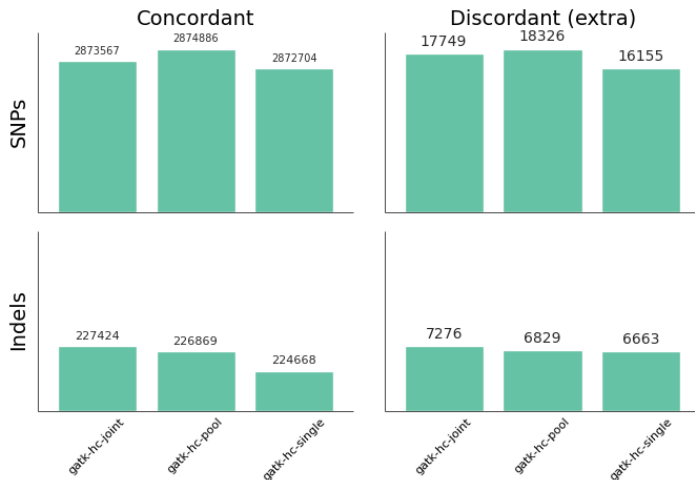
Multiple approaches work well

Incremental joint calling: GATK HaplotypeCaller, FreeBayes, Platypus and samtools



Joint vs batch vs single

single, pooled and joint: GATK HaplotypeCaller



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>

- Coverage: summarize what you can't assess
- Structural: large, complex rearrangements

Sequencing Report: Coverage

Samples included	141-1-2A	141-2-1U	141-2-2U
------------------	----------	----------	----------

Key metrics for dbCMMS v1.0

Sample Id	Cutoff	Avg. Coverage	Avg. Completeness [%]	Diagnostic Yield [%]
141-1-2A	10	143.2874	99.4854	92.249
141-2-1U	10	210.3256	99.667	94.028
141-2-2U	10	193.0433	99.6035	92.5032

<http://www.chanjo.co>

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>

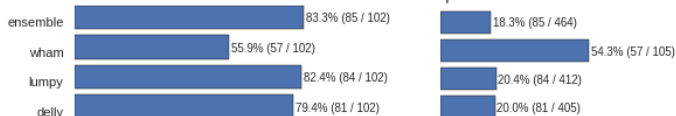
Structural variant evaluation

Deletions

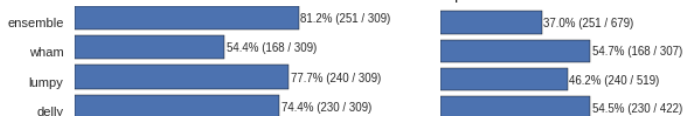
sensitivity

precision

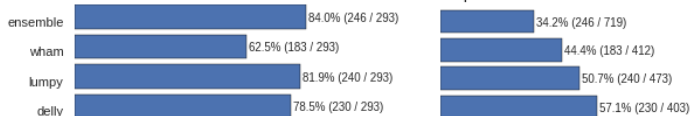
450 to 2000bp



2000 to 5000bp



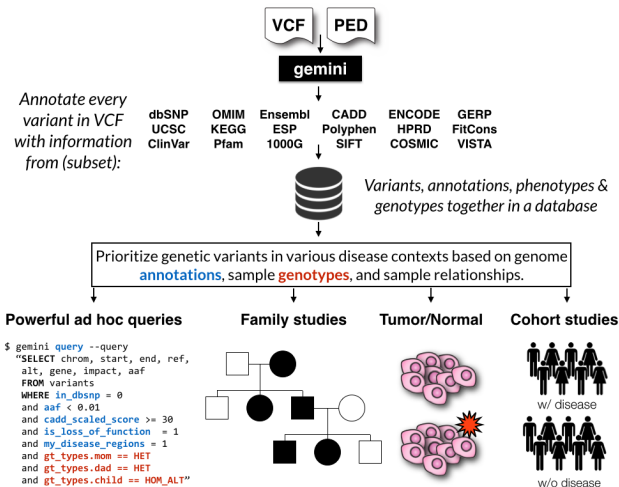
5000 to 25000bp



Variant analysis

- Associate with external annotations
- Manage large numbers of samples
- Query
- Visualize

Analyze: GEMINI



<http://gemini.readthedocs.org>

- PostgreSQL:

- <https://github.com/arq5x/gemini/tree/postgresql>

- Improved genotype representation

- CitusDB:

- <http://www.citusdata.com/>

- PostgreSQL compatible
 - Sharing and replication

- Distributed data schema: Avro + Parquet
- Distributed computation: Spark
- Conversion to and from VCF
- GA4GH:

<http://ga4gh.org/#/fileformats-team>

<http://bdgenomics.org/>

Making bcbio easy to use



John Davey

@johnomics



Following

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

↩ Reply ↻ Retweet ★ Favorite ... More

Automated Install

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

Need a consistent support environment

[Code](#) 22
[Issues](#) 155

Installation Search

We've found 155 issues Sort: Best match ▾

States

Closed	144
Open	11

[Search all of GitHub](#)

Oncofuse installation error #714

Hi @lh312, Sorry for the installation problems, I guess a lot of people have been updating their tools over the academic break. Thanks for figuring out what was wrong, that made it easier to fix it ...

👤 Opened by LH312 24 days ago 💬 2 comments

Mac OS 10.9 installation error #396

👤 Opened by alartin on Apr 13, 2014 💬 2 comments

Installation on Vagrant image fails #713

👤 Opened by ruin 24 days ago 💬 4 comments

Isolated installation download failing #659

👤 Opened by timothee-revil on Nov 10, 2014 💬 14 comments

Connection refused during installation - git cloning #670

👤 Opened by ruin on Nov 25, 2014 💬 2 comments

Installation error #614

Docker lightweight containers



docker

<http://docker.com>

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

- 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>

- Bootstrap from plain AMIs to cluster
- Pull/push data from S3
- Easy interface to start/stop clusters
- Lustre and NFS filesystems

<http://bcb.io/2014/12/19/awsbench/>

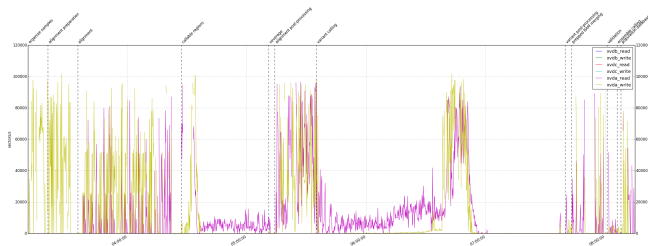
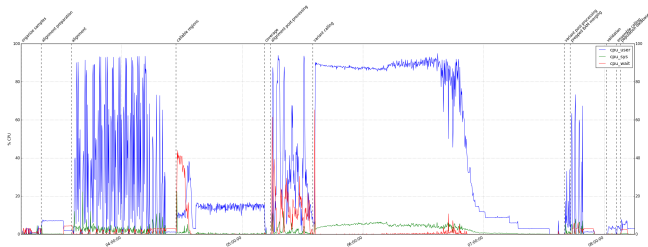
- Code/tools isolated in Docker containers
- Mounted filesystem + Docker for processing
- SLURM scheduler managed with Elasticcluster
- Future targets: Amazon EC2 Container Service

AWS benchmarking

	AWS (Lustre)
Total	4:42
genome data preparation	0:04
alignment preparation	0:12
alignment	0:29
callable regions	0:44
alignment post-processing	0:13
variant calling	2:35
variant post-processing	0:05
prepped BAM merging	0:03
validation	0:05

100X cancer tumor/normal exome on 64 cores (2 c3.8xlarge)

Resource usage plots



- bcbio – quality community built variant calling and RNA-seq analyses
- Validation – methods and scaling
- Variant management and analysis
- Ready to run implementation – Docker and AWS

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