

Supporting validated community developed variant calling analyses

Brad Chapman

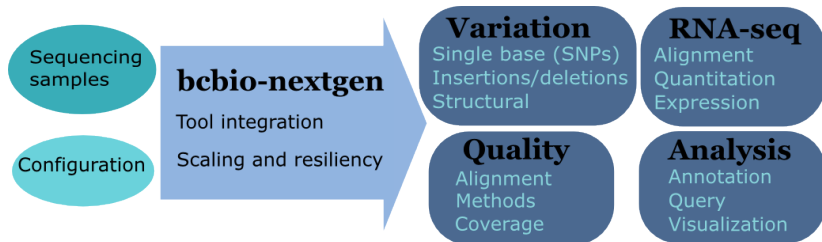
Bioinformatics Core, Harvard School of Public Health

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

<http://j.mp/bcbiolinks>

12 July 2014

Overview

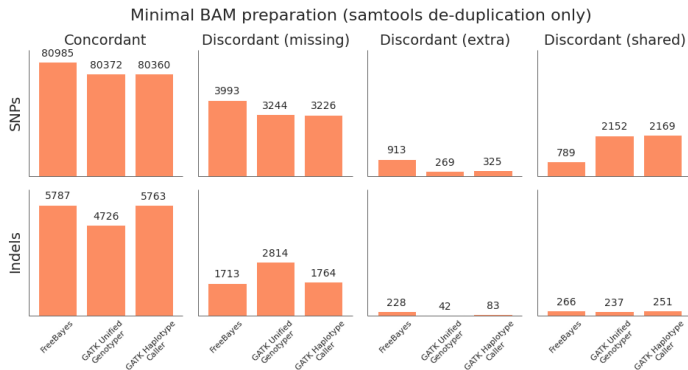


<https://github.com/chapmanb/bcbio-nextgen>

- Aligners: bwa-mem, novoalign, bowtie2
- Variantion: FreeBayes, GATK, 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
- Tool integration
- Validation – outputs + automated evaluation
- Installation of tools and data
- Scaling

Validation > Replication



Genome in a Bottle: <http://www.genomeinabottle.org/>

ICGC-TCGA DREAM: <https://www.synapse.org/#!Synapse:syn312572>

SMAsh: <http://smash.cs.berkeley.edu/>

Make installation easy



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](#) 18

Issues 104

States

Closed 96

Open 8

[Search all of GitHub](#)


Installation

We've found 104 issues

 **Installation** can fail if pypi is blocked

 Opened by [ibelframe](#) 2 days ago

 **Mac OS 10.9 installation** error

 Opened by [alartin](#) on Apr 13  2 comments

 **Update installation.rst**

add --data to dbnspf download


 Opened by [tanglingtung](#) 26 days ago  1 comment

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

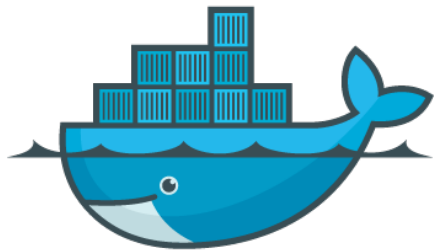
Hi, I encountered an error when installing the latest version of bcbio-nextgen on Ubuntu **installation** halted with a SHA256 mismatch error when it was installing platypus-variant

 Opened by [kennethban](#) 3 days ago  2 comments

 **Installation in arch**

 Opened by [kspham](#) on Jun 12  1 comment

Docker lightweight containers



docker

<http://docker.io>

- 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>

Docker image automation

```
$ ansible-playbook bcbio_vm_aws.yml
$ docker import \
  https://s3.amazonaws.com/bcbio_nextgen/ \
    bcbio-nextgen-docker-image.gz \
  chapmanb/bcbio-nextgen-devel
```

<http://www.ansible.com>

<https://github.com/chapmanb/bcbio-nextgen-vm/tree/master/ansible>

Docker HPC parallelization

bcbio-nextgen-vm
bcbio-nextgen
(workflow and parallel)
IPython parallel

Cluster scheduler
(SLURM, Torque,
SGE, LSF)

Machine 1

Docker Container
bcbio-nextgen
(run tools)
external tools
(bwa, freebayes...)

Machine 2

Docker Container
bcbio-nextgen
(run tools)
external tools
(bwa, freebayes...)

<http://ipython.org/ipython-doc/dev/parallel/index.html>

<https://github.com/roryk/ipython-cluster-helper>

Consistent scaling environment



Amazon challenges

- Cost – spot instances
- Disk – local scratch, no EBS
- Organization – no shared filesystems, S3 push/pull
- Data – reconstitute on minimal machines
- Security – encryption at rest

Amazon approaches

- Clusterk <http://clusterk.com/>
- Arvados <http://arvados.org/>
- Galaxy <http://usegalaxy.org/>

- Community developed variant calling analyses
<https://github.com/chapmanb/bcbio-nextgen>
- Docker: consistent install environment
- Automation: reproducible, understandable builds
- Need for a consistent scaling environment