

# Building community developed open source infrastructure to support large-scale biology research

Brad Chapman

Bioinformatics Core, Harvard School of Public Health

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

<http://j.mp/bcbiolinks>

12 September 2014

## Buffering of crucial functions by paleologous duplicated genes may contribute cyclicity to angiosperm genome duplication

Brad A. Chapman <sup>\*</sup>, <sup>†</sup>, John E. Bowers <sup>\*</sup>, Frank A. Feltus <sup>\*</sup>, and Andrew H. Paterson <sup>\*</sup>, <sup>†</sup>, <sup>‡</sup>, <sup>§</sup>, <sup>¶</sup>

Author Affiliations 

<sup>\*</sup>Plant Genome Mapping Laboratory and Departments of

<sup>†</sup>Plant Biology,

<sup>‡</sup>Genetics, and

<sup>§</sup>Crop and Soil Science, University of Georgia, Athens, GA 30602

# Synthetic biology startup (2004-2009)



<http://www.synthesis.cc/2009/04/on-the-demise-of-condon-devices.html>



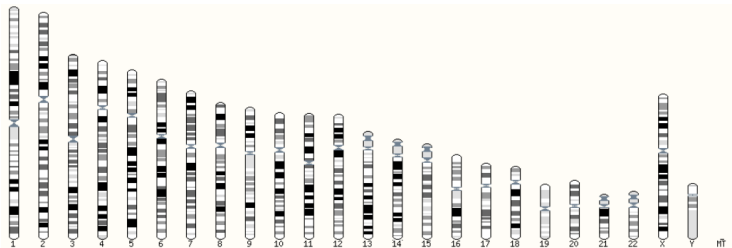
**HARVARD**  
**SCHOOL OF PUBLIC HEALTH**

Powerful ideas for a healthier world

<http://compbio.sph.harvard.edu/chb/>

- Community developed variant calling analyses
- Validation enables science
- Science at scale: 50 to 1500 genomes
- Supporting a community of users
- Software development and science

# Human whole genome sequencing



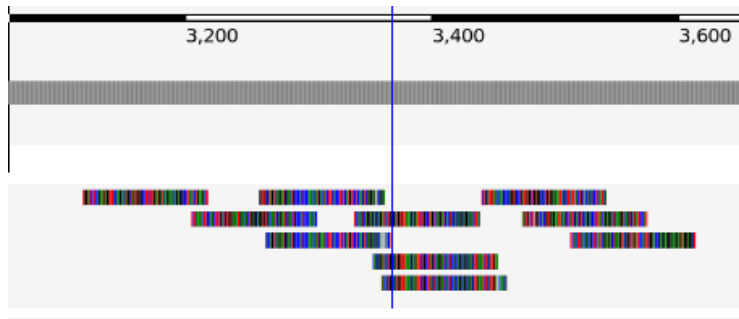
Click on the image above to jump to a chromosome, or click and drag to select a region

## Summary

Assembly	GRCh37.p13 (Genome Reference Consortium Human Reference 37), INSDC Assembly <a href="#">GCA_000001405.14</a> , Feb 2009
Database version	75.37
Base Pairs	3,326,743,047

[http://ensembl.org/Homo\\_sapiens/Location/Genome](http://ensembl.org/Homo_sapiens/Location/Genome)

# High throughput sequencing



# Variant calling

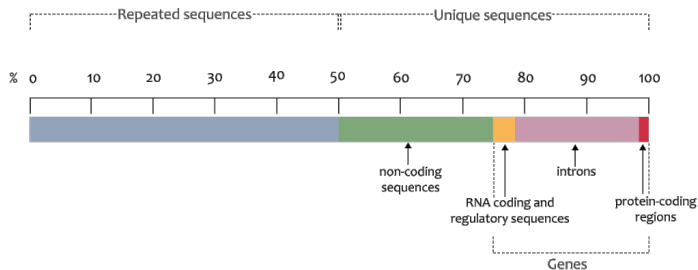


[http://en.wikipedia.org/wiki/SNV\\_calling\\_from\\_NGS\\_data](http://en.wikipedia.org/wiki/SNV_calling_from_NGS_data)



# Scale: exome to whole genome

## The haploid human genome sequence

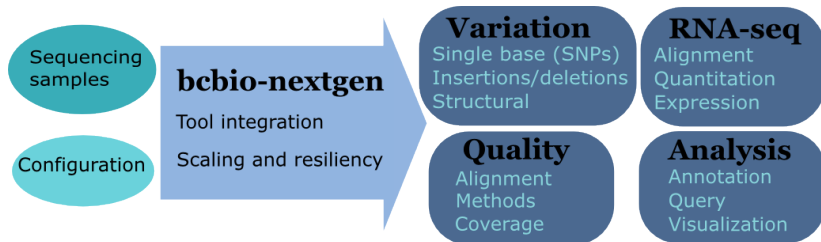


<https://www.flickr.com/photos/119980645@N06/>

# White box software



# 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
- 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

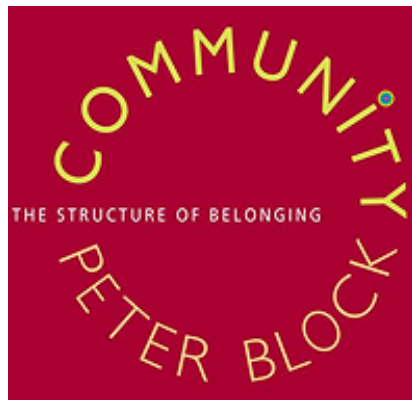
```
#####  
# HugeSeq                                #  
# The Variant Detection Pipeline        #  
#####
```

-- 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>

# 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, it says "Validated, scalable, community developed variant calling and RNA-seq analysis" with a link to <https://bcbio-nextgen.readthedocs.org>. The repository statistics show 2,717 commits, 1 branch, 16 releases, and 18 contributors. The current branch is **master**. The latest commit is titled "Trimming overhaul, removal of decompression of FASTQ files." by user **roryk**, authored 5 hours ago. The commit message is "Trimming overhaul, removal of decompression of FASTQ files." The commit hash is 4249d607ef. The repository has three subdirectories: **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), and **docs** (Disambiguate and fusion fields updated in docs 2 days ago). On the right side, there are links to **Code**, **Issues** (32), **Pull Requests** (5), **Pulse**, **Graphs**, and **Settings**.

chapmanb / **bcbio-nextgen**

Unwatch 33 Unstar 119 Fork 63

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

branch: master bcbio-nextgen / +

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

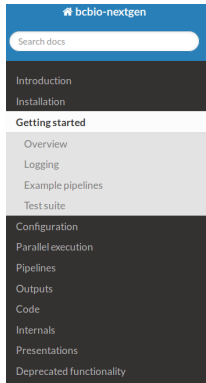
roryk authored 5 hours ago latest commit 4249d607ef

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

Code Issues 32 Pull Requests 5 Pulse Graphs 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
- RNA-seq differential expression
- Structural variations
- Cancer tumor/normal

<http://j.mp/cancer-var-chal>

# Example evaluation

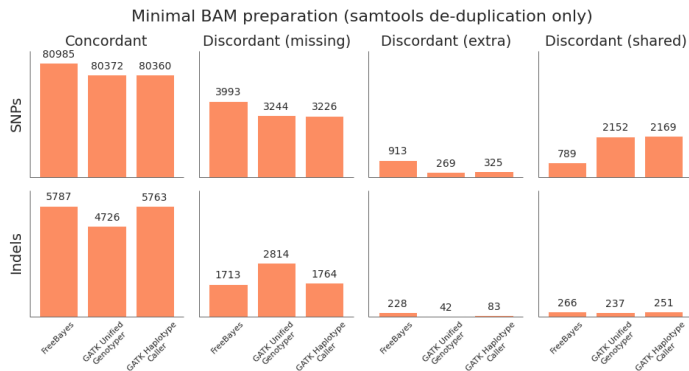
- Variant calling
  - GATK UnifiedGenotyper
  - GATK HaplotypeCaller
  - FreeBayes
- Two preparation methods
  - Full (de-duplication, recalibration, realignment)
  - Minimal (only de-duplication)



Genome in a Bottle  
Consortium

<http://www.genomeinabottle.org/>

# Quantify quality



- Quantification details: <http://j.mp/bcbioeval2>

# 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

# Scaling start point

- Initial pipeline scales with exomes
- 50 whole genomes = 3 months
- Next project: 1500 whole genomes



1500 whole genome scale – 110Tb

```
$ du -sh alz-p3f_2-g5/final
```

```
3.4T  alz-p3f_2-g5/final
```

```
$ ls -lhd *alz* | wc -l
```

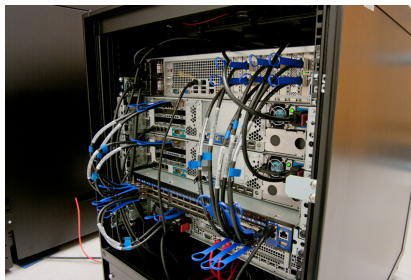
```
31
```

# How?

- Network bandwidth
- Avoid file intermediates
- Parallel alignment
- Parallel genome processing
- Better shared filesystems: Lustre

# Scaling: network bandwidth

## 1 GigE to Infiniband



Dell Genomic Data Analysis Platform; Glen Otero

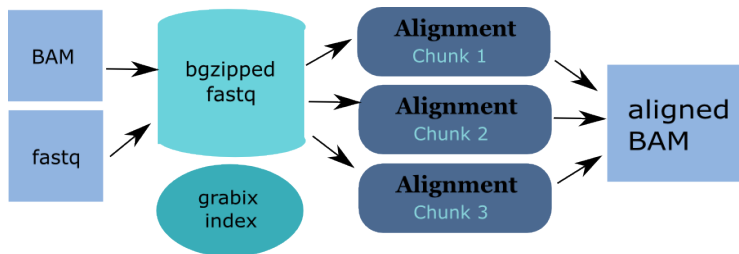
<http://www.dell.com/learn/us/en/555/hpcc/>

[high-performance-computing-life-sciences?c=us&l=en&s=biz&cs=555](http://www.dell.com/learn/us/en/555/hpcc/high-performance-computing-life-sciences?c=us&l=en&s=biz&cs=555)

## Scaling: avoid intermediates

```
("{bwa} mem -M -t {num_cores} -R '{rg_info}' -v 1 "  
"  {ref_file} {fastq_file} {pair_file} "  
"| {samblaster} "  
"| {samtools} view -S -u /dev/stdin "  
"| {sambamba} sort -t {cores} -m {mem} --tmpdir {tmpdir}"  
"  -o {tx_out_file} /dev/stdin")
```

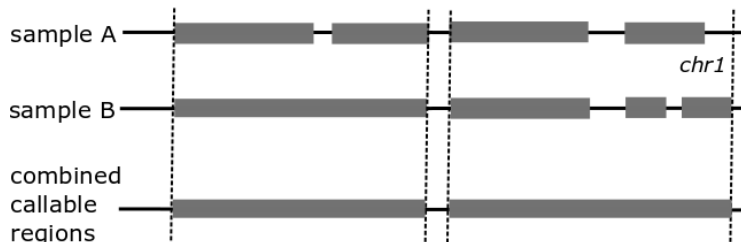
# Scaling: Parallel alignment



<https://github.com/arq5x/grabix>

# Scaling: Parallel by genome

## Selection of genome regions for parallel processing

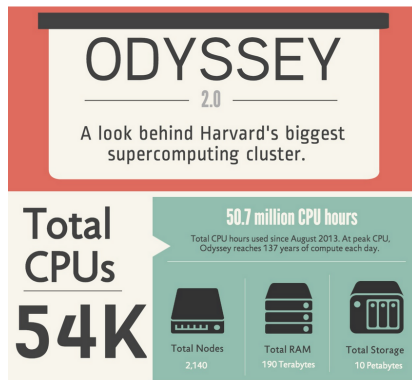


## Scaling: Lustre filesystem

480 cores, 30 samples

Step	Lustre	NFS
alignment	4.5h	6.1h
alignment post-processing	7.0h	20.7h

# Intel + Harvard FAS Research Computing



James Cuff, John Morrissey, Kristina Kermanshahche

<https://rc.fas.harvard.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 [lbeltrame](#) 2 days ago

 Mac OS 10.9 installation error

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

 Update installation.rst

add --data to dbnftp download



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

 SHA256 mismatch for platypus-variant in installation

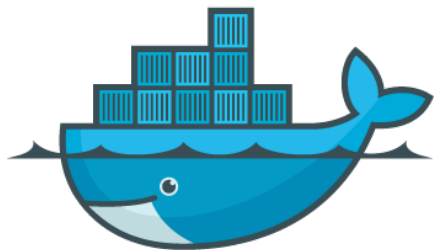
Hi, I encountered an error when installing the latest version of bcbio-nextgen on Ubuntu  
installation halted with a SHA256 mismatch error when it was installing platypus-variant

 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>



<http://software-carpentry.org>

<http://mozillascience.org>



<http://github.com>

<https://bitbucket.org>

**IP[y]:** IPython  
Interactive Computing

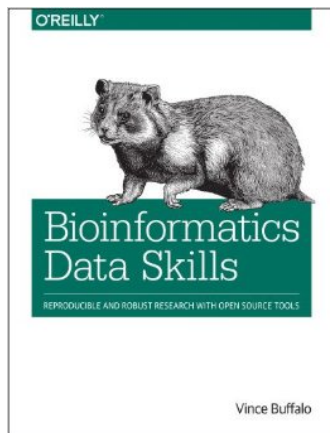


<http://ipython.org>

<http://www.rstudio.com/>



Good practices = good science



<http://shop.oreilly.com/product/0636920030157.do>

# O|B|F



<http://www.open-bio.org>

[http://www.open-bio.org/wiki/BOSC\\_2014](http://www.open-bio.org/wiki/BOSC_2014)

<http://usegalaxy.org>

<https://wiki.galaxyproject.org/Events/GCC2014>

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>

# Coding as a science career

- Wide range of projects
- Collaboration
- Respected
- Help others
- Grow and learn

- Community developed variant calling analyses
- Validation enables science
- Science at scale: 50 to 1500 genomes
- Supporting a community of users
- Software development and science

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