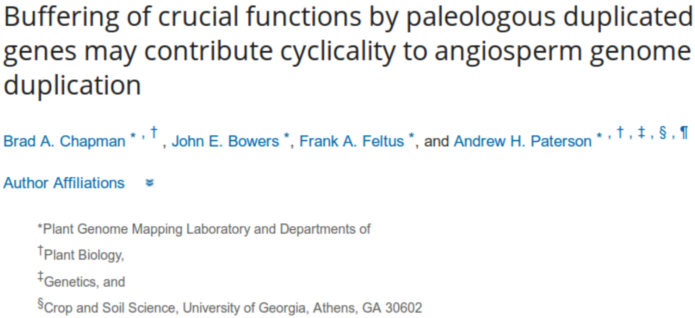
Building community developed open source infrastructure to support large-scale biology research

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

12 September 2014

University of Georgia (1999-2004)



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

Brad A. Chapman[*,†], John E. Bowers[*], Frank A. Feltus[*], and Andrew H. Paterson[*,†,‡,§,¶]

Author Affiliations

[*] Plant Genome Mapping Laboratory and Departments of

[†] Plant Biology,

[‡] Genetics, and

[§] 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

Bioinformatics core – Harvard School of Public Health

HARVARD

SCHOOL OF PUBLIC HEALTH

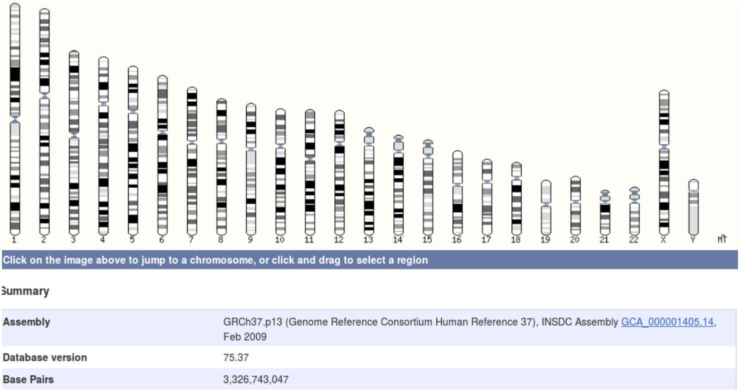
Powerful ideas for a healthier world

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

Summary

- 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



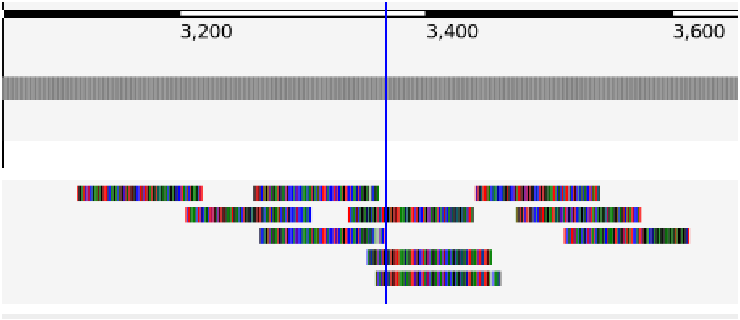
Click on the image above to jump to a chromosome, or click and drag to select a region

Summary

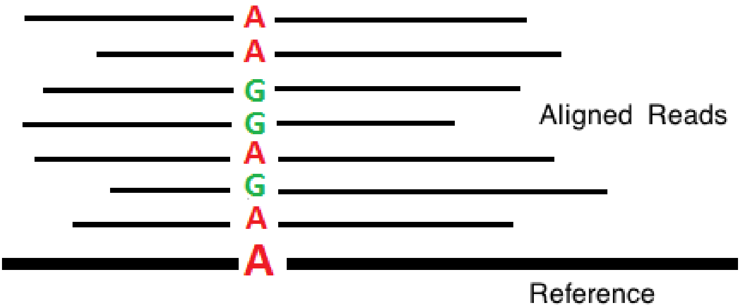
Assembly	GRCh37.p13 (Genome Reference Consortium Human Reference 37), INSDC Assembly GCA_000001405.14, Feb 2009
Database version	75.37
Base Pairs	3,326,743,047

http://ensembl.org/Homo_sapiens/Location/Genome

High throughput sequencing

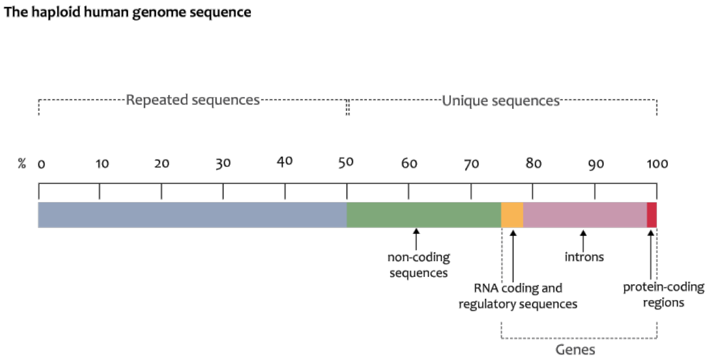


Variant calling



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

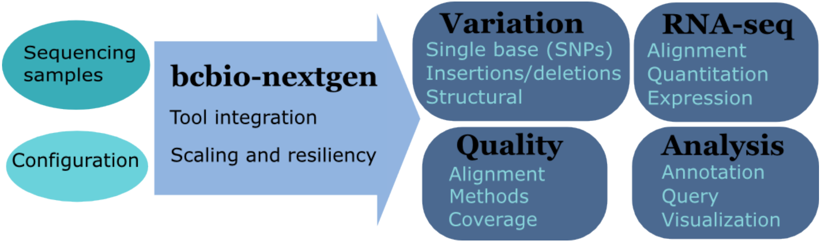
Scale: exome to whole genome



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

White box software

Overview



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

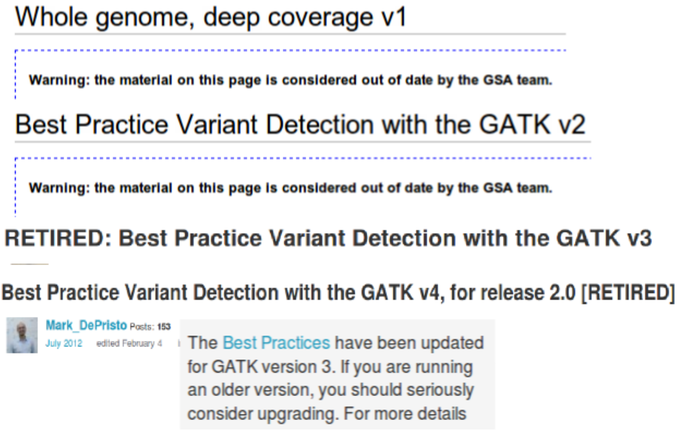
Uses

- 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.

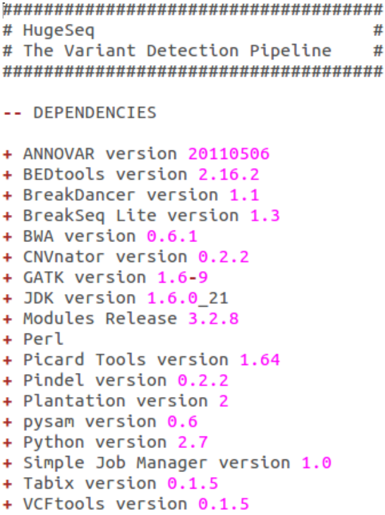
RETIRED: Best Practice Variant Detection with the GATK v3

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



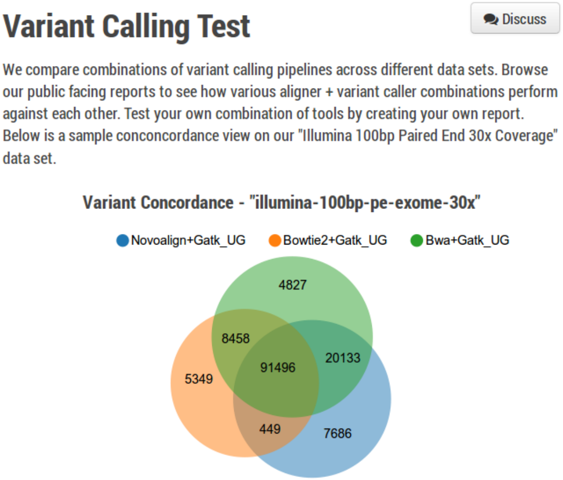
```
#####################################
# HugeSeq                           #
# The Variant Detection Pipeline    #
#####################################

-- 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

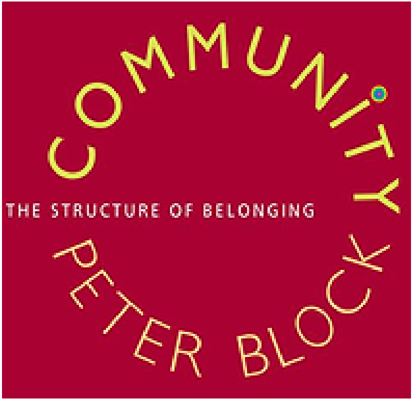
Discuss

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"

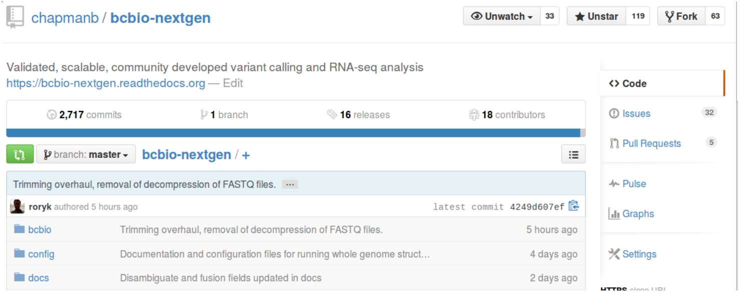
http://www.bioplanet.com/gcat

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

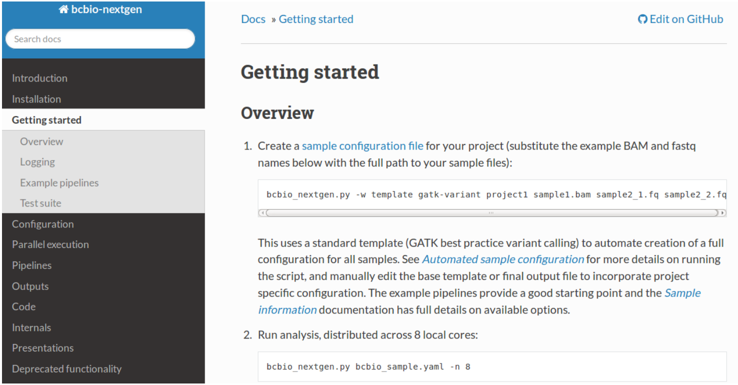
Trimming overhaul, removal of decompression of FASTQ files. ···

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
- RNA-seq differential expression
- Structural variations
- Cancer tumor/normal http://j.mp/cancer-var-chal

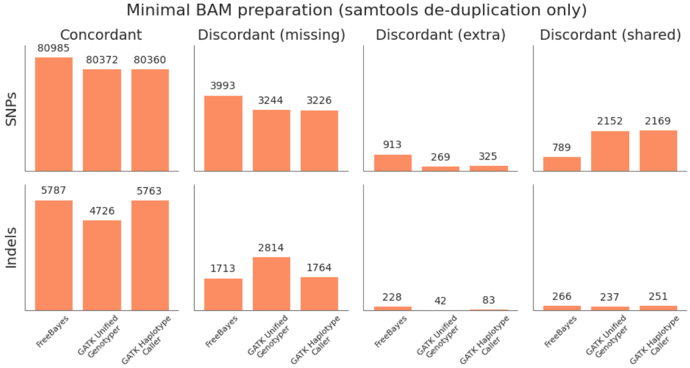
Example evaluation

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

Reference materials

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

End point

1500 whole genome scale – 110Tb

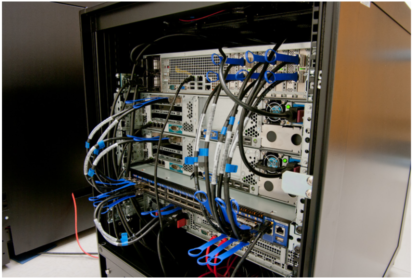
```
$ du -sh alz-p3f_2-g5/final
3.4T  alz-p3f_2-g5/final
$ ls -lhd *alz* | wc -l
31
```

How?

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

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

Scaling: Lustre filesystem

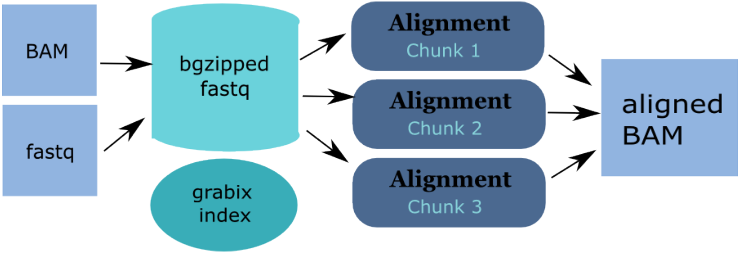
480 cores, 30 samples

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

Scaling: avoid intermediates

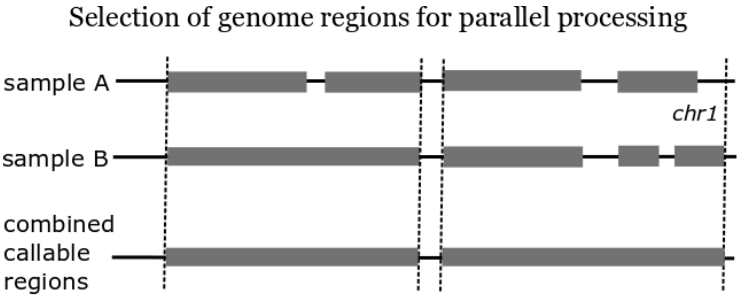
```
("{bwa} mem -M -t {num_cores} -R ’{rg_info}’ -v 1 "
 "  {ref_file} {fastq_file} {pair_file} "
 "| {samblaster} "
 "| {samtools} view -S -u /dev/stdin "
 "| {sambamba} sort -t {cores} -m {mem} --tmpdir {tmpdir}"
 "   -o {tx_out_file} /dev/stdin")
```

Scaling: Parallel alignment

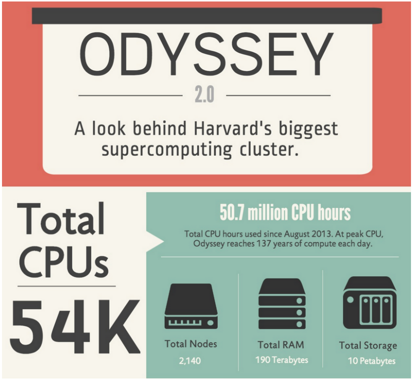


https://github.com/arq5x/grabix

Scaling: Parallel by genome


Selection of genome regions for parallel processing
sample A
chr1
sample B
combined callable regions

Intel + Harvard FAS Research Computing



ODYSSEY

2.0

A look behind Harvard's biggest supercomputing cluster.

Total CPUs

54K

50.7 million CPU hours

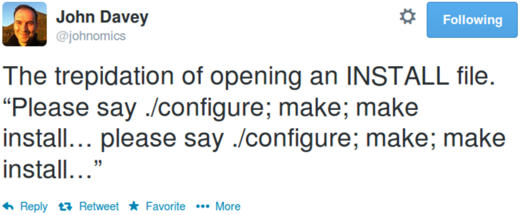
Total CPU hours used since August 2013. At peak CPU, Odyssey reaches 137 years of compute each day.

Total Nodes	Total RAM	Total Storage
2,140	190 Terabytes	10 Petabytes

James Cuff, John Morrissey, Kristina Kermanshahche

https://rc.fas.harvard.edu/

Make installation easy



John Davey
@johnomics

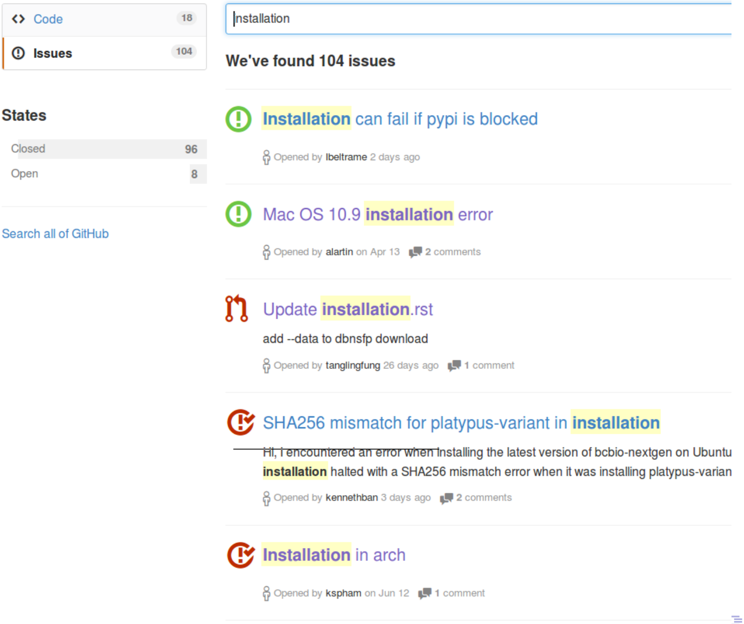
The trepidation of opening an INSTALL file. “Please say ./configure; make; make install… please say ./configure; make; make install…”

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

Search all of GitHub

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**

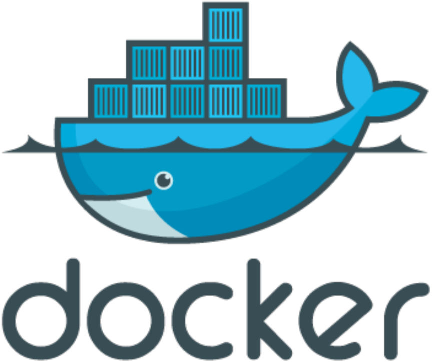
~~Hi, I encountered an error when~~ installing the latest version of bcbio-nextgen on Ubuntu **installation** halted with a SHA256 mismatch error when it was installing platypus-varian

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

Learning good practices

software carpentry

mozilla Science Lab

http://software-carpentry.org
http://mozillascience.org

Revision control

http://github.com
https://bitbucket.org

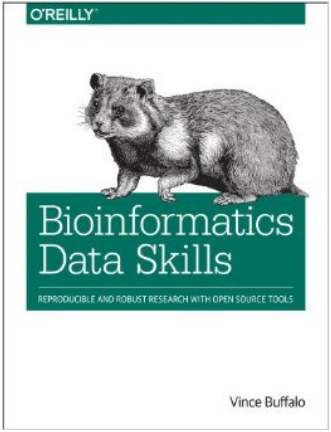
Reproducible environments

IP[y]: IPython
Interactive Computing

RStudio™

http://ipython.org
http://www.rstudio.com/

Good practices = good science



O'REILLY

Bioinformatics Data Skills

REPRODUCIBLE AND ROBUST RESEARCH WITH OPEN SOURCE TOOLS

Vince Buffalo

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

Open Source Communities

O|B|F

Galaxy

http://www.open-bio.org

http://www.open-bio.org/wiki/BOSC_2014

http://usegalaxy.org

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

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

Coding as a science career

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

Summary

- 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