

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

Brad Chapman
Bioinformatics Core, Harvard Chan School

<https://bcb.io>

<http://j.mp/bcbiolinks>

23 September 2015

- My background
- Research scientist at a core facility
- The open source bioinformatics community
- Why you want to work as a research scientist
- How to prepare yourself

Undergrad Michigan State (Ecology)



https://en.wikipedia.org/wiki/Pinus_nigra

Undergrad Michigan State (Plant transformation)



Buffering of crucial functions by paleologous duplicated genes may contribute cyclicity 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>



HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Powerful ideas for a healthier world

<http://bioinformatics.sph.harvard.edu/>

- My background
- Research scientist at a core facility
- The open source bioinformatics community
- Why you want to work as a research scientist
- How to prepare yourself

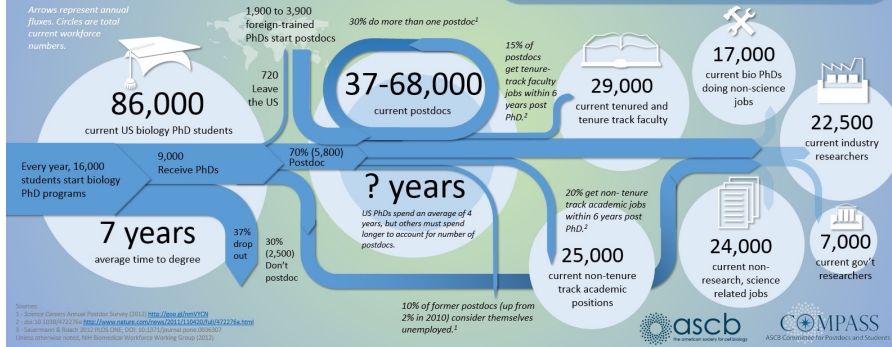
- Work in a support core
- Consulting
- Team of 8 researchers

<http://bioinformatics.sph.harvard.edu/people/>

- Specialize, but also overlap
- Research scientists

Where will a biology PhD take you?

Arrows represent annual fluxes. Circles are total current workforce numbers.



Sources:

- 1 - Science Careers Annual Postdoc Survey (2012). <http://joc.sagepub.com/2012>
- 2 - doi:10.1038/472727a <http://www.nature.com/news/2011/11/24/472727a.html>
- 3 - Sauerbrey & Roach 2012 PLOS ONE, DOI: 10.1371/journal.pone.0036307

Unless otherwise noted, NIH Biomedical Workforce Working Group (2012)

<http://www.sciencedirect.com/science/article/pii/S0968000414001728>

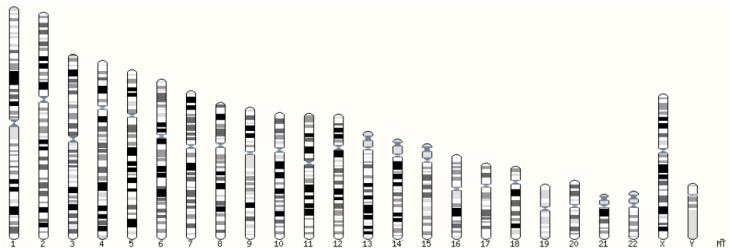
Who we work with?

- Academic Researchers: Harvard Stem Cell Institute, Harvard Medical School, Harvard NeuroDiscovery Center, Massachusetts General Hospital
- Large consortium projects: Cure Alzheimer's, Global Alliance for Genomic Health
- Industry: AstraZeneca, Biogen, Merck
- Startups

- Collaboration with researchers
- Data analysis
- Teaching and training
- Large scale infrastructure development

- Alzheimer's – large populations of affected families
- Cancer treatment – detection of driver mutations, relapse after treatment
- HIV – detection of low frequency drug resistant sub-populations

Human whole genome sequencing



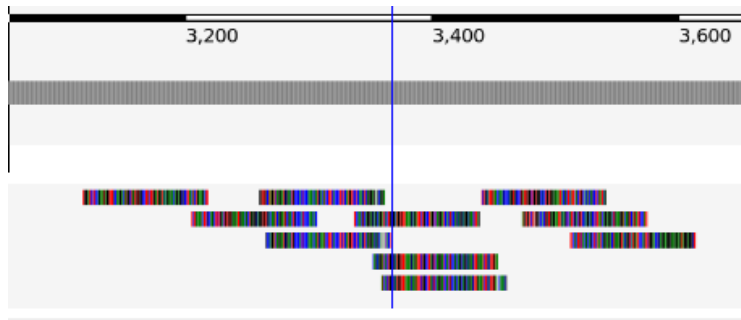
Click on the image above to jump to a chromosome, or click and drag to select a region

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



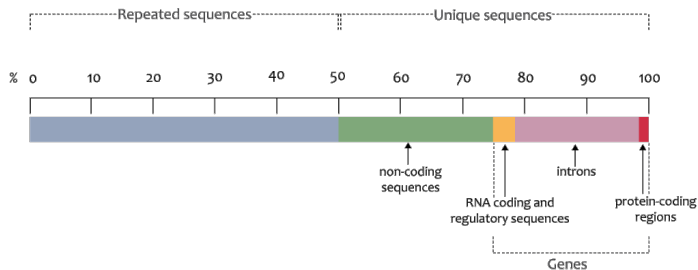
Variant calling



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

Scale: exome to whole genome

The haploid human genome sequence



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

- My background
- Research scientist at a core facility
- The open source bioinformatics community
- Why you want to work as a research scientist
- How to prepare yourself

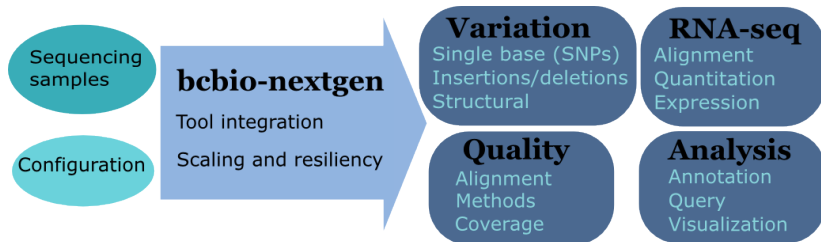
Large scale infrastructure development

- Find shared problems
- Community developed analyses
- Validation
- Scaling
- Supporting a community of users

White box software



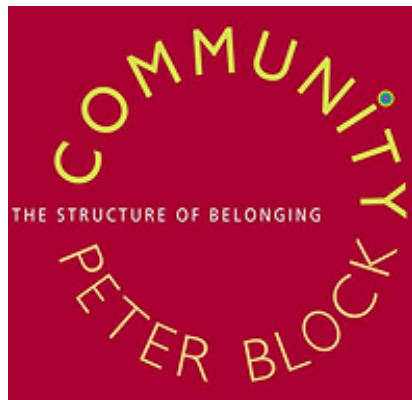
Overview



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

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

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 states: "Validated, scalable, community developed variant calling and RNA-seq analysis" with a link to <https://bcbio-nextgen.readthedocs.org> and an **Edit** button.

Repository statistics are displayed in a bar: **2,717** commits, **1** branch, **16** releases, and **18** contributors.

The current branch is **master**. Below this, a commit message is shown: "Trimming overhaul, removal of decompression of FASTQ files." by user **roryk**, authored 5 hours ago. The latest commit hash is **4249d607ef**.

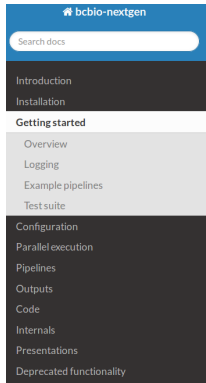
A table lists the repository's structure:

File/Folder	Description	Last Update
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 side, there is a sidebar with links to **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>

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>

- My background
- Research scientist at a core facility
- The open source bioinformatics community
- Why you want to work as a research scientist
- How to prepare yourself

Research scientist as a career – pros

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

Open source communities – pros

- Work on problems with impact
- Large set of peers
- Fortuitous interactions
- Transferable skills

Research scientist – cons

- Less control over overall biological questions
- Juggle more simultaneous projects

- My background
- Research scientist at a core facility
- The open source bioinformatics community
- Why you want to work as a research scientist
- How to prepare yourself



<http://software-carpentry.org>

<http://mozillascience.org>



Atlassian



<http://github.com>

<https://bitbucket.org>

IP[y]: IPython
Interactive Computing

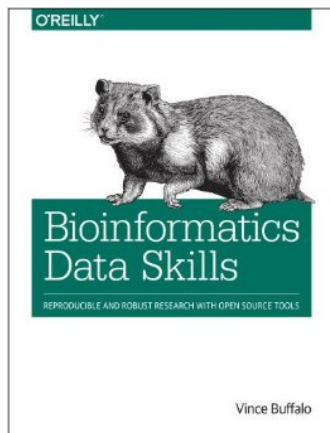


<http://jupyter.org/>

<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://usegalaxy.org>

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

Summary

- Bioinformatics core at Harvard Chan School
- Collaborative research work
- Open source community
- Contribute to public health research

<https://bcb.io>

<https://j.mp/bcbiolinks>