Validated, scalable, community developed variant calling

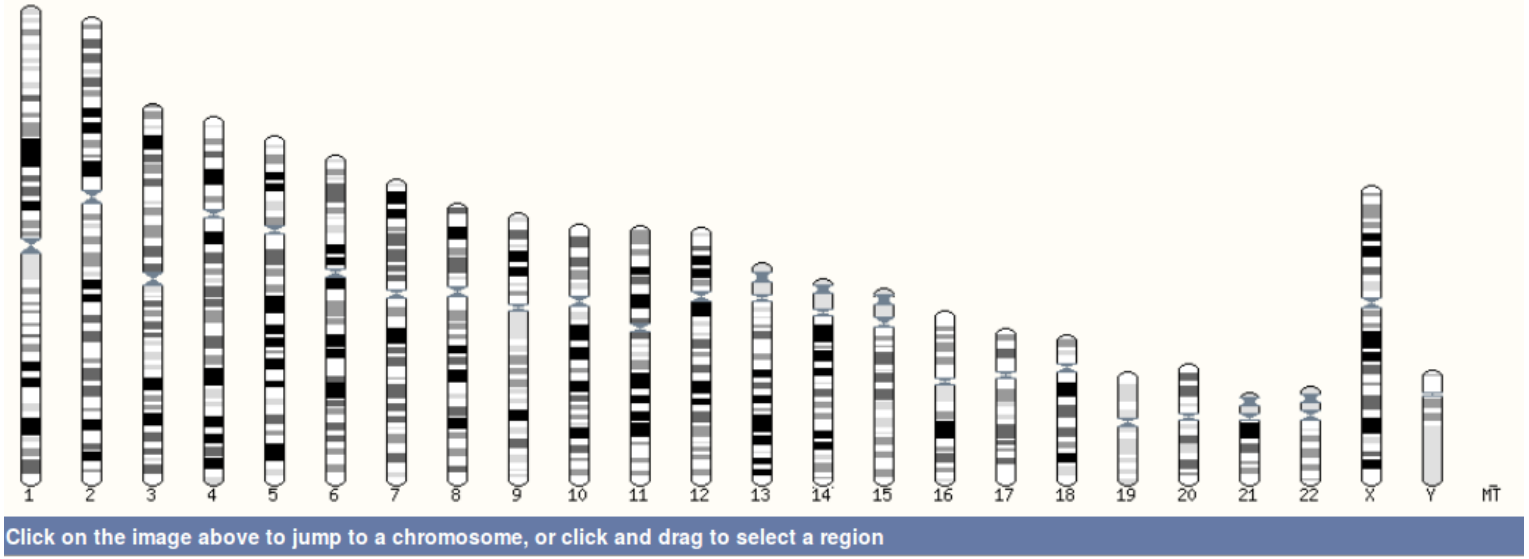
Brad Chapman
Bioinformatics Core, Harvard Chan School
https://github.com/chapmanb/bcbio-nextgen
http://bcb.io
http://j.mp/bcbiolinks

1 April 2015

Acknowledgments

- Harvard Chan School Bioinformatics Core
  http://hsphbio.ghost.io/
- Rudy Tanzi Lab – whole genome scaling
- Harvard FAS Research Computing – infrastructure
- Biogen and Intel – cloud integration
- Wolfson Wohl Cancer Research Centre
- AstraZeneca – cancer variant calling
  https://www.linkedin.com/jobs2/view/40026565

Human whole genome sequencing

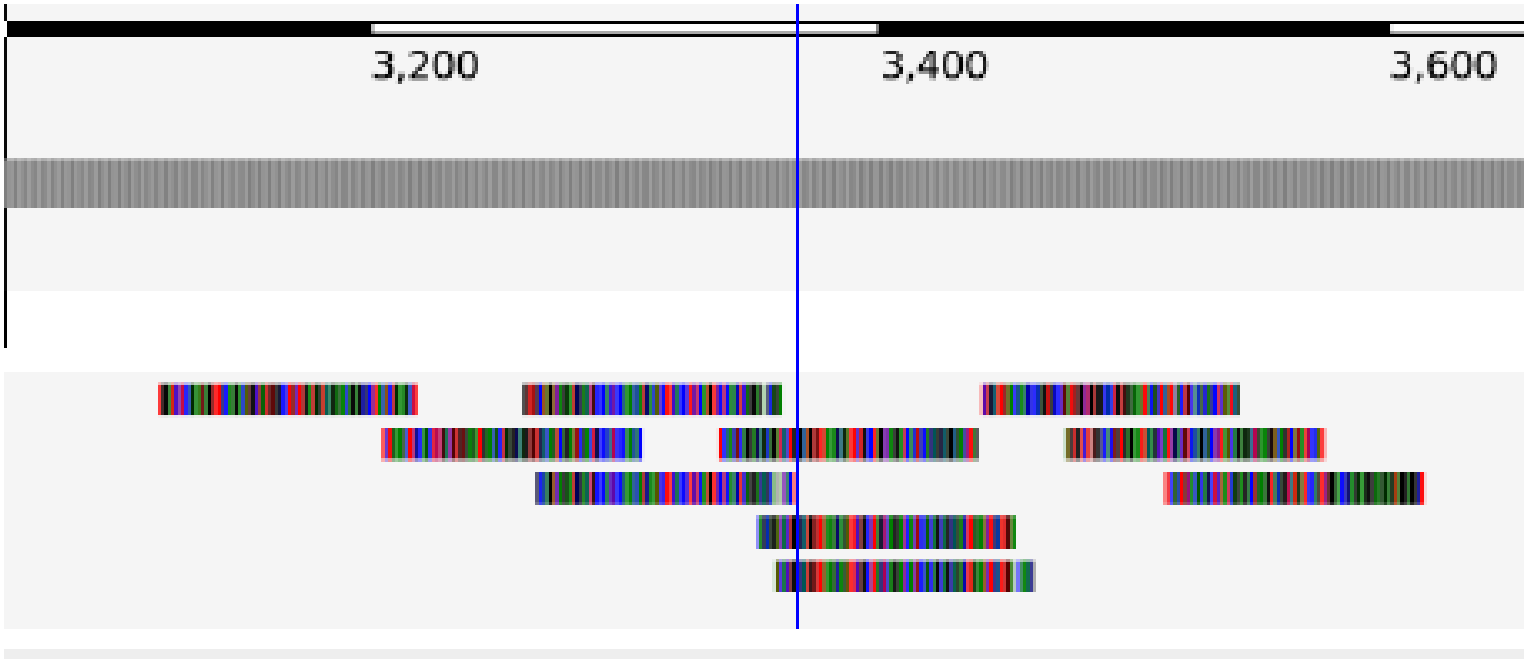


Summary

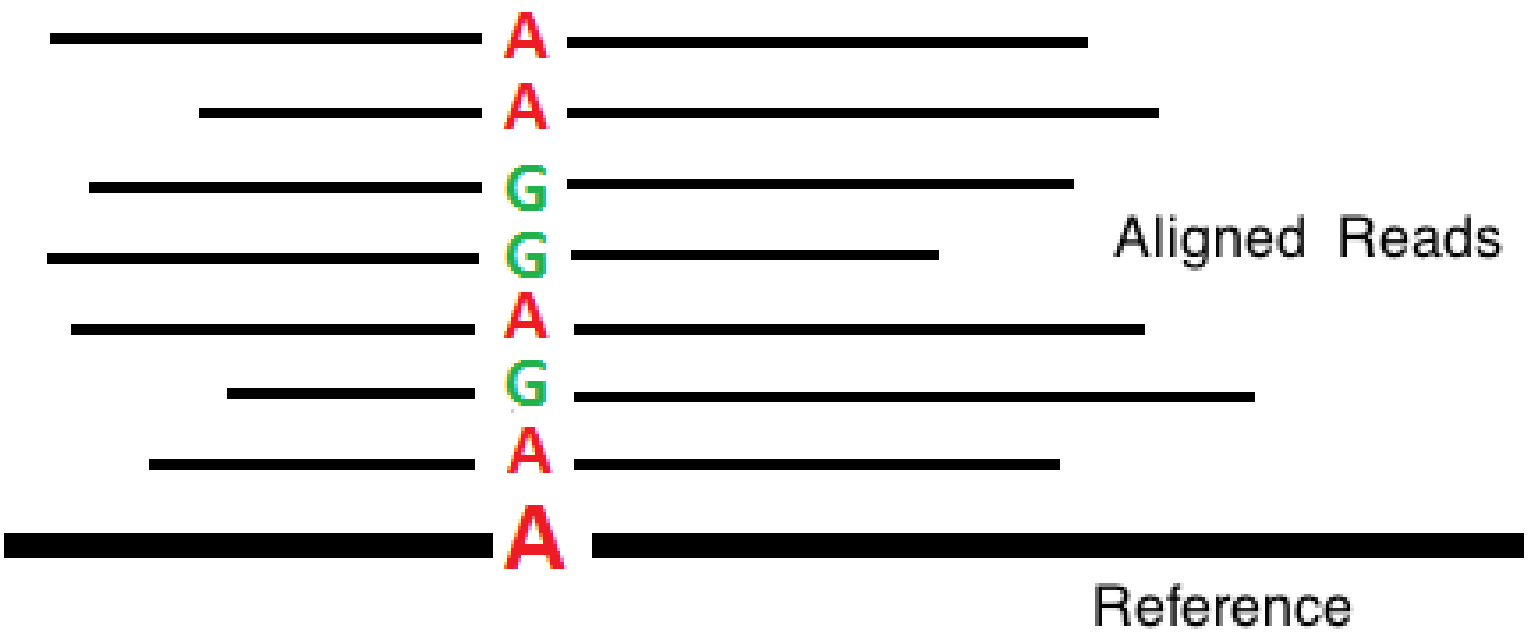
Assembly	GRCh37.p13 (Genome Reference Consortium Human Reference 37), INSDC Assembly GCA_000001405.14, Feb 2009
Database version	75.37
Base Pairs	3,326,743,047

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

High throughput sequencing

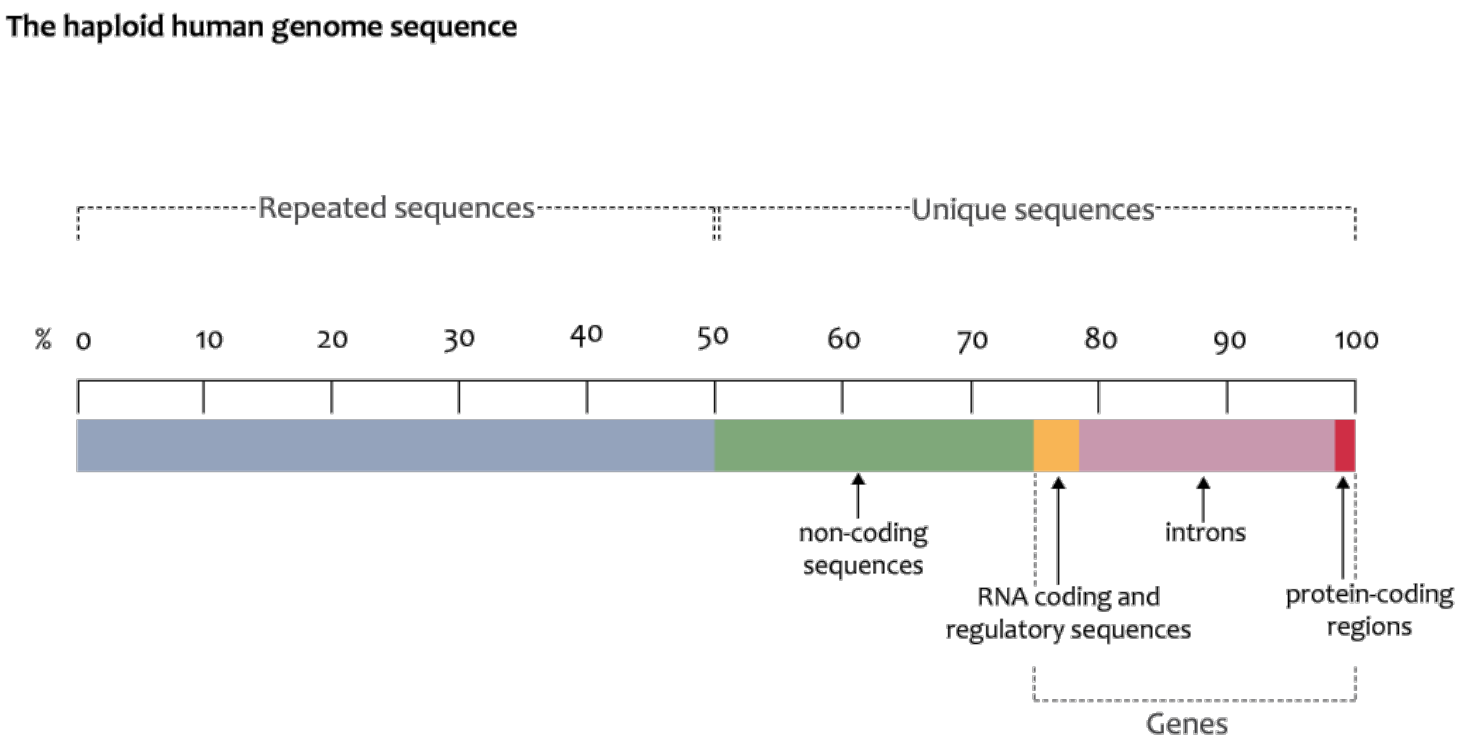


Variant calling



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

Scale: exome to whole genome



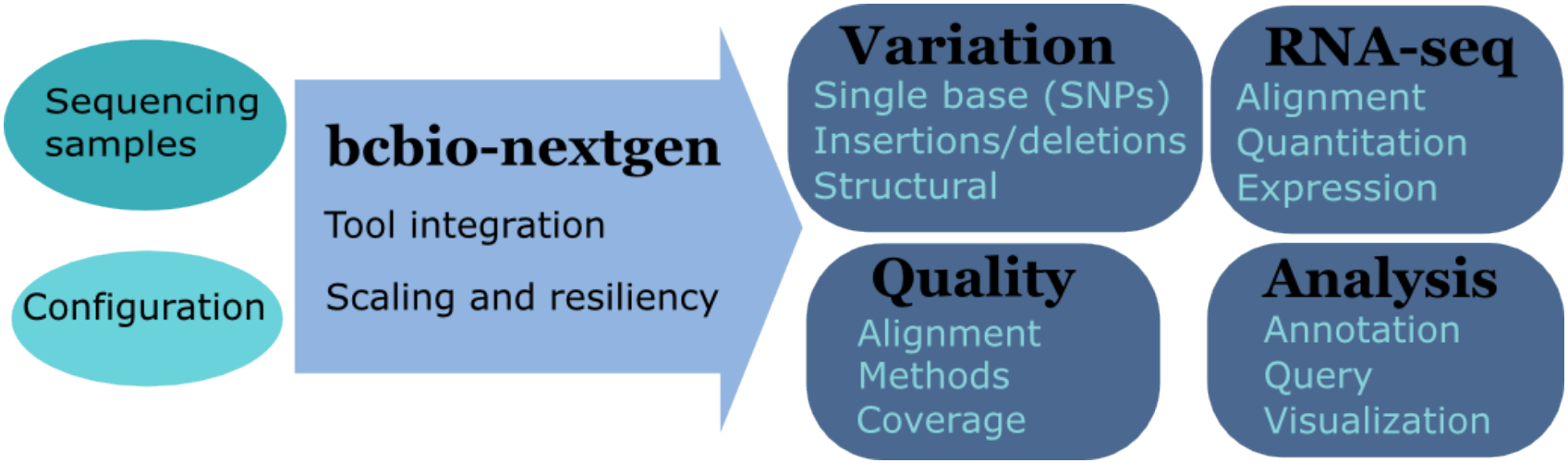
https://www.flickr.com/photos/119980645@N06/

Summary

- Overview of bcbio
- Community development
- Validation
- Docker and Amazon Web Services

White box software

Overview



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

Uses

- Aligners: bwa-mem, novoalign, bowtie2
- Variantion: FreeBayes, GATK, VarDict, MuTecT, Scalpel, SnpEff, VEP, GEMINI, Lumpy, Delly, CNVkit
- RNA-seq: Tophat, STAR, cufflinks, HTSeq
- Quality control: fastqc, bamtools, RNA-SeQC
- Manipulation: bedtools, bcftools, biobambam, sambamba, samblaster, samtools, vcflib, vt

Provides

- 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 baseline functionality

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

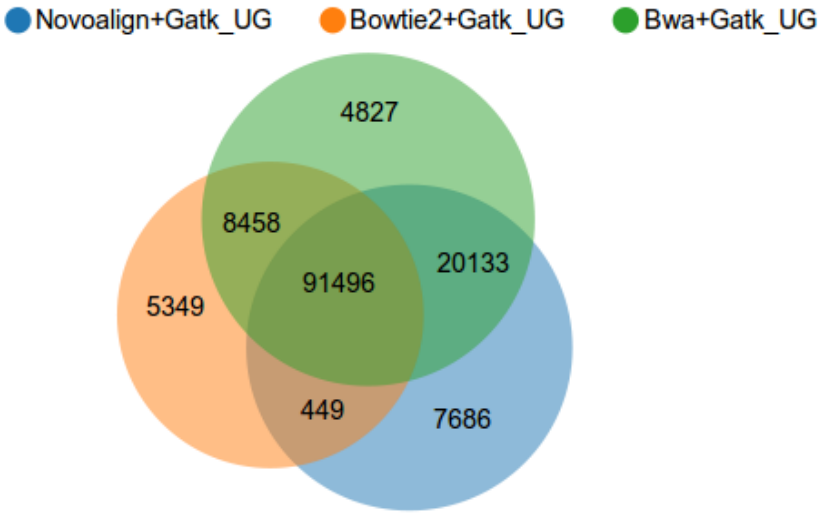
Quality differences between methods

Variant Calling Test

Discuss

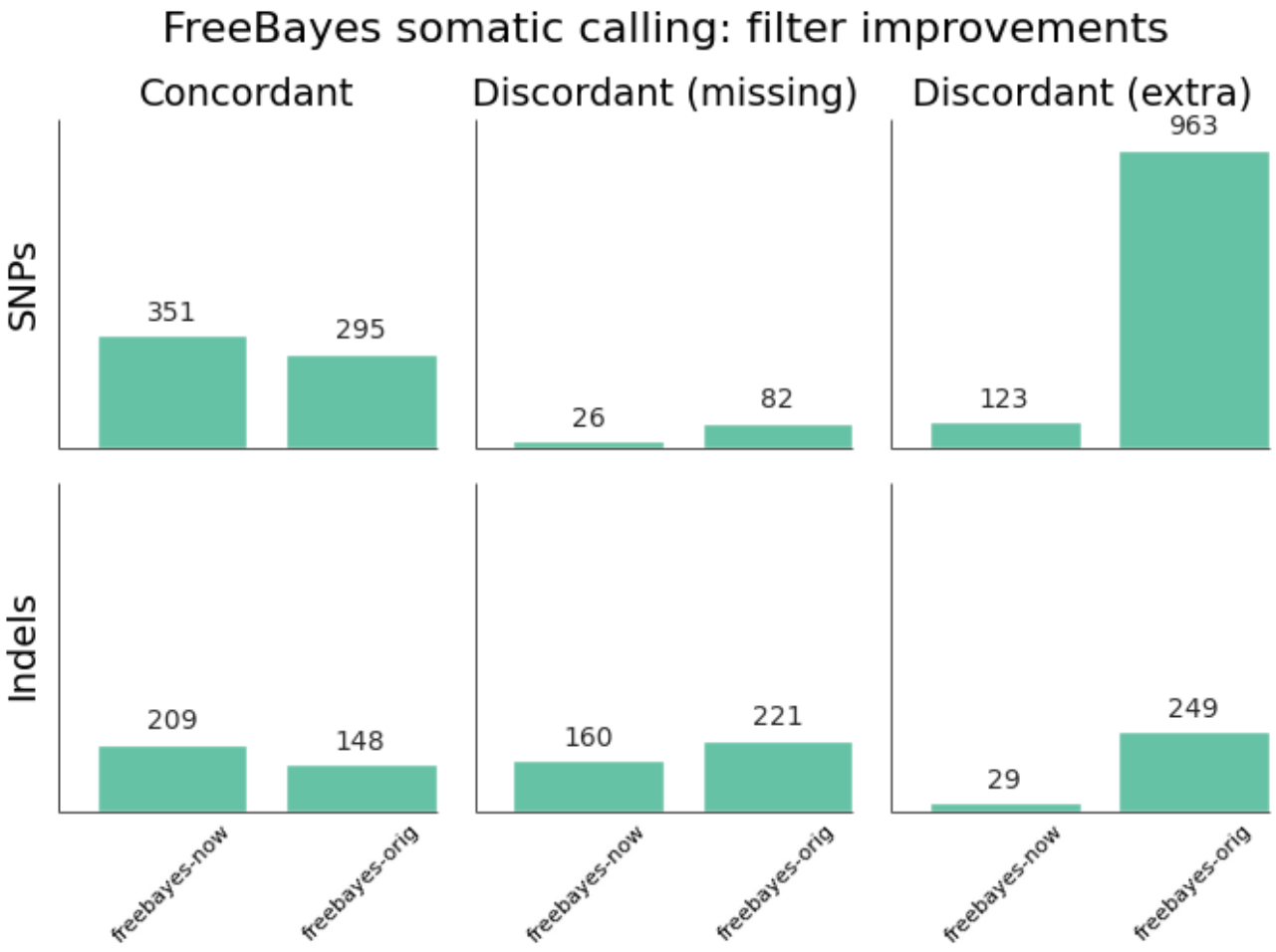
We compare combinations of variant calling pipelines across different data sets. Browse our public facing reports to see how various aligner + variant caller combinations perform against each other. Test your own combination of tools by creating your own report. Below is a sample conconcordance view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

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



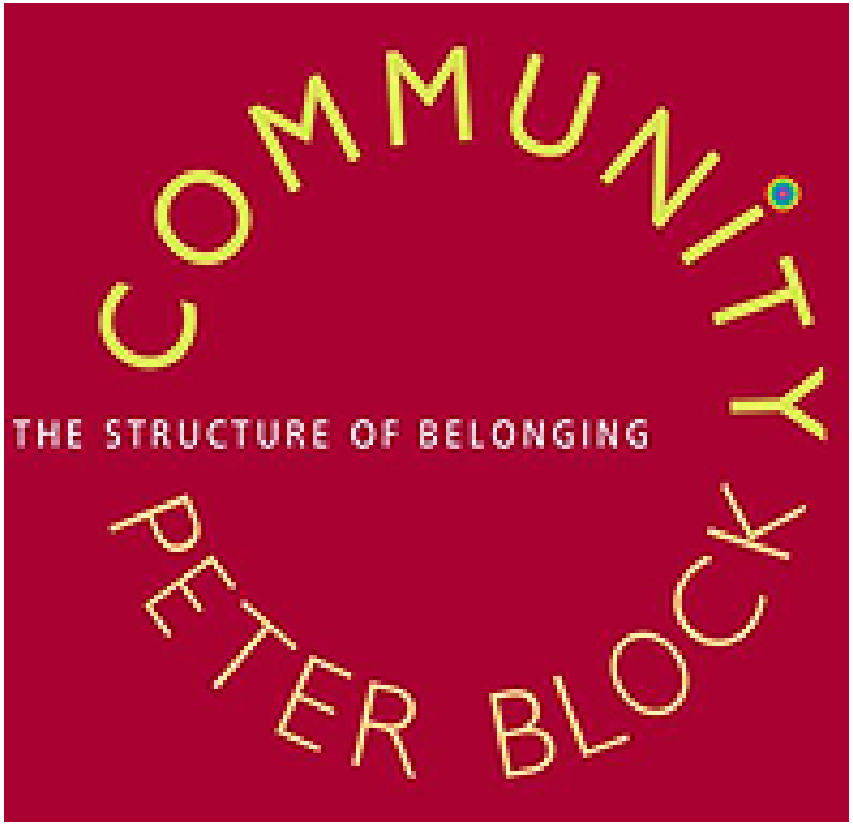
http://www.bioplanet.com/gcat

Benefits of improved filtering



http://j.mp/cancervalpre

Solution



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

Community: contribution

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

Contributors

- Miika Ahdesmaki, AstraZeneca
- Luca Beltrame, IRCCS "Mario Negri" Institute for Pharmacological Research, Milan, Italy
- Alla Bushoy, AstraZeneca
- Guillermo Carrasco, Science for Life Laboratory, Stockholm
- Nick Carriero, Simons Foundation
- Brad Chapman, Harvard Chan Bioinformatics Core
- Saket Choudhary, University Of Southern California
- Peter Cock, The James Hutton Institute
- Matt Edwards, MIT
- Mario Giovacchini, Science for Life Laboratory, Stockholm
- Karl Gutwin, Biogen
- Jeff Hammerbacher, Icahn School of Medicine at Mount Sinai
- John Kern
- Rory Kirchner, Harvard Chan Bioinformatics Core
- Jakub Nowacki, AstraZeneca
- John Morrissey, Harvard Chan Bioinformatics Core
- Lorena Pantano, Harvard Chan Bioinformatics Core
- Brent Pedersen, University of Colorado Denver
- James Porter, The University of Chicago
- Valentine Svensson, Science for Life Laboratory, Stockholm
- Paul Tang, UCSF
- Roman Valls, Science for Life Laboratory, Stockholm
- Kevin Ying, Garvan Institute of Medical Research, Sydney, Australia

Validation

Tests for implementation and methods

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

Reference materials

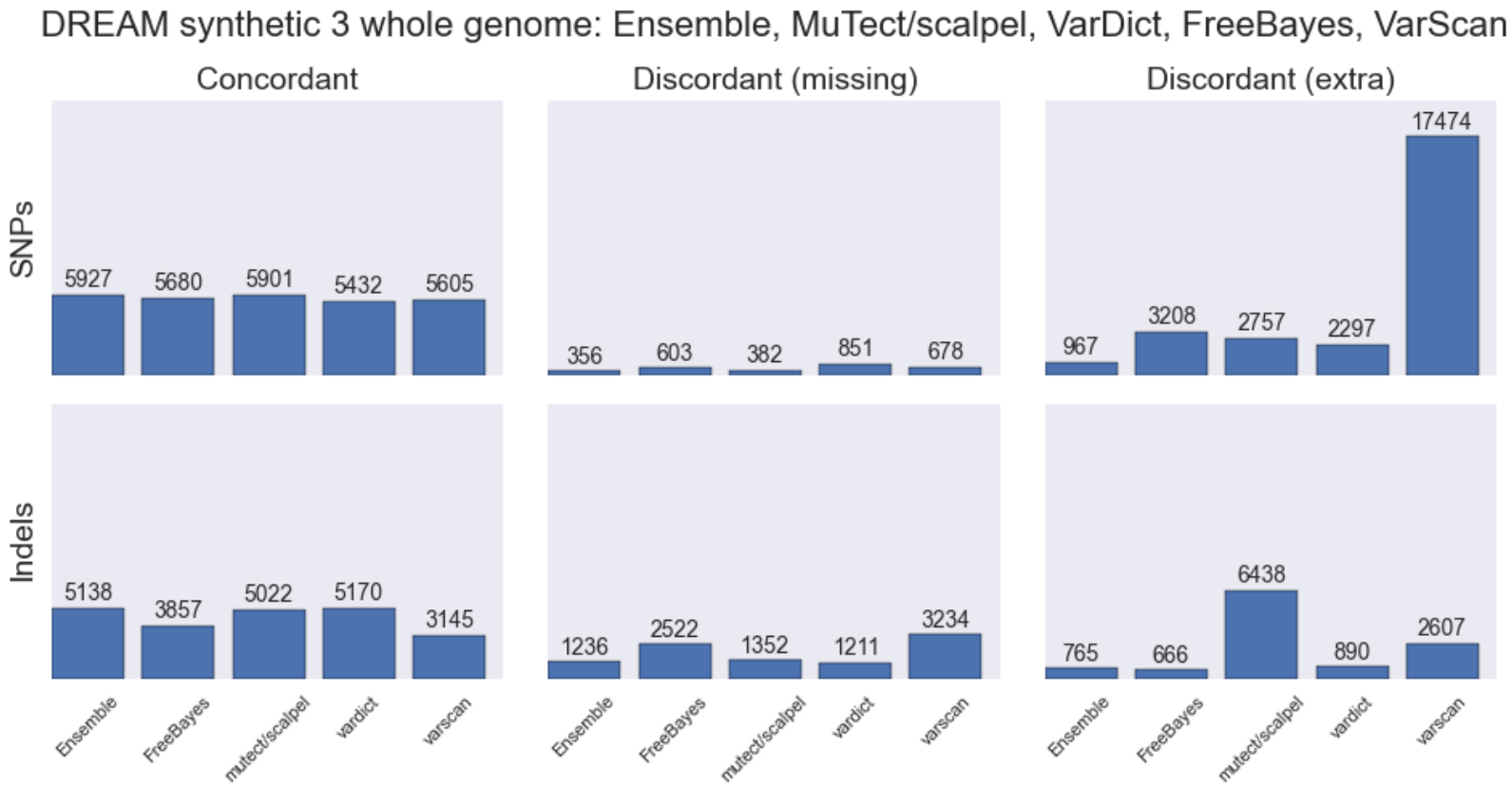
Genome in a Bottle Consortium

Global Alliance for Genomics & Health

ICGC-TCGA DREAM Mutation Calling challenge

http://www.genomeinabottle.org/
http://ga4gh.org/#/benchmarking-team
https://www.synapse.org/#!Synapse:syn312572

Validate and compare caller performance



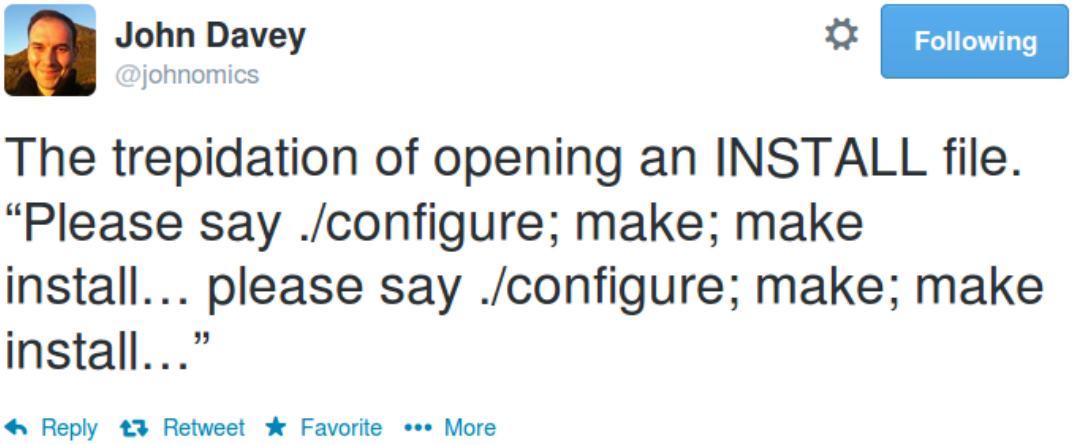
http://bcb.io/2015/03/05/cancerval/

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

Making bcbio easy to use



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

States

Closed 144

Open 11

Search all of GitHub

installation

Search

We've found 155 issues

Sort: **Best match**

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 nuin 24 days ago 4 comments

Isolated **installation** download failing #659

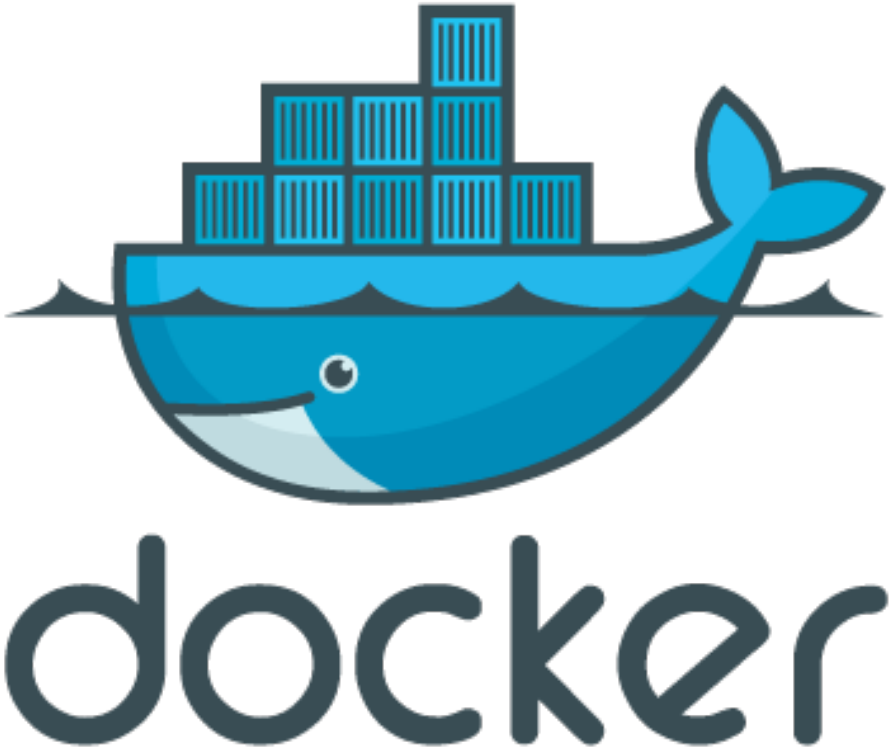
Opened by timothee-revil on Nov 10, 2014 14 comments

Connection refused during **installation** - git cloning #670

Opened by nuin on Nov 25, 2014 2 comments

Installation error #614

Docker lightweight containers



http://docker.com

Docker benefits

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

bcbio + Docker + AWS

- Bootstrap from plain AMIs to cluster
- Pull/push data from S3
- Easy interface to start/stop clusters
- Lustre and encrypted NFS filesystems
- SLURM scheduler managed with Elasticluster

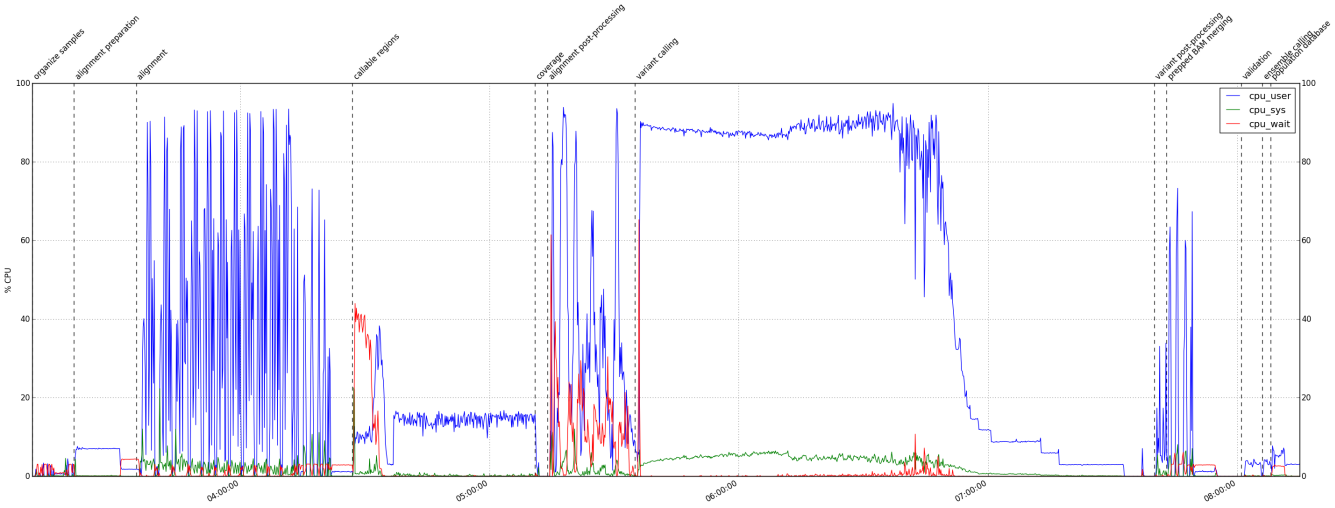
http://bcb.io/2014/12/19/awsbench/

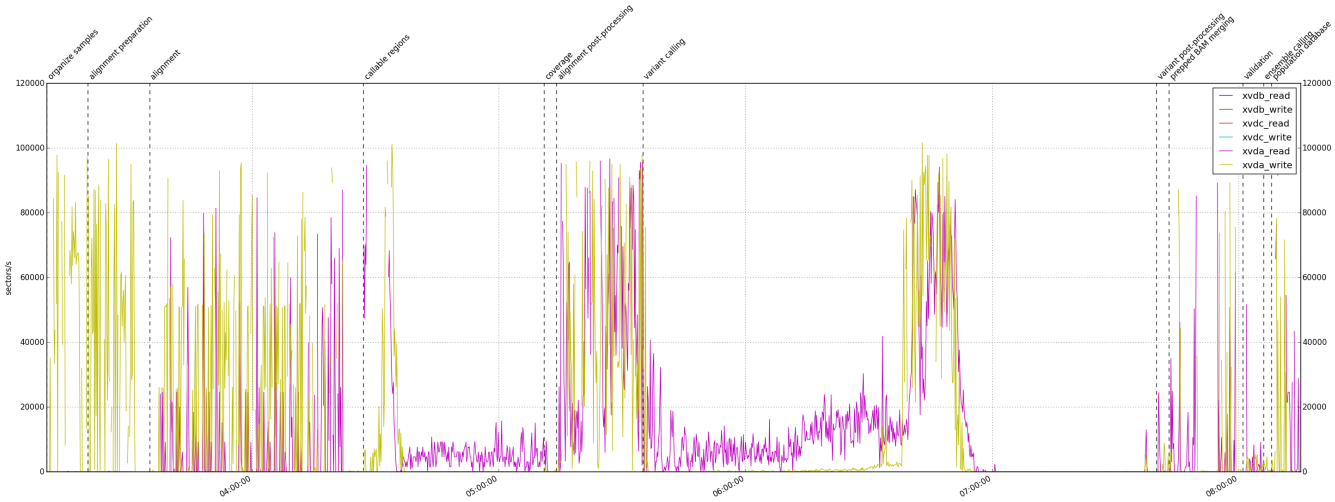
AWS benchmarking

	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





Summary

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

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