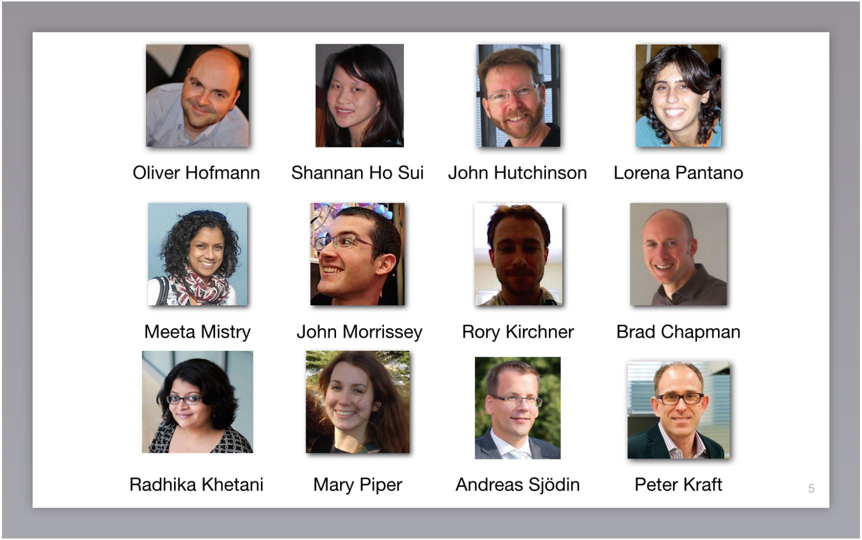
Scalable and Validated Variant Calling Work in the Bioinformatics Core

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

6 February 2015

Bioinformatics Core



Oliver Hofmann

Shannan Ho Sui

John Hutchinson

Lorena Pantano

Meeta Mistry

John Morrissey

Rory Kirchner

Brad Chapman

Radhika Khetani

Mary Piper

Andreas Sjödin

Peter Kraft

What we do

- Project design
- Analysis and consulting
- Teaching and training
- Infrastructure

Find us

FXB 202B

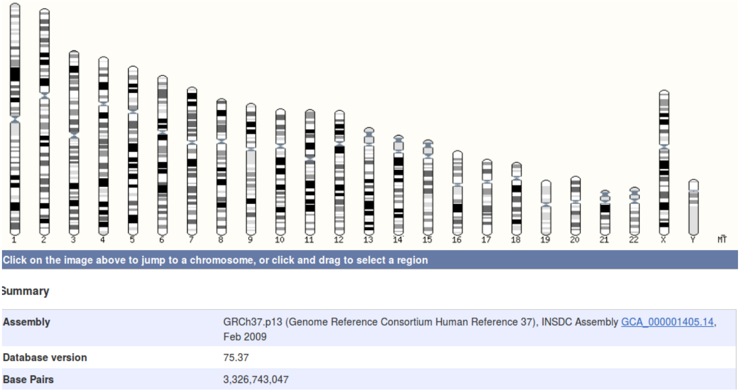
http://hsphbio.ghost.io

http://bioinformatics.hms.harvard.edu

Summary

- What is bcbio?
- Validation
- Support
- Scaling

Human whole genome sequencing



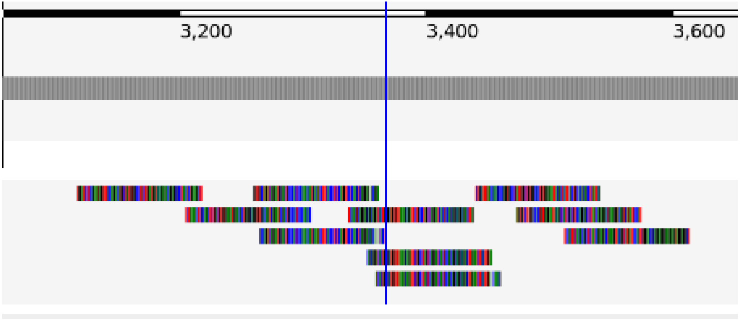
Click on the image above to jump to a chromosome, or click and drag to select a region

Summary

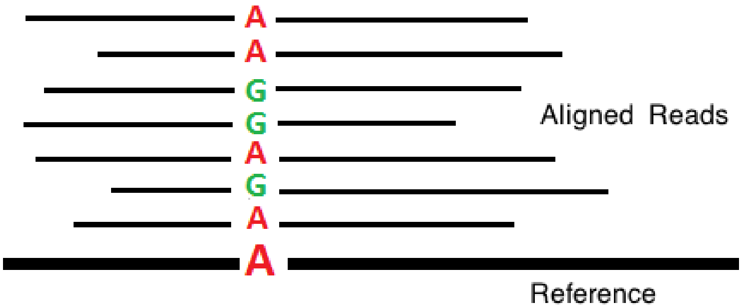
Assembly	GRCh37.p13 (Genome Reference Consortium Human Reference 37), INSDC Assembly GCA_000001405.14, Feb 2009
Database version	75.37
Base Pairs	3,326,743,047

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

High throughput sequencing

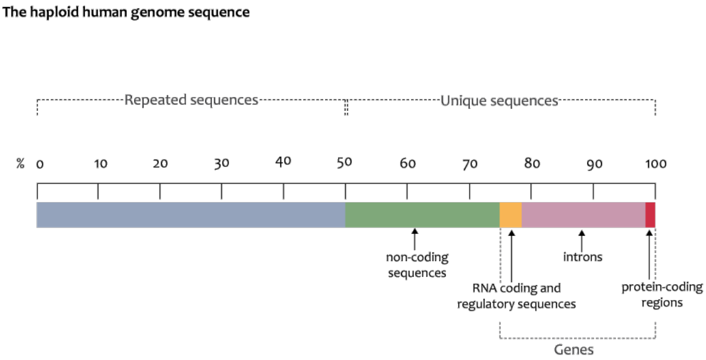


Variant calling



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

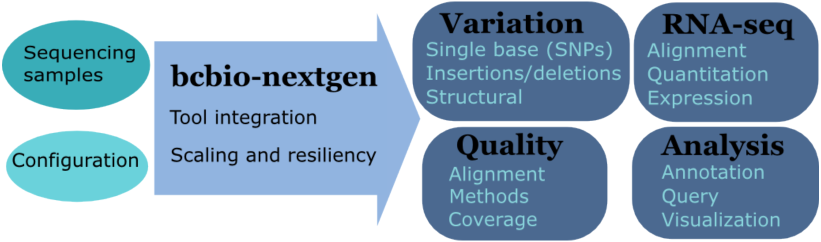
Scale: exome to whole genome



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

White box software

Overview



Sequencing samples

Configuration

bcbio-nextgen

Tool integration

Scaling and resiliency

Variation
Single base (SNPs)
Insertions/deletions
Structural

RNA-seq
Alignment
Quantitation
Expression

Quality
Alignment
Methods
Coverage

Analysis
Annotation
Query
Visualization

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

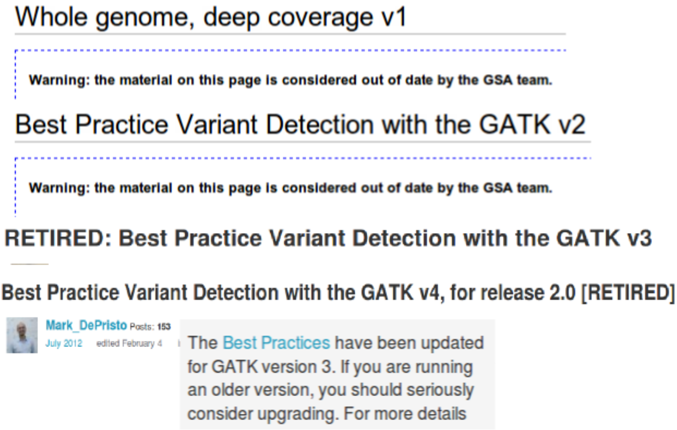
Uses

- 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
- Validation
- Scaling
- Multi-architecture parallel processing

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.

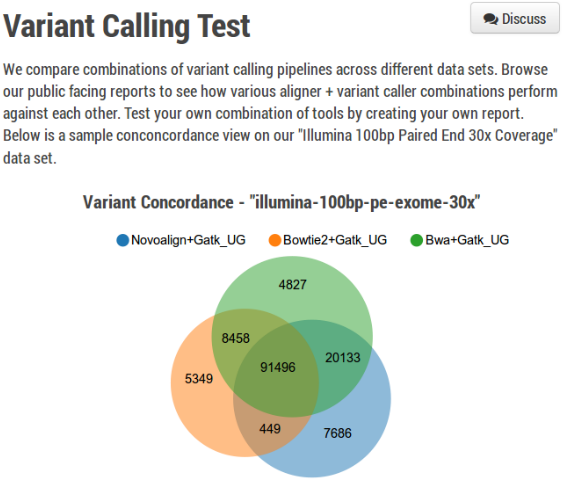
RETIRED: Best Practice Variant Detection with the GATK v3

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

Quality differences between methods



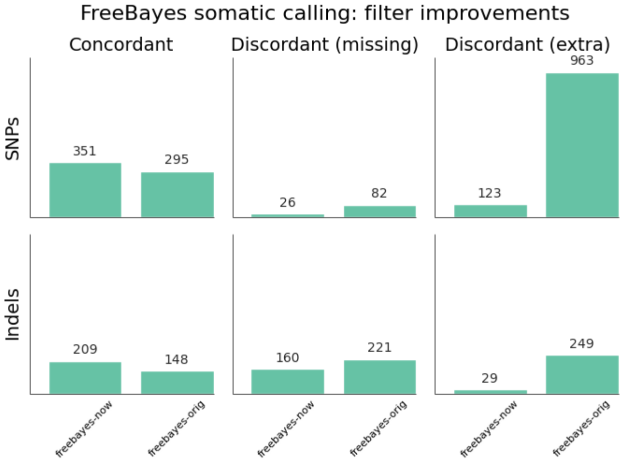
Variant Calling Test

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"

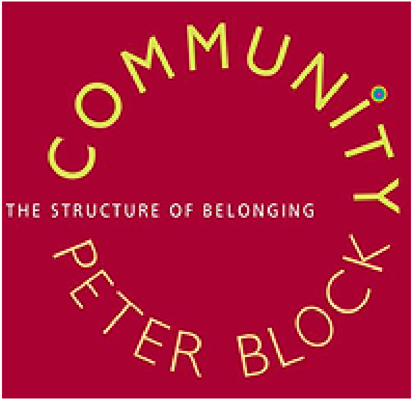
http://www.bioplanet.com/gcat

Benefits of improved filtering



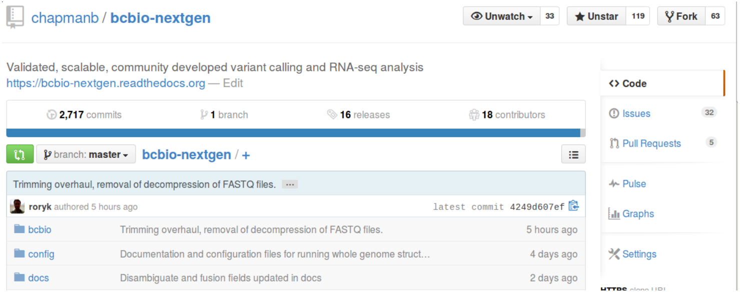
http://j.mp/cancervalpre

Solution



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

Community: contribution



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 / +

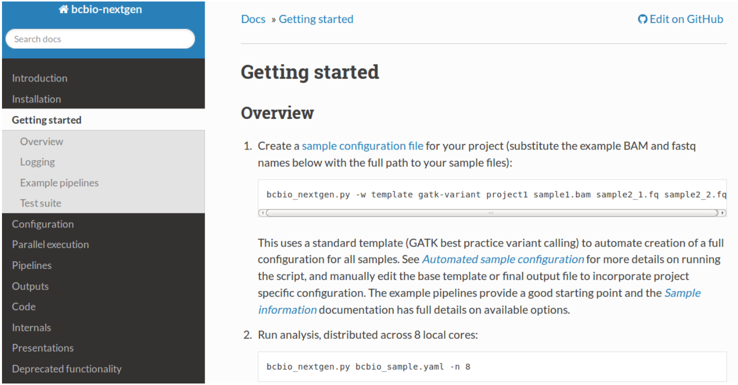
Trimming overhaul, removal of decompression of FASTQ files. …

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

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
```

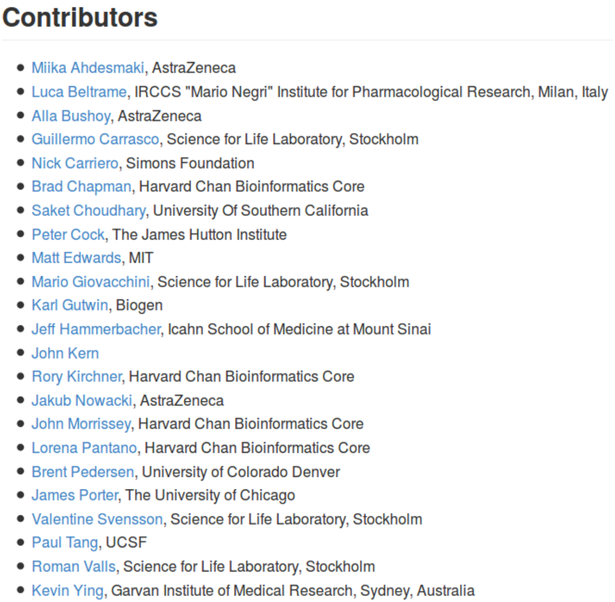
   This uses a standard template (GATK best practice variant calling) to automate creation of a full configuration for all samples. See *Automated sample configuration* for more details on running the script, and manually edit the base template or final output file to incorporate project specific configuration. 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

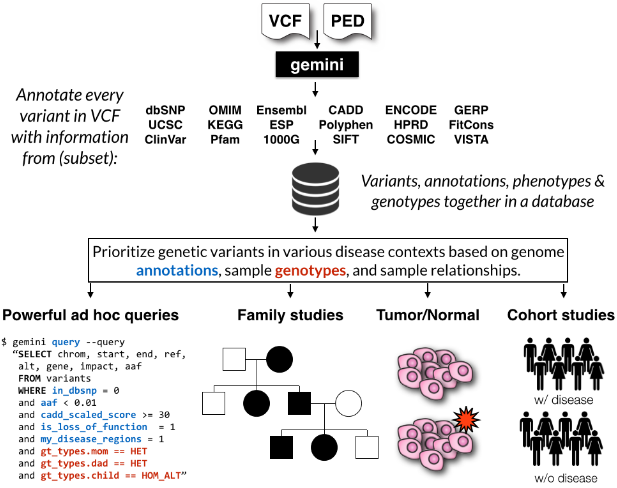
Community



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

Community: GEMINI



VCF
PED
gemini
Annotate every variant in VCF with information from (subset):
dbSNP UCSC ClinVar
OMIM KEGG Pfam
Ensembl ESP 1000G
CADD Polyphen SIFT
ENCODE HPRD COSMIC
GERP FitCons VISTA
Variants, annotations, phenotypes & genotypes together in a database
Prioritize genetic variants in various disease contexts based on genome annotations, sample genotypes, and sample relationships.
Powerful ad hoc queries
Family studies
Tumor/Normal
Cohort studies
w/ disease
w/o disease

```
$ gemini query --query
  “SELECT chrom, start, end, ref,
    alt, gene, impact, aaf
    FROM variants
    WHERE in_dbsnp = 0
    and aaf < 0.01
    and cadd_scaled_score >= 30
    and is_loss_of_function  = 1
    and my_disease_regions = 1
    and gt_types.mom == HET
    and gt_types.dad == HET
    and gt_types.child == HOM_ALT”
```

http://gemini.readthedocs.org

Validation

Tests for implementation and methods

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

Reference materials

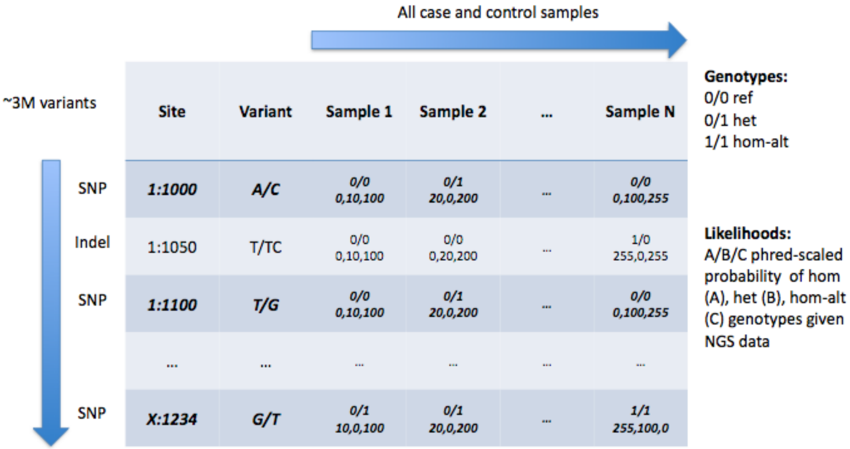
http://www.genomeinabottle.org/

Joint variant calling definitions

- Single sample calling
- Pooled calling
- Joint calling
- Squaring off/backfilling

http://j.mp/bcbiojoint

Squared off VCF



All case and control samples

~3M variants

	Site	Variant	Sample 1	Sample 2	...	Sample N
SNP	***1:1000***	***A/C***	***0/0 0,10,100***	***0/1 20,0,200***	...	***0/0 0,100,255***
Indel	1:1050	T/TC	0/0 0,10,100	0/0 0,20,200	...	1/0 255,0,255
SNP	***1:1100***	***T/G***	***0/0 0,10,100***	***0/1 20,0,200***	...	***0/0 0,100,255***
	...	...	...	...	...	...
SNP	***X:1234***	***G/T***	***0/1 10,0,100***	***0/1 20,0,200***	...	***1/1 255,100,0***

Genotypes:
0/0 ref
0/1 het
1/1 hom-alt

Likelihoods:
A/B/C phred-scaled probability of hom (A), het (B), hom-alt (C) genotypes given NGS data

http://gatkforums.broadinstitute.org/discussion/4150/should-i-analyze-my-samples-alone-or-together

Scaling and analysis flexibility

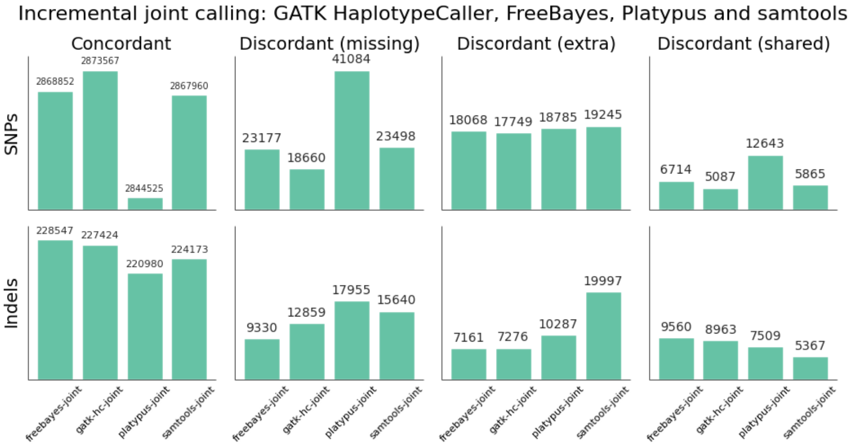
- Parallelize: call samples individually
- Add single new sample to analysis
- Combine existing populations
- Inform calls based on previously known variants

Implementation

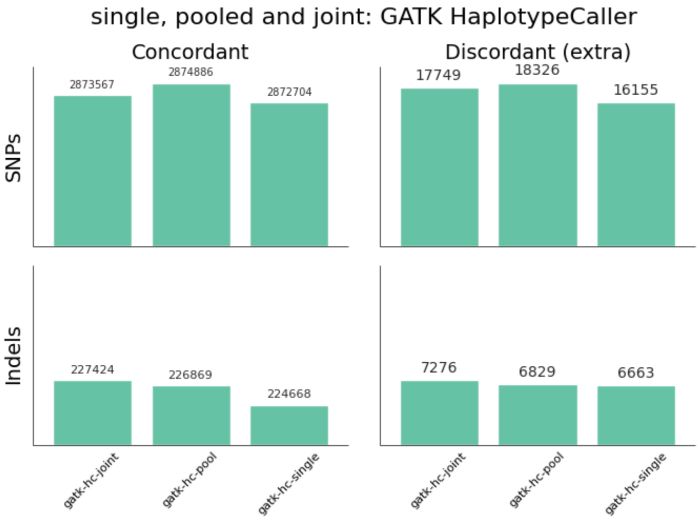
- GATK HaplotypeCaller – gVCFs
- FreeBayes – recalling
- Platypus – recalling
- samtools 1.x – recalling

https://github.com/chapmanb/bcbio.variation.recall

Multiple approaches work well



Joint vs batch vs single



Structural variations

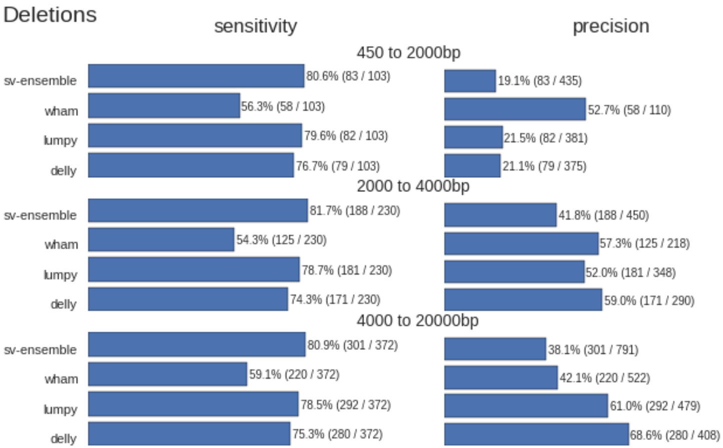
- Goal: identify regions with potential issues
- Rough boundaries for additional analysis
- Ensemble: union of all calls
- Understand sensitivity and precision

http://j.mp/bcbiosv

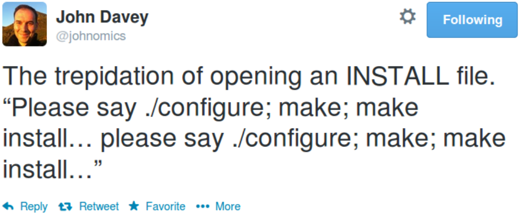
Structural variant callers

- LUMPY https://github.com/arq5x/lumpy-sv
- Delly https://github.com/tobiasrausch/delly
- cn.mops http://www.bioconductor.org/packages/release/bioc/html/cn.mops.html
- CNVkit http://cnvkit.readthedocs.org/
- WHAM https://github.com/jewmanchue/wham

Structural variant evaluation



Making bcbio easy to use



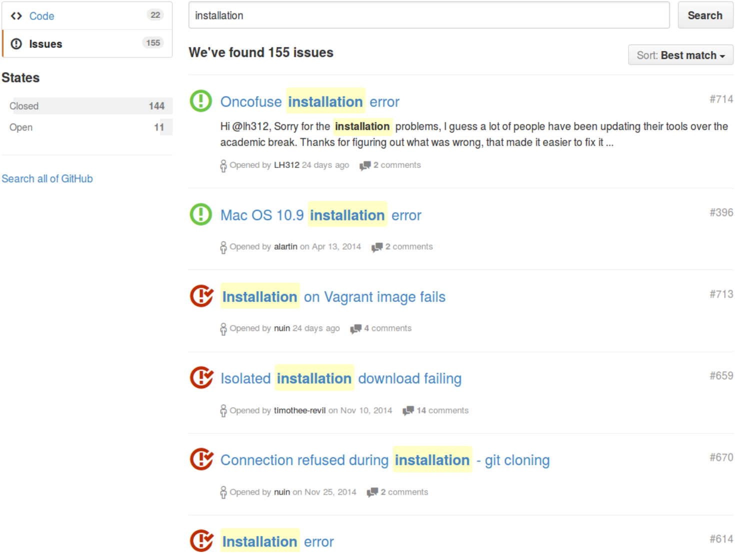
The trepidation of opening an INSTALL file. “Please say ./configure; make; make install… please say ./configure; make; make install…”

Automated Install

We made it easy to install a large number of biological tools.

Good or bad idea?

Need a consistent support environment



States

Closed	144
Open	11

Search all of GitHub

We've found 155 issues

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

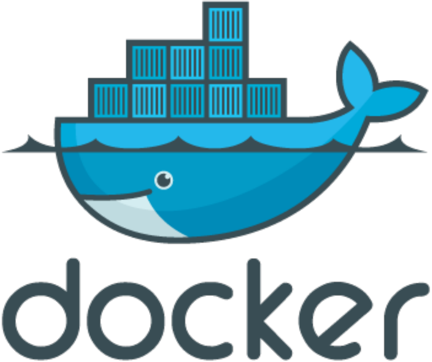
Opened by timothee-revil on Nov 10, 2014 14 comments

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 + Amazon Web Services

- Ready to run
- Easy interface to start/stop clusters
- Pull/push data from encrypted S3
- Lustre and encrypted NFS filesystems

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

Harvard HPC

- Odyssey at FAS
  `https://rc.fas.harvard.edu/`

- Orchestra at HMS
  `https://rc.hms.harvard.edu/`

Scaling: Start point

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

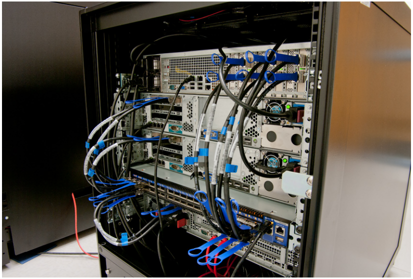
Scaling: End point

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
```

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: shared filesystem

480 cores, 30 samples

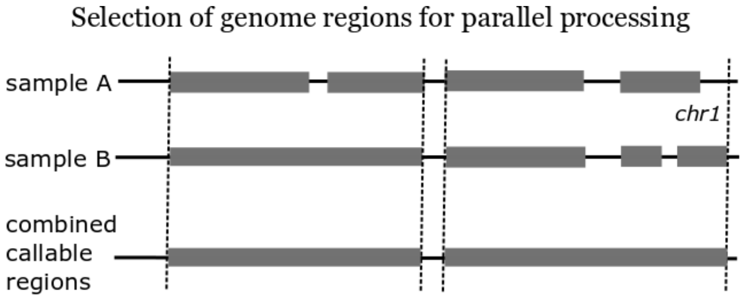
Step	Lustre	NFS
alignment	4.5h	6.1h
alignment post-processing	7.0h	20.7h

James Cuff, John Morrissey (FAS)
Kristina Kermanshahche (Intel)

Scaling: avoid intermediates

```
("{bwa} mem -M -t {num_cores} -R ’{rg_info}’ -v 1 "
 "  {ref_file} {fastq_file} {pair_file} "
 "| {samblaster} "
 "| {samtools} sort -@ {cores} -m {mem} -T {tmp_file}"
 "   -o {tx_out_file} /dev/stdin")
```

Scaling: Parallel by genome

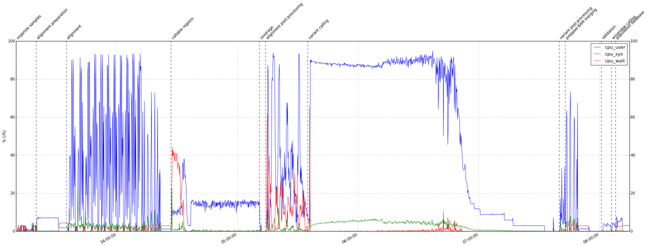

Selection of genome regions for parallel processing
sample A
chr1
sample B
combined callable regions

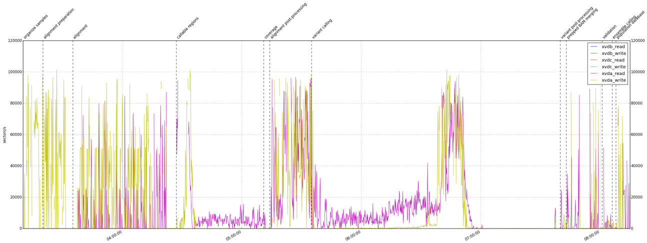
Scaling: AWS benchmarking

	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)

Scaling: Resource usage plots





Summary

- bcbio – community built variant calling and RNA-seq analyses
- Validation – measure quality = good science
- Support – AWS and local HPC
- Scaling – diverse teams

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