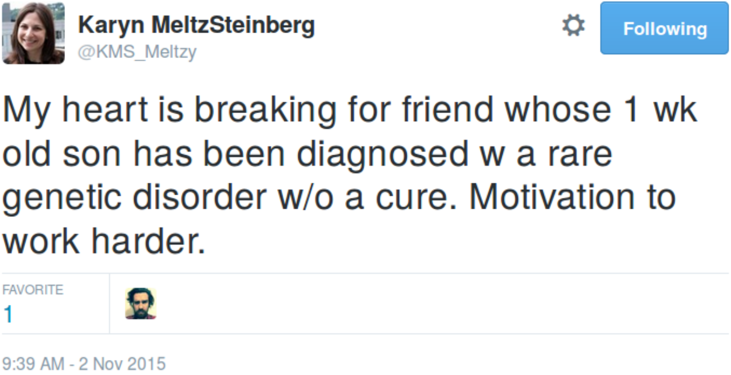
Improving support and distribution of validated analysis tools

Brad Chapman
Bioinformatics Core, Harvard Chan School
https://bcb.io
http://j.mp/bcbiolinks

9 November 2015

We need to do science faster



Karyn MeltzSteinberg
@KMS_Meltzy

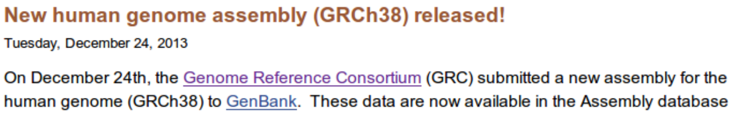
My heart is breaking for friend whose 1 wk old son has been diagnosed w a rare genetic disorder w/o a cure. Motivation to work harder.

FAVORITE
1

9:39 AM - 2 Nov 2015

https://twitter.com/KMS_Meltzy/status/661206070308794368

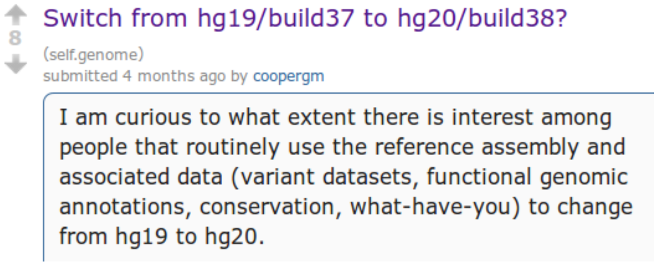
We need to incorporate improvements faster



New human genome assembly (GRCh38) released!

Tuesday, December 24, 2013

On December 24th, the Genome Reference Consortium (GRC) submitted a new assembly for the human genome (GRCh38) to GenBank. These data are now available in the Assembly database



8

Switch from hg19/build37 to hg20/build38?

(self.genome)

submitted 4 months ago by coopergm

I am curious to what extent there is interest among people that routinely use the reference assembly and associated data (variant datasets, functional genomic annotations, conservation, what-have-you) to change from hg19 to hg20.

https://www.reddit.com/r/genome/comments/3b3s3t/switch_from_hg19build37_to_hg20build38/

Daily bioinformatics work

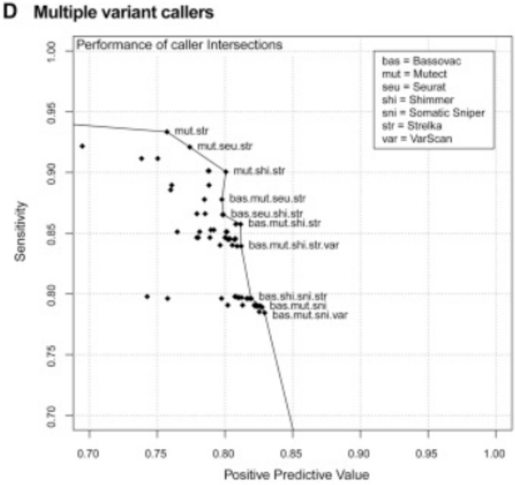
- Install tools
- Put tools together
- Test and validate
- Improve algorithms
- Scale
- Read literature
- Do biology

Standard analyses not routine

> *Four major genome centers predicted single-nucleotide variants (SNVs) for The Cancer Genome Atlas (TCGA) lung cancer samples, but only 31.0% (1,667/5,380) of SNVs were identified by all four.*

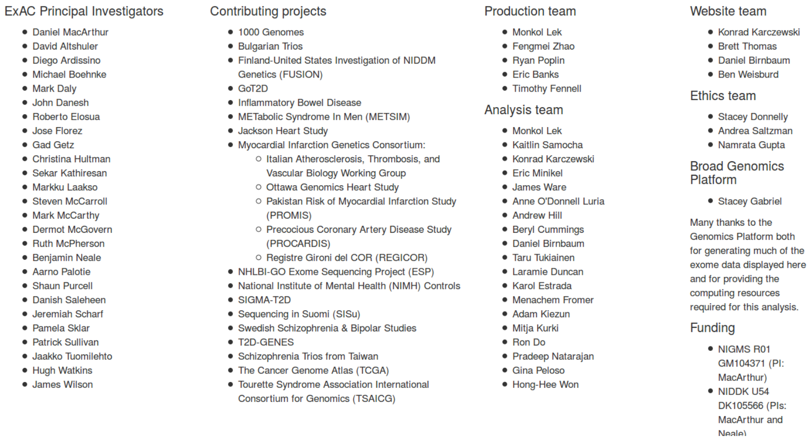
http://www.nature.com/nmeth/journal/vaop/ncurrent/full/nmeth.3407.html

Combining analyses



http://www.cell.com/cell-systems/abstract/S2405-4712%2815%2900113-1

Working together produces great things



ExAC Principal Investigators

- Daniel MacArthur
- David Altshuler
- Diego Ardissino
- Michael Boehnke
- Mark Daly
- John Danesh
- Roberto Elosua
- Jose Florez
- Gad Getz
- Christina Hultman
- Sekar Kathiresan
- Markku Laakso
- Steven McCarroll
- Mark McCarthy
- Dermot McGovern
- Ruth McPherson
- Benjamin Neale
- Aarno Palotie
- Shaun Purcell
- Danish Saleheen
- Jeremiah Scharf
- Pamela Sklar
- Patrick Sullivan
- Jaakko Tuomilehto
- Hugh Watkins
- James Wilson

Contributing projects

- 1000 Genomes
- Bulgarian Trios
- Finland-United States Investigation of NIDDM Genetics (FUSION)
- GoT2D
- Inflammatory Bowel Disease
- METabolic Syndrome In Men (METSIM)
- Jackson Heart Study
- Myocardial Infarction Genetics Consortium:
  - Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group
  - Ottawa Genomics Heart Study
  - Pakistan Risk of Myocardial Infarction Study (PROMIS)
  - Precocious Coronary Artery Disease Study (PROCARDIS)
  - Registre Gironi del COR (REGICOR)
- NHLBI-GO Exome Sequencing Project (ESP)
- National Institute of Mental Health (NIMH) Controls
- SIGMA-T2D
- Sequencing in Suomi (SISu)
- Swedish Schizophrenia & Bipolar Studies
- T2D-GENES
- Schizophrenia Trios from Taiwan
- The Cancer Genome Atlas (TCGA)
- Tourette Syndrome Association International Consortium for Genomics (TSAICG)

Production team

- Monkol Lek
- Fengmei Zhao
- Ryan Poplin
- Eric Banks
- Timothy Fennell

Analysis team

- Monkol Lek
- Kaitlin Samocha
- Konrad Karczewski
- Eric Minikel
- James Ware
- Anne O'Donnell Luria
- Andrew Hill
- Beryl Cummings
- Daniel Birnbaum
- Taru Tukiainen
- Laramie Duncan
- Karol Estrada
- Menachem Fromer
- Adam Kiezun
- Mitja Kurki
- Ron Do
- Pradeep Natarajan
- Gina Peloso
- Hong-Hee Won

Website team

- Konrad Karczewski
- Brett Thomas
- Daniel Birnbaum
- Ben Weisburd

Ethics team

- Stacey Donnelly
- Andrea Saltzman
- Namrata Gupta

Broad Genomics Platform

- Stacey Gabriel

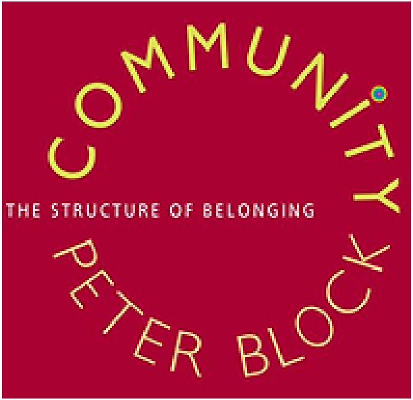
Many thanks to the Genomics Platform both for generating much of the exome data displayed here and for providing the computing resources required for this analysis.

Funding

- NIGMS R01 GM104371 (PI: MacArthur)
- NIDDK U54 DK105566 (PIs: MacArthur and Neale)

http://exac.broadinstitute.org/about

Solution



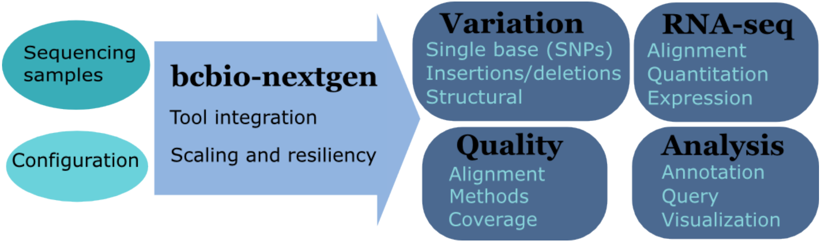
http://www.amazon.com/Community-Structure-Belonging-Peter-Block/dp/1605092770

Large scale infrastructure development

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

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, novoalign, bowtie2
- Variantion: FreeBayes, GATK, VarDict, MuTecT, Scalpel, SnpEff, VEP, GEMINI, Lumpy, Manta, CNVkit
- RNA-seq: Tophat, STAR, cufflinks, Sailfish, HISAT2
- Quality control: fastqc, bamtools, RNA-SeQC
- Manipulation: bedtools, bcftools, biobambam, sambamba, samblaster, samtools, vcflib, vt

Provides

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

We made a pipeline – so what?

> *There have been a number of previous efforts to create publicly available analysis pipelines for high throughput sequencing data. Examples include Omics-Pipe, bcbio-nextgen, TREVA and NGSane. These pipelines offer a comprehensive, automated process that can analyse raw sequencing reads and produce annotated variant calls. However, the main audience for these pipelines is the research community. Consequently, there are many features required by clinical pipelines that these examples do not fully address. Other groups have focused on improving specific features of clinical pipelines. The Churchill pipeline uses specialised techniques to achieve high performance, while maintaining reproducibility and accuracy. However it is not freely available to clinical centres and it does not try to improve broader clinical aspects such as detailed quality assurance reports, robustness, reports and specialised variant filtering. The Mercury pipeline offers a comprehensive system that addresses many clinical needs: it uses an automated workflow system (Valence) to ensure robustness, abstract computational resources and simplify customisation of the pipeline. Mercury also includes detailed coverage reports provided by ExCID, and supports compliance with US privacy laws (HIPAA) when run on DNANexus, a cloud computing platform specialised for biomedical users. Mercury offers a comprehensive solution for clinical users, however it does not achieve our desired level of transparency, modularity and simplicity in the pipeline specification and design. Further, Mercury does not perform specialised variant filtering and prioritisation that is specifically tuned to the needs of clinical users.*

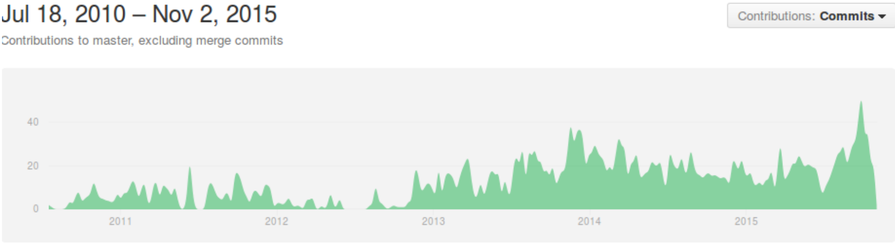
http://www.genomemedicine.com/content/7/1/68

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

Community: sustainability

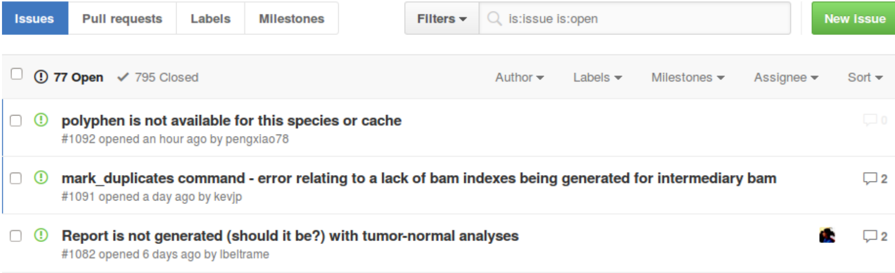


Jul 18, 2010 – Nov 2, 2015

Contributions to master, excluding merge commits

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

Community: support



□ ⓘ 77 Open ✓ 795 Closed

□ ⓘ polyphen is not available for this species or cache
#1092 opened an hour ago by pengxiao78 — 0

□ ⓘ mark_duplicates command - error relating to a lack of bam indexes being generated for intermediary bam
#1091 opened a day ago by kevjp — 2

□ ⓘ Report is not generated (should it be?) with tumor-normal analyses
#1082 opened 6 days ago by lbeltrame — 2

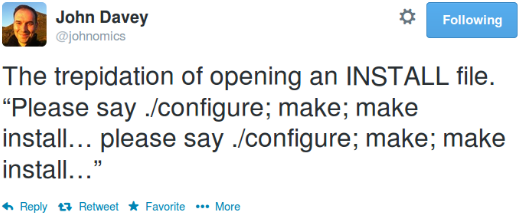
https://bcbio-nextgen.readthedocs.org

Parts of bcbio

What can we replace?

- Installation
- Infrastructure – runs on your cluster
- Tool integration
- Validation – stability
- Rapid development – new improvements

Installation



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?

bcbio + Docker + AWS

- bcbio tools + code in Docker containers
- Bootstrap from plain AMIs to cluster
- Pull/push data from S3
- Lustre and encrypted NFS filesystems

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

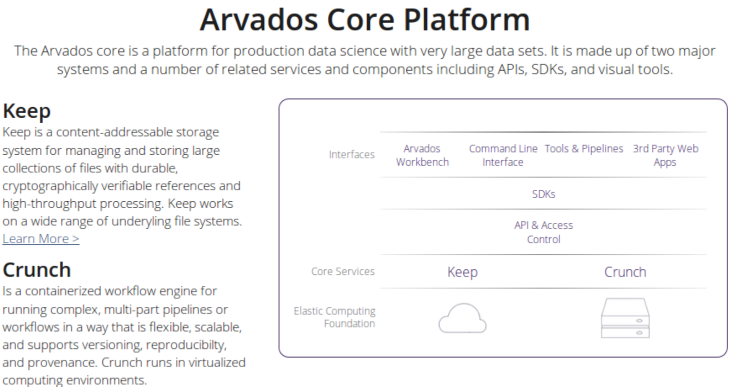
https://github.com/chapmanb/bcbio-nextgen-vm

Common Workflow Language

- Standard way to describe workflows
- Explicit markup of input/output files
- Automatically generated by bcbio
- Run on multiple infrastructures
- Community

https://github.com/chapmanb/bcbio-nextgen/tree/master/cwl

Infrastructure



Arvados Core Platform

The Arvados core is a platform for production data science with very large data sets. It is made up of two major systems and a number of related services and components including APIs, SDKs, and visual tools.

Keep

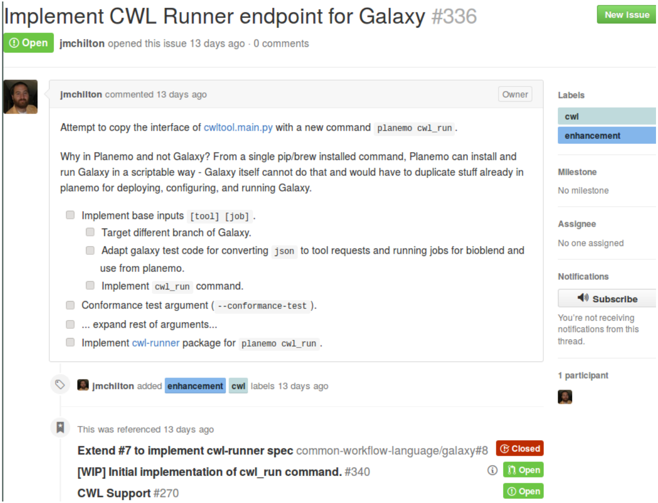
Keep is a content-addressable storage system for managing and storing large collections of files with durable, cryptographically verifiable references and high-throughput processing. Keep works on a wide range of underlying file systems. Learn More >

Crunch

Is a containerized workflow engine for running complex, multi-part pipelines or workflows in a way that is flexible, scalable, and supports versioning, reproducibilty, and provenance. Crunch runs in virtualized computing environments.

https://arvados.org/

Infrastructure



Implement CWL Runner endpoint for Galaxy #336

New Issue

ⓘ Open **jmchilton** opened this issue 13 days ago · 0 comments

jmchilton commented 13 days ago — Owner

Attempt to copy the interface of cwltool.main.py with a new command `planemo cwl_run`.

Why in Planemo and not Galaxy? From a single pip/brew installed command, Planemo can install and run Galaxy in a scriptable way - Galaxy itself cannot do that and would have to duplicate stuff already in planemo for deploying, configuring, and running Galaxy.

- □ Implement base inputs `[tool] [job]`.
  - □ Target different branch of Galaxy.
  - □ Adapt galaxy test code for converting `json` to tool requests and running jobs for bioblend and use from planemo.
  - □ Implement `cwl_run` command.
- □ Conformance test argument (`--conformance-test`).
- □ ... expand rest of arguments...
- □ Implement cwl-runner package for `planemo cwl_run`.

jmchilton added **enhancement** **cwl** labels 13 days ago

This was referenced 13 days ago

Extend #7 to implement cwl-runner spec common-workflow-language/galaxy#8 — Closed

[WIP] Initial implementation of cwl_run command. #340 — Open

CWL Support #270 — Open

Labels

cwl

enhancement

Milestone

No milestone

Assignee

No one assigned

Notifications

Subscribe

You're not receiving notifications from this thread.

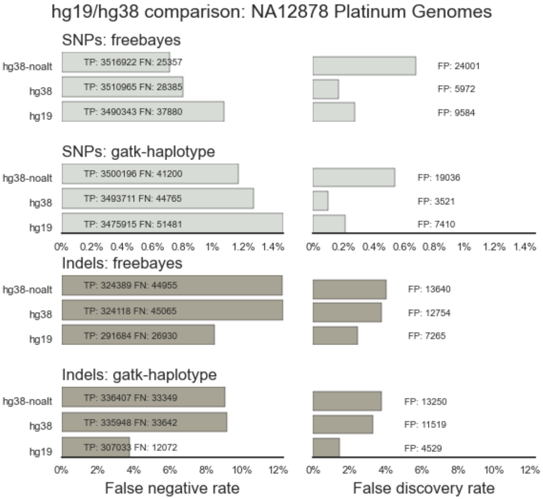
1 participant

https://github.com/galaxyproject/planemo

Vision: pluggable components

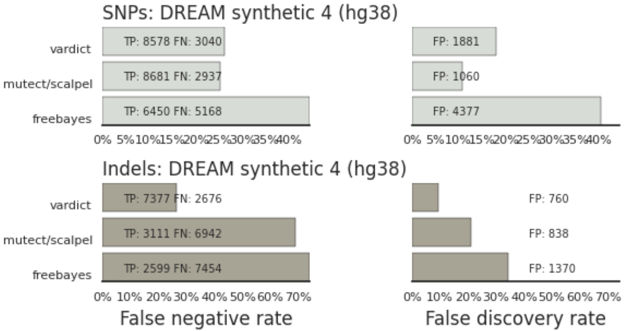
- CWL description + platforms for run
- Docker containers with tools + code
- Mix and match implementations
- Do research and development and production in same environment

Practical: Human build 38 validation



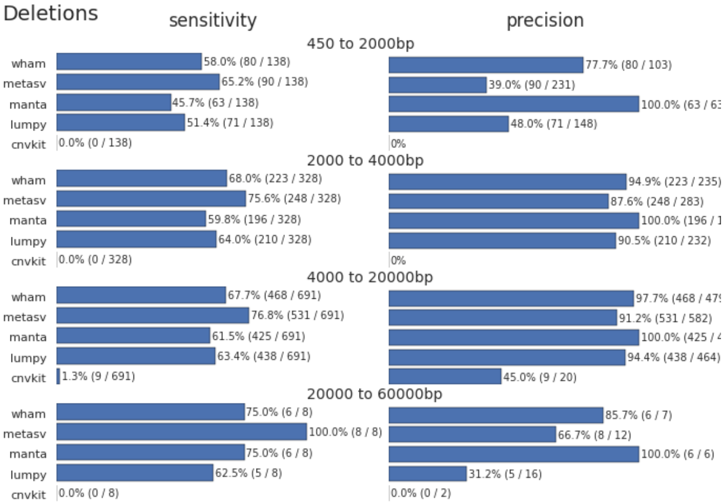
http://bcb.io/2015/09/17/hg38-validation/

Practical: cancer validation



https://github.com/bcbio/bcbio.github.io/blob/master/_posts/2015-10-05-vardict-filtering.md

Practical: structural variant calling



Summary

- Do more science faster
- Community – integrate, not re-implement
- Docker + CWL enables integration
- Let's talk about ways to work together