Interoperable community developed variant calling with bcbio and the Common Workflow Language

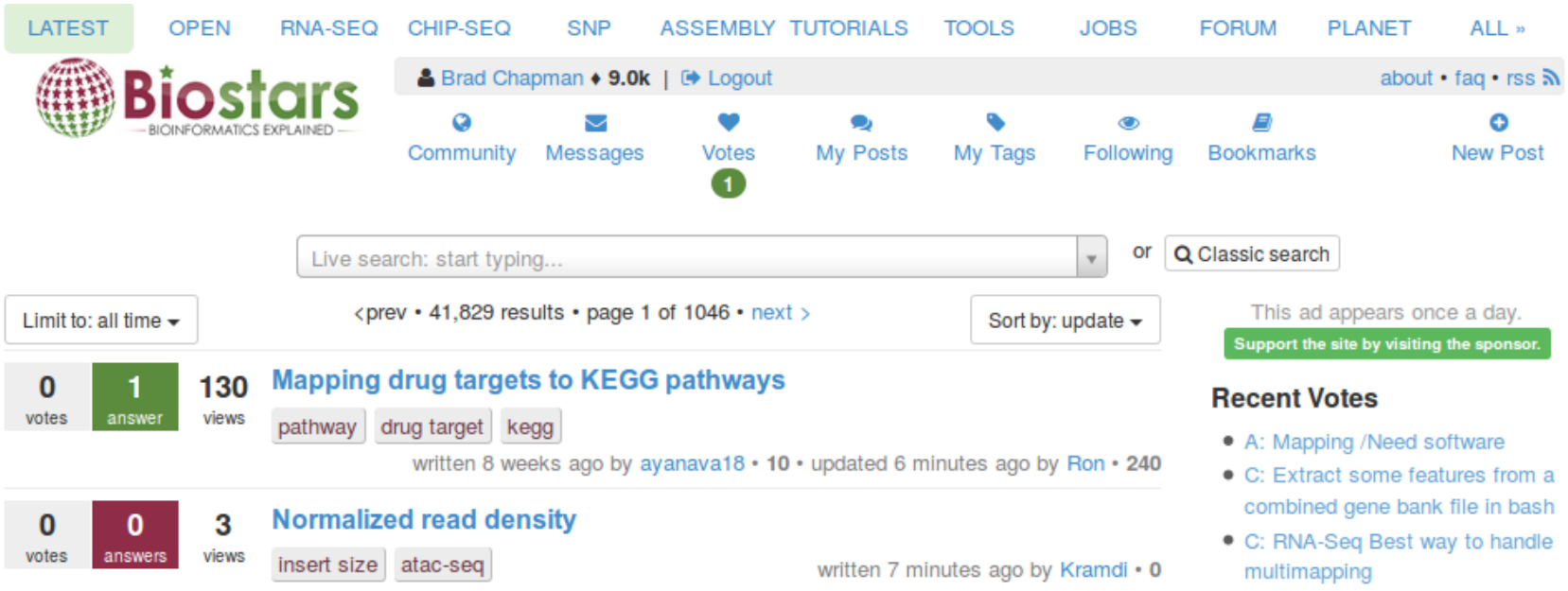
Brad Chapman
Bioinformatics Core, Harvard Chan School
https://bcb.io
http://j.mp/bcbiolinks

3 November 2016

Overview

- Open source communities
- bcbio: community developed analyses
- Value of variant validation
- Interoperable infrastructure

Supporting bioinformatics



https://www.biostars.org/

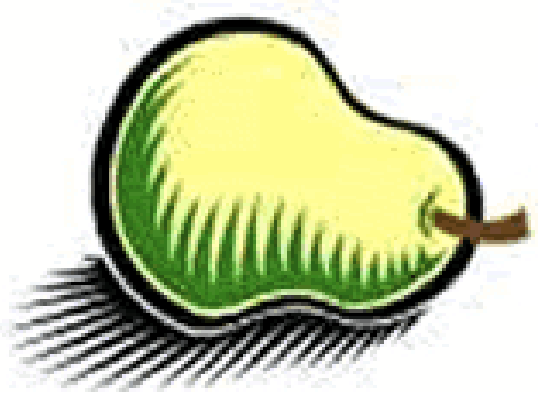
Open source community

OBF

Page Discussion

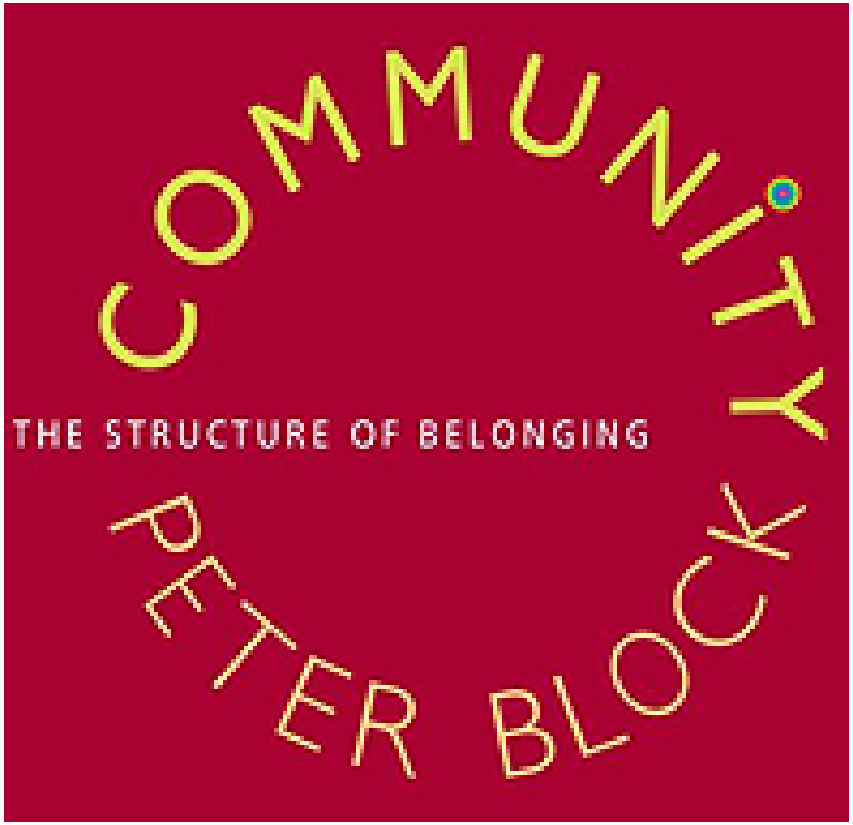
BOSC 2016

Main Page
Projects
News
BOSC
OBF Board
Join



http://www.open-bio.org/wiki/BOSC_2016

Build communities



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

Working together produces great things

Principal Investigators

Daniel MacArthur
Aarno Palotie
Andres Metspalu
Anne Remes
Adolfo Correa
Andre Franke
Ann Pulver
Ben Glaser
Ben Neale
Bong-Jo Kim
Carlos Pato
Carlos A Aguilar Salinas
Christina Hultman
Christine M. Albert
Christopher Haiman
Cliceria Gonzalez
Colin Palmer
Craig Hanis
Dan Roden
Dan Turner
Dana Dabelea
Daniel Chasman
Danish Saleheen
David Altshuler
David Goldstein
Dawood Darbar
Dermot McGovern
Diego Ardissino
Donald Bowden
Emelia J. Benjamin
Erkki Vartiainen
Erwin Bottinger
Gad Getz
George Kirov
Gil Atzmon
Harlan M. Krumholz
Harry Sokol
Heribert Schunkert
Hilkka Soininen
Hugh Watkins
Jaakko Kaprio
Jaana Suvisaari
James Meigs
James Wilson
Jaspal Kooner
Jaume Marrugat
Jeanette Erdmann
Jeremiah Scharf
John Barnard
John Chambers
John D. Rioux
Jose Florez
Josée Dupuis
Judy Cho
Juliana Chan
Kyong Soo Park
Leif Groop
Lorena Orozco
Lori Bonnycastle
Maija Wessman
Mark Daly
Mark McCarthy
Markku Laakso
Martti Färkkilä
Matthew Bown
Matthew Harms
Matti Holi
Michael Boehnke
Michael O'Donovan
Michael Owen
Mikko Hiltunen
Mikko Kallela
Mina Chung
Ming Tsuang
Moore Shoemaker
Nazneen Rahman
Nilesh Samani
Olle Melander
Pamela Sklar
Patrick T. Ellinor
Patrick Sullivan
Peter Nilsson
Phil Zeitler
Ramnik Xavier
Ravindranath Duggirala
Rinse Weersma
Roberto Elosua
Ronald Ma
Ruth Loos
Ruth McPherson
Samuli Ripatti
Sekar Kathiresan
Seppo Koskinen
Soo Heon Kwak
Stephen Glatt
Steve McCarroll
Steven A. Lubitz
Subra Kugathasan
Tai Shyong
Tariq Ahmad
Teresa Tusie Luna
Terho Lehtimäki
Tim Spector
Tõnu Esko
Tuomi Tiinamaija
Veikko Salomaa
Yik Ying Teo
Young Jin Kim

Contributing projects

1000 Genomes
1958 Birth Cohort
ALSGEN
Alzheimer's Disease Sequencing Project (ADSP)
Atrial Fibrillation Genetics Consortium (AFGen)
Estonian Genome Center, University of Tartu (EGCUT)
Bulgarian Trios
Finland-United States Investigation of NIDDM Genetics (FUSION)
Finnish Twin Cohort Study
FINN-ADGEN
FINRISK
Framingham Heart Study
Génome Québec - Genizon Biobank
Genomic Psychiatry Cohort
GoT2D
Genotype-Tissue Expression Project (GTEx)
Health2000
Inflammatory Bowel Disease:
- Helsinki University Hospital Finland
- NIDDK IBD Genetics Consortium
- Quebec IBD Genetics Consortium

Jackson Heart Study
Kuopio Alzheimer Study
LifeLines Cohort
MESTA
METabolic Syndrome In Men (METSIM)
Finnish Migraine Study
Myocardial Infarction Genetics Exome Sequencing Consortium (MIGenExS):
- Leicester Exome Seq
- Ottawa Genomics Heart Study
- Pakistan Risk of Myocardial Infarction Study (PROMIS)
- Precocious Coronary Artery Disease Study (PROCARDIS)
- Registre Gironi del COR (REGICOR)

National Institute of Mental Health (NIMH) Controls
NHLBI-GO Exome Sequencing Project (ESP)
NHLBI TOPMed
North GerMIF Exomes
Schizophrenia Trios from Taiwan
Sequencing Initiative Suomi (SiSu)
SIGMA-T2D
South GerMIF Exomes
Swedish Schizophrenia & Bipolar Studies
T2D-GENES
- GoDARTS

T2D-SEARCH

Production team

Eric Banks
Charlotte Tolonen
Dave Shiga
Fengmei Zhao
Jose Soto
Kathleen Tibbetts
Laura Gauthier
Monkol Lek
Ryan Poplin
Sam Novod
Valentin Ruano-Rubio
Yossi Farjoun

Analysis team

Konrad Karczewski
Laurent Francioli
Kristen Laricchia
Monkol Lek
Anne O'Donnell Luria
Beryl Cummings
Daniel Birnbaum
Eric Minikel
James Ware
Kaitlin Samocha
Laramie Duncan
Cotton Seed
Tim Poterba
Mark Daly
Ben Neale

Website team

Ben Weisburd
Konrad Karczewski
Matthew Solomonson
Daniel Birnbaum

Ethics team

Jessica Altöldi
Andrea Saltzman
Molly Schleicher
Namrata Gupta
Stacey Donnelly

Broad Genomics Platform

Stacey Gabriel
Steven Ferriera
Susanna Hamilton

Funding

NIGMS R01 GM104371
(PI: MacArthur)

NIDDK U54 DK105566
(PIs: MacArthur and Neale)

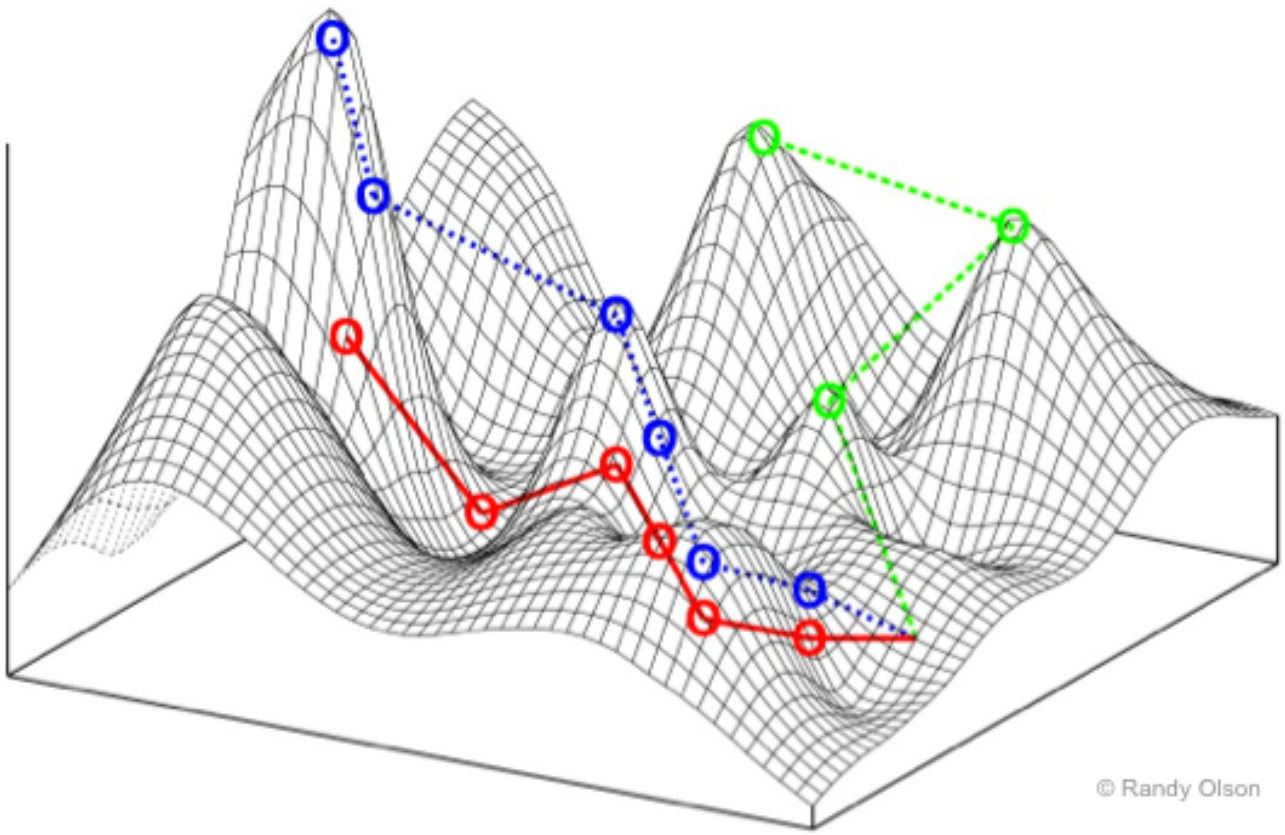
The vast majority of the data storage, computing resources, and human effort used to generate this call set were donated by the Broad Institute.

http://gnomad.broadinstitute.org/about

Challenge: many communities

https://galaxyproject.org/
http://www.cbioportal.org/
https://www.synapse.org/

Challenge: open source communities not yet optimal



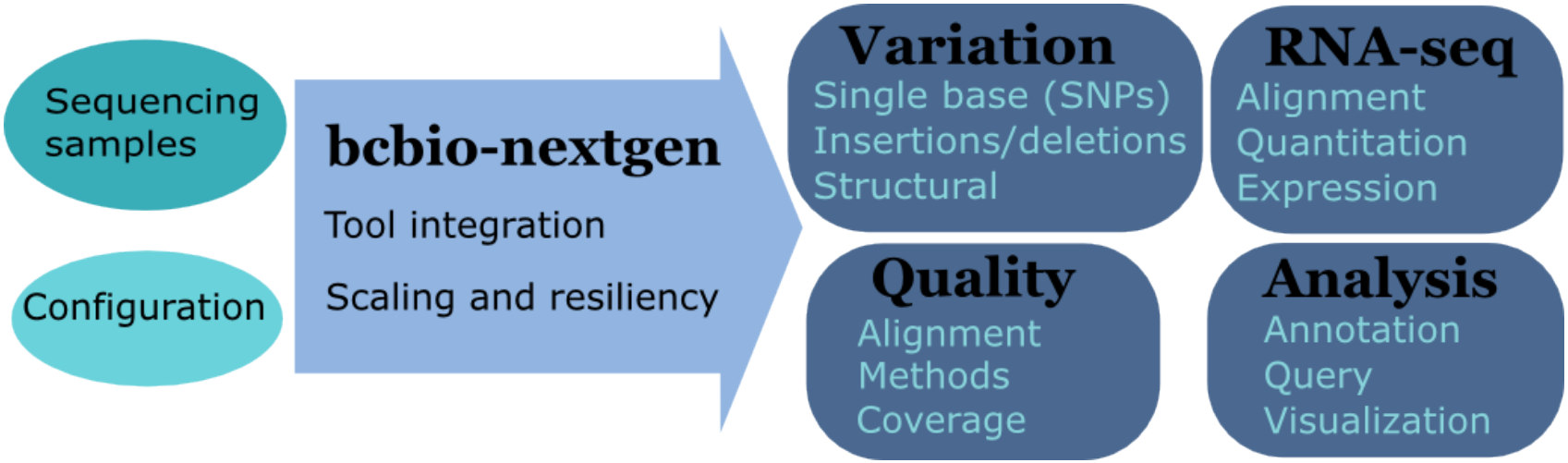
https://en.wikipedia.org/wiki/Fitness_landscape

Large scale infrastructure development

- Shared problems – academic, industry, startups
- Community developed analyses
- Validation
- Scaling
- Supporting a community of users

White box software

Overview



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

High level configuration

```
- analysis: variant2
  genome_build: hg38
  algorithm:
    aligner: bwa
    mark_duplicates: true
    recalibrate: false
    realign: false
    variantcaller: [gatk-haplotype, freebayes, vardict]
    ensemble:
      numpass: 2
    svcaller: [lumpy, manta]
```

https://bcbio-nextgen.readthedocs.org/en/latest/contents/configuration.html

Uses

- Aligners: bwa, novoalign, bowtie2, HISAT2
- Variantion: FreeBayes, GATK, VarDict, MuTecT2, Scalpel, SnpEff, VEP, GEMINI, Lumpy, Manta, CNVkit, WHAM
- RNA-seq: Tophat, STAR, Cufflinks, Sailfish
- Quality control: FastQC, samtools, Qualimap, MultiQC
- Manipulation: bedtools, bcftools, biobambam, picard, sambamba, samblaster, samtools, vcflib, vt

Provides

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

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

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.

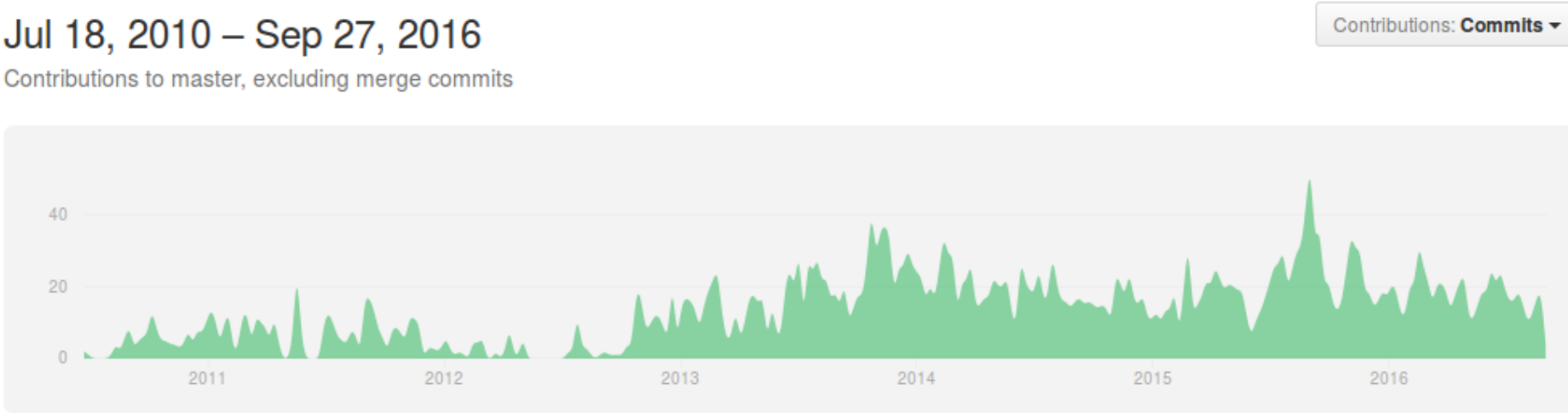
RETIRED: Best Practice Variant Detection with the GATK v3

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

Community: sustainability



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

Community: support

95 Open ✓ 1,215 Closed Author Labels Milestones Assignee Sort

update yaml templates
#1575 opened 3 minutes ago by saboswell

HG38 and Gemini 7
#1573 opened a day ago by matthdsm

Test run error 2
#1572 opened 4 days ago by firatuyulur

vep annotation fields + hgvs 7
#1571 opened 4 days ago by matthdsm

how to force bam to stream directly to bwa? 2
#1567 opened 6 days ago by brentp

Would it be possible to run the QC stage in parallel? 14
#1556 opened 14 days ago by NeillGibson

consider samtools depth to replace sambamba | bedtools in callable 21
#1549 opened 18 days ago by brentp

https://bcbio-nextgen.readthedocs.org

Community: contribution

chapmanb / **bcbio-nextgen**

Unwatch 74 | Unstar 342 | Fork 172

Code | Issues 95 | Pull requests 4 | Projects 0 | Pulse | Graphs | Settings

Validated, scalable, community developed variant calling, RNA-seq and small RNA analysis https://bcbio-nextgen.readthedocs.org — Edit

5,060 commits | **2** branches | **35** releases | **43** contributors | MIT

Branch: **master** | New pull request | Create new file | Upload files | Find file | Clone or download

chapmanb Docs: how to change logging directory location — Latest commit `f91a5e8` 3 hours ago

artwork	add logo to README and docs	a year ago
bcbio	VEP: add support for HGVS with clinical_reporting	8 hours ago
config	Update resources to match new bgzipped BED files	9 days ago
docs	Docs: how to change logging directory location	3 hours ago

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

Community: documentation

bcbio-nextgen

Search docs

Introduction
Installation
Getting started
Overview
Logging
Example pipelines
Test suite
Configuration
Parallel execution
Pipelines
Outputs
Code
Internals
Presentations
Deprecated functionality

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

Supported analysis types

- ⊟ **Pipelines**
  - ⊟ **Germline variant calling**
    - Basic germline calling
    - Population calling
  - Cancer variant calling
  - Structural variant calling
  - RNA-seq
  - single-cell RNA-seq
  - smallRNA-seq
  - ChIP-seq

https://bcbio-nextgen.readthedocs.org/en/latest/contents/pipelines.html

Value of validation

- Integration tests for pipelines
- Unbiased algorithm comparisons
- Baseline for improving methods

Reference materials

Genome in a Bottle Consortium

Global Alliance for Genomics & Health

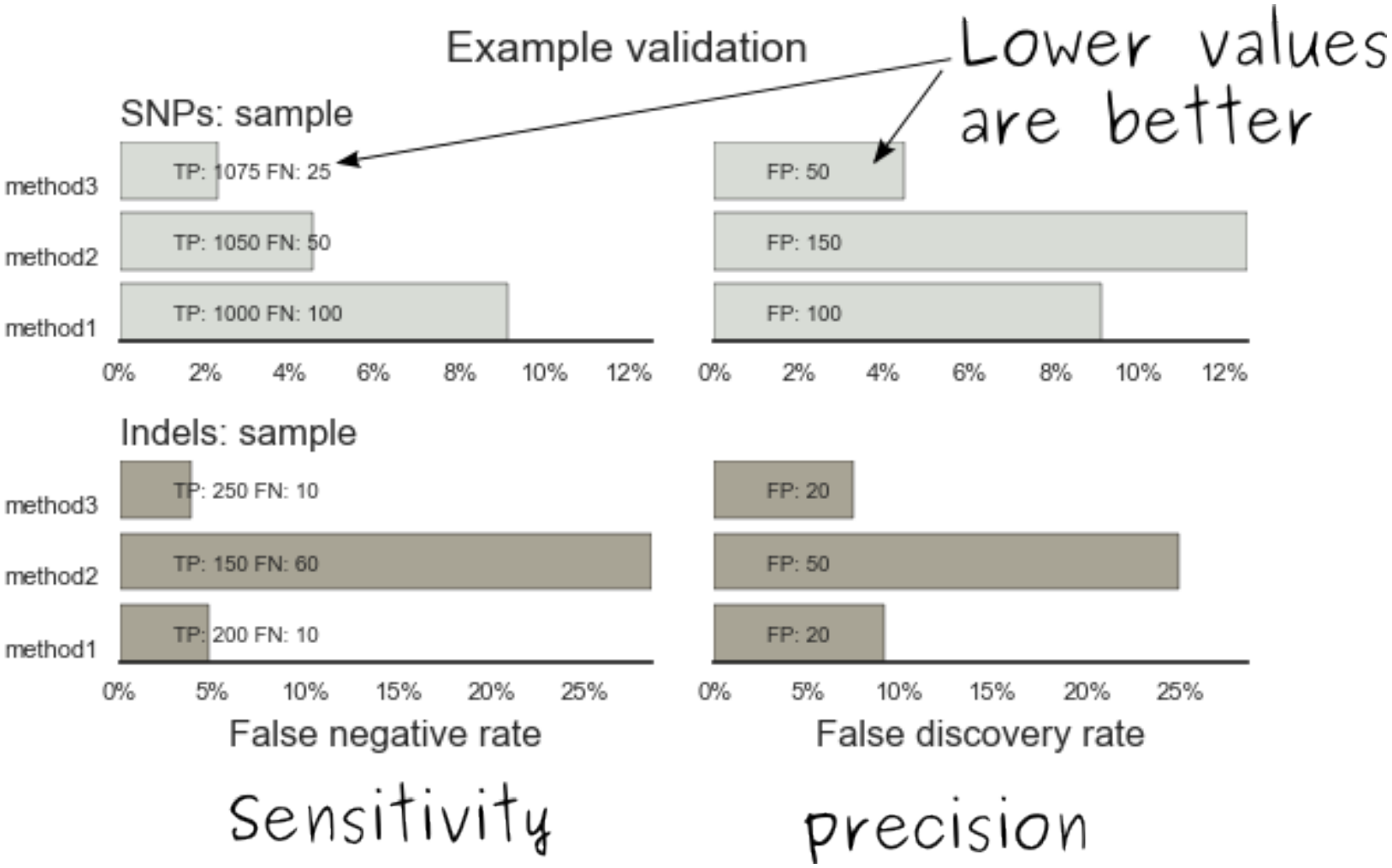
ICGC-TCGA DREAM Mutation Calling challenge

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

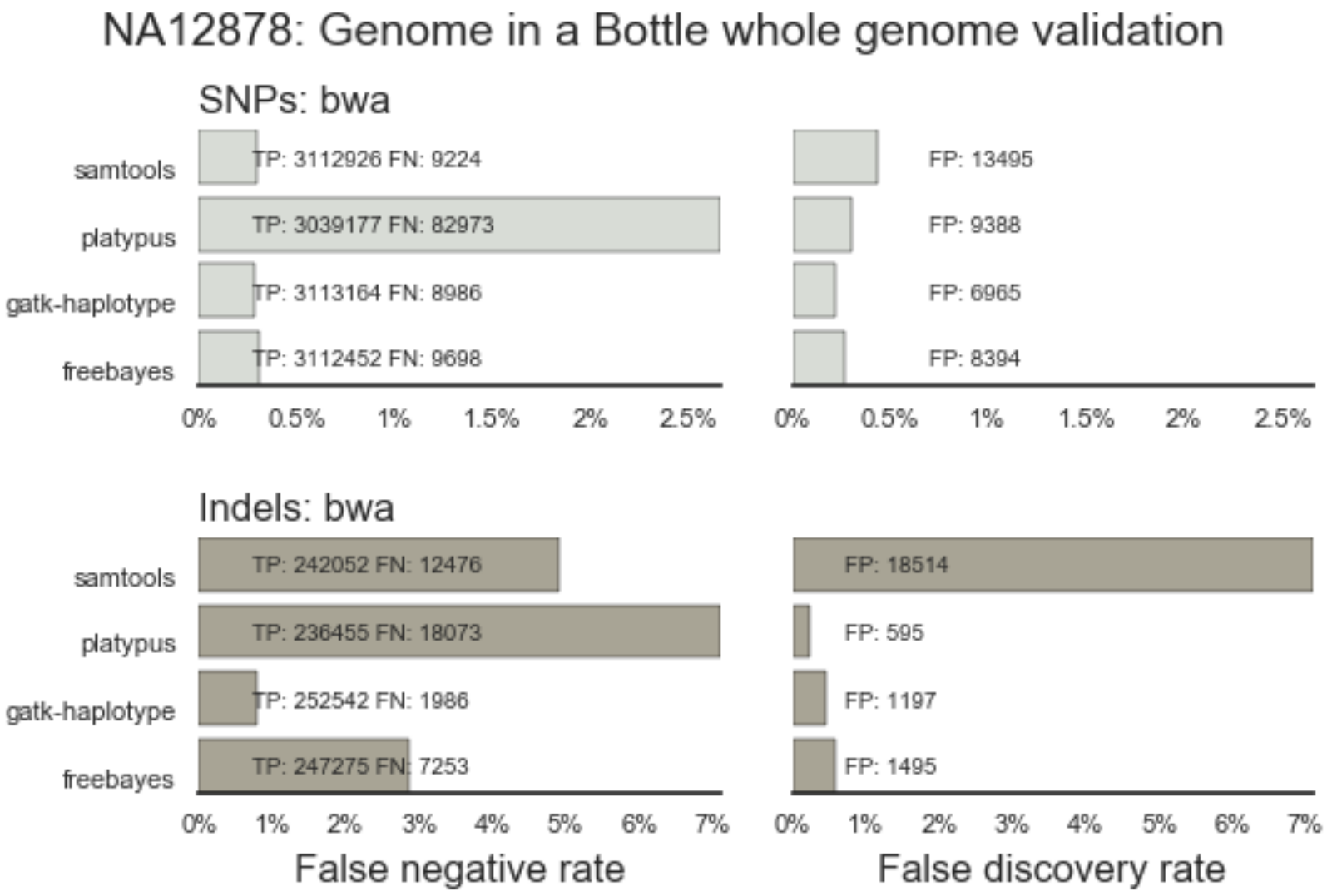
GATK and Genome in a Bottle

- Collaboration with GATK methods development
- Compare HaplotypeCaller to other methods
- Germline validation
- Genome in a Bottle reference materials
  - NA12878 – Caucasian
  - NA24385 – Ashkenazim Jewish
  - NA24631 – Chinese

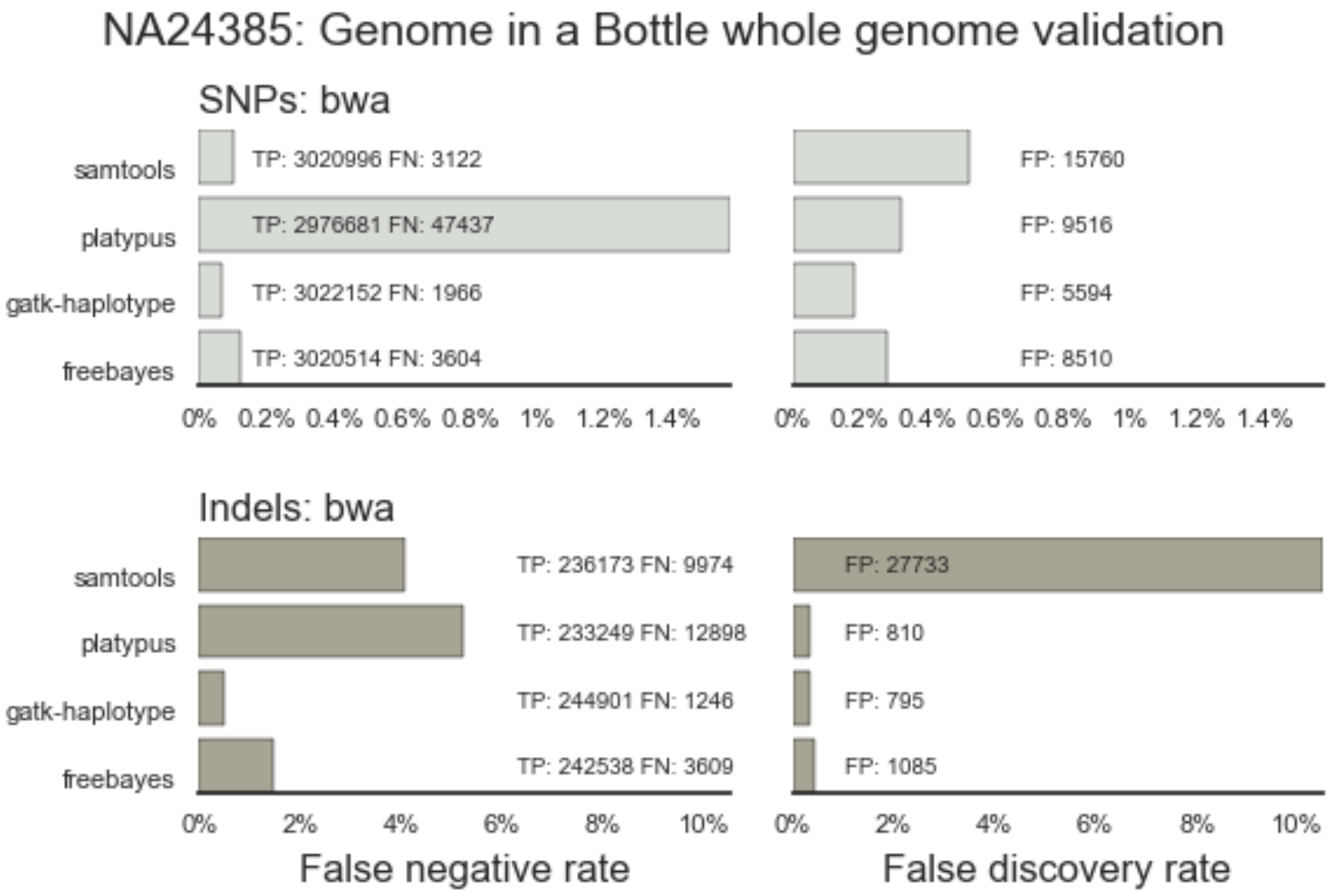
Validation graphs



NA12878


NA12878: Genome in a Bottle whole genome validation
SNPs: bwa
samtools
platypus
gatk-haplotype
freebayes
TP: 3112926 FN: 9224
TP: 3039177 FN: 82973
TP: 3113164 FN: 8986
TP: 3112452 FN: 9698
FP: 13495
FP: 9388
FP: 6965
FP: 8394
0%
0.5%
1%
1.5%
2%
2.5%
Indels: bwa
TP: 242052 FN: 12476
TP: 236455 FN: 18073
TP: 252542 FN: 1986
TP: 247275 FN: 7253
FP: 18514
FP: 595
FP: 1197
FP: 1495
0%
1%
2%
3%
4%
5%
6%
7%
False negative rate
False discovery rate

NA24385


NA24385: Genome in a Bottle whole genome validation
SNPs: bwa
samtools
platypus
gatk-haplotype
freebayes
TP: 3020996 FN: 3122
TP: 2976681 FN: 47437
TP: 3022152 FN: 1966
TP: 3020514 FN: 3604
FP: 15760
FP: 9516
FP: 5594
FP: 8510
0% 0.2% 0.4% 0.6% 0.8% 1% 1.2% 1.4%
0% 0.2% 0.4% 0.6% 0.8% 1% 1.2% 1.4%
Indels: bwa
samtools
platypus
gatk-haplotype
freebayes
TP: 236173 FN: 9974
TP: 233249 FN: 12898
TP: 244901 FN: 1246
TP: 242538 FN: 3609
FP: 27733
FP: 810
FP: 795
FP: 1085
0% 2% 4% 6% 8% 10%
0% 2% 4% 6% 8% 10%
False negative rate
False discovery rate

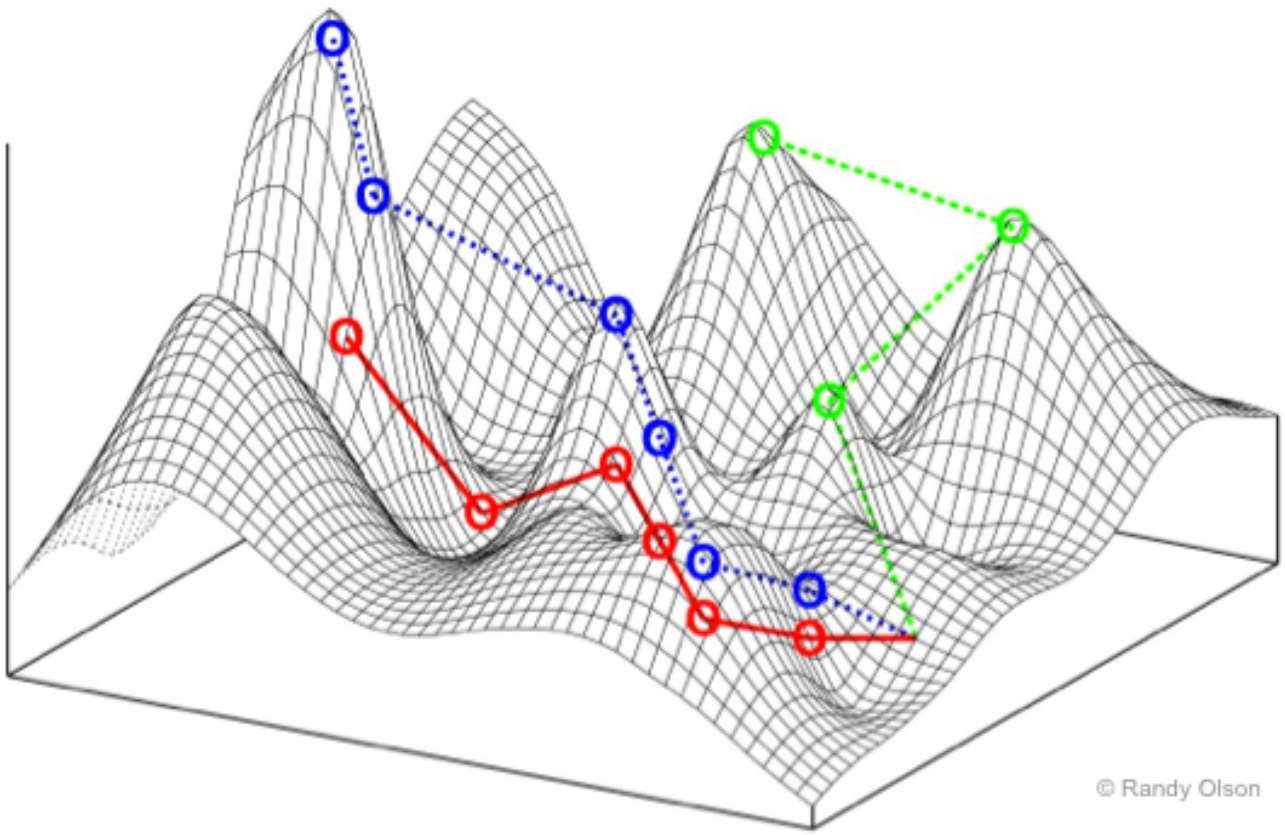
Validation results

- Good performance for GATK HaplotypeCaller
- Other good performing callers: FreeBayes
- Consistency across diverse samples
- Identify potential problem areas for tuning
  - samtools Indel false positive rates
  - Platypus SNP sensitivity
- PrecisionFDA: https://precision.fda.gov/

Infrastructure Goals

- Free, open source, community developed
- Welcoming to contributions
- Local machines
- Clusters: SLURM, SGE, Torque, PBS, LSF
- Clouds: Amazon, Google, Azure
- Clinical environments
- User interface for researchers
- Integrate with LIMS
- Accessible to the general public

Challenge: open source communities not yet optimal



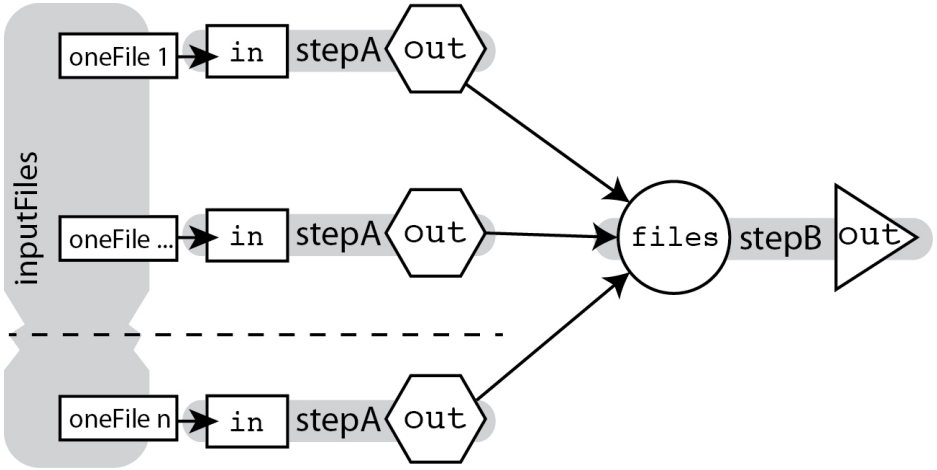
https://en.wikipedia.org/wiki/Fitness_landscape

Better abstractions = more interoperability

https://bcbio-nextgen.readthedocs.io/en/latest/contents/cwl.html

Workflow Description Language (WDL)

```
workflow myWorkflowName {
  call task_A
  call task_B
}
task task_A { ... }
task task_B { ... }
```



https://software.broadinstitute.org/wdl/

Common Workflow Language (CWL)

Workflow	pipeline-se-narrow.cwl		
Sub-workflow 1	01-qc-se.cwl		
Step 1	extract.cwl	extract.py	
Step 2	count.cwl	count.py	
Step 3	fastqc.cwl	fastqc	
Sub-workflow 2	02-trim.cwl		
...			

http://www.commonwl.org/

https://f1000research.com/slides/5-1617

Abstraction > Implementation

$$WDL \leftrightarrow CWL$$

CWL in bcbio

- Start with high level configuration file
- Generate CWL
- Run CWL:
  - Any infrastructure that supports CWL
  - Generated CWL
  - Docker or local bcbio installation
  - Genome data

https://bcbio-nextgen.readthedocs.io/en/latest/contents/cwl.html

Why use a workflow abstraction?

- Integrate with multiple platforms
  - Arvados
  - Toil
  - Cromwell
  - Galaxy
  - Nextflow
  - Seven Bridges
  - DNAnexus
- Stop maintaining bcbio specific infrastructure
- Focus on hard biological problems

Arvados

Welcome to the Arvados Project

The Arvados community is dedicated to building a new generation of open source distributed computing software for bioinformatics, data science, and production analysis using massive data sets.

https://arvados.org/
https://cloud.curoverse.com/

Toil

- UCSC NIH Big Data to Knowledge Center for Translational Genomics
- Supports CWL via conversion to internal Python workflow description
- Local HPC support: SLURM, SGE
- Cloud: AWS + spot instances

http://toil.readthedocs.io/en/latest/

Cromwell via conversion to WDL

- bcbio workflow abstractions supported in WDL
  - Tasks, workflows, nested workflows
  - Scatter based parallelization
  - Grouping/batching of samples
- Work in progress CWL to WDL converter based on cwl2wdl
- Happy to collaborate

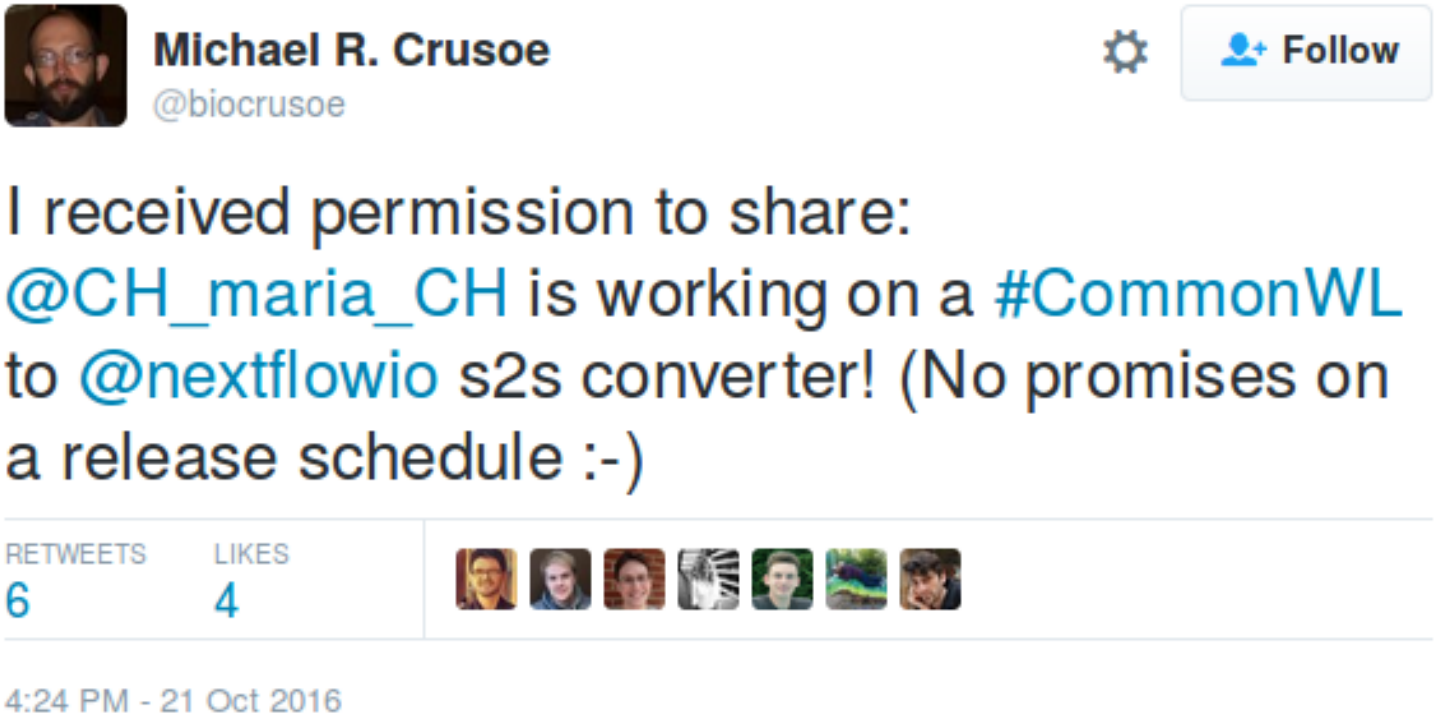
https://github.com/broadinstitute/cromwell https://github.com/chapmanb/bcbio-nextgen/blob/master/scripts/utils/cwltool2wdl.py

Galaxy

- CWL support in progress
- Supports subset of CWL – tool definitions
- Needs workflow support

https://github.com/galaxyproject/planemo
https://f1000research.com/posters/5-2567

Nextflow



https://www.nextflow.io/

Seven Bridges

- Currently supports CWL v2 + extensions
- Moving to CWL v1.0
- External runner, Bunny, supports v1.0

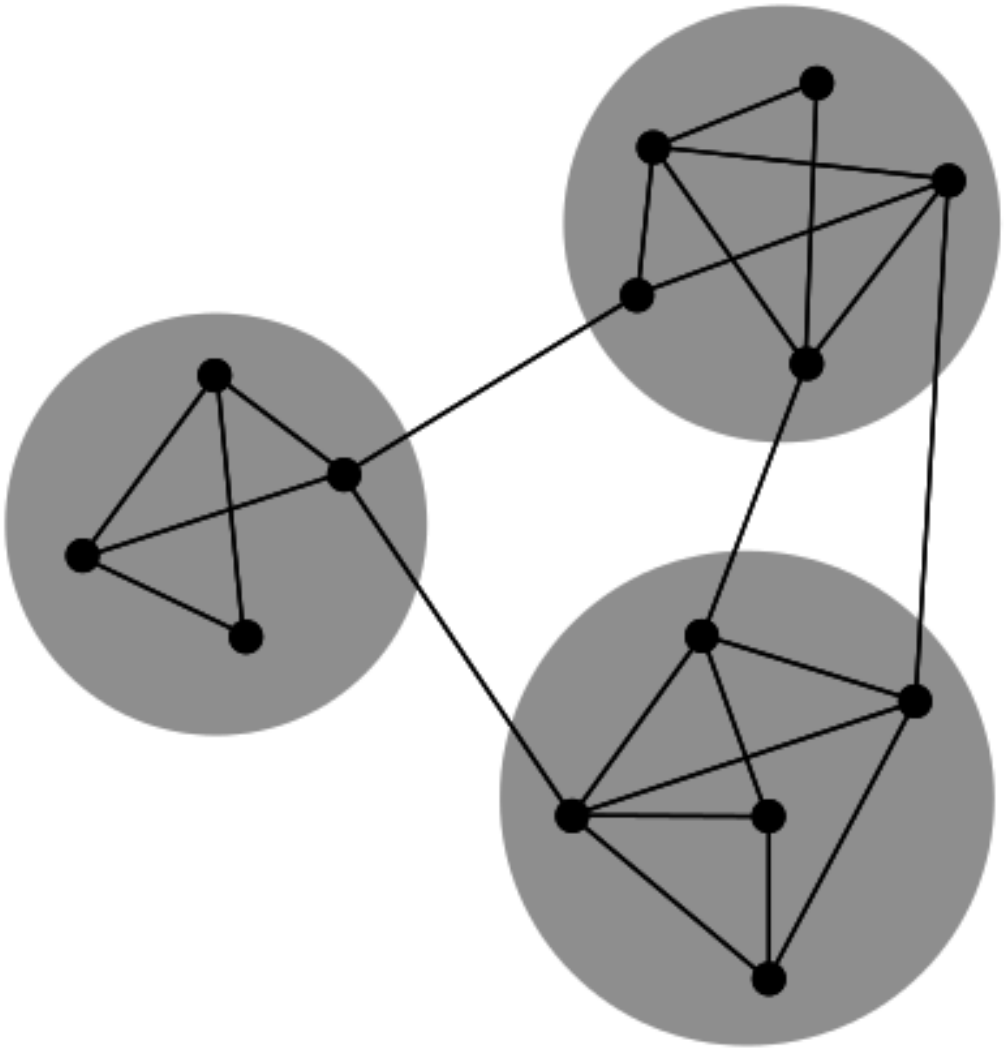
https://github.com/rabix/bunny
http://docs.sevenbridges.com/docs/sdk-overview

DNAnexus

- Integration of bcbio Docker containers
- Single node runs
- Convert CWL to DNAnexus API for distributed

https://www.dnanexus.com/developer-resources

Connections



By jham3 - Own work, CC BY-SA 3.0,

https://commons.wikimedia.org/w/index.php?curid=17125894

Summary

- bcbio community developed resources
- Value of validation
  - Germline calling with Genome in a Bottle
- Interoperable infrastructure
  - Importance of abstractions
  - Common Workflow Language
  - Practical interoperability

http://bcb.io