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

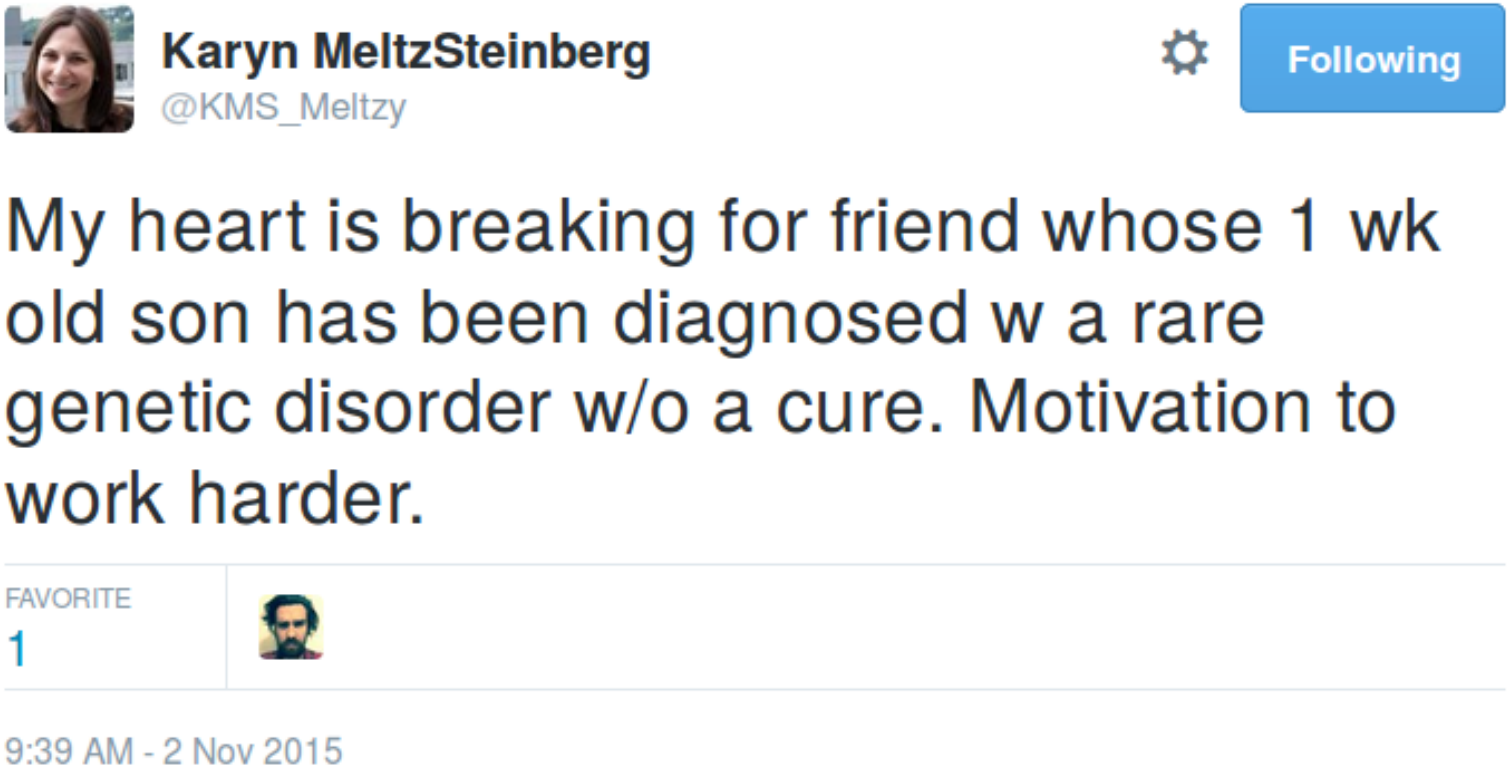
Brad Chapman
Bioinformatics Core, Harvard Chan School
https://bcb.io
http://j.mp/bcbiolinks

28 September 2016

Overview

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

We need to do science faster



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

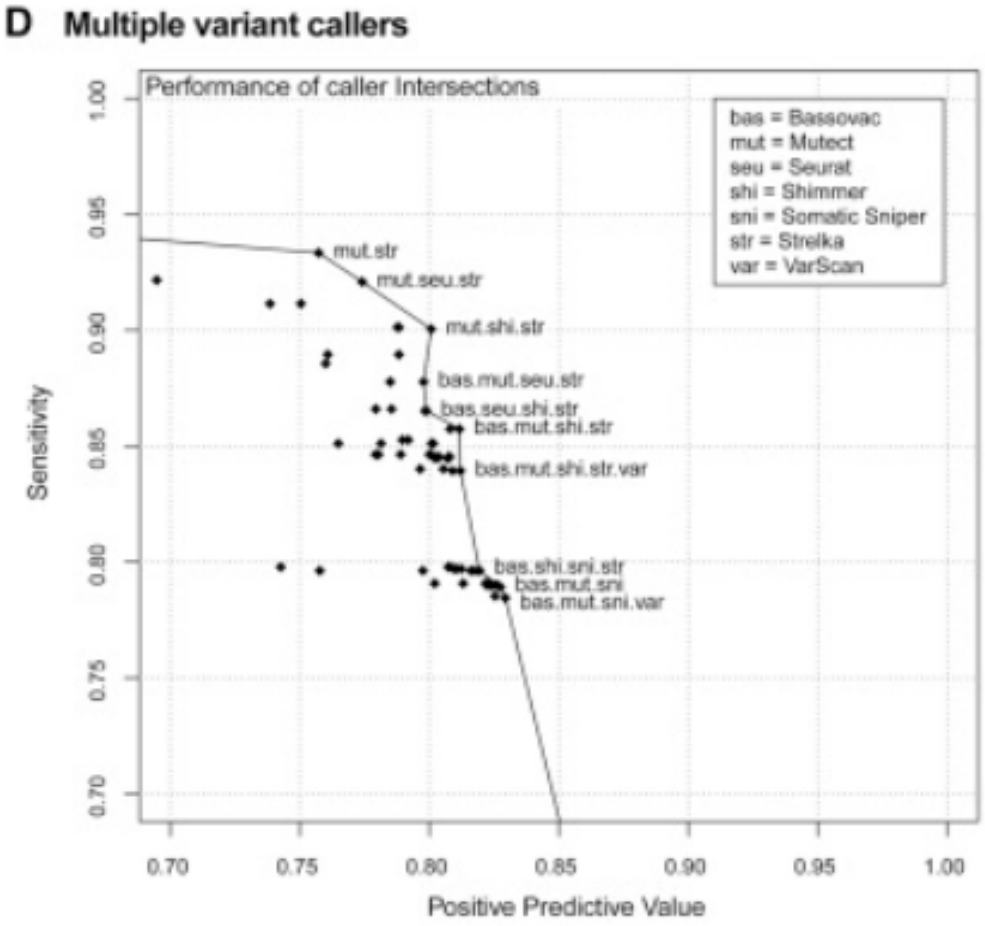
- Install tools
- Put tools together
- Test and validate
- Scale
- Improve algorithms
- Read literature
- Do biology

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 = better results



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

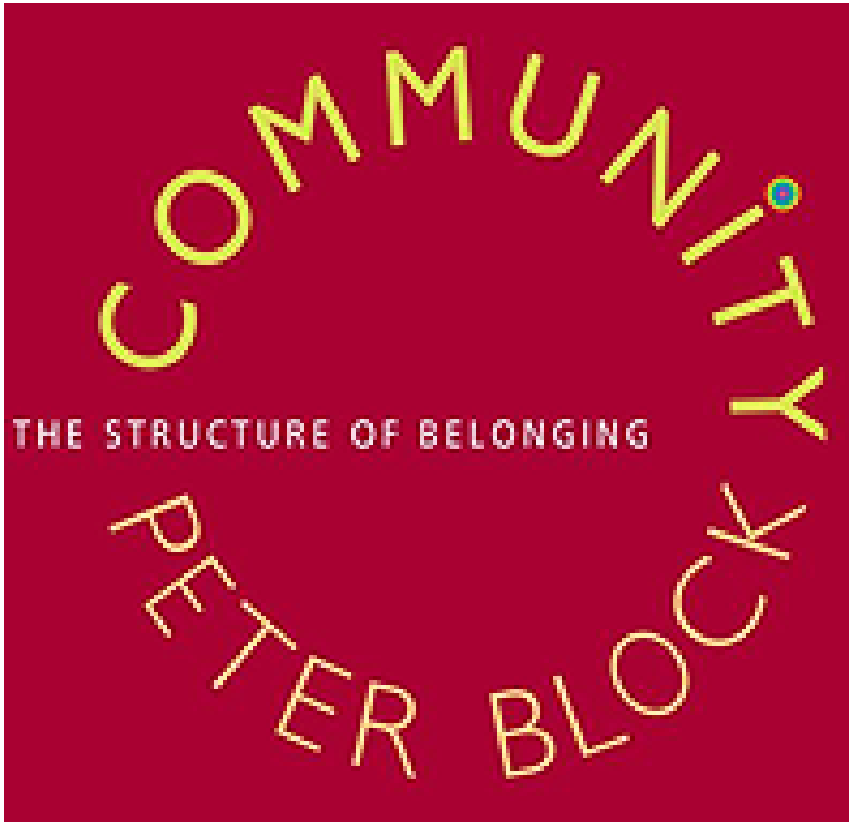
Many thanks to the Genomics Platform both for generating much of the exome data displayed here and for providing the computing resources required for this analysis.

Funding

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

http://exac.broadinstitute.org/about

Solution



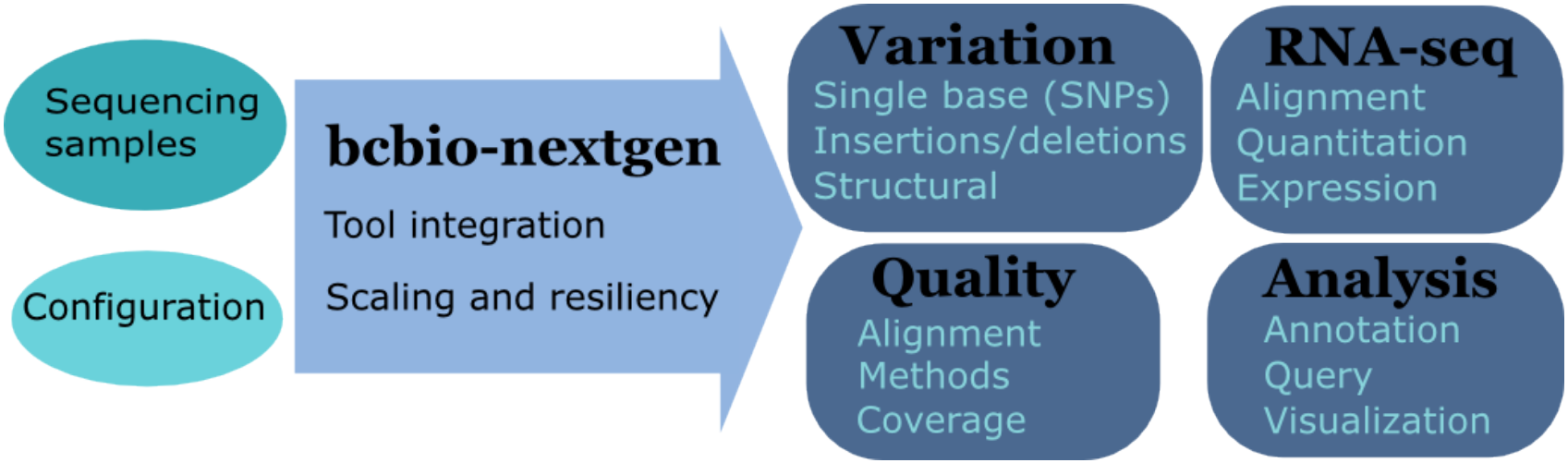
http://www.amazon.com/Community-Structure-Belonging-Peter-Block/dp/1605092770

Large scale infrastructure development

- 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

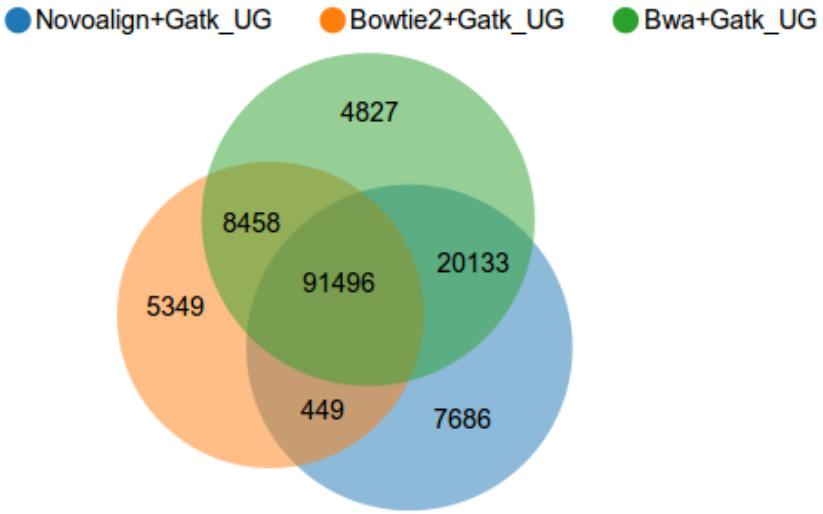
Quality differences between methods

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

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.

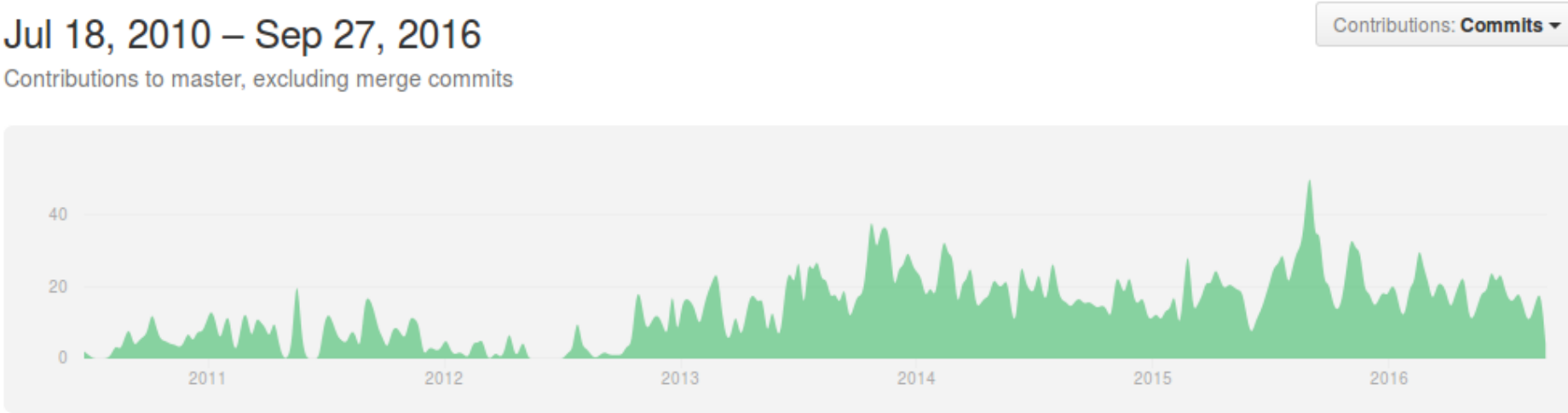
RETIRED: Best Practice Variant Detection with the GATK v3

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

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

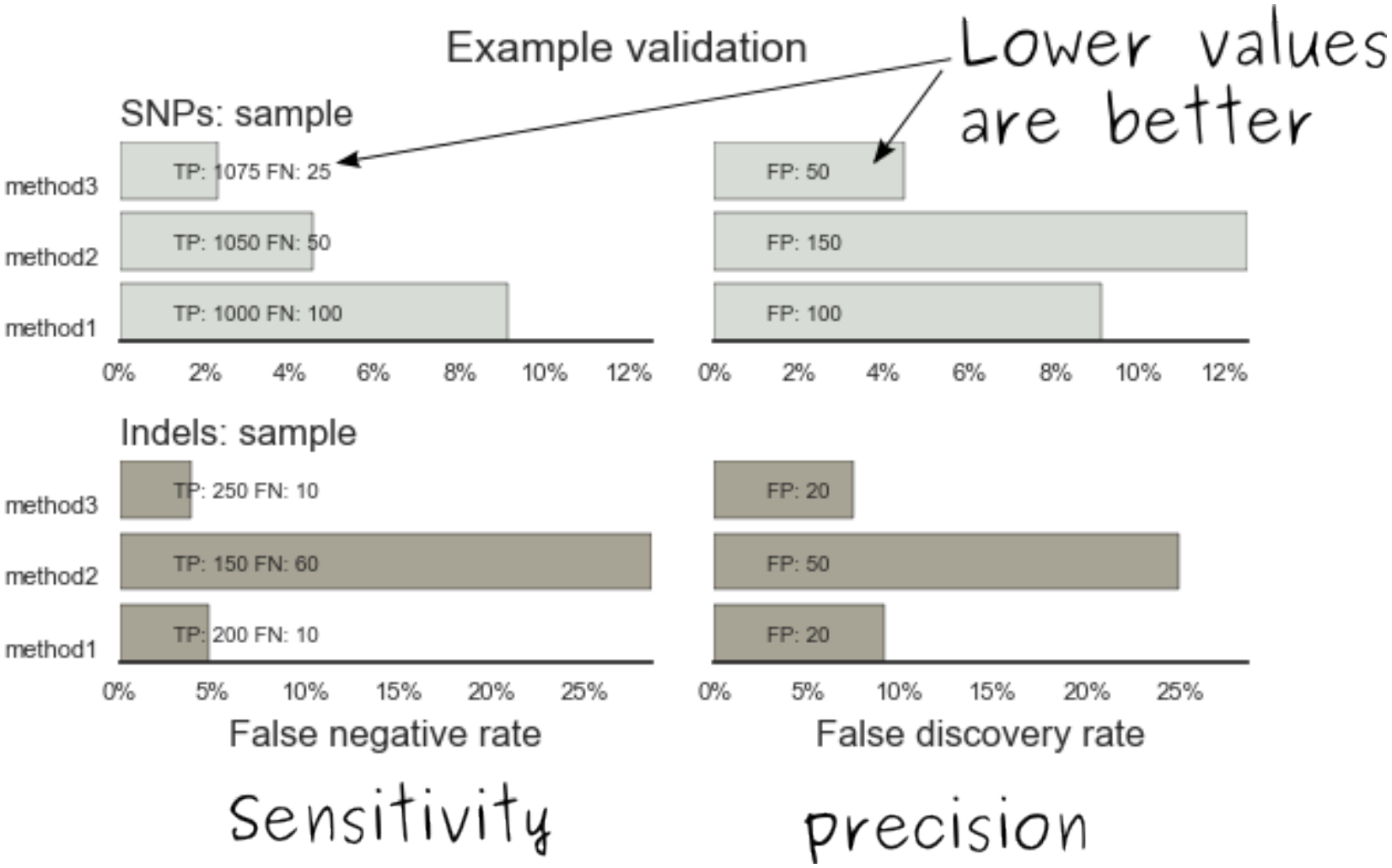
ICGC-TCGA DREAM Mutation Calling challenge

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

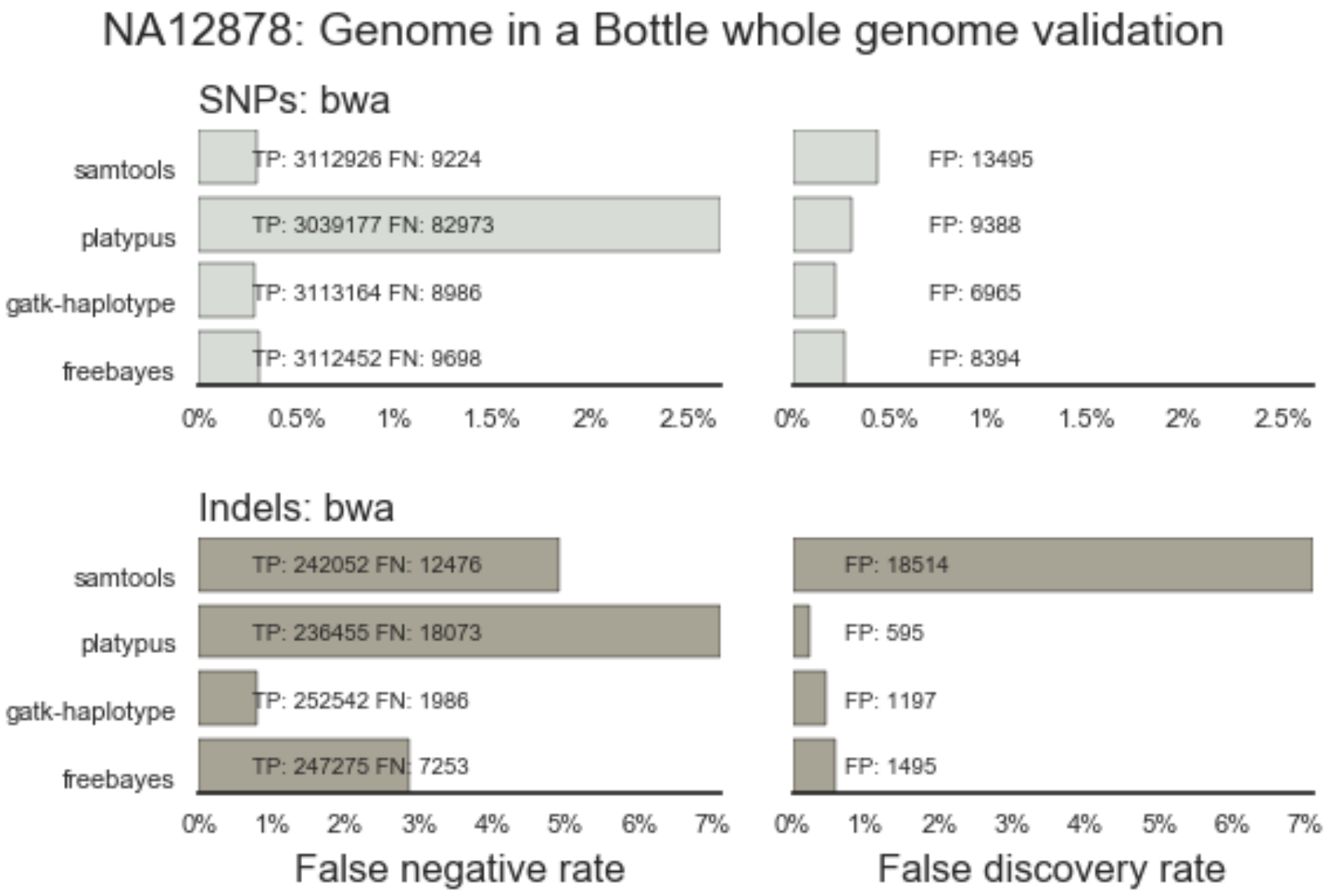
GATK and Genome in a Bottle

- 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

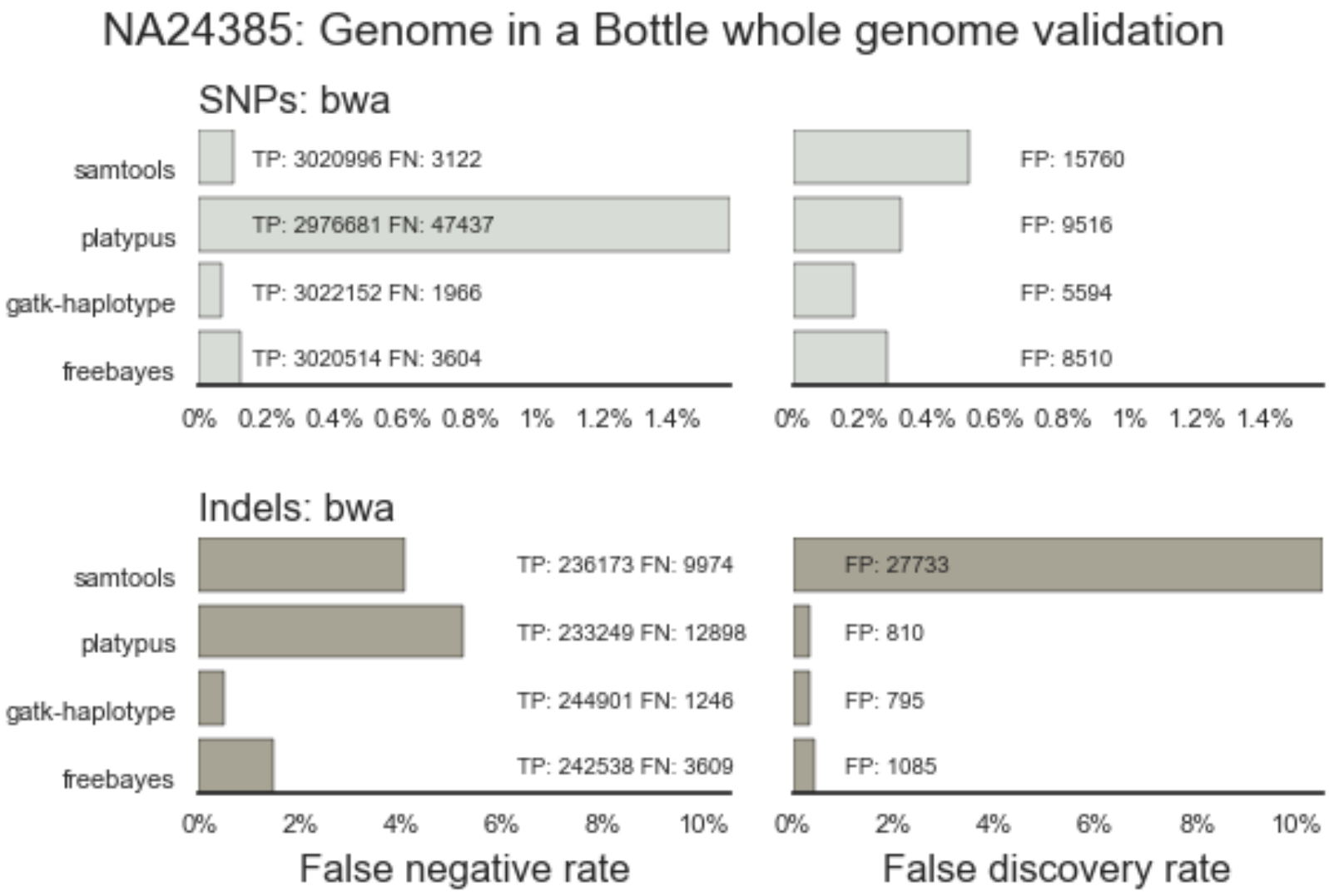
Validation graphs



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

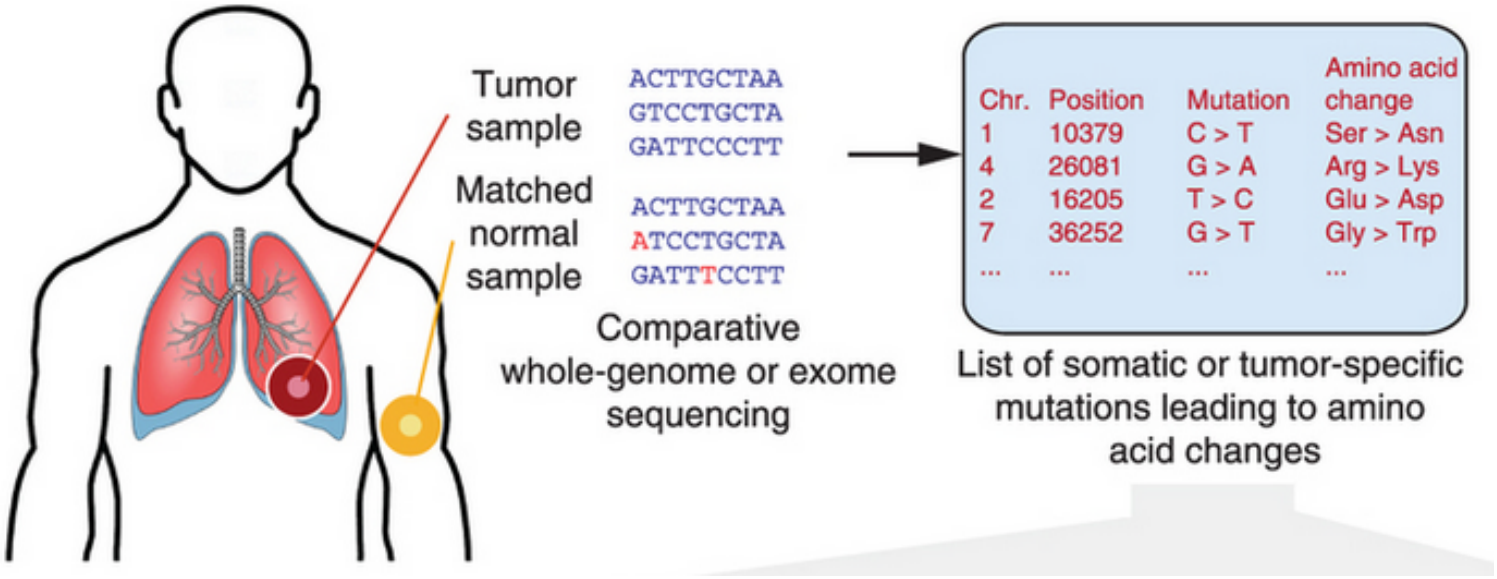
NA24385


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

Conclusions

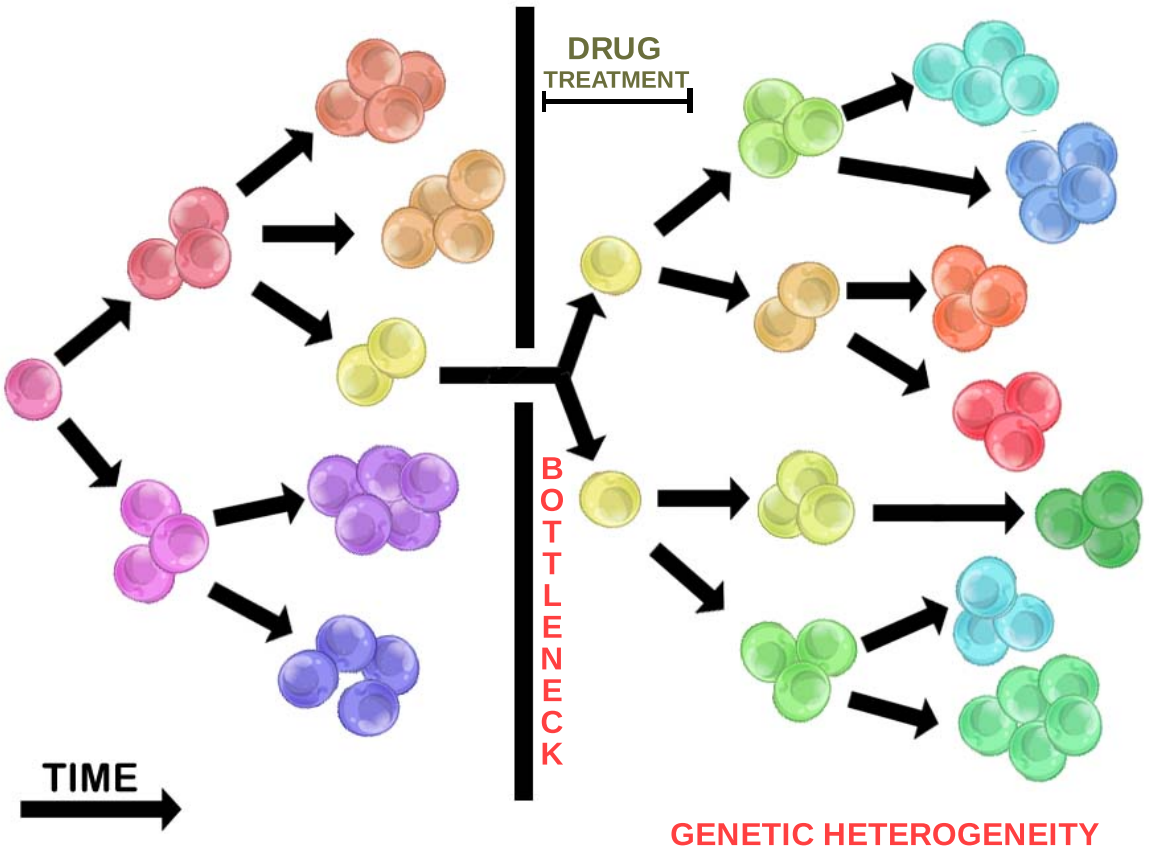
- Good performance for GATK HaplotypeCaller
- Other good performing callers like FreeBayes
- Consistency across diverse samples
- Identify potential problem areas for tuning
  - samtools Indel false positive rates
  - Platypus SNP sensitivity
- PrecisionFDA: https://precision.fda.gov/

Cancer somatic calling



http://www.nature.com/nmeth/journal/v10/n8/fig_tab/nmeth.2562_F1.html

Cancer heterogeneity



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

VarDict

- AstraZeneca
- Germline + Cancer calling
- SNP + Insertion/Deletions
- Whole genome + exome
- Also works on deep targeted data

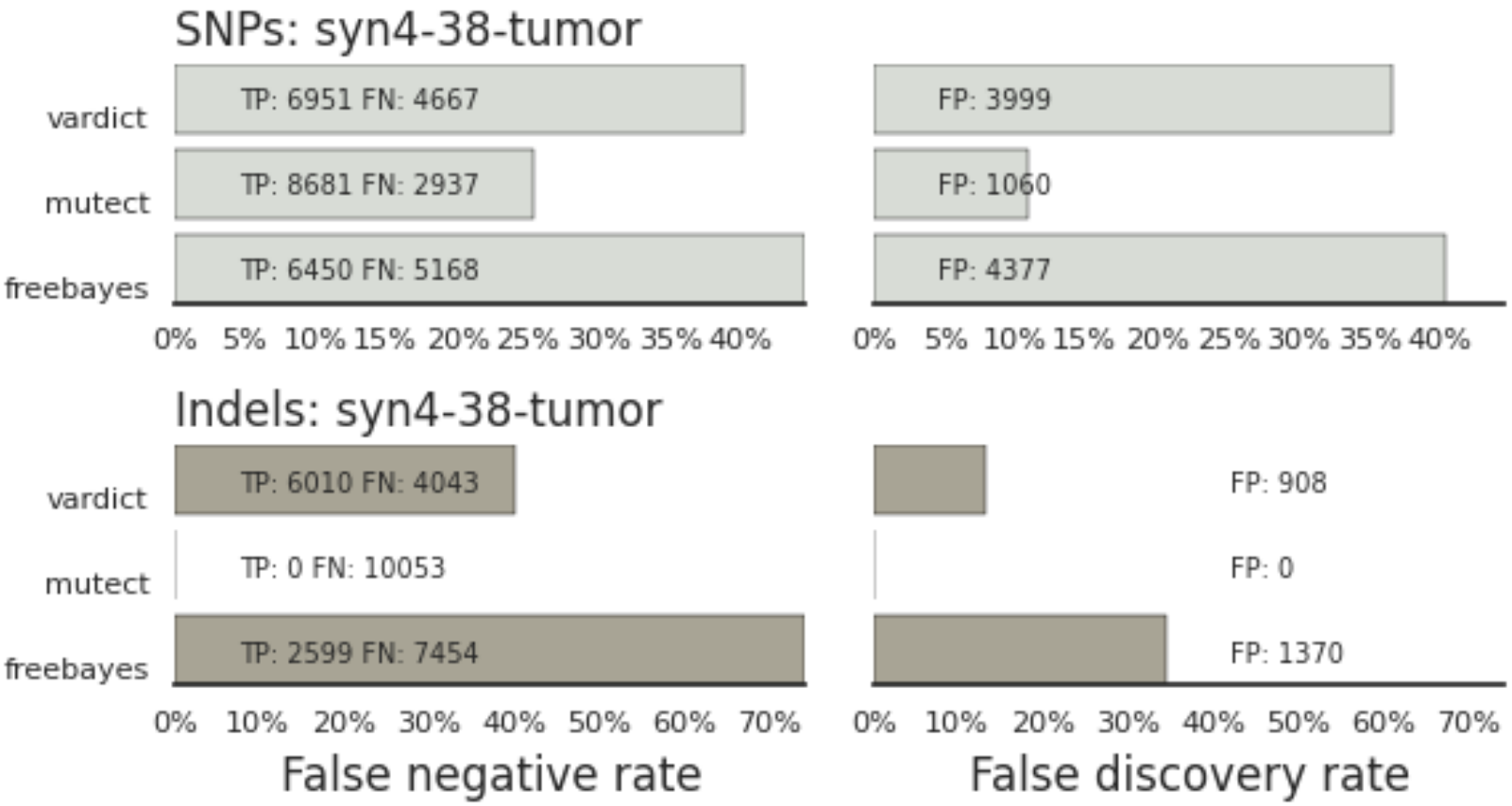
https://github.com/AstraZeneca-NGS/VarDictJava
http://nar.oxfordjournals.org/content/early/2016/04/07/nar.gkw227.full

DREAM synthetic dataset 4

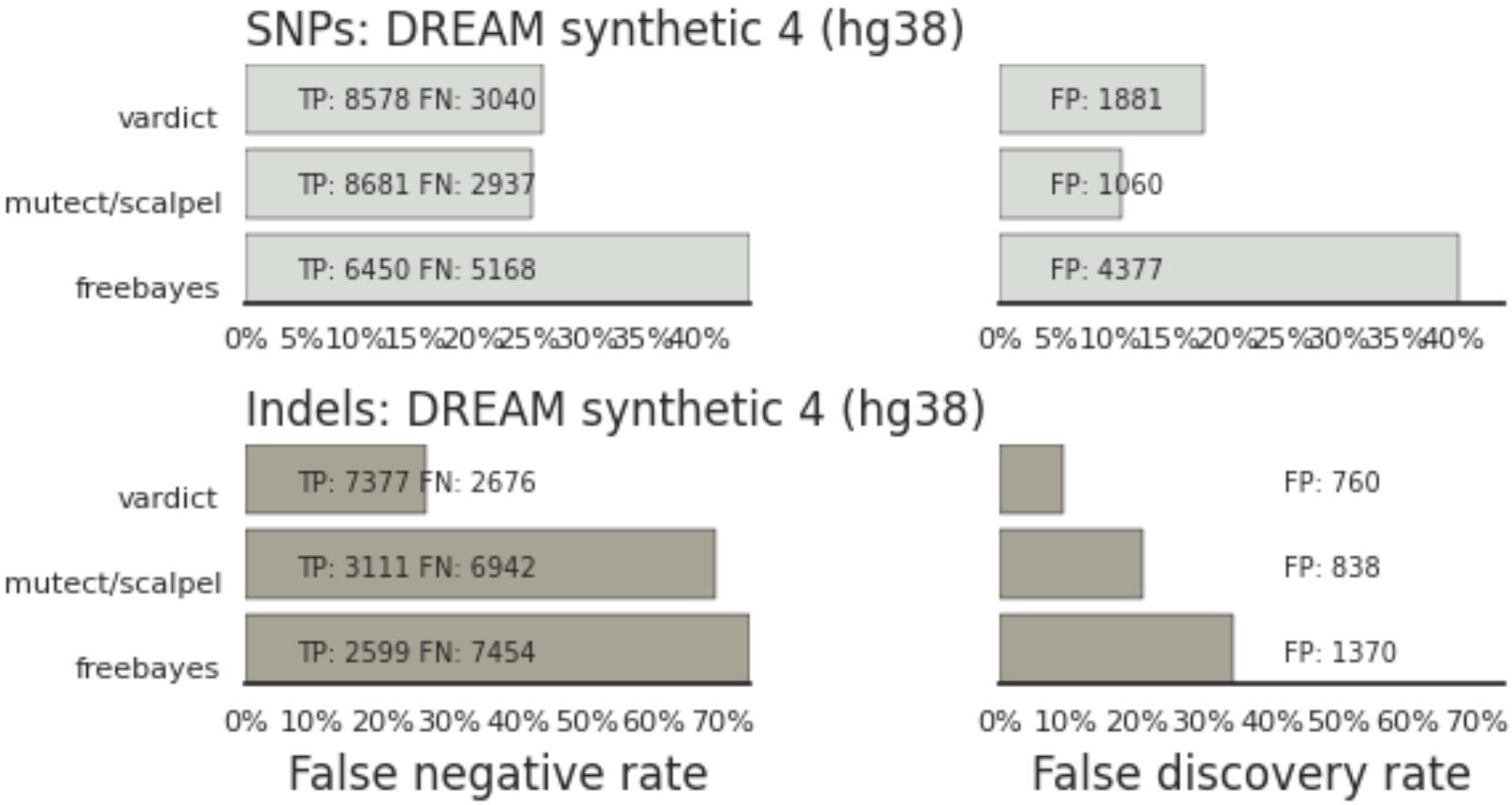
in silico 3	*in silico 4*
BWA Backtrack	BWA MEM
SNV, SV (deletions, duplications, insertions, inversions) & INDEL	SNV, SV (deletions, duplications, inversions) & INDEL
100%	80%
50%, 33%, 20%	50%, 35% (effectively 30% and 15% due to cellularity)
Female	Male
HCC1143 BL from TCGA Benchmark 4	CPCG0102R (Provided by ICGC)

https://www.synapse.org/#!Synapse:syn312572/wiki/62018

VarDict sensivitity/precision before



VarDict sensivitity/precision after

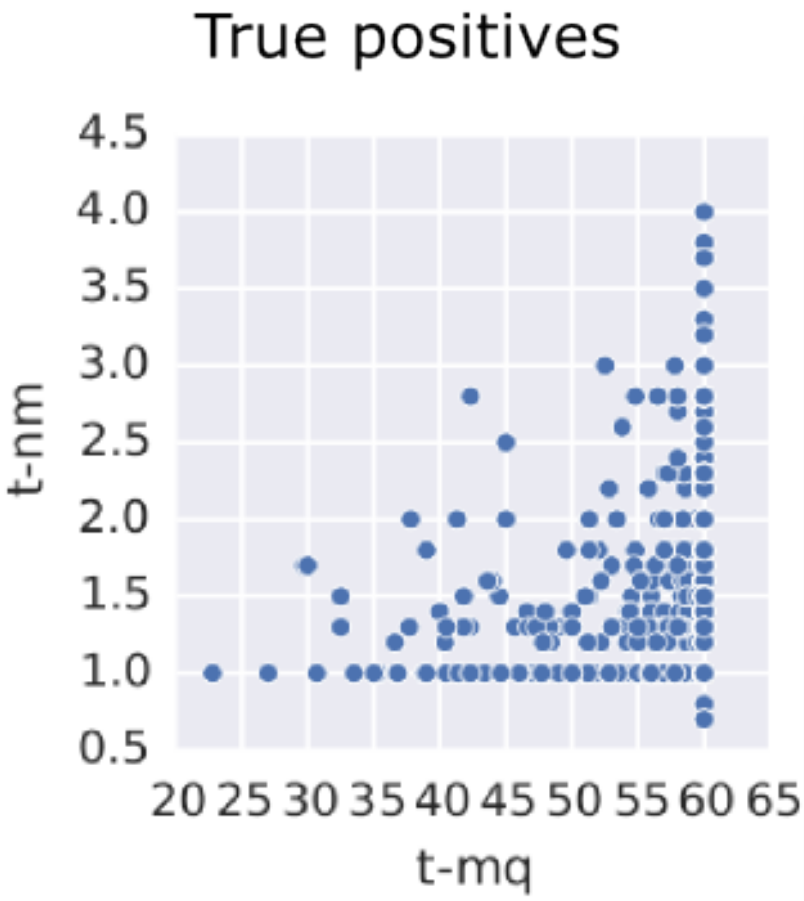


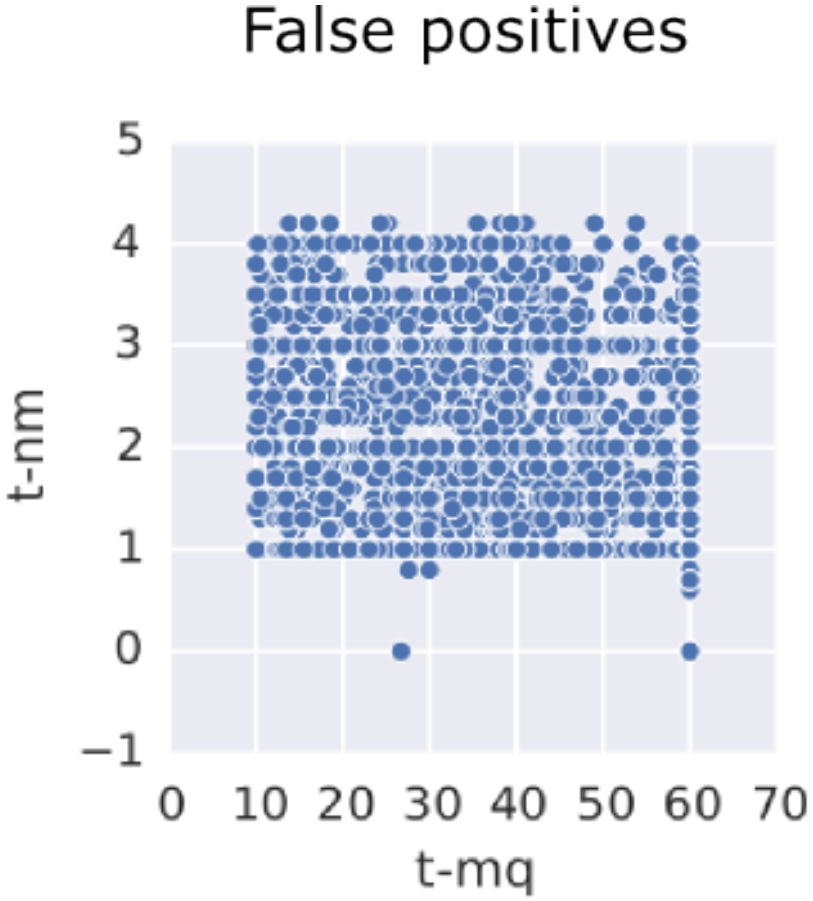
How? Filter summary

```
((AF * DP < 6) &&
 ((MQ < 55.0 && NM > 1.0) ||
  (MQ < 60.0 && NM > 2.0) ||
  (DP < 10) ||
  (QUAL < 45)))
```

http://bcb.io/2016/04/04/vardict-filtering/

Example filter: mapping quality and number of mismatches

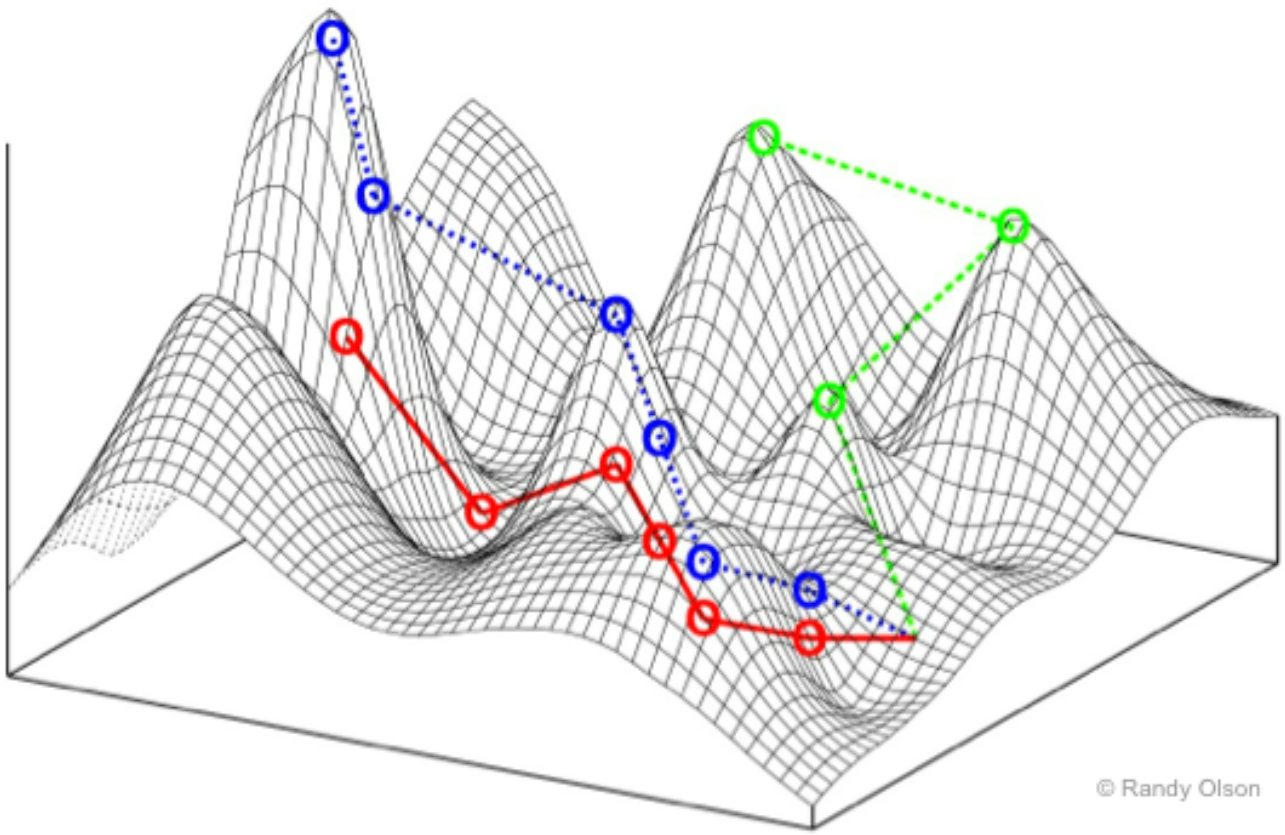




Infrastructure Goals

- 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

Open source communities not yet optimal



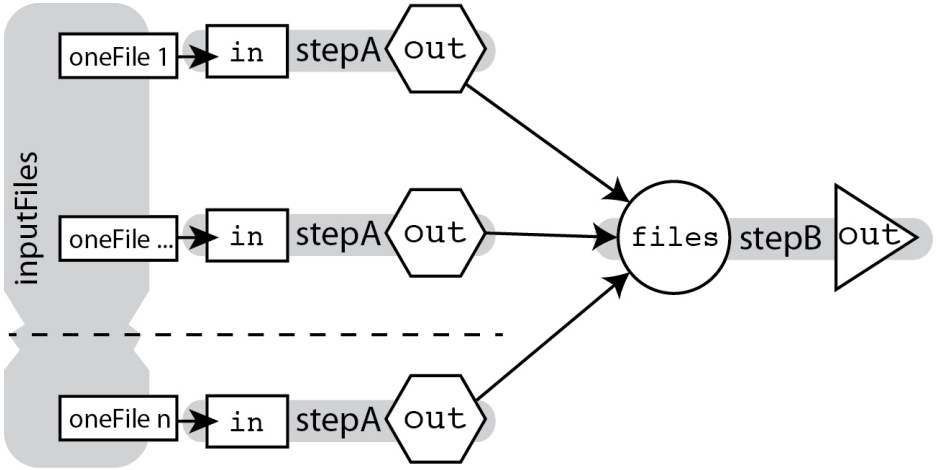
https://en.wikipedia.org/wiki/Fitness_landscape

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 directly
- 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 https://arvados.org/
  - Toil http://toil.readthedocs.io/en/latest/
  - Seven Bridges https://www.sbgenomics.com/
- Stop maintaining bcbio specific infrastructure
- Focus on hard biological problems

Producing 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/chapmanb/bcbio-nextgen/blob/master/scripts/utils/cwltool2wdl.py

Summary

- bcbio community developed resources
- Value of validation
  - GATK Genome in a Bottle Validation
  - Improve low frequency cancer calling
- Interoperable infrastructure
  - CWL and WDL integration

http://bcb.io