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

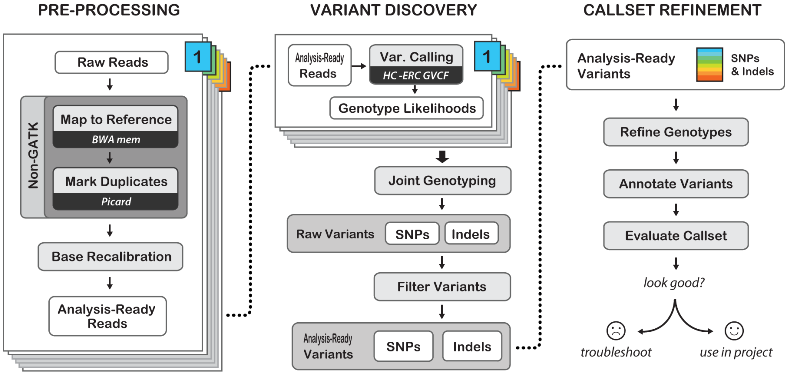
Brad Chapman
Bioinformatics Core, Harvard Chan School
https://bcb.io
http://j.mp/bcbiolinks

1 May 2017

Overview

- Development challenges and open source communities
- bcbio: community developed analyses
- Value of variant validation
- Interoperable infrastructure on the Cancer Genomics Cloud

I want to build a variant calling pipeline



Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

https://software.broadinstitute.org/gatk/best-practices/

Feature support burden

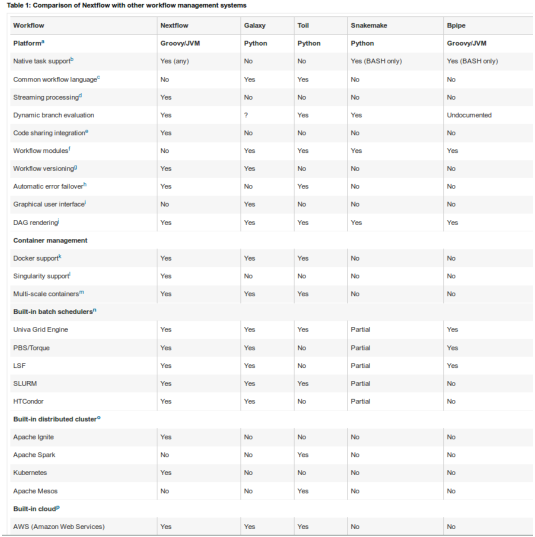
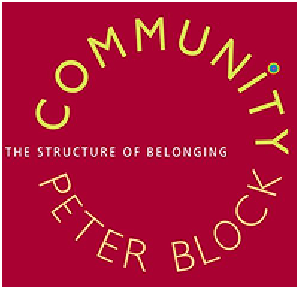


Table 1: Comparison of Nextflow with other workflow management systems

Workflow	Nextflow	Galaxy	Toil	Snakemake	Bpipe
Platform[a]	**Groovy/JVM**	**Python**	**Python**	**Python**	**Groovy/JVM**
Native task support[b]	Yes (any)	No	No	Yes (BASH only)	Yes (BASH only)
Common workflow language[c]	No	Yes	Yes	No	No
Streaming processing[d]	Yes	No	No	No	No
Dynamic branch evaluation	Yes	?	Yes	Yes	Undocumented
Code sharing integration[e]	Yes	No	No	No	No
Workflow modules[f]	No	Yes	Yes	Yes	Yes
Workflow versioning[g]	Yes	Yes	No	No	No
Automatic error failover[h]	Yes	No	Yes	No	No
Graphical user interface[i]	No	Yes	No	No	No
DAG rendering[j]	Yes	Yes	Yes	Yes	Yes
Container management					
Docker support[k]	Yes	Yes	Yes	No	No
Singularity support[l]	Yes	No	No	No	No
Multi-scale containers[m]	Yes	Yes	Yes	No	No
Built-in batch schedulers[n]					
Univa Grid Engine	Yes	Yes	Yes	Partial	Yes
PBS/Torque	Yes	Yes	No	Partial	Yes
LSF	Yes	Yes	No	Partial	Yes
SLURM	Yes	Yes	Yes	Partial	No
HTCondor	Yes	Yes	No	Partial	No
Built-in distributed cluster[o]					
Apache Ignite	Yes	No	No	No	No
Apache Spark	No	No	Yes	No	No
Kubernetes	Yes	No	No	No	No
Apache Mesos	No	No	Yes	No	No
Built-in cloud[p]					
AWS (Amazon Web Services)	Yes	Yes	Yes	No	No

http://www.nature.com/nbt/journal/v35/n4/full/nbt.3820.html

Build open source communities



COMMUNITY

THE STRUCTURE OF BELONGING

PETER BLOCK

BOSC 2017

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

http://www.open-bio.org/wiki/BOSC_2017

Challenge: many communities

CANCER GENOMICS CLOUD
SEVEN BRIDGES

cBioPortal
for Cancer Genomics

Synapse

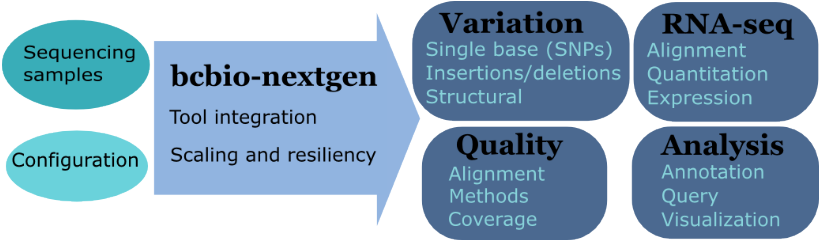
http://www.cancergenomicscloud.org/
http://www.cbioportal.org/
https://www.synapse.org/

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, cnvkit]
```

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.*

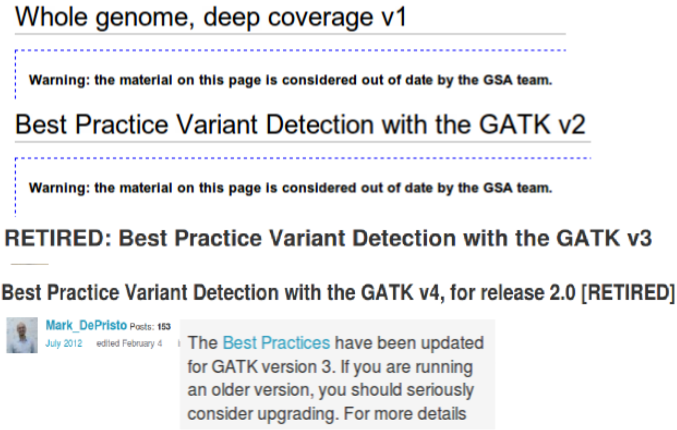
http://www.genomemedicine.com/content/7/1/68

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.

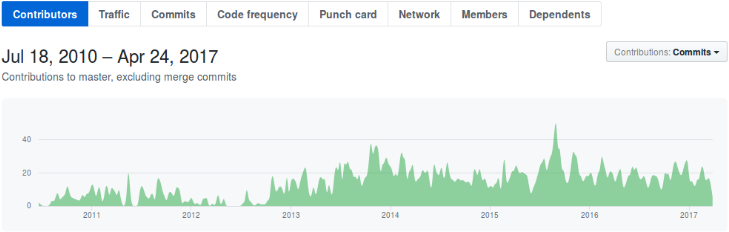
RETIRED: Best Practice Variant Detection with the GATK v3

Best Practice Variant Detection with the GATK v4, for release 2.0 [RETIRED]

Mark_DePristo Posts: 153
July 2012 edited February 4

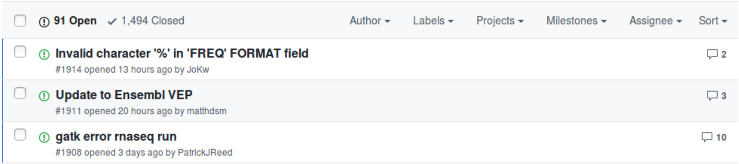
The Best Practices have been updated for GATK version 3. If you are running an older version, you should seriously consider upgrading. For more details

Community: sustainability and support



Jul 18, 2010 – Apr 24, 2017

Contributions to master, excluding merge commits



□ 91 Open ✓ 1,494 Closed

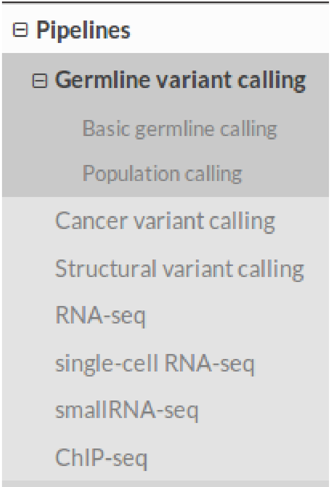
□ Invalid character '%' in 'FREQ' FORMAT field 2
#1914 opened 13 hours ago by JoKw

□ Update to Ensembl VEP 3
#1911 opened 20 hours ago by matthdsm

□ gatk error rnaseq run 10
#1908 opened 3 days ago by PatrickJReed

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

Supported analysis types



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

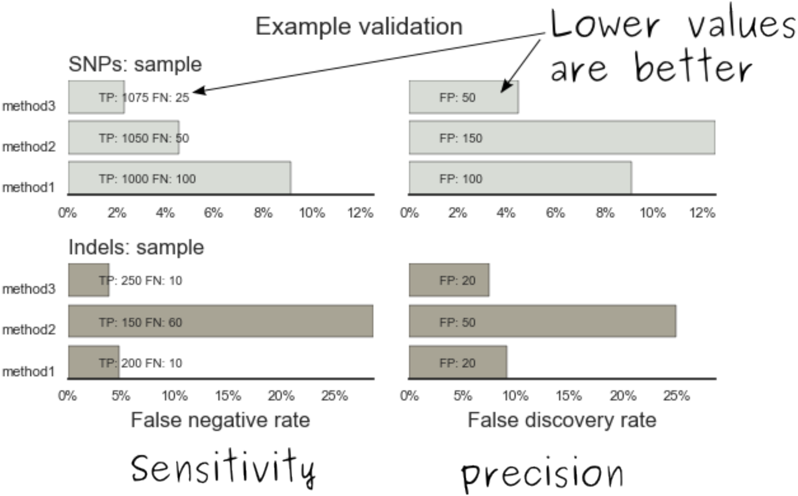
ICGC-TCGA DREAM Mutation Calling challenge

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

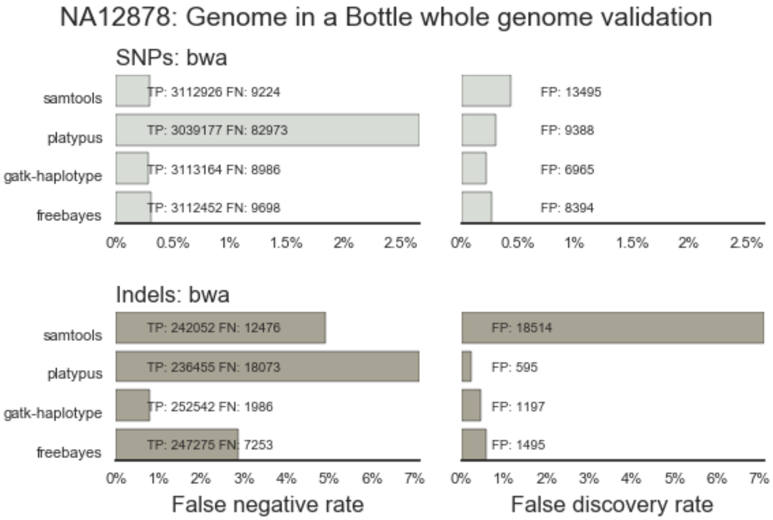
GATK and Genome in a Bottle

- 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

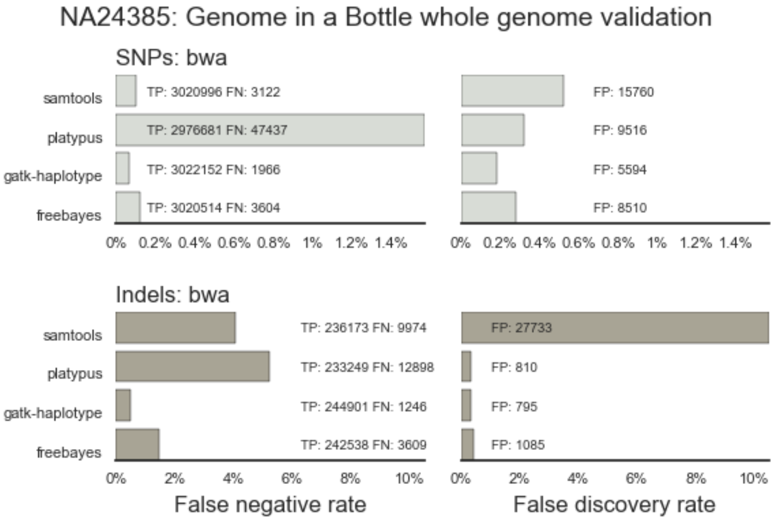
Validation graphs



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
samtools
platypus
gatk-haplotype
freebayes
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

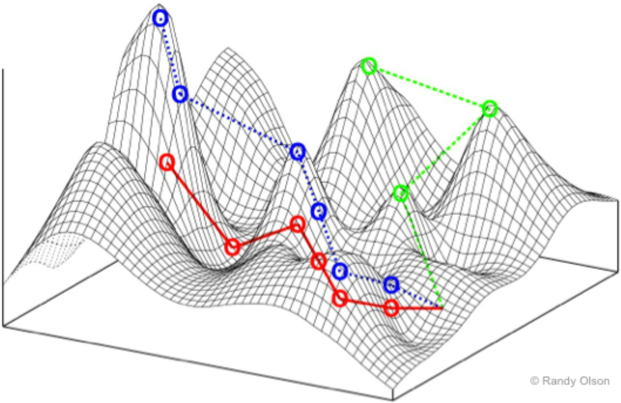
Validation results

- 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

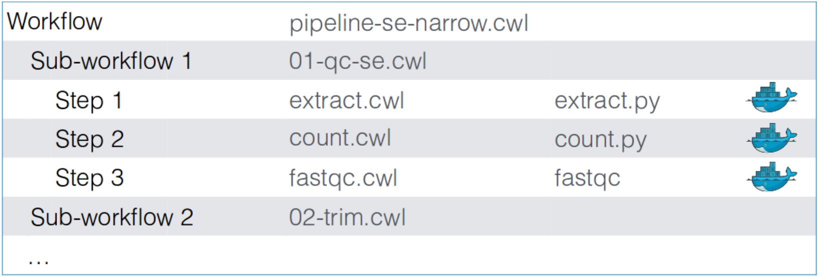
Better abstractions = more interoperability

COMMON WORKFLOW LANGUAGE

docker

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

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

CWL in bcbio

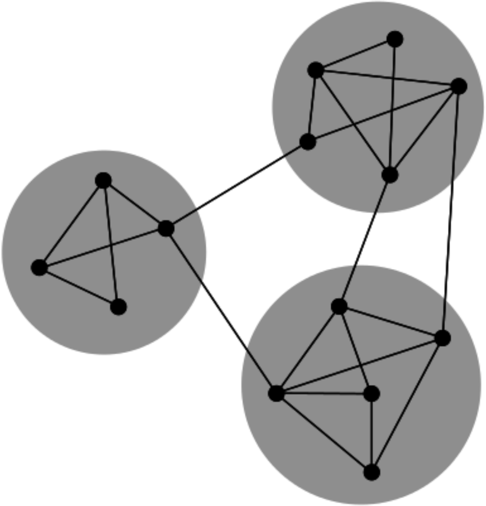
- Start with high level configuration file
- Generate CWL
- Run CWL:
  - Any infrastructure that supports CWL
  - Generated CWL
  - Docker or local bcbio installation
  - Genome data

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

Why use a workflow abstraction?

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

Connections



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

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

Practical example

- Build CWL using bcbio + CGC
- Use publicly available Simons Genome Diversity Project sample

Summary

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

`http://bcb.io`