

bcbio: validated, scalable, community developed
variant calling, RNA-seq and small RNA analysis

Brad Chapman
Bioinformatics Core, Harvard Chan School

<https://bcb.io>

<http://j.mp/bcbiolinks>

23 March 2016

- Goal: open source community developed sequencing analysis
- Overview of bcbio validated variant calling
- Science
 - Human build 38 + HLA
 - Cancer calling of low frequency variants
 - Structural variation
- Move to community workflow infrastructure

We need to do science faster



Karyn MeltzSteinberg

@KMS_Meltzy



Following

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



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

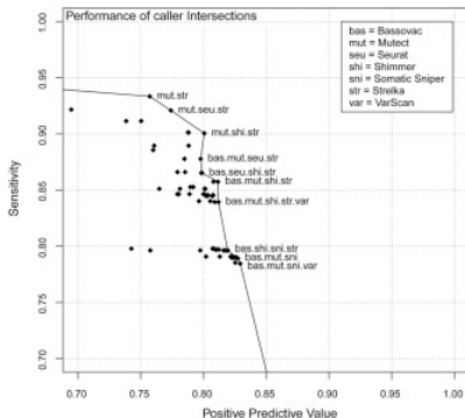
Standard analyses not routine

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

D Multiple variant callers



<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 Ardisino
- 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 Klezun
- 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

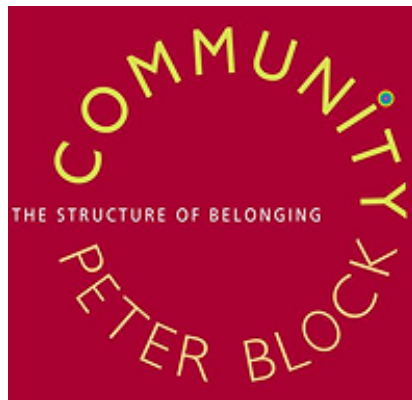
Many thanks to the Genomics Platform both for generating much of the exome data displayed here and for providing the computing resources required for this analysis.

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>

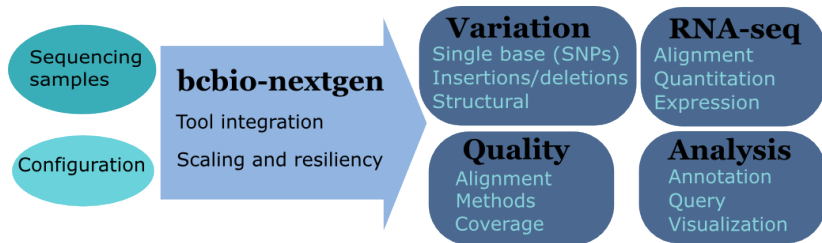
Large scale infrastructure development

- 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>

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

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

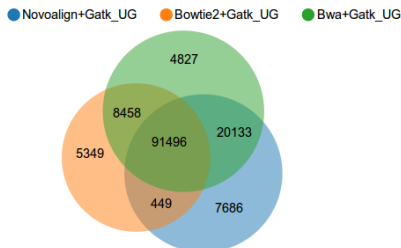
Quality differences between methods

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 concordance 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>

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

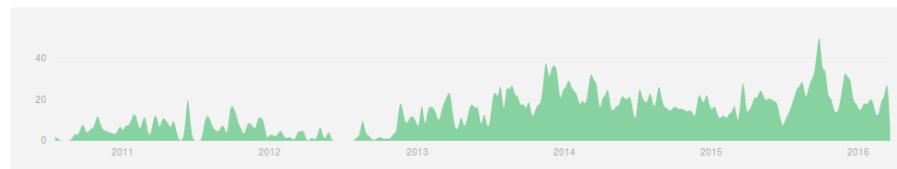
The [Best Practices](#) have been updated for GATK version 3. If you are running an older version, you should seriously consider upgrading. For more details

Community: sustainability

Jul 18, 2010 – Mar 22, 2016















Contributions to master, excluding merge commits

Contributions: **Commits** ▾



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

Community: support

<input type="checkbox"/>	 85 Open  954 Closed	Author ▾	Labels ▾	Milestones ▾	Assignee ▾	Sort ▾
<input type="checkbox"/>	 Looking for summary.pdf and error rendering R report #1282 opened 21 hours ago by pfujita					 1
<input type="checkbox"/>	 100.000 X covered region included for Freebayes variant calling? #1281 opened a day ago by NeillGibson					 2
<input type="checkbox"/>	 mutect config file missing error. #1280 opened 2 days ago by mortunco					 3
<input type="checkbox"/>	 bammerge stuck in qsub interactive session #1279 opened 4 days ago by razZ0r					 1
<input type="checkbox"/>	 (p)bgzip and multithreading/indexing #1278 opened 4 days ago by schelhorn					 2
<input type="checkbox"/>	 pbspro controller start fails #1276 opened 6 days ago by razZ0r					 32

<https://bcbio-nextgen.readthedocs.org>

Community: contribution

chapmanb / bcbio-nextgen

Unwatch 66 Unstar 282 Fork 154

Code Issues 85 Pull requests 4 Pulse Graphs Settings

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

4,552 commits 2 branches 32 releases 39 contributors

Branch: master New pull request

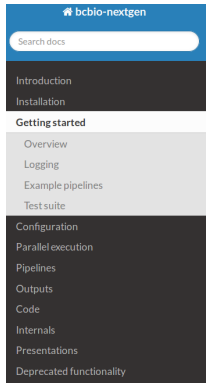
New file Upload files Find file HTTPS <https://github.com/chapmanb/bcbio-nextgen> Download ZIP

chapmanb Docker: new error logging when not CWL driven Latest commit b820ca7 25 minutes ago

artwork	add logo to README and docs	10 months ago
bcbio	Docker: new error logging when not CWL driven	25 minutes ago
config	Start of work to simplify bcbio_system file	15 hours ago
docs	Refine single-cell RNA-seq under correct header.	21 hours ago
scripts	Fix name for genbank conversion script.	2 days ago
tests	Add initial implementation of a single cell RNA-seq pipeline	6 days ago

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

Community: documentation



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

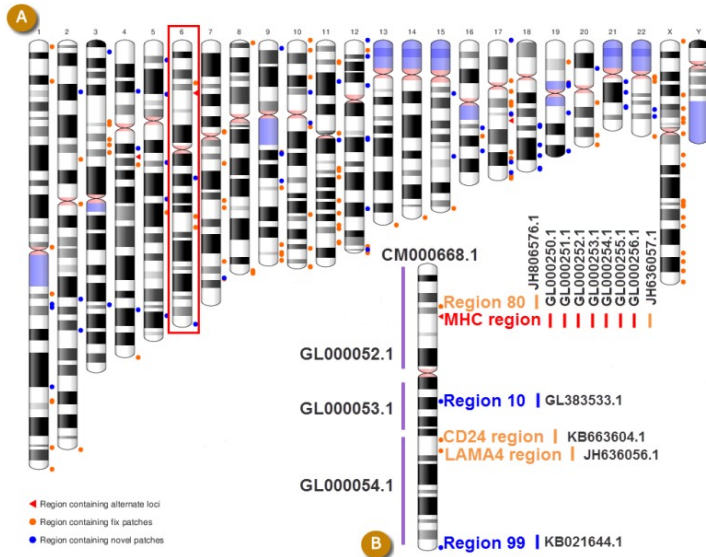
smallRNA-seq

ChIP-seq

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

- **Human build 38 + HLA**
- Low frequency somatic calling
- Structural variation

GRCh37/hg19

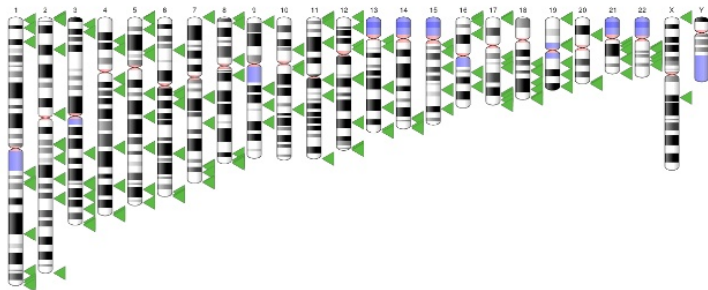


<http://www.ncbi.nlm.nih.gov/books/NBK153600/?report=reader>

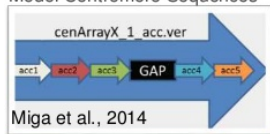
GRCh38 – graph based, many more alternative loci

Excitement about GRCh38

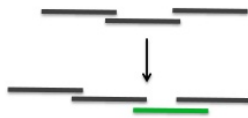
Alt loci



Model Centromere Sequences



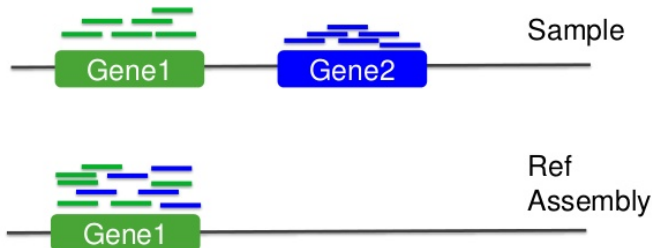
DPYD
GGAACGCAG
GGAACACAG
R->C



<http://www.slideshare.net/GenomeRef/transitioning-to-grch38>

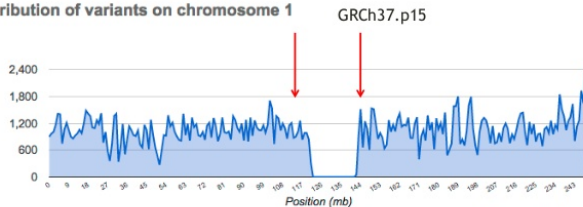
GRCh38 – advantage for variant calling

Reference assembly influence

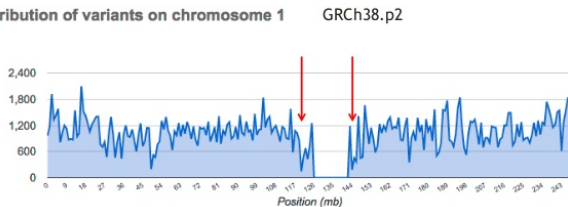


Avoiding collapsed repeats

Distribution of variants on chromosome 1



Distribution of variants on chromosome 1



<http://www.slideshare.net/kmsteinberg/>

the-importance-of-high-quality-reference-genome-assemblies-to-personal-and-medical-genomics

Comparison

- Build 37 and 38
- Validation sets: Genome in a Bottle, Illumina Platinum Genomes
- Lift-over methods: CrossMap/LiftOver, NCBI Remap
- 38 builds: with/without alternative alleles
- Variant callers: FreeBayes, GATK
HaplotypeCaller

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



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>

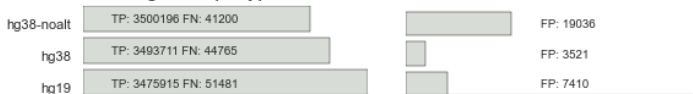
<https://www.synapse.org/#!Synapse:syn312572>

hg19/hg38 comparison: NA12878 Platinum Genomes

SNPs: freebayes



SNPs: gatk-haplotype



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: freebayes



Indels: gatk-haplotype



0% 2% 4% 6% 8% 10% 12%

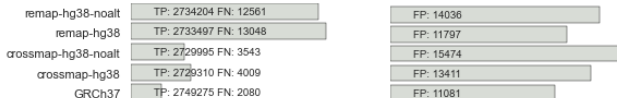
0% 2% 4% 6% 8% 10% 12%

False negative rate

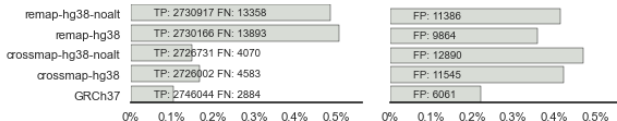
False discovery rate

GRCh37/hg38 comparison: NA12878 Genome in a Bottle

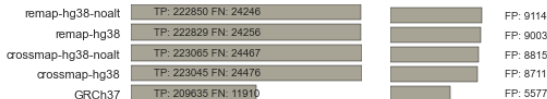
SNPs: freebayes



SNPs: gatk-haplotype



Indels: freebayes



Indels: gatk-haplotype



False negative rate

False discovery rate

- SNPs: build 38 more sensitive
- SNPs: build 38 reduces false positives
- Indels: build 38 detected more
- Indels: work on sensitivity and precision

Need conversion approaches for resources not yet available on build 38

- CrossMap:

<http://crossmap.sourceforge.net/>

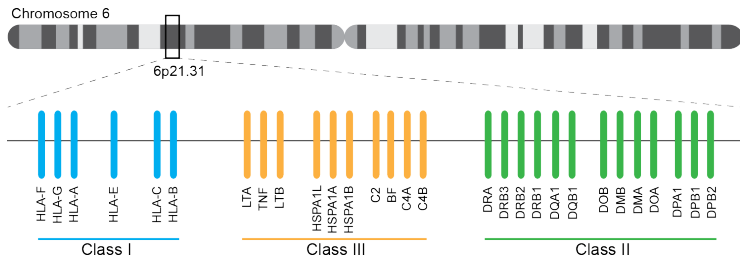
- NCBI remap:

<http://www.ncbi.nlm.nih.gov/genome/tools/remap>

- Both performed well

- NCBI remap has additional sensitivity, but requires tuning

Major histocompatibility complex (MHC) – HLAs



<http://www.ebi.ac.uk/ipd/imgt/hla/>

<http://sciscogenetics.com/technology/human-leukocyte-antigen-complex/>

Alignment: bwa alternative allele support

Read: ATCAGCATC

```
ALT ctg 1:      TGAAA---CGAATGCAAATGGTCAATCAGCATCGAACTAGTCACAT
                ||||| (high div) ||||| (novel ins) |||||
Chromosome: GCGTACATGATACGAATCgGCATCATGGTC-----CTAGTCACATCGTAATC
                ||||| ||||| (novel ins) |||||
ALT ctg 2:      TGATACGAATCgcCATCATGGTCAATCgcAgCGAACTAGTCACAT
```

4 potential hits: **ATCAGCATC** > **ATCgGCATC** > **ATCgcCATC** > **ATCgcAgC**

2 hit groups: {**ATCAGCATC**, **ATCgcAgC**} and {**ATCgGCATC**, **ATCgcCATC**}

Hits considered in mapQ: **ATCAGCATC** and **ATCgGCATC** (best from each group)

In the output SAM: **ATCgGCATC** as the primary SAM line with mapQ=0

ATCAGCATC as a supplementary line with mapQ>0

ATCgcAgC as a supplementary line with mapQ>0

ATCgcCATC in an XA tag, not as a separate line

<https://github.com/lh3/bwa/blob/master/README-alt.md>

HLA typing

- 1000 genomes: build 38 + IMGT/HLA-3.18.0
- bwa mem extracts HLA reads
- Map reads only to HLA sequences
- OptiType: Call HLA types

<https://github.com/lh3/bwa/blob/master/README-alt.md#hla-typing>

<https://github.com/FRED-2/OptiType>

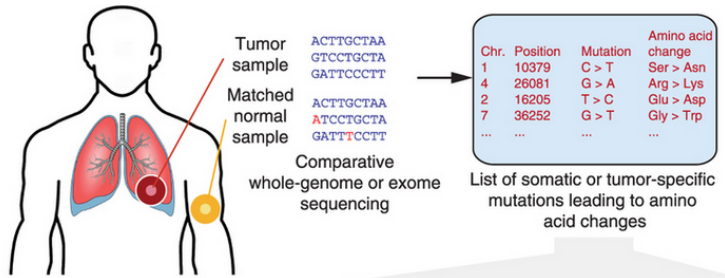
- Omixon example data
- Exome (1000 genomes) and deep targeted data
- P-group resolution
- HLA type I calls (A, B, C)
- Great results across exome and targeted

<http://www.omixon.com/hla-typing-example-data/>

<https://gist.github.com/chapmanb/8f994618a7fc5e88f893>

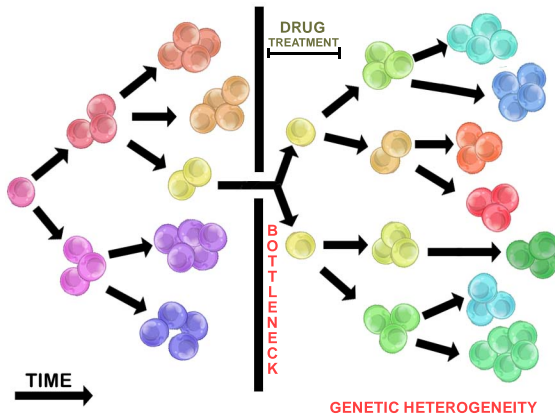
- Human build 38 + HLA
- **Low frequency somatic calling**
- Structural variation

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

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

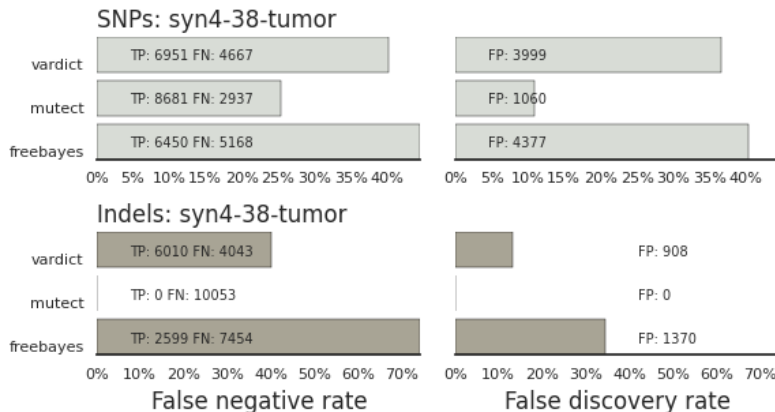
<https://github.com/AstraZeneca-NGS/VarDictJava>

DREAM synthetic dataset 4

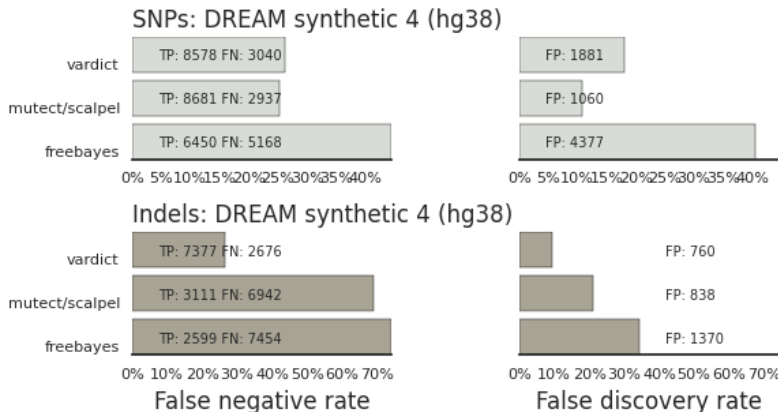
<i>in silico 3</i>	<i>in silico 4</i>
BWA Backtrack	BWA MEM
SNV, SV (deletions, duplications, insertions, inversions) & INDEL	SNV, SV (deletions, duplications, inversions) & INDEL
100%	80%
50%, 33%, 20%	50%, 35% (effectively 30% and 15% due to cellularity)
Female	Male
HCC1143 BL from TCGA Benchmark 4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

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

VarDict sensitivity/precision before



VarDict sensitivity/precision after

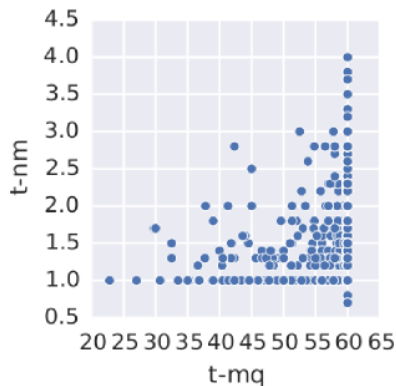


How? Filter summary

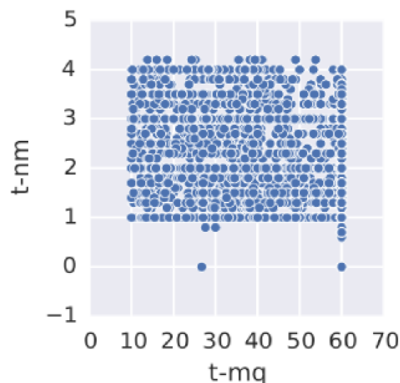
```
((AF * DP < 6) &&  
  ((MQ < 55.0 && NM > 1.0) ||  
   (MQ < 60.0 && NM > 2.0) ||  
   (DP < 10) ||  
   (QUAL < 45)))
```

Filter: mapping quality and number of mismatches

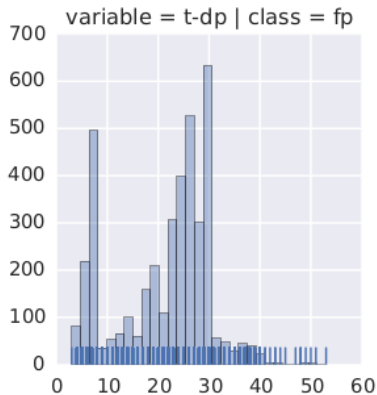
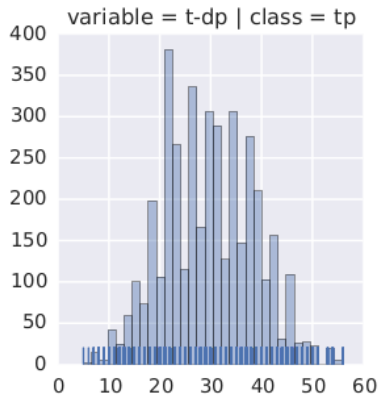
True positives



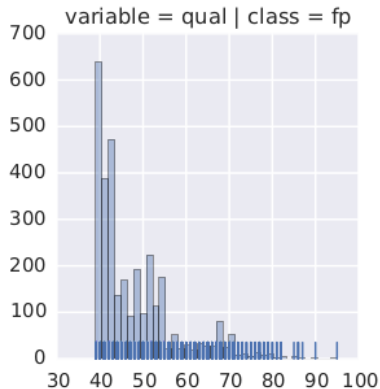
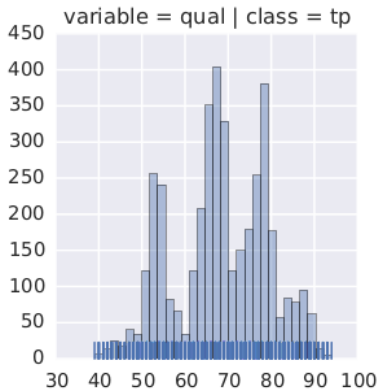
False positives



Filter: low depth



Filter: low quality

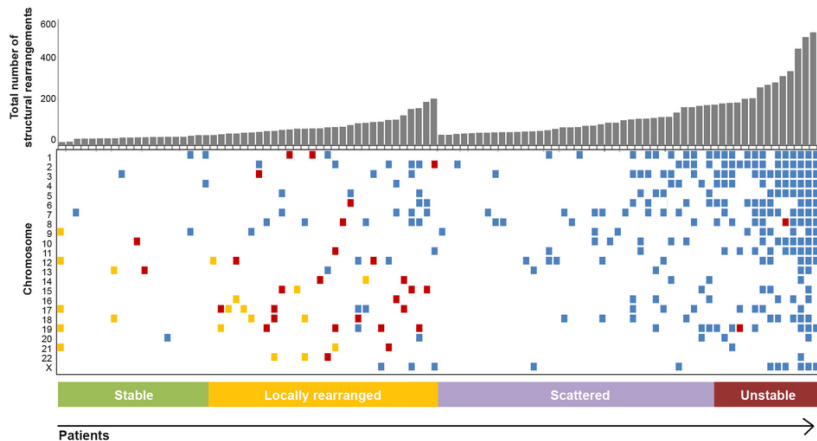


How can we improve?

- Incorporate machine learning methods
- Generalize with additional datasets
- AML31: <http://aml31.genome.wustl.edu/>

- Human build 38 + HLA
- Low frequency somatic calling
- **Structural variation**

Structural variants critical in cancer

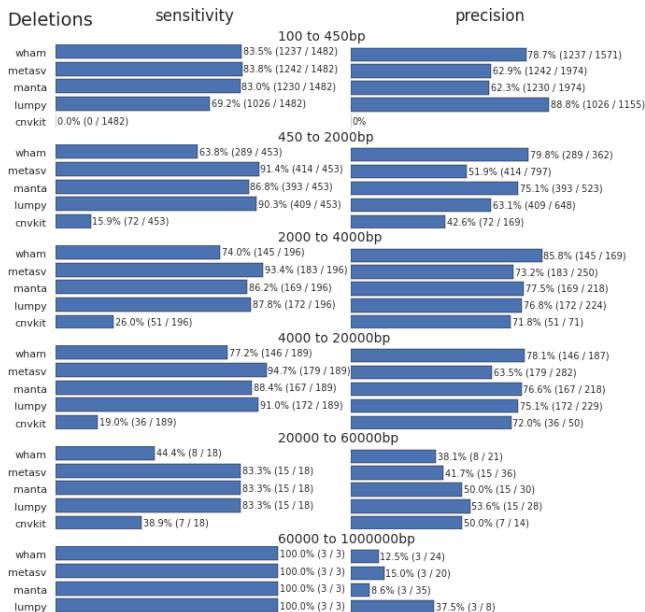


<http://www.nature.com/nature/journal/v518/n7540/full/nature14169.html>

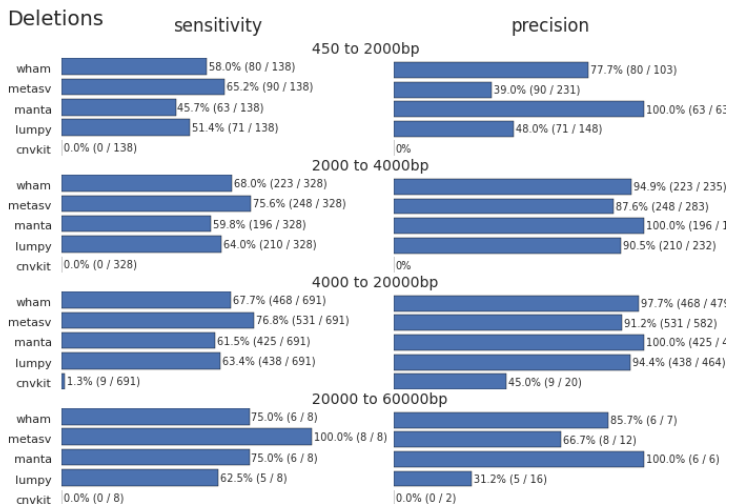
Improvements in speed, sensitivity and precision

- Lumpy: <https://github.com/arq5x/lumpy-sv>
- Manta: <https://github.com/Illumina/manta>
- CNVkit: <https://github.com/etal/cnvkit>
- WHAM: <https://github.com/zeeev/wham>
- MetaSV: <https://github.com/bioinform/metasv>

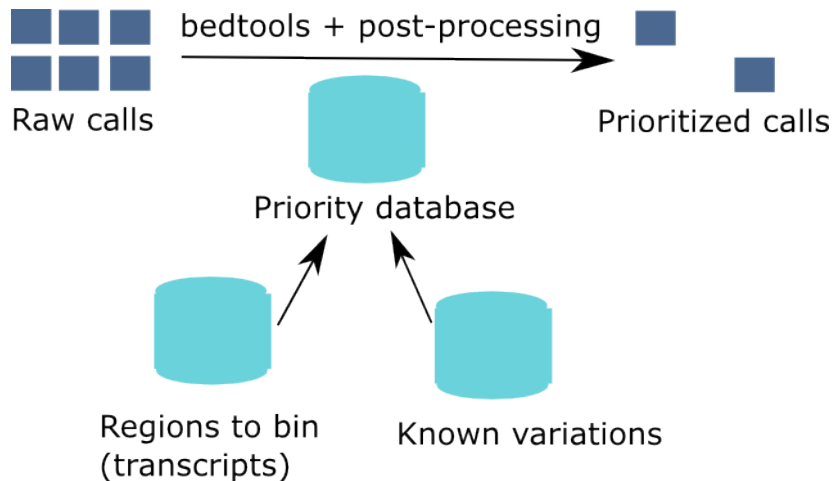
Results: Germline deletions



Results: Somatic deletions

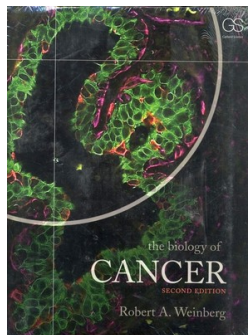


Prioritize in previously known regions



Public cancer variant databases

- CIViC: <https://civic.genome.wustl.edu>
- IntOGen: <http://www.intogen.org>



<http://www.amazon.com/The-Biology-Cancer-Robert-Weinberg/dp/0815340761>

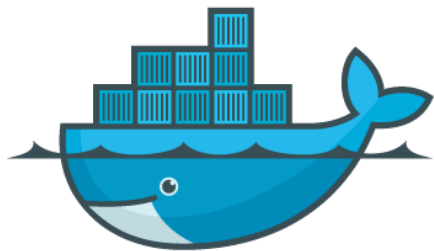
- Runs locally on a single machine with multiple cores
- Also works in parallel on compute clusters (LSF, SGE, Slurm, Torque/PBS)
- Pre-installed and maintained on Odyssey and Orchestra
- Uses custom IPython-based backend for parallelization



COMMON
WORKFLOW
LANGUAGE

<http://www.commonwl.org/>

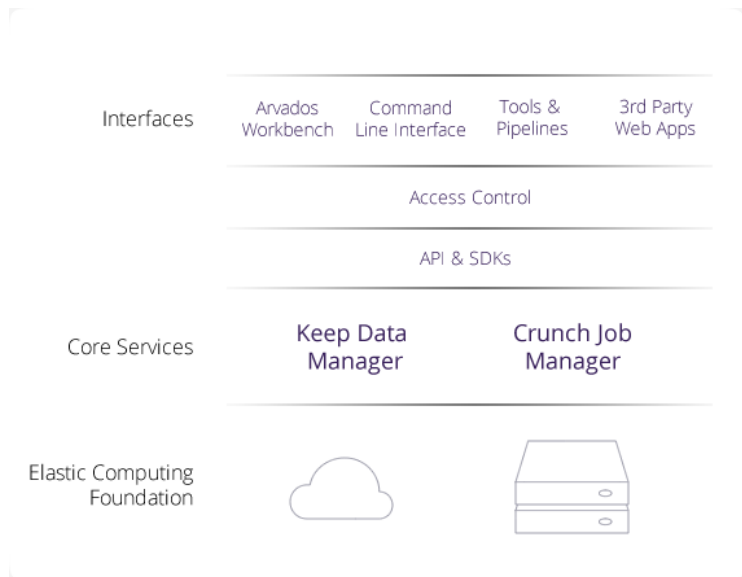
Docker: code and tool isolation



docker

<http://docker.com>

Arvados: provenance and reproducibility



<https://arvados.org>

Toil: multiple CWL workflow runners

- From NIH Big Data to Knowledge Center for Translation Genomics at UCSC
- Alternative workflow system
- Supports CWL
- Lightweight with cluster integration

<http://toil.readthedocs.org/en/latest/>

<https://github.com/BD2KGenomics/toil>

Summary

- Open source community sequencing analysis
- bcbio validated variant calling
- Science
 - Human build 38 + HLA
 - Cancer calling of low frequency variants
 - Structural variation
- Move to community infrastructure

<http://bcb.io>