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

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

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?
8
↓ (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/

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

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**
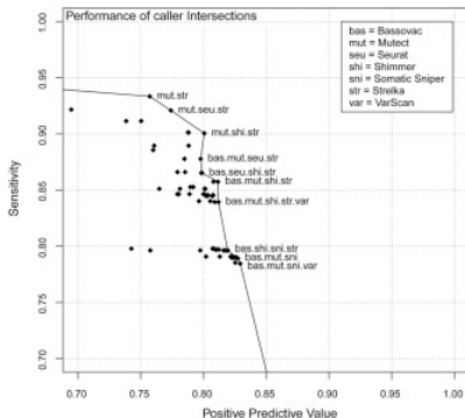
Working together produces great things

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
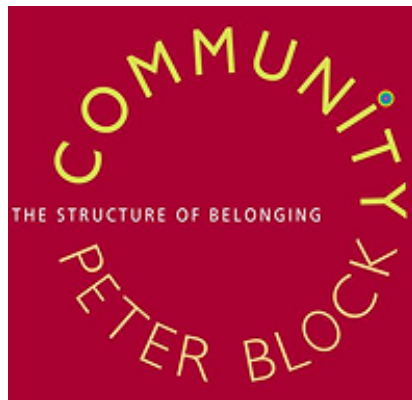- Namrata Gupta

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)
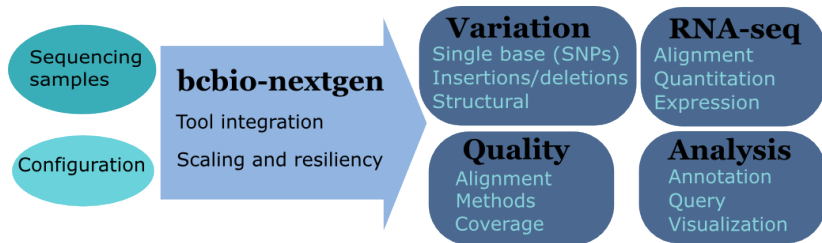- NIDDK U54 DK105566 (PIs: MacArthur and Neale)

http://exac.broadinstitute.org/about

Solution

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

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, HISAT2
- Variantion: 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

http://www.bioplanet.com/gcat
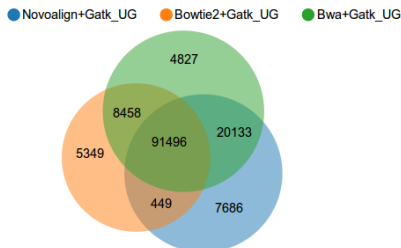
We made a pipeline – so what?

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

Community: sustainability

Jul 18, 2010 – Mar 22, 2016

Contributions to master, excluding merge commits

Contributions: **Commits** ▾

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

Community: support

Community: contribution

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

Community: documentation

https://bcbio-nextgen.readthedocs.org

Supported analysis types

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

GRCh37/hg19
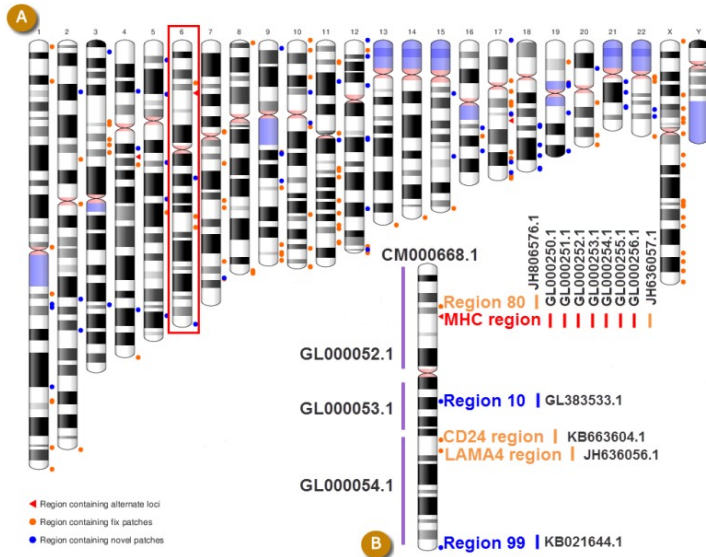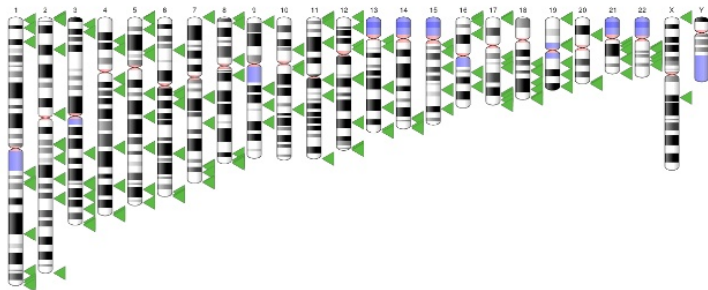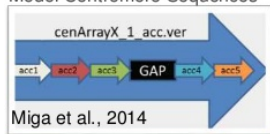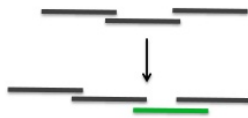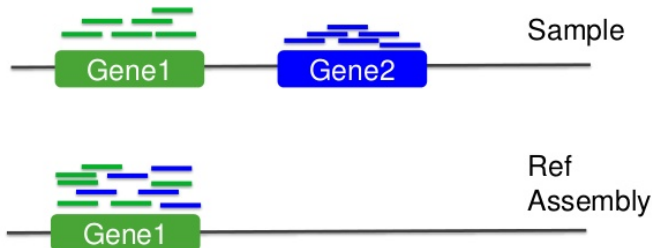
Avoiding collapsed repeats
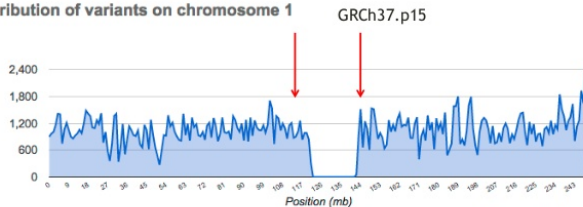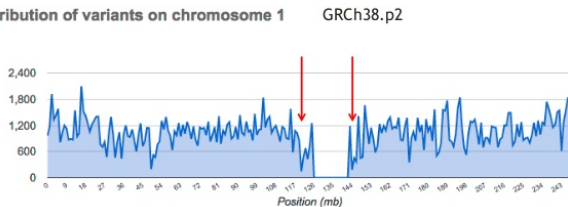
- Build 37 and 38
- Validation sets: Genome in a Bottle, Illumina Platinum Genomes
- Lift-over methods: CrossMap/LiftOver, NCBI Remap
- 38 builds: with/without alternative alleles
- Variant callers: FreeBayes, GATK HaplotypeCaller

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

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

hg19/hg38 comparison: NA12878 Platinum Genomes
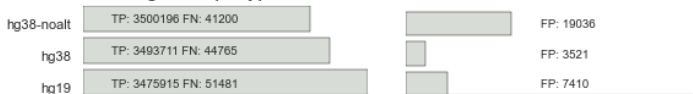
GRCh37/hg38 comparison: NA12878 Genome in a Bottle
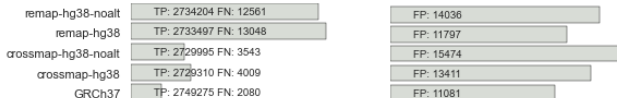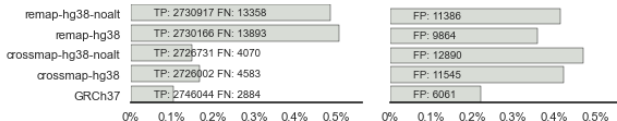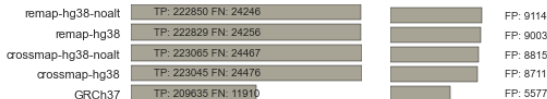
- SNPs: build 38 more sensitive
- SNPs: build 38 reduces false positives
- Indels: build 38 detected more
- Indels: work on sensitivity and precision

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/
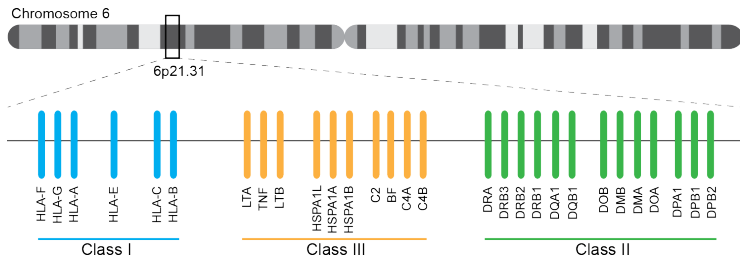
Alignment: bwa alternative allele support

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

- 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

Comparative whole-genome or exome sequencing

List of somatic or tumor-specific mutations leading to amino acid changes
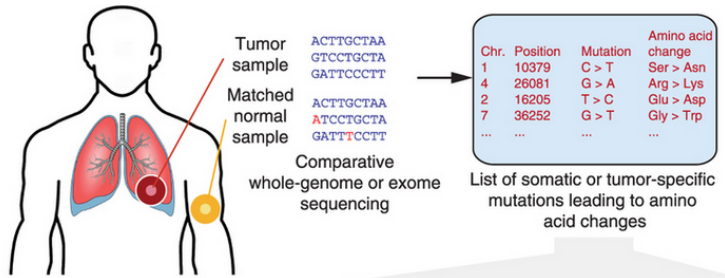
http://www.nature.com/nmeth/journal/v10/n8/fig_tab/nmeth.2562_F1.html

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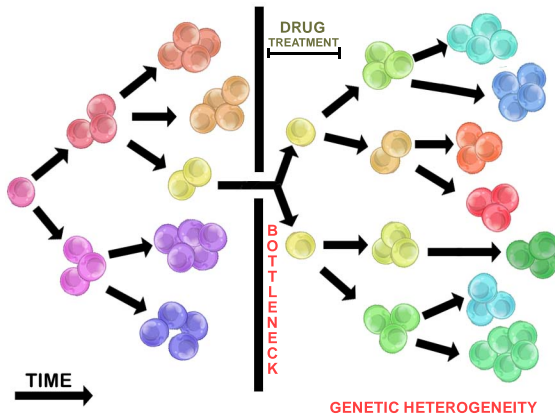
- AstraZeneca
- SNP + Insertion/Deletions
- Whole genome + exome
- Also works on deep targeted data

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

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

SNPs: syn4-38-tumor

vardict	TP: 6951 FN: 4667
mutect	TP: 8681 FN: 2937
freebayes	TP: 6450 FN: 5168

	FP: 3999
	FP: 1060
	FP: 4377

0% 5% 10% 15% 20% 25% 30% 35% 40%   0% 5% 10% 15% 20% 25% 30% 35% 40%

Indels: syn4-38-tumor

vardict	TP: 6010 FN: 4043
mutect	TP: 0 FN: 10053
freebayes	TP: 2599 FN: 7454

	FP: 908
	FP: 0
	FP: 1370

0% 10% 20% 30% 40% 50% 60% 70%   0% 10% 20% 30% 40% 50% 60% 70%

False negative rate   False discovery rate

VarDict sensitivity/precision after

SNPs: DREAM synthetic 4 (hg38)

vardict — TP: 8578 FN: 3040 — FP: 1881

mutect/scalpel — TP: 8681 FN: 2937 — FP: 1060

freebayes — TP: 6450 FN: 5168 — FP: 4377

0% 5%10%15%20%25%30%35%40%

Indels: DREAM synthetic 4 (hg38)

vardict — TP: 7377 FN: 2676 — FP: 760

mutect/scalpel — TP: 3111 FN: 6942 — FP: 838

freebayes — TP: 2599 FN: 7454 — FP: 1370

0% 10% 20%30% 40%50%60% 70%

False negative rate

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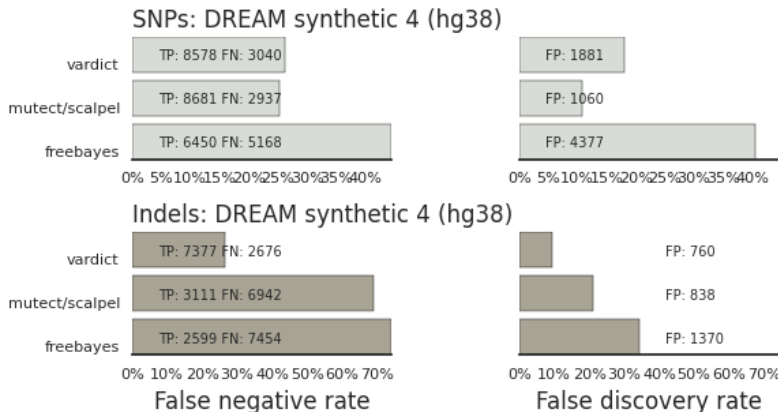
```
((AF * DP < 6) &&
 ((MQ < 55.0 && NM > 1.0) ||
  (MQ < 60.0 && NM > 2.0) ||
  (DP < 10) ||
  (QUAL < 45)))
```

Filter: mapping quality and number of mismatches

True positives
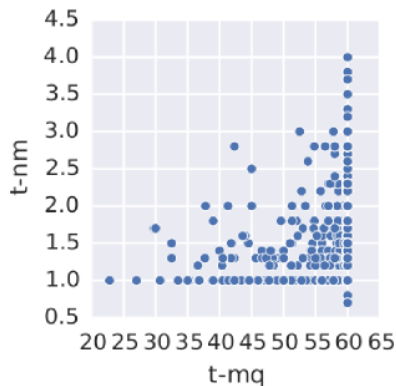
False positives
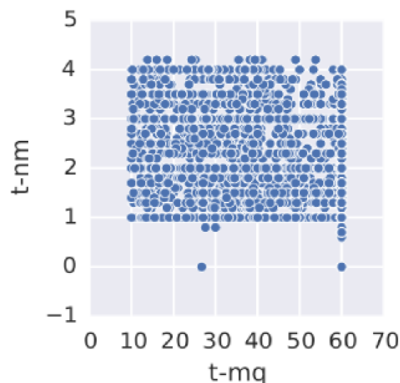
Filter: low depth
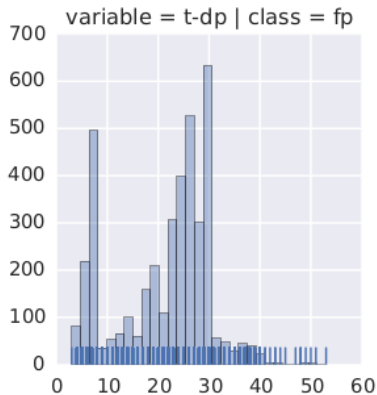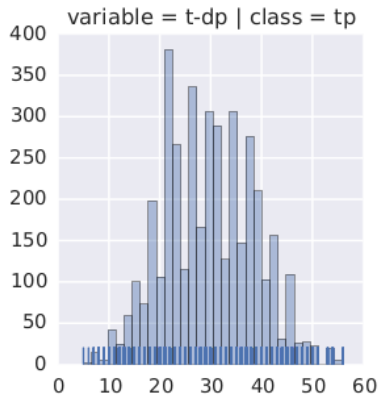
Filter: low quality
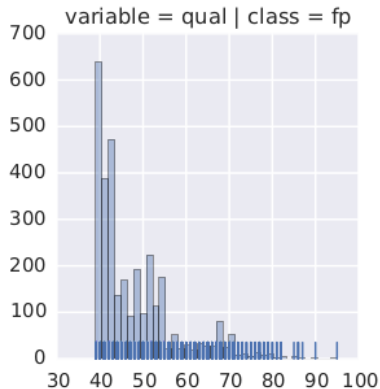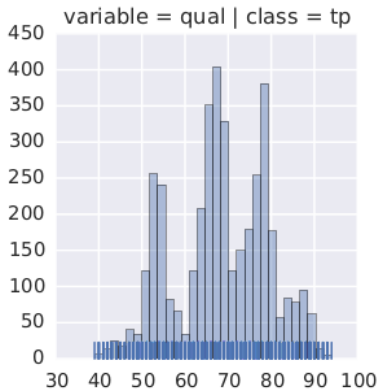
- Incorporate machine learning methods
- Generalize with additional datasets
- AML31: http://aml31.genome.wustl.edu/

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

Structural variants critical in cancer
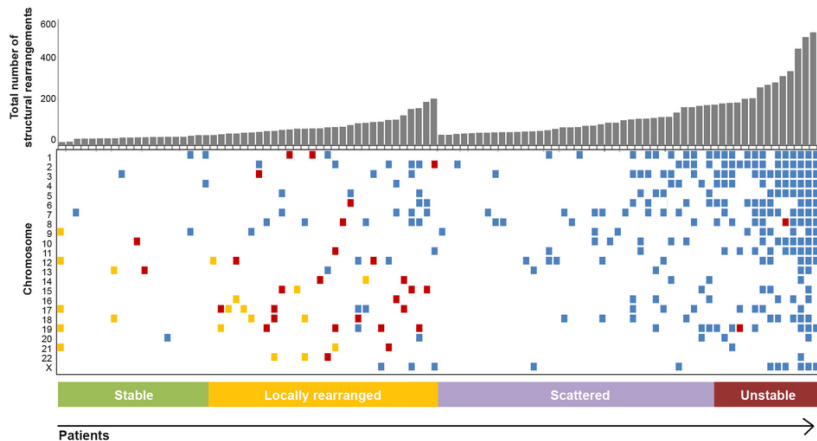
- Lumpy: https://github.com/arq5x/lumpy-sv
- Manta: https://github.com/Illumina/manta
- CNVkit: https://github.com/etal/cnvkit
- WHAM: https://github.com/zeeev/wham
- MetaSV: https://github.com/bioinform/metasv

Results: Germline deletions

Deletions   sensitivity   precision

100 to 450bp

	sensitivity	precision
wham	83.5% (1237 / 1482)	78.7% (1237 / 1571)
metasv	83.8% (1242 / 1482)	62.9% (1242 / 1974)
manta	83.0% (1230 / 1482)	62.3% (1230 / 1974)
lumpy	69.2% (1026 / 1482)	88.8% (1026 / 1155)
cnvkit	0.0% (0 / 1482)	0%

450 to 2000bp

	sensitivity	precision
wham	63.8% (289 / 453)	79.8% (289 / 362)
metasv	91.4% (414 / 453)	51.9% (414 / 797)
manta	86.8% (393 / 453)	75.1% (393 / 523)
lumpy	90.3% (409 / 453)	63.1% (409 / 648)
cnvkit	15.9% (72 / 453)	42.6% (72 / 169)

2000 to 4000bp

	sensitivity	precision
wham	74.0% (145 / 196)	85.8% (145 / 169)
metasv	93.4% (183 / 196)	73.2% (183 / 250)
manta	86.2% (169 / 196)	77.5% (169 / 218)
lumpy	87.8% (172 / 196)	76.8% (172 / 224)
cnvkit	26.0% (51 / 196)	71.8% (51 / 71)

4000 to 20000bp

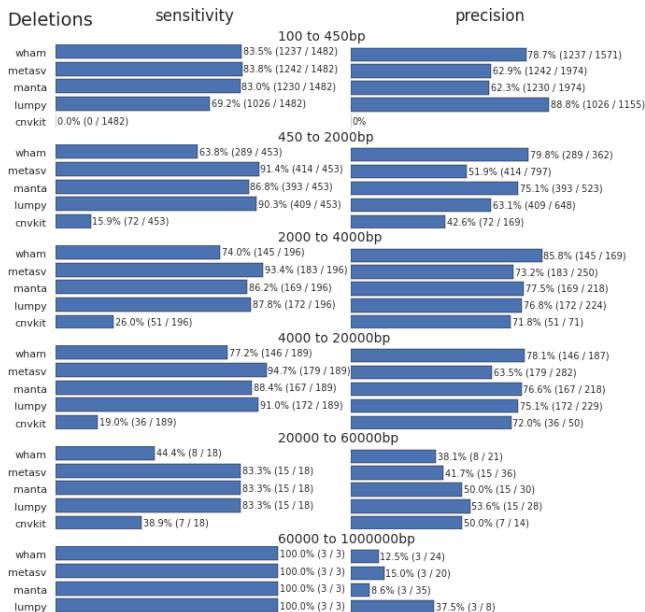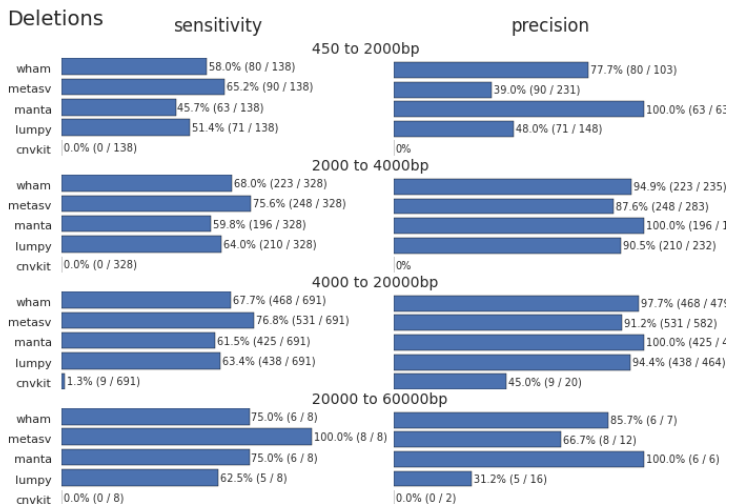|	sensitivity	precision
wham	77.2% (146 / 189)	78.1% (146 / 187)
metasv	94.7% (179 / 189)	63.5% (179 / 282)
manta	88.4% (167 / 189)	76.6% (167 / 218)
lumpy	91.0% (172 / 189)	75.1% (172 / 229)
cnvkit	19.0% (36 / 189)	72.0% (36 / 50)

20000 to 60000bp

	sensitivity	precision
wham	44.4% (8 / 18)	38.1% (8 / 21)
metasv	83.3% (15 / 18)	41.7% (15 / 36)
manta	83.3% (15 / 18)	50.0% (15 / 30)
lumpy	83.3% (15 / 18)	53.6% (15 / 28)
cnvkit	38.9% (7 / 18)	50.0% (7 / 14)

60000 to 1000000bp

	sensitivity	precision
wham	100.0% (3 / 3)	12.5% (3 / 24)
metasv	100.0% (3 / 3)	15.0% (3 / 20)
manta	100.0% (3 / 3)	8.6% (3 / 35)
lumpy	100.0% (3 / 3)	37.5% (3 / 8)

Results: Somatic deletions
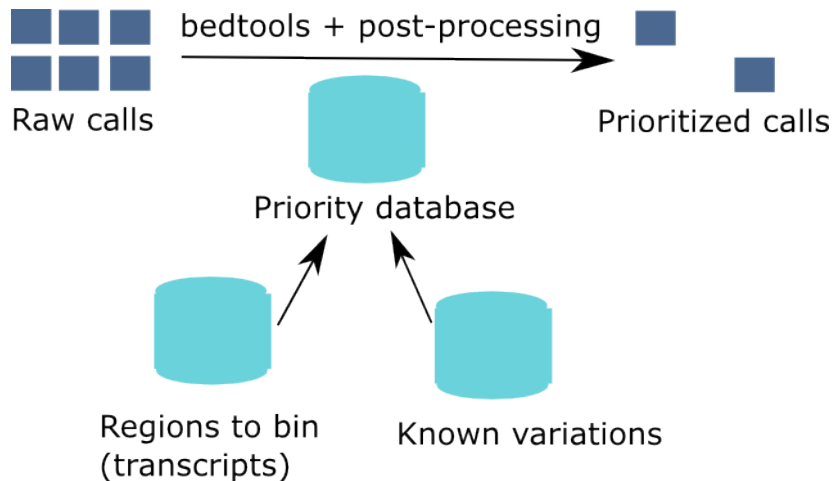
Public cancer variant databases

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

http://www.amazon.com/The-Biology-Cancer-Robert-Weinberg/dp/0815340761
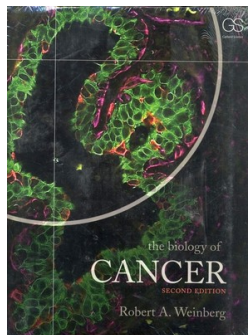
- Runs locally on a single machine with multiple cores
- Also works in parallel on compute clusters (LSF, SGE, Slurm, Torque/PBS)
- Pre-installed and maintained on Odyssey and Orchestra
- Uses custom IPython-based backend for parallelization
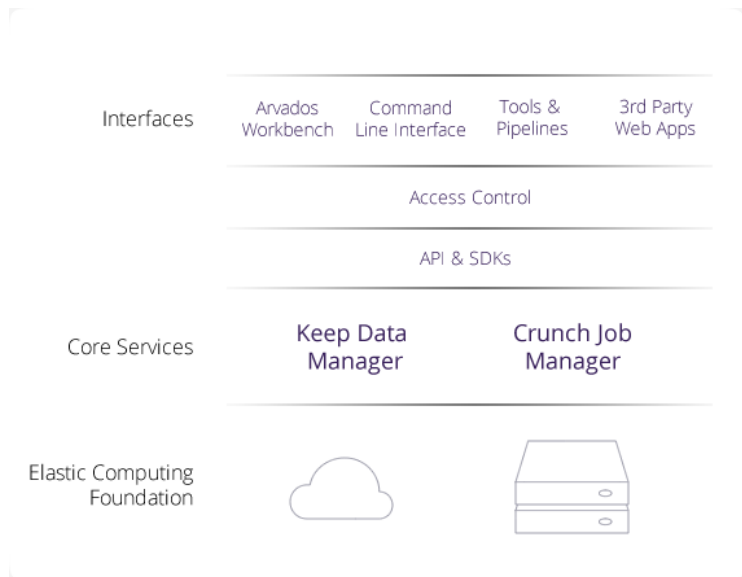
COMMON
WORKFLOW
LANGUAGE

http://www.commonwl.org/

http://docker.com
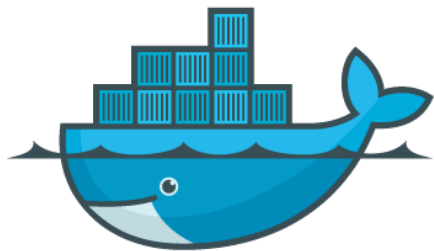
Arvados: provenance and reproducibility

https://arvados.org

- 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

- 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