Overview of 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

11 January 2016

- Motivate for using open source community resources
- Overview of bcbio validated variant calling
- Integration goals with Arvados
- Science
  - Human build 38 + HLA
  - Cancer calling of low frequency variants
  - Structural variation

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

↑
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.

Daily bioinformatics work

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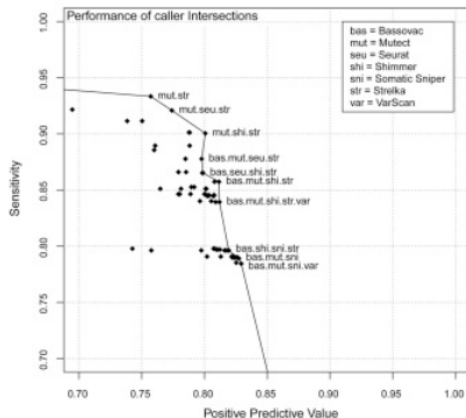- Improve algorithms
- Scale
- Read literature
- Do biology

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

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
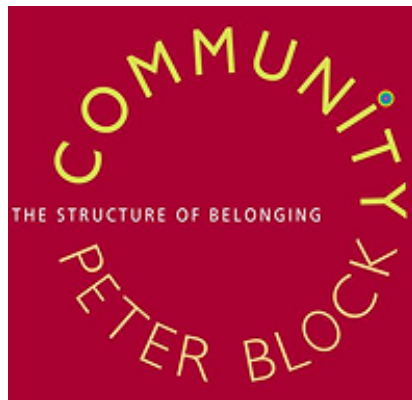
Broad Genomics Platform
- Stacey Gabriel

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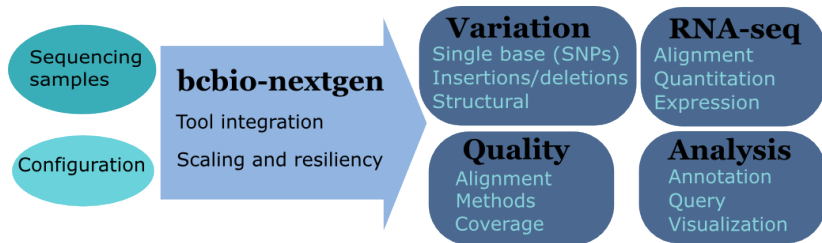- NIDDK U54 DK105566 (PIs: MacArthur and Neale)

http://exac.broadinstitute.org/about

Solution

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

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
- Tool integration
- Validation – outputs + automated evaluation
- Scaling
- Installation of tools and data

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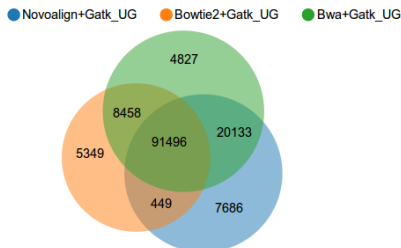
We made a pipeline – so what?

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

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 – Nov 2, 2015

Contributions to master, excluding merge commits

Contributions: **Commits**
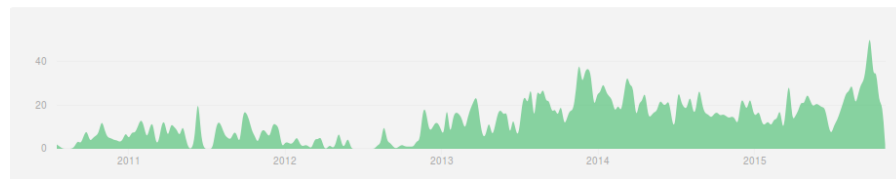
https://github.com/chapmanb/bcbio-nextgen

Community: support

Community: contribution

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

Community: documentation

https://bcbio-nextgen.readthedocs.org

- Replace bcbio internal workflow with CWL
- bcbio generates CWL
- Arvados/Crunch runs generated CWL with bcbio Docker container
- Plug in sections of bcbio code into Arvados
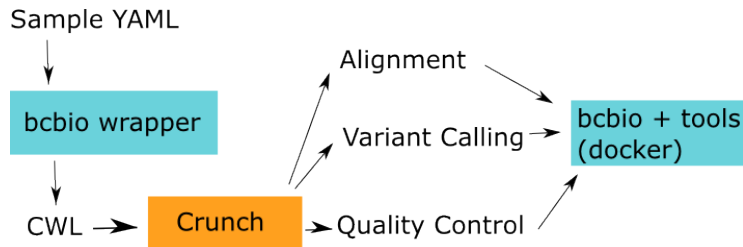
- Standard command line interface to call bcbio sub-function

- Docker container (bcbio/bcbio) with all tools installed

- Generates CWL for small pipelines, with sample-level parallelization

- Runs with cwltool

https://github.com/chapmanb/bcbio-nextgen/tree/master/cwl

- Run with CWL enabled Crunch
- Work on ways to integrate bcbio data with Keep
- Provide standard pipelines in bcbio as part of Arvados
- Plug in sections of bcbio code to interoperate with other tools

- Advanced parallelization (split samples, genomic regions)
- Port remaining bcbio functionality (define inputs/outputs for each step)
- Testing with other CWL implementations (Galaxy, Seven Bridges)

- **Human build 38 + HLA**
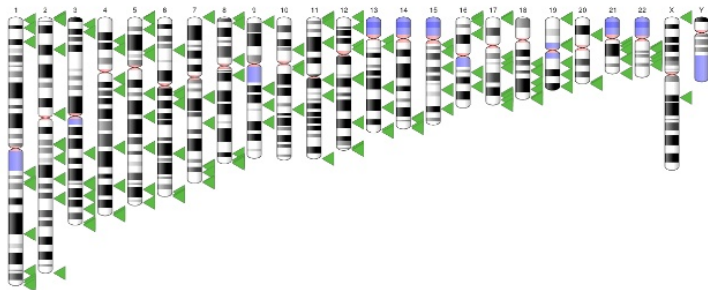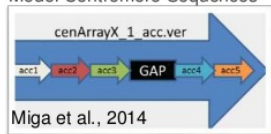- Low frequency somatic calling
- Structural variation
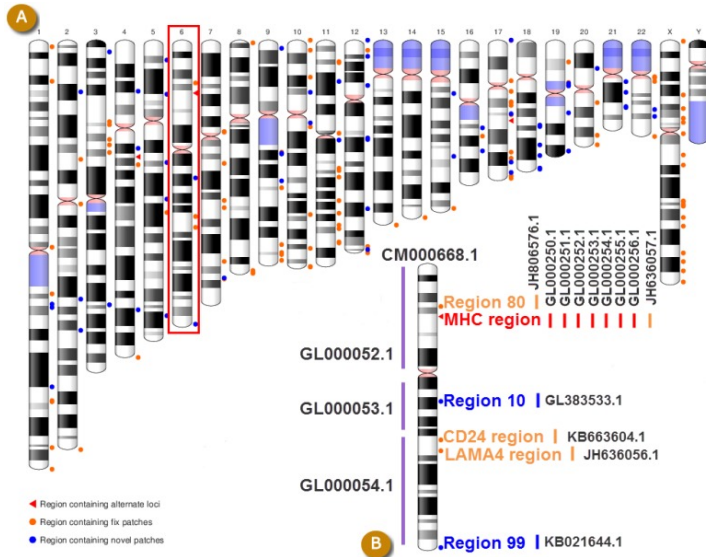
GRCh38 – graph based, many more alternative loci
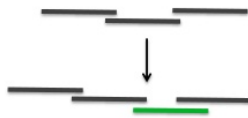
Reference assembly influence

Sample

Gene1 Gene2

Ref
Assembly

Gene1
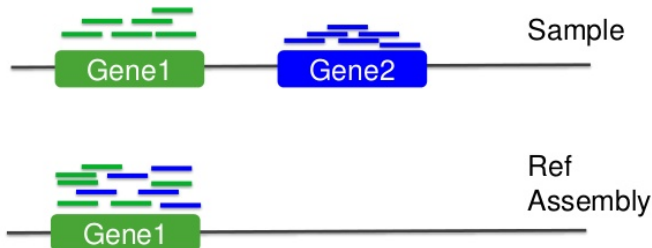
Personalis

Avoiding collapsed repeats
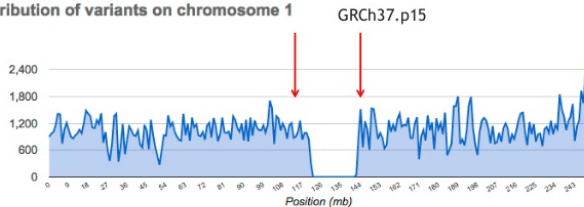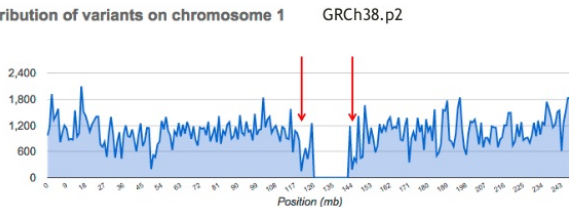
Comparison

- 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

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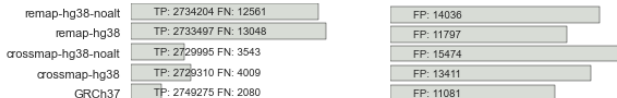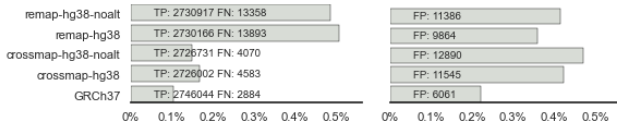
- SNPs: build 38 more sensitive
- SNPs: build 38 reduces false positives
- Indels: build 38 detected more
- Indels: work on sensitivity and precision

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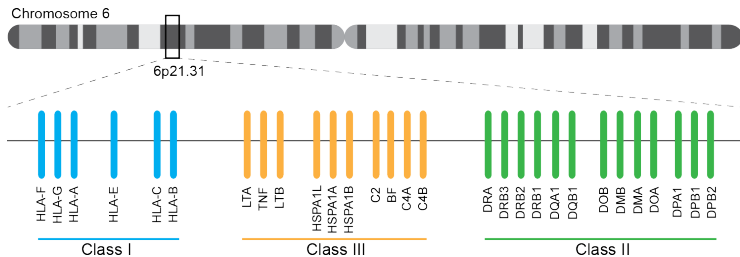
http://sciscogenetics.com/technology/human-leukocyte-antigen-complex/

Alignment: bwa alternative allele support


```
          Read: ATCAGCATC

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

       4 potential hits: ATCAGCATC > ATCgGCATC > ATCgcCATC > ATCgcCAgC
         2 hit groups: {ATCAGCATC,ATCgcCAgC} 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
                          ATCgcCAgC as a supplementary line with mapQ>0
                          ATCgcCATC in an XA tag, not as a separate line
```
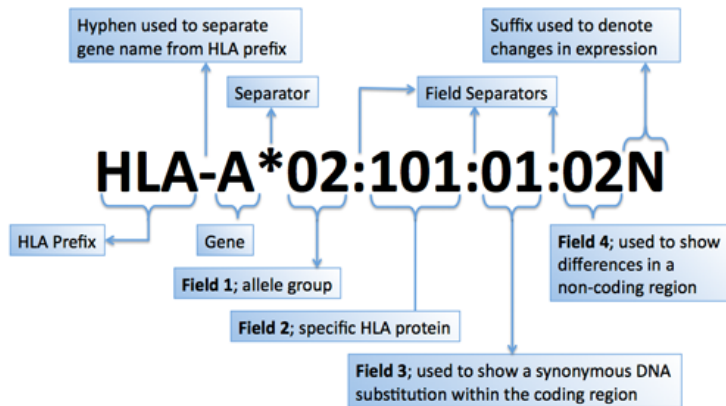
https://github.com/lh3/bwa/blob/master/README-alt.md

- 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

HLA nomenclature

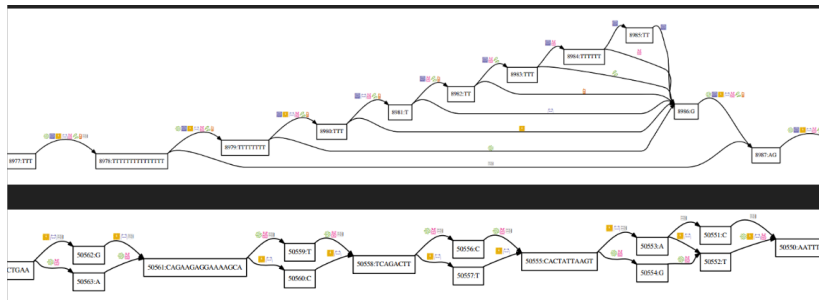© SGE Marsh 04/10

https://www.ebi.ac.uk/ipd/imgt/hla/
http://hla.alleles.org/alleles/p_groups.html

Validations

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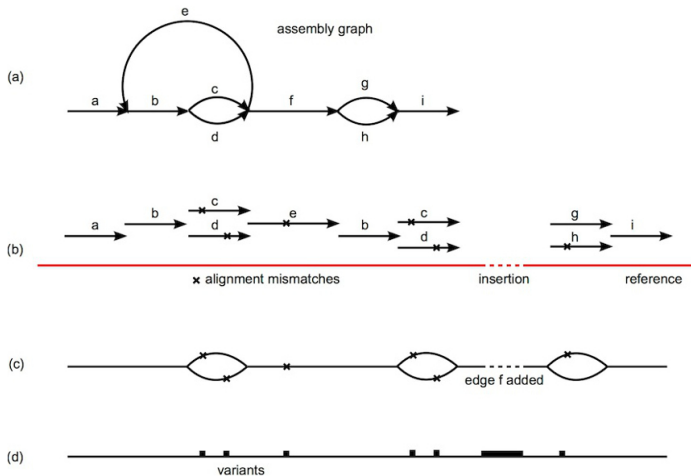- Great results across exome and targeted

http://www.omixon.com/hla-typing-example-data/
https://gist.github.com/chapmanb/8f994618a7fc5e88f893

HLA complexity

Genome graphs and variation

vg – tools for working with variant graphs
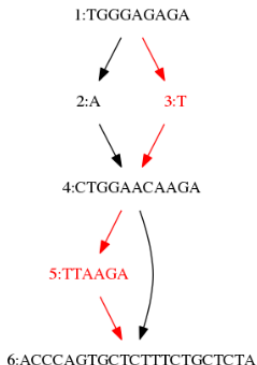
vg – future variant calling
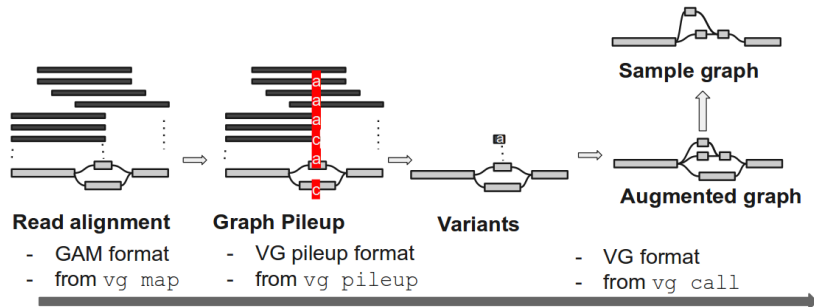
- Human build 38 + HLA
- **Low frequency somatic calling**
- Structural variation

Cancer somatic calling

Comparative whole-genome or exome sequencing

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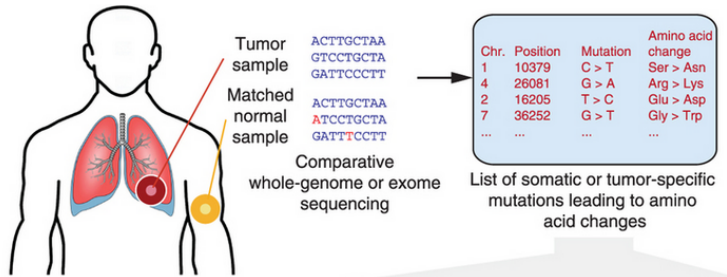
http://www.nature.com/nmeth/journal/v10/n8/fig_tab/nmeth.2562_F1.html

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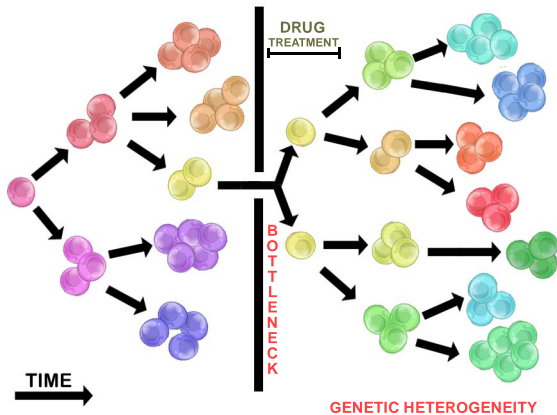
http://en.wikipedia.org/wiki/Tumour_heterogeneity

- AstraZeneca
- SNP + Insertion/Deletions
- Works on very deep targeted data

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

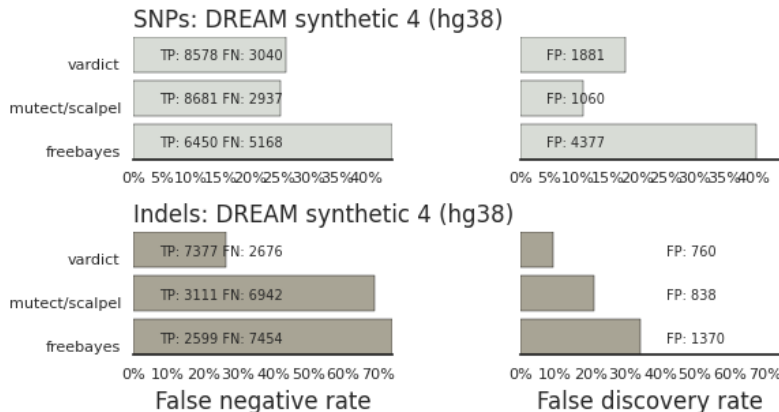| in silico 3	in silico 4
BWA Backtrack	BWA MEM
SNV, SV (deletions, duplications, insertions, inversions) & INDEL	SNV, SV (deletions, duplications, inversions) & INDEL
100%	80%
50%, 33%, 20%	50%, 35% (effectively 30% and 15% due to cellularity)
Female	Male
HCC1143 BL from TCGA Benchmark 4	CPCG0102R (Provided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

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%

False negative rate   False discovery rate
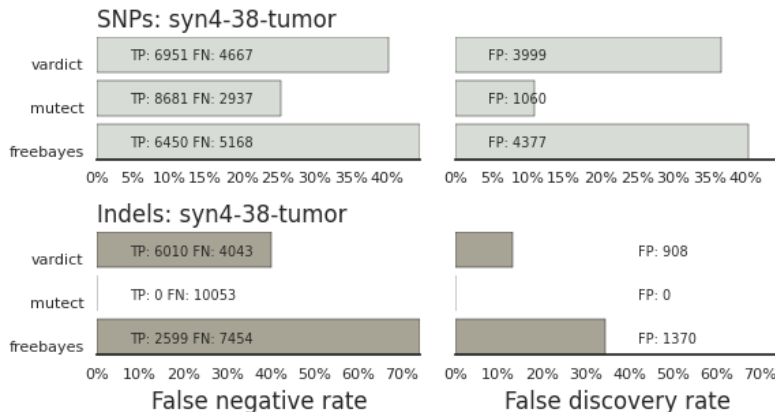
VarDict sensivitity/precision after
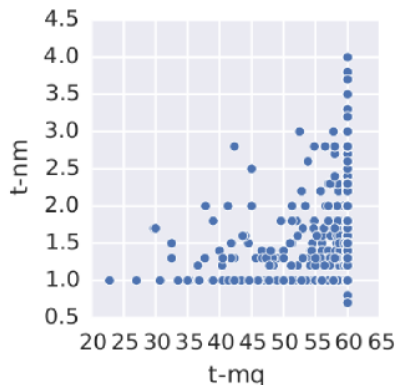
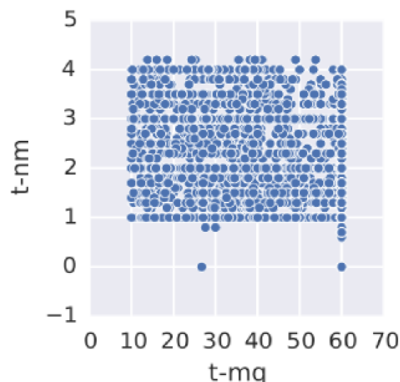```
((AF * DP < 6) &&
 ((MQ < 55.0 && NM > 1.0) ||
  (MQ < 60.0 && NM > 2.0) ||
  (DP < 10) ||
  (QUAL < 45)))
```

# Filter: mapping quality and number of mismatches

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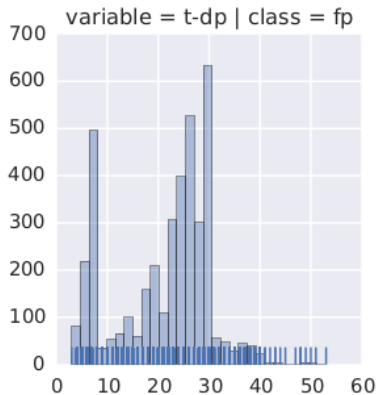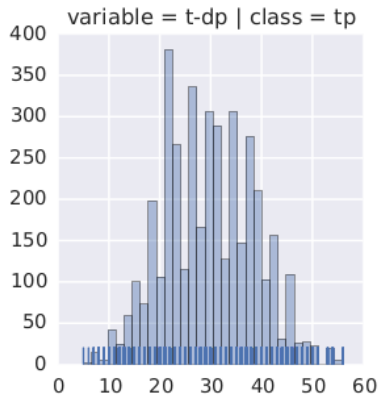
# Filter: low quality
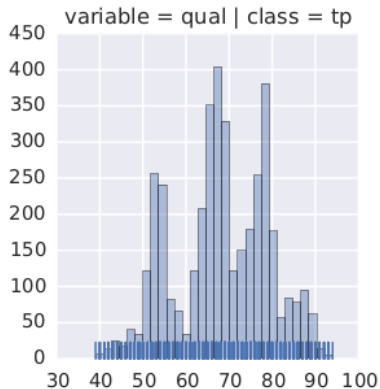
- Incorporate machine learning methods
- Generalize with additional datasets
- AML31: http://aml31.genome.wustl.edu/

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

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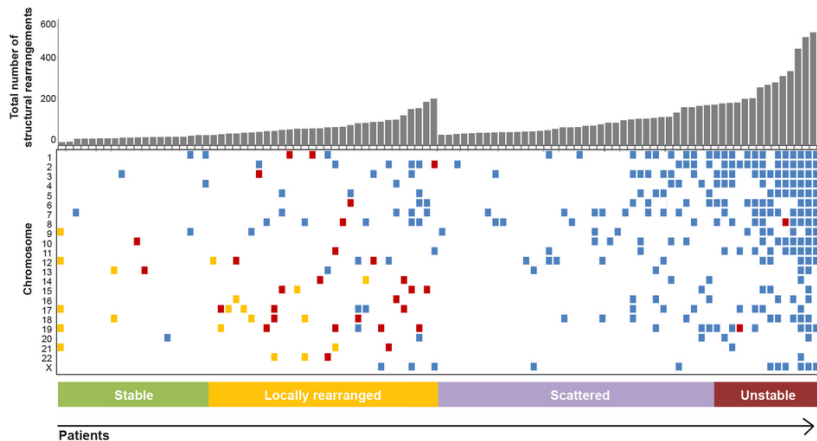
- Lumpy: https://github.com/arq5x/lumpy-sv
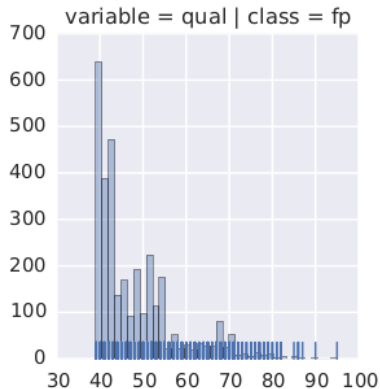- Manta: https://github.com/Illumina/manta
- CNVkit: https://github.com/etal/cnvkit
- WHAM: https://github.com/zeeev/wham
- MetaSV: https://github.com/bioinform/metasv

# Results: Germline deletions

# Results: Somatic deletions

# Results: Somatic insertions



Inversions

sensitivity

precision

### 450 to 2000bp

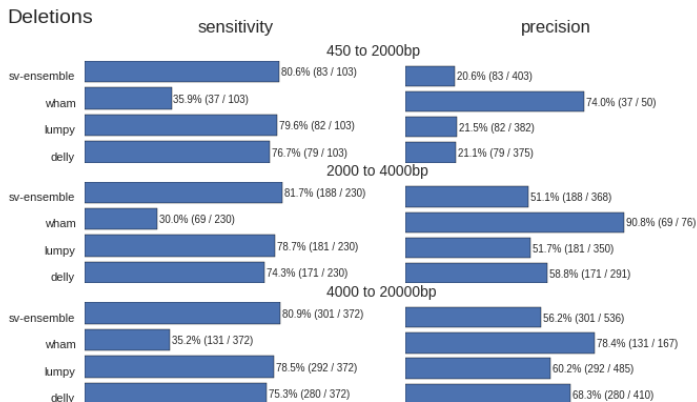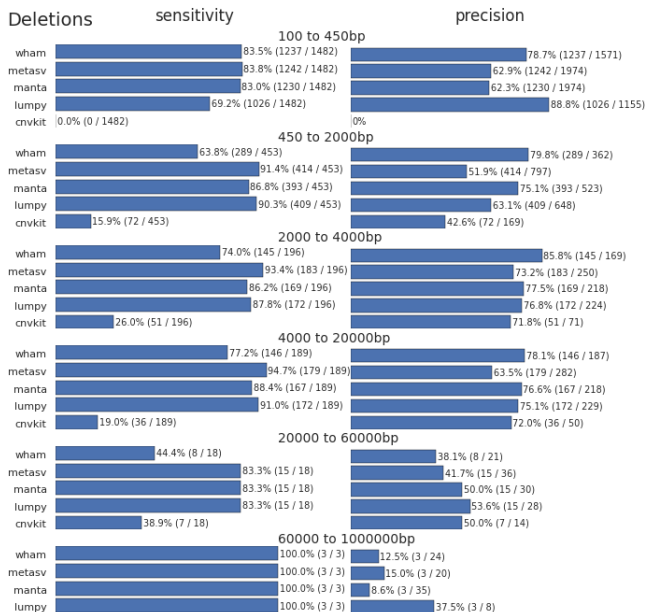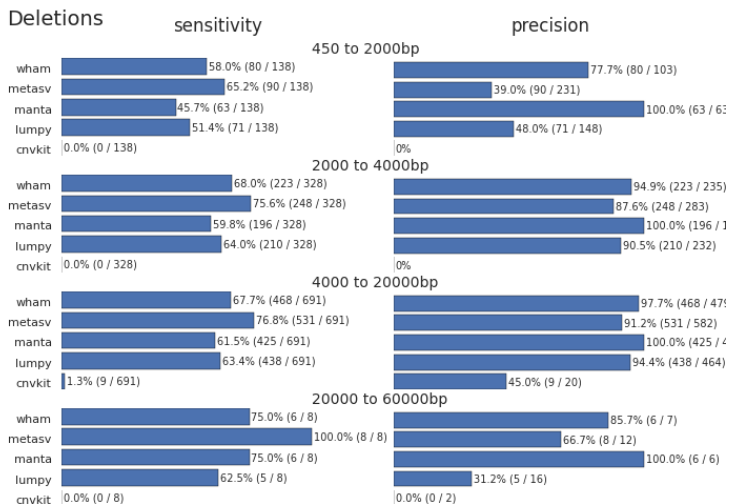| | sensitivity | precision |
|---|---|---|
| wham | 60.3% (35 / 58) | 92.1% (35 / 38) |
| metasv | 72.4% (42 / 58) | 95.5% (42 / 44) |
| manta | 55.2% (32 / 58) | 100.0% (32 / 32 |
| lumpy | 58.6% (34 / 58) | 100.0% (34 / 34 |
| cnvkit | 0.0% (0 / 58) | 0% |

### 2000 to 4000bp

| | sensitivity | precision |
|---|---|---|
| wham | 70.8% (114 / 161) | 100.0% (114 / 1 |
| metasv | 76.4% (123 / 161) | 100.0% (123 / 1 |
| manta | 69.6% (112 / 161) | 100.0% (112 / 1 |
| lumpy | 65.8% (106 / 161) | 100.0% (106 / 1 |
| cnvkit | 0.0% (0 / 161) | 0% |

### 4000 to 20000bp

| | sensitivity | precision |
|---|---|---|
| wham | 72.2% (564 / 781) | 99.6% (564 / 56 |
| metasv | 84.4% (659 / 781) | 98.9% (659 / 66 |
| manta | 71.7% (560 / 781) | 98.9% (560 / 56 |
| lumpy | 65.9% (515 / 781) | 99.8% (515 / 51 |
| cnvkit | 0.0% (0 / 781) | 0% |

### 20000 to 60000bp

| | sensitivity | precision |
|---|---|---|
| wham | 58.6% (17 / 29) | 94.4% (17 / 18) |
| metasv | 69.0% (20 / 29) | 95.2% (20 / 21) |
| manta | 58.6% (17 / 29) | 100.0% (17 / 17 |
| lumpy | 44.8% (13 / 29) | 100.0% (13 / 13 |
| cnvkit | 0.0% (0 / 29) | 0% |

# 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
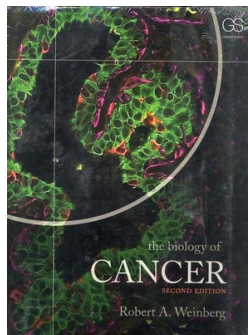
- Open source community resources
- bcbio validated variant calling
- Science
  - Human build 38 + HLA
  - Cancer calling of low frequency variants
  - Structural variation

http://bcb.io