Learning variant calling with validated, scalable, community developed tools

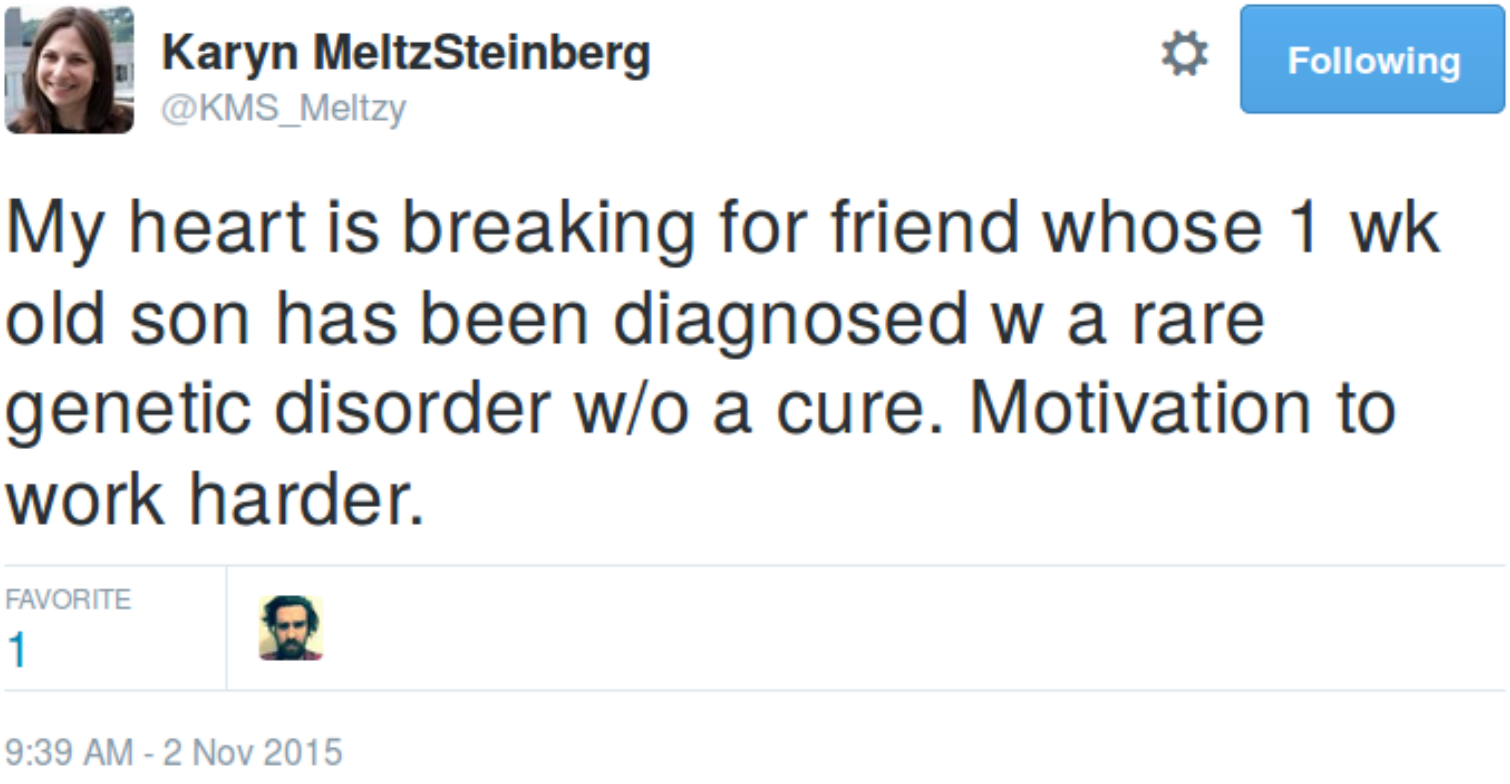
Brad Chapman
Bioinformatics Core, Harvard Chan School
https://bcb.io
http://j.mp/bcbiolinks

18 November 2015

Outline

- Motivate for using open source community resources
- Overview of bcbio validated variant calling
- Science
  - Human build 38
  - Cancer calling of low frequency variants
  - Structural variation
- Practical calling example

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

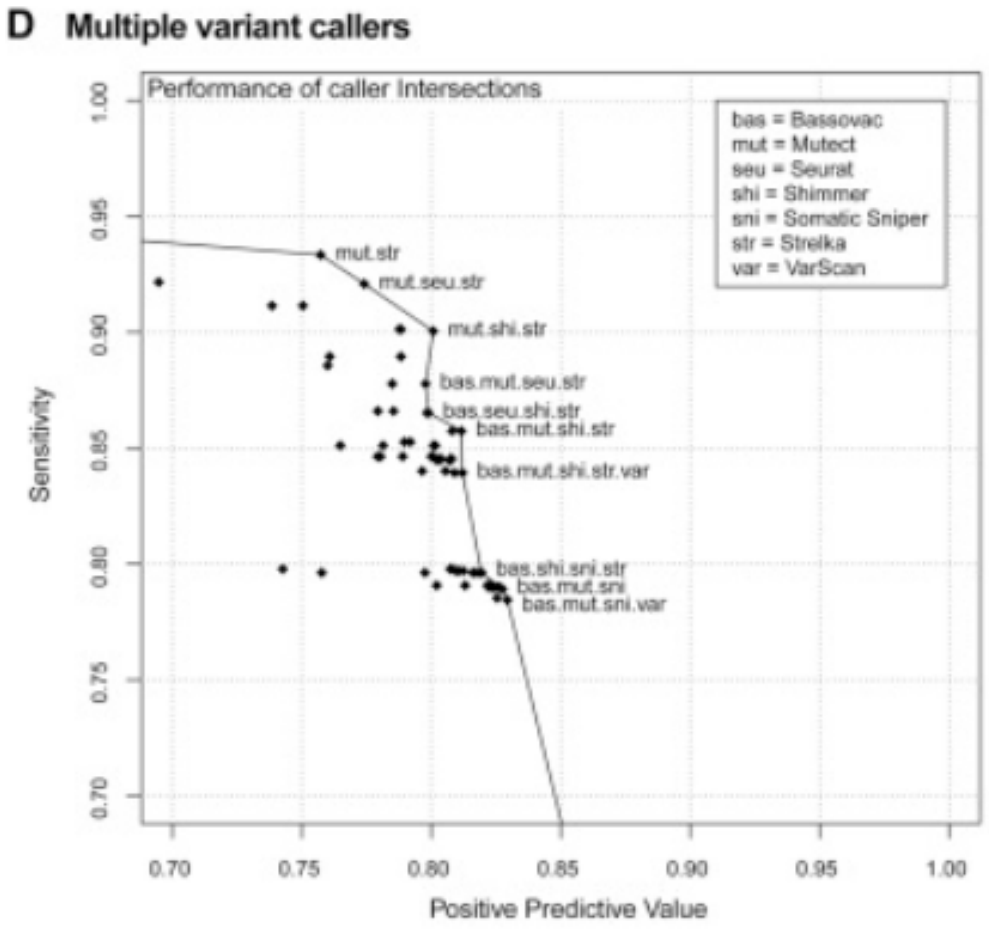
- Install tools
- Put tools together
- Test and validate
- Improve algorithms
- Scale
- Read literature
- Do biology

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



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

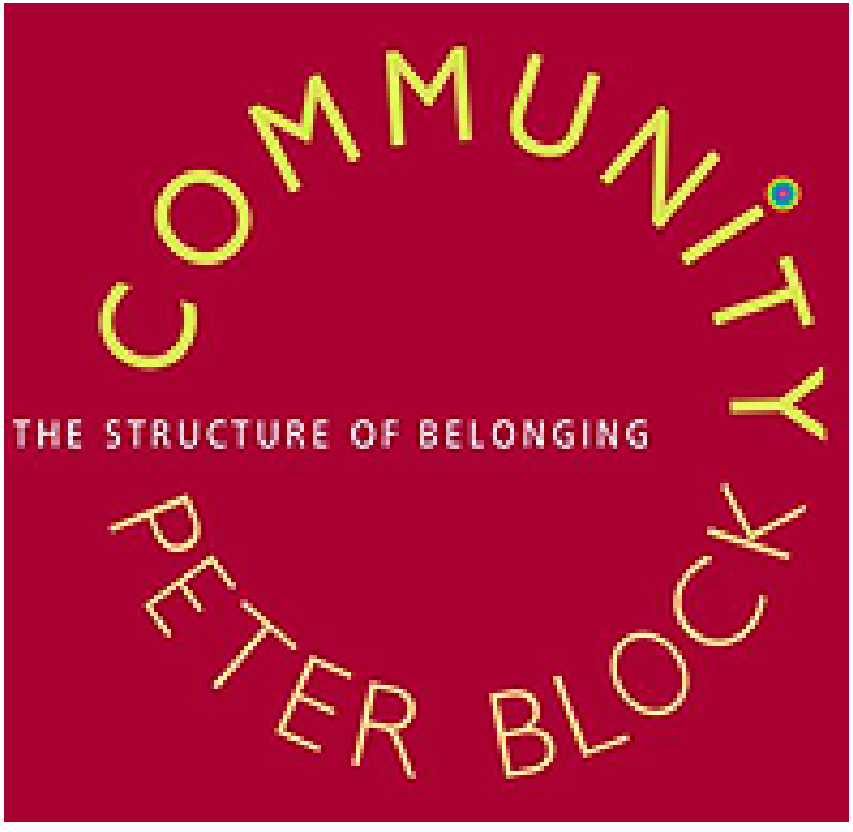
Many thanks to the Genomics Platform both for generating much of the exome data displayed here and for providing the computing resources required for this analysis.

Funding

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

http://exac.broadinstitute.org/about

Solution



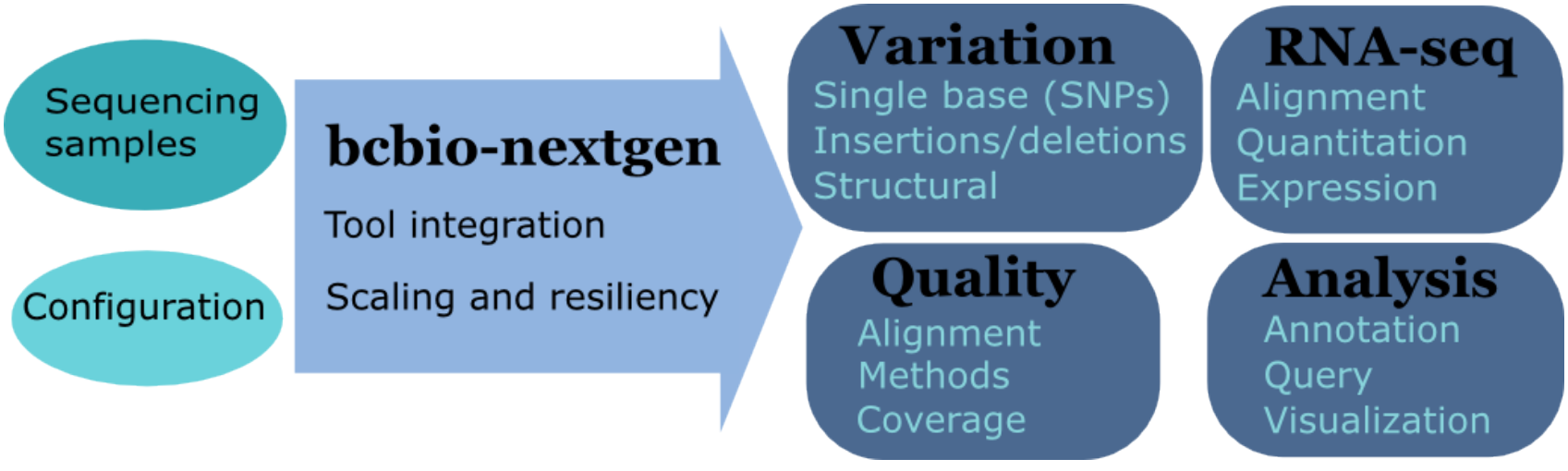
http://www.amazon.com/Community-Structure-Belonging-Peter-Block/dp/1605092770

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

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, RNA-SeQC
- Manipulation: bedtools, bcftools, biobambam, sambamba, samblaster, samtools, vcflib, vt

Provides

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

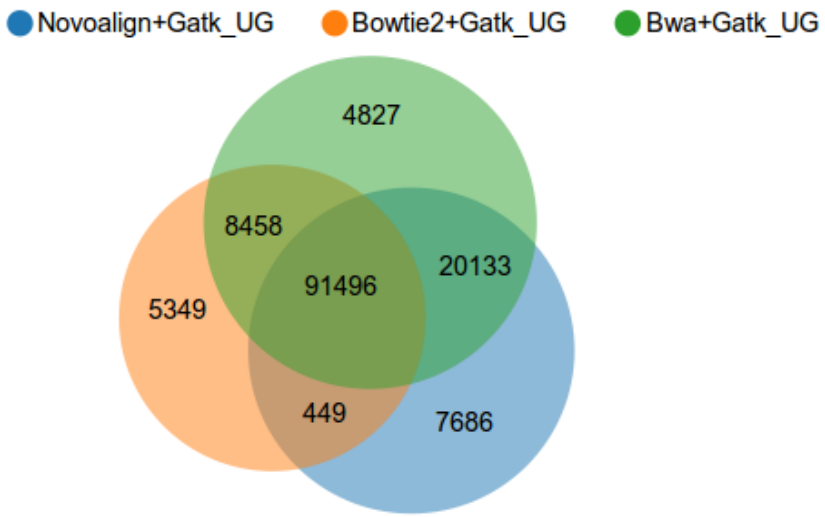
Quality differences between methods

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.

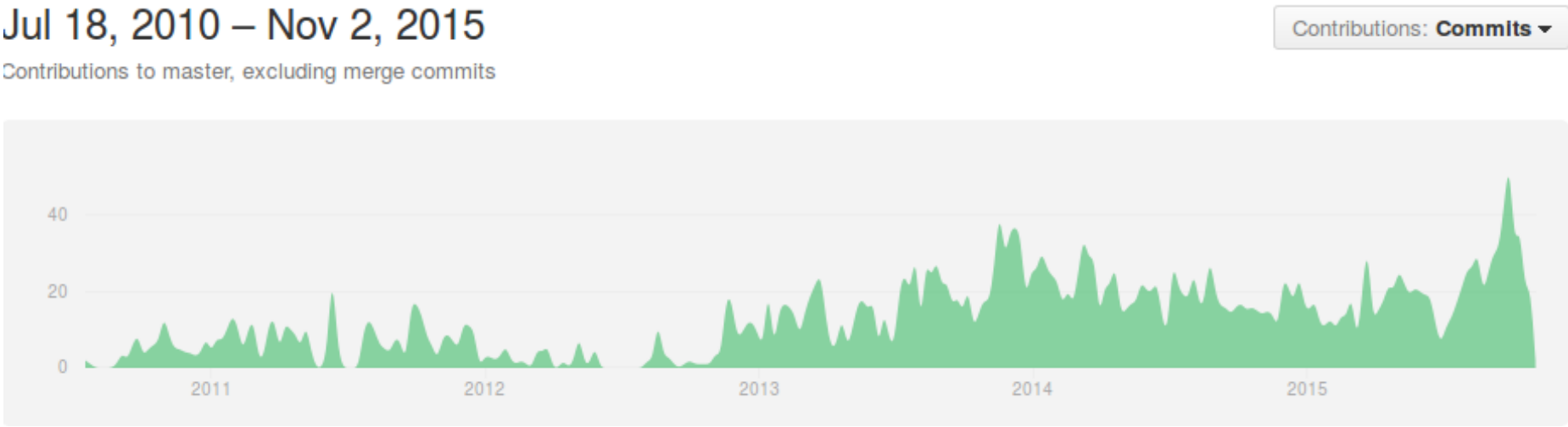
RETIRED: Best Practice Variant Detection with the GATK v3

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

Mark_DePristo Posts: 153
July 2012 edited February 4

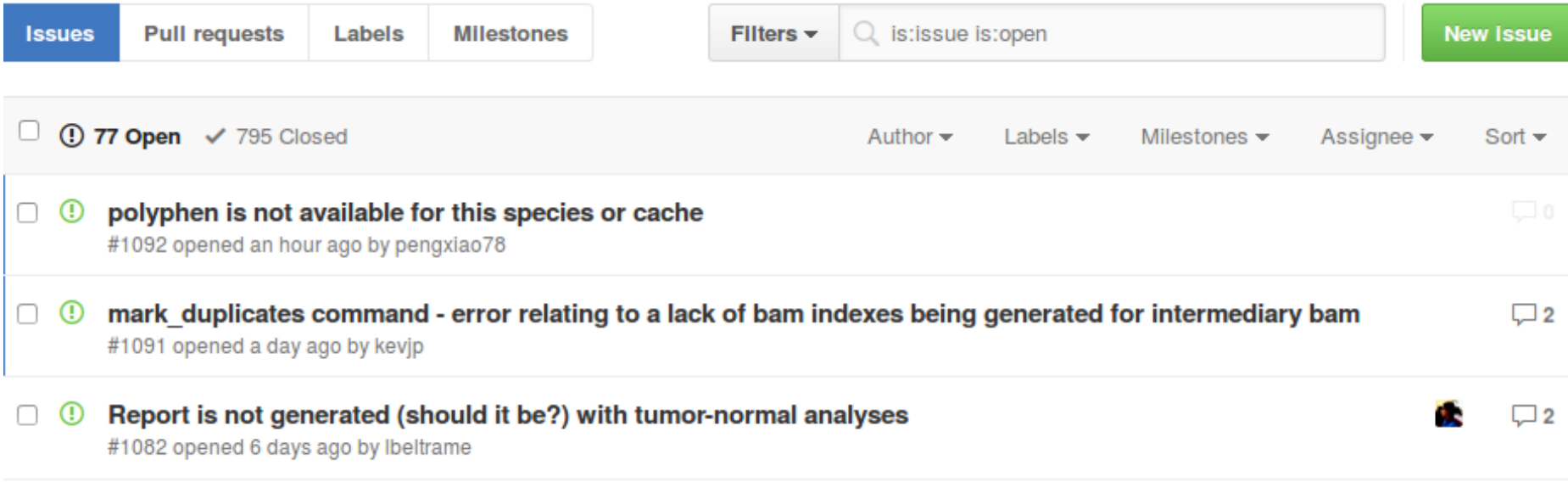
The Best Practices have been updated for GATK version 3. If you are running an older version, you should seriously consider upgrading. For more details

Community: sustainability



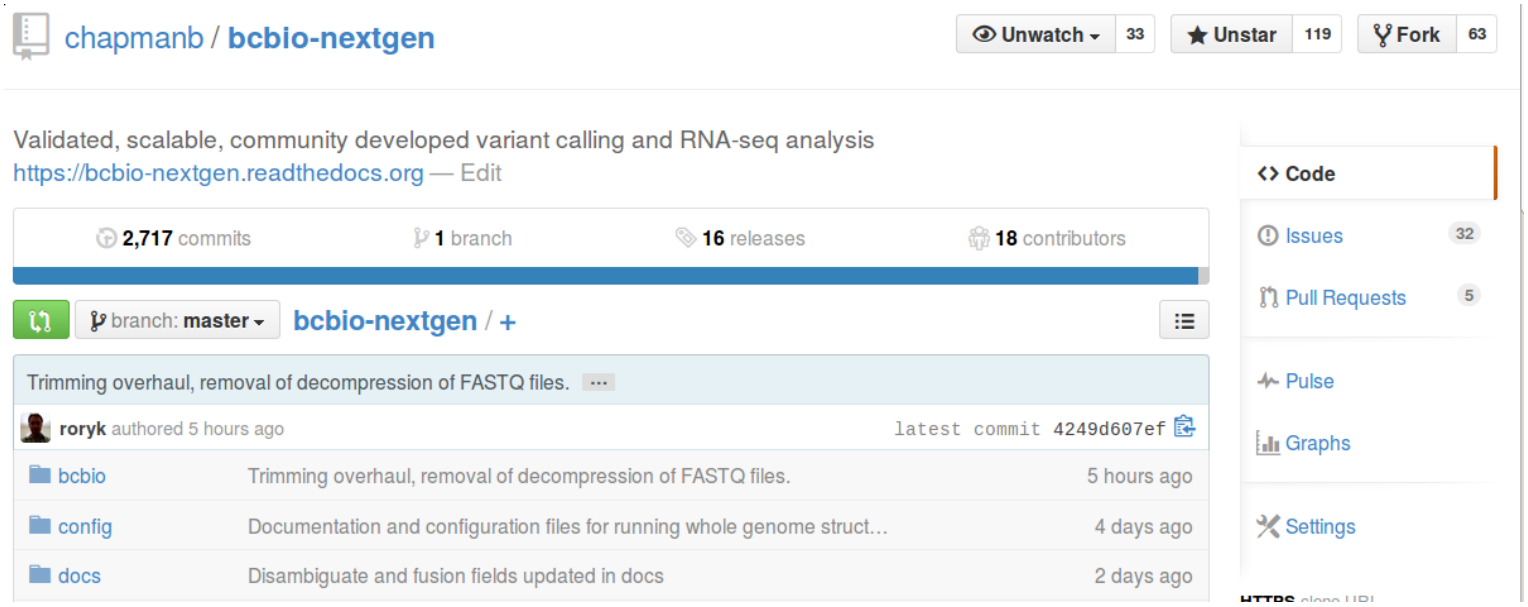
https://github.com/chapmanb/bcbio-nextgen

Community: support



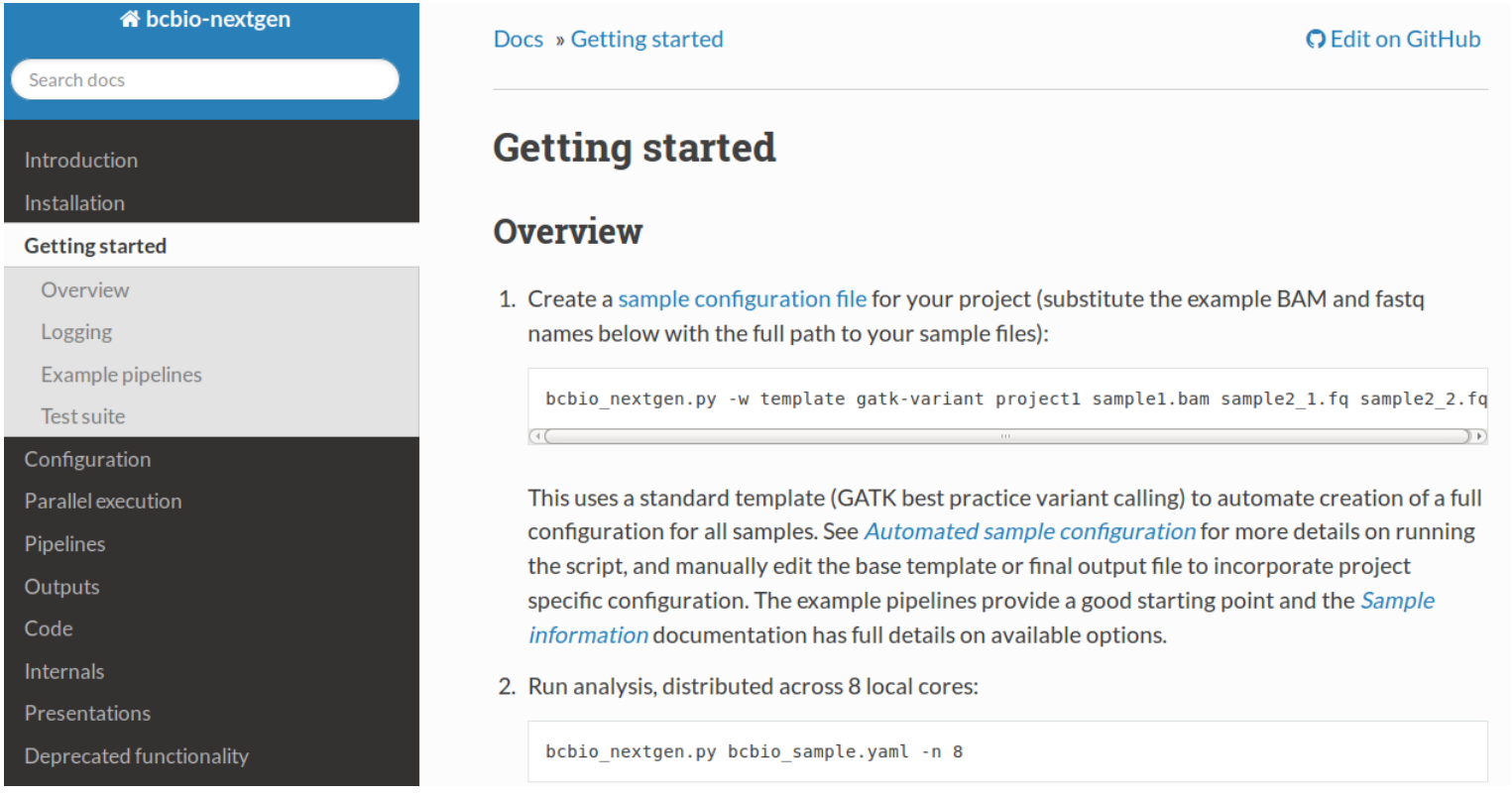
https://bcbio-nextgen.readthedocs.org

Community: contribution



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

Community: documentation

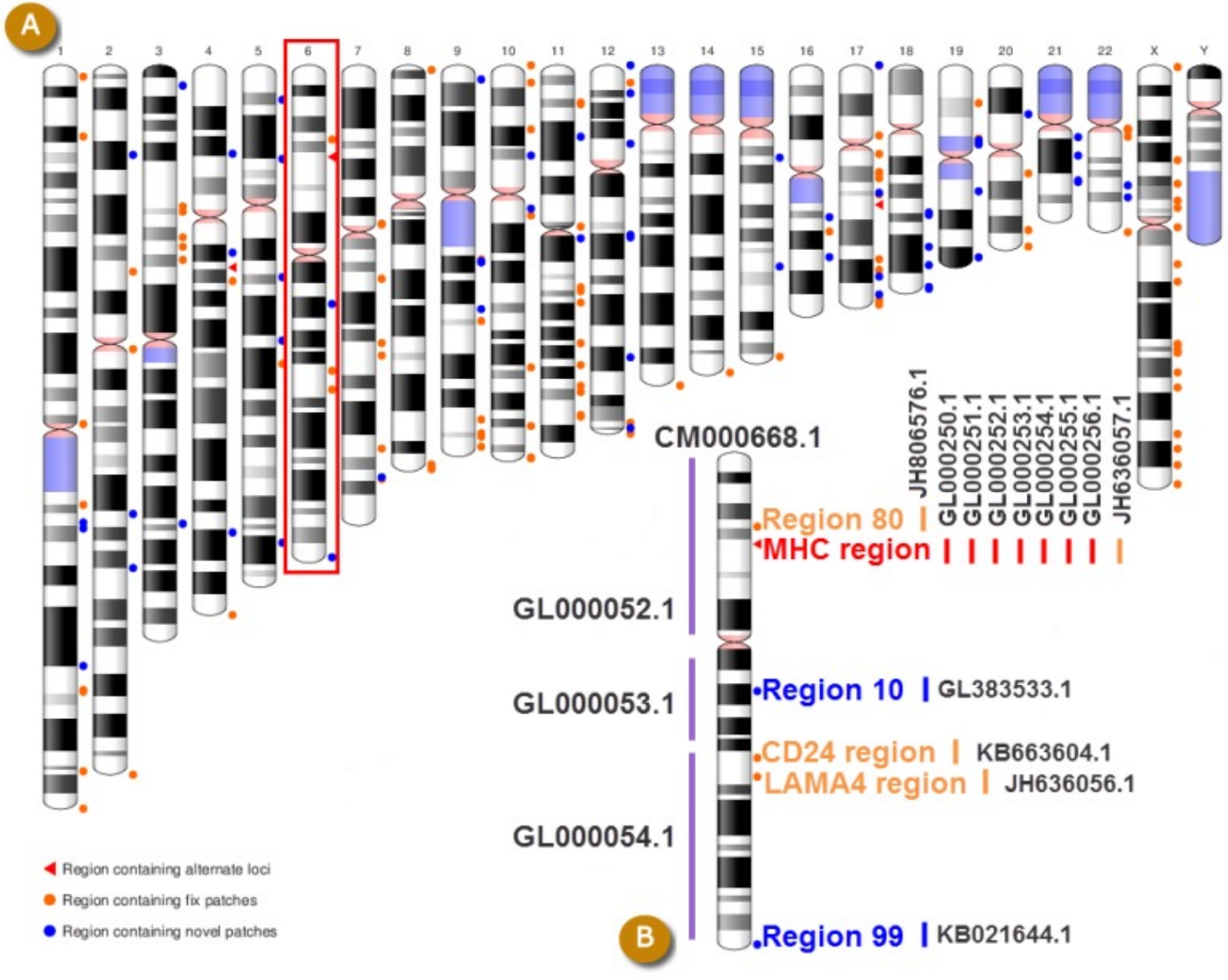


https://bcbio-nextgen.readthedocs.org

Outline: Science

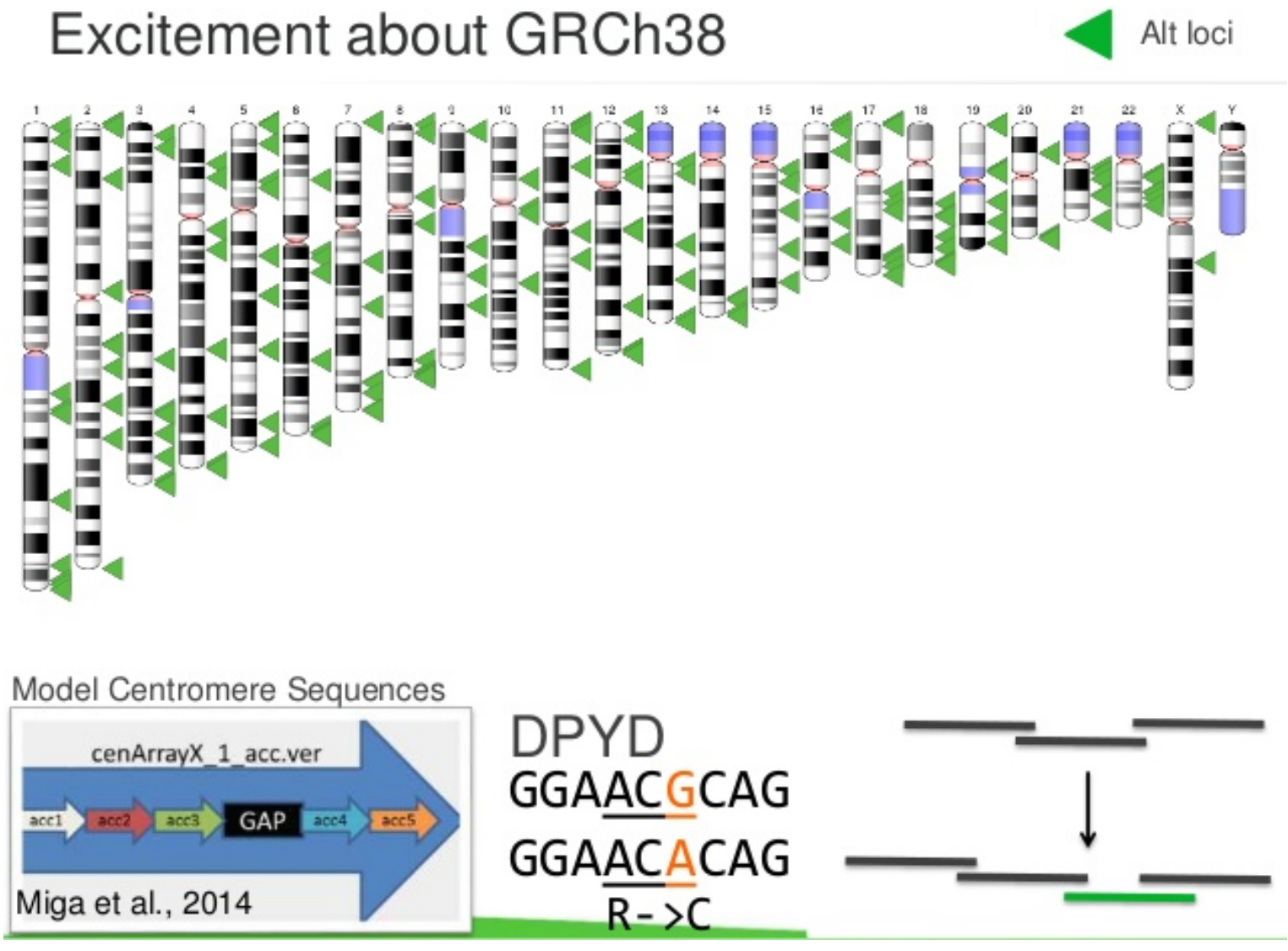
- **Human build 38**
- Low frequency somatic calling
- Structural variation

Currently: GRCh37/hg19



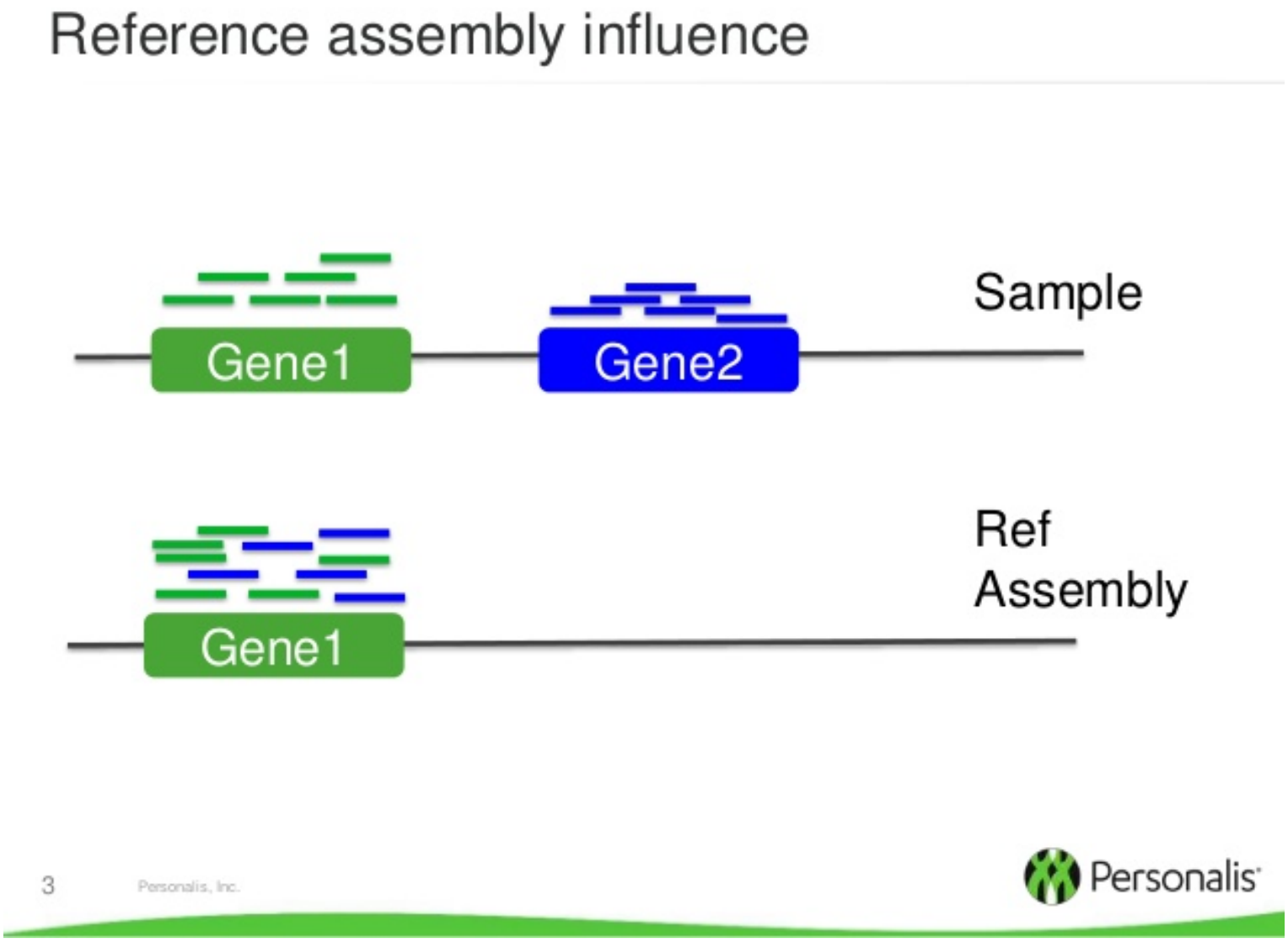
http://www.ncbi.nlm.nih.gov/books/NBK153600/?report=reader

GRCh38 – graph based, many more alternative loci



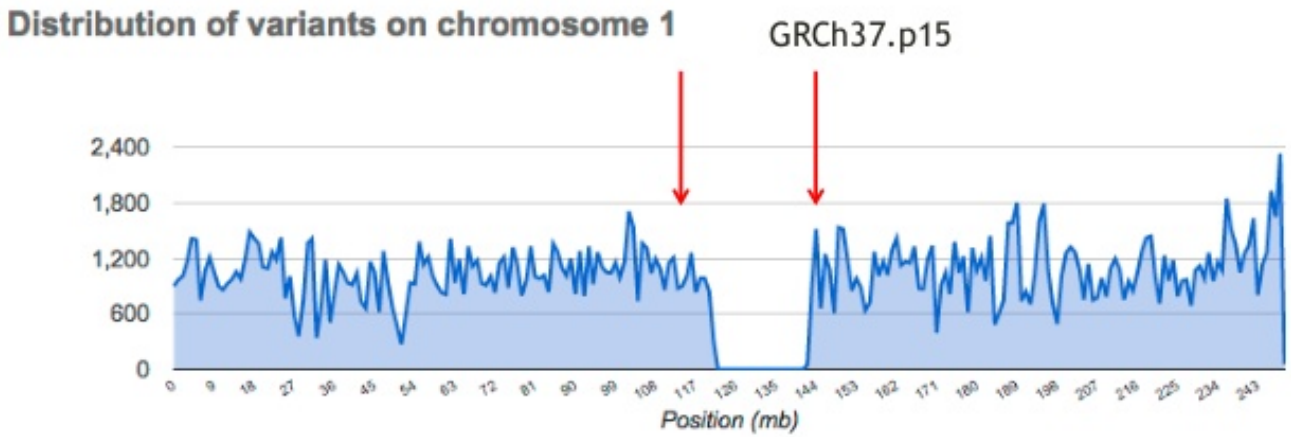
http://www.slideshare.net/GenomeRef/transitioning-to-grch38

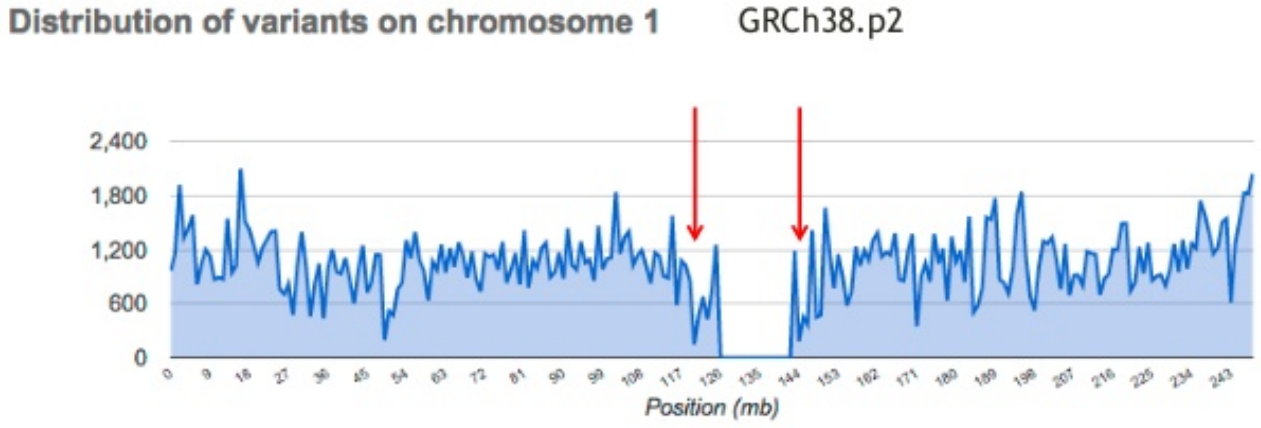
GRCh38 – advantage for variant calling



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

Avoiding collapsed repeats





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/

Reference materials

Genome in a Bottle Consortium

Global Alliance for Genomics & Health

ICGC-TCGA DREAM Mutation Calling challenge

http://www.genomeinabottle.org/

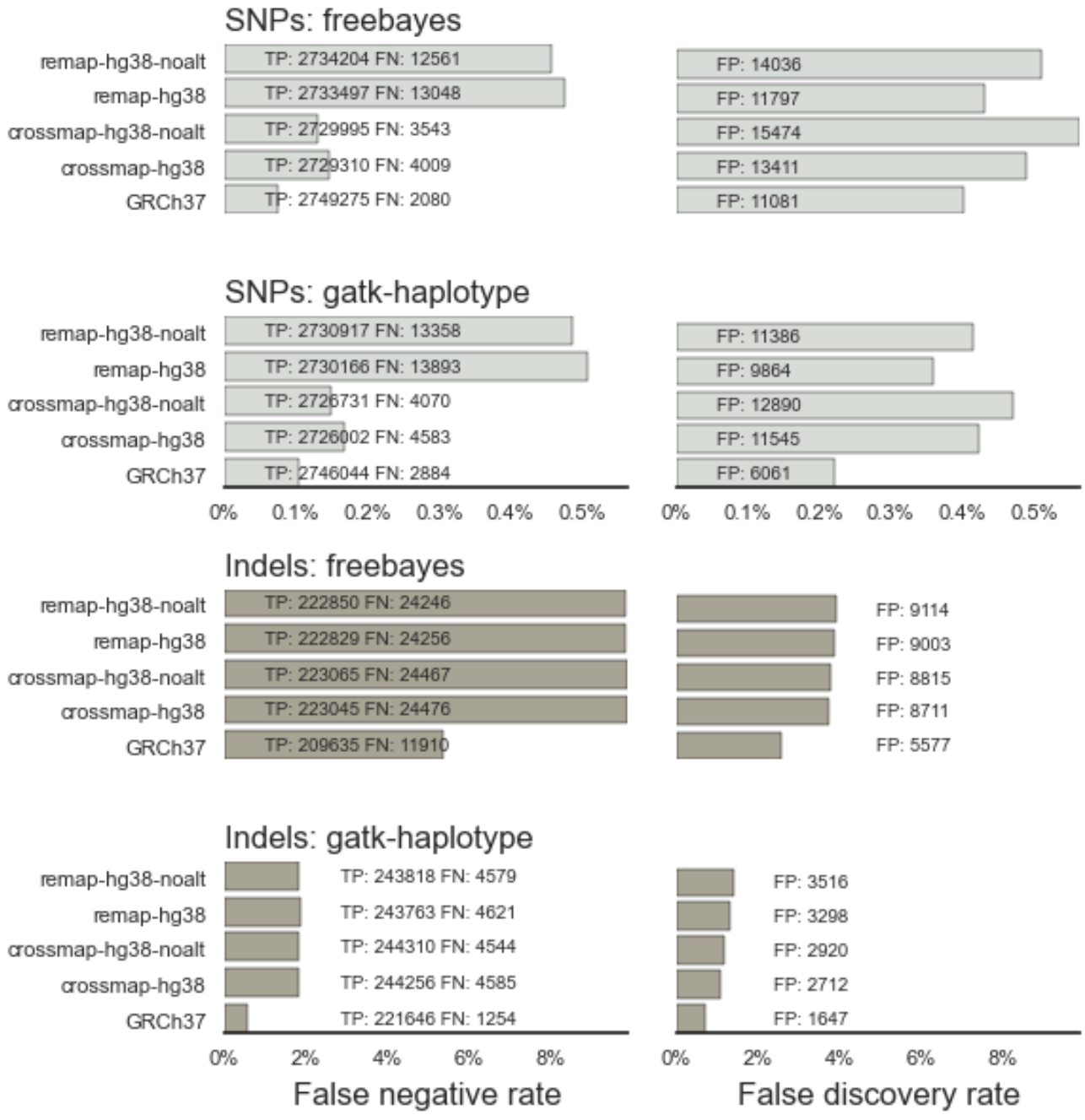
http://ga4gh.org/#/benchmarking-team

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

**

hg19/hg38 comparison: NA12878 Platinum Genomes

SNPs: freebayes


GRCh37/hg38 comparison: NA12878 Genome in a Bottle
SNPs: freebayes
remap-hg38-noalt
remap-hg38
crossmap-hg38-noalt
crossmap-hg38
GRCh37
TP: 2734204 FN: 12561
TP: 2733497 FN: 13048
TP: 2729995 FN: 3543
TP: 2729310 FN: 4009
TP: 2749275 FN: 2080
FP: 14036
FP: 11797
FP: 15474
FP: 13411
FP: 11081
SNPs: gatk-haplotype
TP: 2730917 FN: 13358
TP: 2730166 FN: 13893
TP: 2726731 FN: 4070
TP: 2726002 FN: 4583
TP: 2746044 FN: 2884
FP: 11386
FP: 9864
FP: 12890
FP: 11545
FP: 6061
0%
0.1%
0.2%
0.3%
0.4%
0.5%
Indels: freebayes
TP: 222850 FN: 24246
TP: 222829 FN: 24256
TP: 223065 FN: 24467
TP: 223045 FN: 24476
TP: 209635 FN: 11910
FP: 9114
FP: 9003
FP: 8815
FP: 8711
FP: 5577
Indels: gatk-haplotype
TP: 243818 FN: 4579
TP: 243763 FN: 4621
TP: 244310 FN: 4544
TP: 244256 FN: 4585
TP: 221646 FN: 1254
FP: 3516
FP: 3298
FP: 2920
FP: 2712
FP: 1647
0%
2%
4%
6%
8%
False negative rate
False discovery rate

Small variant results

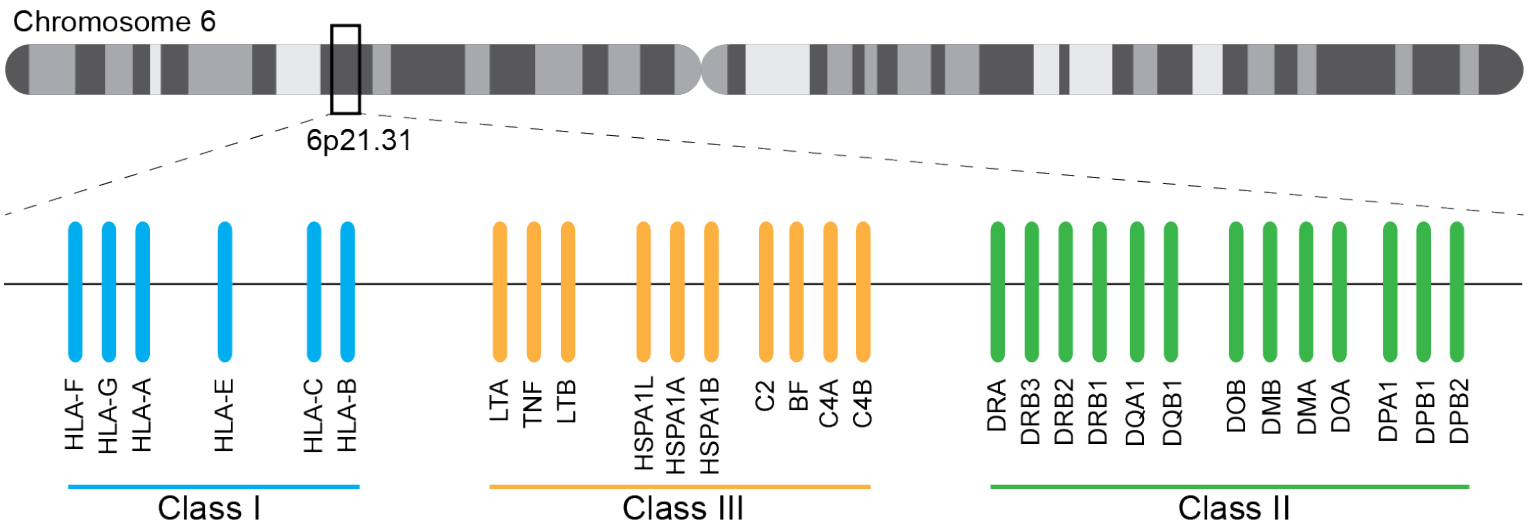
- SNPs: build 38 more sensitive
- SNPs: build 38 reduces false positives
- Indels: build 38 detected more
- Indels: work on sensitivity and precision

Remapping results

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 needs 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
```

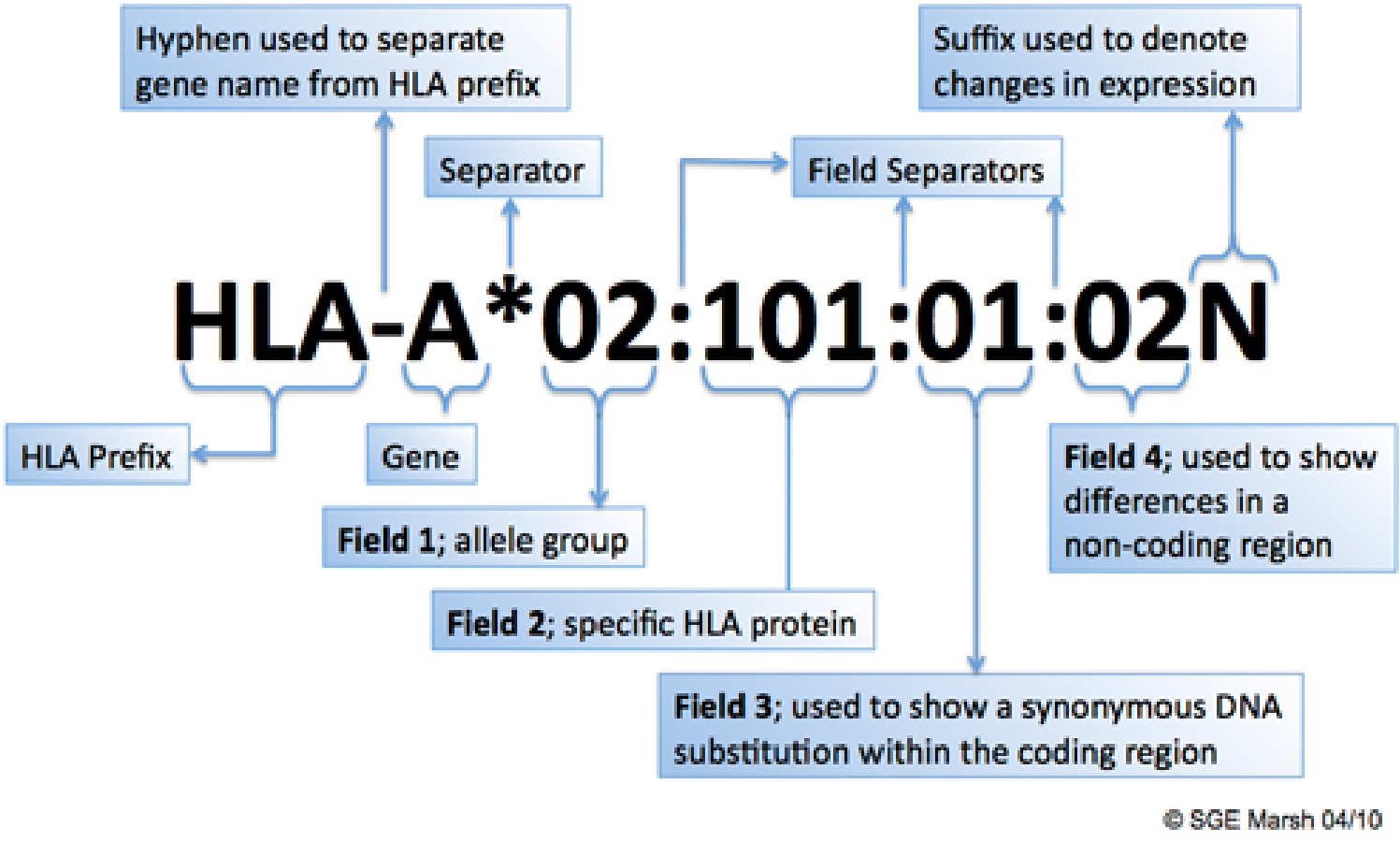
https://github.com/lh3/bwa/blob/master/README-alt.md

HLA typing

- bwakit implementation
- 1000 genomes: build 38 + IMGT/HLA-3.18.0
- bwa extracts HLA reads
- fermi de novo assembly
- Remap assemblies back to HLA choices
- Call HLA types

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

HLA nomenclature



https://www.ebi.ac.uk/ipd/imgt/hla/

http://hla.alleles.org/alleles/p_groups.html

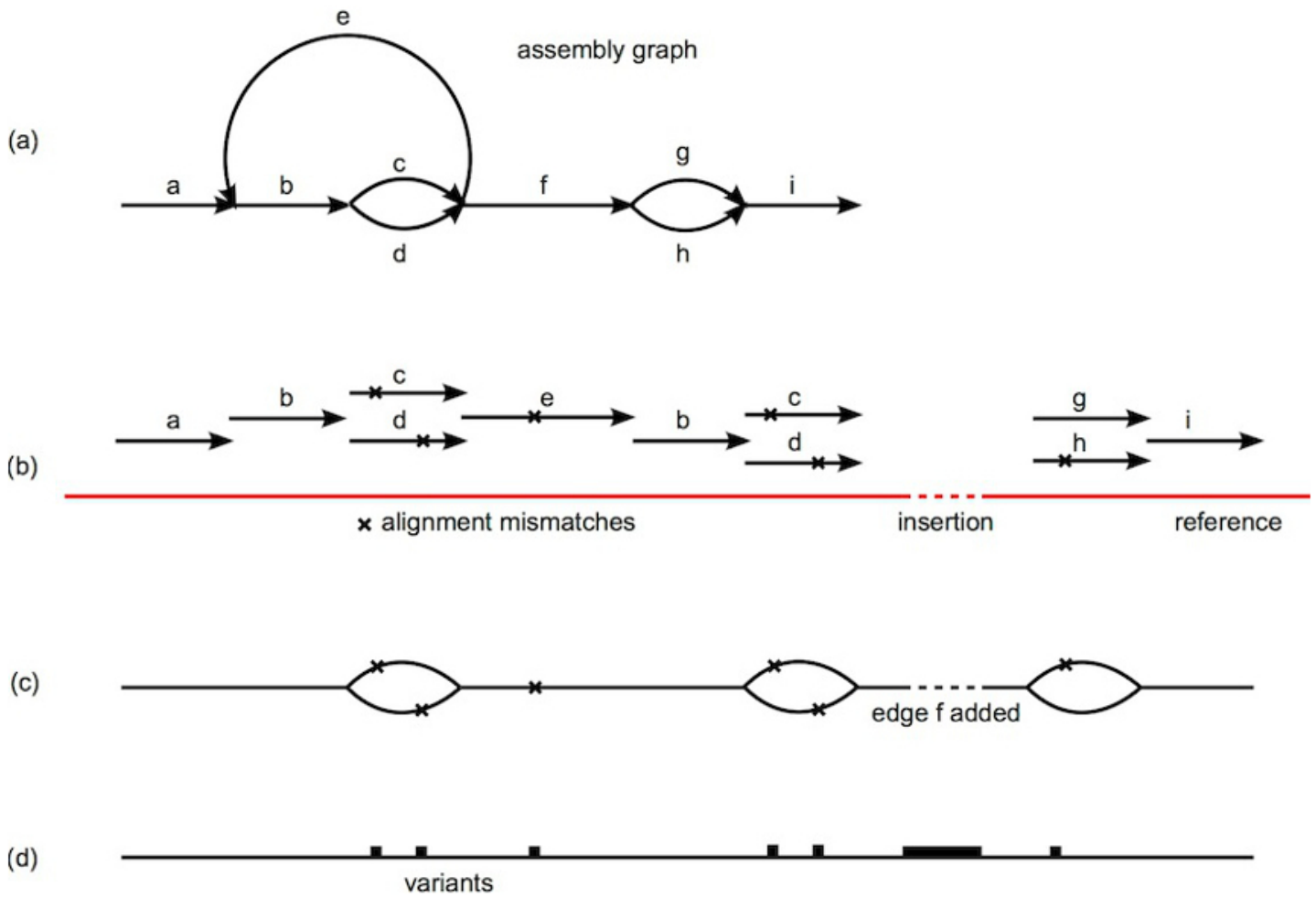
Validations

- Omixon example data
- bwakit calls on exome and deep targeted data
- P-group resolution
- Good results for exome
- Assembly problems with deep targeted

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

https://gist.github.com/chapmanb/8e2a18c7bbbee3167395

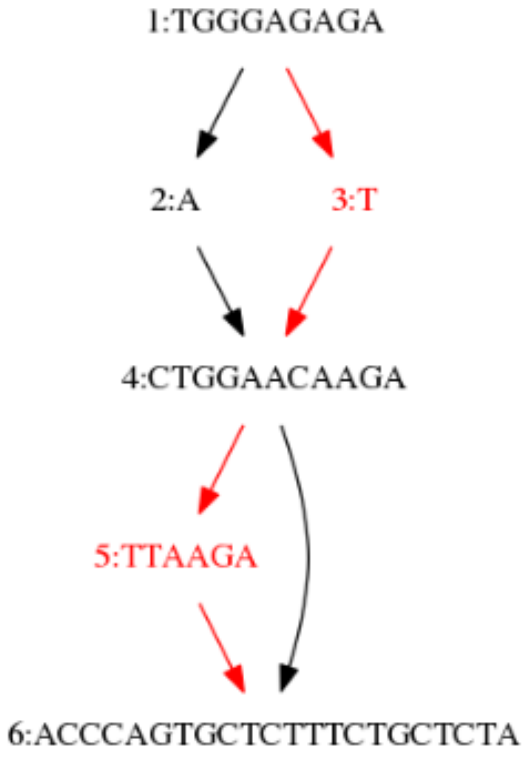
Genome graphs and variation



http://www.nature.com/ng/journal/v46/n12/fig_tab/ng.3121_SF6.html

vg – tools for working with variant graphs

POS	ID	REF	ALT
10	.	A	T
21	.	A	ATTAAGA
...			

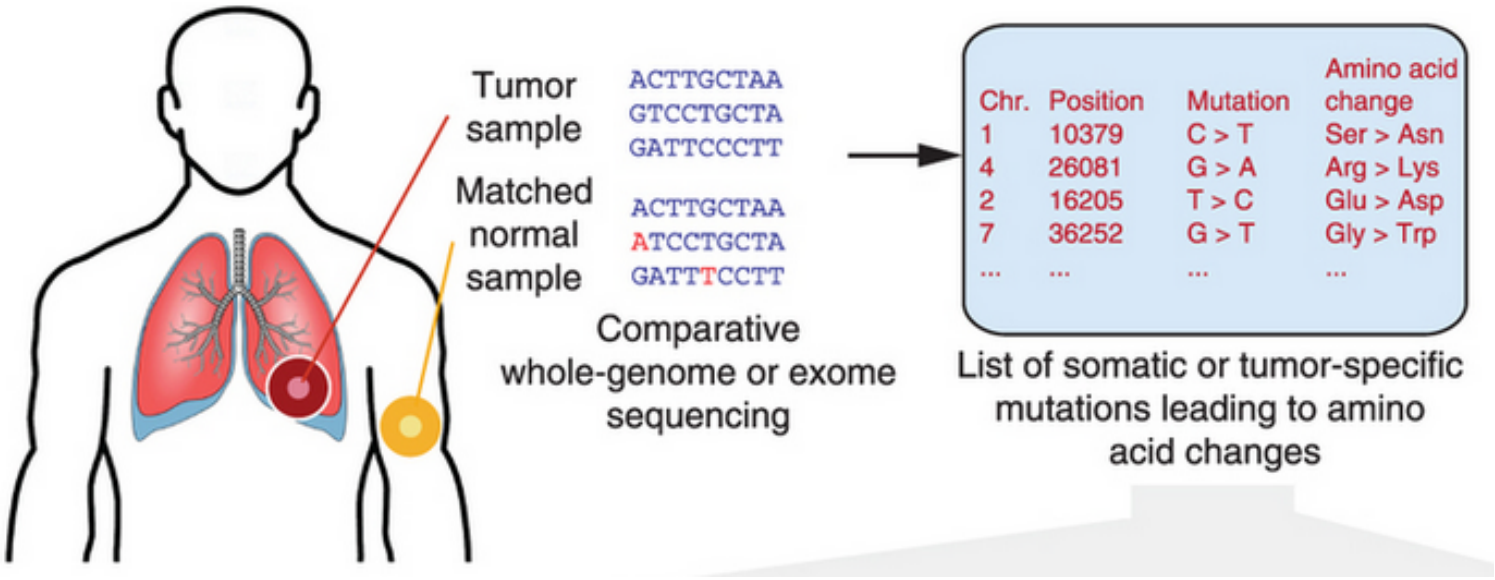


https://docs.google.com/presentation/d/1bbl2zY4qWQ0yYBHhoVuXb79HdgajRotIUa_VEn3kTpI

Outline: science

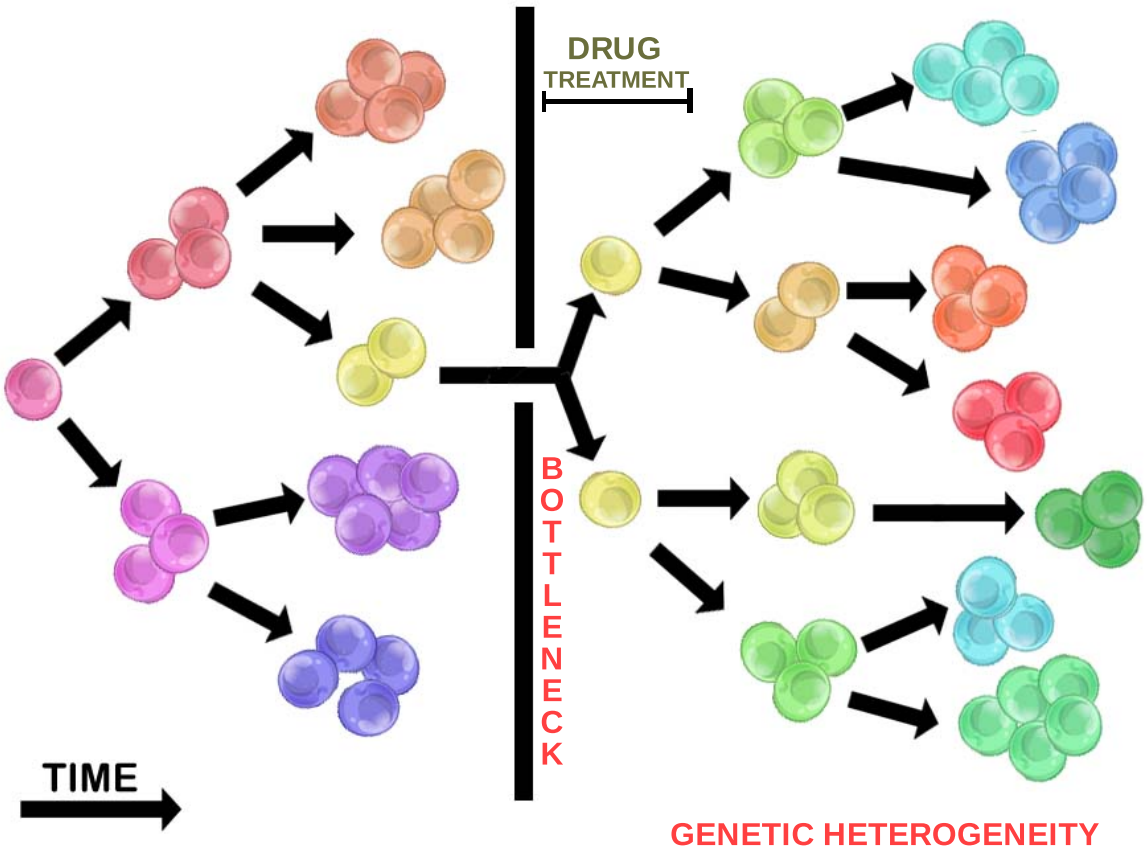
- Human build 38
- **Low frequency somatic calling**
- Structural variation

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

VarDict

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

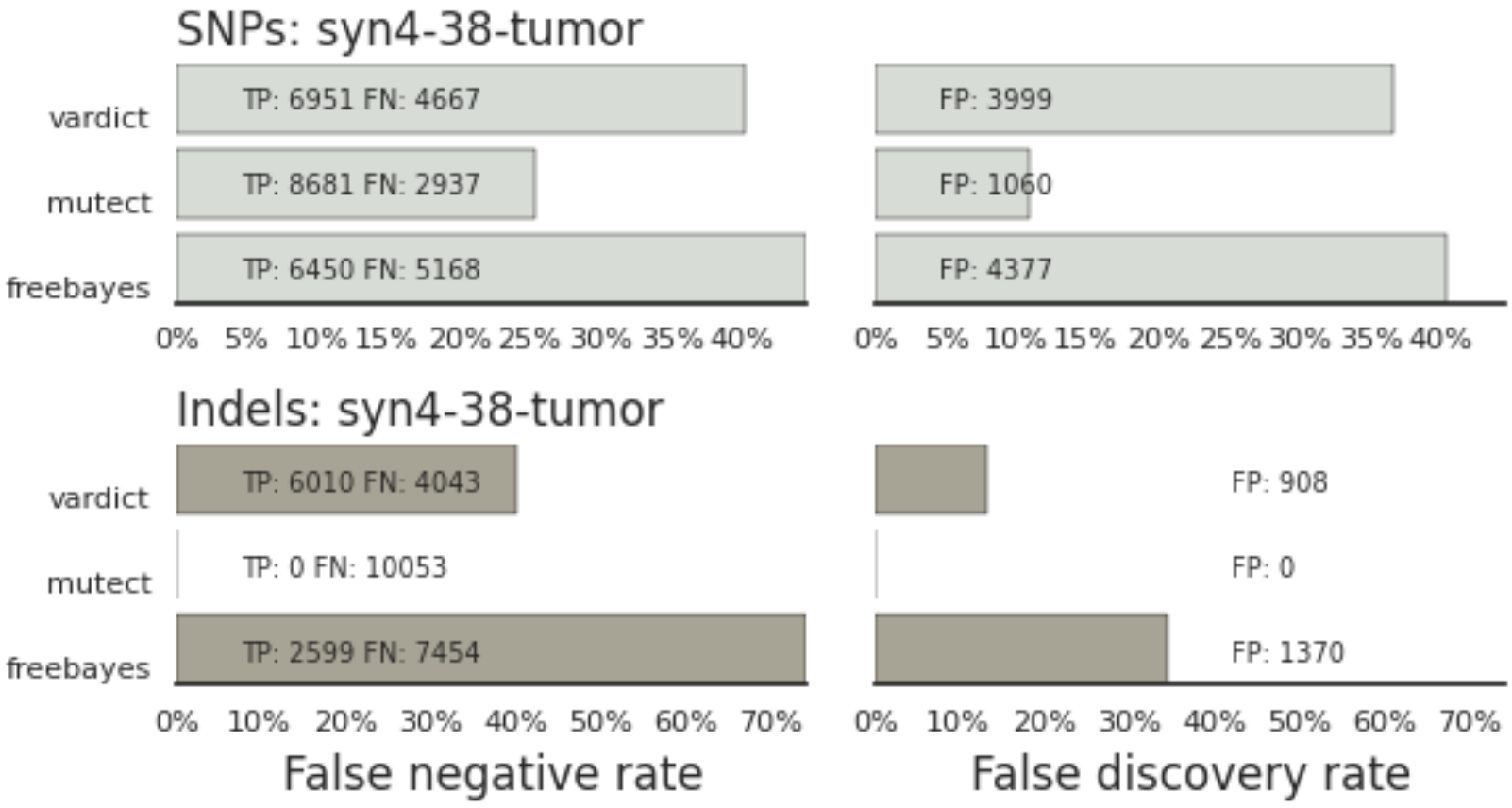
https://github.com/AstraZeneca-NGS/VarDictJava

DREAM synthetic dataset 4

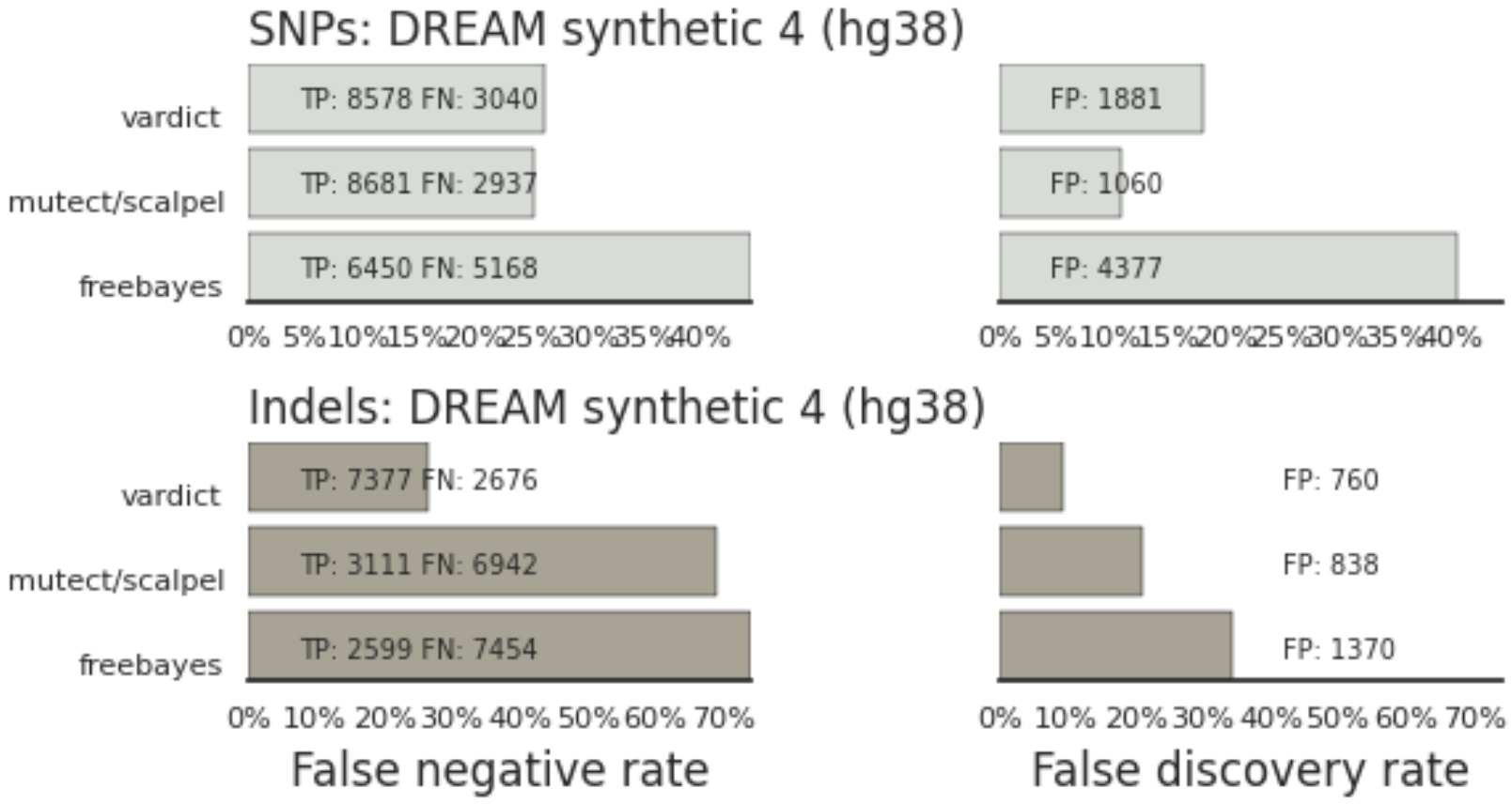
in silico 3	*in silico* 4
BWA Backtrack	BWA MEM
SNV, SV (deletions, duplications, insertions, inversions) & INDEL	SNV, SV (deletions, duplications, inversions) & INDEL
100%	80%
50%, 33%, 20%	50%, 35% (effectively 30% and 15% due to cellularity)
Female	Male
HCC1143 BL from TCGA Benchmark 4	CPCG0102R (Provided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

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

VarDict sensivitity/precision before



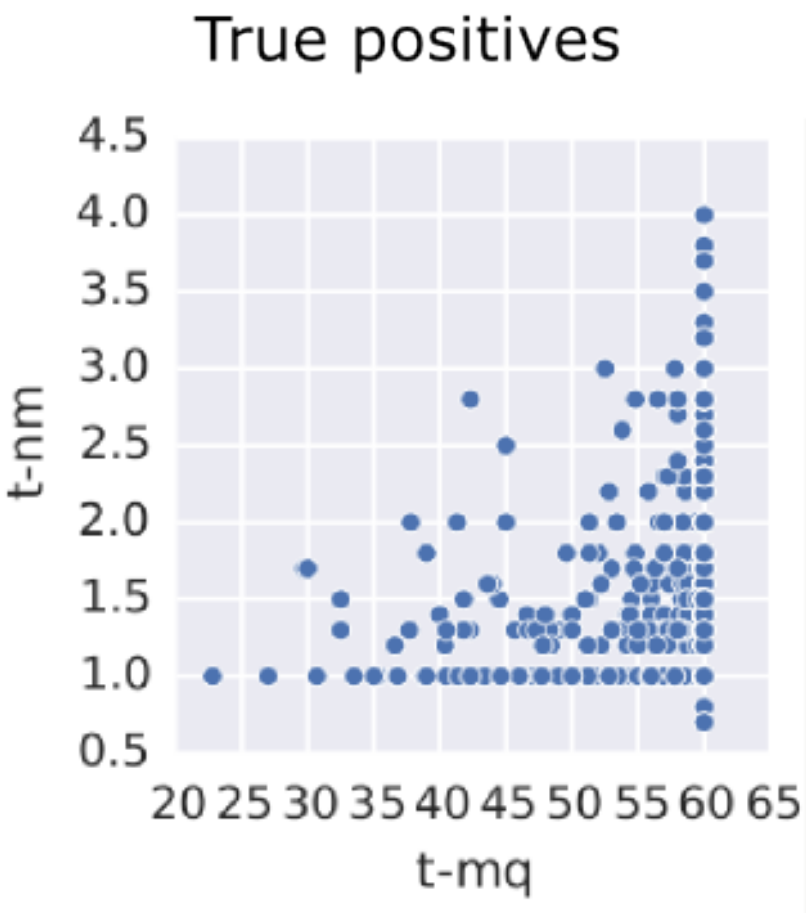
VarDict sensivitity/precision after

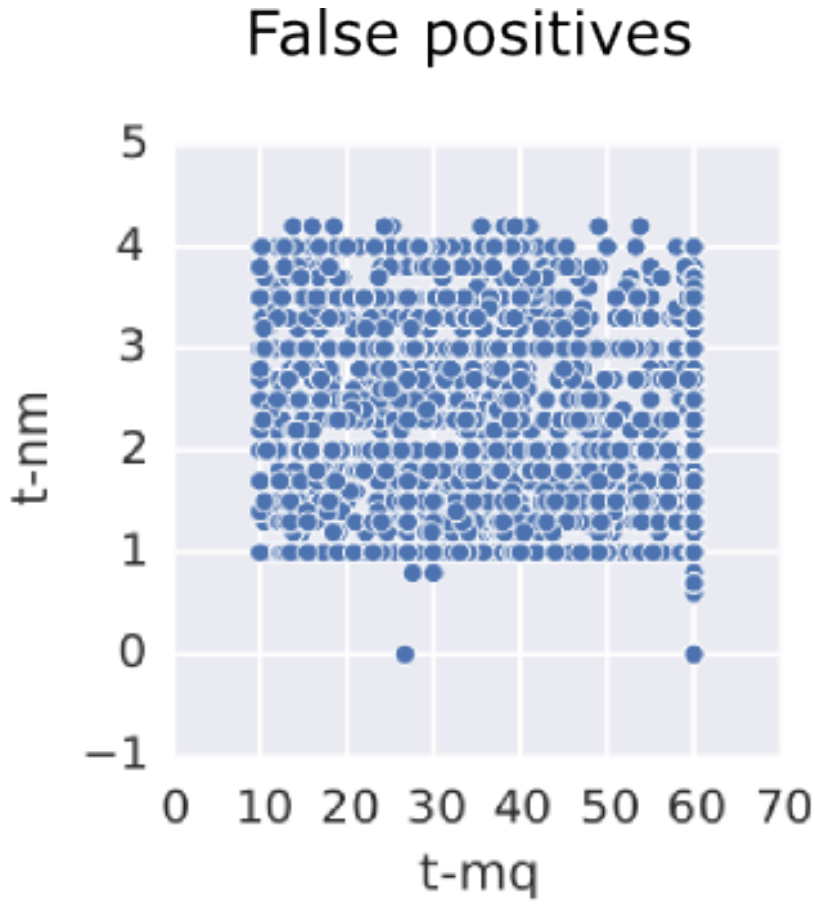


How? Filter summary

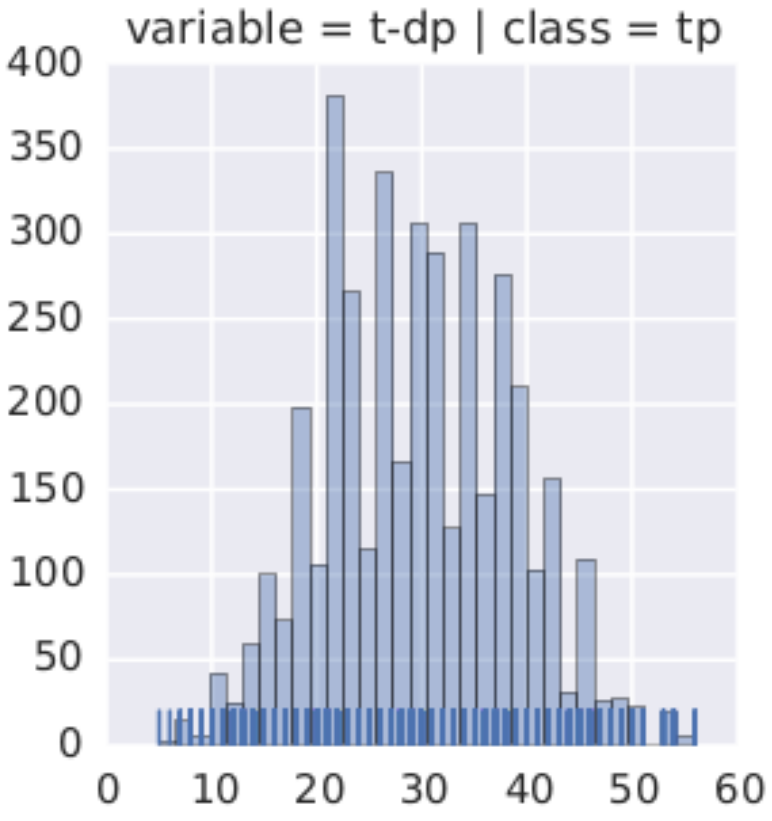
```
((AF * DP < 6) &&
 ((MQ < 55.0 && NM > 1.0) ||
  (MQ < 60.0 && NM > 2.0) ||
  (DP < 10) ||
  (QUAL < 45)))
```

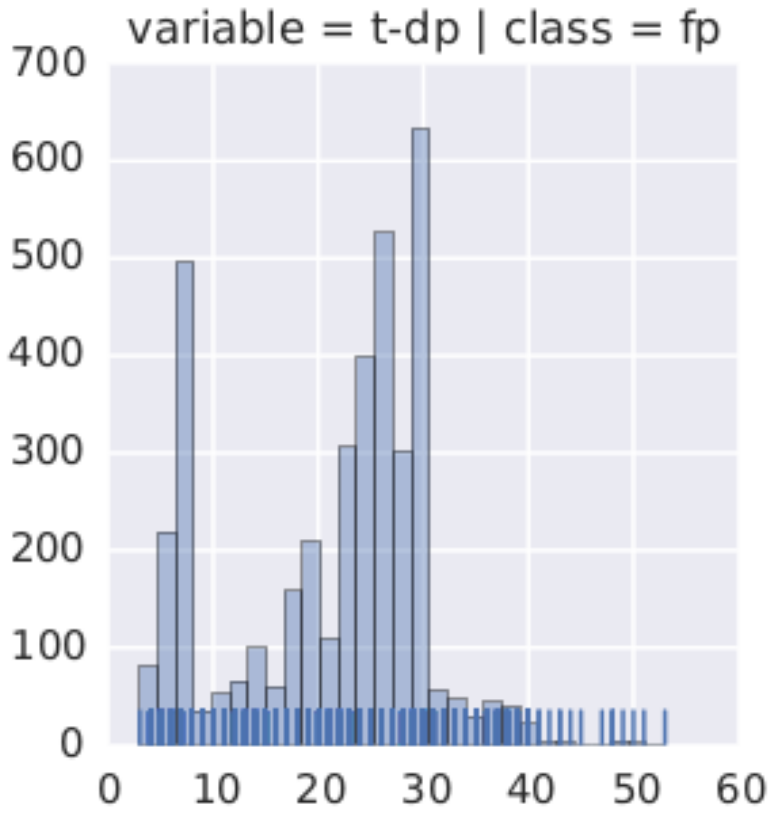
Filter: mapping quality and number of mismatches



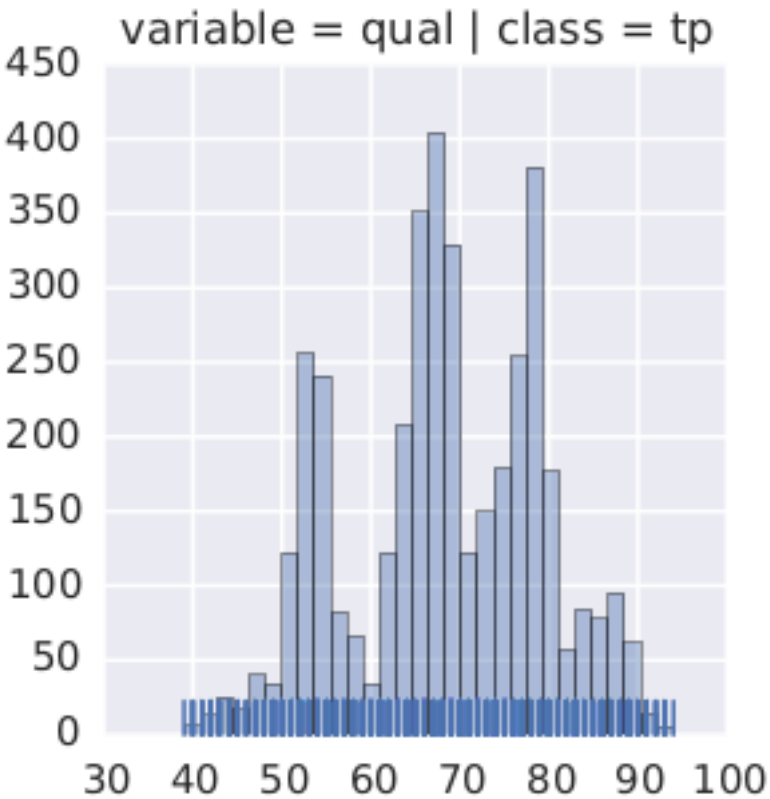


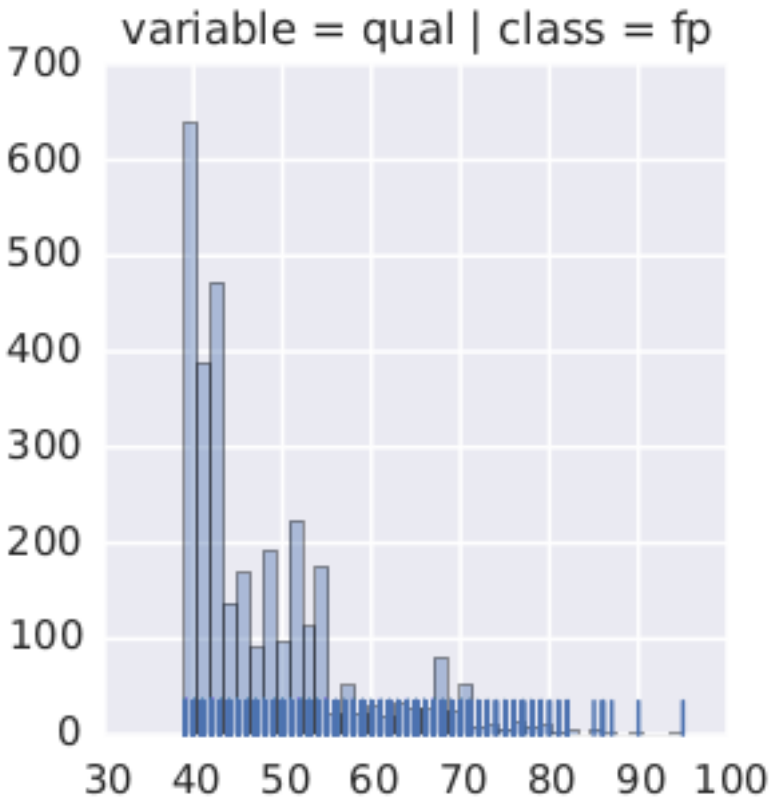
Filter: low depth





Filter: low quality





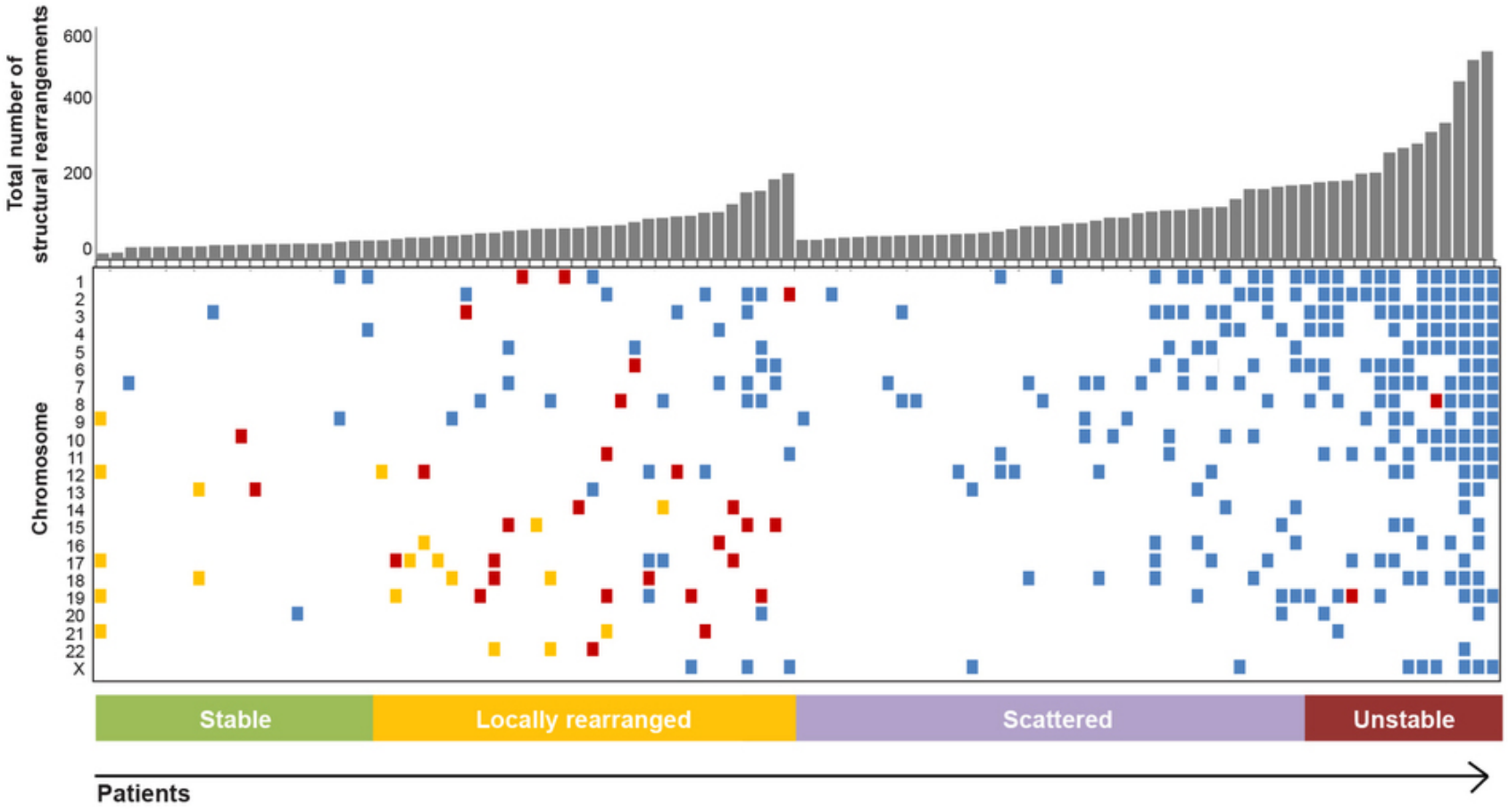
How can we improve?

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

Outline: science

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

Structural variants critical in cancer

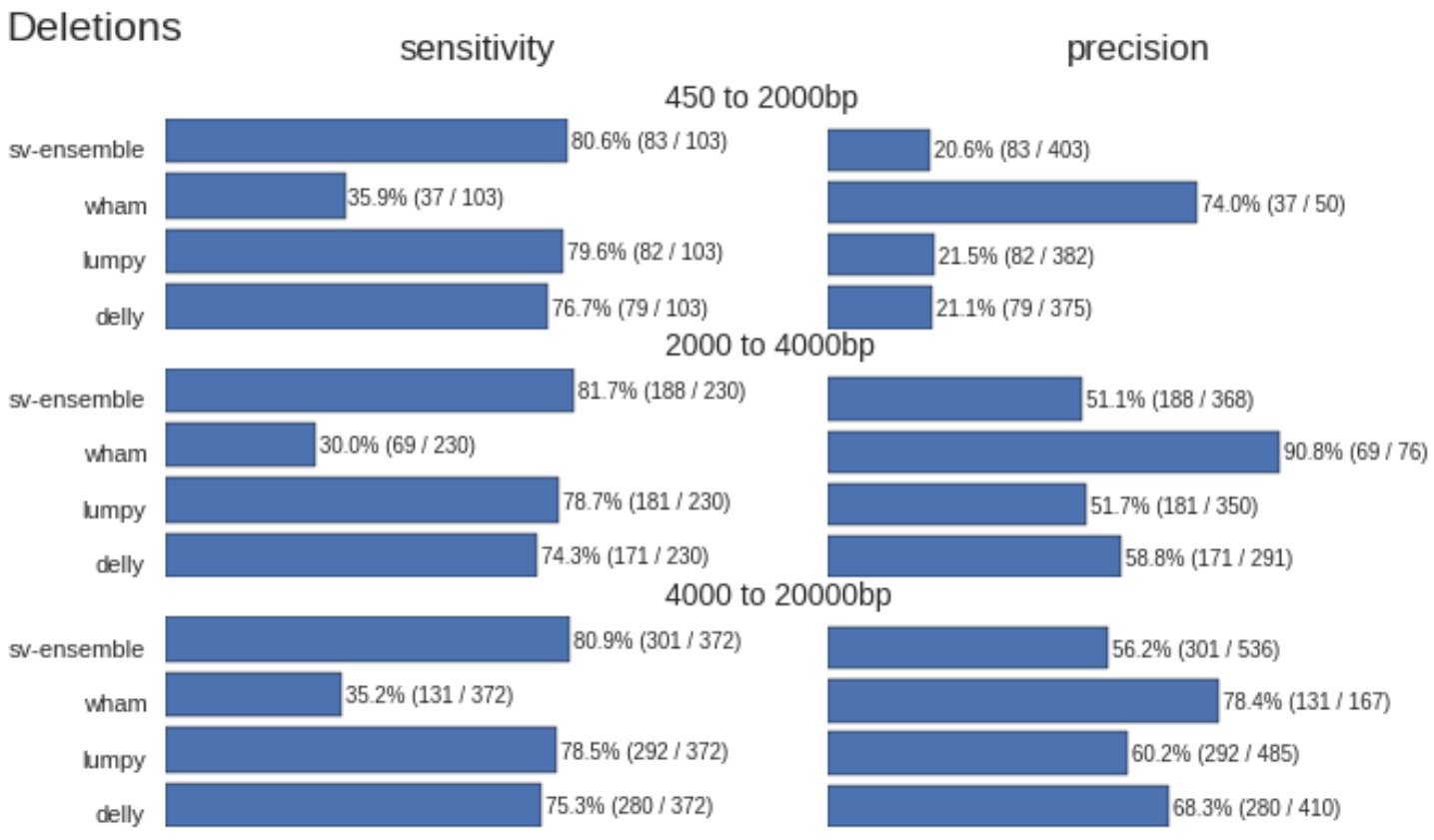


http://www.nature.com/nature/journal/v518/n7540/full/nature14169.html

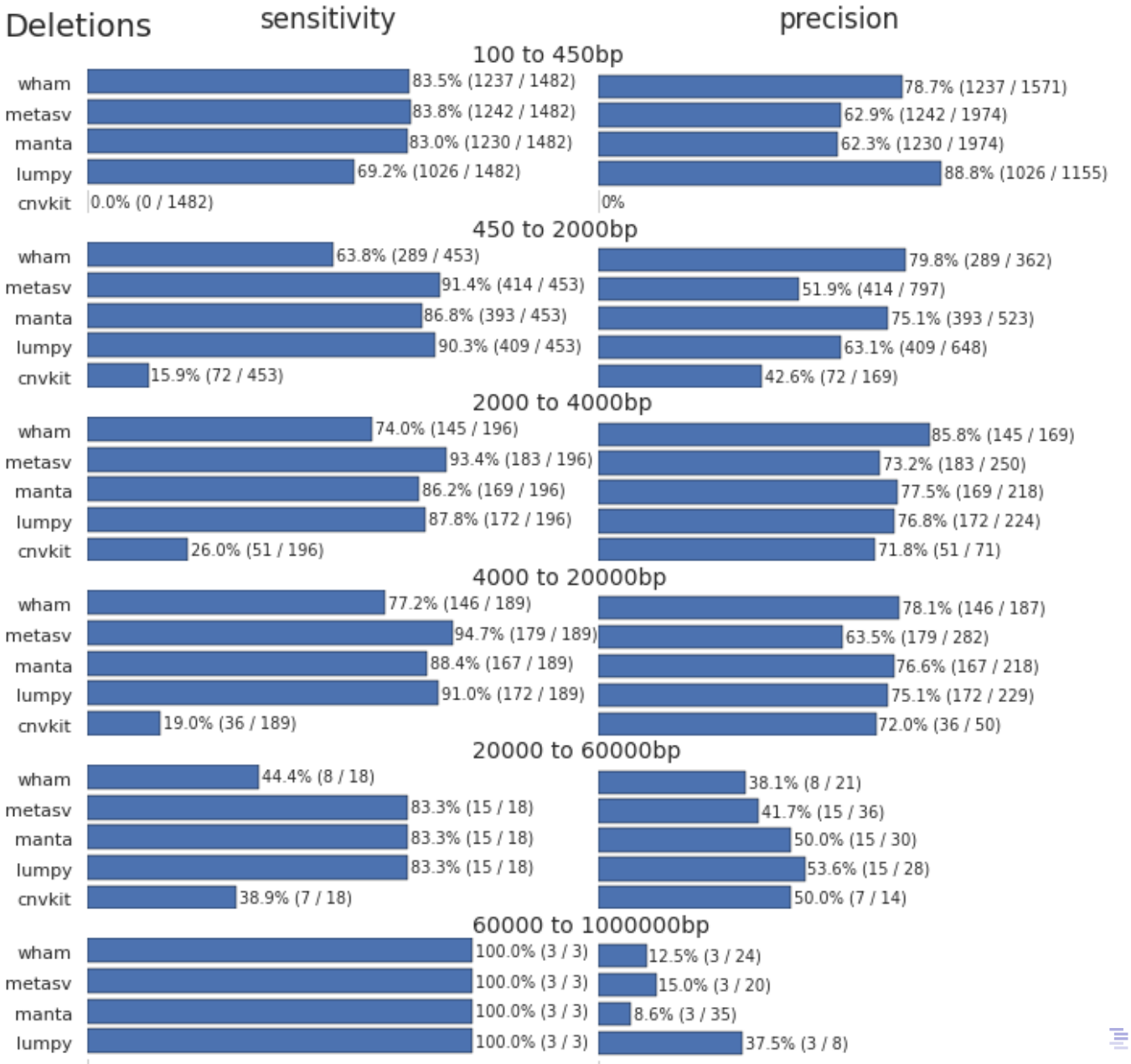
Improvements in speed, sensitivity and precision

- 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

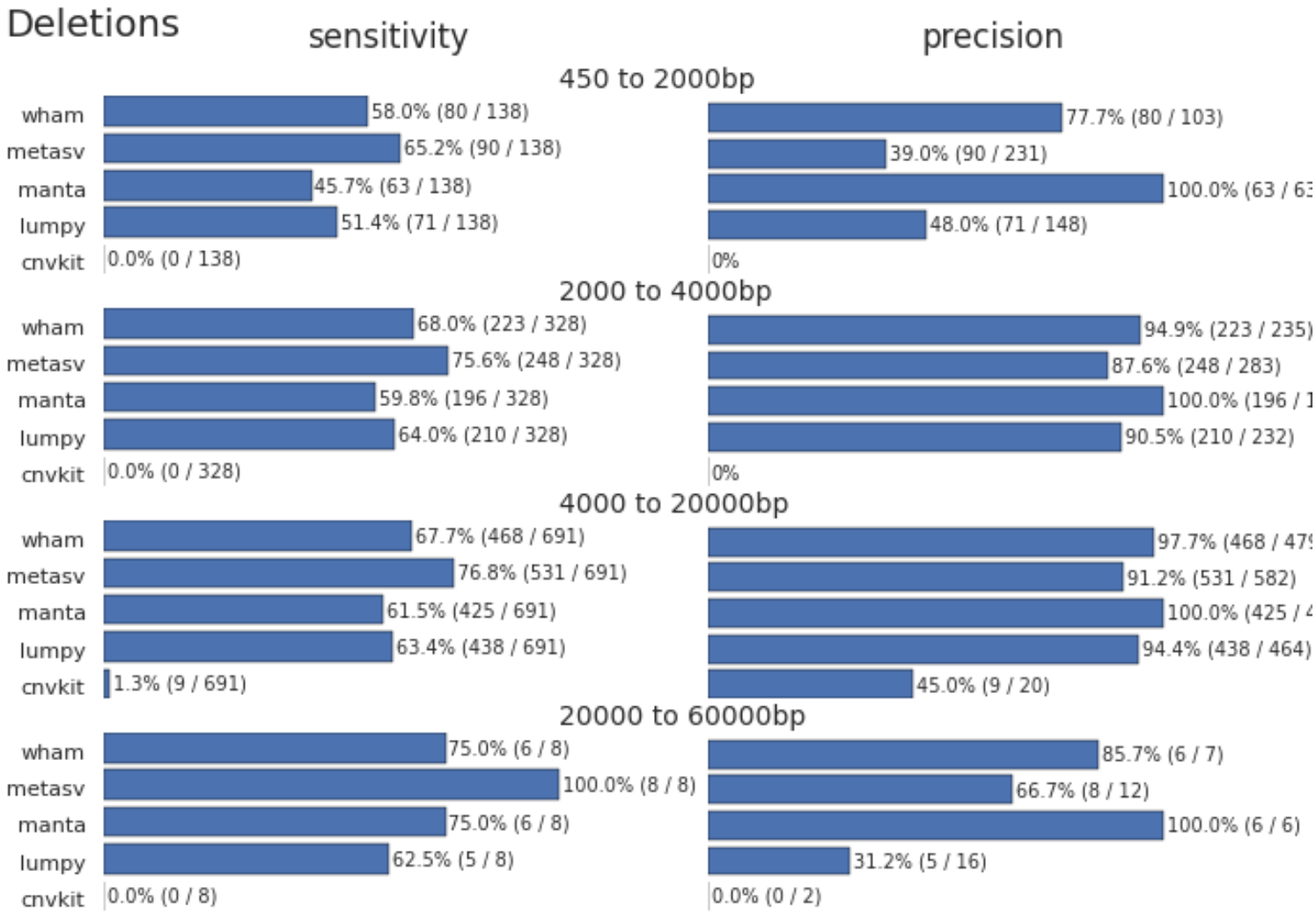
Last year: Somatic deletions



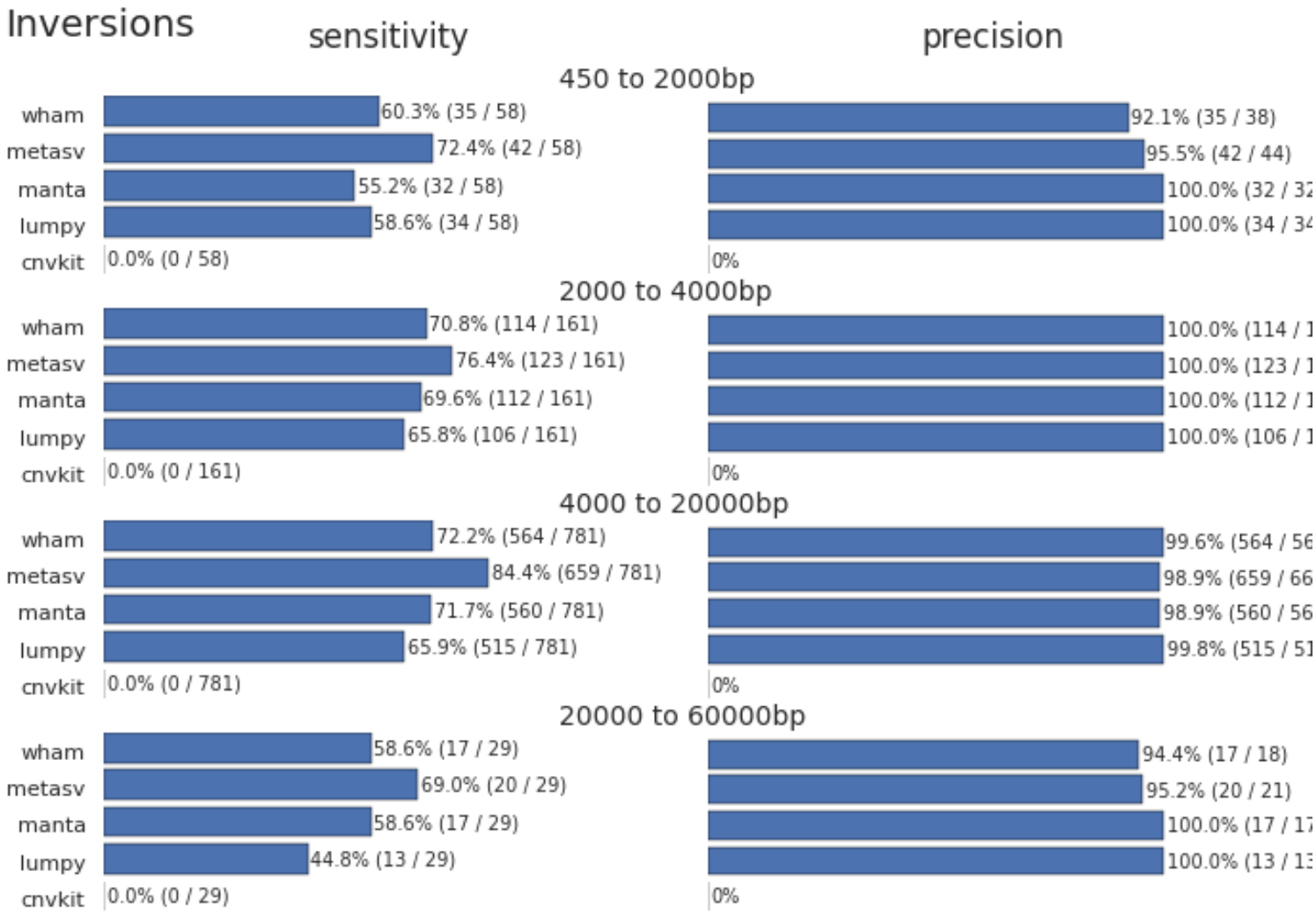
Results: Germline deletions



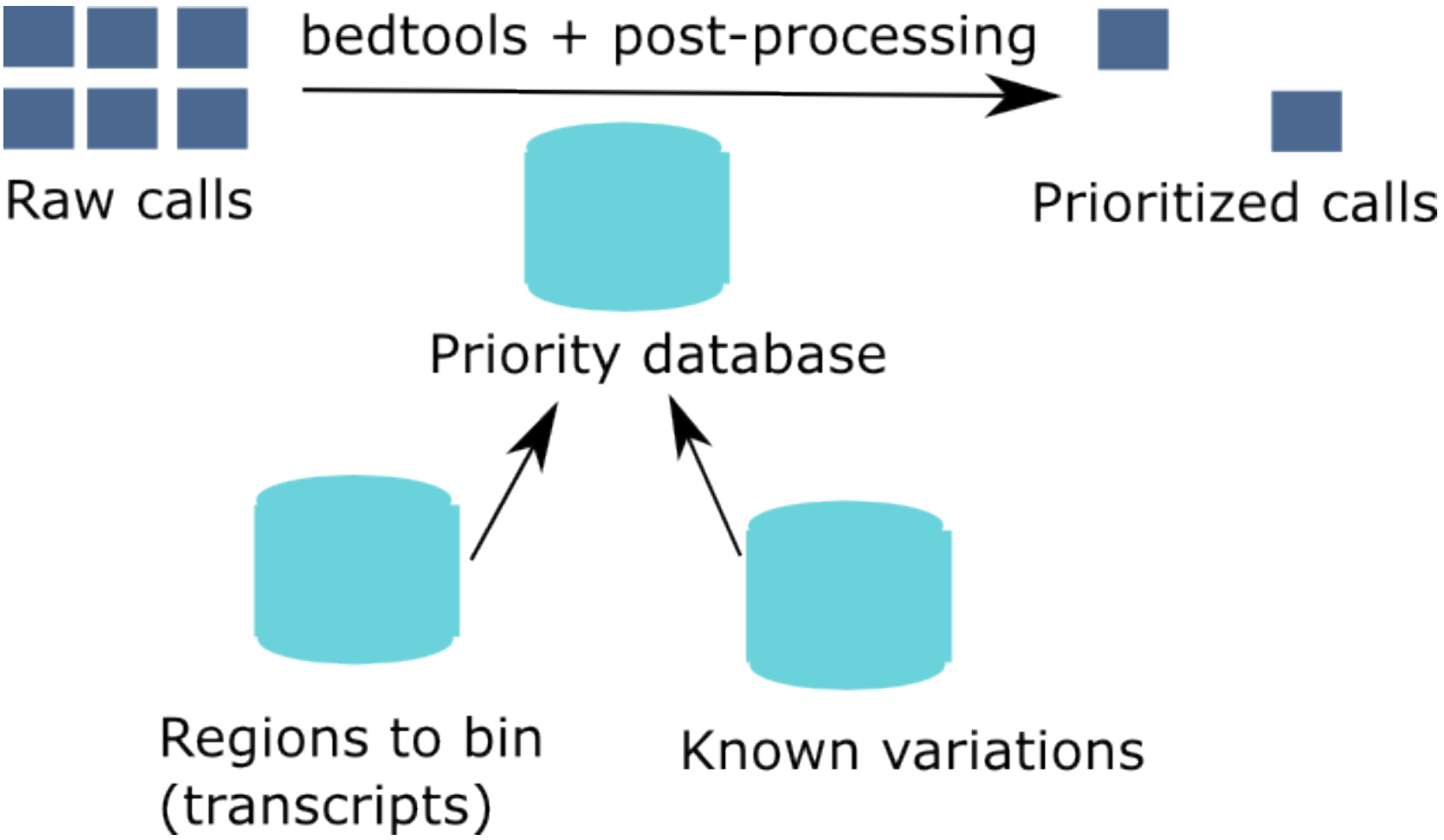
Results: Somatic deletions



Results: Somatic insertions

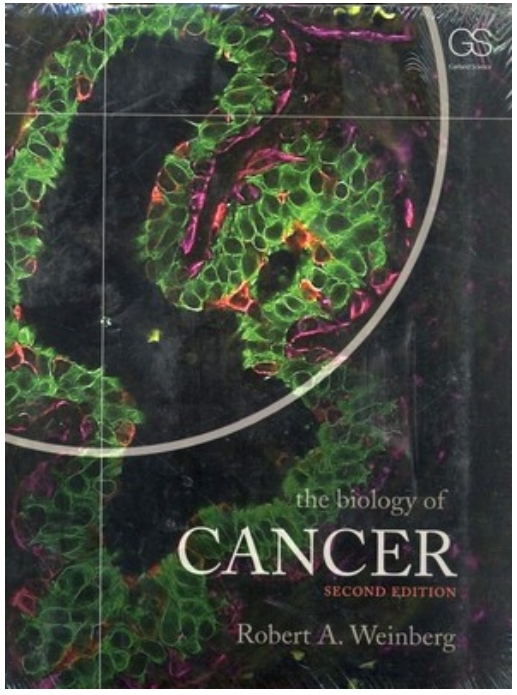


Prioritize in previously known regions



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

Summary

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

http://bcb.io

Practical overview

- Small dataset – single chromosome, exome
- Cancer sample from DREAM synthetic dataset 3
- Call against build 38
- Structrural variants

bcbio configuration file

```
---
details:
  - analysis: variant2
    genome_build: hg38
    algorithm:
      aligner: bwa
      mark_duplicates: true
      recalibrate: false
      realign: false
      variantcaller: [vardict, mutect, freebayes]
      ensemble:
        numpass: 2
      svcaller: [cnvkit, lumpy, manta]
```

https://bcbio-nextgen.readthedocs.org/en/latest/contents/configuration.html

bcbio template file – CSV

```
samplename,description,batch,phenotype,sex,variant_regions
sample1,ERR256785,batch1,normal,female,/path/to/regions.bed
sample2,ERR256786,batch1,tumor,,/path/to/regions.bed
```

https://bcbio-nextgen.readthedocs.org/en/latest/contents/configuration.html#automated-sample-configuration

Template to full configuration

```
bcbio_nextgen.py -w template \
    tumor-paired.yaml project1.csv \
    sample1.bam sample2_1.fq sample2_2.fq
```

https://bcbio-nextgen.readthedocs.org/en/latest/contents/configuration.html#automated-sample-configuration

Running bcbio

```
bcbio_nextgen.py bcbio_sample.yaml -n 8
```

https://bcbio-nextgen.readthedocs.org/en/latest/contents/testing.html