bcbio validation: build 38, low frequency somatic variants, structural variations

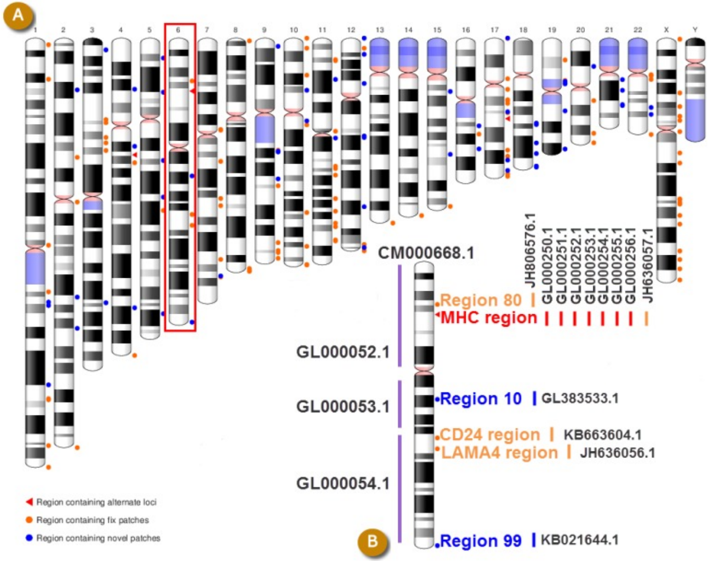
Brad Chapman
Bioinformatics Core, Harvard Chan School
https://bcb.io
http://j.mp/bcbiolinks

5 November 2015

Outline

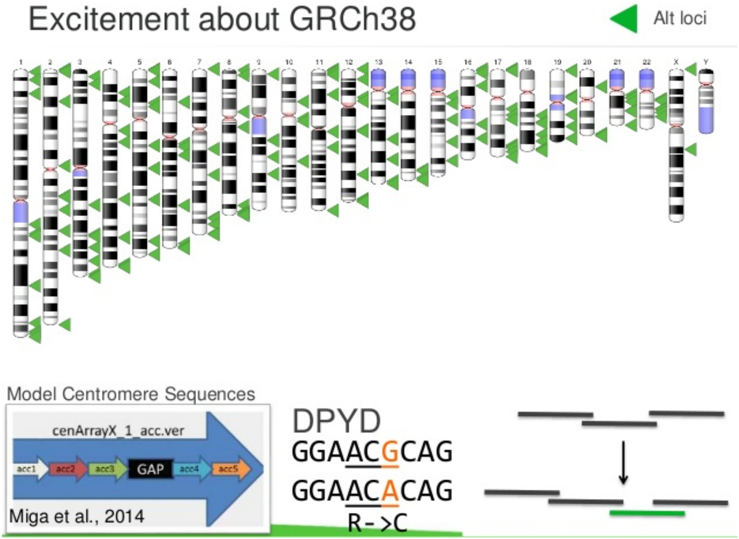
- **Human build 38**
- Low frequency somatic calling
- Structural variation

Currently: GRCh37/hg19



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

GRCh38 – graph based, many more alternative loci



Excitement about GRCh38

Alt loci

Model Centromere Sequences

cenArrayX_1_acc.ver

Miga et al., 2014

DPYD

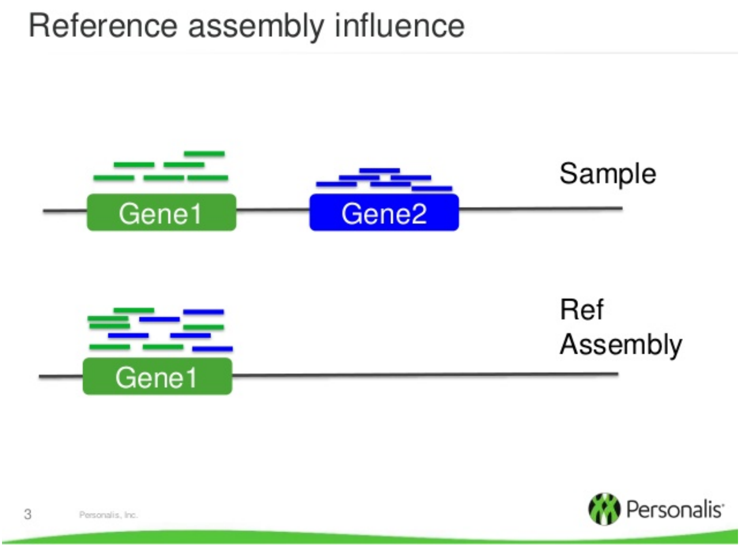
GGAACGCAG

GGAACACAG

R->C

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

GRCh38 – advantage for variant calling



Reference assembly influence

Gene1

Gene2

Sample

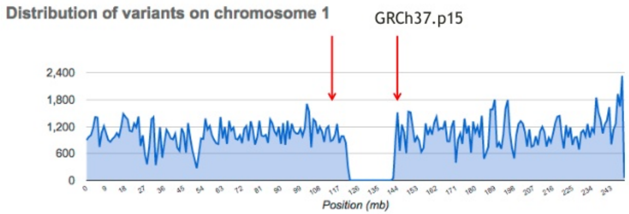
Gene1

Ref Assembly

Personalis

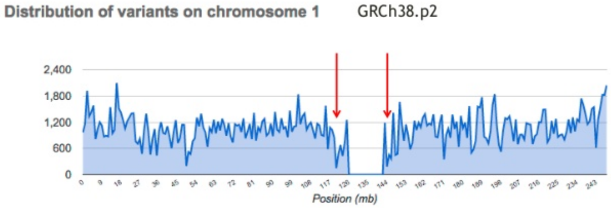
http://www.slideshare.net/GenomeRef/transitioning-to-grch38

Avoiding collapsed repeats



Distribution of variants on chromosome 1

GRCh37.p15



Distribution of variants on chromosome 1

GRCh38.p2

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

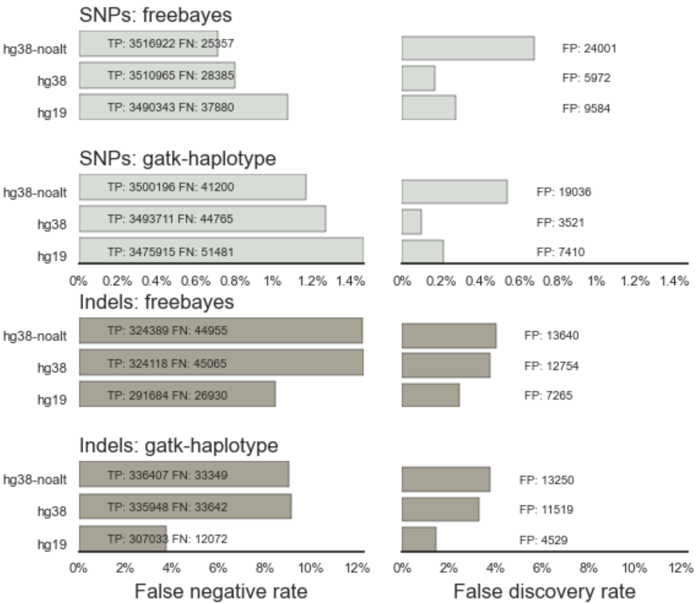
Genome in a Bottle Consortium

Global Alliance for Genomics & Health

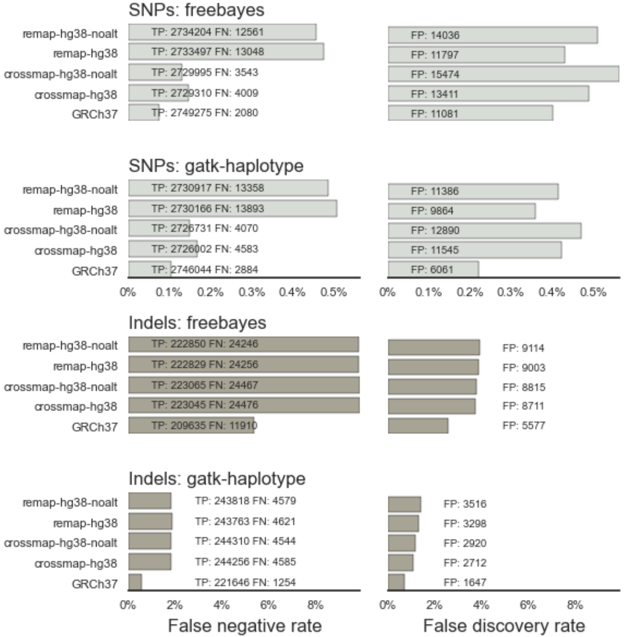
ICGC-TCGA DREAM Mutation Calling challenge

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

hg19/hg38 comparison: NA12878 Platinum Genomes



GRCh37/hg38 comparison: NA12878 Genome in a Bottle



Small variant results

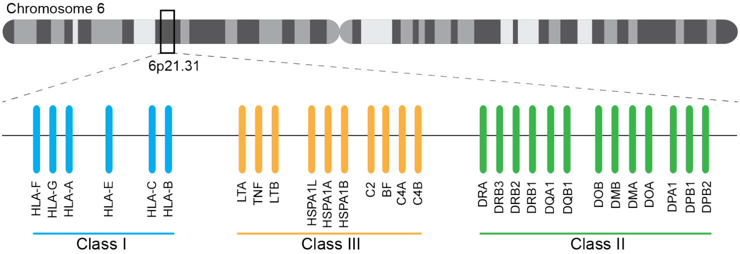
- SNPs: build 38 more sensitive
- SNPs: build 38 reduces false positives
- Indels: build 38 detected more
- Indels: work on sensitivity and precision

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

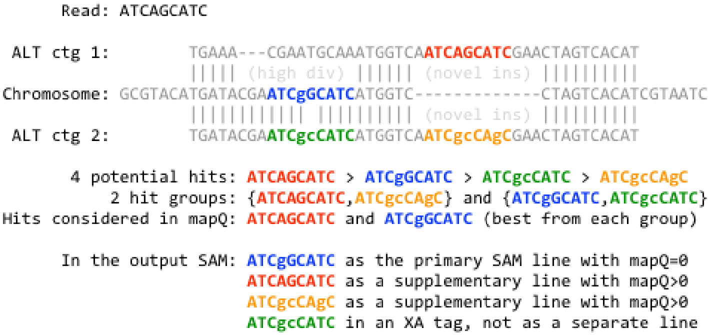
Major histocompatibility complex (MHC) – HLAs



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
```

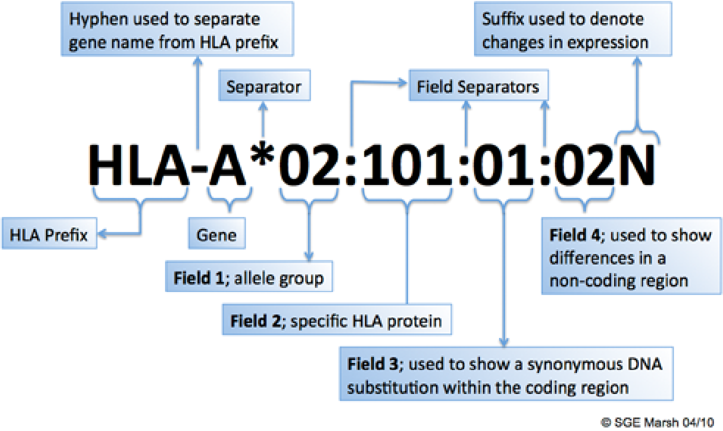
https://github.com/lh3/bwa/blob/master/README-alt.md

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


Hyphen used to separate gene name from HLA prefix
Separator
Field Separators
Suffix used to denote changes in expression
HLA-A*02:101:01:02N
HLA Prefix
Gene
Field 1; allele group
Field 2; specific HLA protein
Field 3; used to show a synonymous DNA substitution within the coding region
Field 4; used to show differences in a non-coding region
© SGE Marsh 04/10

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

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

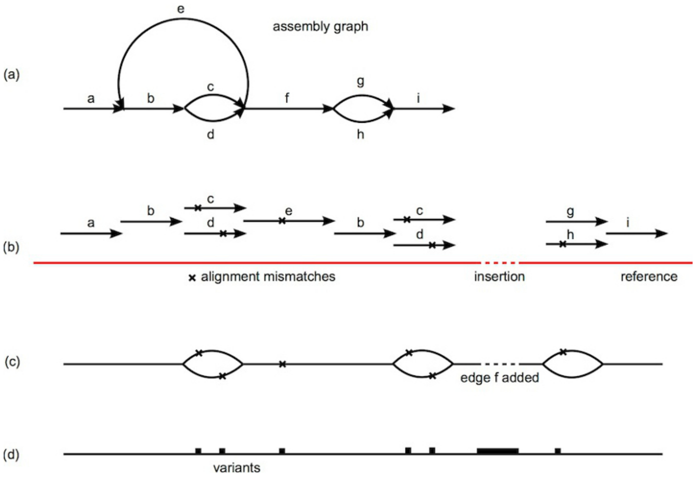
Validations

- 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

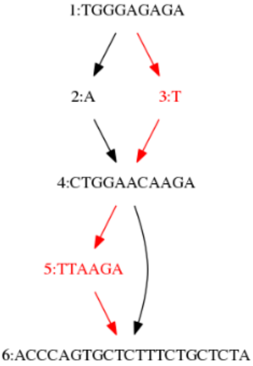
Genome graphs and variation



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
...			

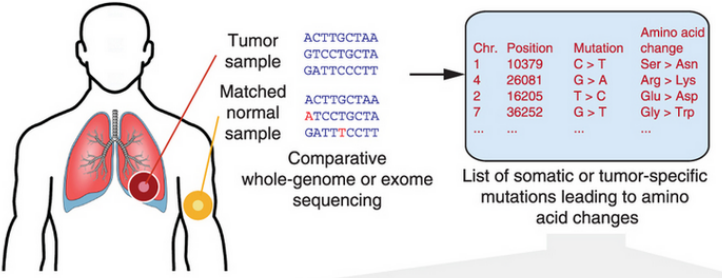


https://docs.google.com/presentation/d/1bbl2zY4qWQ0yYBHhoVuXb79HdgajRotIUa_VEn3kTpI

Outline

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

Cancer somatic calling



Tumor sample

ACTTGCTAA
GTCCTGCTA
GATTCCCTT

Matched normal sample

ACTTGCTAA
ATCCTGCTA
GATTTCCTT

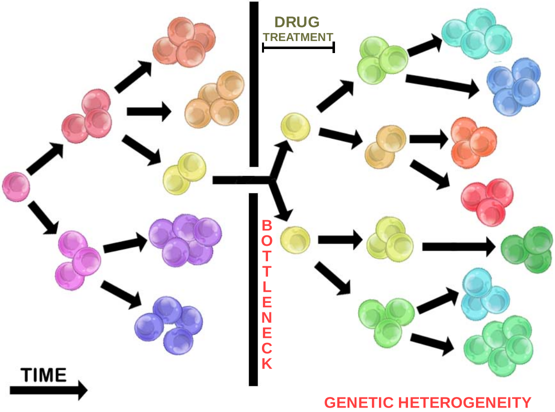
Comparative whole-genome or exome sequencing

Chr.	Position	Mutation	Amino acid change
1	10379	C > T	Ser > Asn
4	26081	G > A	Arg > Lys
2	16205	T > C	Glu > Asp
7	36252	G > T	Gly > Trp
...	...	...	...

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



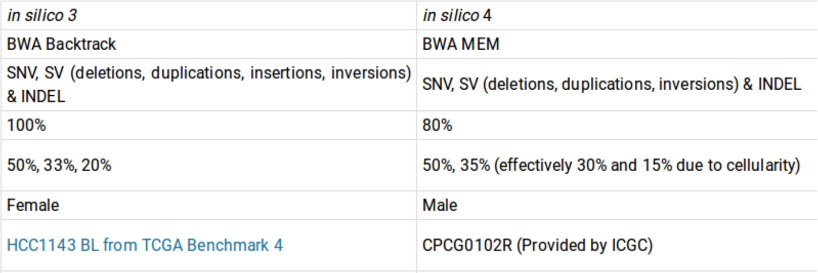
http://en.wikipedia.org/wiki/Tumour_heterogeneity

VarDict

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

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

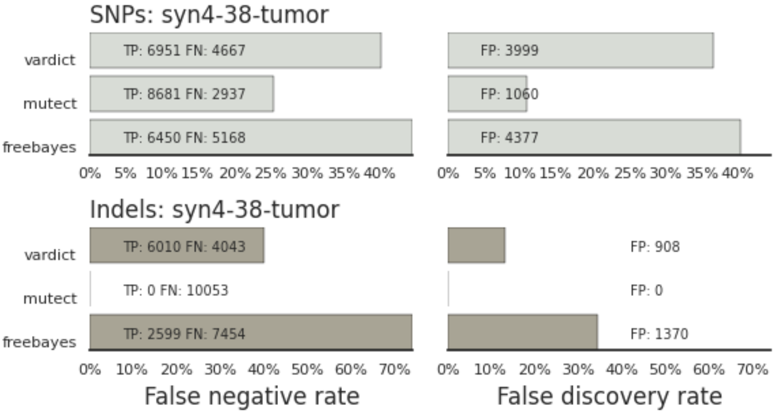
DREAM synthetic dataset 4



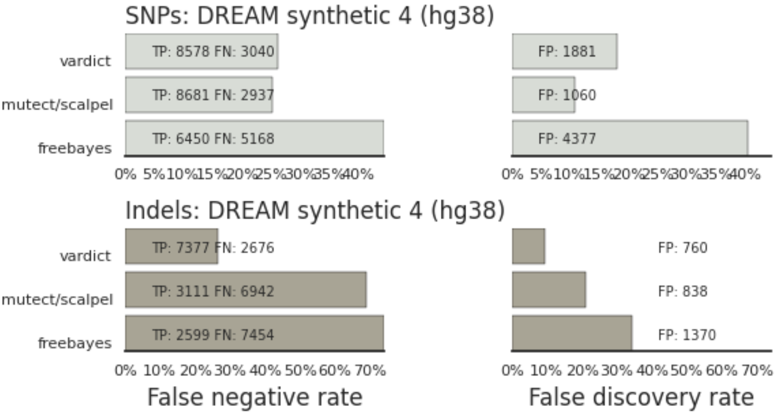
in silico 3	*in silico* 4
BWA Backtrack	BWA MEM
SNV, SV (deletions, duplications, insertions, inversions) & INDEL	SNV, SV (deletions, duplications, inversions) & INDEL
100%	80%
50%, 33%, 20%	50%, 35% (effectively 30% and 15% due to cellularity)
Female	Male
HCC1143 BL from TCGA Benchmark 4	CPCG0102R (Provided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

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

VarDict sensivitity/precision before



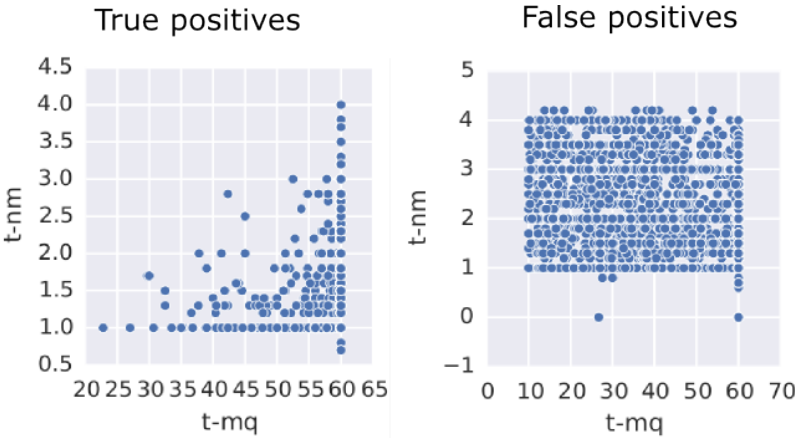
VarDict sensivitity/precision after



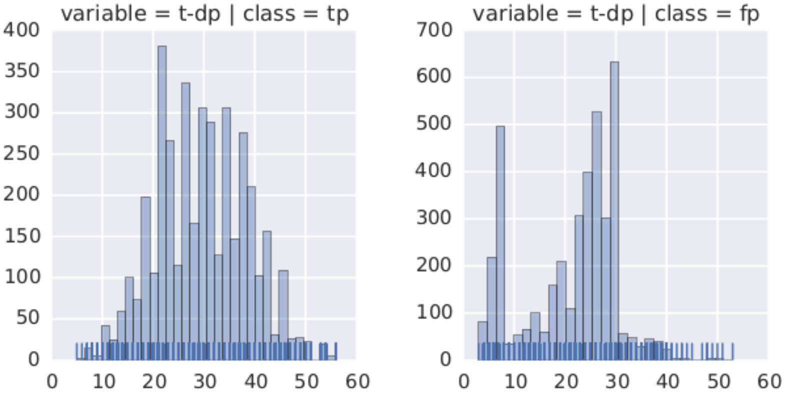
How? Filter summary

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

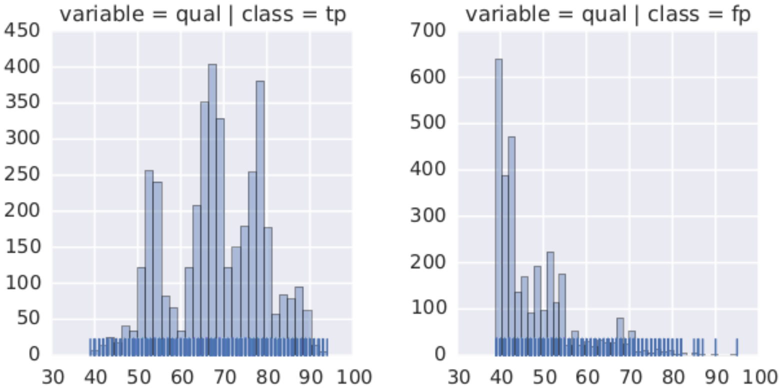
Filter: mapping quality and number of mismatches



Filter: low depth



Filter: low quality



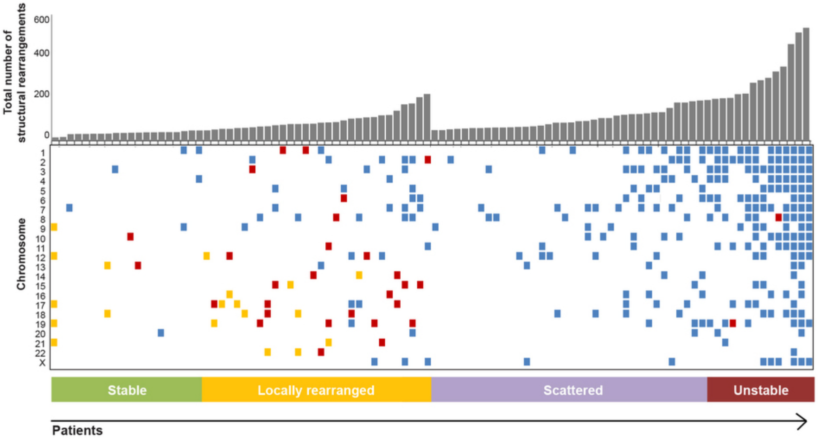
How can we improve?

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

Outline

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

Structural variants critical in cancer

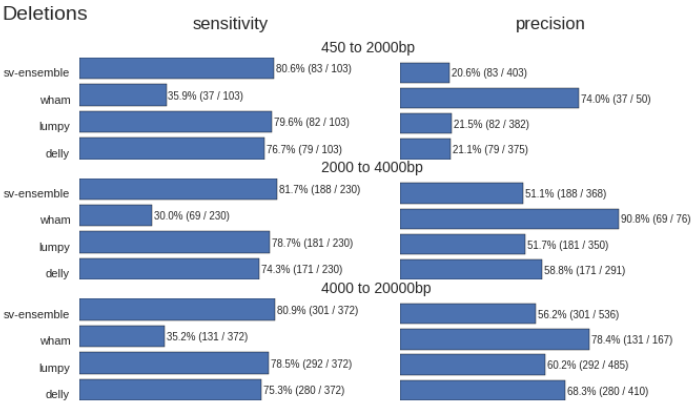


http://www.nature.com/nature/journal/v518/n7540/full/nature14169.html

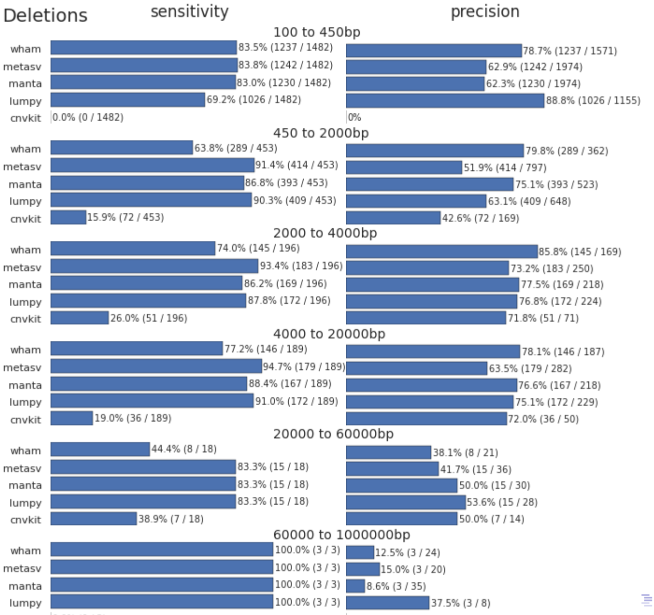
Improvements in speed, sensitivity and precision

- 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

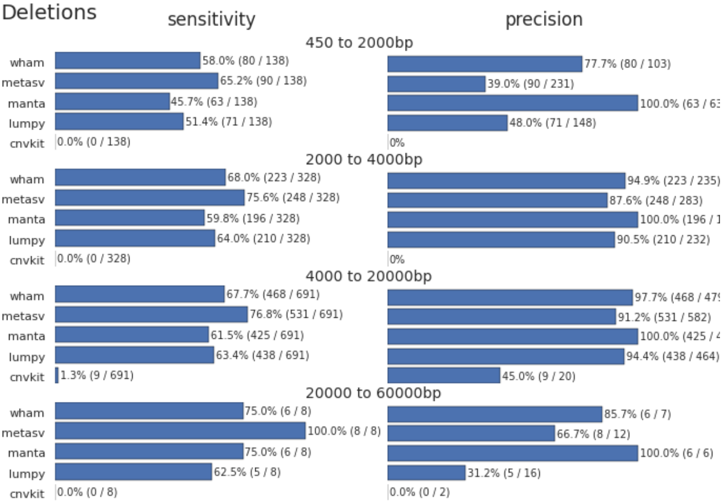


Results: Germline deletions

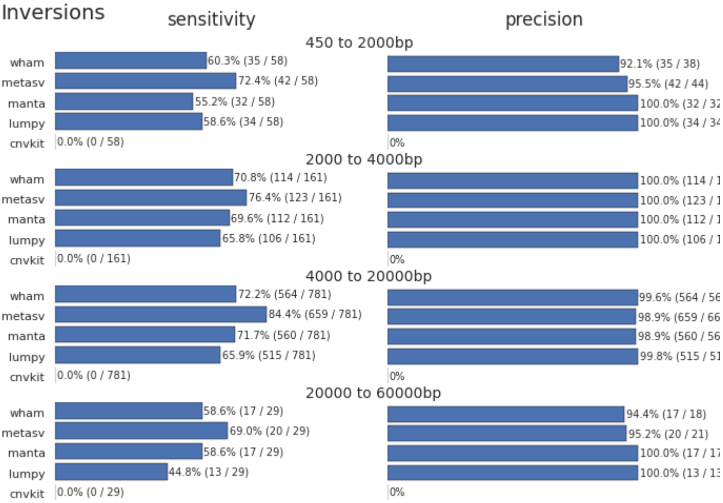


]

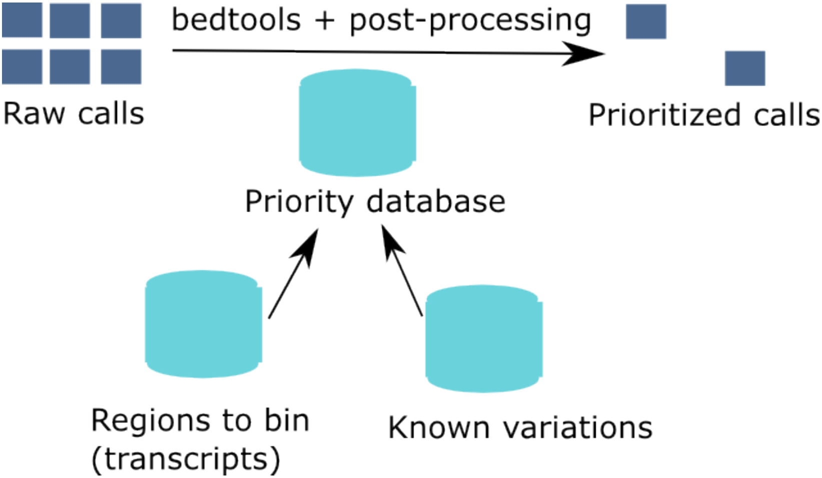
Results: Somatic deletions



Results: Somatic insertions



Prioritize in previously known regions



bedtools + post-processing

Raw calls

Prioritized calls

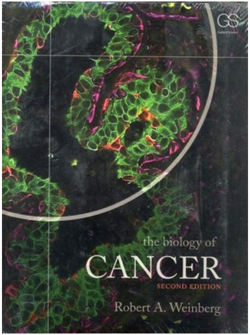
Priority database

Regions to bin (transcripts)

Known variations

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

- Demonstrate current validation work in bcbio
- Human build 38
- HLA typing
- Low frequency cancer calling
- Structural variations + prioritization

http://bcb.io