Variant calling: tools, validation, genomes and outputs

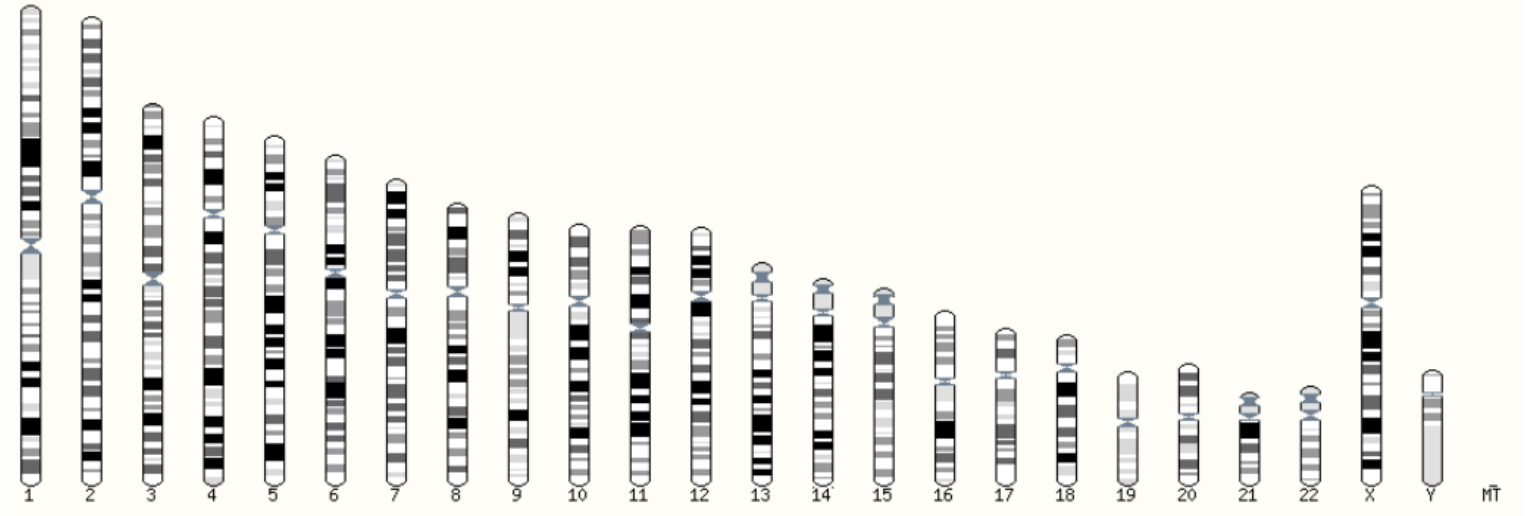
Brad Chapman
Bioinformatics Core, Harvard Chan School
https://github.com/chapmanb/bcbio-nextgen
http://bcb.io
http://j.mp/bcbiolinks

19 May 2015

Outline

- **Variant calling made easy**
- Tools
- Validation
- Post-calling annotation
- Genomes and graphs
- Understanding outputs
- Automating everything – bcbio

Human whole genome sequencing



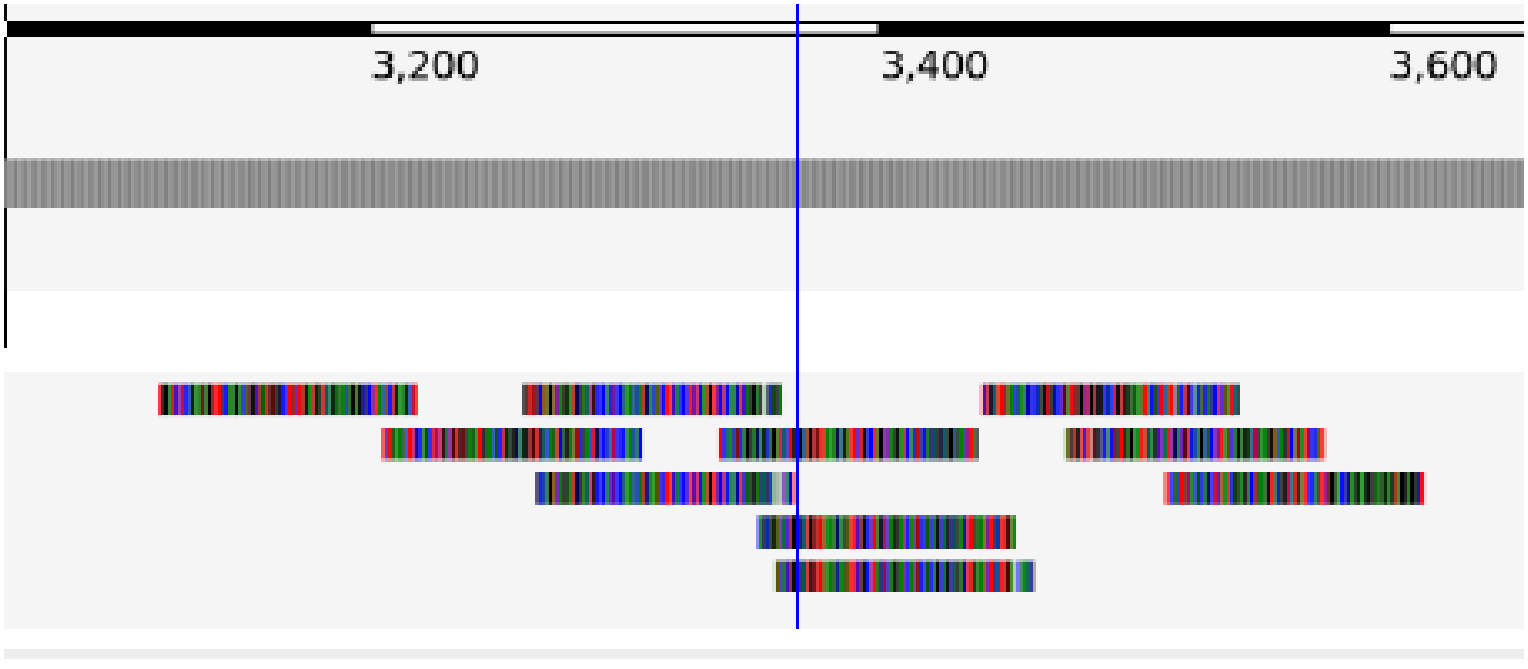
Click on the image above to jump to a chromosome, or click and drag to select a region

Summary

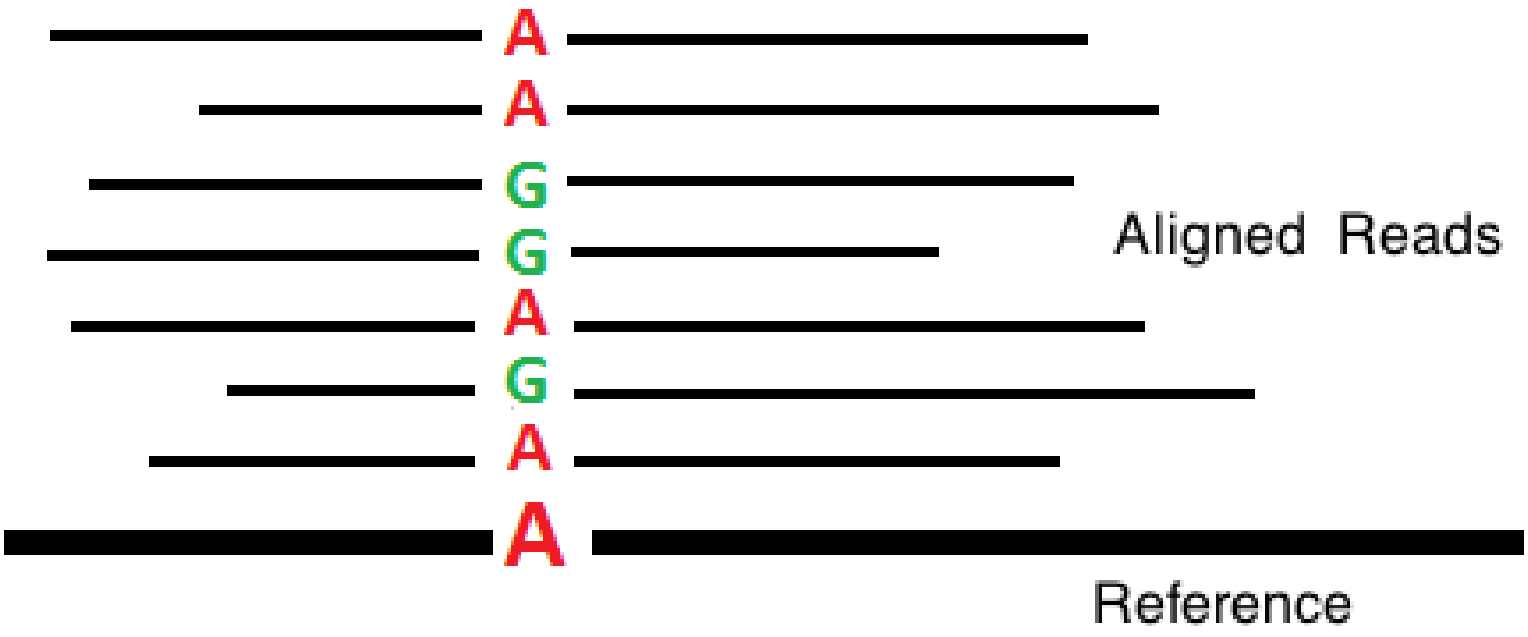
Assembly	GRCh37.p13 (Genome Reference Consortium Human Reference 37), INSDC Assembly GCA_000001405.14, Feb 2009
Database version	75.37
Base Pairs	3,326,743,047

http://ensembl.org/Homo_sapiens/Location/Genome

High throughput sequencing

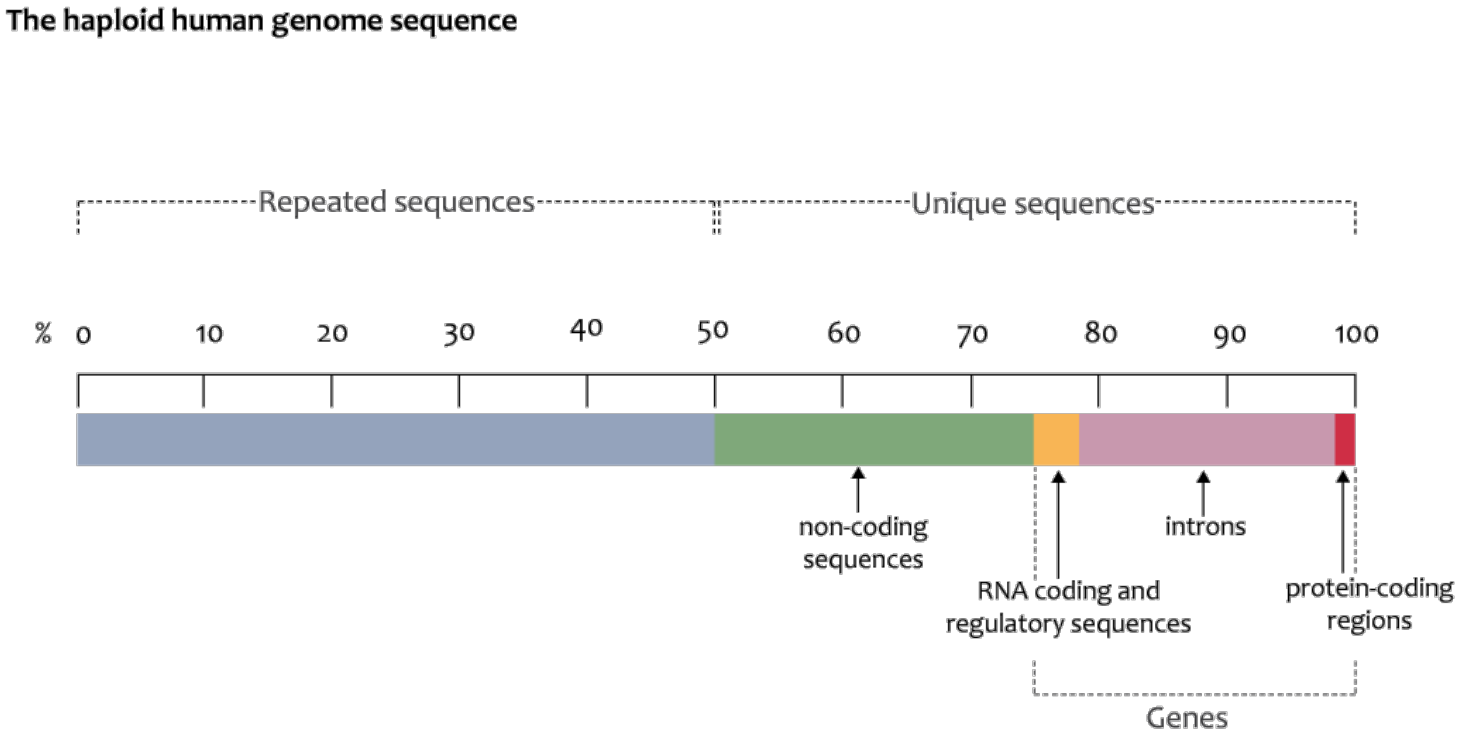


Variant calling



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

Scale: exome to whole genome

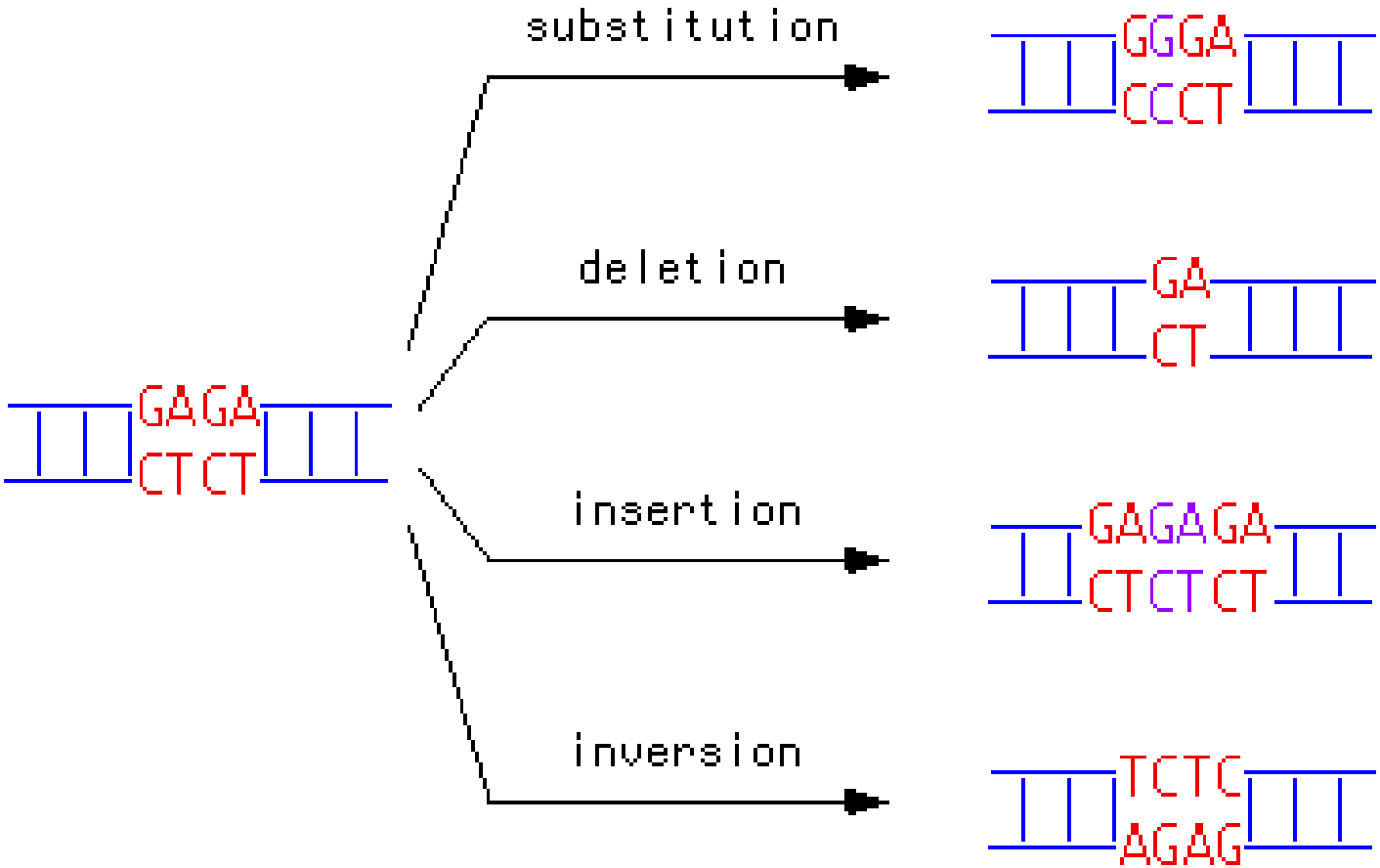


https://www.flickr.com/photos/119980645@N06/

Outline

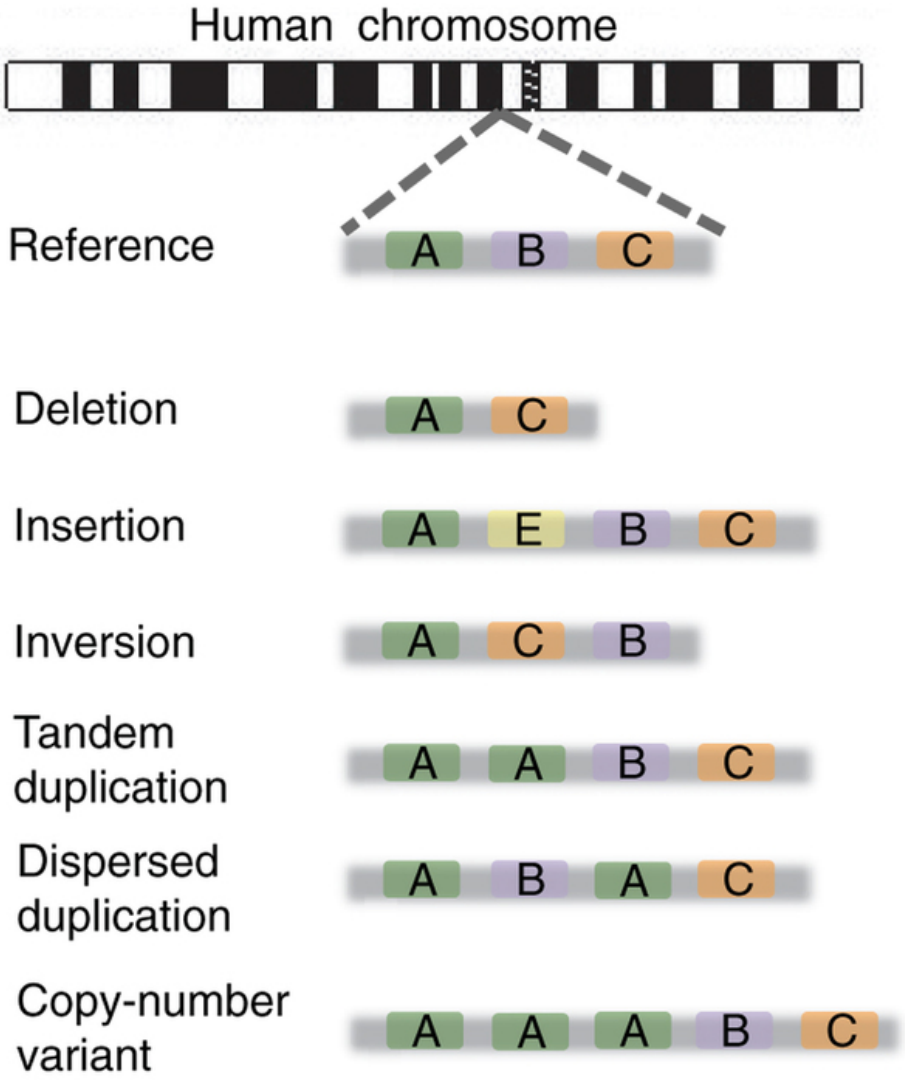
- Variant calling made easy
- **Tools**
- Validation
- Post-calling annotation
- Genomes and graphs
- Understanding outputs
- Automating everything – bcbio

SNPs and Indels



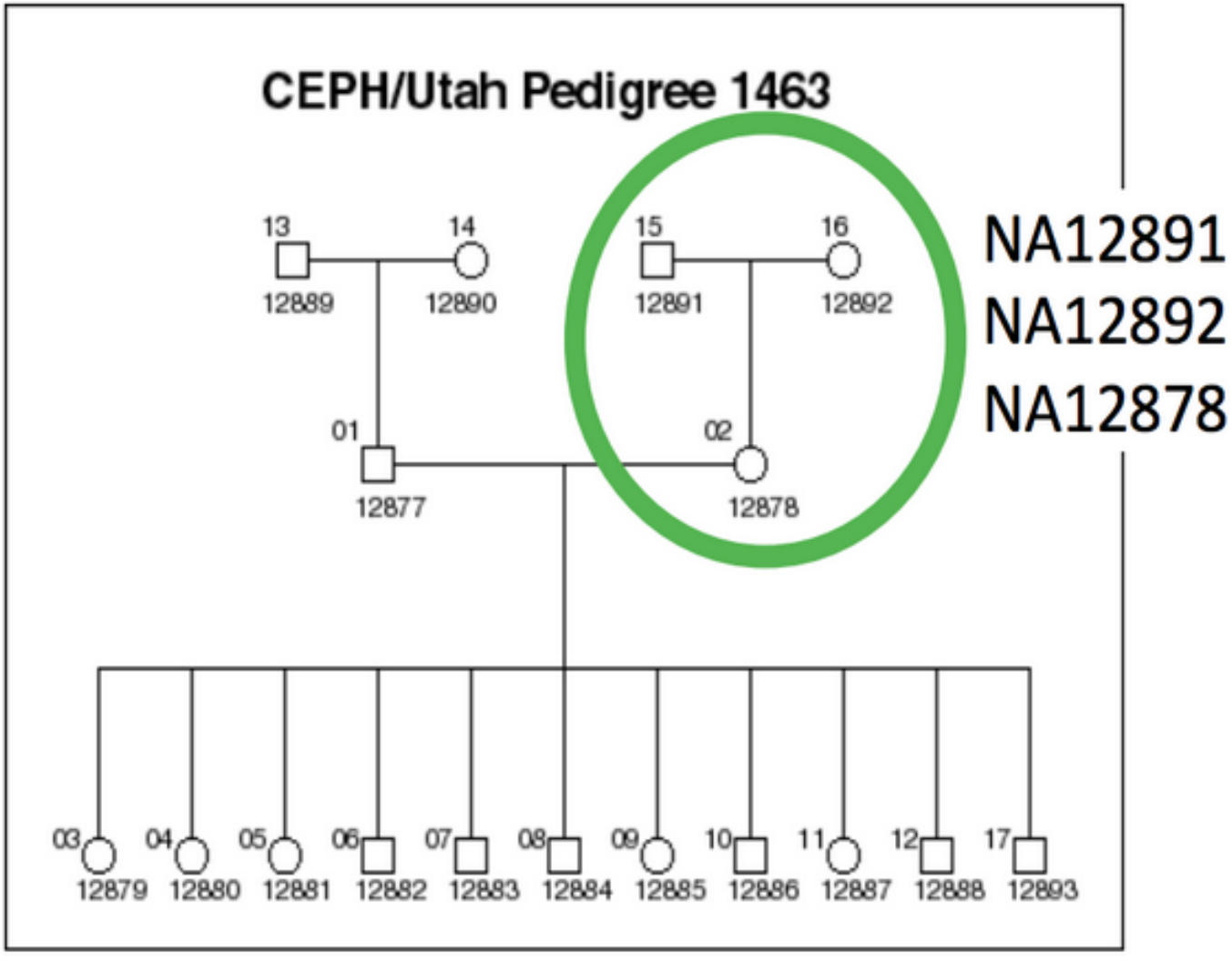
http://carolguze.com/text/442-2-mutations.shtml

Structural variations



http://www.nature.com/nmeth/journal/v9/n2/full/nmeth.1858.html

Germline population calling



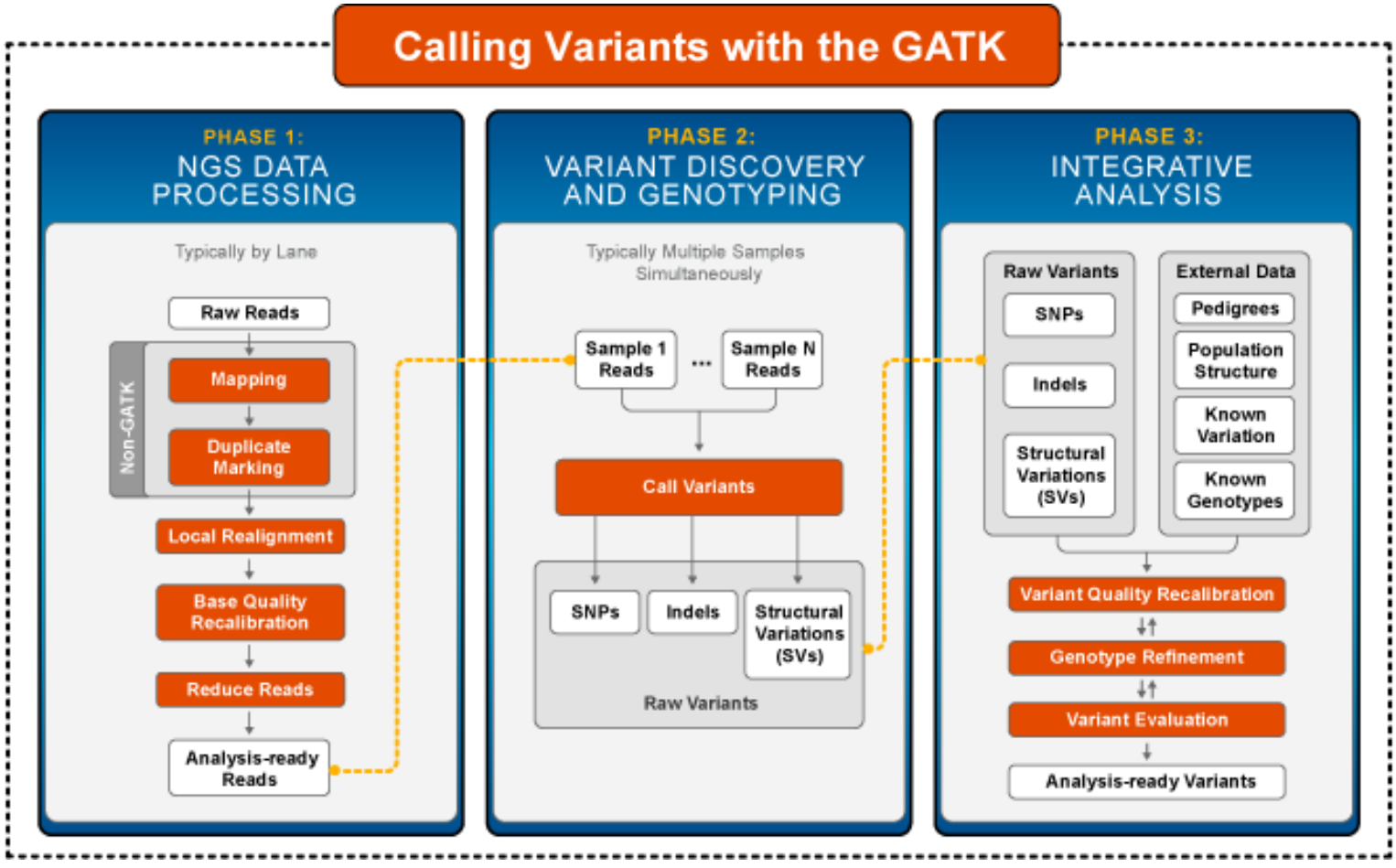
http://blog.goldenhelix.com/grudy/the-state-of-ngs-variant-calling-dont-panic/

Genome Analysis Toolkit (GATK)

The Genome Analysis Toolkit or GATK is a software package developed at the Broad Institute to analyze high-throughput sequencing data. The toolkit offers a wide variety of tools, with a primary focus on variant discovery and genotyping as well as strong emphasis on data quality assurance. Its robust architecture, powerful processing engine and high-performance computing features make it capable of taking on projects of any size.

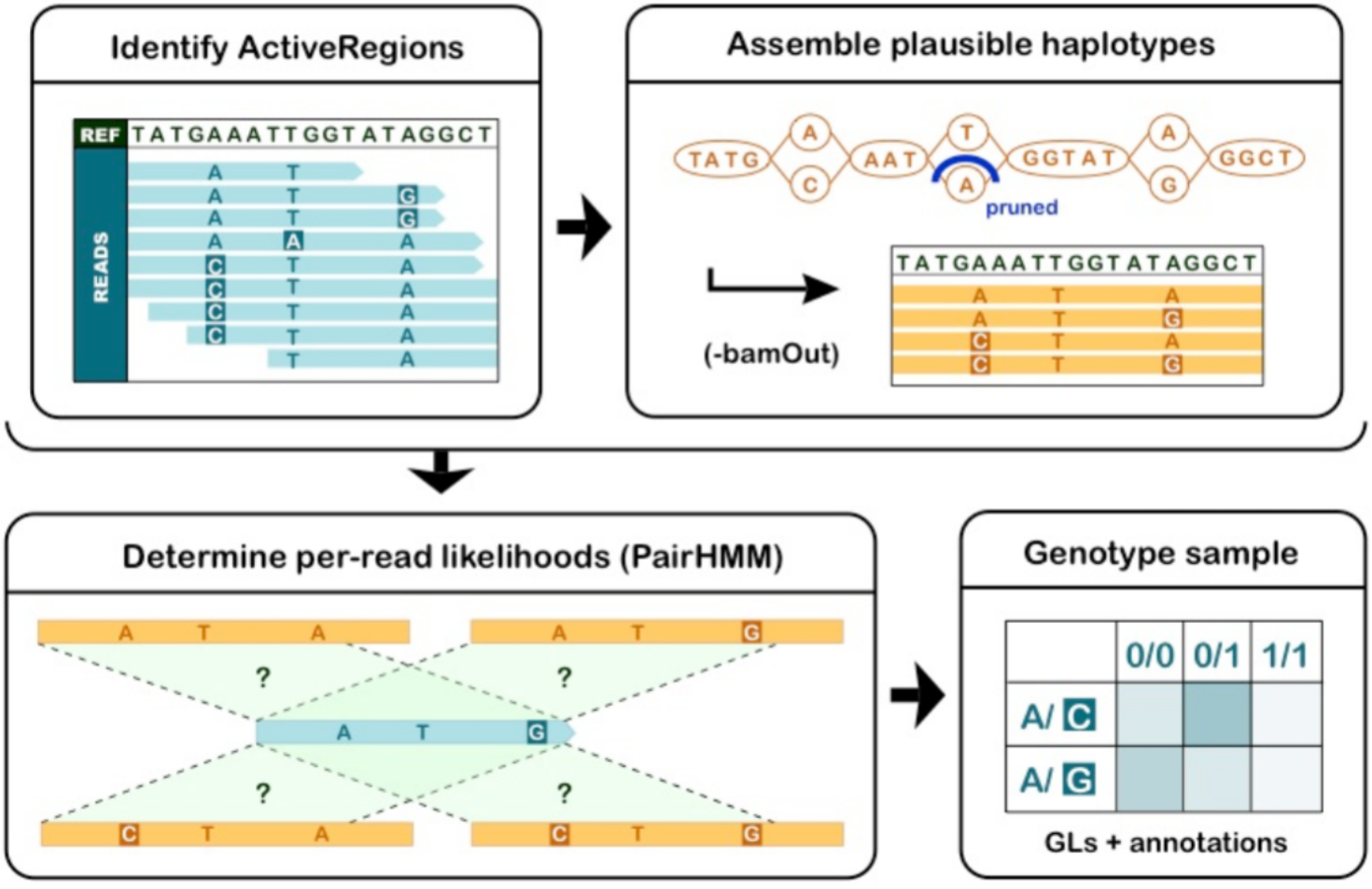
https://www.broadinstitute.org/gatk/

GATK Best Practices



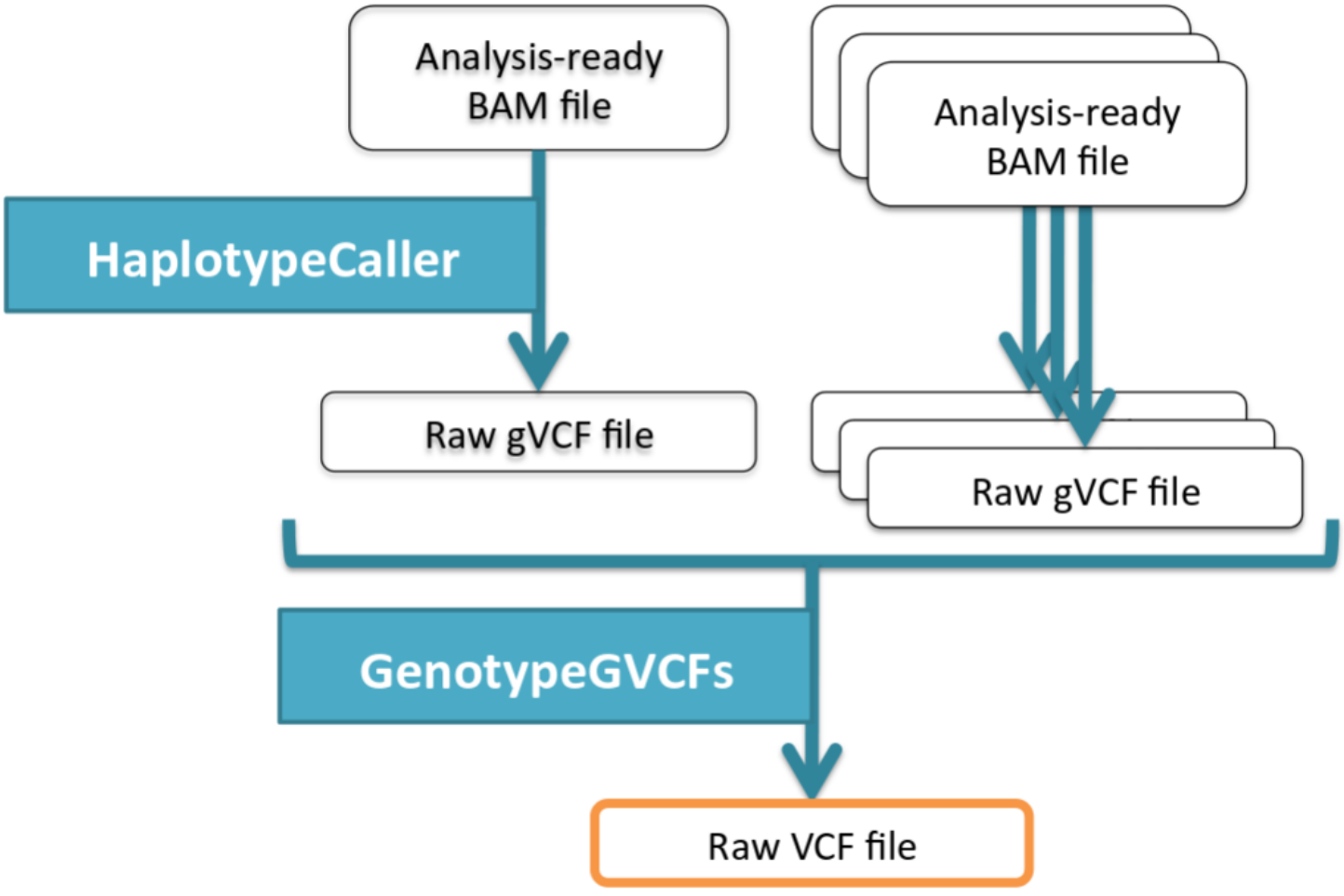
https://www.broadinstitute.org/gatk/guide/best-practices

HaplotypeCaller



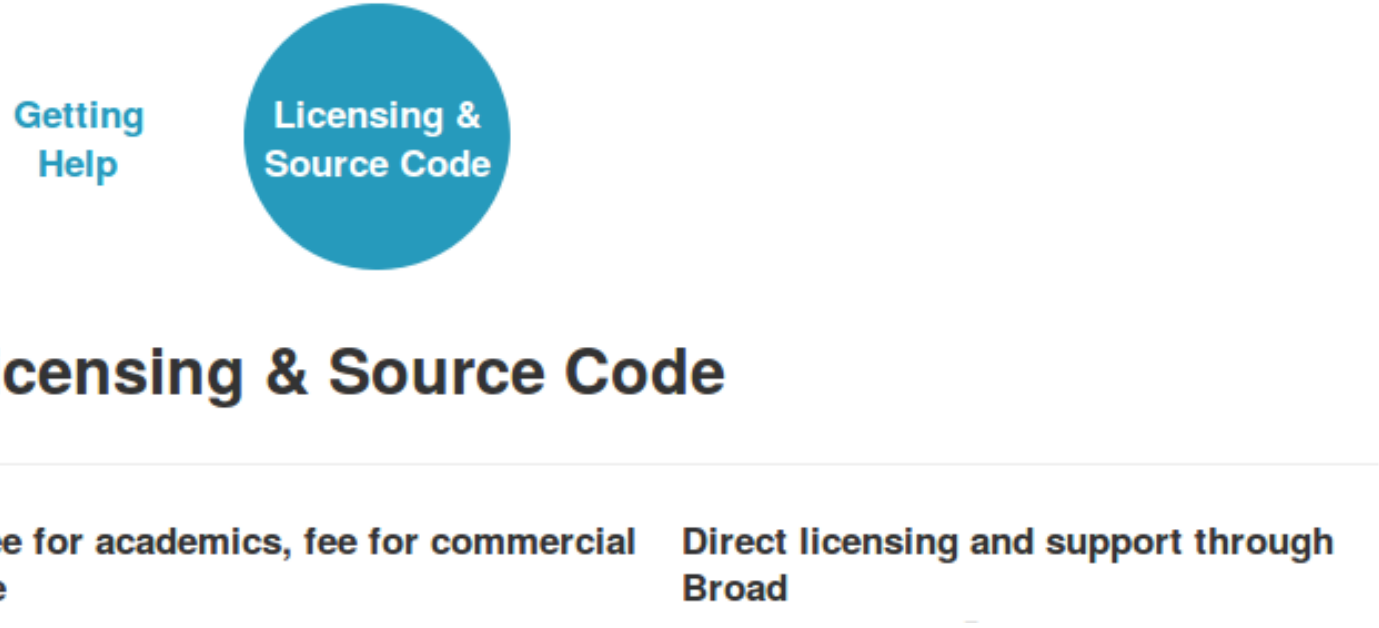
http://gatkforums.broadinstitute.org/discussion/5464/workshop-presentations-2015-uk-4-20-24

Joint calling on large populations



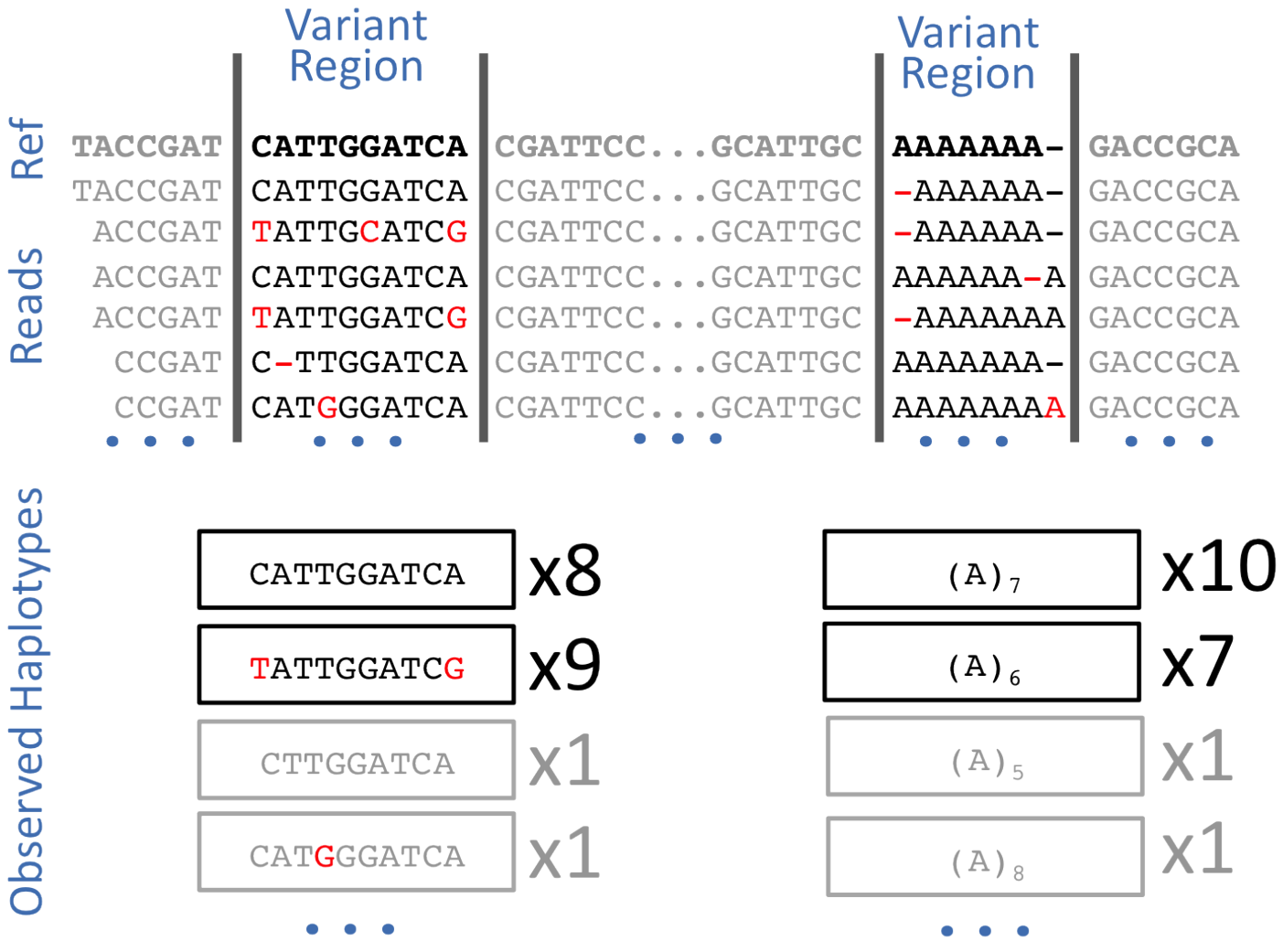
http://gatkforums.broadinstitute.org/discussion/5464/workshop-presentations-2015-uk-4-20-24

Support and availability



https://github.com/broadgsa

FreeBayes



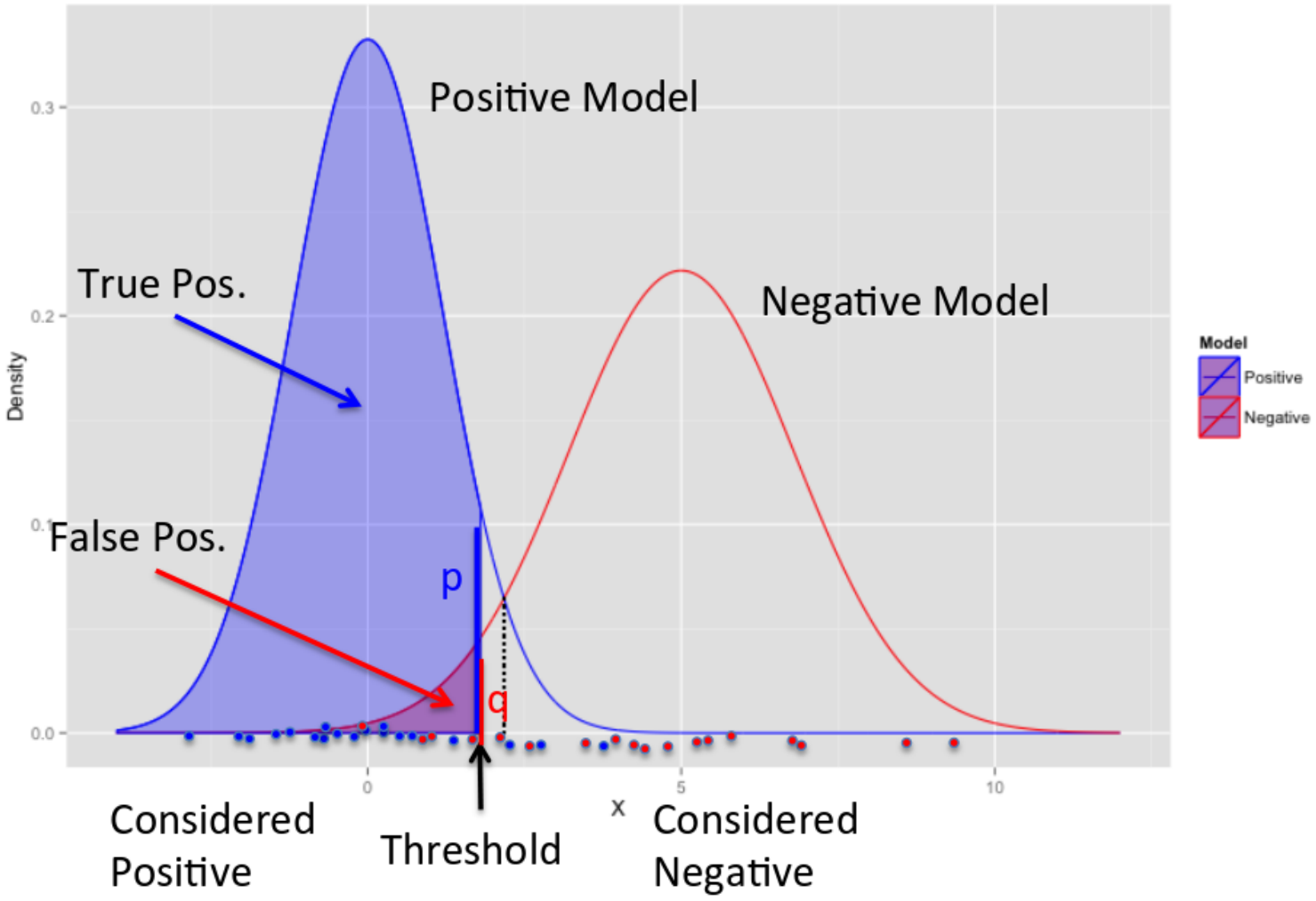
https://github.com/ekg/freebayes

Filtering – hard cutoffs

```
filters = ('((AC[0] / AN) <= 0.5 && DP < 4 && %QUAL < 20) || '
           '(DP < 13 && %QUAL < 10) || '
           '((AC[0] / AN) > 0.5 && DP < 4 && %QUAL < 50)')
```

http://bcb.io/2014/05/12/wgs-trio-variant-evaluation/

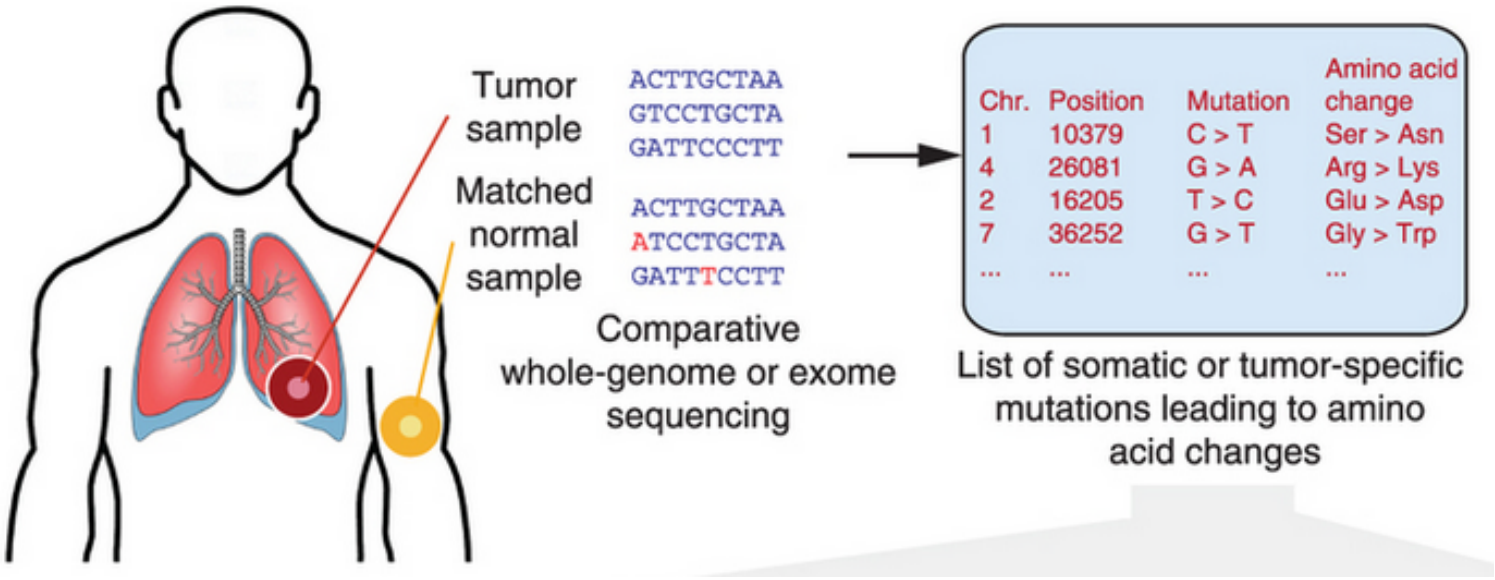
Filtering – Variant Quality Score Recalibration



VQSLOD(x) = Log(p(x)/q(x))

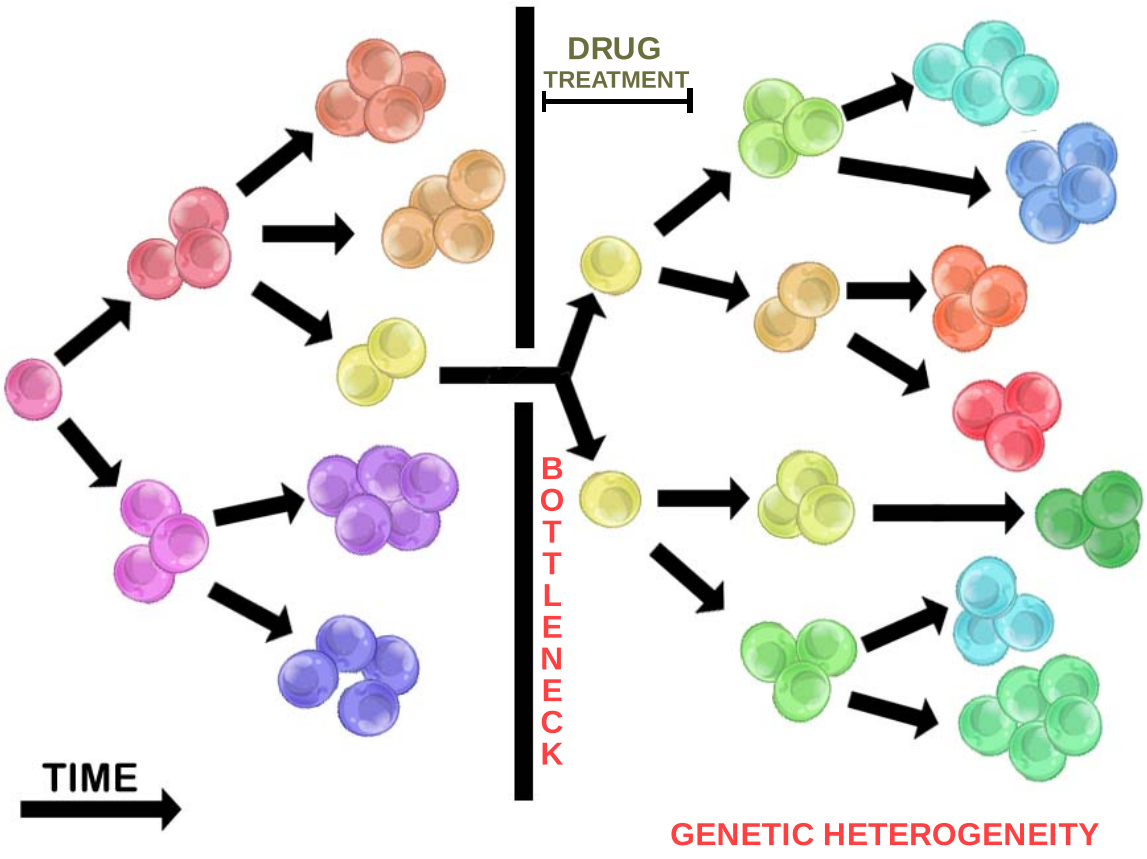
http://gatkforums.broadinstitute.org/discussion/5464/workshop-presentations-2015-uk-4-20-24

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

MuTect

- Broad GATK UnifiedGenotyper based
- SNP only

https://www.broadinstitute.org/cancer/cga/mutect

VarDict

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

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

Outline

- Variant calling made easy
- Tools
- **Validation**
- Post-calling annotation
- Genomes and graphs
- Understanding outputs
- Automating everything – bcbio

Reference materials

Genome in a Bottle Consortium

Global Alliance for Genomics & Health

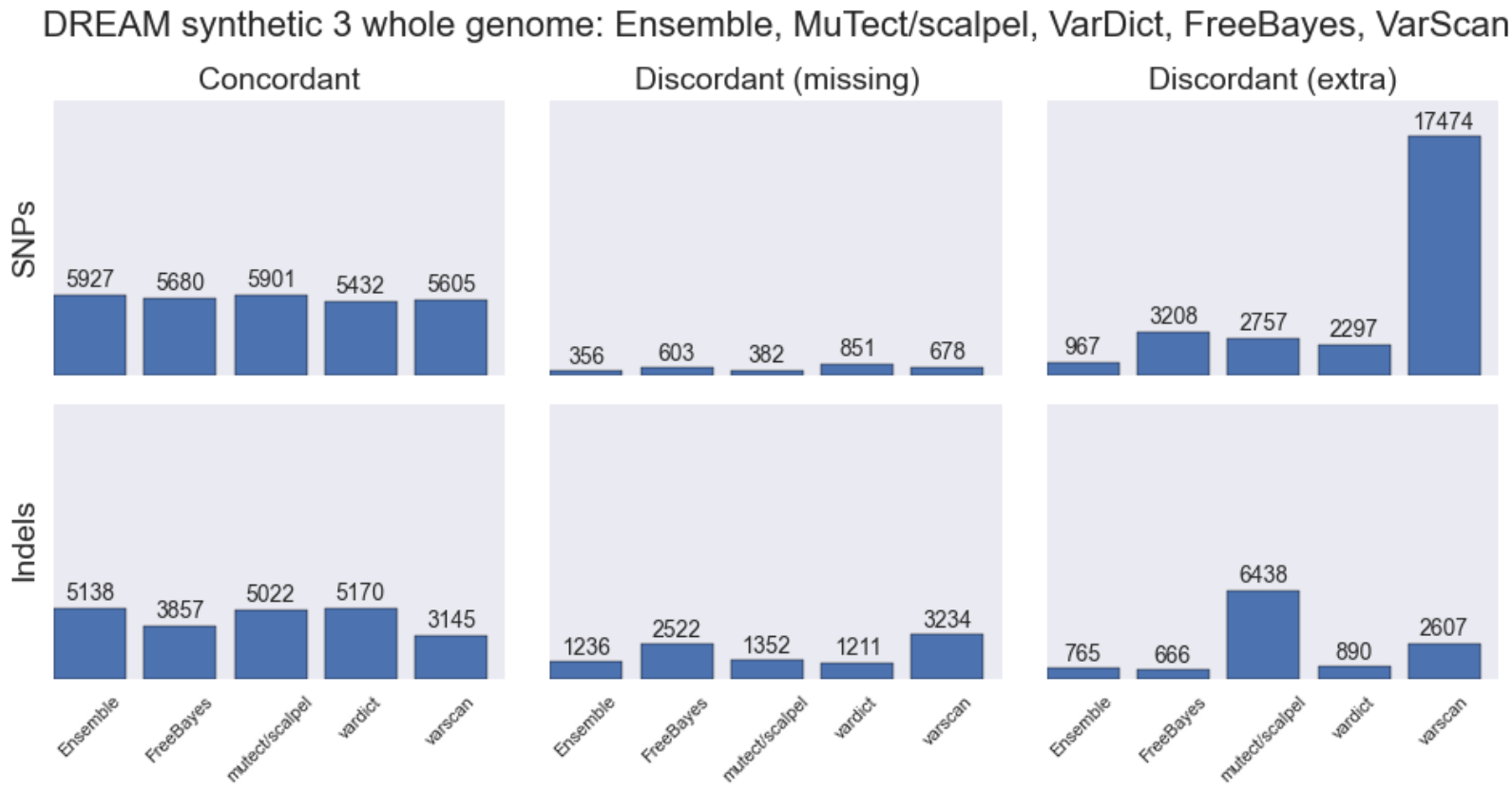
ICGC-TCGA DREAM Mutation Calling challenge

http://www.genomeinabottle.org/

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

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

Validate and compare caller performance

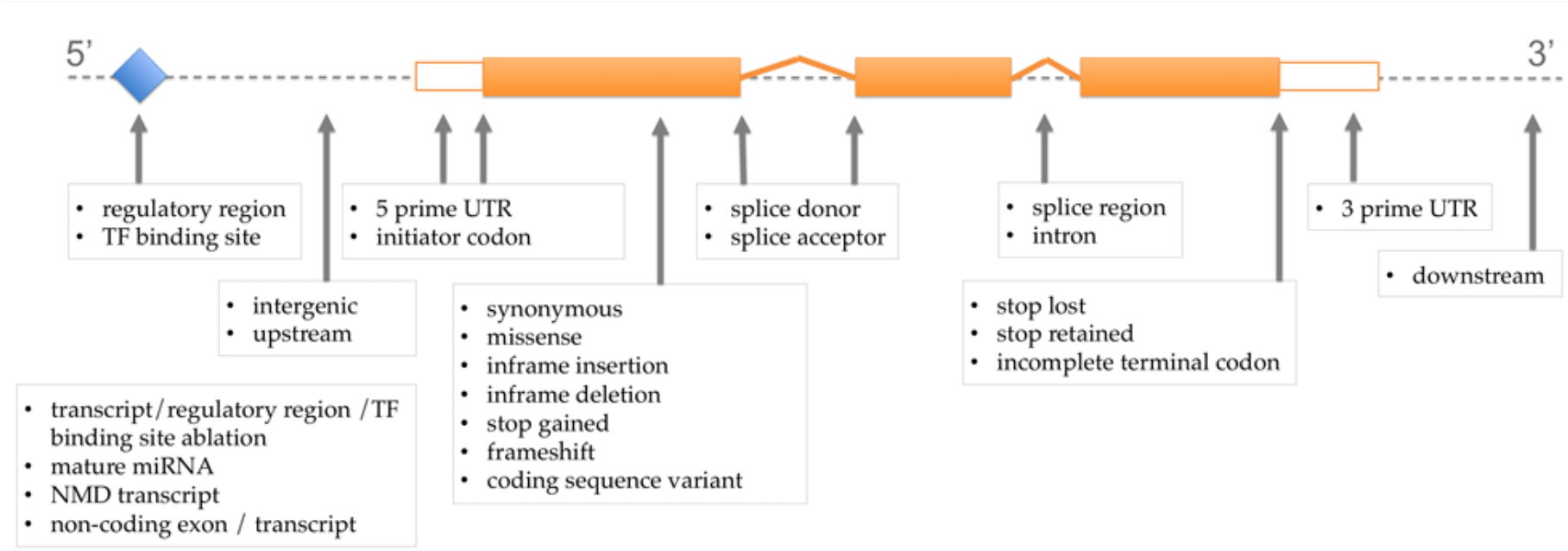


http://bcb.io/2015/03/05/cancerval/

Outline

- Variant calling made easy
- Tools
- Validation
- **Post-calling annotation**
- Genomes and graphs
- Understanding outputs
- Automating everything – bcbio

Effects prediction



http://www.ensembl.org/info/genome/variation/predicted_data.html

Tools for effects predictions

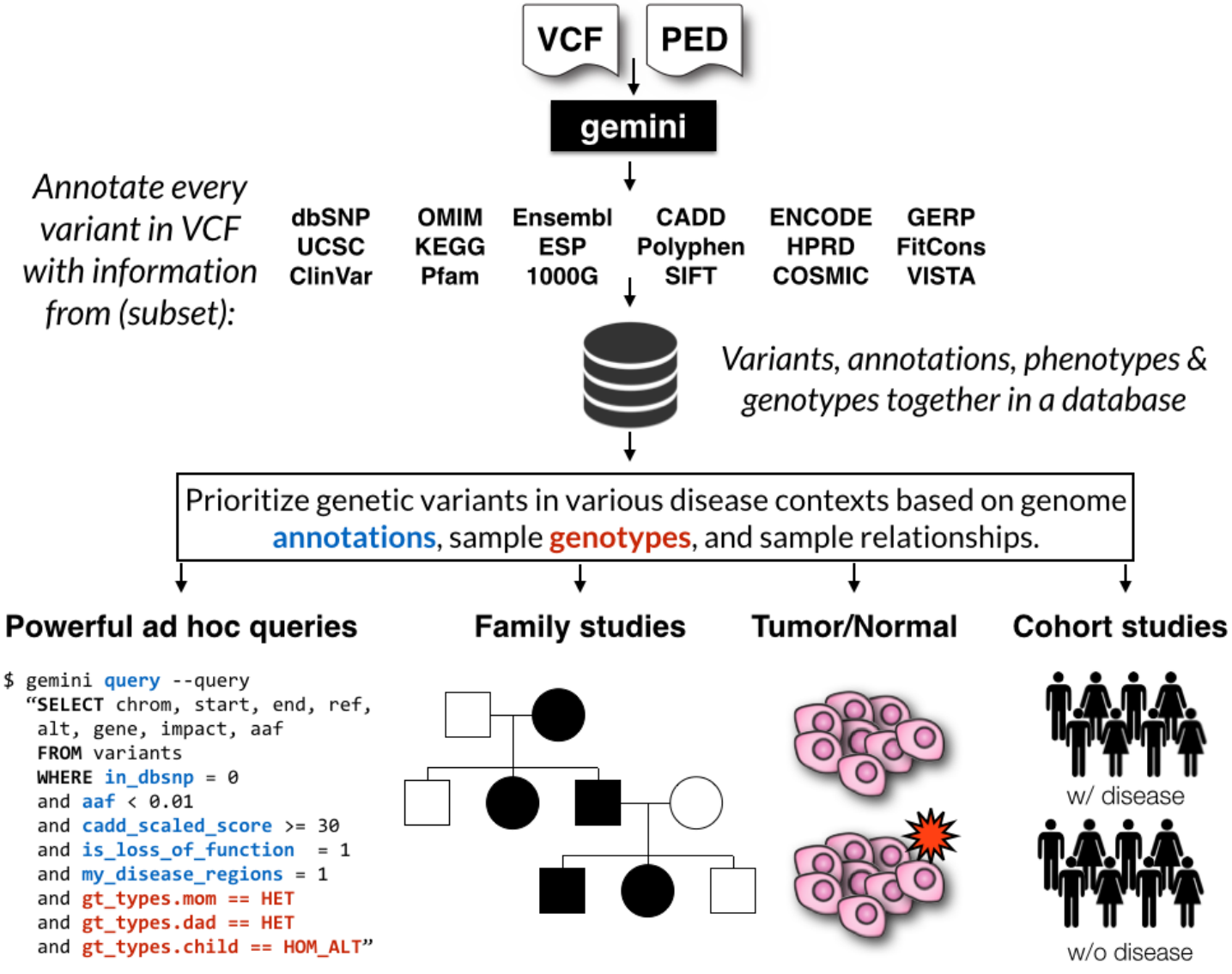
- snpEff

http://snpeff.sourceforge.net/

- Variant Effect Predictor (VEP) from Ensembl

http://www.ensembl.org/info/docs/tools/vep/index.html

Annotation and analysis – GEMINI

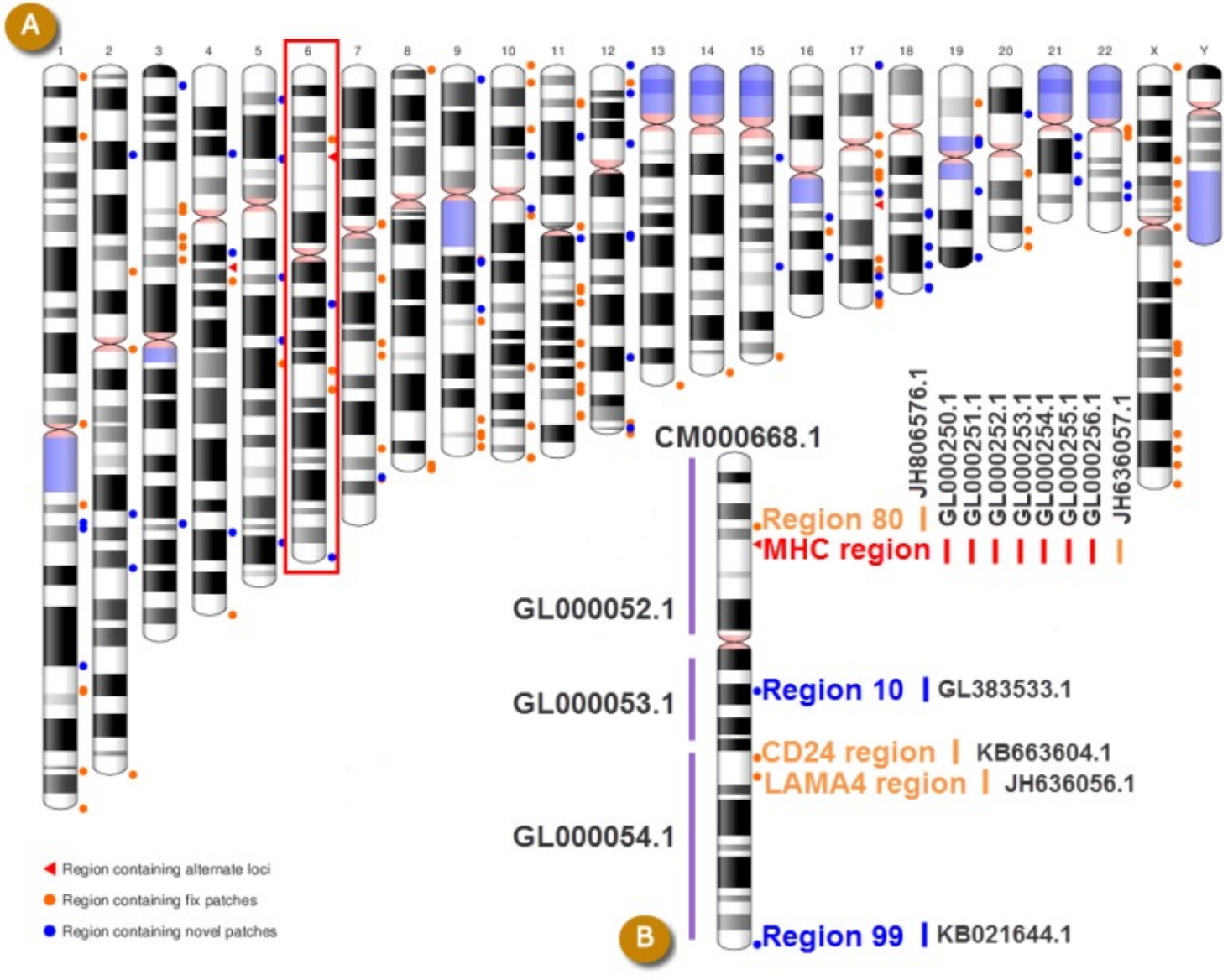


https://github.com/arq5x/gemini

Outline

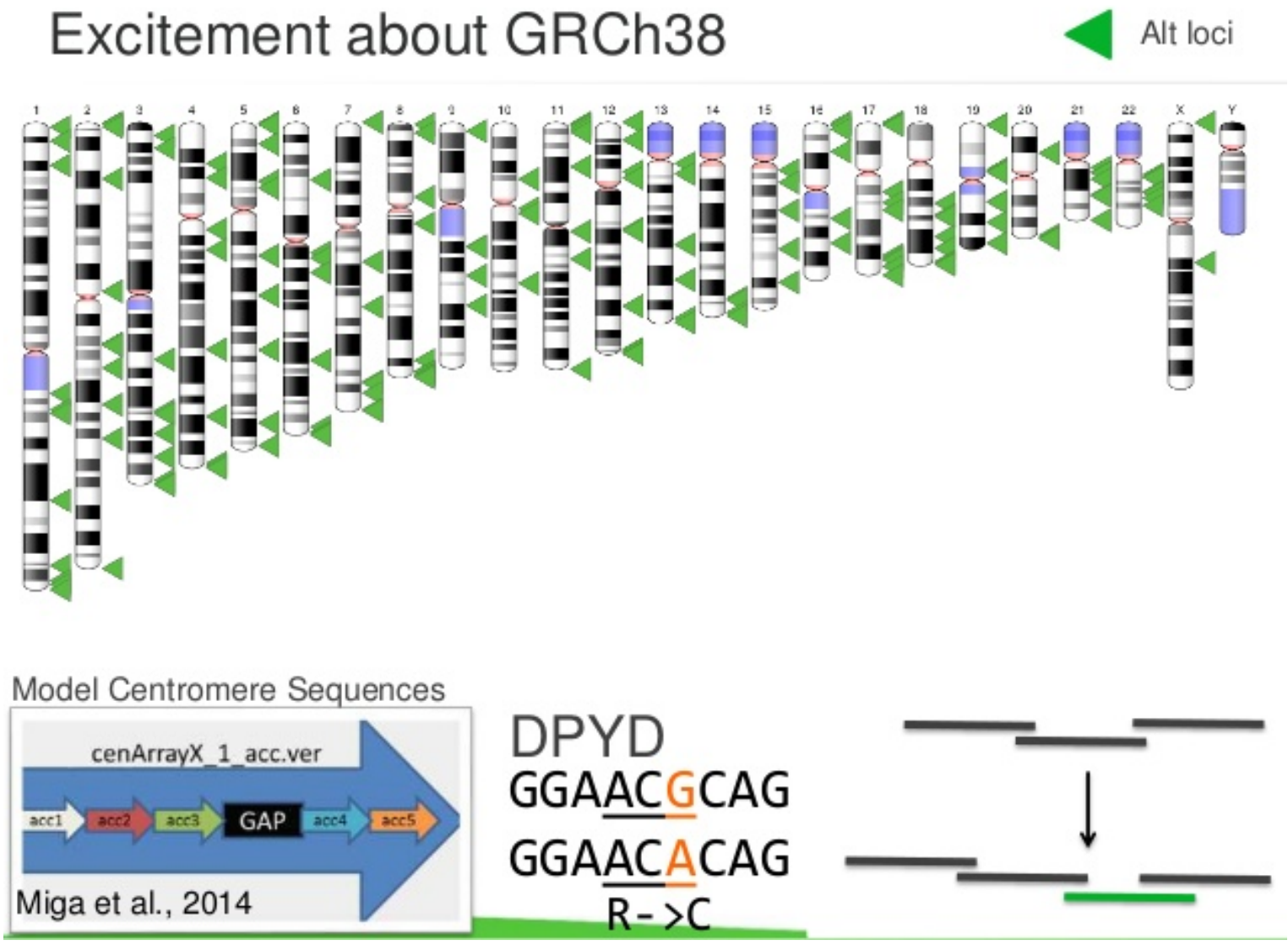
- Variant calling made easy
- Tools
- Validation
- Post-calling annotation
- **Genomes and graphs**
- Understanding outputs
- Automating everything – bcbio

Currently: GRCh37/hg19



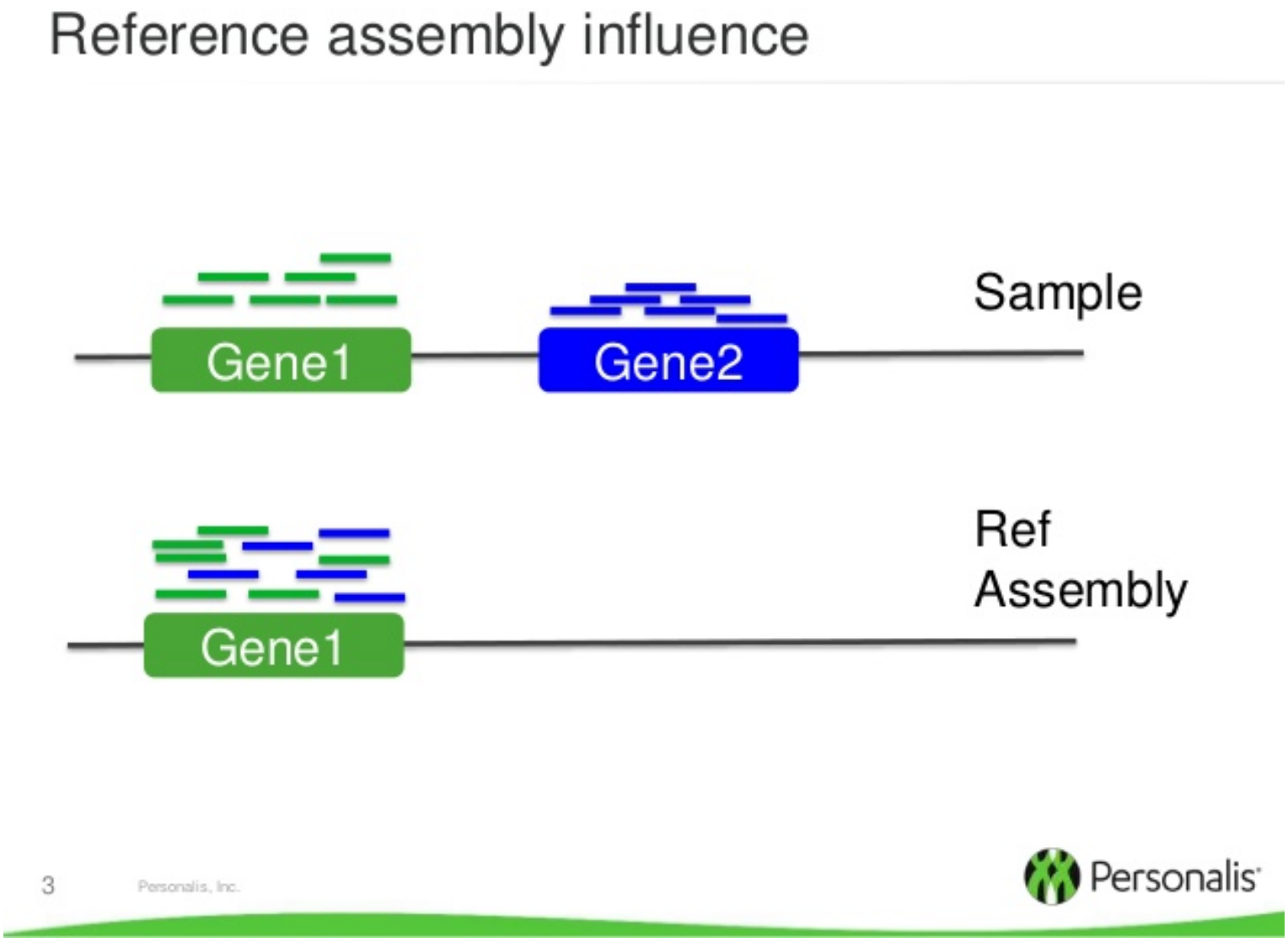
http://www.ncbi.nlm.nih.gov/books/NBK153600/?report=reader

GRCh38 – graph based, many more alternative loci



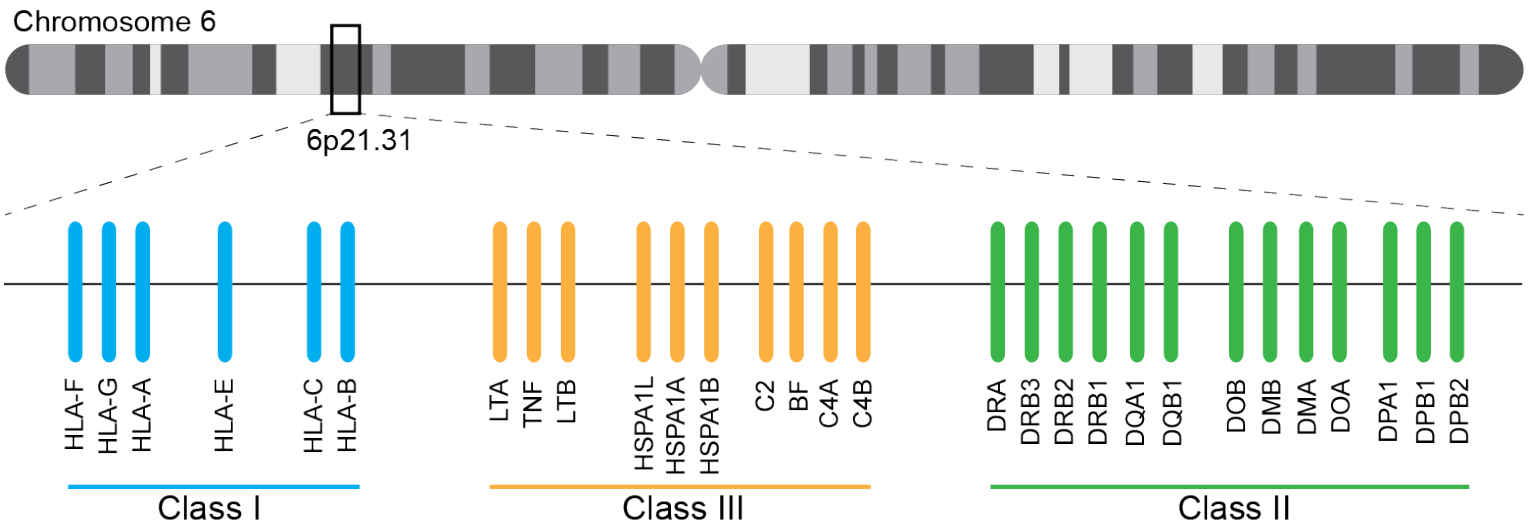
http://www.slideshare.net/GenomeRef/transitioning-to-grch38

GRCh38 – advantage for variant calling



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

Major histocompatibility complex (MHC) – HLAs



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

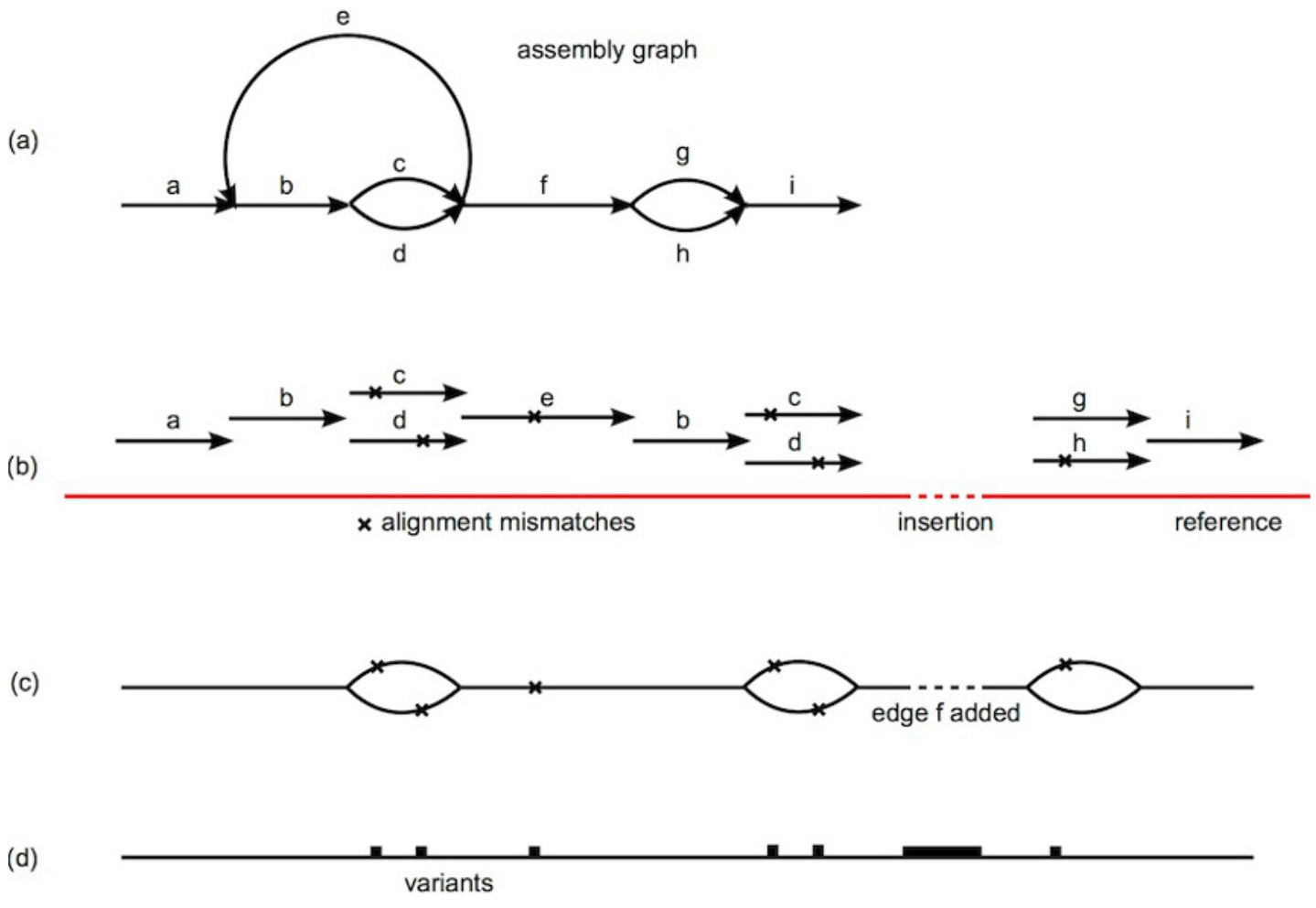
http://sciscogenetics.com/technology/human-leukocyte-antigen-complex/

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

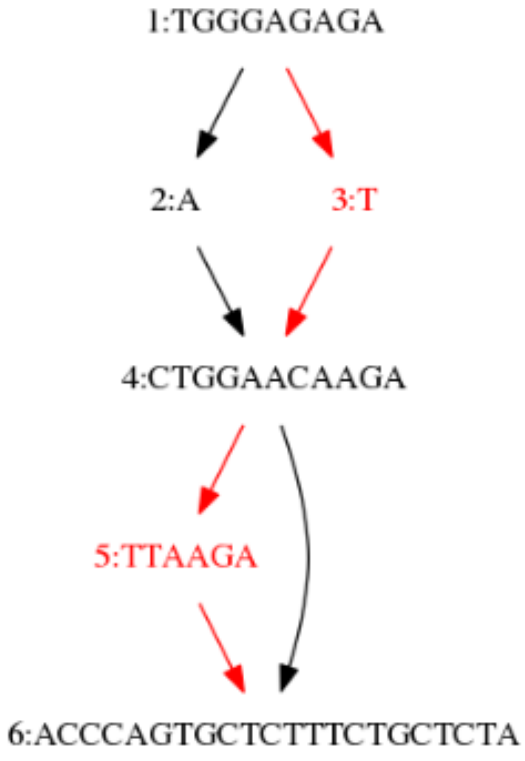
Genome graphs and variation



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
...			

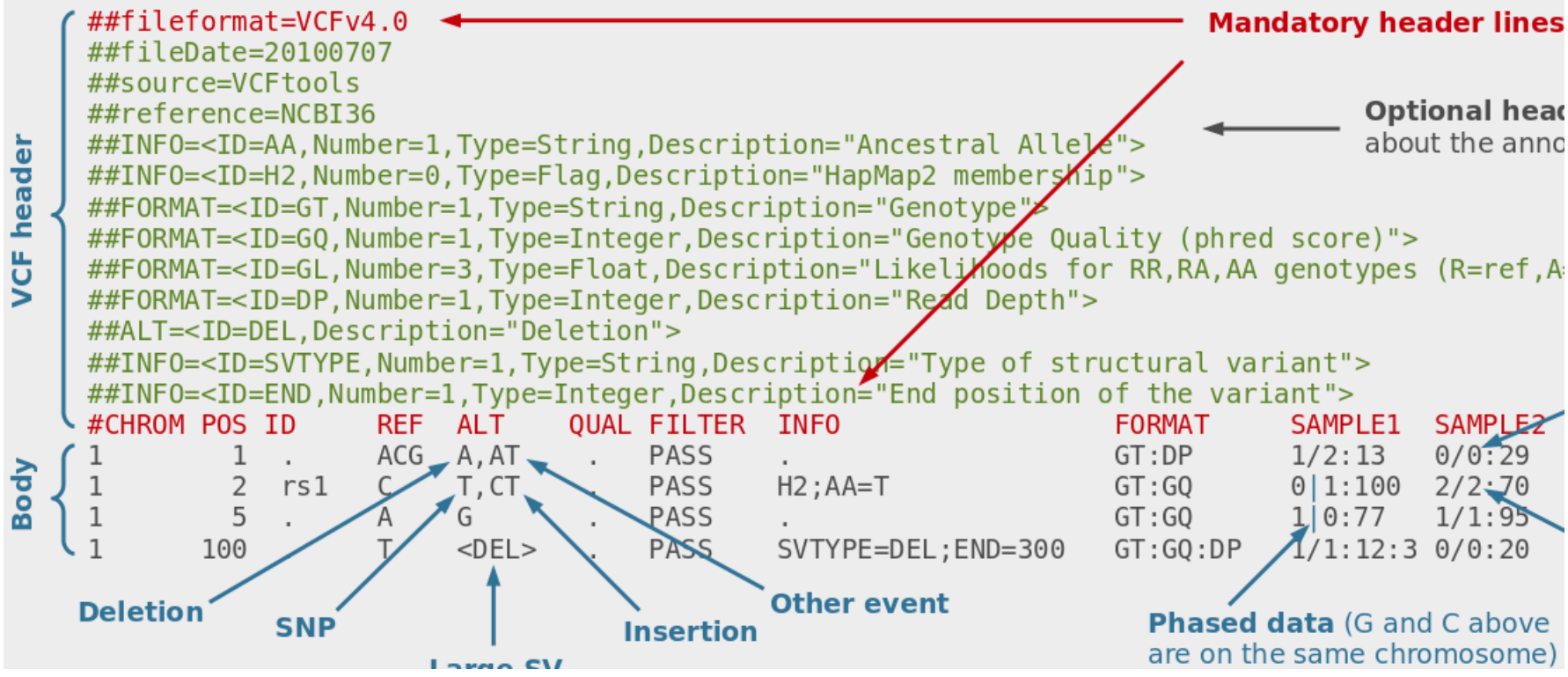


https://docs.google.com/presentation/d/1bbl2zY4qWQ0yYBHhoVuXb79HdgajRotIUa_VEn3kTpI

Outline

- Variant calling made easy
- Tools
- Validation
- Post-calling annotation
- Genomes and graphs
- **Understanding outputs**
- Automating everything – bcbio

VCF – overview



http://vcftools.sourceforge.net/VCF-poster.pdf

VCF – representations

Types of variants

SNPs

Alignment
```
ACGT
ATGT
```

VCF representation

POS	REF	ALT
2	C	T

Insertions

Alignment
```
AC-GT
ACTGT
```

VCF representation

POS	REF	ALT
2	C	CT

Deletions

Alignment
```
ACGT
A--T
```

VCF representation

POS	REF	ALT
1	ACG	A

Complex events

Alignment
```
ACGT
A-TT
```

VCF representation

POS	REF	ALT
1	ACG	AT

Large structural variants

VCF representation

POS	REF	ALT	INFO
100	T	<DEL>	SVTYPE=DEL;END=300

http://vcftools.sourceforge.net/VCF-poster.pdf

Learning to read VCFs

- Step by step guide from Broad

https://www.broadinstitute.org/gatk/guide/article?id=1268

- Specification

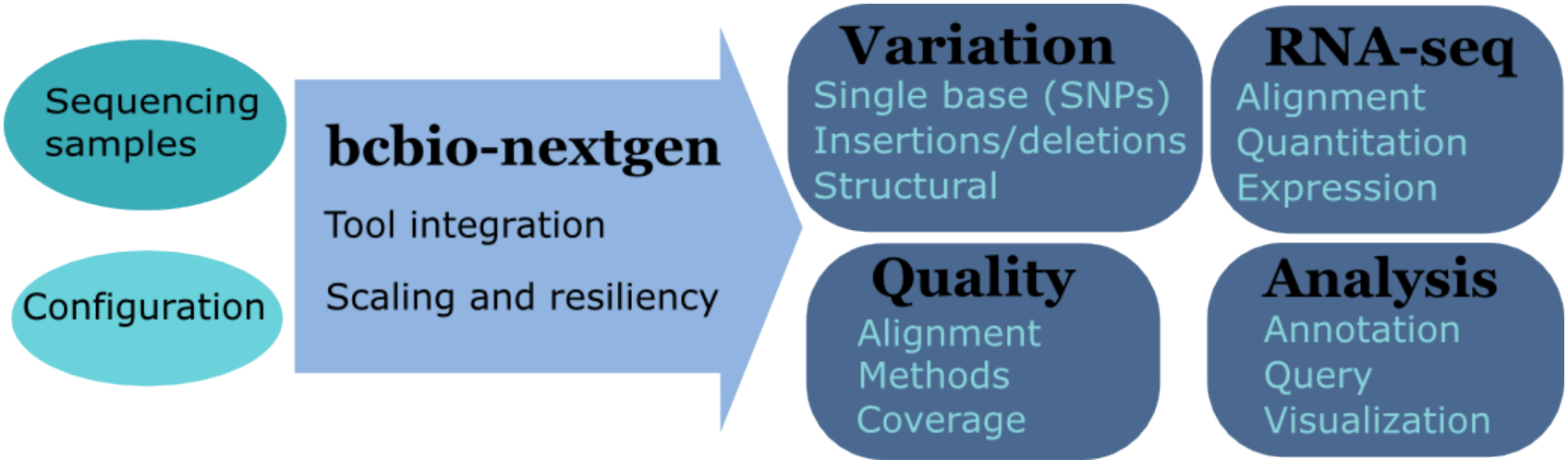
http://samtools.github.io/hts-specs/

Outline

- Variant calling made easy
- Tools
- Validation
- Post-calling annotation
- Genomes and graphs
- Understanding outputs
- **Automating everything – bcbio**

White box software

Overview



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

Uses

- Aligners: bwa-mem, novoalign, bowtie2
- Variantion: FreeBayes, GATK, VarDict, MuTecT, Scalpel, SnpEff, VEP, GEMINI, Lumpy, Delly, CNVkit
- RNA-seq: Tophat, STAR, cufflinks, HTSeq
- Quality control: fastqc, bamtools, RNA-SeQC
- Manipulation: bedtools, bcftools, biobambam, sambamba, samblaster, samtools, vcflib, vt

Provides

- Community – collected set of expertise
- Validation – outputs + automated evaluation
- Scaling
- Ready to run parallel processing on AWS
- Local installation of tools and data

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

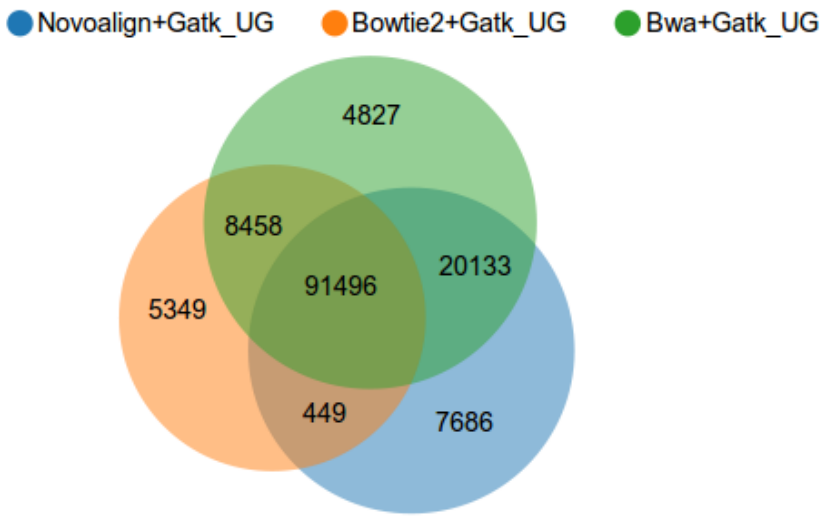
Quality differences between methods

Variant Calling Test

Discuss

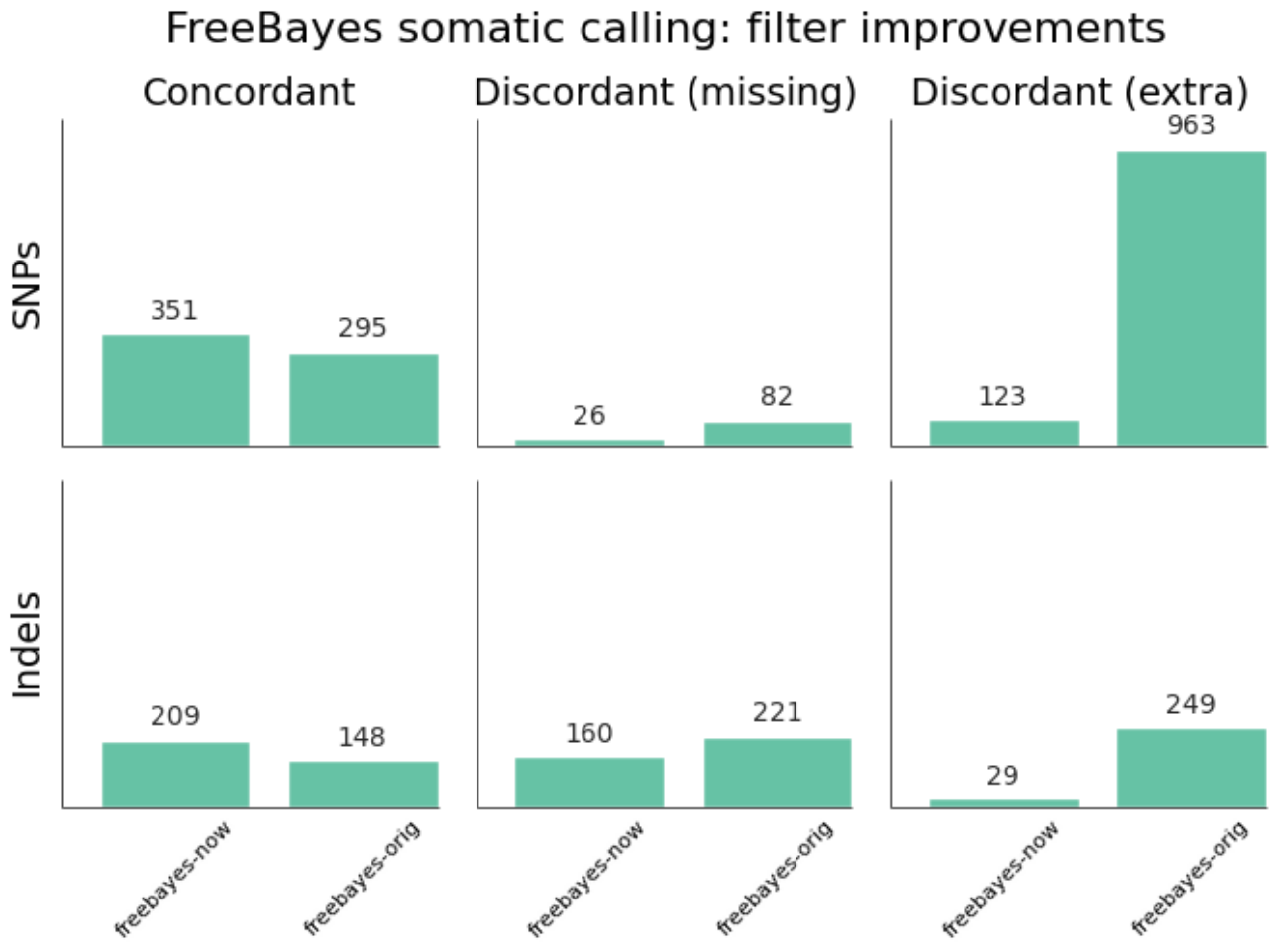
We compare combinations of variant calling pipelines across different data sets. Browse our public facing reports to see how various aligner + variant caller combinations perform against each other. Test your own combination of tools by creating your own report. Below is a sample conconcordance view on our "Illumina 100bp Paired End 30x Coverage" data set.

Variant Concordance - "illumina-100bp-pe-exome-30x"



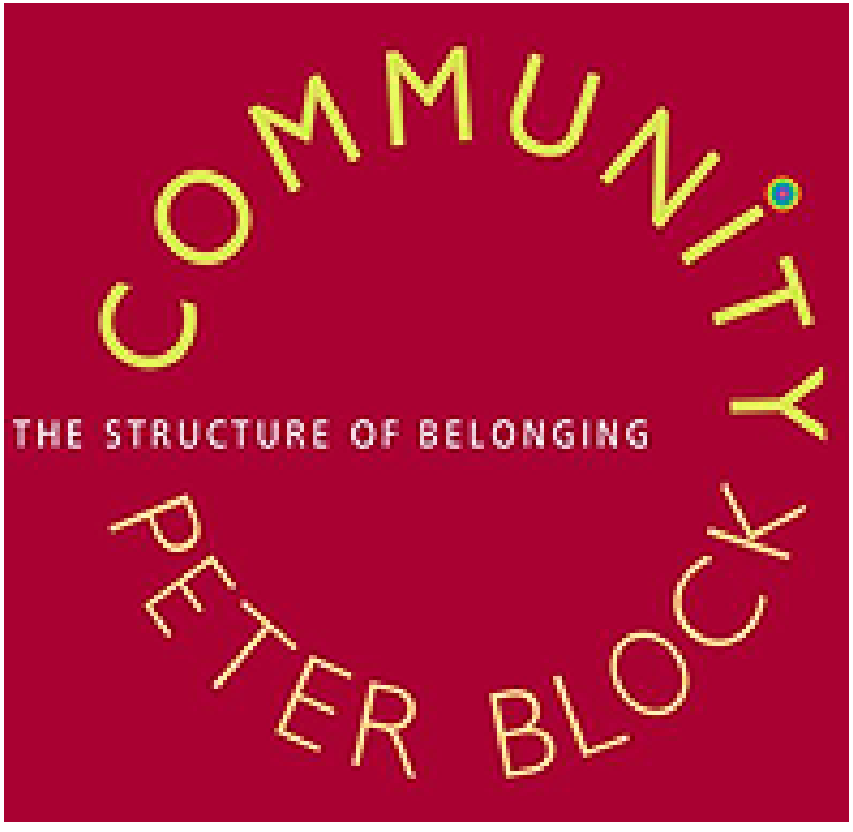
http://www.bioplanet.com/gcat

Benefits of improved filtering



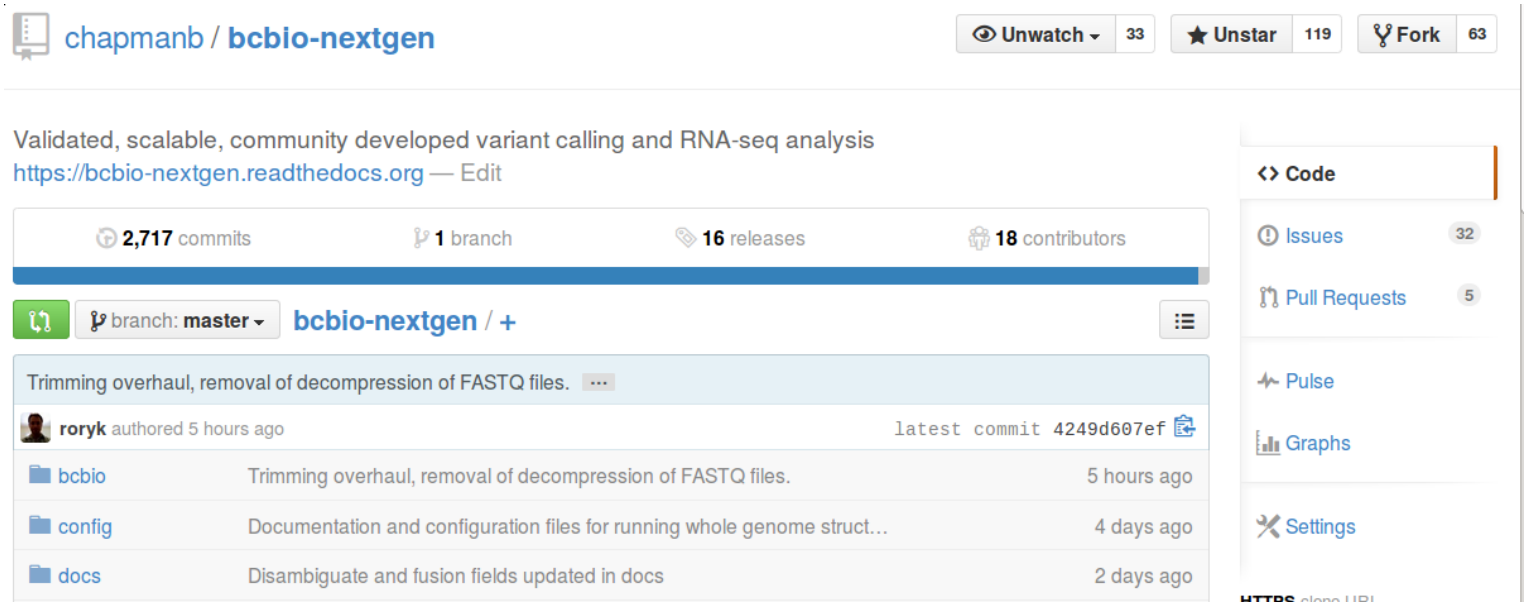
http://j.mp/cancervalpre

Solution



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

Community: contribution



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