Variant calling with validated community developed tools

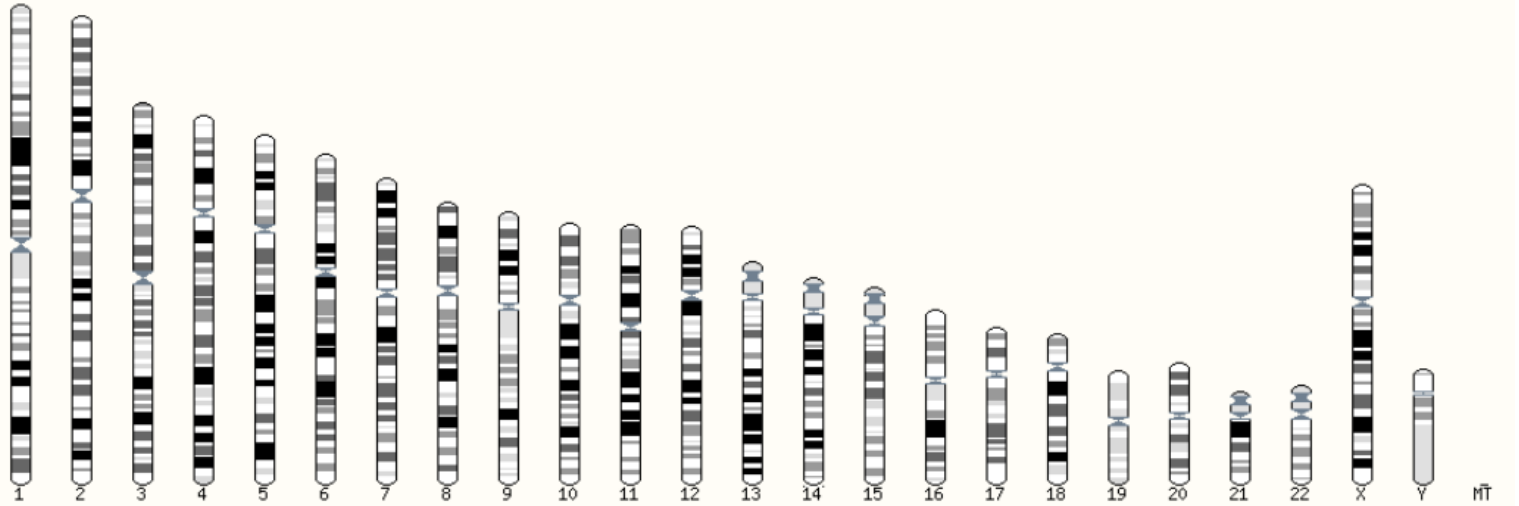
Brad Chapman
Bioinformatics Core, Harvard Chan School
http://j.mp/bcbiolinks

10 October 2018

Outline

- Overview of variant calling tools
- bcbio: open source, validated, community built
- Practical example: Personal Genome Project
- Cancer calling of low frequency variants

Human whole genome sequencing



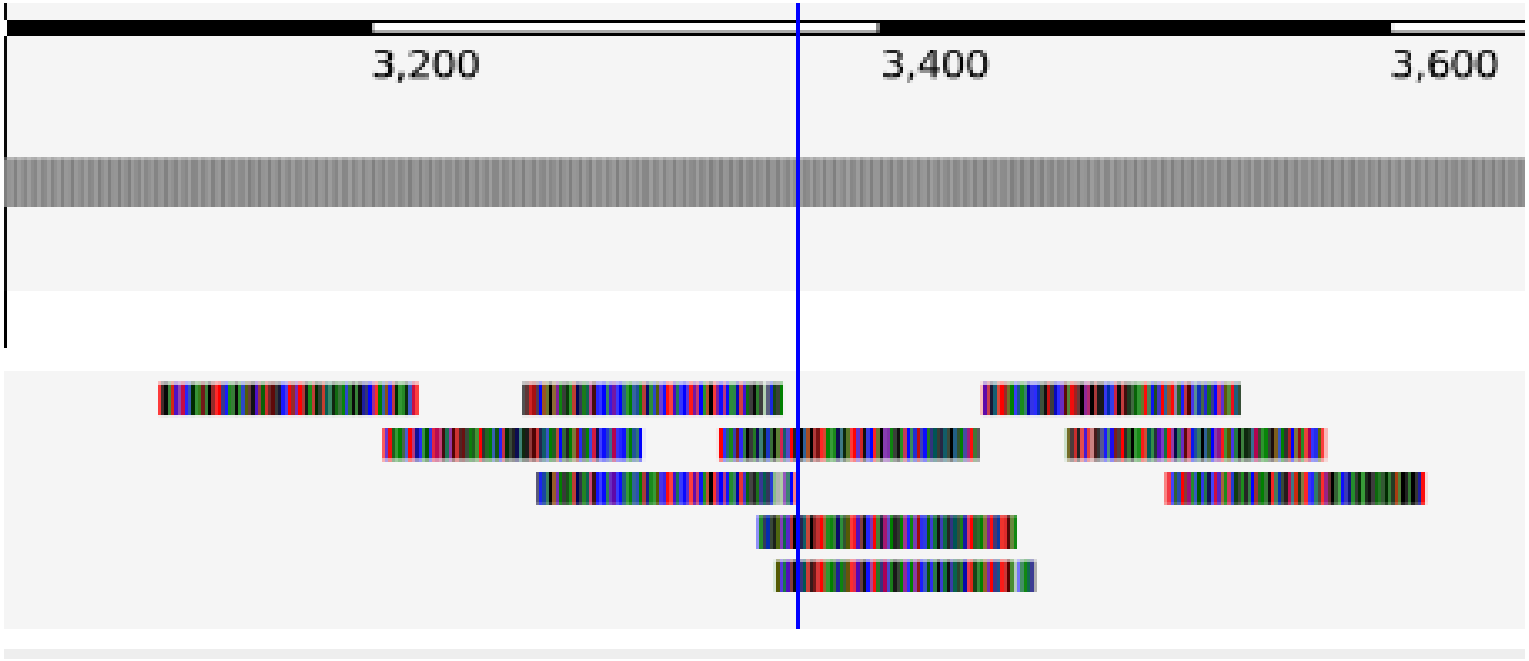
Click on the image above to jump to a chromosome, or click and drag to select a region

Summary

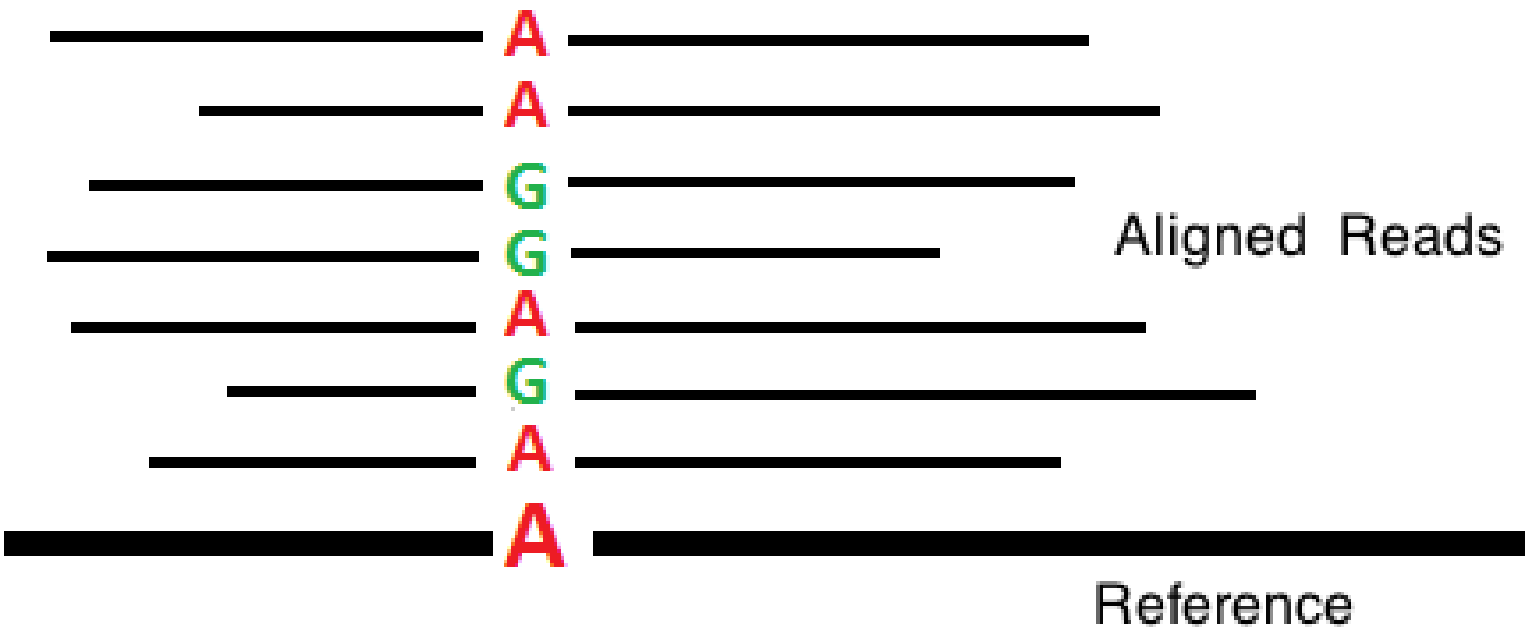
Assembly	GRCh37.p13 (Genome Reference Consortium Human Reference 37), INSDC Assembly GCA_000001405.14, Feb 2009
Database version	75.37
Base Pairs	3,326,743,047

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

High throughput sequencing

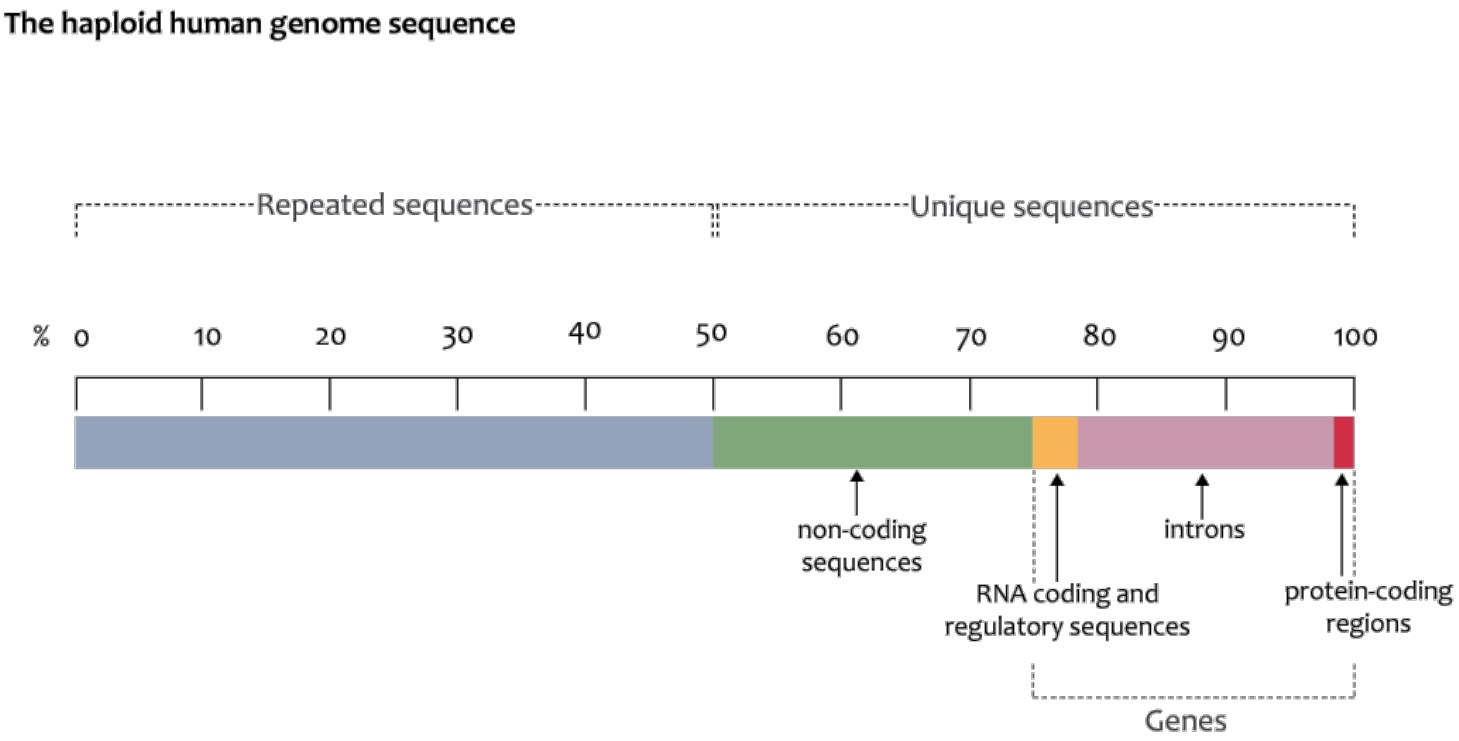


Variant calling



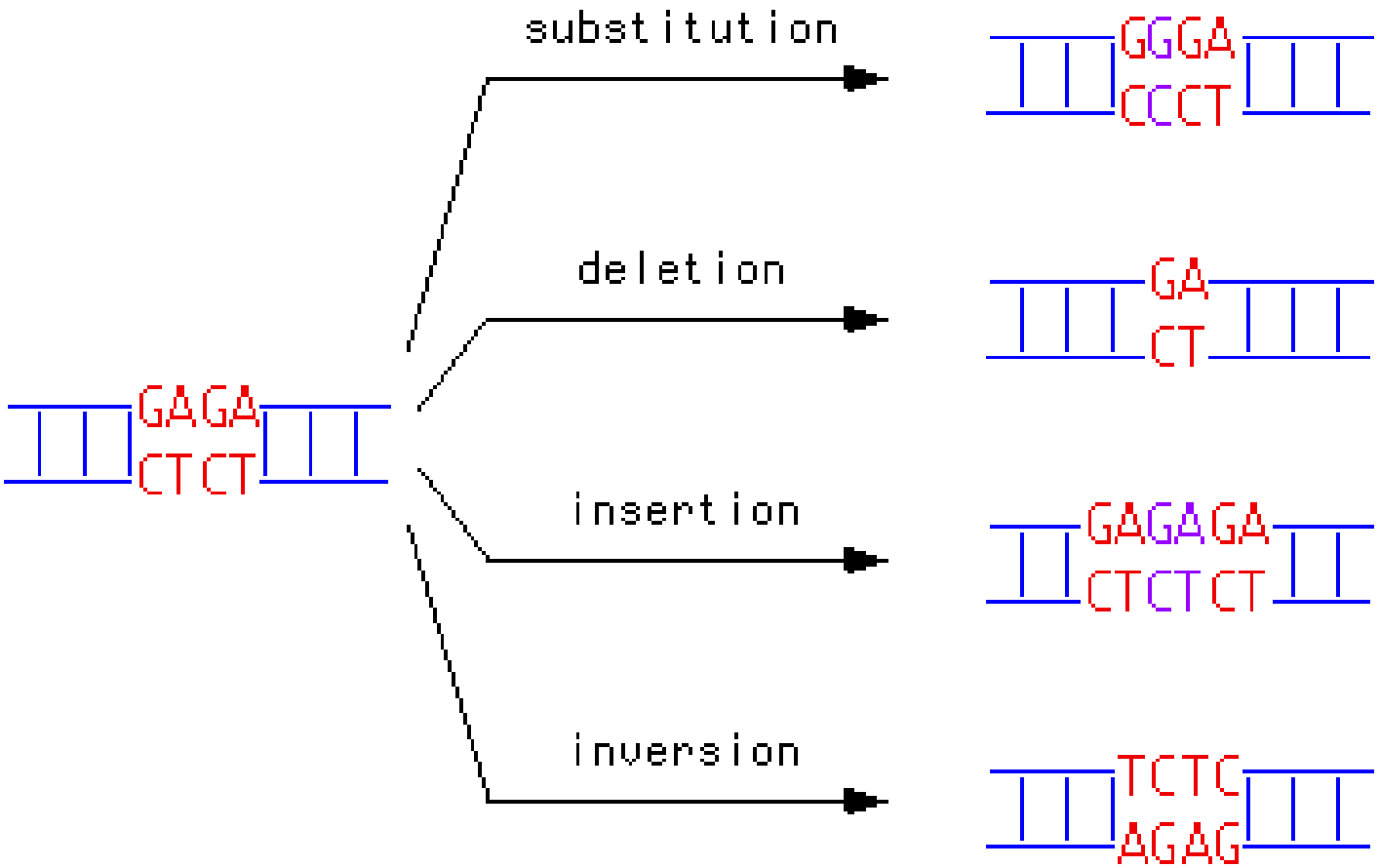
http://en.wikipedia.org/wiki/SNV_calling_from_NGS_data

Scale: exome to whole genome



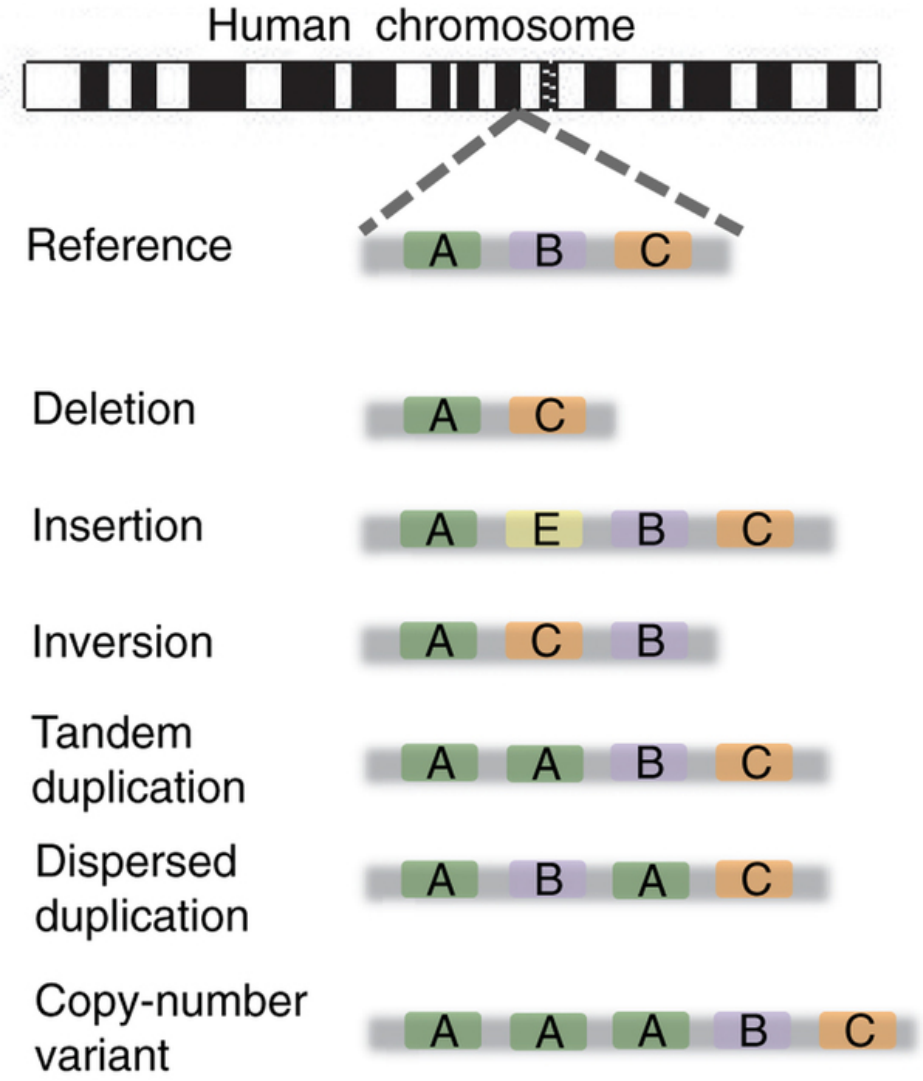
https://www.flickr.com/photos/119980645@N06/

SNPs and Indels



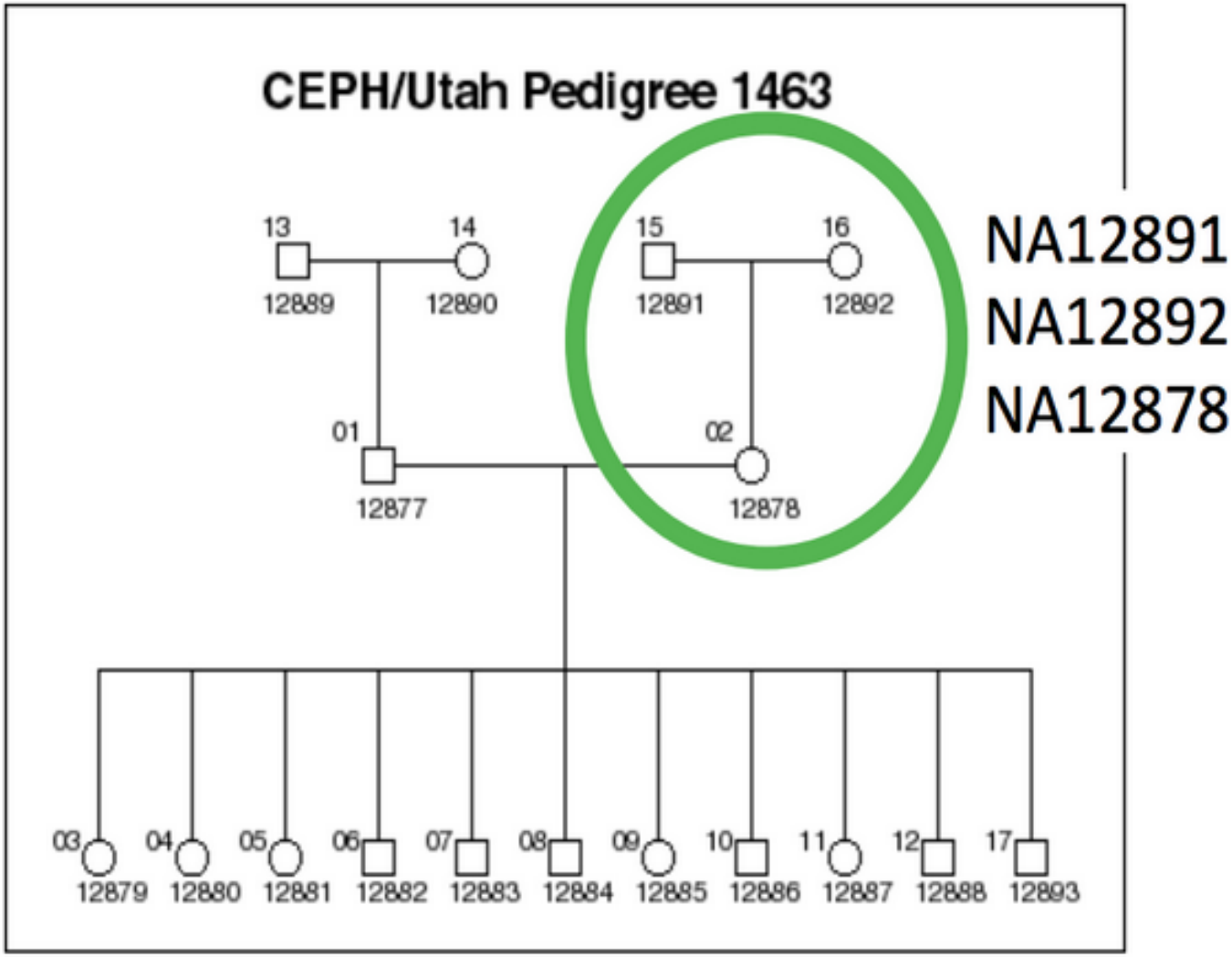
http://carolguze.com/text/442-2-mutations.shtml

Structural variations



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

Germline population calling



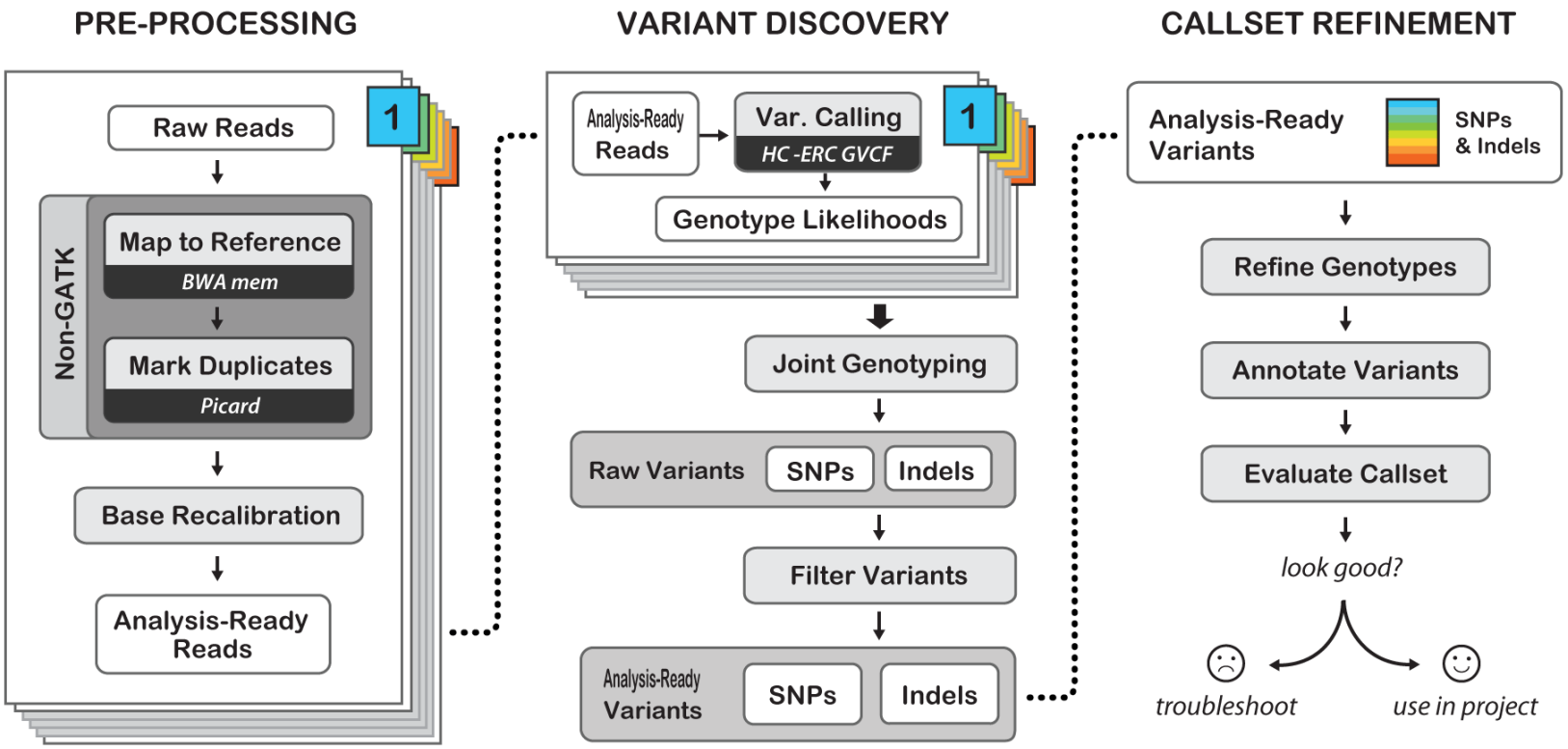
http://blog.goldenhelix.com/grudy/the-state-of-ngs-variant-calling-dont-panic/

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/

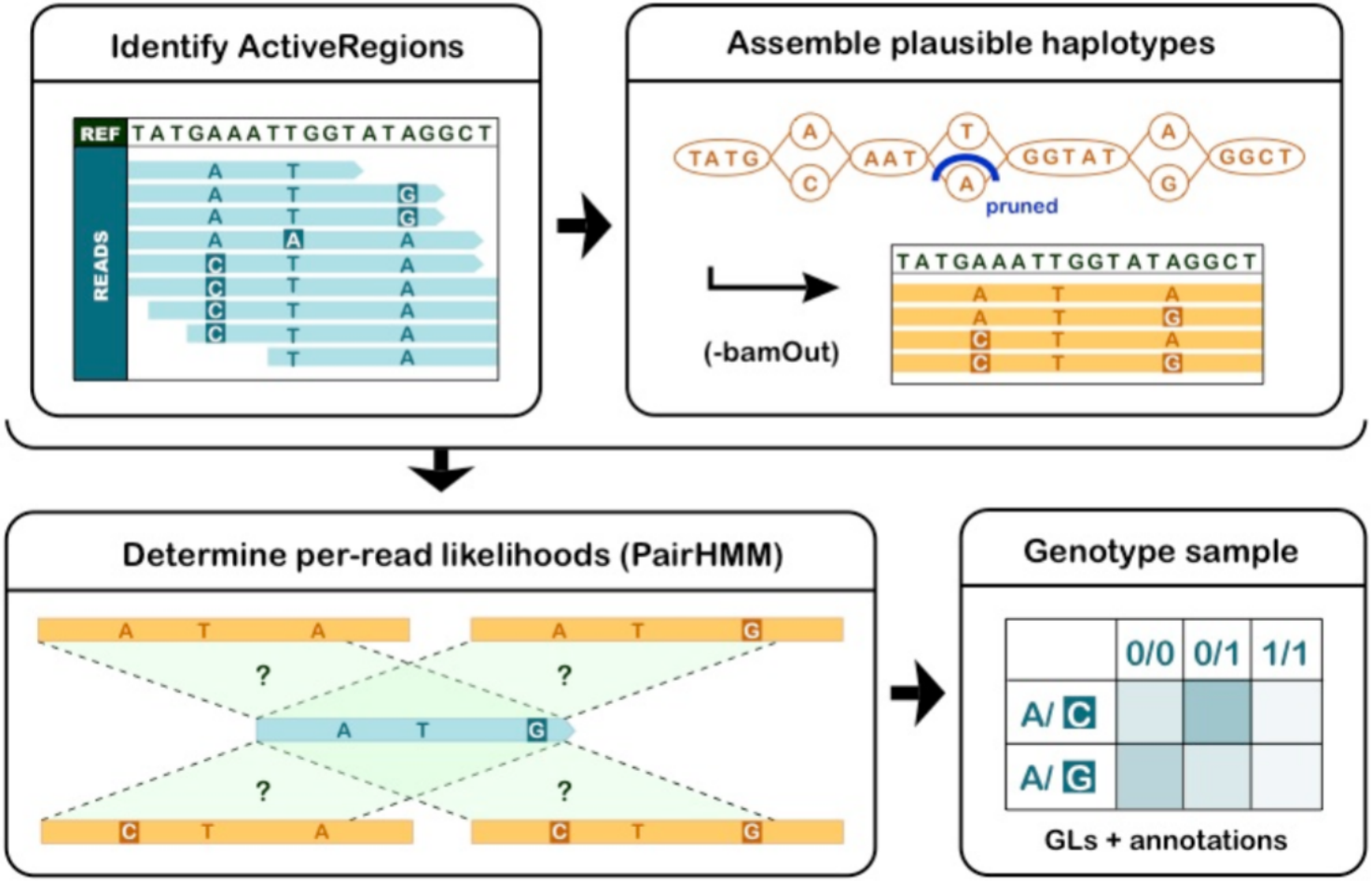
GATK Best Practices



Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

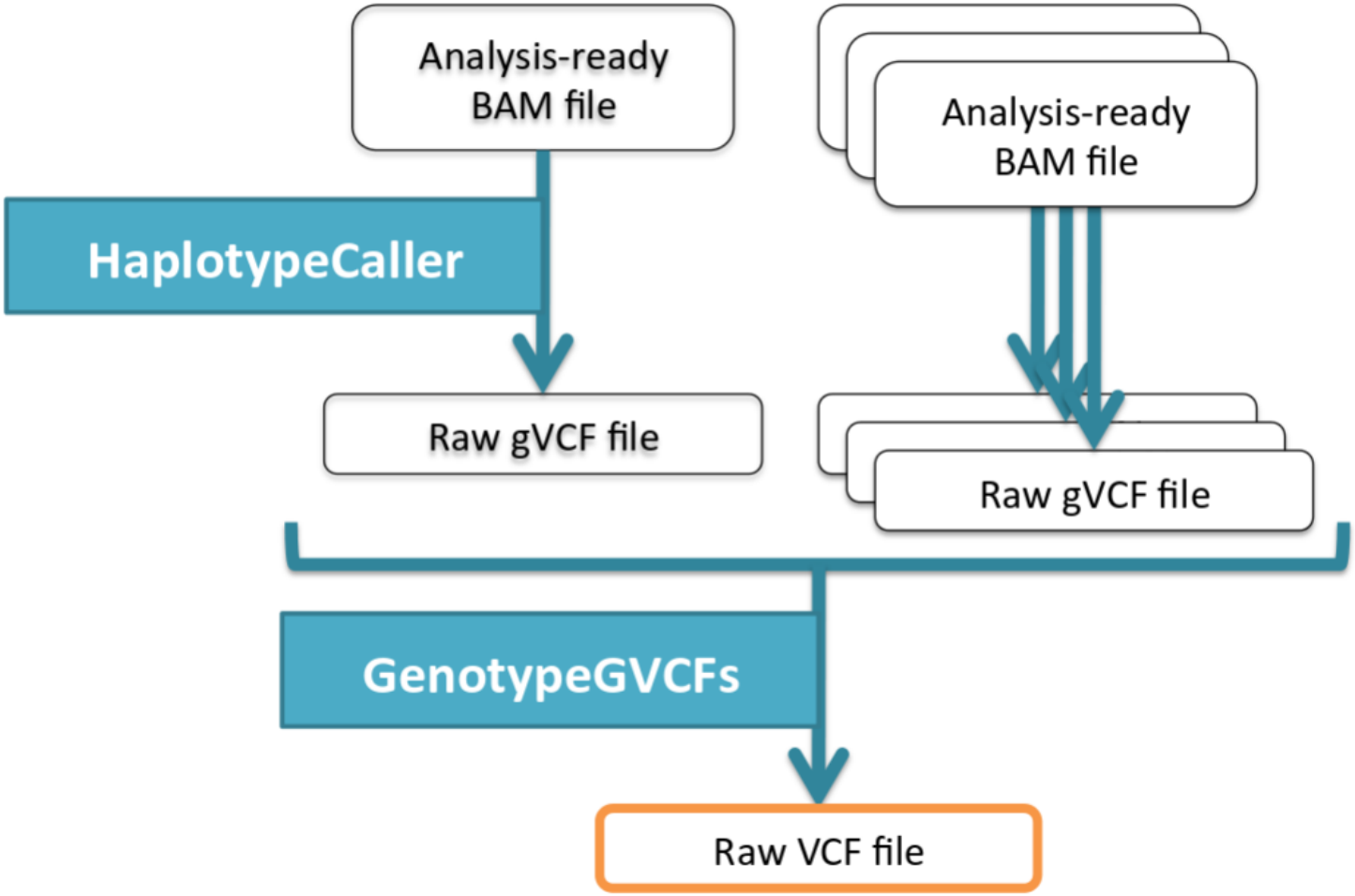
https://software.broadinstitute.org/gatk/best-practices/

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

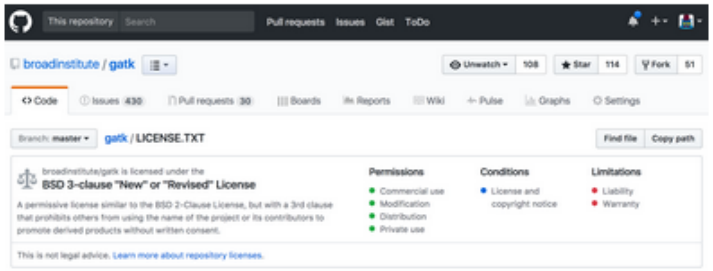
GATK4 now open source for all uses

GATK4 is completely open source

Posted by Geraldine_VdAuwera on 24 May 2017 (11)

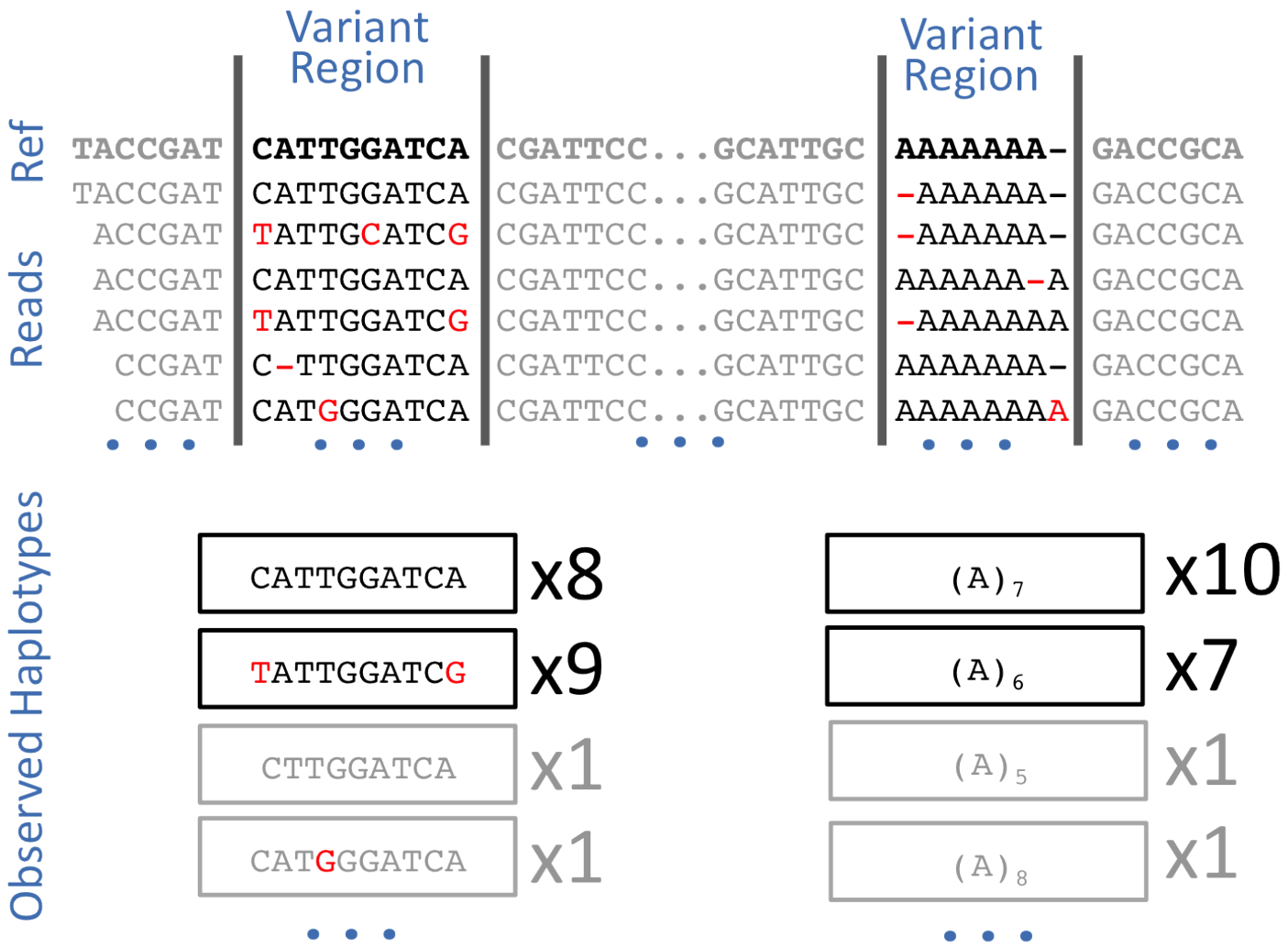
This is one of two posts announcing the imminent beta release of GATK4; for a technical description of features, see this other post.

"Wait, what?" Yes, you read that right, we're moving GATK4 to a fully open source license -- specifically, BSD 3-clause. And to be clear, this applies to all of GATK4. Not just the core framework (which, little known fact, has always been open source), but all the tools that were previously "protected", including HaplotypeCaller, the new CNV discovery tools, everything. The whole enchilada.



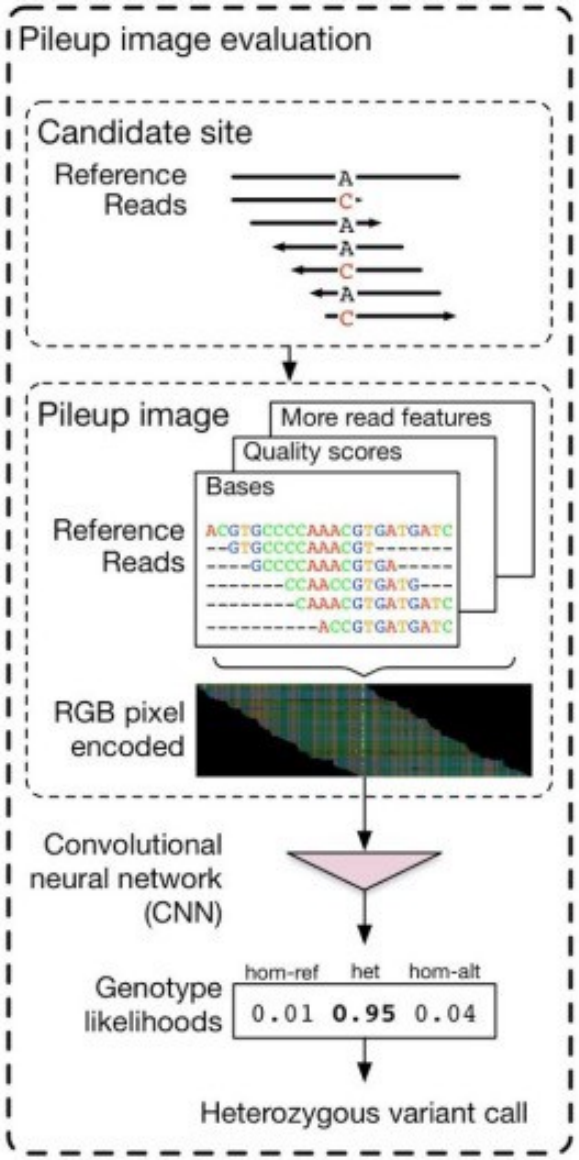
https://software.broadinstitute.org/gatk/blog?id=9645

FreeBayes



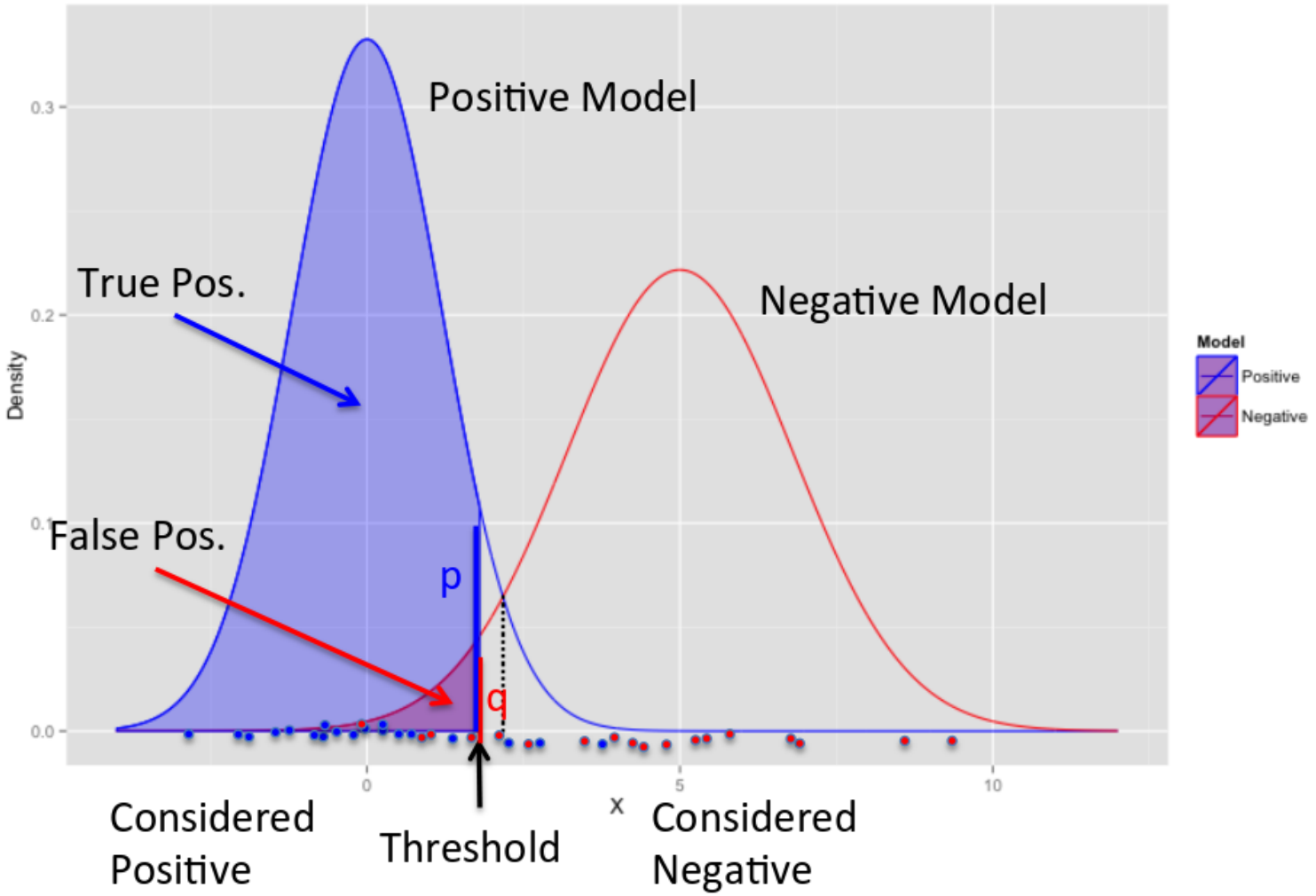
https://github.com/ekg/freebayes

DeepVariant



https://github.com/google/deepvariant

Filtering – Variant Quality Score Recalibration



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

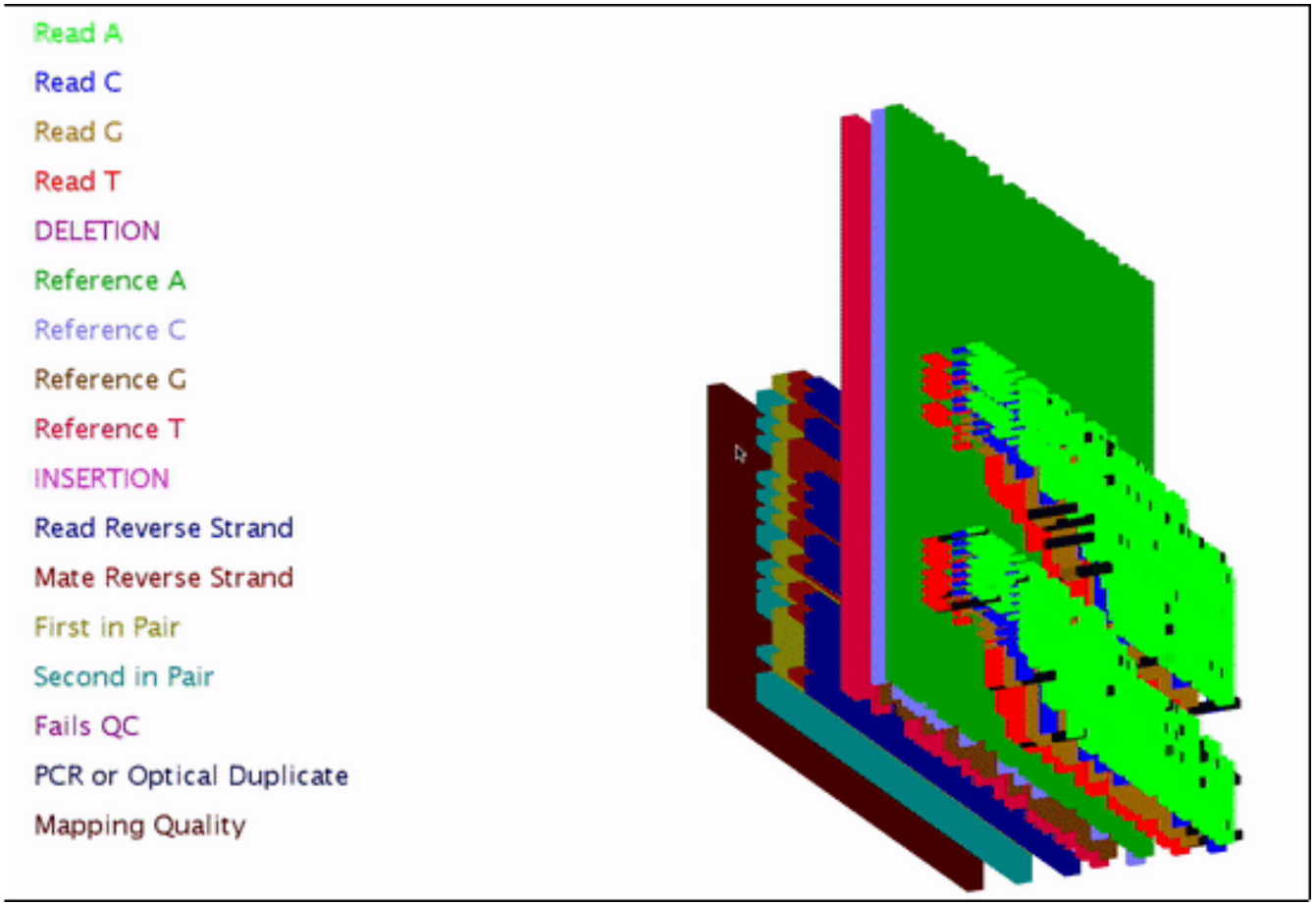
Filtering – hard cutoffs

```
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 – GATK and Deep Learning

Convolutional Neural Networks

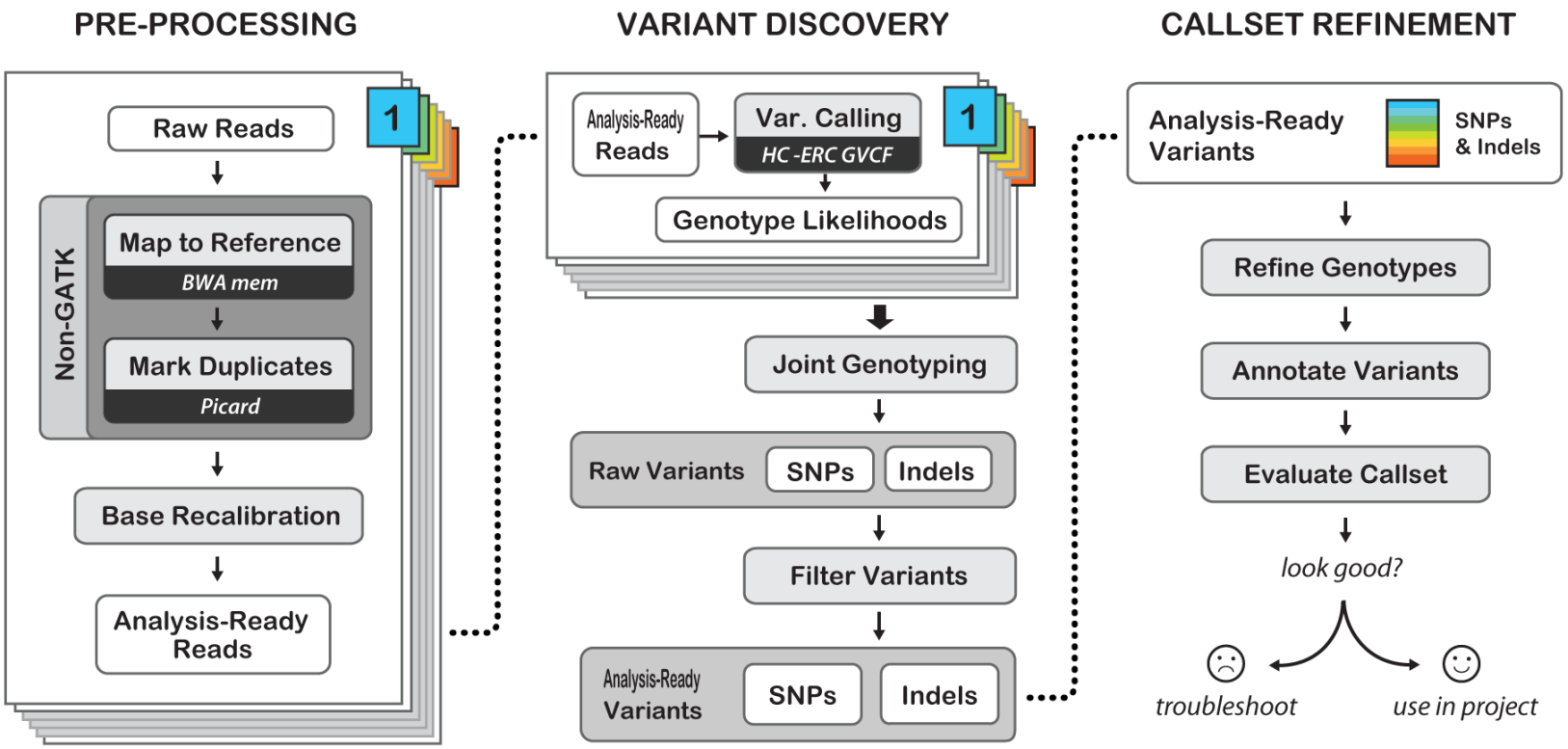


https://gatkforums.broadinstitute.org/gatk/discussion/10996/deep-learning-in-gatk4

Outline

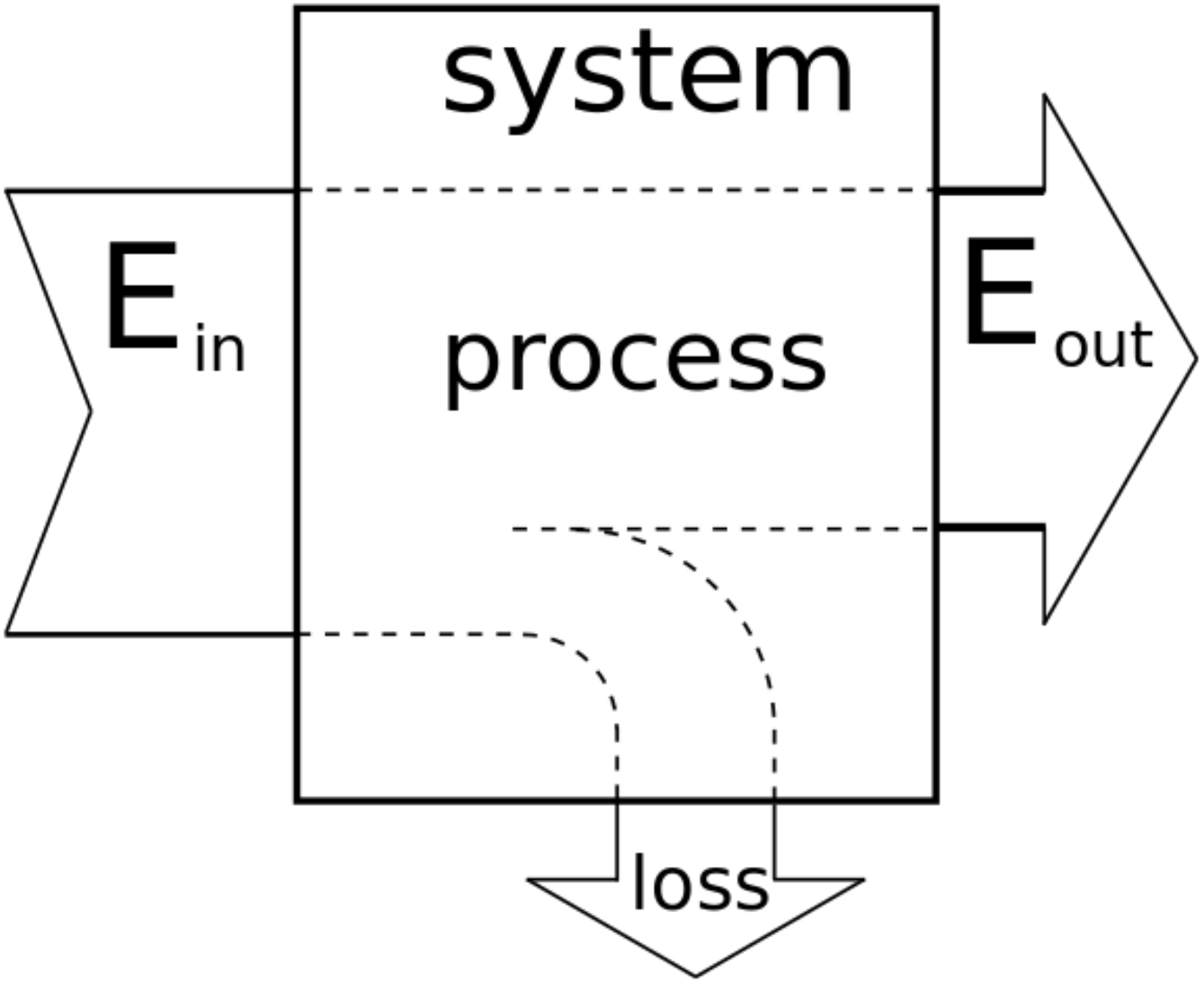
- Overview of variant calling tools
- **bcbio: open source, validated, community built**
- Practical example: Personal Genome Project
- Cancer calling of low frequency variants

You want to build a variant calling pipeline



Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

https://software.broadinstitute.org/gatk/best-practices/

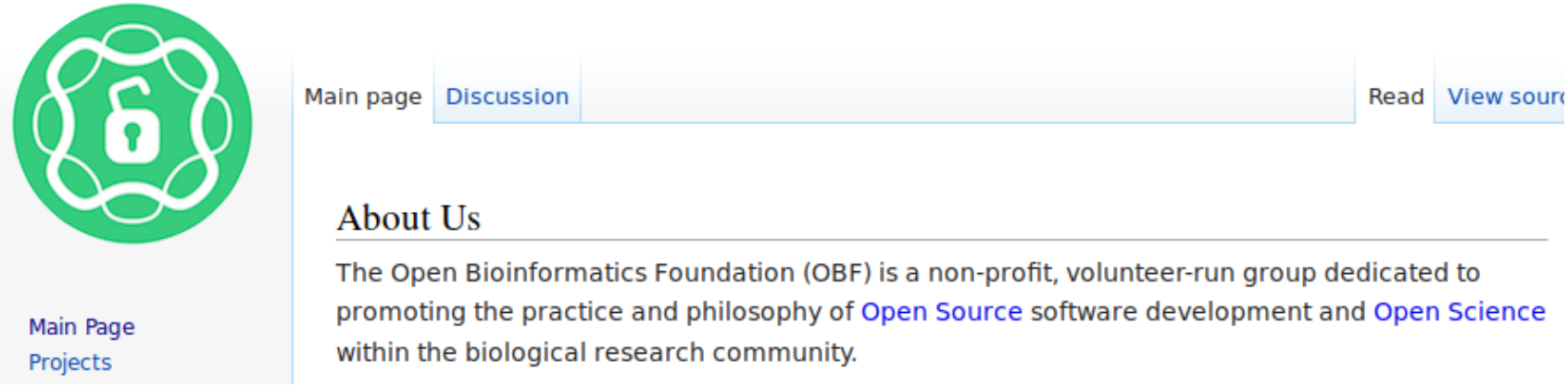


https://commons.wikimedia.org/wiki/File:Efficiency_diagram_by_Zureks.svg

Barriers to implementing yourself

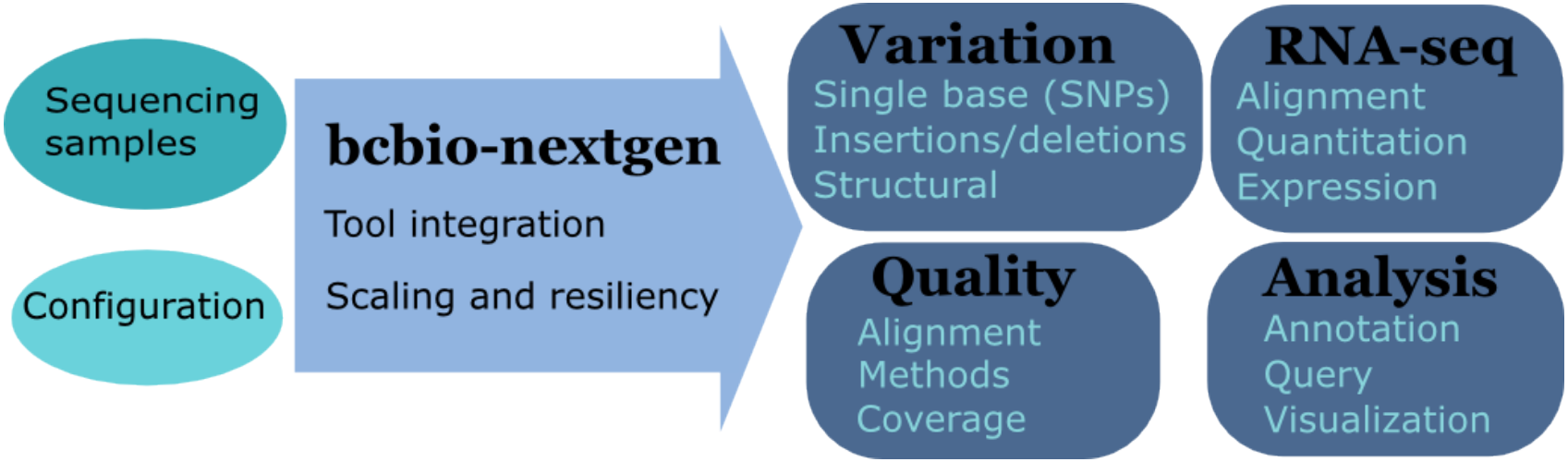
- Changing tools
- Feature support burden
- Multi-platform interoperability
- Validation

Build open source communities



https://www.open-bio.org

Overview



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

Supported analysis types

- ⊟ **Pipelines**
  - ⊟ **Germline variant calling**
    - Basic germline calling
    - Population calling
  - Cancer variant calling
  - Structural variant calling
  - RNA-seq
  - single-cell RNA-seq
  - smallRNA-seq
  - ChIP-seq

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

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

Best Practice Variant Detection with the GATK v2

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 Administrator, GSA Member admin

July 2012 edited February 4 in Methods and Workflows

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

GATK 4.0 will be released Jan 9, 2018

Posted by Geraldine_VdAuwera on 16 Oct 2017

Feature support burden

Table 1: Comparison of Nextflow with other workflow management systems

Workflow	Nextflow	Galaxy	Toil	Snakemake	Bpipe
Platform[a]	**Groovy/JVM**	**Python**	**Python**	**Python**	**Groovy/JVM**
Native task support[b]	Yes (any)	No	No	Yes (BASH only)	Yes (BASH only)
Common workflow language[c]	No	Yes	Yes	No	No
Streaming processing[d]	Yes	No	No	No	No
Dynamic branch evaluation	Yes	?	Yes	Yes	Undocumented
Code sharing integration[e]	Yes	No	No	No	No
Workflow modules[f]	No	Yes	Yes	Yes	Yes
Workflow versioning[g]	Yes	Yes	No	No	No
Automatic error failover[h]	Yes	No	Yes	No	No
Graphical user interface[i]	No	Yes	No	No	No
DAG rendering[j]	Yes	Yes	Yes	Yes	Yes
Container management					
Docker support[k]	Yes	Yes	Yes	No	No
Singularity support[l]	Yes	No	No	No	No
Multi-scale containers[m]	Yes	Yes	Yes	No	No
Built-in batch schedulers[n]					
Univa Grid Engine	Yes	Yes	Yes	Partial	Yes
PBS/Torque	Yes	Yes	No	Partial	Yes
LSF	Yes	Yes	No	Partial	Yes
SLURM	Yes	Yes	Yes	Partial	No
HTCondor	Yes	Yes	No	Partial	No
Built-in distributed cluster[o]					
Apache Ignite	Yes	No	No	No	No
Apache Spark	No	No	Yes	No	No
Kubernetes	Yes	No	No	No	No
Apache Mesos	No	No	Yes	No	No
Built-in cloud[p]					
AWS (Amazon Web Services)	Yes	Yes	Yes	No	No

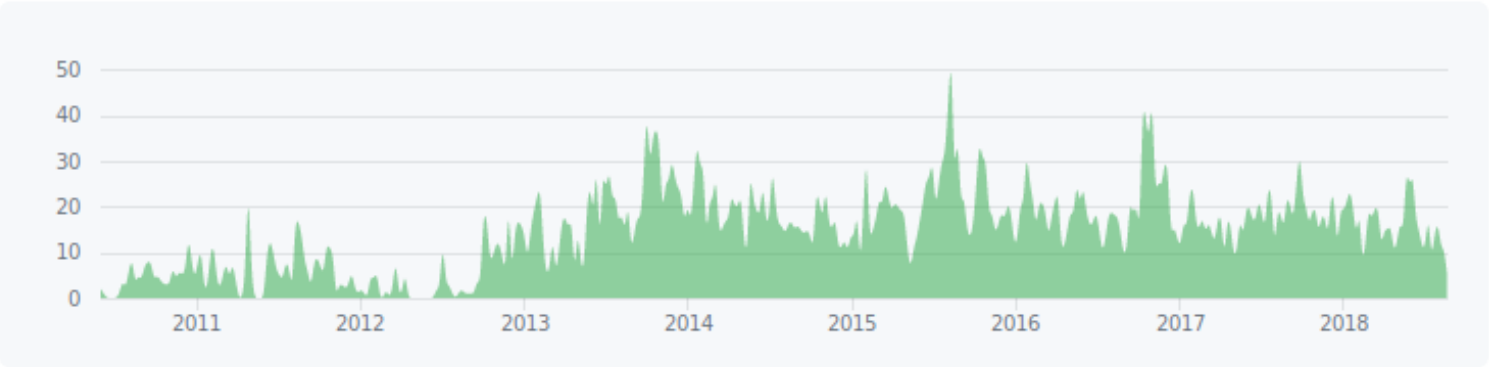
http://www.nature.com/nbt/journal/v35/n4/full/nbt.3820.html

Community: sustainability and support

Jul 18, 2010 – Oct 9, 2018

Contributions: **Commits**

Contributions to master, excluding merge commits



181 Open ✓ 1,935 Closed | Author | Labels | Projects | Milestones | Assignee | Sort

- **variant2 pipeline (via slurm) failing during align_prep** 3
  #2534 opened a day ago by jimmybgammyknee
- **PURPLE: purity & ploidy estimator support in bcbio** 5
  #2532 opened 5 days ago by pdiakumis
- **Support for lossy CRAM in bcbio WGrS variant calling pipeline** 6
  #2531 opened 5 days ago by WimSpee
- **smallrnaseq fastqc step failure** 3
  #2525 opened 14 days ago by mshadbolt

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

Infrastructure Goals

- Local machines
- Clusters: SLURM, SGE, Torque, PBS, LSF
- Clouds: Amazon, Google, Azure
- Clinical environments
- User interface for researchers
- Integrate with LIMS
- Accessible to the general public

Mike Lin Retweeted

DNAnexus, Inc. @dnanexus · 13 Jun 2013
#BigData Parking: "There's no reason **to move data** outside the #cloud. You can do **analysis** right there." ow.ly/m14Ke **#genomics**

Stuart Watt @morungos · 4 Mar 2014
Big upcoming change in **genomics**: **data** sets are now too large **to** download for **analysis**. **Move** code **to** the **data**, not vice versa #ibcretreat2014

Rob Schaefer @CSciBio · Jul 17
huge problem: moving **analysis** to the data, not the other way around. **@ewanbirney** #ISAG2017 #BigData

Aaron Quinlan
@aaronquinlan

Following

This is the only way genomic research can scale.

Javier Quilez @jaquol
Laura Clarke: do not download the data, bring the analysis to the data @laurastephen #gi2017

6:54 PM · 1 Nov 2017

Why do we transfer data around?

- Lots of work to setup and configure an analysis
- Hard to port scalable analysis to new environment

Many great workflow systems: Nexflow

```
#!/usr/bin/env nextflow

cheers=Channel.from "Bonjour","Ciao","Hello","Hola"

process sayHello {
  input:
  val x from cheers

  """
  echo $x world!
  """
}
```

Nextflow

Data-driven computational pipelines

Nextflow enables scalable and reproducible scientific workflows using software containers. It allows the adaptation of pipelines written in the most common scripting languages.

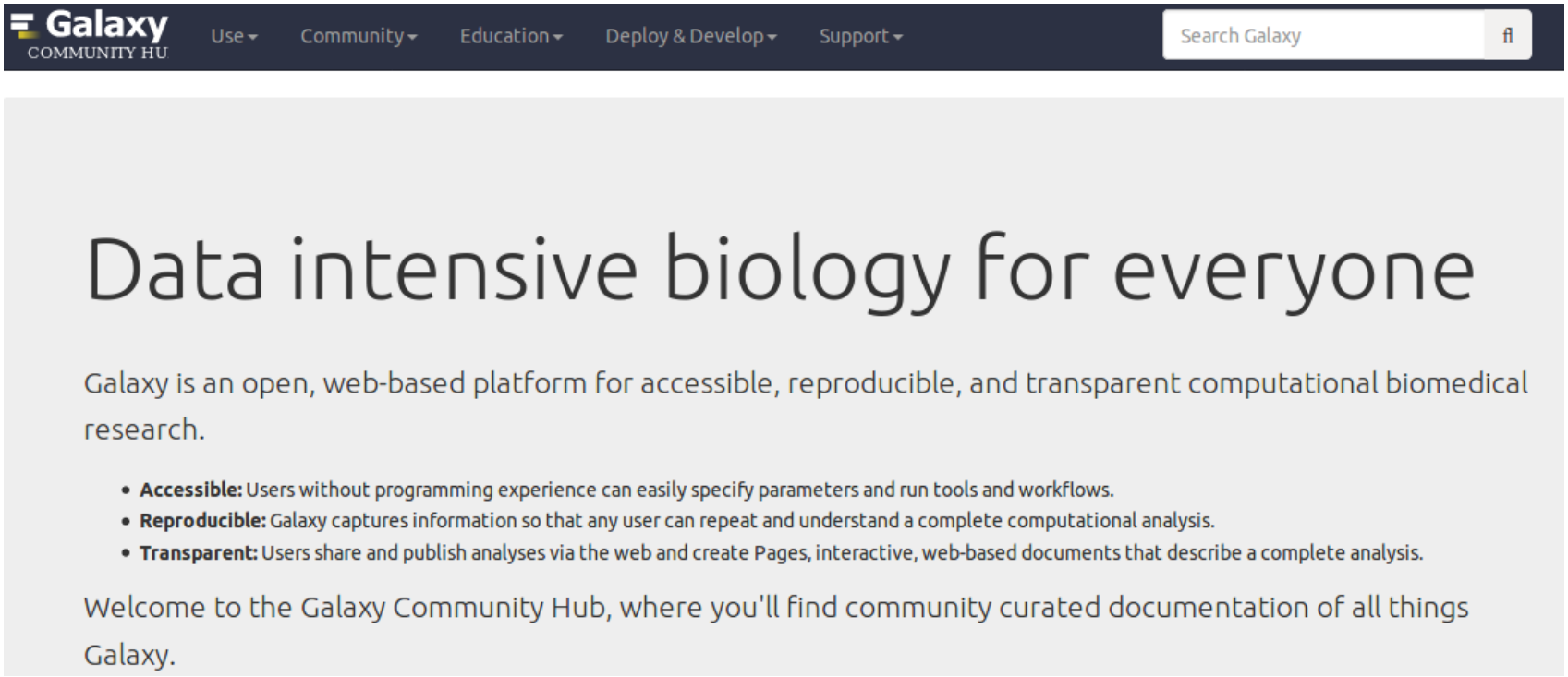
Its fluent DSL simplifies the implementation and the deployment of complex parallel and reactive workflows on clouds and clusters.

Find out more

Zero config | Polyglot | Concurrency | Scale easily

http://nextflow.io/

Many great workflow systems: Galaxy



http://galaxyproject.org/

Many great workflow systems: Snakemake

Snakemake Tutorial

This tutorial introduces the text-based workflow system Snakemake. Snakemake follows the GNU Make paradigm: workflows are defined in terms of rules that define how to create output files from input files. Dependencies between the rules are determined automatically, creating a DAG (directed acyclic graph) of jobs that can be automatically parallelized.

Snakemake sets itself apart from existing text-based workflow systems in the following way. Hooking into the Python interpreter, Snakemake offers a definition language that is an extension of Python with syntax to define rules and workflow specific properties. This allows to combine the flexibility of a plain scripting language with a pythonic workflow definition. The Python language is

https://snakemake.readthedocs.io

But, many workflow systems

Existing Workflow systems

Michael R. Crusoe edited this page 8 hours ago · 141 revisions

Computational Data Analysis Workflow Systems

An incomplete list

176. Reflow: a language and runtime for distributed, integrated data processing in the cloud https://github.com/grailbio/reflow
177. Resolwe: an open source dataflow package for Django framework https://github.com/genialis/resolwe
178. Yahoo! Pipes (historical) https://en.wikipedia.org/wiki/Yahoo!_Pipes
179. Walrus https://github.com/fjukstad/walrus
180. Apache Beam https://beam.apache.org/
181. CLOSHA https://closha.kobic.re.kr/ https://www.bioexpress.re.kr/go_tutorial http://docplayer.net/19700397-Closha-manual-ver1-1-kobic-korean-bioinformation-center-kogun82-kribb-re-kr-2016-05-08-bioinformatics-workflow-management-system-in-bio-express.html

https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems

We'll never agree on one system

- Advantages and disadvantages to each
- Familiarity and teaching
- Personal preference

So we can't easily share workflows

- Single workflow system allows coordinated groups
- Create barrier to sharing externally
- Hard to mix and match components between workflow environments
- How can we do better?

Better abstractions = more interoperability

https://bcbio-nextgen.readthedocs.io/en/latest/contents/cwl.html

Common Workflow Language (CWL)

Workflow	pipeline-se-narrow.cwl		
Sub-workflow 1	01-qc-se.cwl		
Step 1	extract.cwl	extract.py	
Step 2	count.cwl	count.py	
Step 3	fastqc.cwl	fastqc	
Sub-workflow 2	02-trim.cwl		
...			

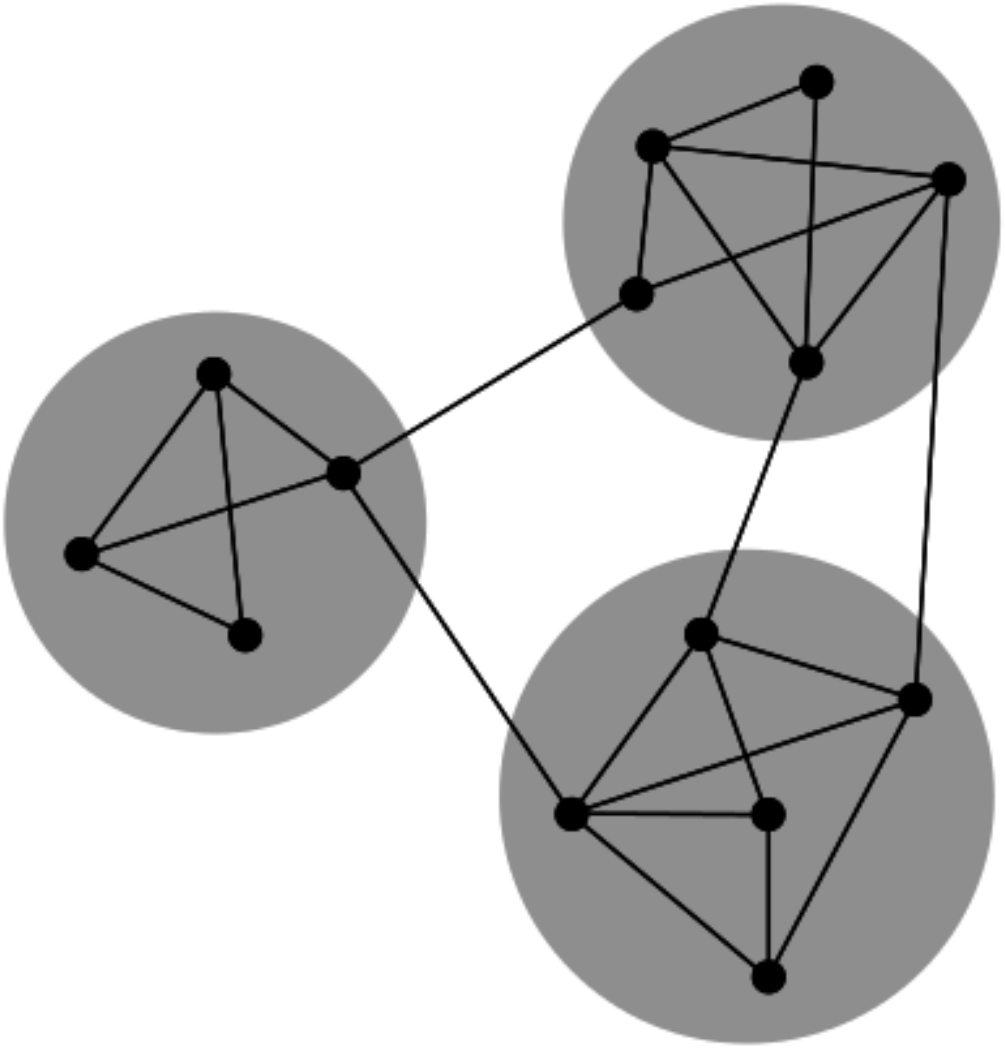
http://www.commonwl.org/

https://f1000research.com/slides/5-1617

Why use a workflow abstraction?

- Integrate with multiple platforms
  - Cromwell – HPC, Cloud, local
  - Arvados
  - DNAnexus
  - Seven Bridges + Cancer Genomics Cloud
- Stop maintaining bcbio specific infrastructure
- Focus on hard biological problems

Connections



By jham3 - Own work, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=17125894

Outline

- Overview of variant calling tools
- bcbio: open source, validated, community built
- **Practical example: Personal Genome Project**
- Cancer calling of low frequency variants

CWL in bcbio

- Start with high level configuration file
- Generate CWL
- Run, on any infrastructure that supports CWL
  - Generated CWL
  - Docker or local bcbio installation
  - Genome data

https://bcbio-nextgen.readthedocs.io/en/latest/contents/cwl.html

Practical example: Personal Genome Project

The Personal Genome Project

The Personal Genome Project, initiated in 2005, is a vision and coalition of projects across the world dedicated to creating public genome, health, and trait data. Sharing data is critical to scientific progress, but has been hampered by traditional research practices. The PGP approach is to invite willing participants to publicly share their personal data for the greater good.

http://www.personalgenomes.org/us

Whole genome sequencing data plus metadata

Public Profile -- huD57BBF

Real Name
James L Vick

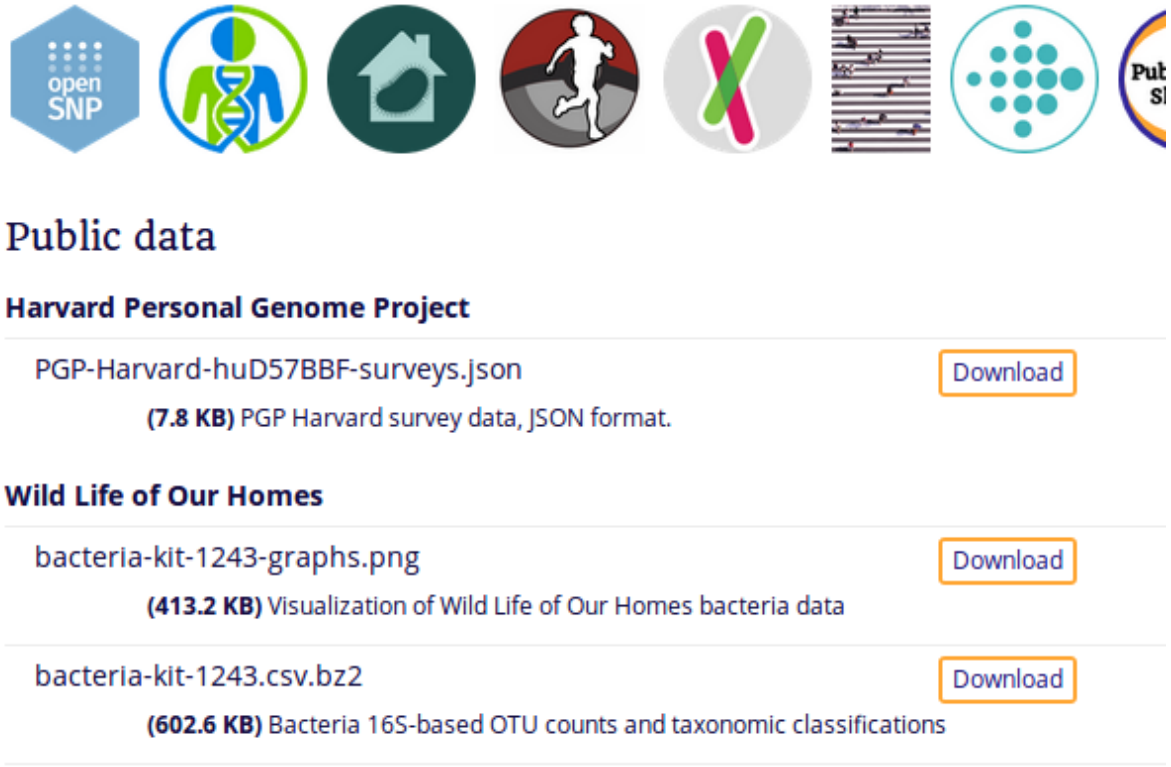
Personal Health Records

Demographic Information

Date of Birth	1949-04-30 (69 years old)
Gender	Male
Weight	165lbs (75kg)
Height	5ft 10in (177cm)
Blood Type	O+
Race	White

https://my.pgp-hms.org/profile/huD57BBF

Rich set of associated data



https://www.openhumans.org/member/jameslvick/

Template: describe your analysis

```
- files: huD57BBF.bam
  description: huD57BBF
  analysis: variant
  genome_build: hg38
  algorithm:
    aligner: bwa
    variantcaller: gatk-haplotype
    svcaller: [manta, lumpy, cnvkit]
    hlacaller: optitype
```

https://github.com/bcbio/bcbio_validation_workflows

Local filesystem environment

```
local:
  ref: biodata/collections
  inputs:
    - biodata/regions
    - biodata/pgp
resources:
  default:
    cores: 8
    memory: 3500M
    jvm_opts: [-Xms750m, -Xmx3500m]
```

Equivalent on a remote platform

```
arvados:
  reference: su92l-4zz18-3p00f79y4p535ia
  input: [su92l-4zz18-ihm3wrgyuwcmsx1]
resources:
  default: {cores: 16, memory: 3500M,
            jvm_opts: [-Xms1g, -Xmx3500m]}
```

Arvados pipeline run

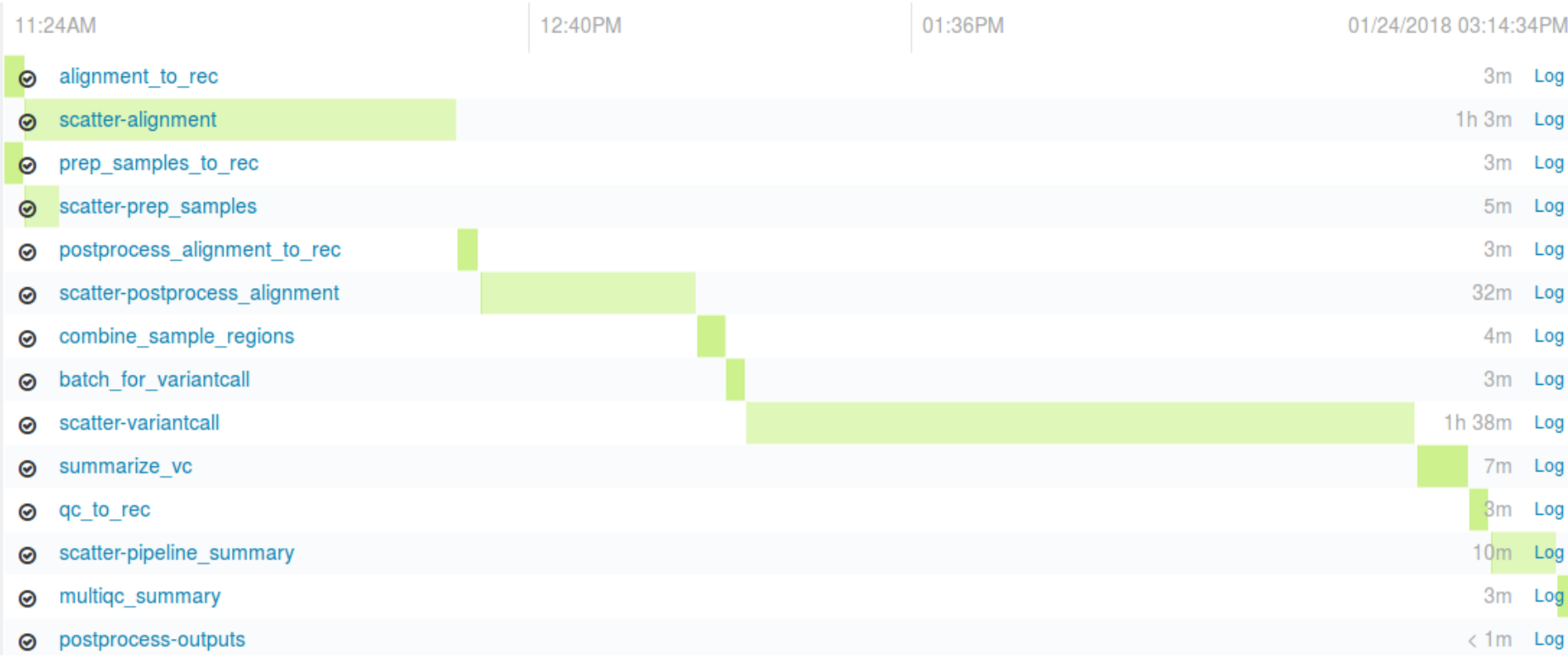
postprocess_variants	Complete	1h15m / 1h15m (1.0×)
concat_batch_variantcalls	Complete	1m / 1m (1.0×)
variantcall_batch_region_3	Complete	4h1m / 4h1m (1.0×)
variantcall_batch_region	Complete	3h43m / 3h43m (1.0×)
summarize_sv	Complete	0m13s / 0m13s (1.0×)
detect_sv	Complete	2h4m / 2h4m (1.0×)
variantcall_batch_region_2	Complete	2h50m / 2h50m (1.0×)
detect_sv_2	Complete	46m / 46m (1.0×)
detect_sv_3	Complete	11m / 11m (1.0×)

Run on DNAnexus platform

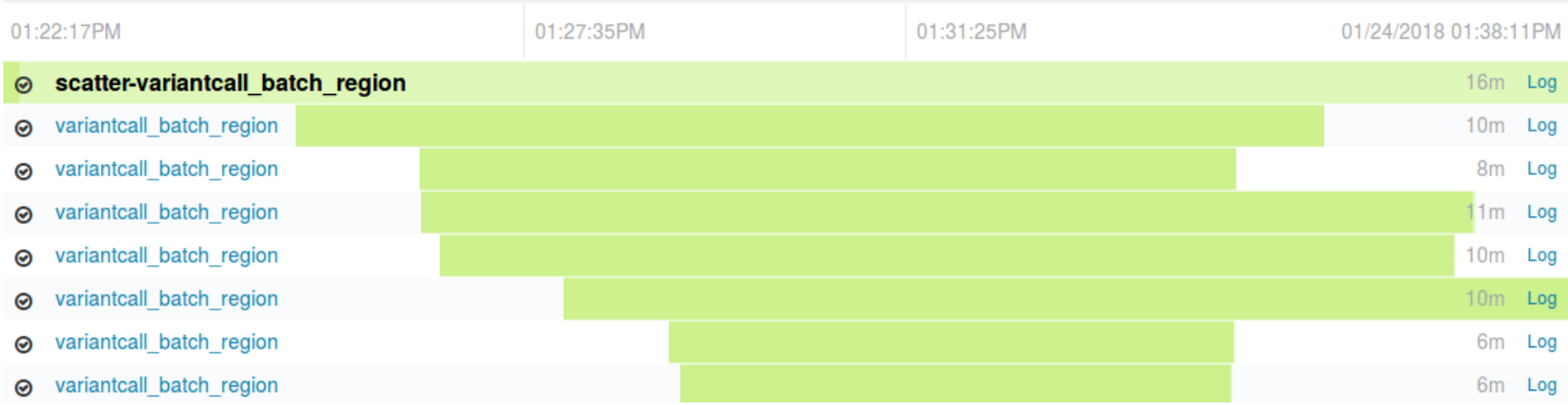
```
dnanexus:
  project: PGP
  ref:
    project: bcbio_resources
    folder: /reference_genomes
  inputs:
    - /data/input
resources:
  default:
    cores: 8
    memory: 3500M
    jvm_opts: [-Xms750m, -Xmx3500m]
```

https://platform.dnanexus.com

DNAnexus monitoring: align, variant call, QC



Variant calling parallelization: per region



Outline

- Overview of variant calling tools
- bcbio: open source, validated, community built
- **Practical example: Personal Genome Project**
- Cancer calling of low frequency variants

- ApoE https://www.snpedia.com/index.php/APOE
- Two variants, on chromosome 19, that impact risk of Alzheimer's disease and cholesterol metabolism

rs429358	rs7412	Name
C	T	ε1
T	T	ε2
T	C	ε3
C	C	ε4

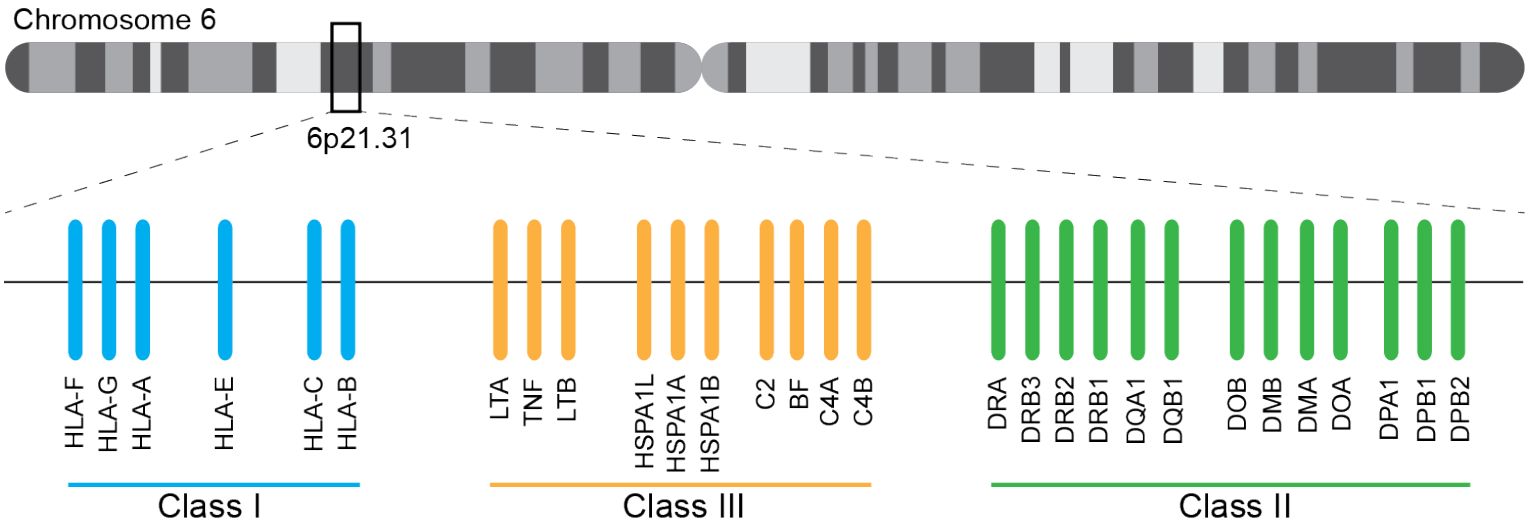
- Apo-ε1/ε1 gs267 rs429358(C;C) rs7412(T;T) the rare **missing allele**
- Apo-ε1/ε2 gs271 (C;T) (T;T)
- Apo-ε1/ε3 gs270 (C;T) (C;T) ambiguous with ε2/ε4
- Apo-ε1/ε4 gs272 (C;C) (C;T)
- Apo-ε2/ε2 gs268 (T;T) (T;T)
- Apo-ε2/ε3 gs269 (T;T) (C;T)
- Apo-ε2/ε4 gs270 (C;T) (C;T) ambiguous with ε1/ε3
- Apo-ε3/ε3 gs246 (T;T) (C;C) the most common
- Apo-ε3/ε4 gs141 (C;T) (C;C)
- Apo-ε4/ε4 gs216 (C;C) (C;C) ~11x increased Alzheimer's risk

ApoE analysis

```
$ tabix huD57BBF-gatk-haplotype.vcf.gz
    chr19:44908684-44908684
chr19   44908684        rs429358        T       C
1116.80    PASS
ANN=C|missense_variant|MODERATE|APOE|c.388T>C|p.Cys130Arg
GT:AD:DP:GQ:MMQ:PL      1/1:0,26:26:78:60:1145,78,0
$ tabix huD57BBF-gatk-haplotype.vcf.gz
     chr19:44908822-44908822
```

http://bit.ly/pgp-analysis

Major histocompatibility complex (MHC) – HLAs



http://www.ebi.ac.uk/ipd/imgt/hla/
http://sciscogenetics.com/technology/human-leukocyte-antigen-complex/

HLA typing

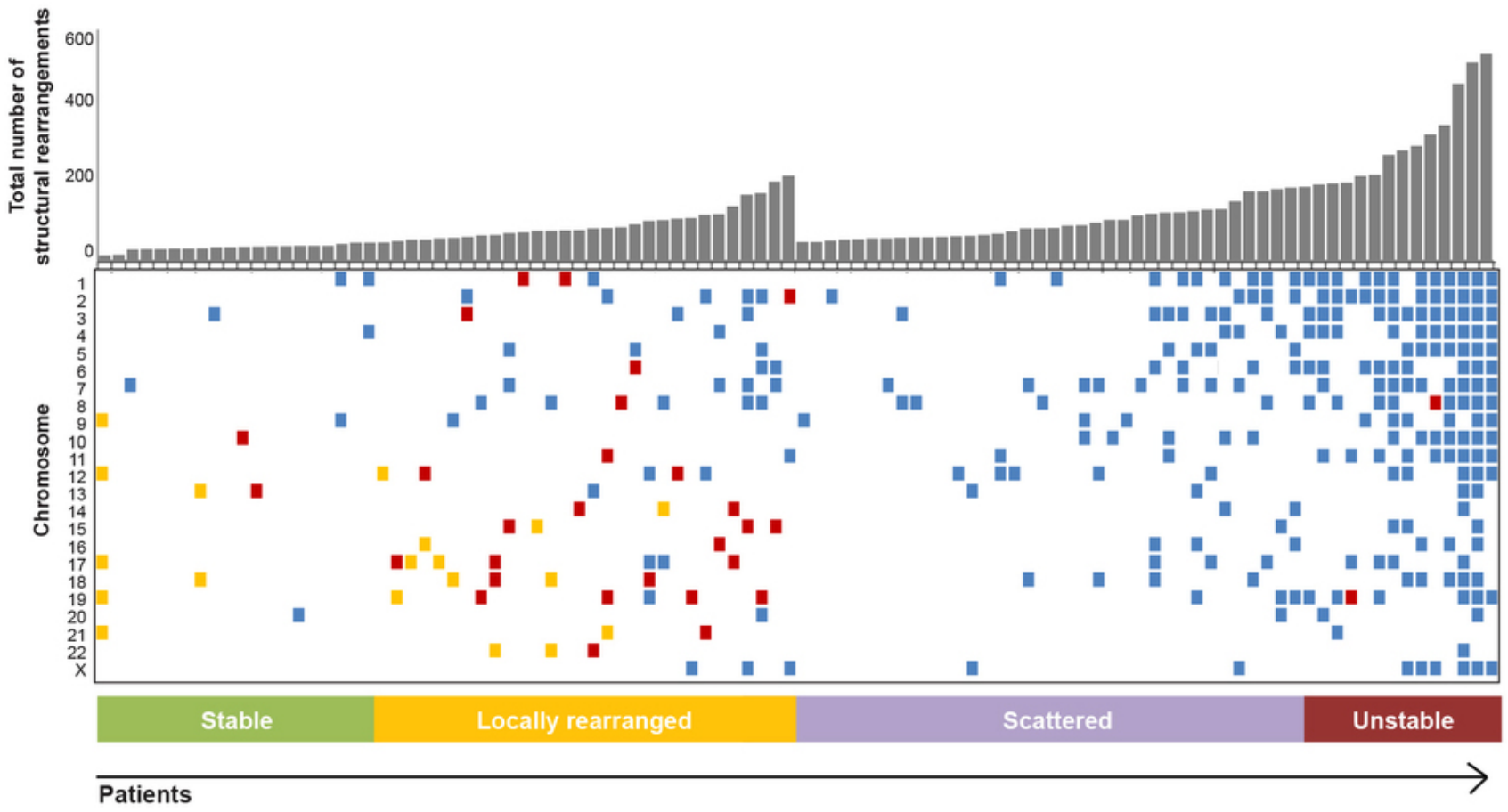
- 1000 genomes: build 38 + IMGT/HLA-3.18.0
- bwa mem extracts HLA reads
- Map reads only to HLA sequences
- OptiType: Call HLA types

https://github.com/lh3/bwa/blob/master/README-alt.md\#hla-typing
https://github.com/FRED-2/OptiType

HLA outputs

```
HLA-A*11:01;HLA-A*24:02
HLA-B*27:05;HLA-B*55:01
HLA-C*07:02;HLA-C*07:02
```

Structural variants critical – pancreatic cancer example



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

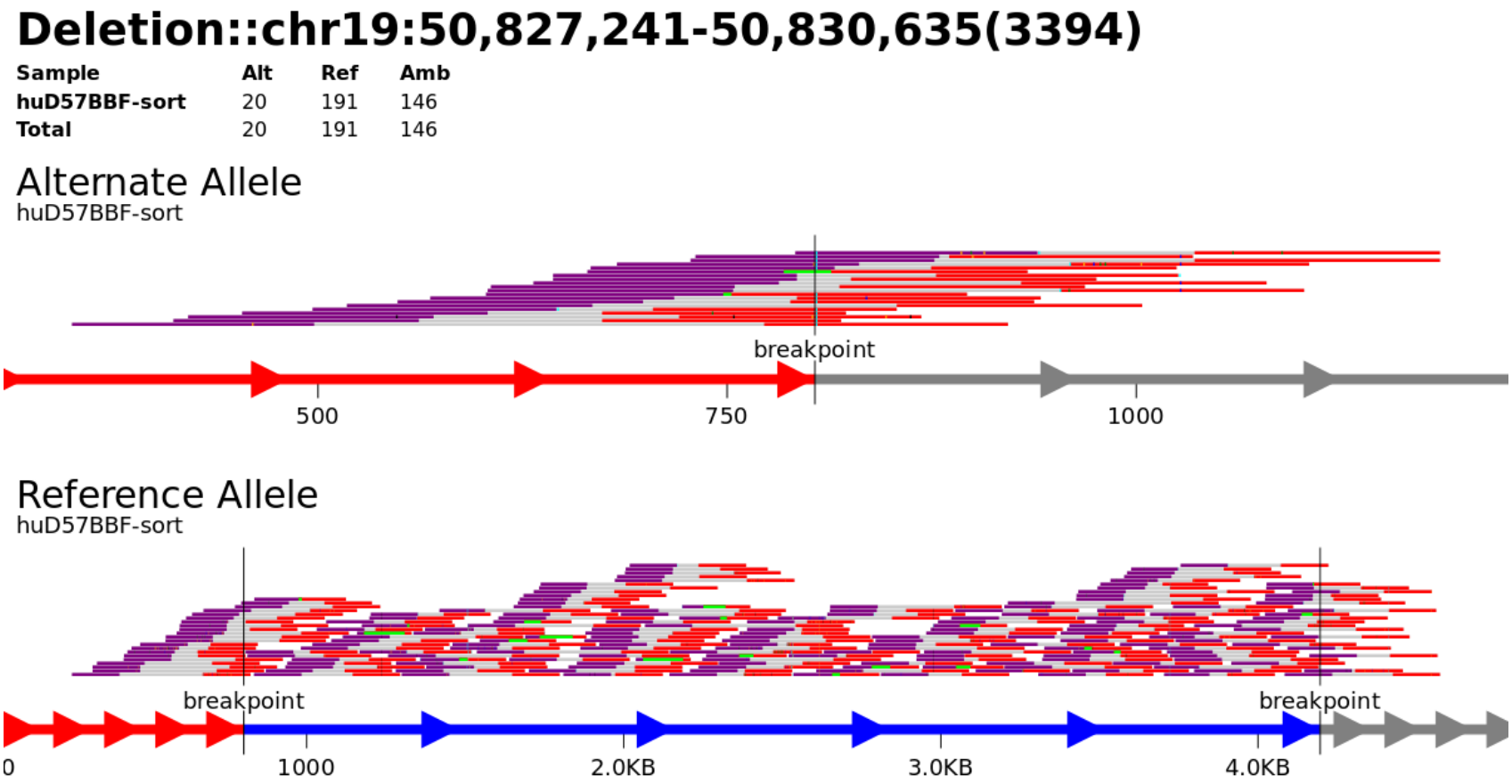
Tools used

- Manta: https://github.com/Illumina/manta
  Split and paired end reads
- Lumpy: https://github.com/arq5x/lumpy-sv
  Split and paired ends reads
- CNVkit: https://github.com/etal/cnvkit
  Read depth based

Example deletion call – 3 callers

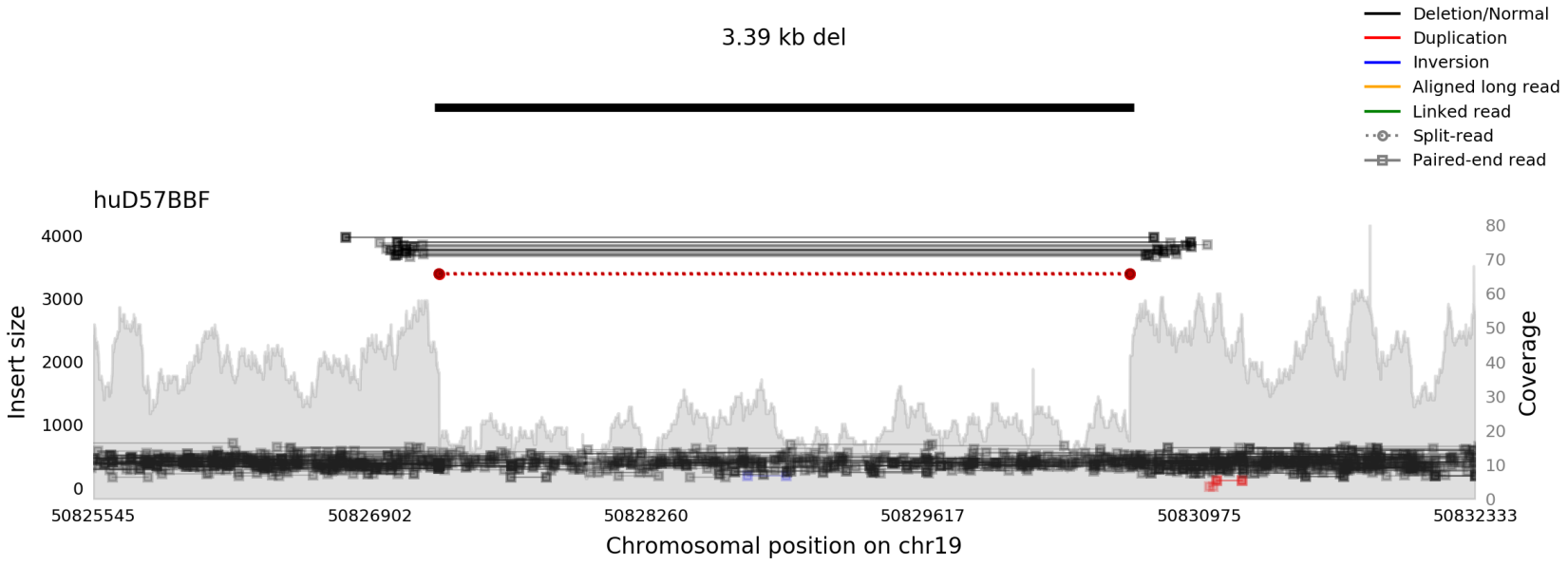
```
chr19   50827242        MantaDEL:67020:0:1:0:0:0
T   <DEL>   658.0 PASS
END=50830636;SVTYPE=DEL;SVLEN=-3394;
ANN=<DEL>|bidirectional_gene_fusion|HIGH|AC011523.2&KLK15|
ENSG00000267968&ENSG00000174562|gene_variant|
GT:FT:GQ:PL:PR:SR       0/1:PASS:504:708,0,501:18,16:23,12
```

Viewing deletion – svviz



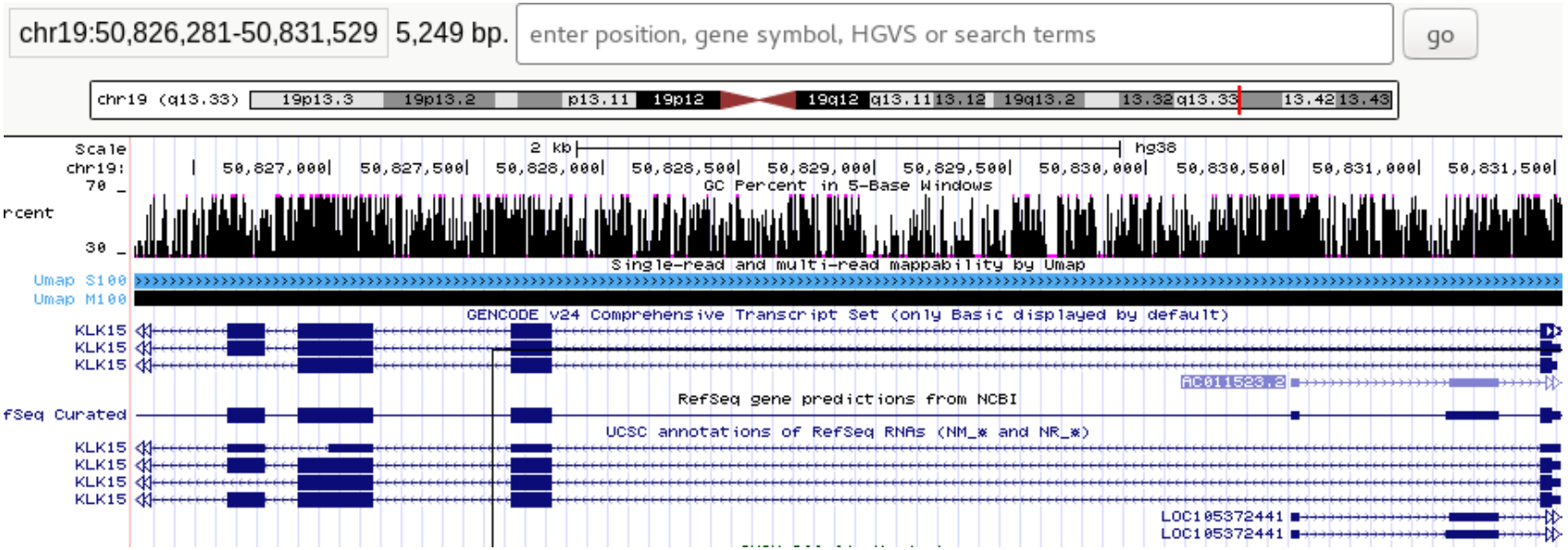
http://svviz.readthedocs.io

Viewing deletion – SV-plaudit



https://github.com/jbelyeu/SV-plaudit

Genomic region with deletion – KLK15



http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38

KLK15 known function

KLK15

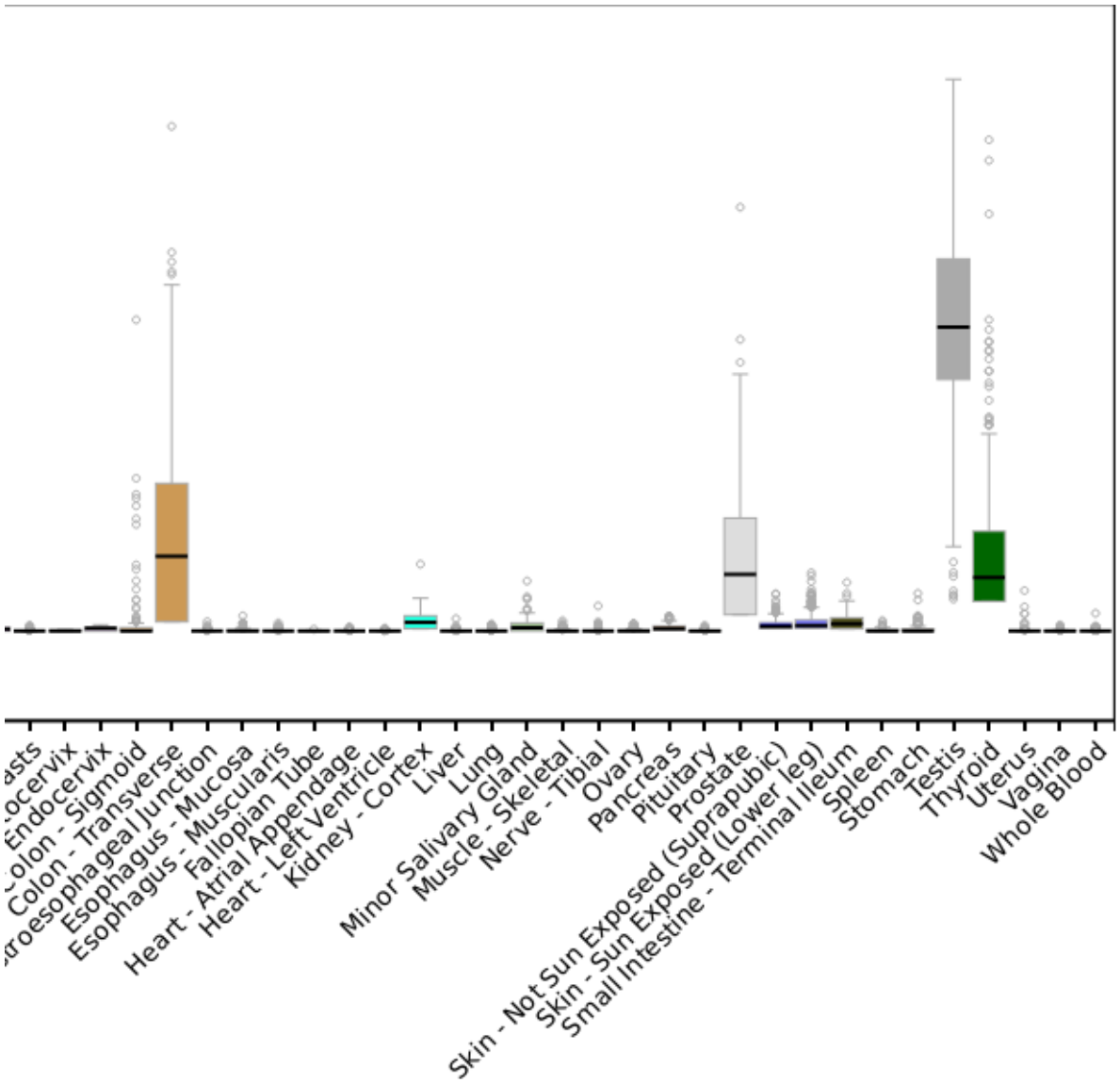
From Wikipedia, the free encyclopedia

Kallikrein-15 is a protein that in humans is encoded by the *KLK15* gene.[5][6][7][8][9]

Kallikreins are a subgroup of serine proteases having diverse physiological functions. Growing evidence suggests that many kallikreins are implicated in carcinogenesis and some have potential as novel cancer and other disease biomarkers. This gene is one of the fifteen kallikrein subfamily members located in a cluster on chromosome 19. I prostate cancer, this gene has increased expression, which indicates its possible use as a diagnostic or prognostic marker for prostate cancer. The gene contains multiple polyadenylation sites and alternative splicing results in multiple transcript variants encoding distinct isoforms.[9]

https://en.wikipedia.org/wiki/KLK15

Tissue specific gene expression



https://www.gtexportal.org/home/gene/ENSG00000174562.9

Self reported conditions

Conditions

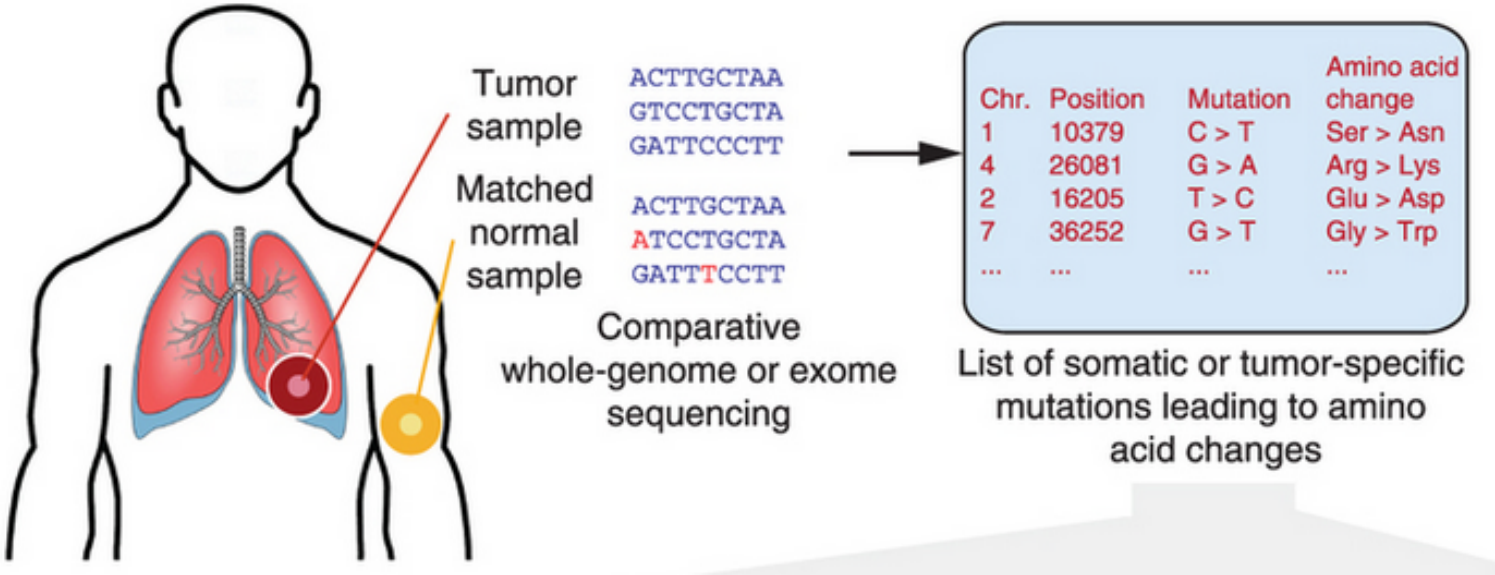
Name	Start Date
Benign Prostatic Hypertrophy (BPH)	1998-01-01
Heart murmur	2005-01-01
High Cholesterol	2000-01-01
Thyroid Nodule	2006-01-01

https://my.pgp-hms.org/profile/huD57BBF

Outline

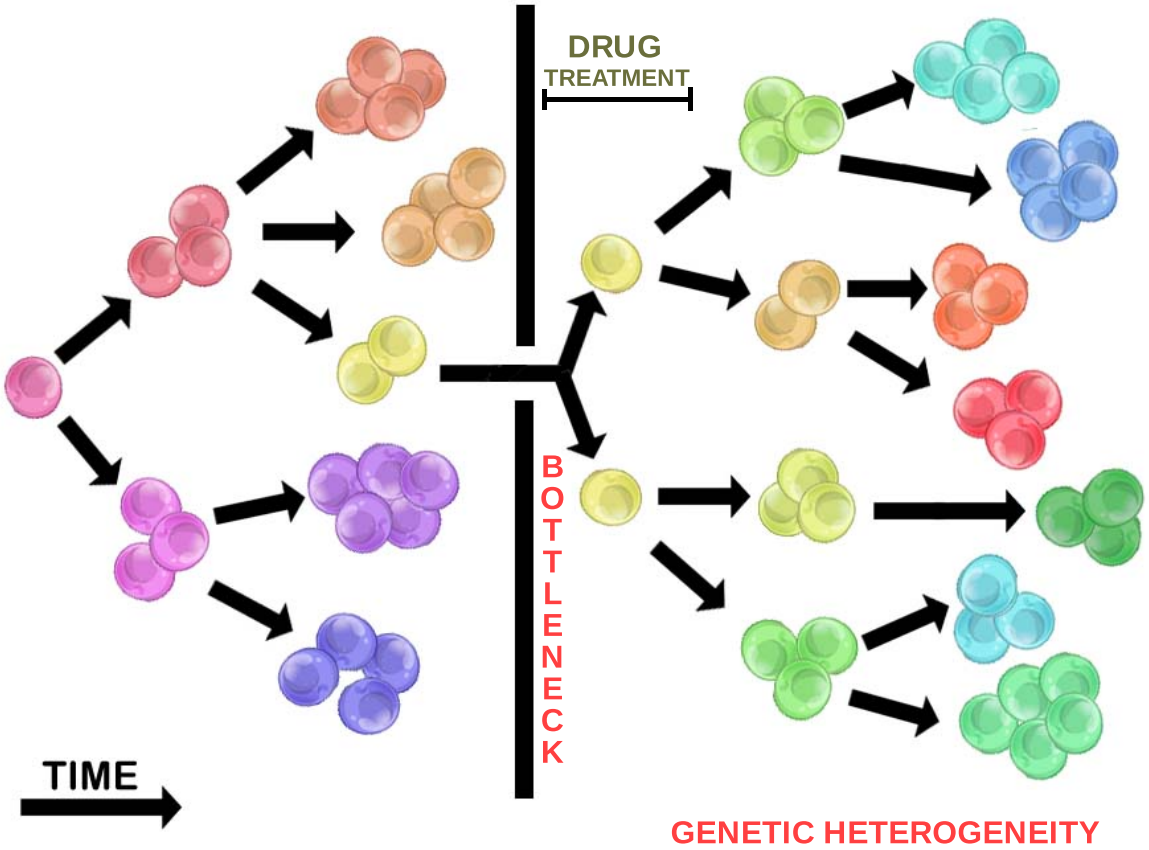
- Overview of variant calling tools
- bcbio: open source, validated, community built
- Practical example: Personal Genome Project
- **Cancer calling of low frequency variants**

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

https://github.com/AstraZeneca-NGS/VarDictJava
http://nar.oxfordjournals.org/content/early/2016/04/07/nar.gkw227.full

Validation: key component of bcbio

- Pre-built workflows with known outputs
- Cover multiple cases: germline, somatic, low frequency, FFPE, structural variants
- Large collections of diverse workflows

https://github.com/bcbio/bcbio_validation_workflows

Value of validation

- Integration tests for pipelines
- Unbiased algorithm comparisons
- Baseline for improving methods
- Automated tests for platforms

Reference materials

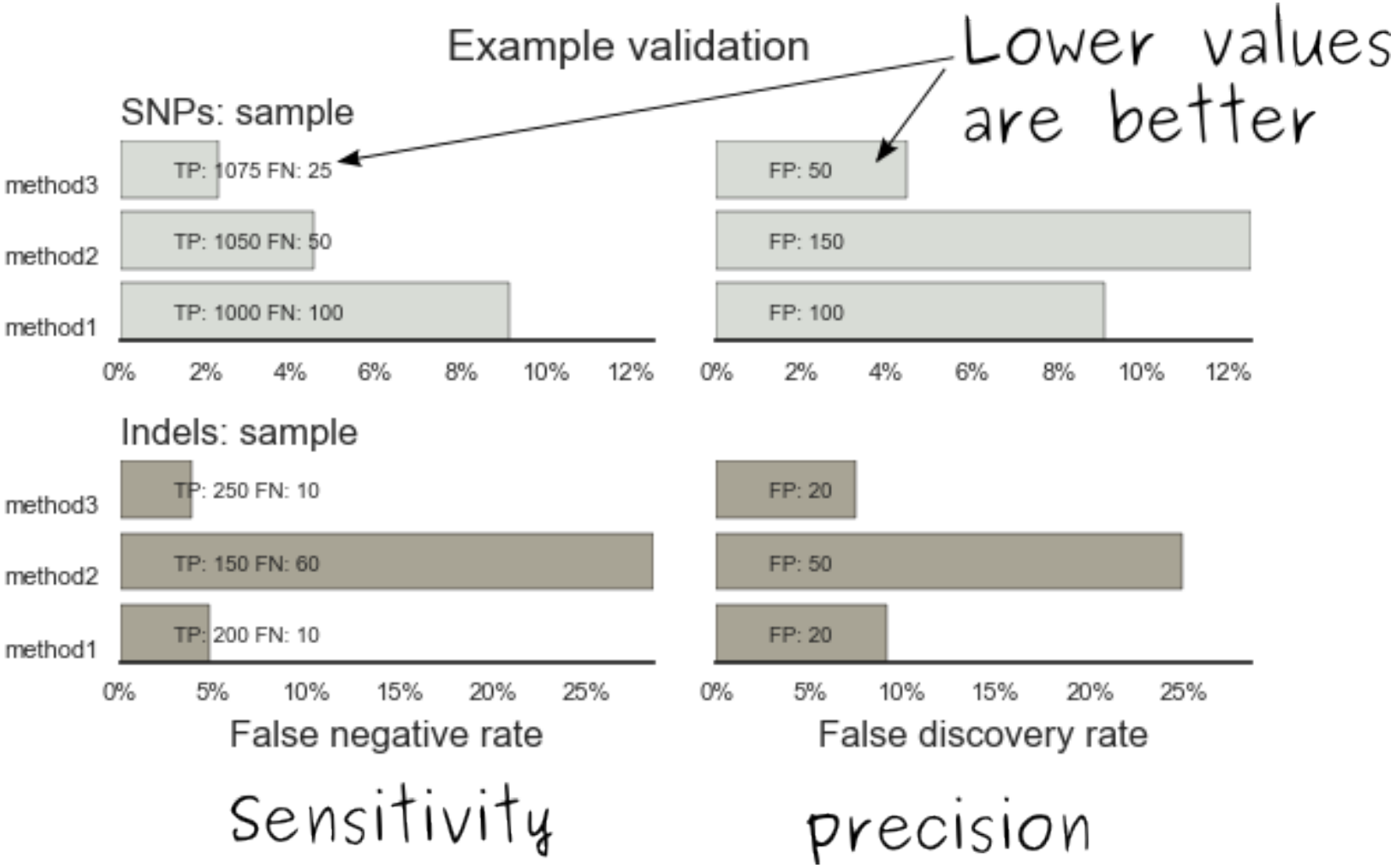
Genome in a Bottle Consortium

Global Alliance for Genomics & Health

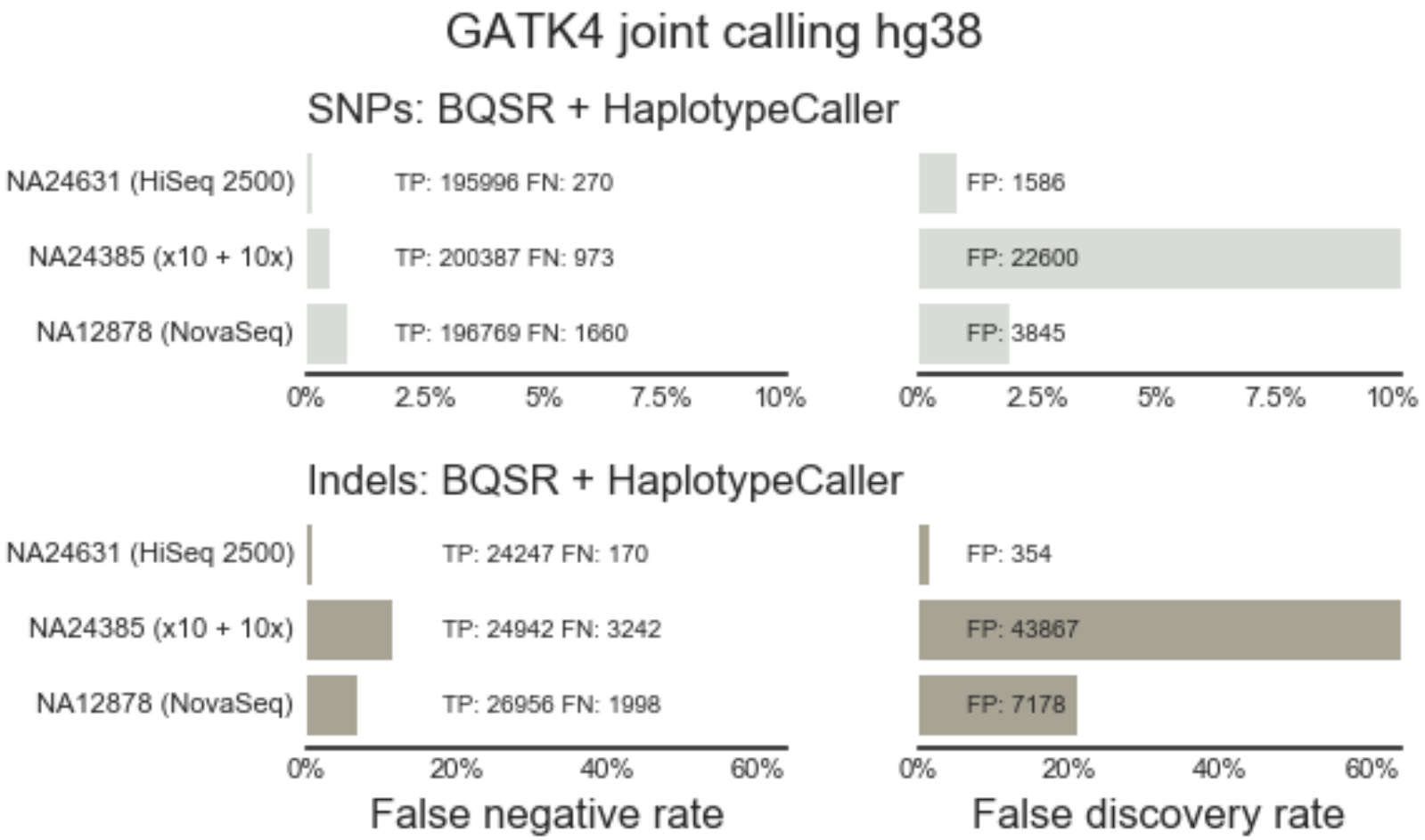
ICGC-TCGA DREAM Mutation Calling challenge

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

Validation graphs

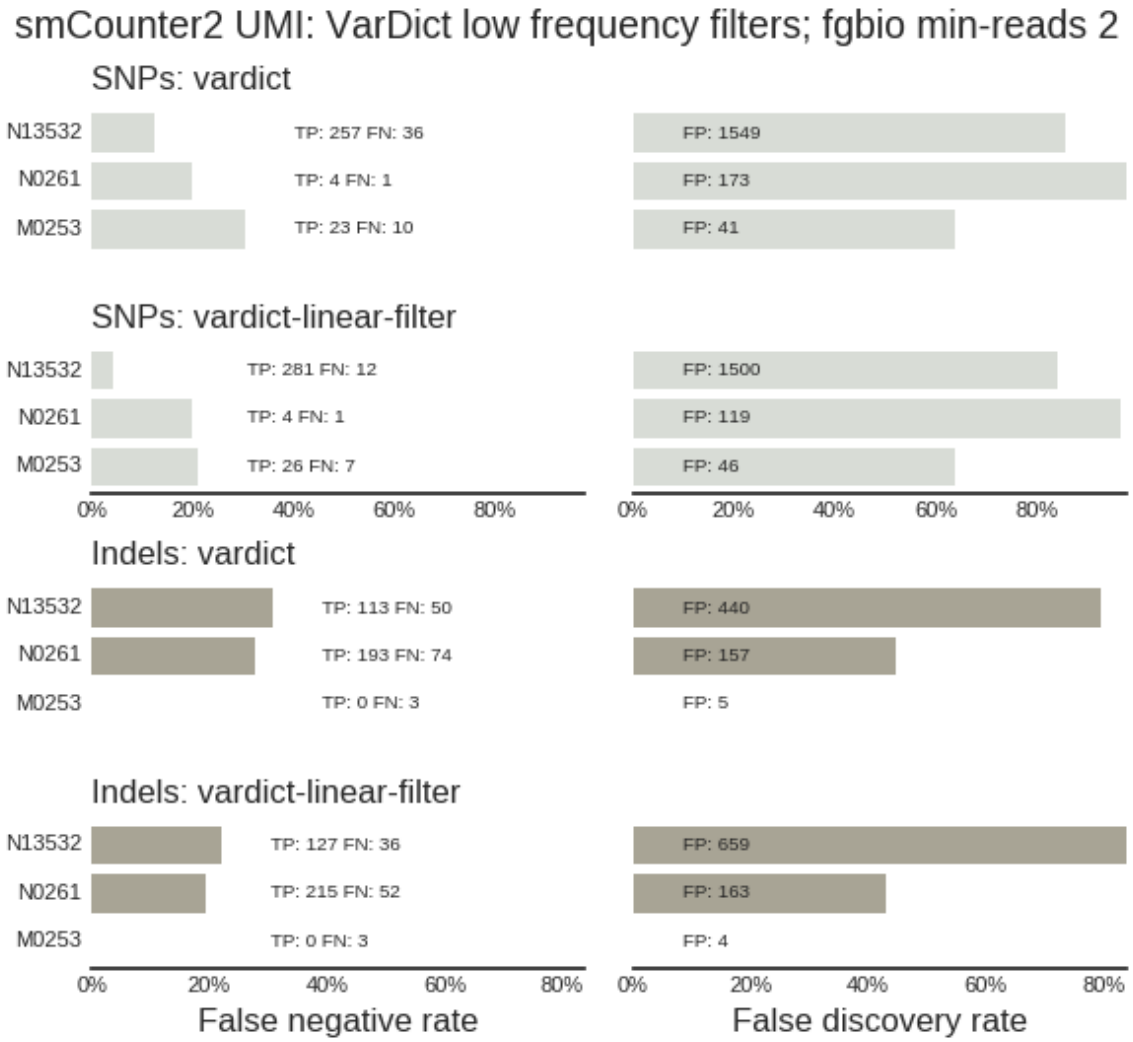


NA12878, NA24385, NA24631 GATK4 joint calling



https://github.com/bcbio/bcbio_validations/tree/master/gatk4

Low frequency variants: VarDict



https://github.com/bcbio/bcbio_validations/tree/master/somatic-lowfreq

Takeaways

- Overview of variant calling tools
- bcbio: open source, validated, community built
- Science = collaboration and re-use
- How to run bcbio analyses where you want them
- Interpreting variant calling outputs
- We can build better things together