# Personal Genome Project: Hackathon 1.0 example projects

Brad Chapman
Bioinformatics Core, Harvard Chan School
http://bit.ly/pgp-resources-1
http://bit.ly/pgp-analysis

21 July 2018

# What we'll do

- Understand the types of data in the personal genome project
- Learn how to query and find a genome to analyze
- Analyses you can do with a single genome

# Types of analyses

- Looking at small variants for traits
- HLA typing: the adaptive immune system
- Structural variants: larger events

# Outline

- **Overview of the Personal Genome Project and Data**
- Identify participants of interest
- Overview of human variant data analysis
- Example of looking at small variant data: ApoE
- Additional analyses with BAM reads:
  - HLA typing
  - Structural variant analysis
- Platform for data analysis: CWL, Arvados, bcbio

# 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**

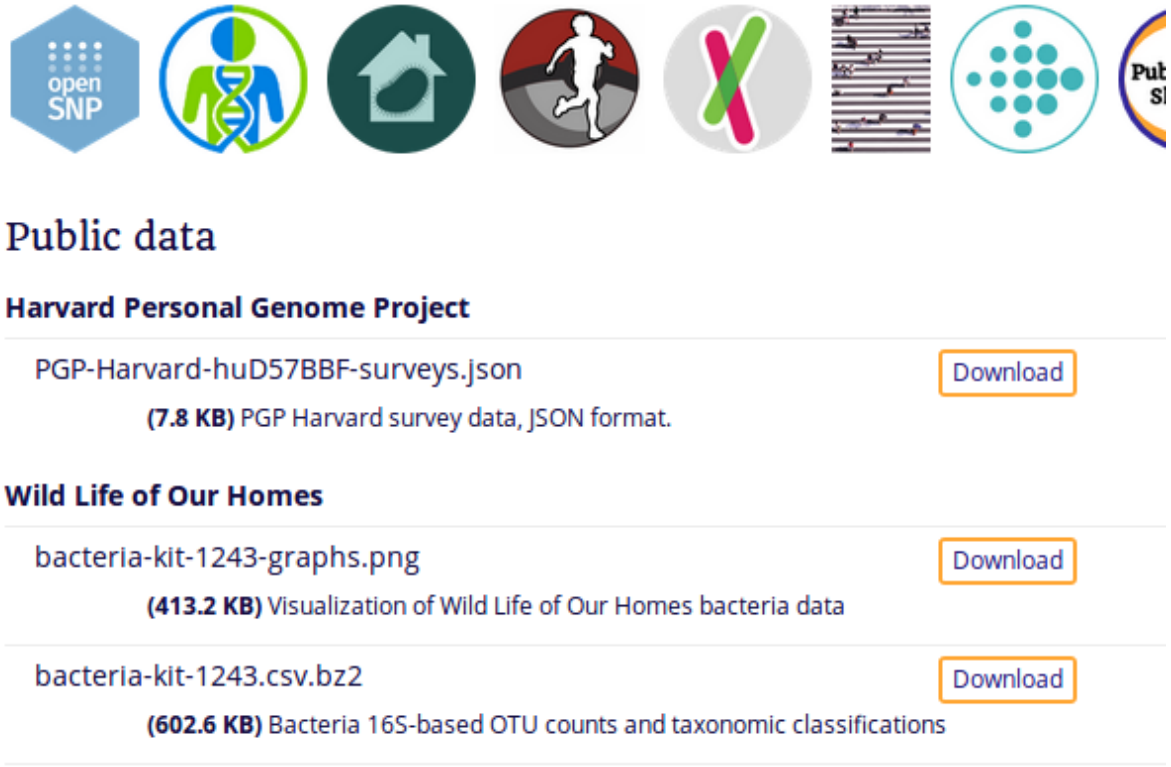
Personal Health Records

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/

# Collections of data in PGP

- Processed data: variants
  - Per participant portable ready to use VCFs
- Raw data: reads
  - Per participant BAM files of reads

http://bit.ly/pgp-analysis

https://github.com/bcbio/bcbio_validation_workflows/tree/master/pgp#find-bam-and-vcf-files-in-arvados-collections

# Outline

- Overview of the Personal Genome Project and Data
- **Identify participants of interest**
- Overview of human variant data analysis
- Example of looking at small variant data: ApoE
- Additional analyses with BAM reads:
  - HLA typing
  - Structural variant analysis
- Platform for data analysis: CWL, Arvados, bcbio

# Find a participant of interest

- Untap SQL database:
  https://github.com/abeconnelly/untap
- Participants plus associated metadata
- Regularly updated with new participants

https://collections.su92l.arvadosapi.com/c=2210f7ee07fc1c8b926e5db28eff9635-3284/_/html/index.html?disposition=inline

# Query for participant

- Example query and selection of participant

http://bit.ly/pgp-analysis

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

# Command line

```
$ cd /mnt/work/pgp/examples
$ bcbiovm_python \
    scripts/extract_veritas_pgp.py \
    untap.db

huD57BBF 53Gb No demographics
 [u'Family Tree DNA', u'Veritas Genetics',
  u'23andMe']
```

# Query, SQL to python pandas dataframe

```
query = ("SELECT uploaded_data.human_id, date, name "
         "FROM uploaded_data WHERE "
         "data_type == 'Veritas Genetics' AND "
         "uploaded_data.name GLOB '*VCF'")
conn = sqlite3.connect(sys.argv[1])
df = pd.read_sql_query(query, conn)
```

http://bit.ly/pgp-analysis

# Other example queries

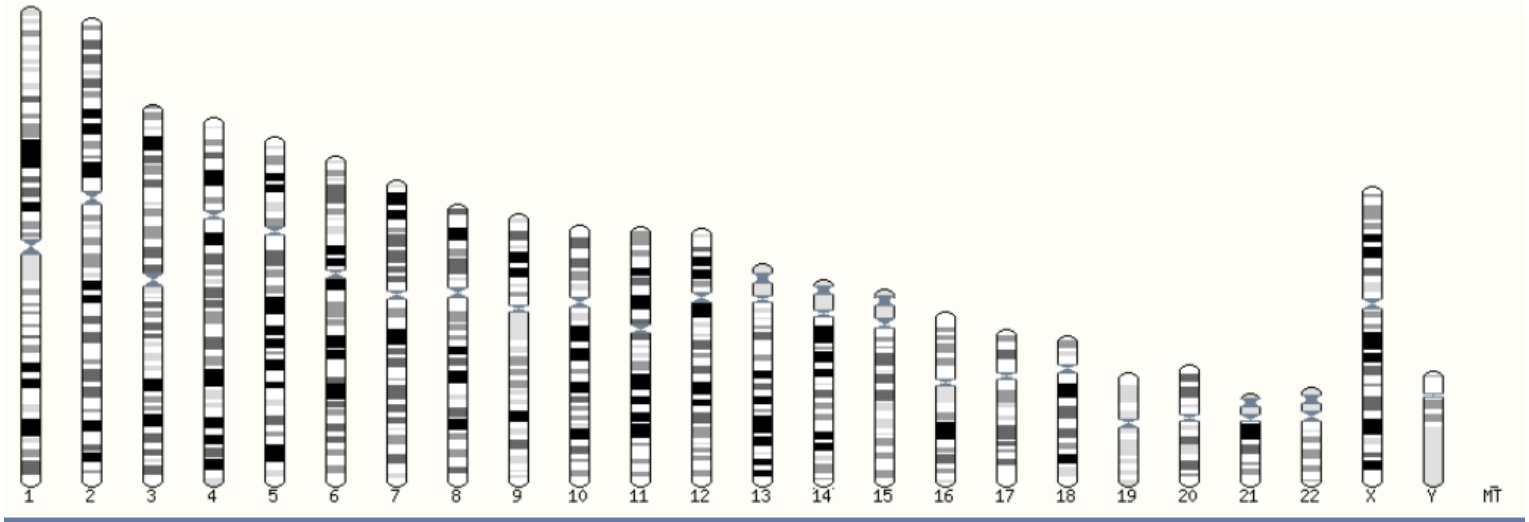
- Jupyter notebooks
- Summarize Age, Bloodtype, Ethnicity, Gender

https://github.com/swzCuroverse/PGPGraphics

# Outline

- Overview of the Personal Genome Project and Data
- Identify participants of interest
- **Overview of human variant data analysis**
- Example of looking at small variant data: ApoE
- Additional analyses with BAM reads:
  - HLA typing
  - Structural variant analysis
- Platform for data analysis: CWL, Arvados, bcbio

# Human whole genome sequencing



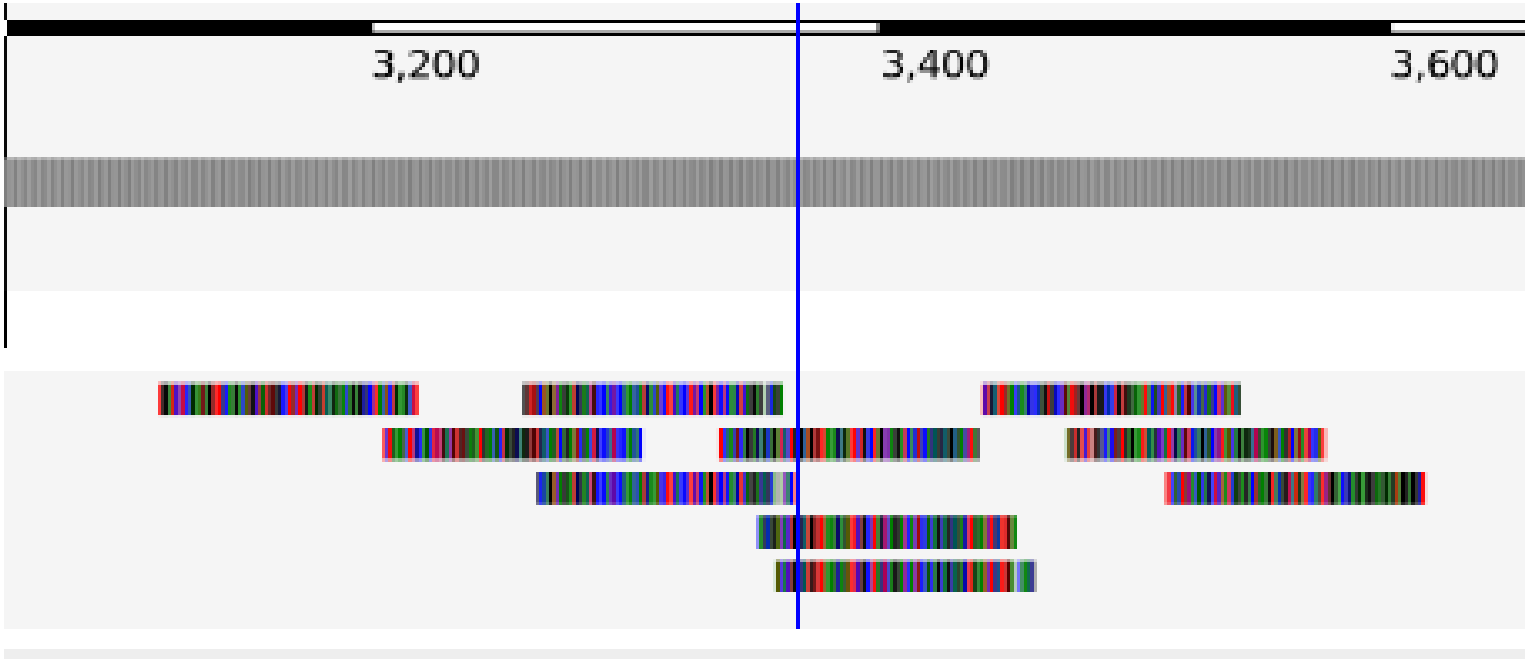

Click on the image above to jump to a chromosome, or click and drag to select a region

Summary

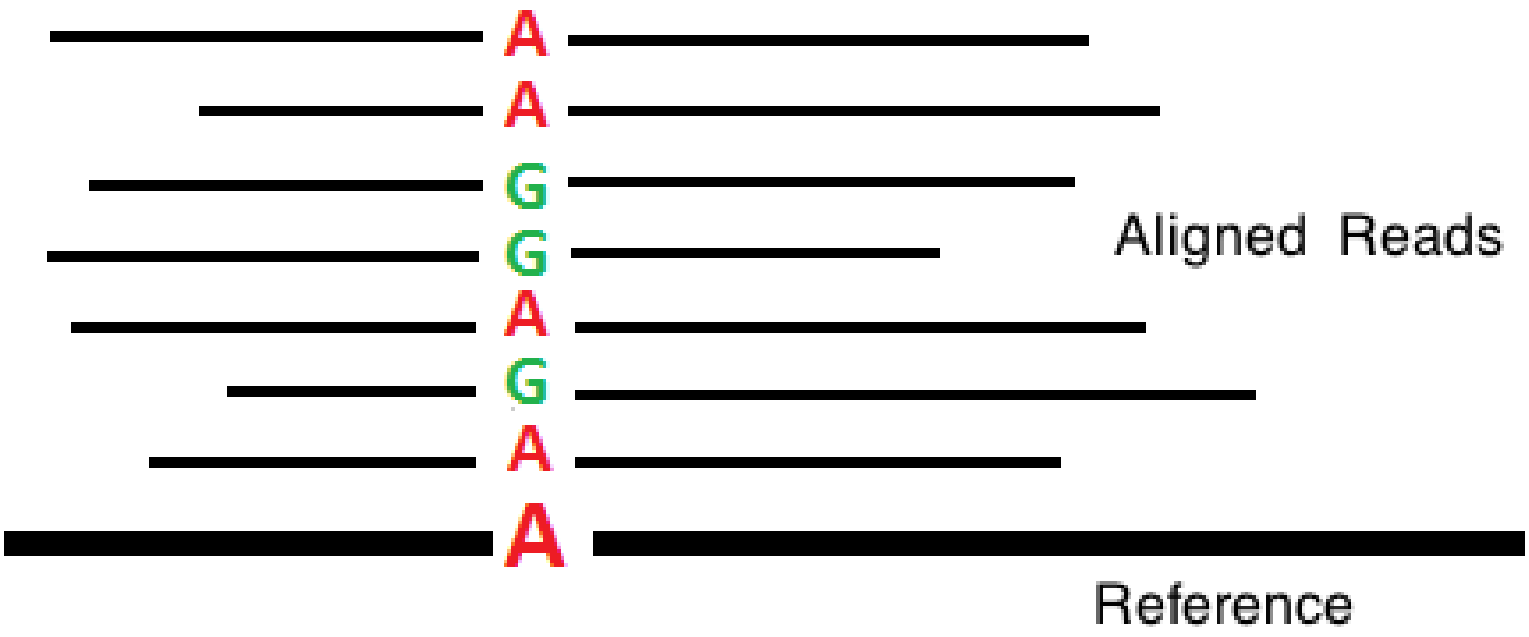
| | |
|---|---|
| **Assembly** | GRCh37.p13 (Genome Reference Consortium Human Reference 37), INSDC Assembly GCA_000001405.14, Feb 2009 |
| **Database version** | 75.37 |
| **Base Pairs** | 3,326,743,047 |

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

# High throughput sequencing

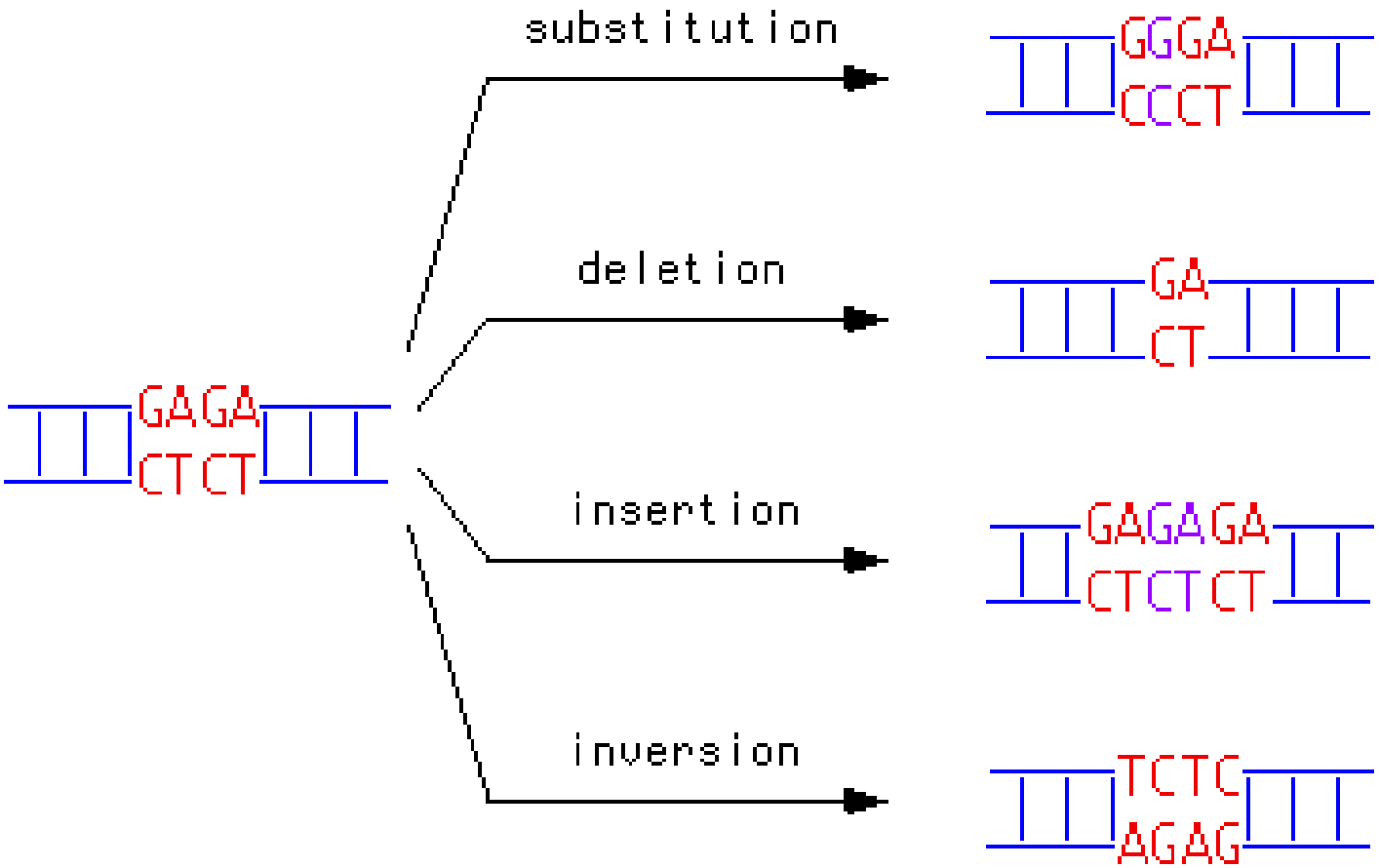


# Variant calling




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

# SNPs and Indels



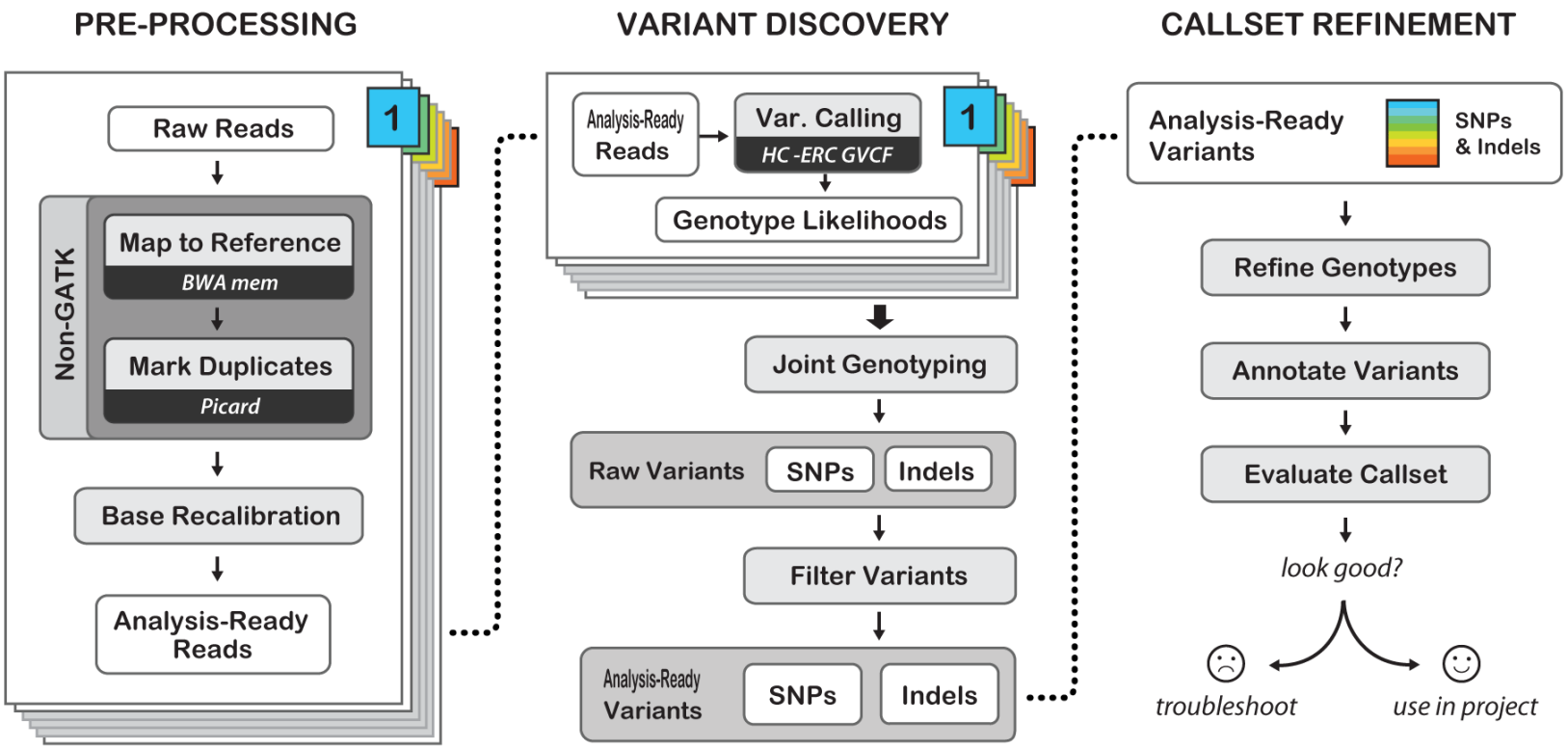

http://carolguze.com/text/442-2-mutations.shtml

# 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/

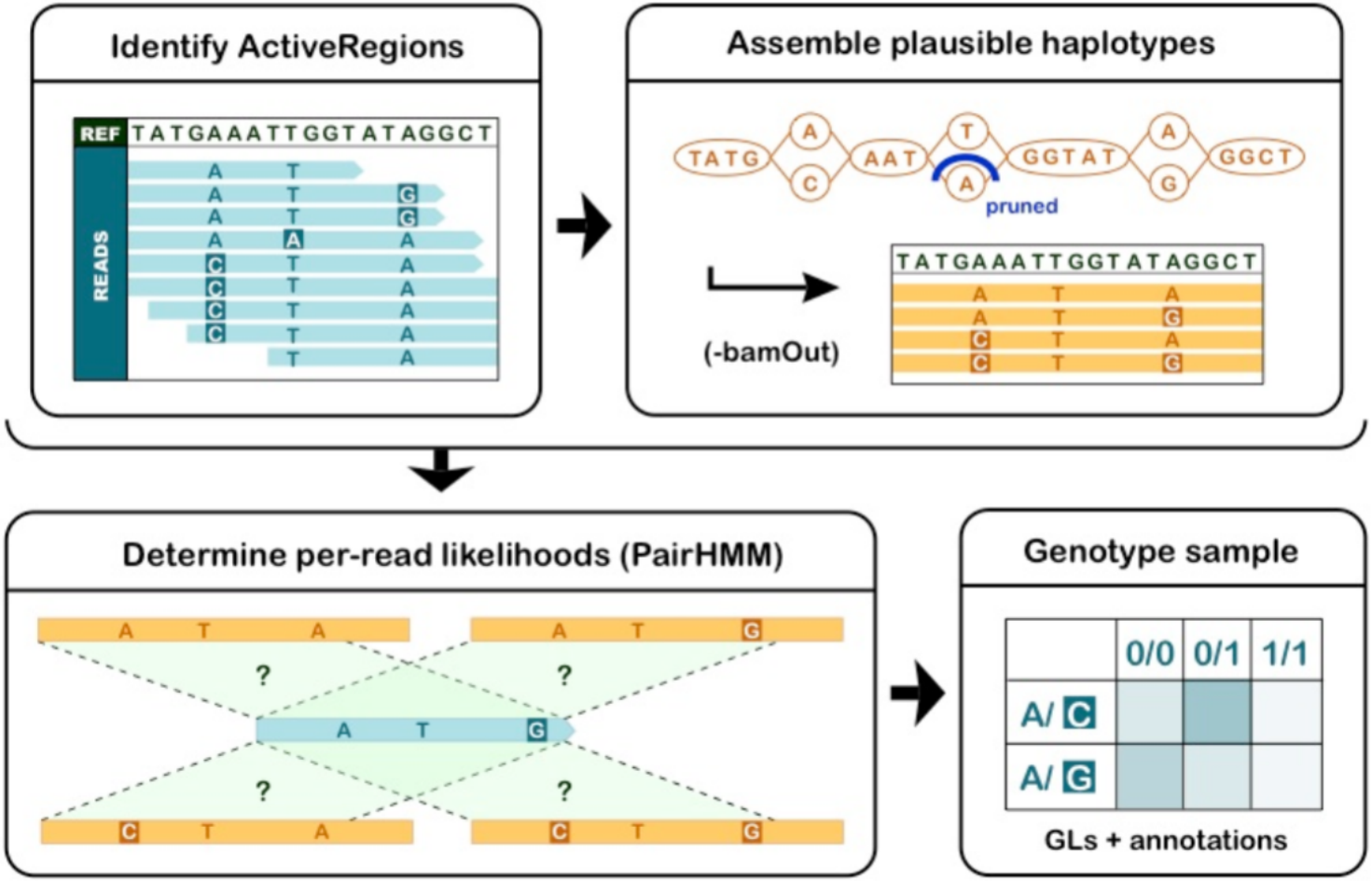
# GATK Best Practices




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

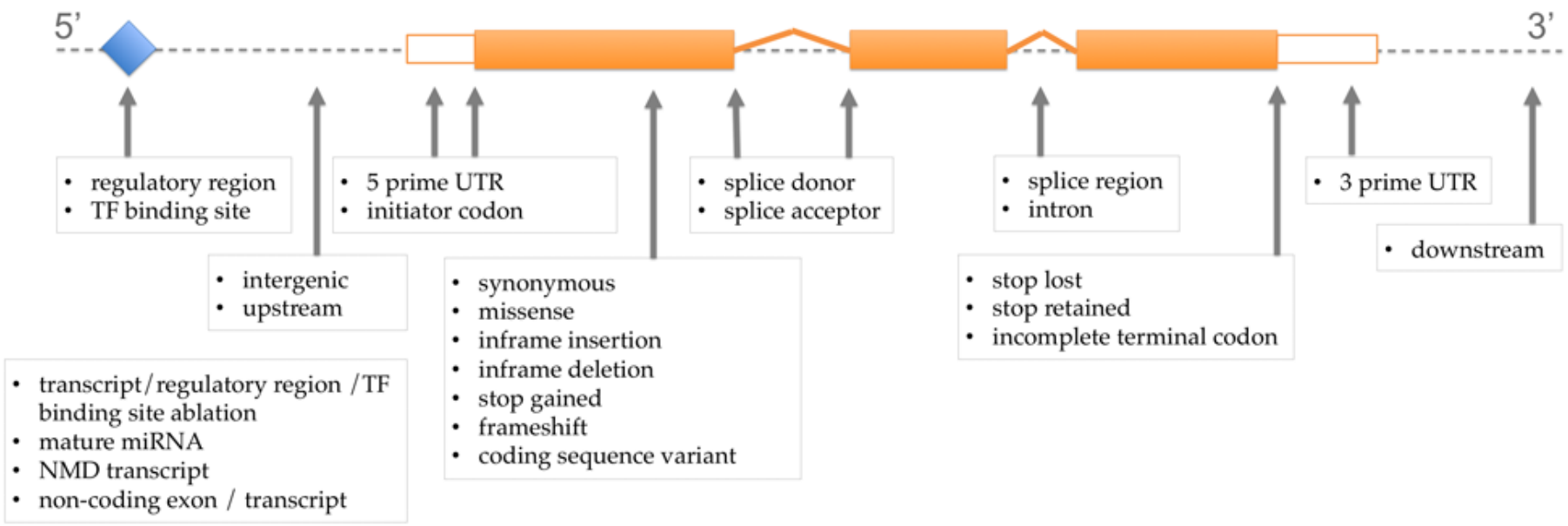
https://software.broadinstitute.org/gatk/best-practices/

# HaplotypeCaller



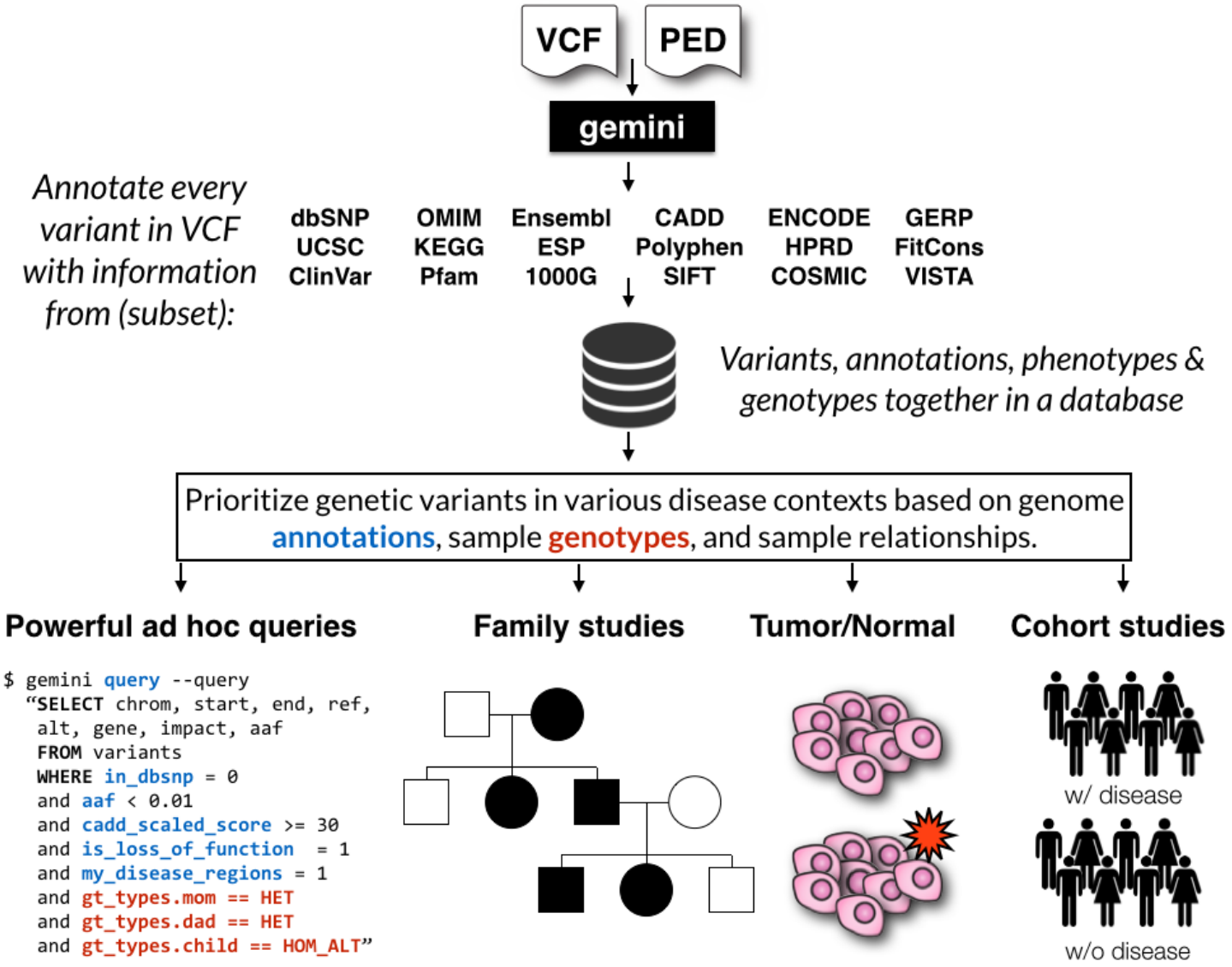

http://gatkforums.broadinstitute.org/discussion/5464/workshop-presentations-2015-uk-4-20-24

# Effects prediction



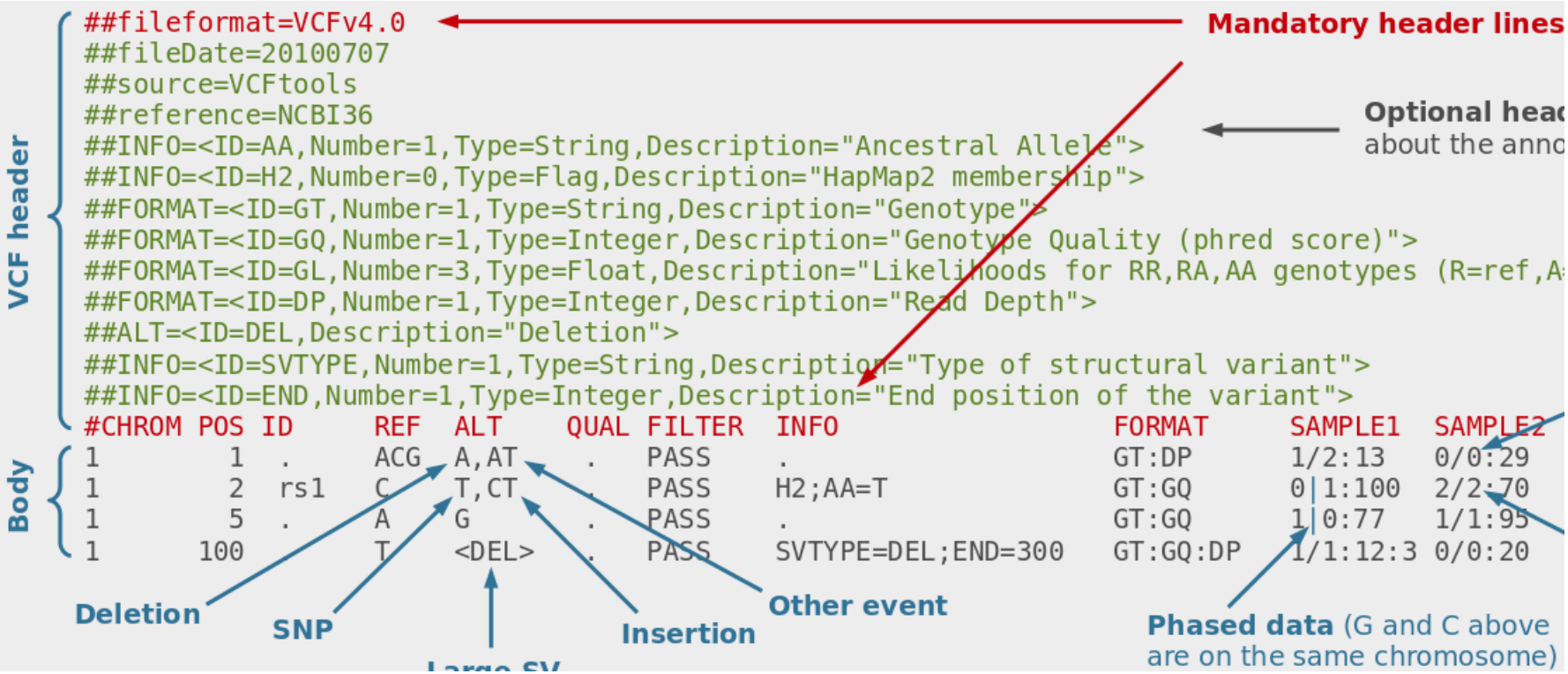

http://www.ensembl.org/info/genome/variation/predicted_data.html

# Annotation and analysis – GEMINI




https://github.com/arq5x/gemini

# 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

- Overview of the Personal Genome Project and Data
- Identify participants of interest
- Overview of human variant data analysis
- **Example of looking at small variant data: ApoE**
- Additional analyses with BAM reads:
  - HLA typing
  - Structural variant analysis
- Platform for data analysis: CWL, Arvados, bcbio

# Examine existing variation files

- Portable VCFs with small variant data
- Hosted as data collection with standard wget retrieval
- Also downloaded on work machines for PGP event: /mnt/work/pgp/vcf

https://workbench.su92l.arvadosapi.com/collections/su92l-4zz18-2rwb81xy8f1eh42

- 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

# Outline

- Overview of the Personal Genome Project and Data
- Identify participants of interest
- Overview of human variant data analysis
- Example of looking at small variant data: ApoE
- **Additional analyses with BAM reads**:
  - HLA typing
  - Structural variant analysis
- Platform for data analysis: CWL, Arvados, bcbio

# Performing additional analyses

- Raw files of reads in BAM format
- Also hosted as data collection by participant
- Demonstrate using open platforms for performing additional data analyses

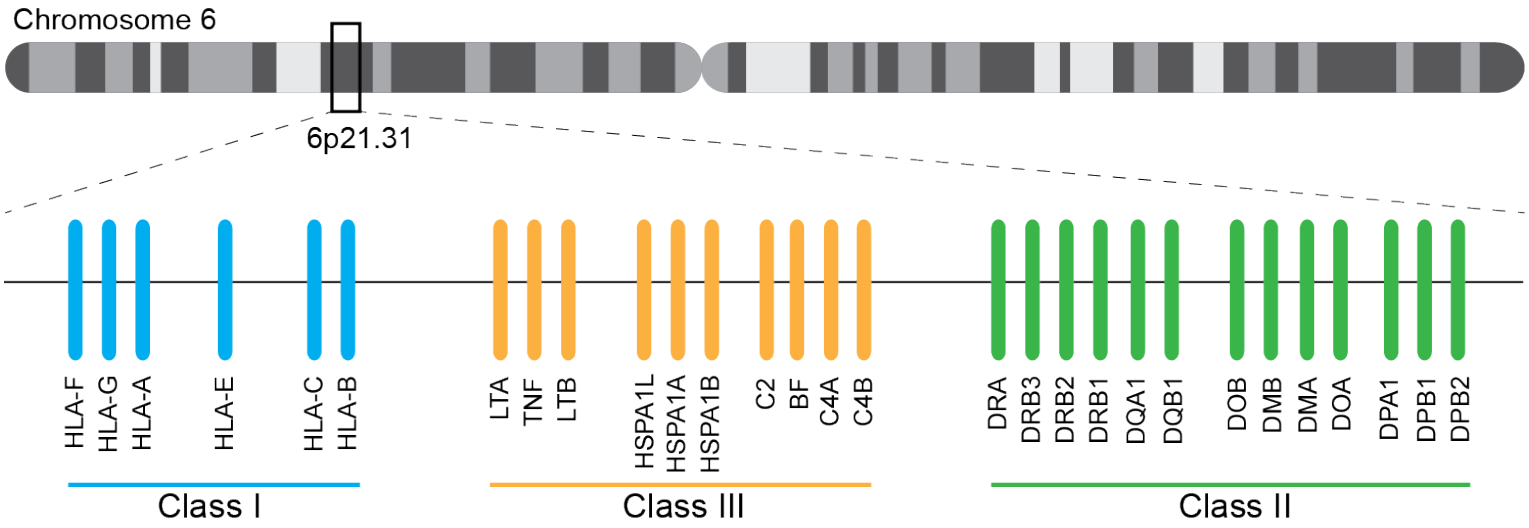
https://workbench.su92l.arvadosapi.com/collections/su92l-4zz18-1rqqi0kpkfmfite

# Major histocompatibility complex (MHC) – HLA

- human leukocyte antigen (HLA)
- Apative immune system
- Cell surface display and recognition
- Organ transplants, Cancer immunotherapy

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

# HLA – complex and repetitive




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

# HLA typing

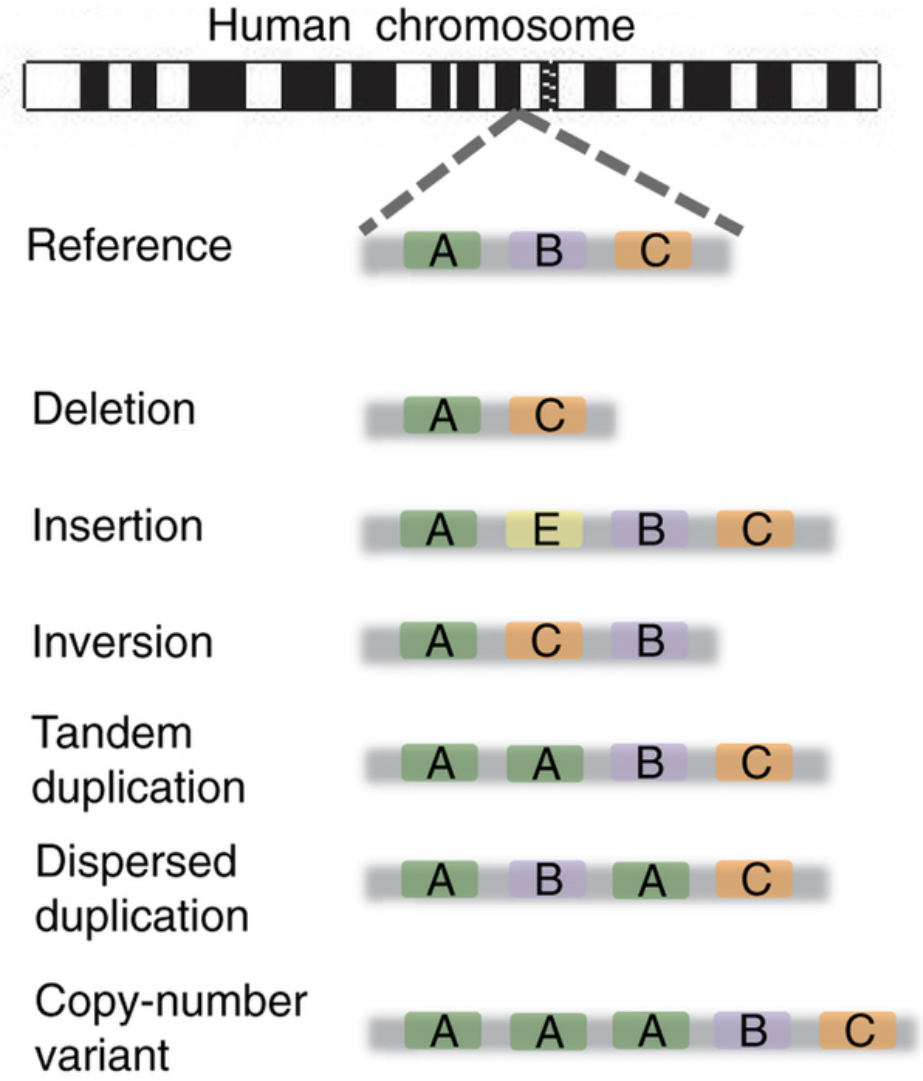
- 1000 genomes: build 38 + IMGT/HLA-3.18.0
- bwa mem extracts HLA reads
- Map reads only to HLA sequences
- OptiType: Call HLA types

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 variations




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

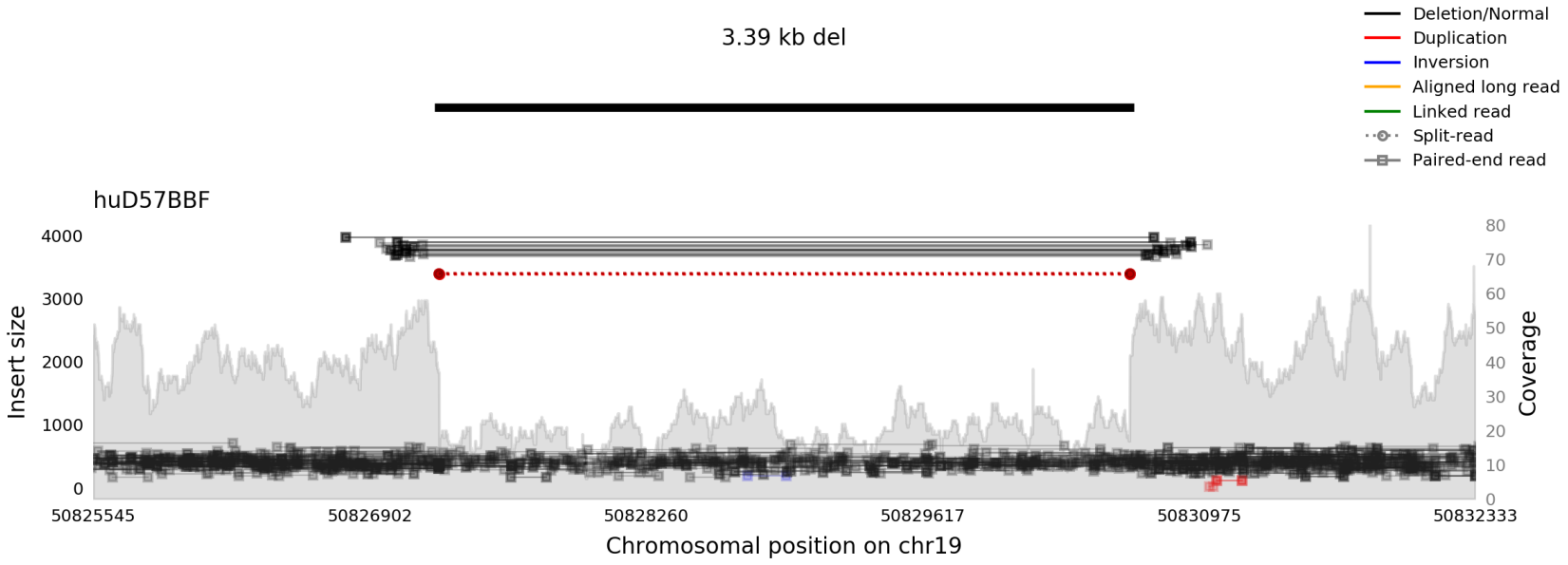
# Tools used

- 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

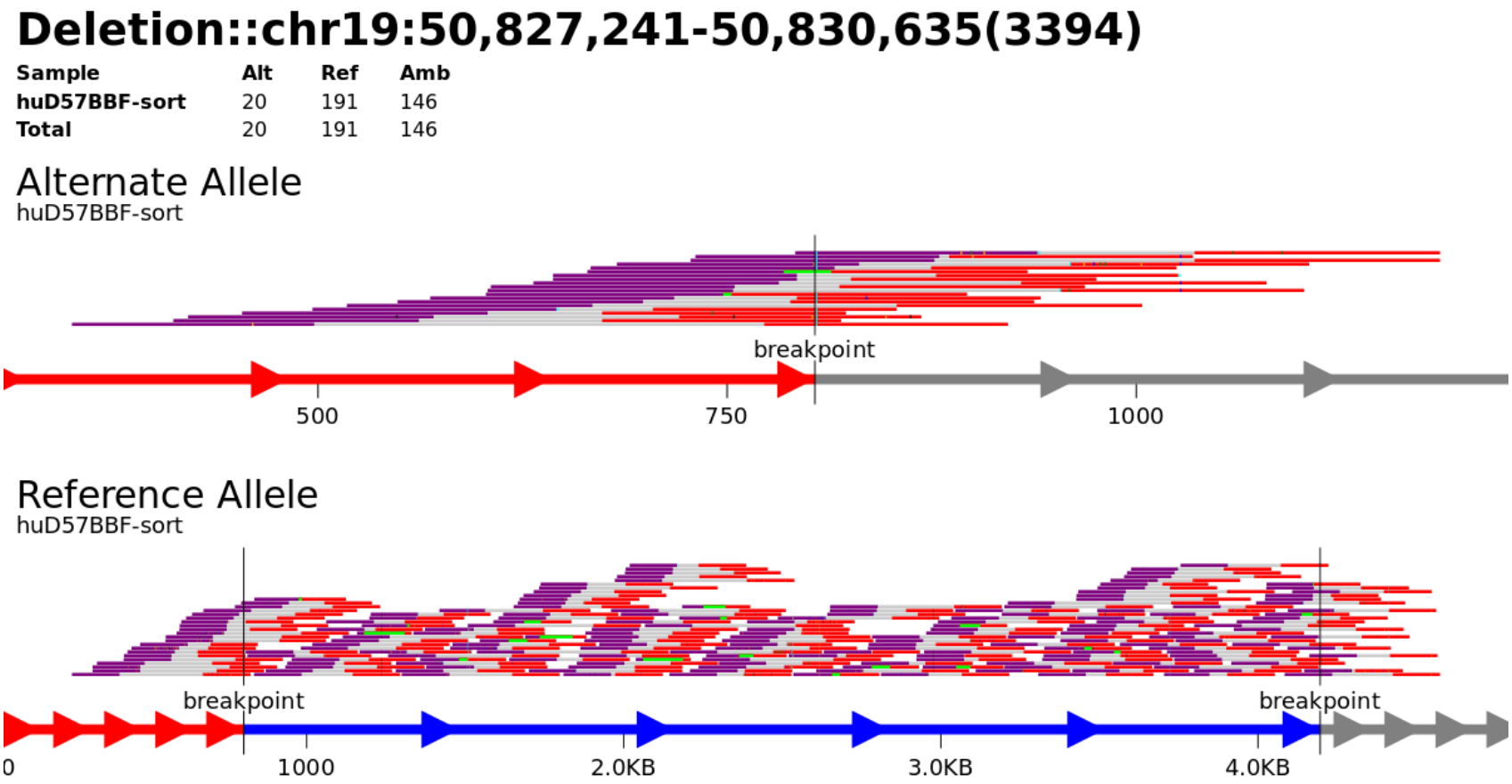
```
chr19   50827242        MantaDEL:67020:0:1:0:0:0
T   <DEL>   658.0 PASS
END=50830636;SVTYPE=DEL;SVLEN=-3394;
ANN=<DEL>|bidirectional_gene_fusion|HIGH|AC011523.2&KLK15|
ENSG00000267968&ENSG00000174562|gene_variant|
GT:FT:GQ:PL:PR:SR       0/1:PASS:504:708,0,501:18,16:23,12
```

# Viewing deletion – SV-plaudit



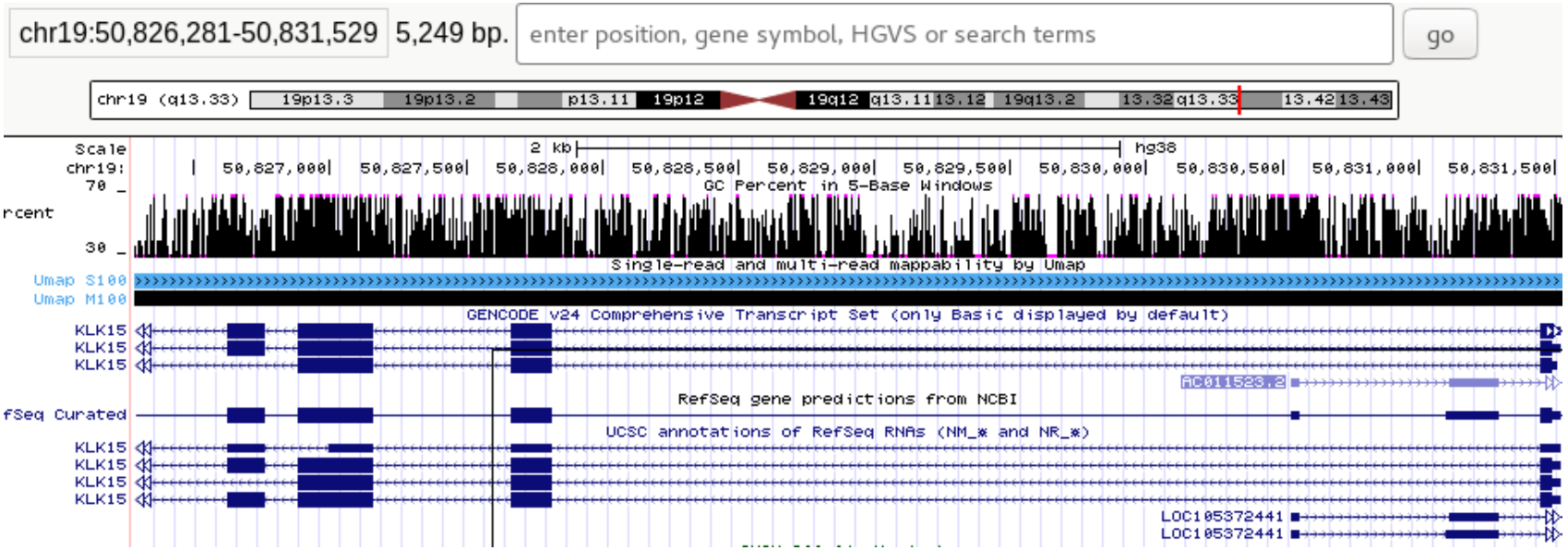

https://github.com/jbelyeu/SV-plaudit

# Viewing deletion – svviz




http://svviz.readthedocs.io

# Genomic region with deletion – KLK15




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

# KLK15 known function

## KLK15

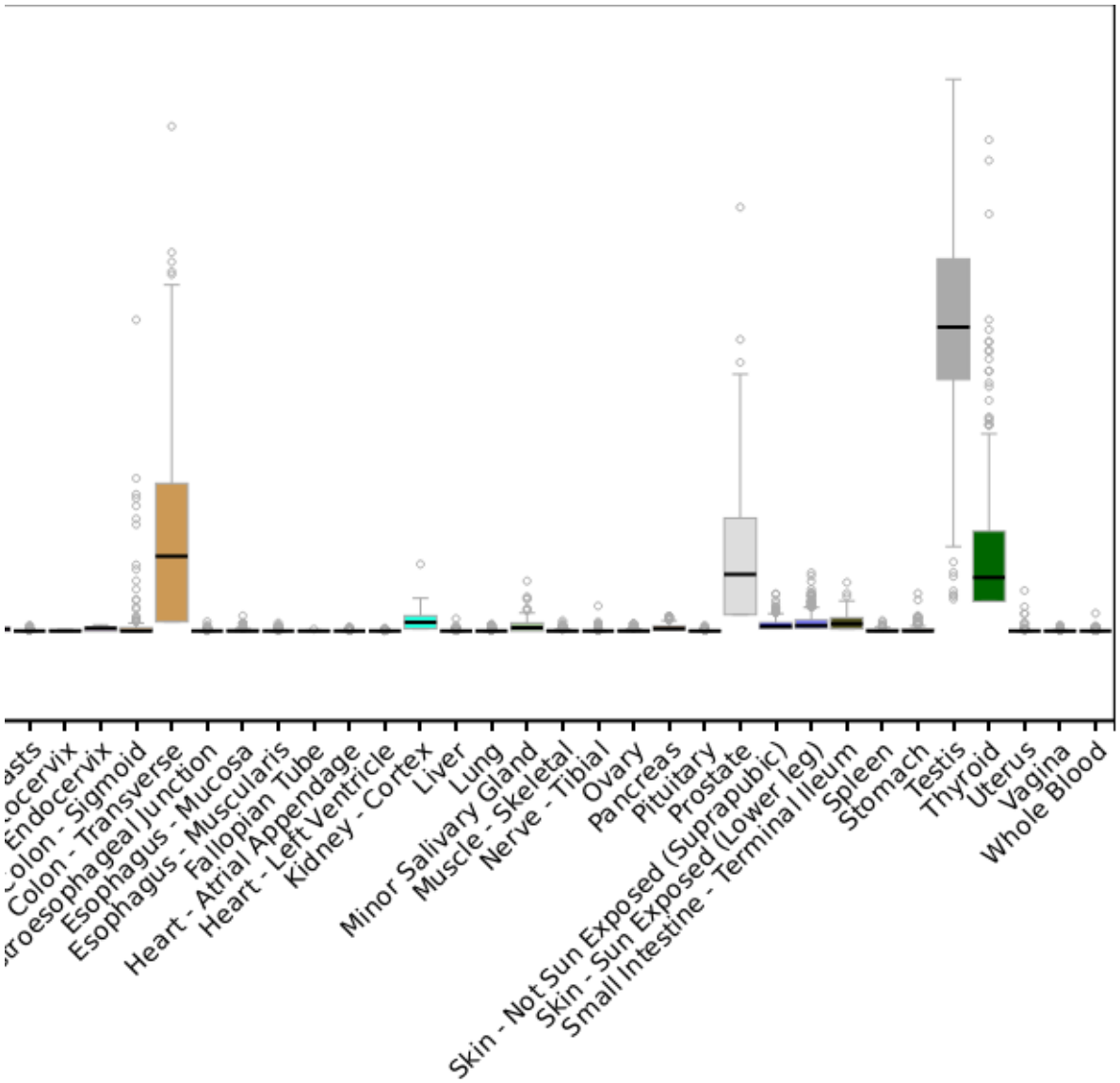
From Wikipedia, the free encyclopedia

**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

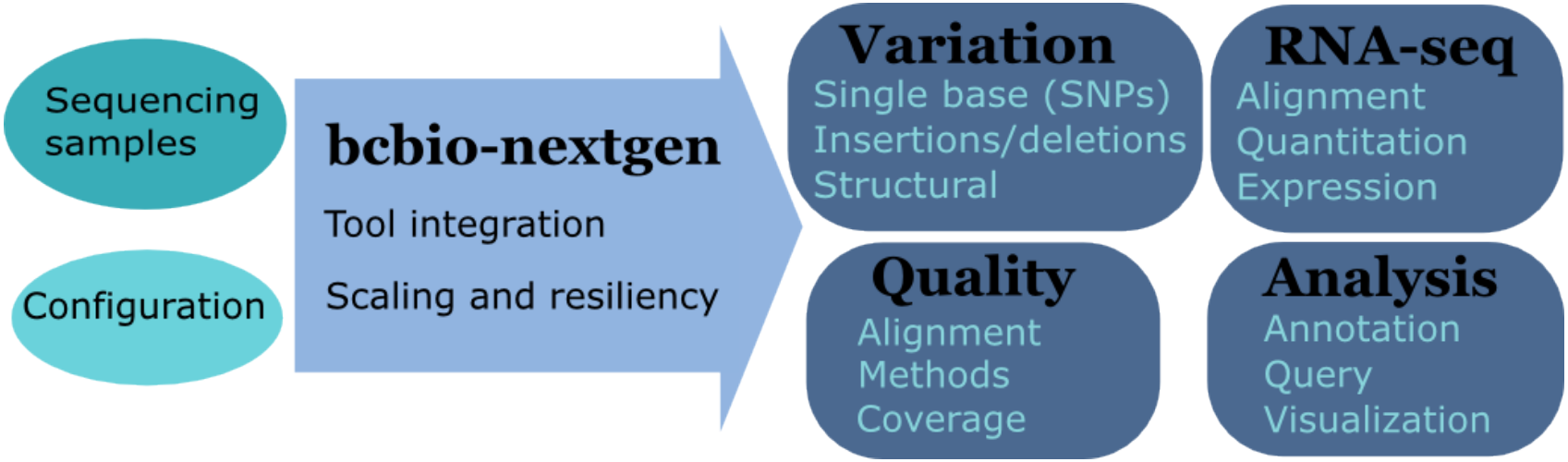
| Name | Start Date |
|---|---|
| Benign Prostatic Hypertrophy (BPH) | 1998-01-01 |
| Heart murmur | 2005-01-01 |
| High Cholesterol | 2000-01-01 |
| Thyroid Nodule | 2006-01-01 |

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

# Outline

- Overview of the Personal Genome Project and Data
- Identify participants of interest
- Overview of human variant data analysis
- Example of looking at small variant data: ApoE
- Additional analyses with BAM reads:
  - HLA typing
  - Structural variant analysis
- **Platform for data analysis: CWL, Arvados, bcbio**

# Open source community analysis




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

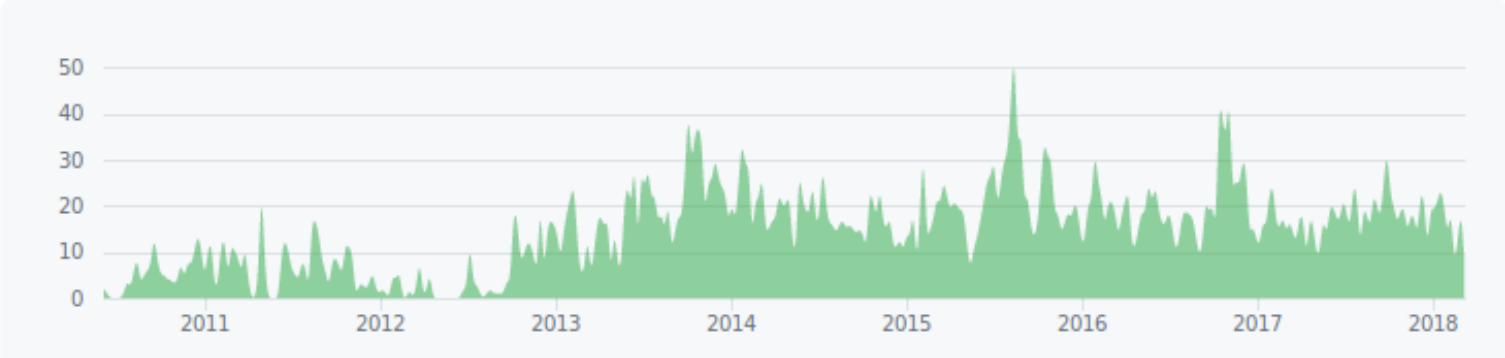
https://bcbio-nextgen.readthedocs.org/en/latest/contents/pipelines.html

# Community: sustainability and support

## Jul 18, 2010 – Apr 25, 2018

Contributions: **Commits**

Contributions to master, excluding merge commits




**Filters** is:issue is:open **Labels** **Milestones** **New issue**

**137 Open** 1,848 Closed Author Labels Projects Milestones Assignee Sort

**mirge 2.0 error** 3
#2379 opened 7 hours ago by mshadbolt

**Salmon quant.sf expression files are not combined, but the featurecount and stringtie files are combined?** 1
#2378 opened 15 hours ago by WimSpee

**Run Strelka2: Uncaught exception occurred**

# 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

# 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

# Arvados + Curoverse + Veritas

## Welcome to the Arvados Project

The Arvados community is dedicated to building a new generation of open source distributed computing software for bioinformatics, data science, and production analysis using massive data sets.

https://arvados.org/

# Why use a workflow abstraction?

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

# 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

# bcbio-vm: CWL wrapper

- bcbio-like interface integrating with external tools
- Install wrapper plus supported runners

```
conda install -c conda-forge -c bioconda bcbio-nextgen-vm
```

https://github.com/bcbio/bcbio-nextgen-vm
https://bioconda.github.io/

# 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

# Describe the platform resources

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

# Build Common Workflow Language description

```
bcbio_vm.py cwl --systemconfig bcbio_system-arvados.yaml \
  pgp_sv_hla.yaml
```

# Launch analysis

```
bcbio_vm.py cwlrun arvados pgp_sv_hla-workflow -- \
  --project-uuid su92l-j7d0g-eoibug3nrwg8ysj
```

https://workbench.su92l.arvadosapi.com/projects/su92l-j7d0g-eoibug3nrwg8ysj

# 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×) |

https://workbench.su92l.arvadosapi.com/container_requests/su92l-xvhdp-iprauko4kegv1kz

# Summary

- Overview of the Personal Genome Project and Data
- Identify participants of interest
- Overview of human variant data analysis
- Example of looking at small variant data: ApoE
- Additional analyses with BAM reads:
  - HLA typing
  - Structural variant analysis
- Platform for data analysis: CWL, Arvados, bcbio

# Next steps

- Work through examples to get started
- Propose your own projects building off these ideas
- Brainstorm new research ideas from PGP data
- Help us improve data access and organization
- Improve documentation and resources