Personal Genome Project analysis examples

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

26 April 2018

Outline

- **Overview of the Personal Genome Project**
- Identifying participants of interest
- Finding and examining variant data
- Finding raw read data
- Platforms for data analysis: CWL, Arvados, bcbio
- Running an interoperable analysis on PGP data
- Running structural variant and HLA analyses

Explain PGP and available data

- ToDo: existing slides we can use?

Outline

- Overview of the Personal Genome Project
- **Identifying participants of interest**
- Finding and examining variant data
- Finding raw read data
- Platforms for data analysis: CWL, Arvados, bcbio
- Running an interoperable analysis on PGP data
- Examine structural variant and HLA results

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

Outline

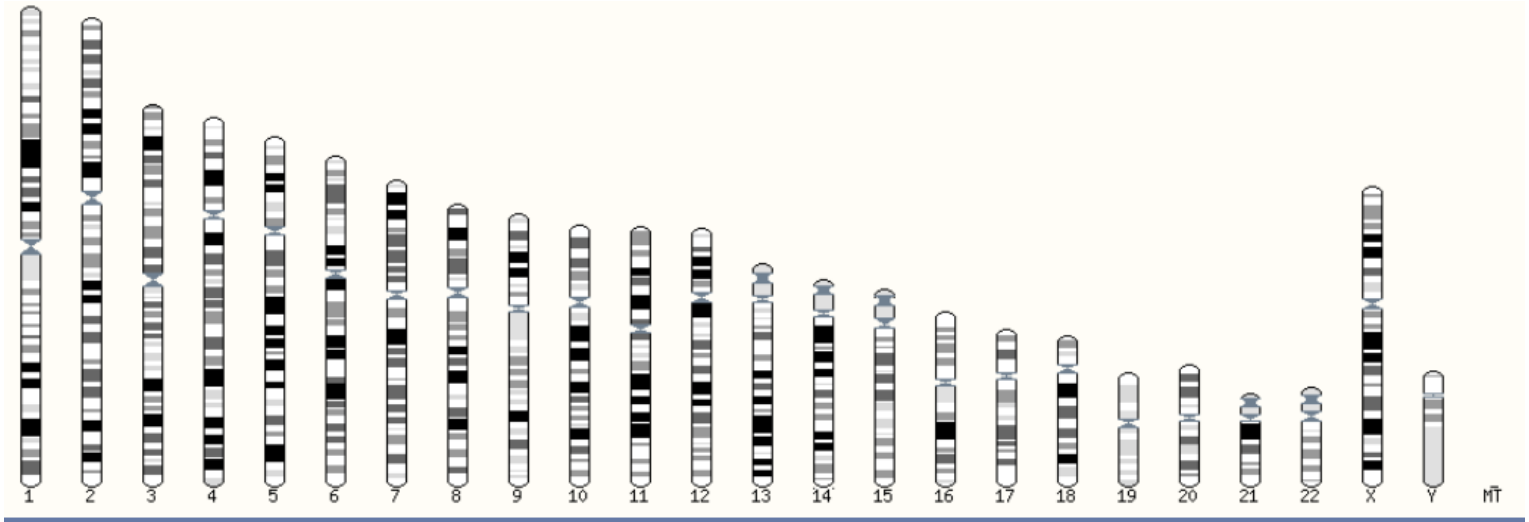
- Overview of the Personal Genome Project
- Identifying participants of interest
- **Finding and examining variant data**
- Finding raw read data
- Platforms for data analysis: CWL, Arvados, bcbio
- Running an interoperable analysis on PGP data
- Examine structural variant and HLA results

Examine existing variation files

- Portable VCFs with small variant data
- Hosted as data collection with standard wget retrieval

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

Human whole genome sequencing



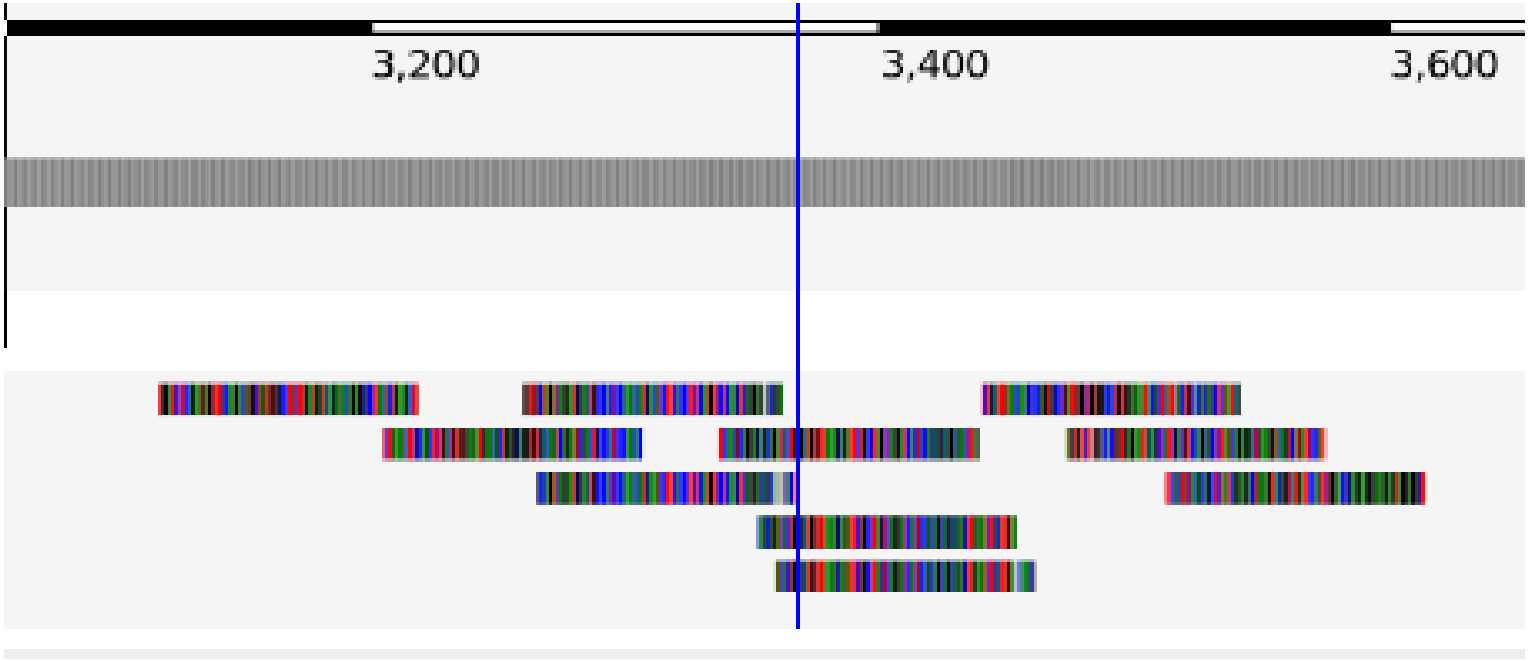
Click on the image above to jump to a chromosome, or click and drag to select a region

Summary

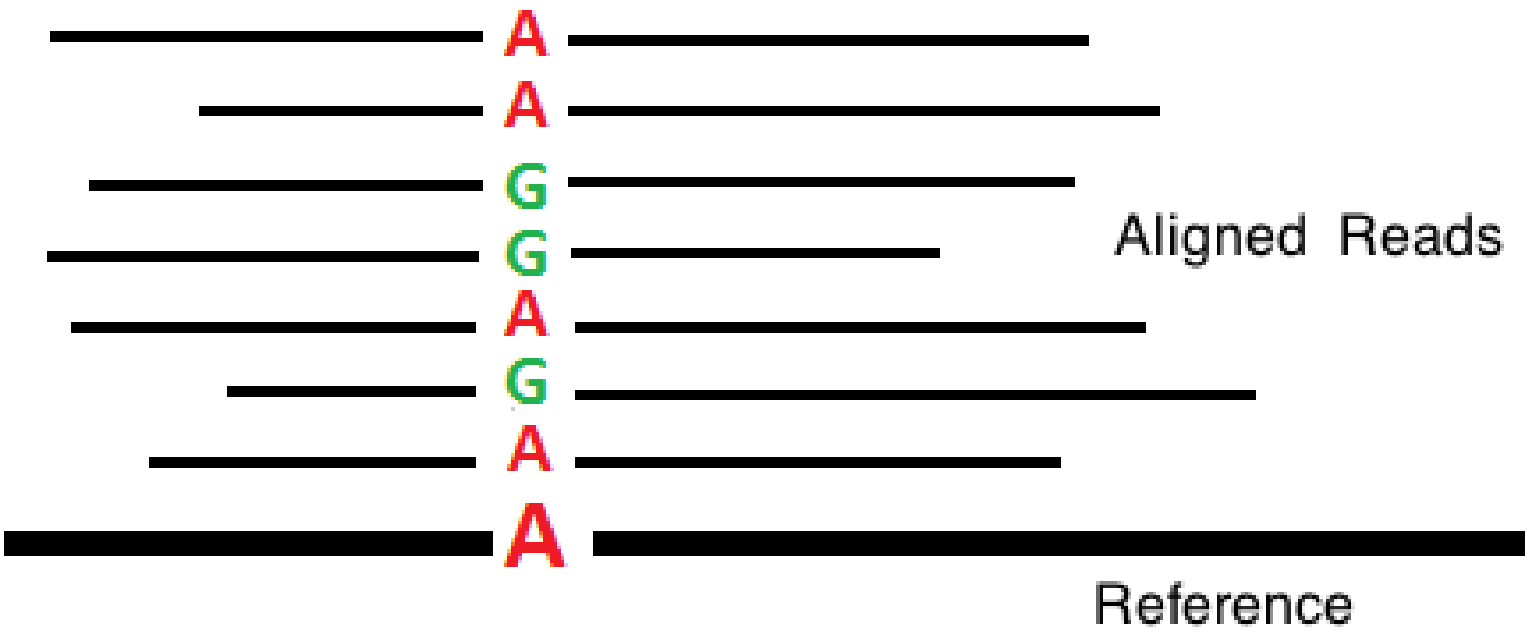
Assembly	GRCh37.p13 (Genome Reference Consortium Human Reference 37), INSDC Assembly GCA_000001405.14, Feb 2009
Database version	75.37
Base Pairs	3,326,743,047

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

High throughput sequencing

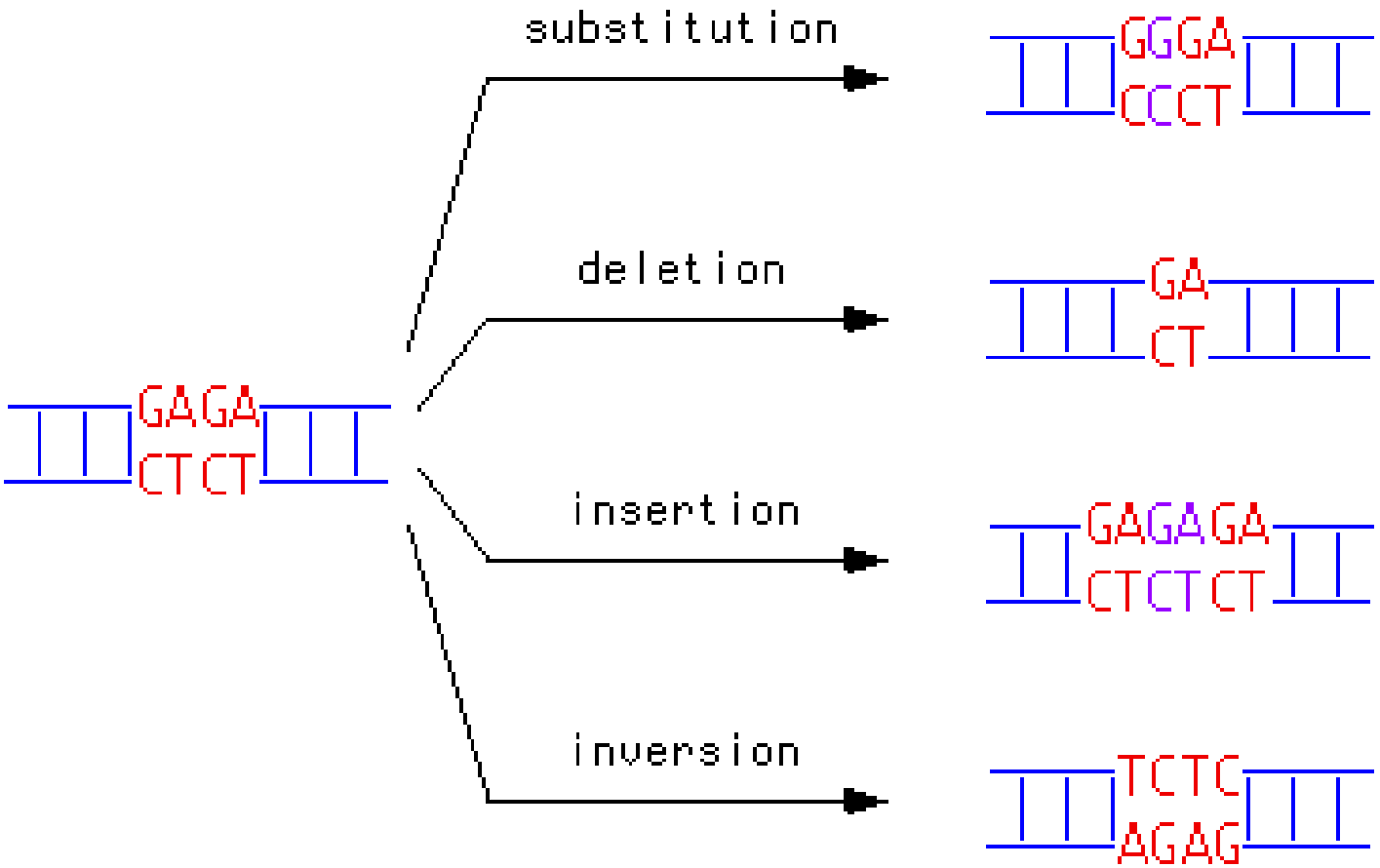


Variant calling



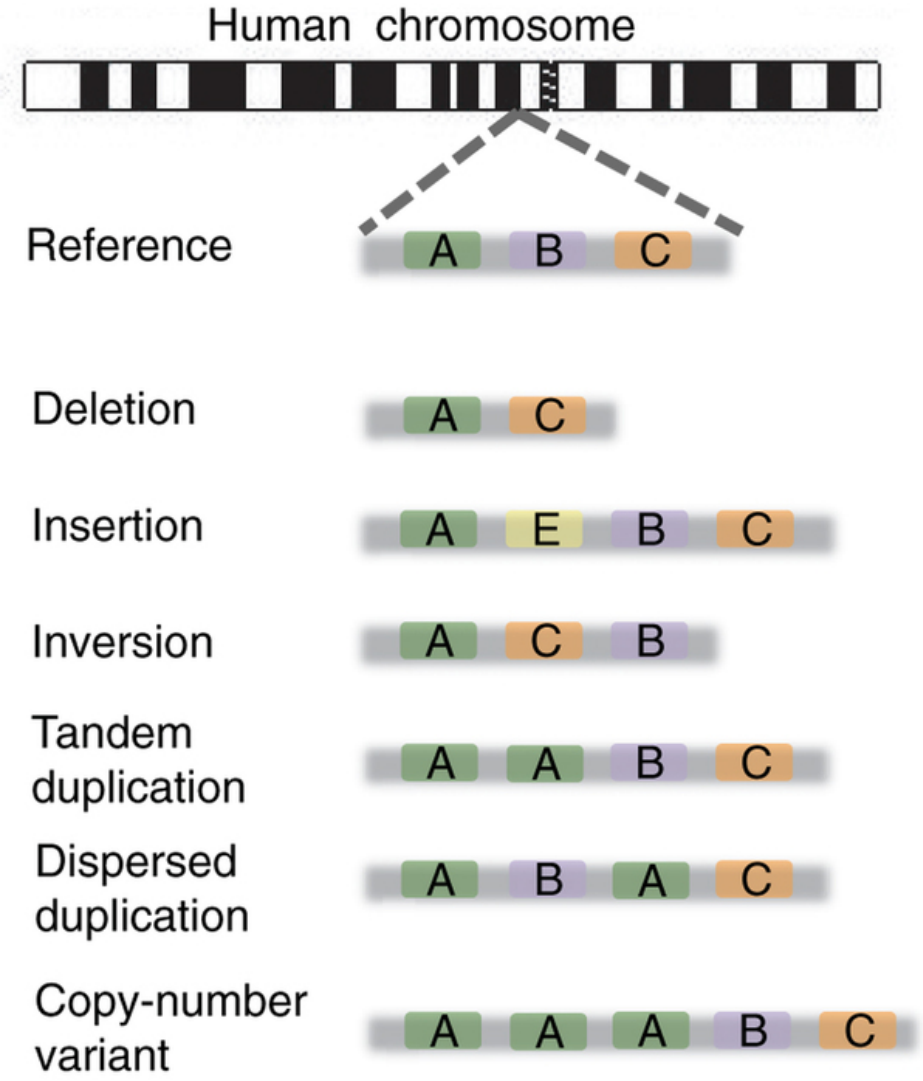
http://en.wikipedia.org/wiki/SNV_calling_from_NGS_data

SNPs and Indels



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

Structural variations



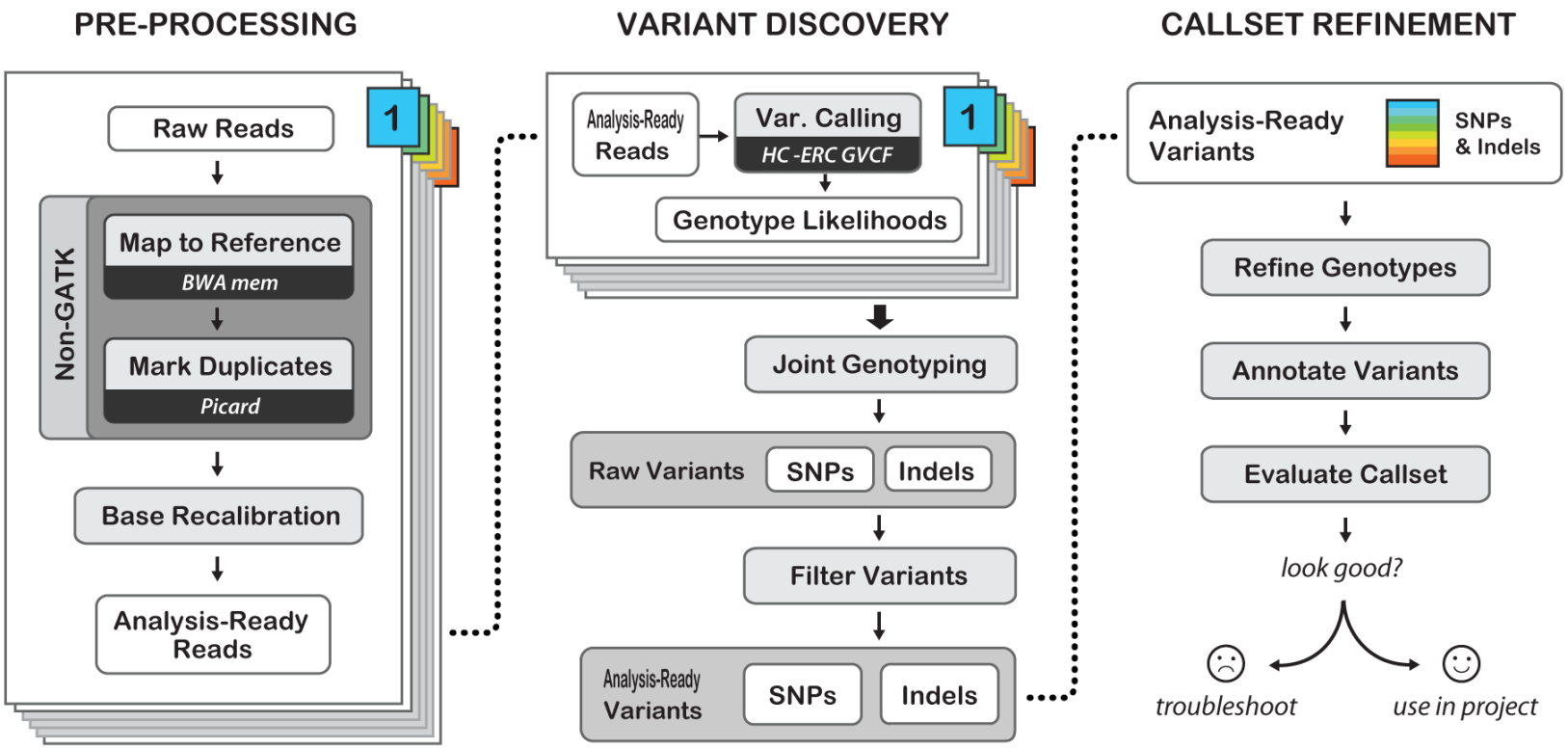
http://www.nature.com/nmeth/journal/v9/n2/full/nmeth.1858.html

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/

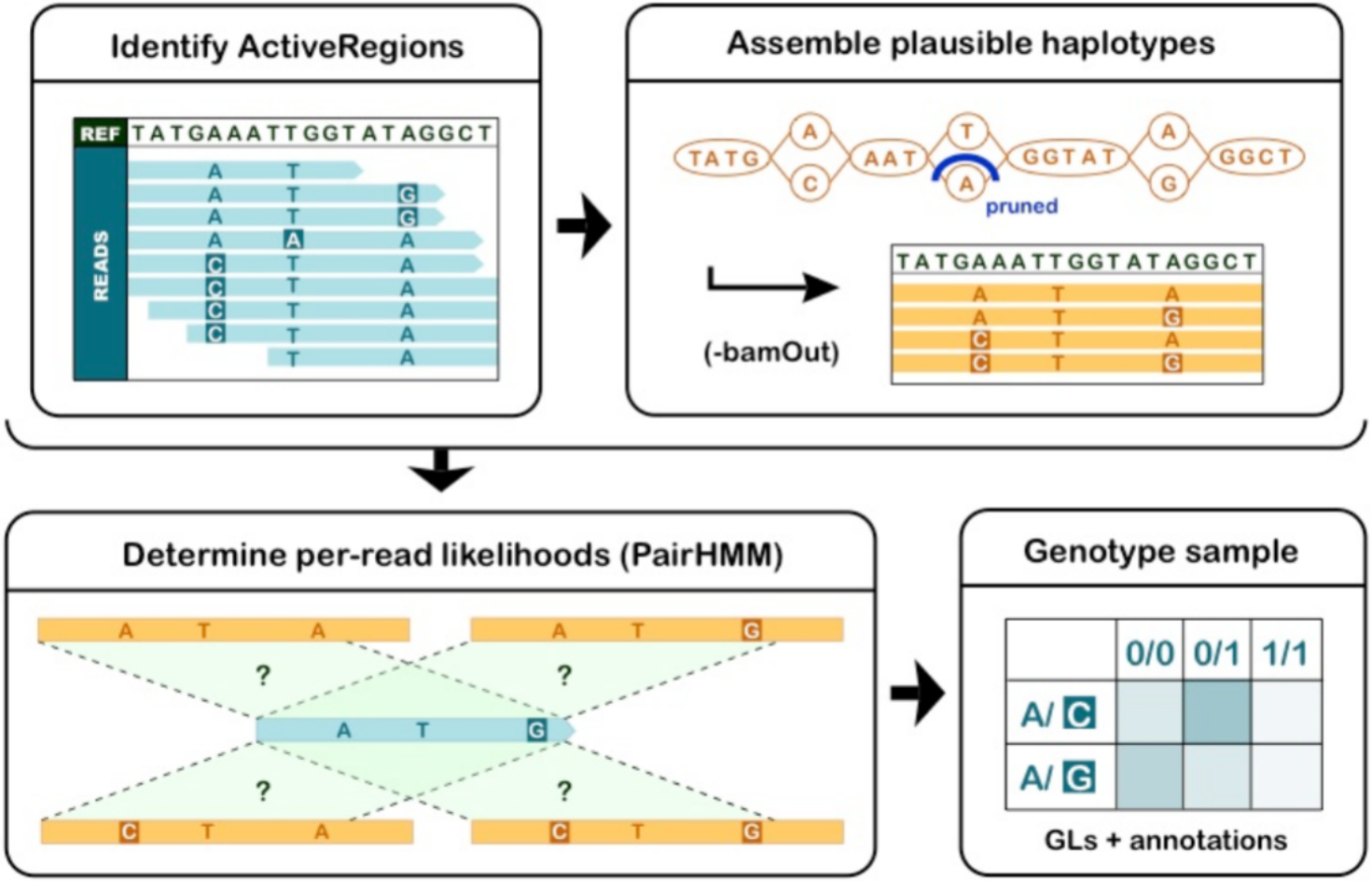
GATK Best Practices



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

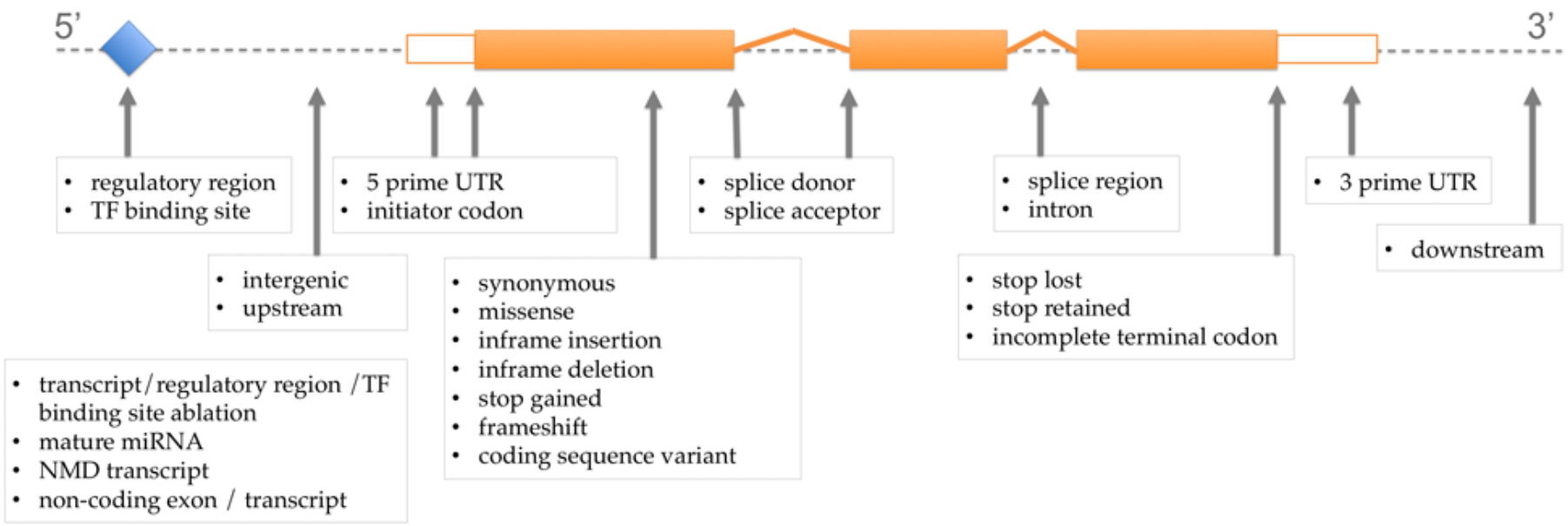
https://software.broadinstitute.org/gatk/best-practices/

HaplotypeCaller



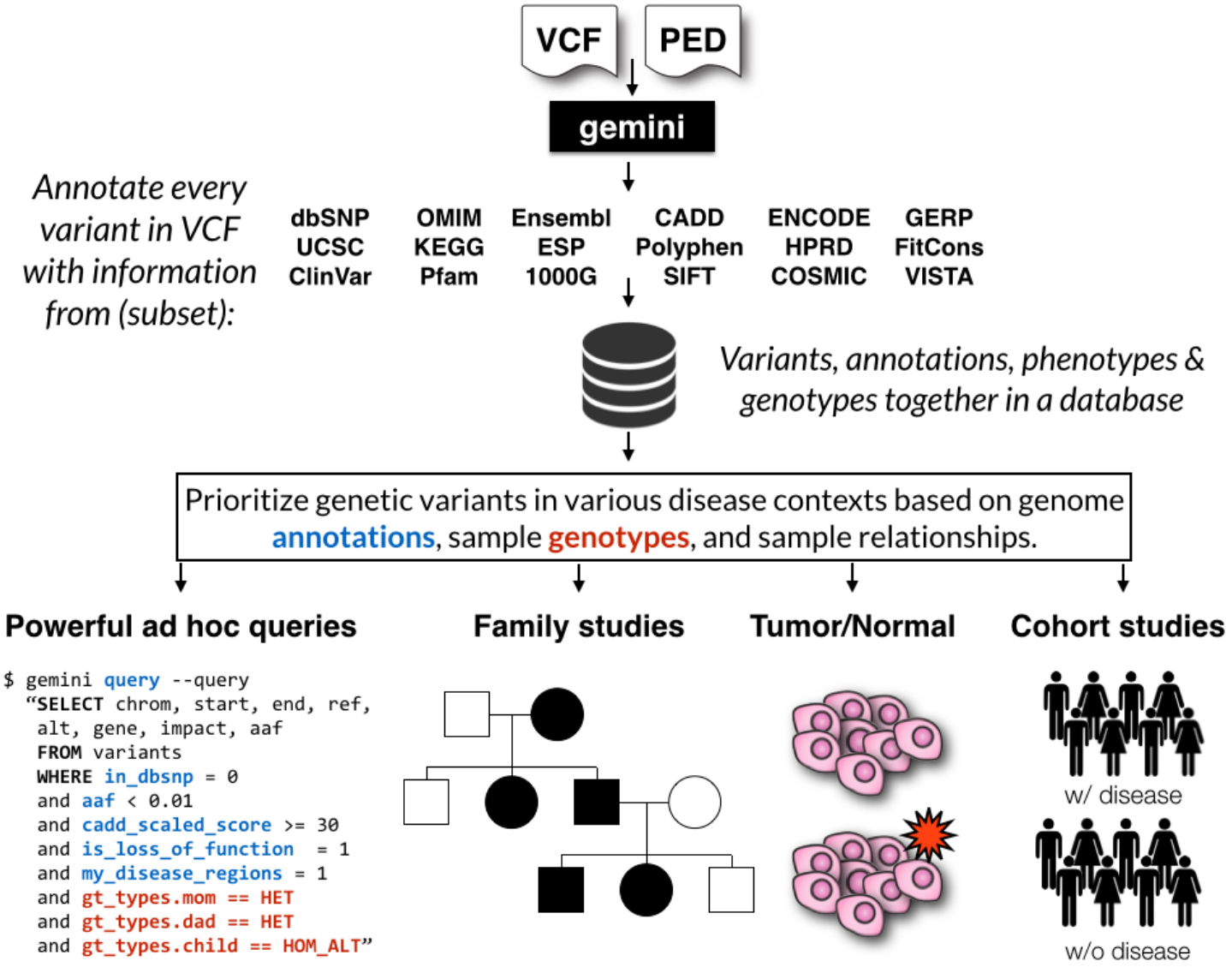
http://gatkforums.broadinstitute.org/discussion/5464/workshop-presentations-2015-uk-4-20-24

Effects prediction



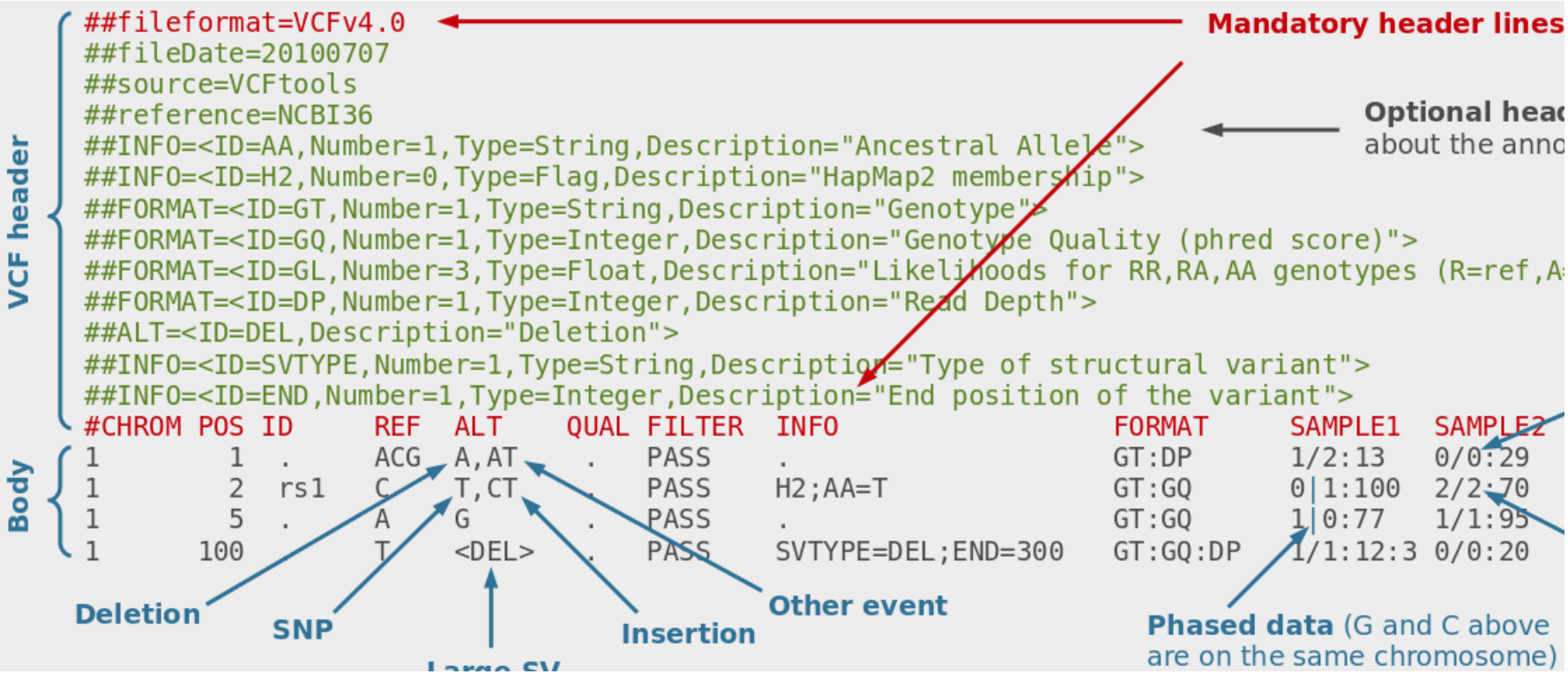
http://www.ensembl.org/info/genome/variation/predicted_data.html

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/

- 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

- Query and outcomes

http://bit.ly/pgp-analysis

Outline

- Overview of the Personal Genome Project
- Identifying participants of interest
- Finding and examining variant data
- **Finding raw read data**
- Platforms for data analysis: CWL, Arvados, bcbio
- Running an interoperable analysis on PGP data
- Examine structural variant and HLA results

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

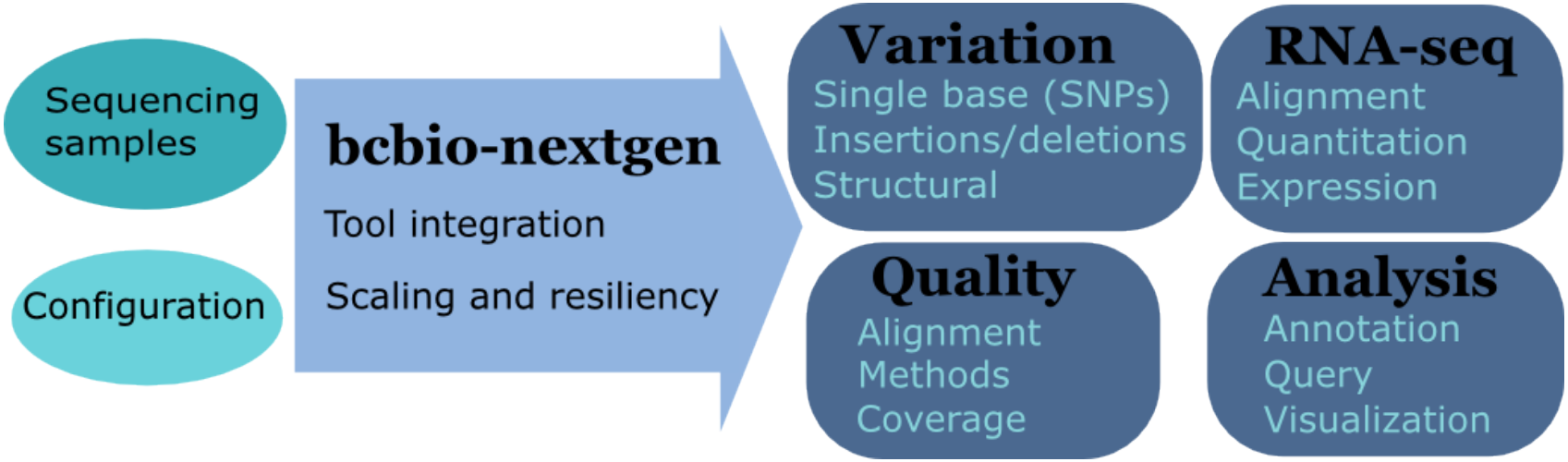
Outline

- Overview of the Personal Genome Project
- Identifying participants of interest
- Finding and examining variant data
- Finding raw read data
- **Platforms for data analysis: CWL, Arvados, bcbio**
- Running an interoperable analysis on PGP data
- Examine structural variant and HLA results

Build open source communities

https://gccbosc2018.sched.com/

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

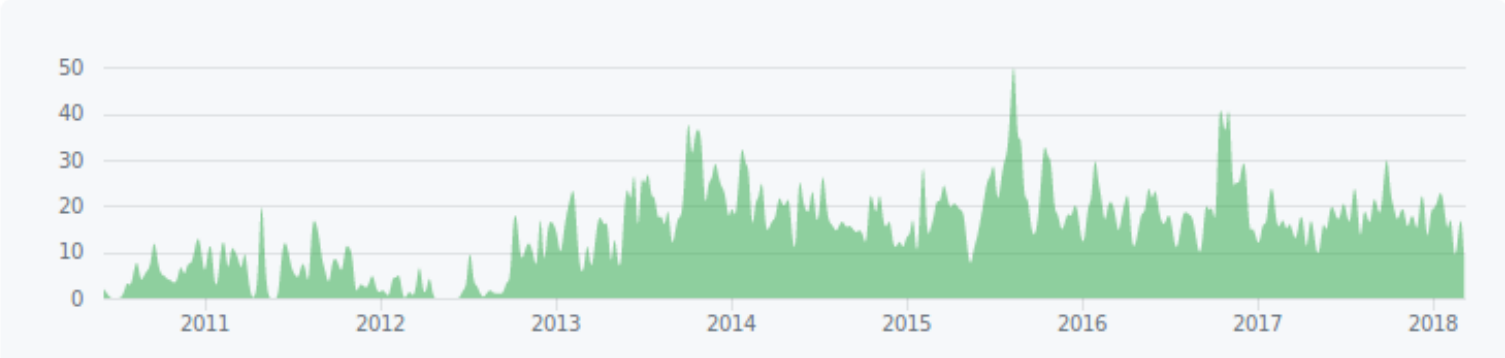
https://bcbio-nextgen.readthedocs.org/en/latest/contents/pipelines.html

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

Many great 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

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
  - Rabix Bunny – local
  - Toil – HPC, local
  - DNAnexus – AWS, Azure
  - Seven Bridges + Cancer Genomics Cloud
- Stop maintaining bcbio specific infrastructure
- Focus on hard biological problems

Unique goals with CWL

- Multiple concurrent production environments
  - HPC
  - Platforms (Arvados, DNAnexus, SevenBridges)
  - Direct on Cloud (AWS, GCP, Azure)
- Coordinated release and update process
  - Workflow
  - Tools in containers
  - Reference data

Outline

- Overview of the Personal Genome Project
- Identifying participants of interest
- Finding and examining variant data
- Finding raw read data
- Platforms for data analysis: CWL, Arvados, bcbio
- **Running an interoperable analysis on PGP data**
- Examine structural variant and HLA results

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

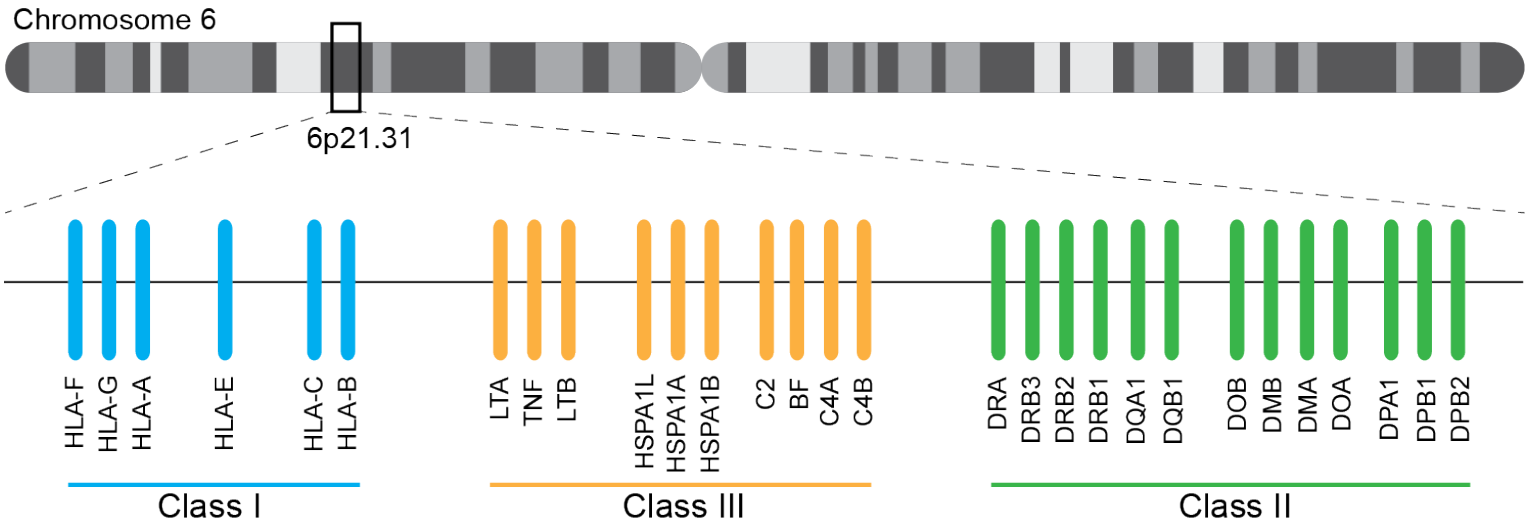
```
bcbio_vm.py cwlrun arvados pgp_sv_hla-workflow -- \
  --project-uuid su92l-j7d0g-eoibug3nrwg8ysj
```

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

Outline

- Overview of the Personal Genome Project
- Identifying participants of interest
- Finding and examining variant data
- Finding raw read data
- Platforms for data analysis: CWL, Arvados, bcbio
- Running an interoperable analysis on PGP data
- **Examine structural variant and HLA results**

Major histocompatibility complex (MHC) – HLAs



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

HLA typing

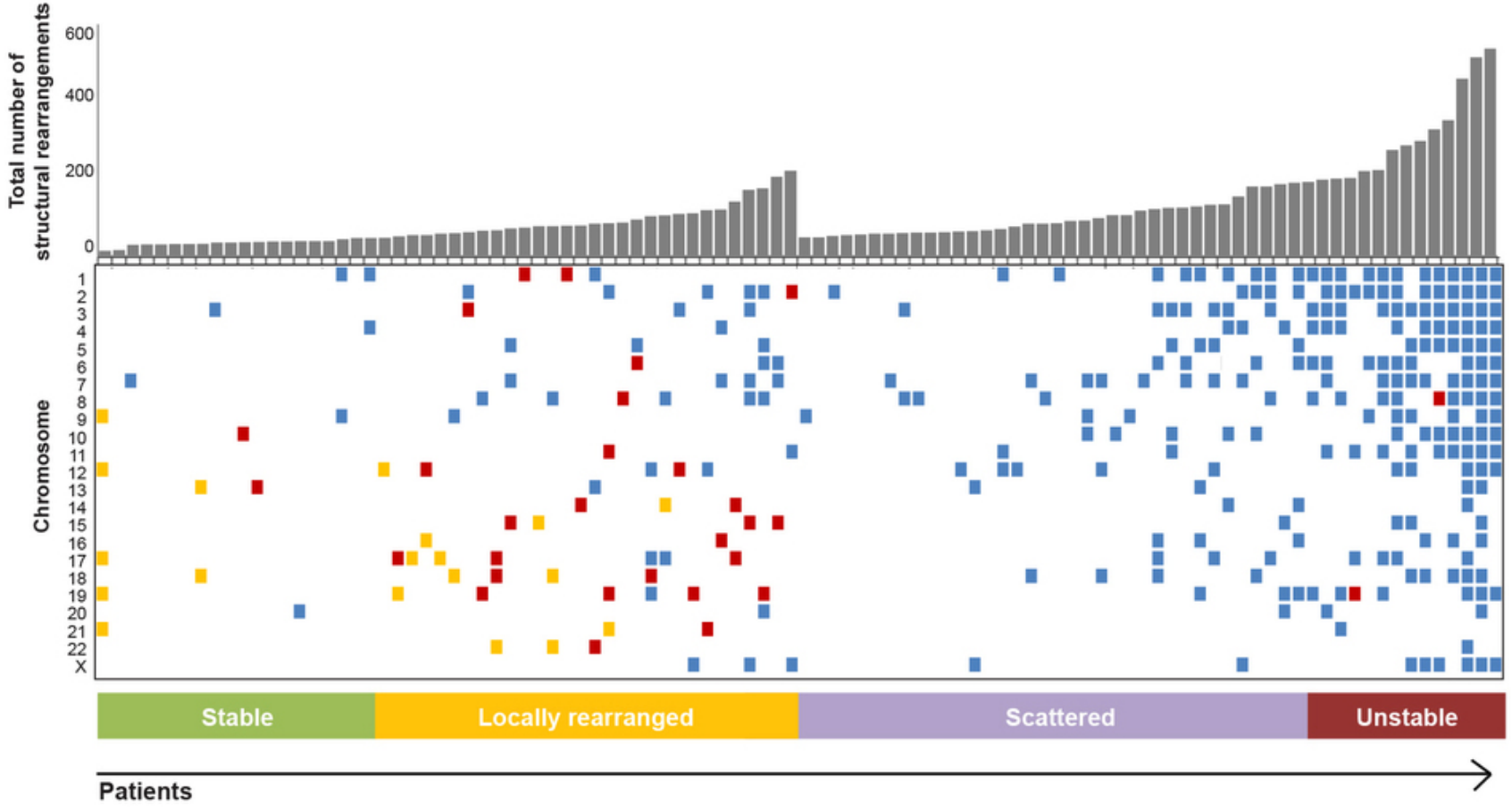
- 1000 genomes: build 38 + IMGT/HLA-3.18.0
- bwa mem extracts HLA reads
- Map reads only to HLA sequences
- OptiType: Call HLA types

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

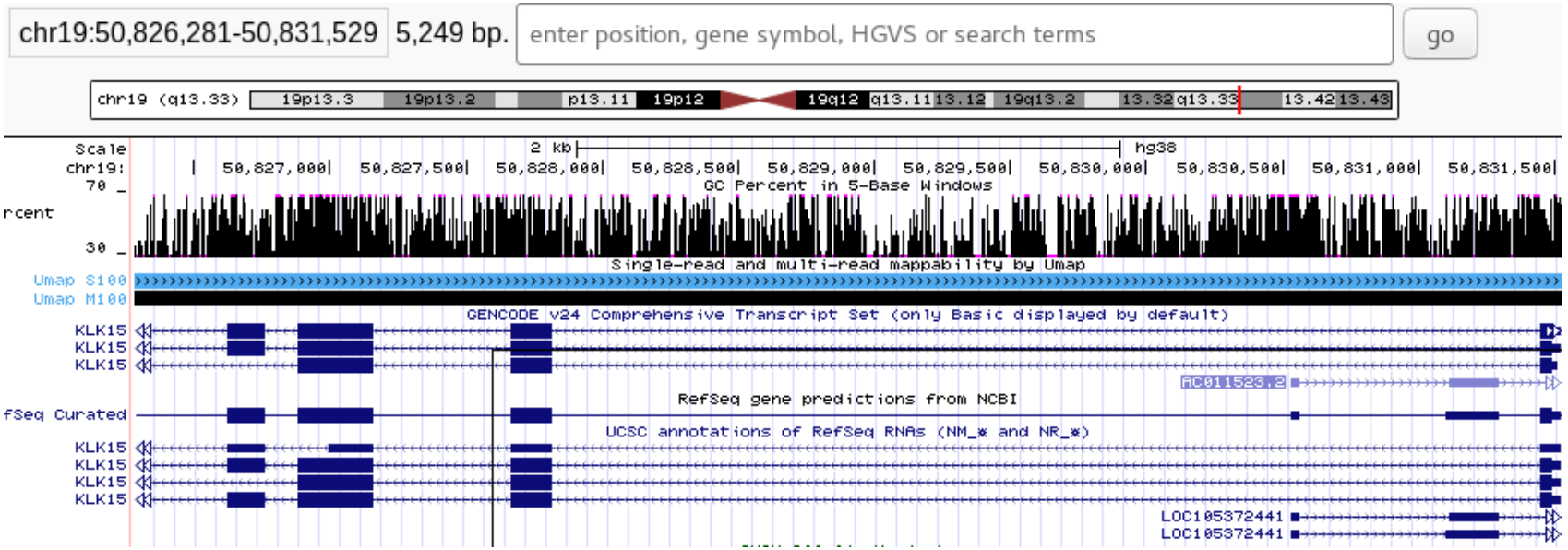
Tools used

- 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
```

Genomic region with deletion – KLK15



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

KLK15 known function

KLK15

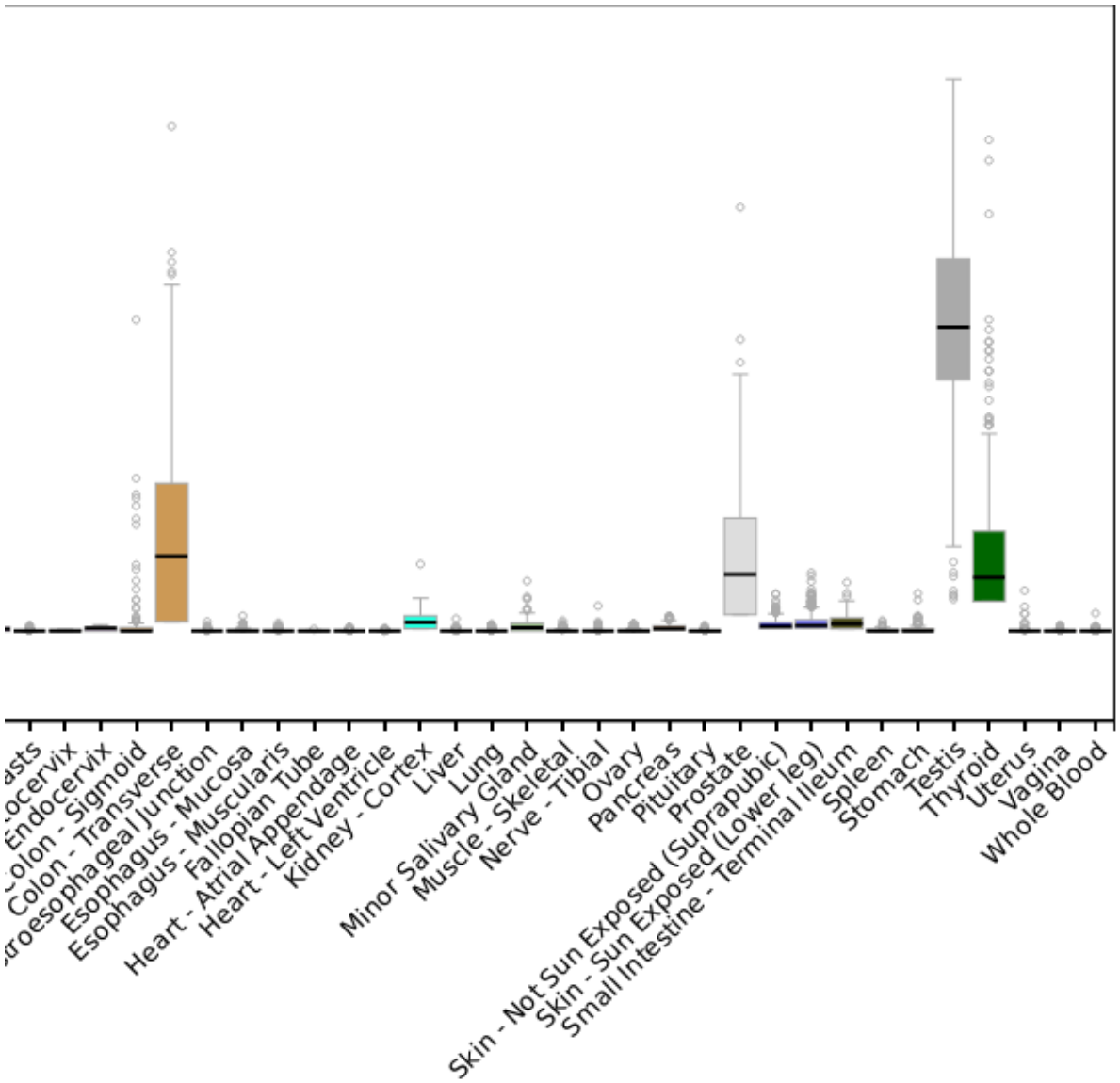
From Wikipedia, the free encyclopedia

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

Summary

- Overview of the Personal Genome Project
- Identifying participants of interest
- Finding and examining variant data
- Finding raw read data
- Platforms for data analysis: CWL, Arvados, bcbio
- Running an interoperable analysis on PGP data
- Running structural variant and HLA analyses