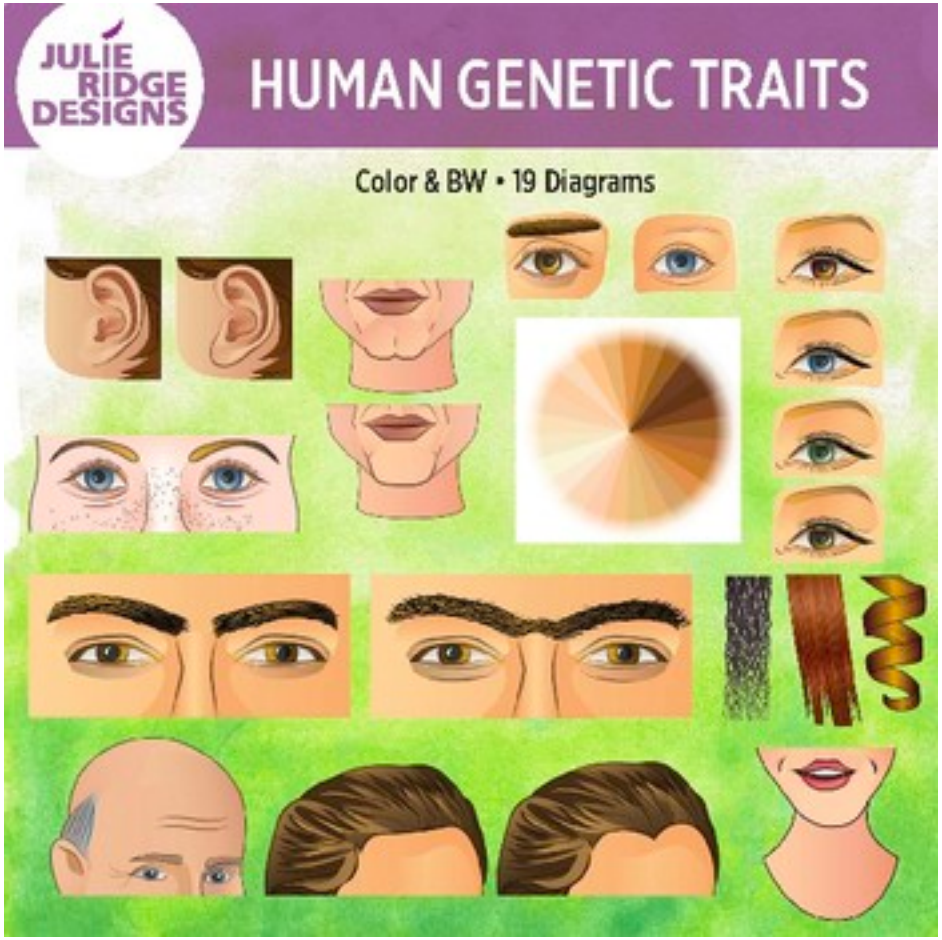
# Mutation Identification in Genomics

- What is a genetic disease?
- What is a genome, exome and Gene Panel
- What is Variation
  - Somatic vs Germline
  - SNVs, Indels and Structural Variation

- Is there an easy way to run all those command line programs?
  - BioHPC Astrocyte

# Genetic Traits

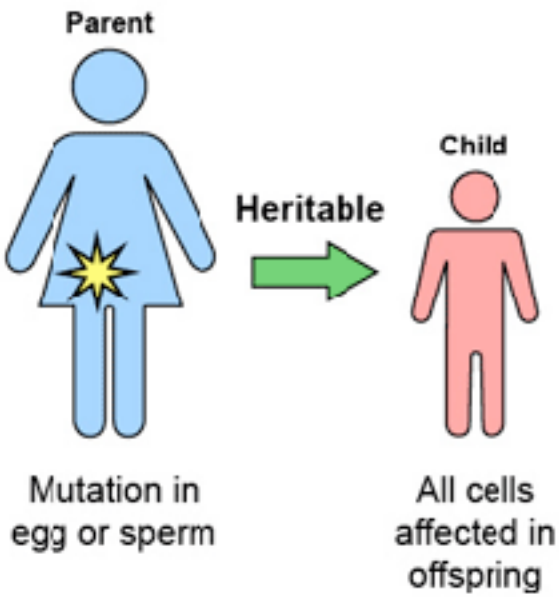
- A phenotype is an individual's observable traits, such as height, eye color, and blood type.
- The genetic contribution to the phenotype is called the genotype.
- Some traits are largely determined by the genotype, while other traits are largely determined by environmental factors.



# Genetic Disease

- A genetic disorder is a genetic problem caused by one or more abnormalities in the genome.
- A single-gene disorder is the result of a single mutated gene.
- Autosomal dominant disorders occur with only one mutated copy of the gene.
- Recessive disorders require both copies are mutated.
- X-linked dominant disorders are caused by mutations in genes on the X chromosome.
- Mitochondrial disease, also known as maternal inheritance, applies to genes encoded by mitochondrial DNA.
- Genetic disorders may also be complex, multifactorial, or polygenic, meaning they are associated with the effects of multiple genes in combination with lifestyles and environmental factors.

# Acquired vs Inherited Variation

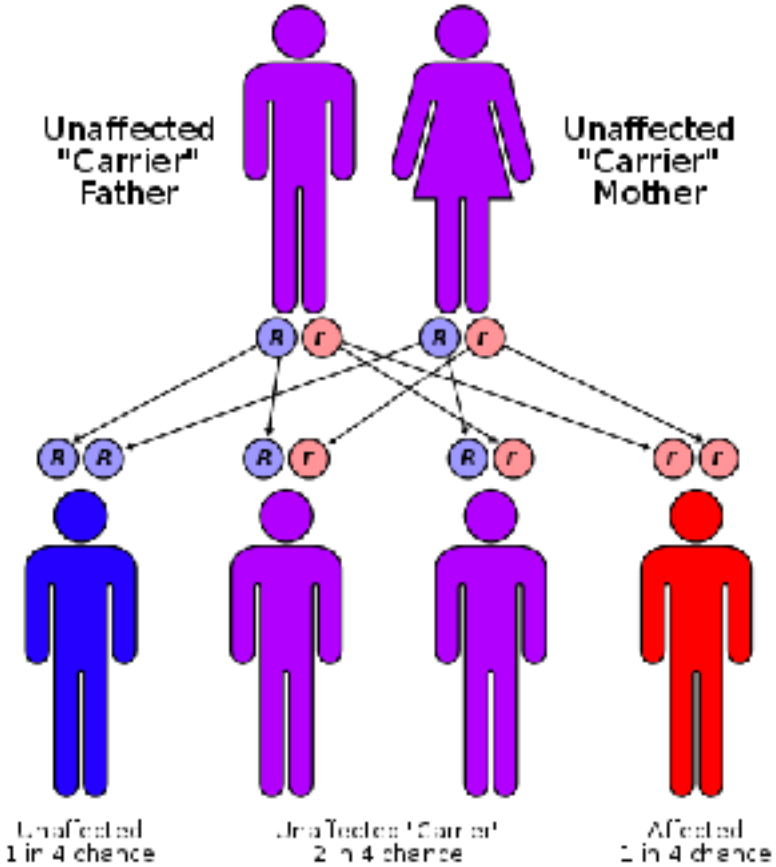



Germline

Somatic

# Mendelian Disease

| Disorder prevalence (approximate) | |
|---|---|
| **Autosomal dominant** | |
| Familial hypercholesterolemia | 1 in 500 |
| Polycystic kidney disease | 1 in 1250 |
| Neurofibromatosis type I | 1 in 2,500 |
| Hereditary spherocytosis | 1 in 5,000 |
| Marfan syndrome | 1 in 4,000 |
| Huntington's disease | 1 in 15,000 |
| **Autosomal recessive** | |
| Sickle cell anaemia | 1 in 625 |
| Cystic fibrosis | 1 in 2,000 |
| Tay-Sachs disease | 1 in 3,000 |
| Phenylketonuria | 1 in 12,000 |
| Mucopolysaccharidoses | 1 in 25,000 |
| Lysosomal acid lipase deficiency | 1 in 40,000 |
| Glycogen storage diseases | 1 in 50,000 |
| Galactosemia | 1 in 57,000 |
| **X-linked** | |
| Duchenne muscular dystrophy | 1 in 7,000 |
| Hemophilia | 1 in 10,000 |



# Somatic/Mosaic Disease

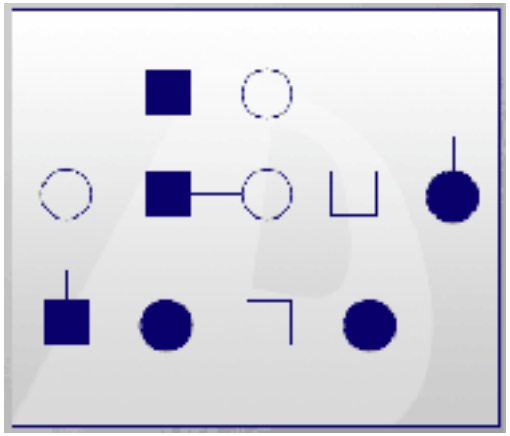
- Acquired diseases are caused by acquired mutations in a gene or group of genes that occur during a person's life.
- These include many cancers, as well as some forms of neurofibromatosis.
- Mosaicism, involves the presence of two or more populations of cells with different genotypes in one individual, who has developed from a single fertilized egg.
- Intersex conditions can be caused by mosaicism where some cells in the body have XX and others XY chromosomes
- Other endogenous factors can also lead to mosaicism including mobile elements, DNA polymerase slippage, and unbalanced chromosomal segregation.
- Exogenous factors include nicotine and UV radiation

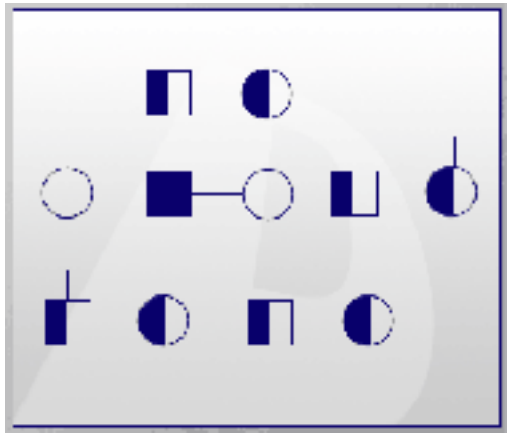
# Complex Disease

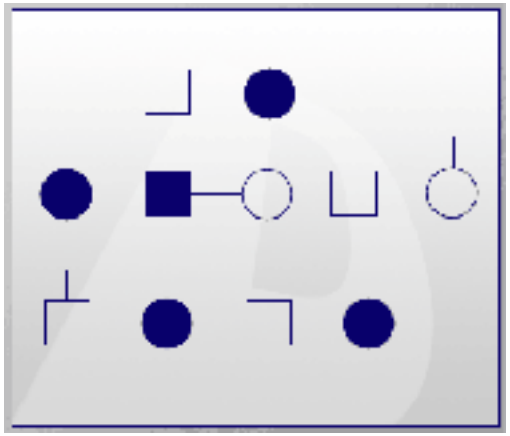
- Complex diseases are caused by a combination of genetic, environmental, and lifestyle factors, most of which have not yet been identified.
- Some examples include Alzheimer's disease, scleroderma, asthma, Parkinson's disease, multiple sclerosis, osteoporosis, connective tissue diseases, kidney diseases, autoimmune diseases, etc

# Pedigrees

- Identification of disease causing variation was originally done using pedigrees (multigenerational family studies)







# Genome

- A genome is the entire set of genetic material for an organism.
- The human genome consists of about 3 billion base pairs of DNA across 23 pairs of chromosomes.
- More than 99 percent of the human genome is the same in all people.
- That means that differences in less than 1 percent of our genome accounts for the vast diversity of humans across the globe.

# Exome

- The exome is a subset of the genome that contains protein coding genes.
- Exons are also referred to as the coding region of a gene
- The exons of all our genes make up approximately 1.5% of our genome and are collectively referred to as the “exome”.
- There are some important DNA sequences that are not contained within the exome in noncoding DNA that have important biological functions, such as regulating the coding regions of the genome.

# Gene Panels

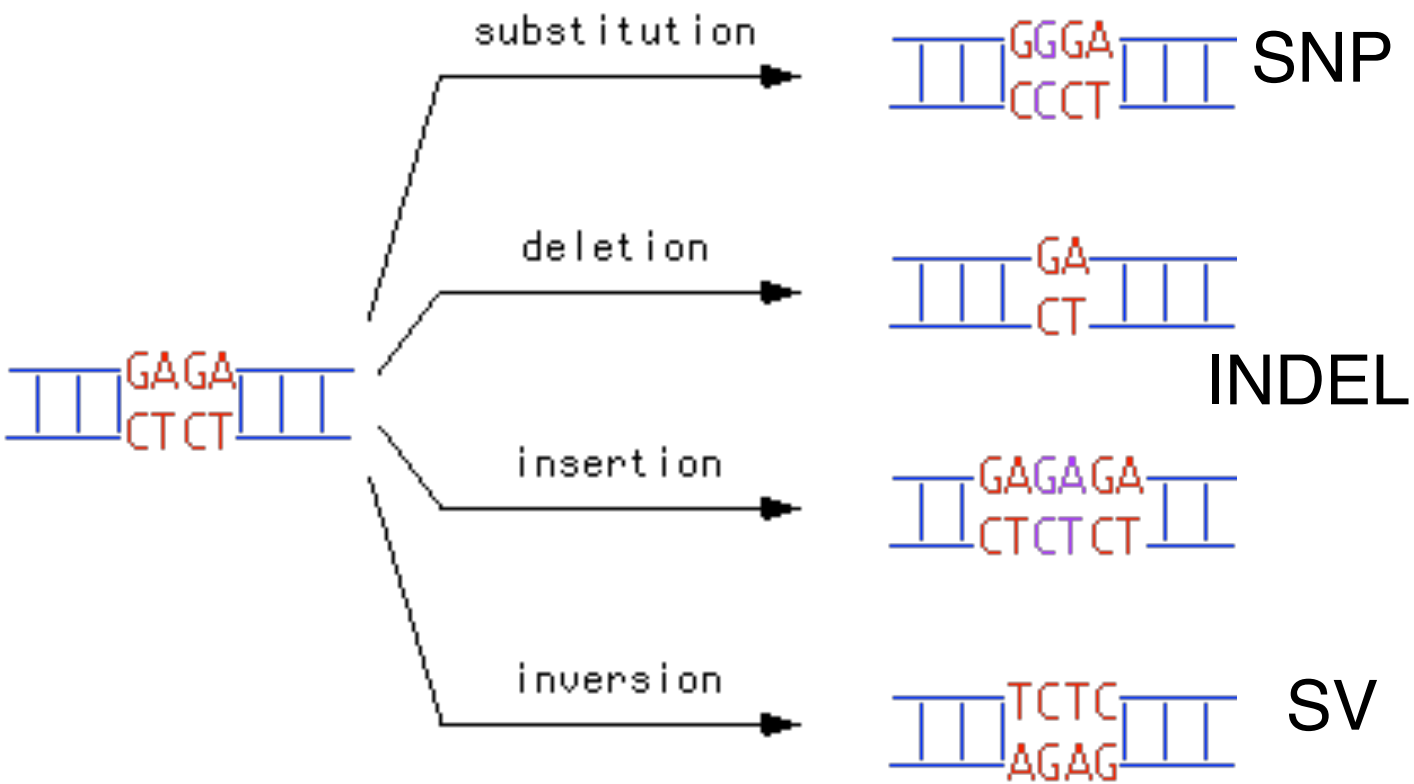
- A gene panel is a gene subset of the exome
- It contains a subset of exons for a select group of genes
- Gene Panels are useful if you need to do deep sequencing > 1000X
- Many clinical tumor tests use gene panels.

# Pros and Cons of WGS vs Targeted Panels

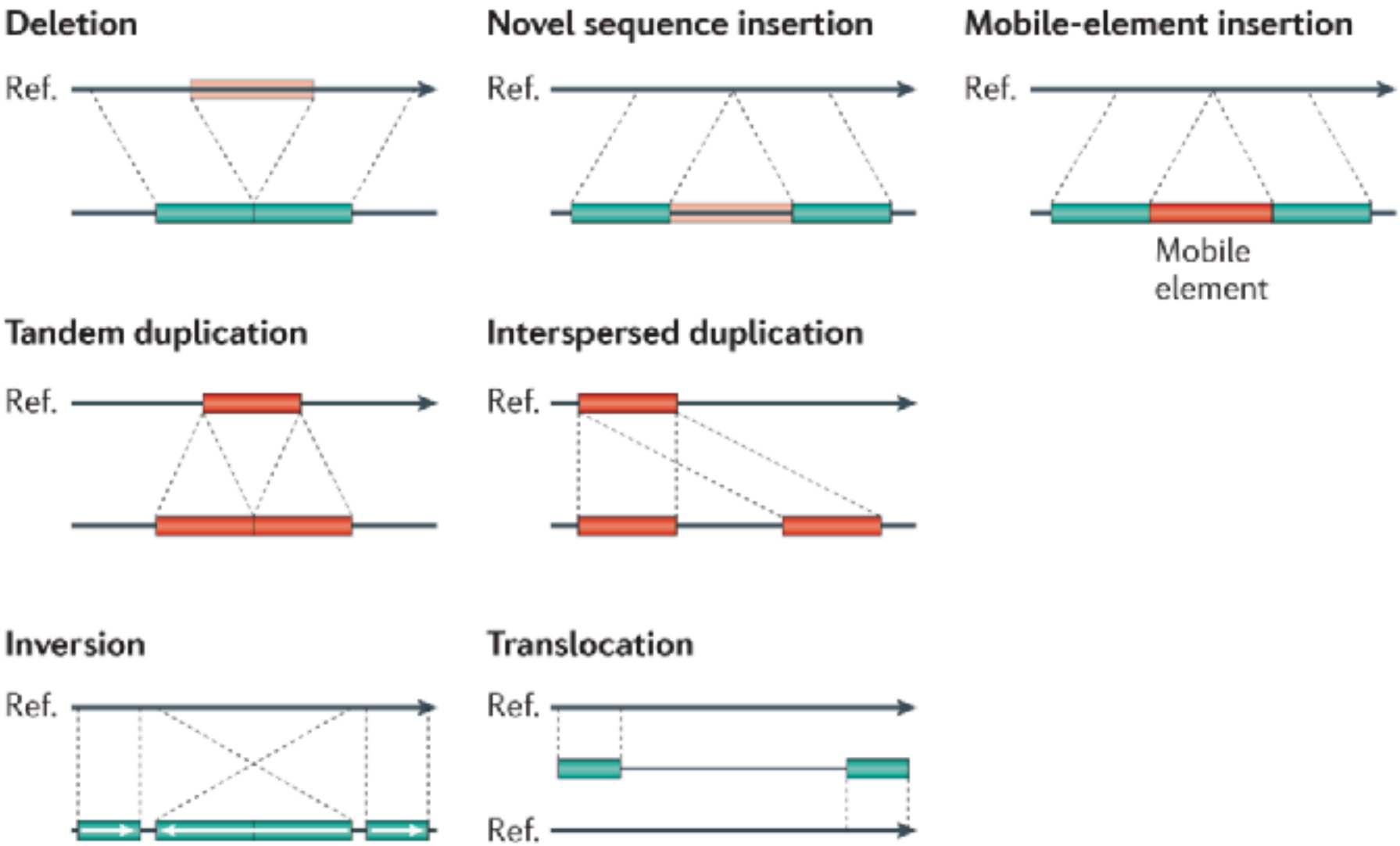
- Whole Genome Sequence can better predict large structural changes including CNV, large Indels, etc
- Whole Genome has more uniform coverage of the protein coding regions
- ~$1300 30-40X coverage

- Targeted panes are cheaper
- Whole Exome Sequencing costs ~$500 for 100X coverage
- In somatic/mosaic conditions you might need > 1000X coverage.
- Generate less data to store and analyze

# Types of Variation

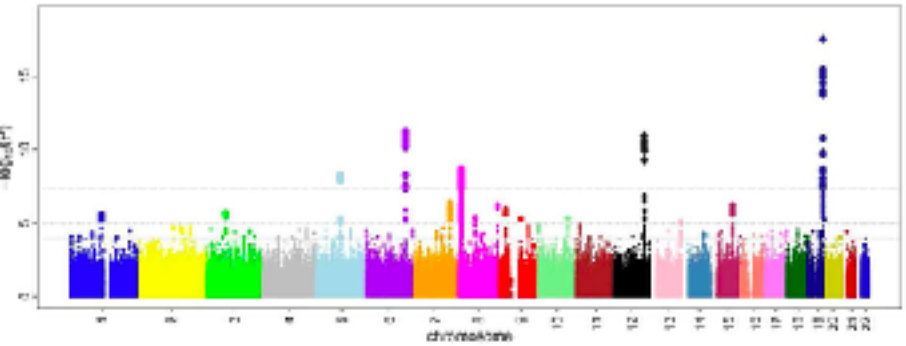


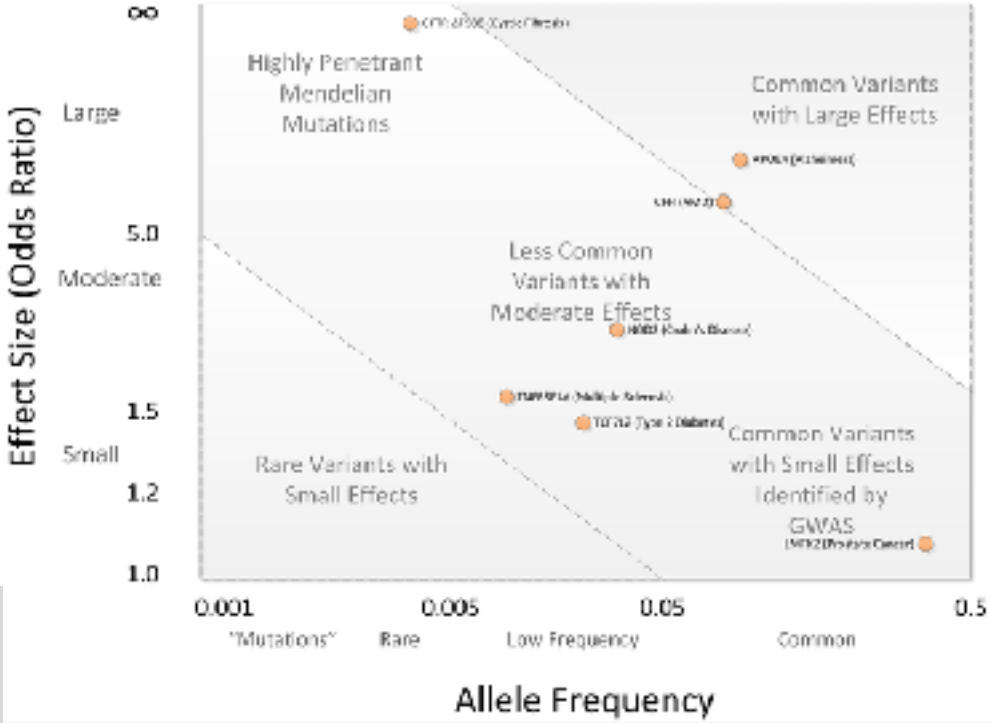
# Types of Structural Variation



# GWAS

- Genome Wide Association studies examines associations between single-nucleotide polymorphisms (SNPs) and traits using statistical methods like Fisher Exact Test
- Often these associations have varying contributions to the trait (effect size).





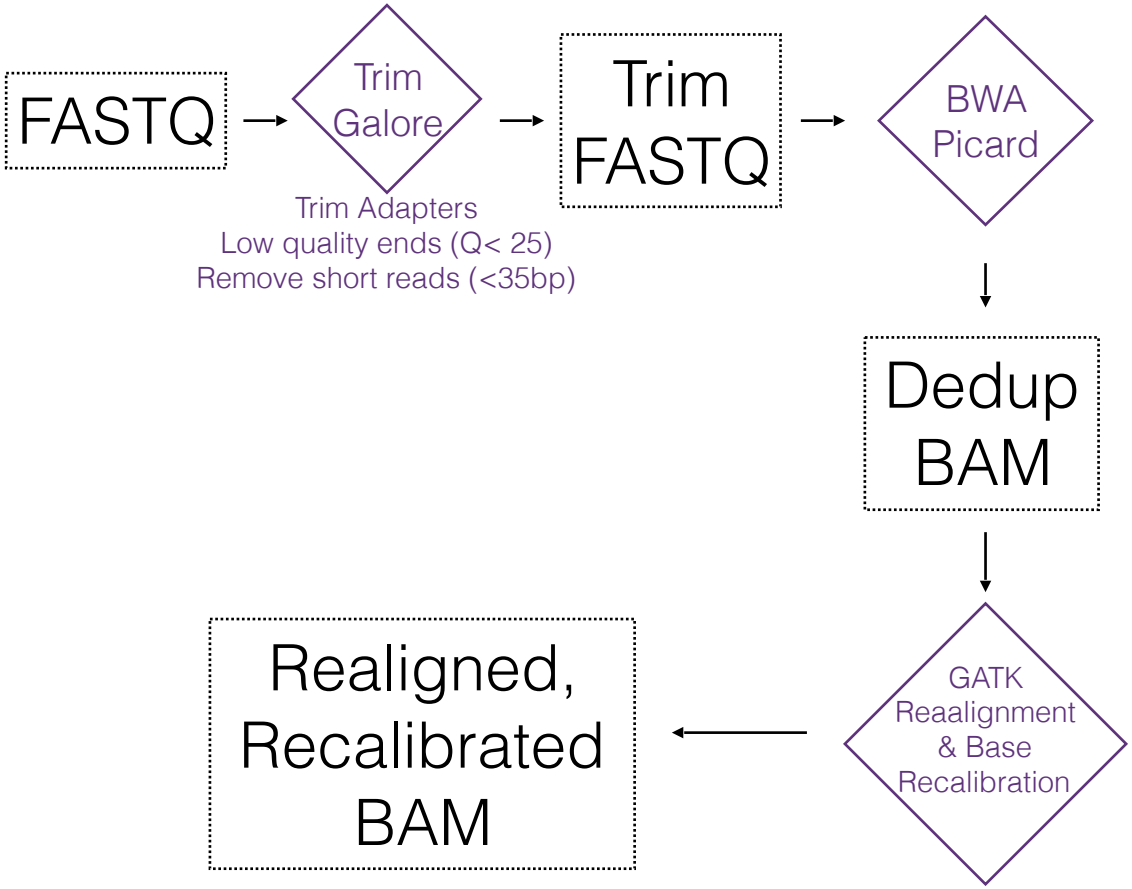
# PheWAS

- Phenome-wide association studies (PheWAS) is a quantitative method to determine disease associations can we make with a given gene?
- This is in contrast to GWAS which aims to identify associations, PheWAS aims to explain the cause and effect.
- For example, given a single nucleotide polymorphism (SNP) identified by GWAS (SNP: rs17234657) and association with infection, one may conclude that the SNP increases susceptibility of the host.
- In contrast, with PheWAS new putative associations may be identified through interrogation of phenomic markers within the EHR. Hence, an alternative mechanism is identified, where rs17234657 is found to be associated with an increase in autoimmune disease and the treatment used (immunosuppressive medication) is the cause of the infection.

# Large Reference Populations

- HapMap
  - The International HapMap Project was an organization that aimed to develop a haplotype map (HapMap) of the human genome using SNP genotyping arrays
- 1000G
  - The 1000 Genomes project aimed to sequence using NGS > 1000 genomes in “pure” and “ad-mixture” human populations to identify human variation across the genome
- ExAC
  - ExAC collected the SNP and Indel calls in ~ 26K genomes/exomes to accumulation prevelence in the population studied in many genomes projects
- gnomAD
  - The Genome Aggregation Database (gnomAD) is a resource of aggregate genomes and aimed to harmonize both exome and genome sequencing data from over 120K exomes and 15K genomes.

# Alignment Workflows

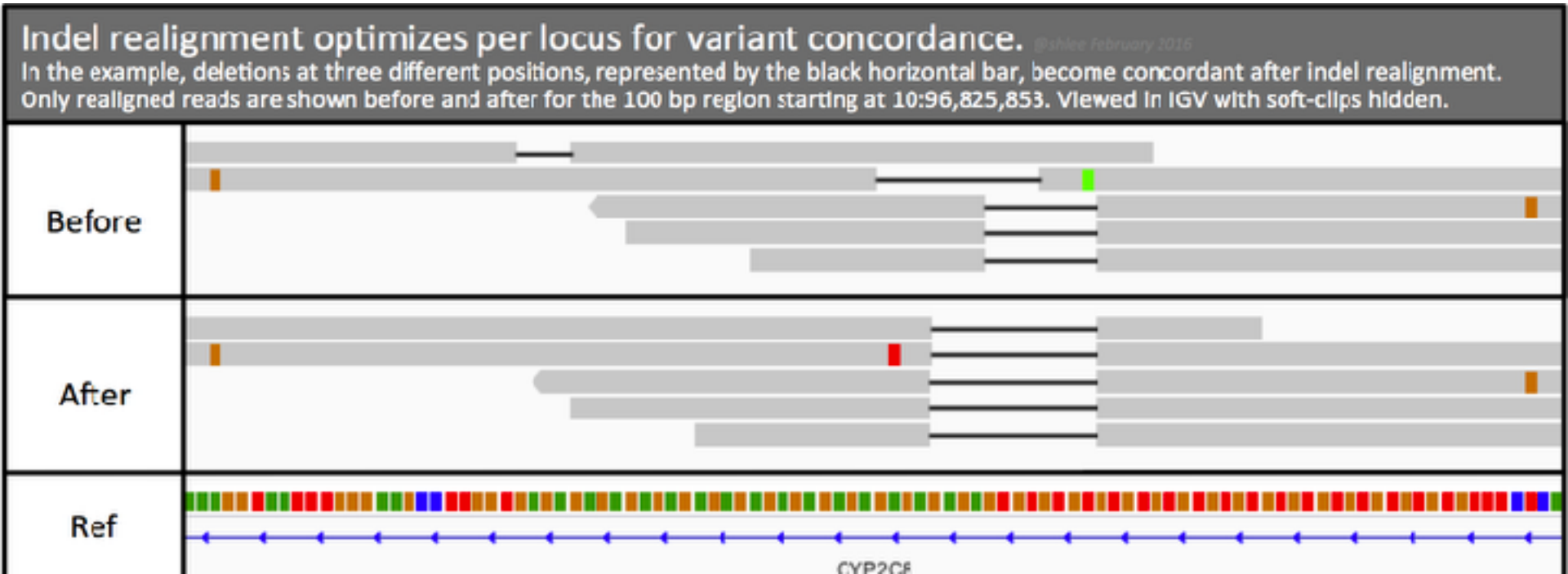


# Why are we so worried about sequence duplication?

- When DNA is sequenced, PCR is used to amplify sequence library to ensure that only DNA with “a known adapter” is sequenced.
- Since PCR has a small error rate, “early errors” can be amplified and could skew your results
- We remove duplicates to remove potential noise.
- Although in my experience in deep sequencing removing duplicates doesn’t really change downstream results

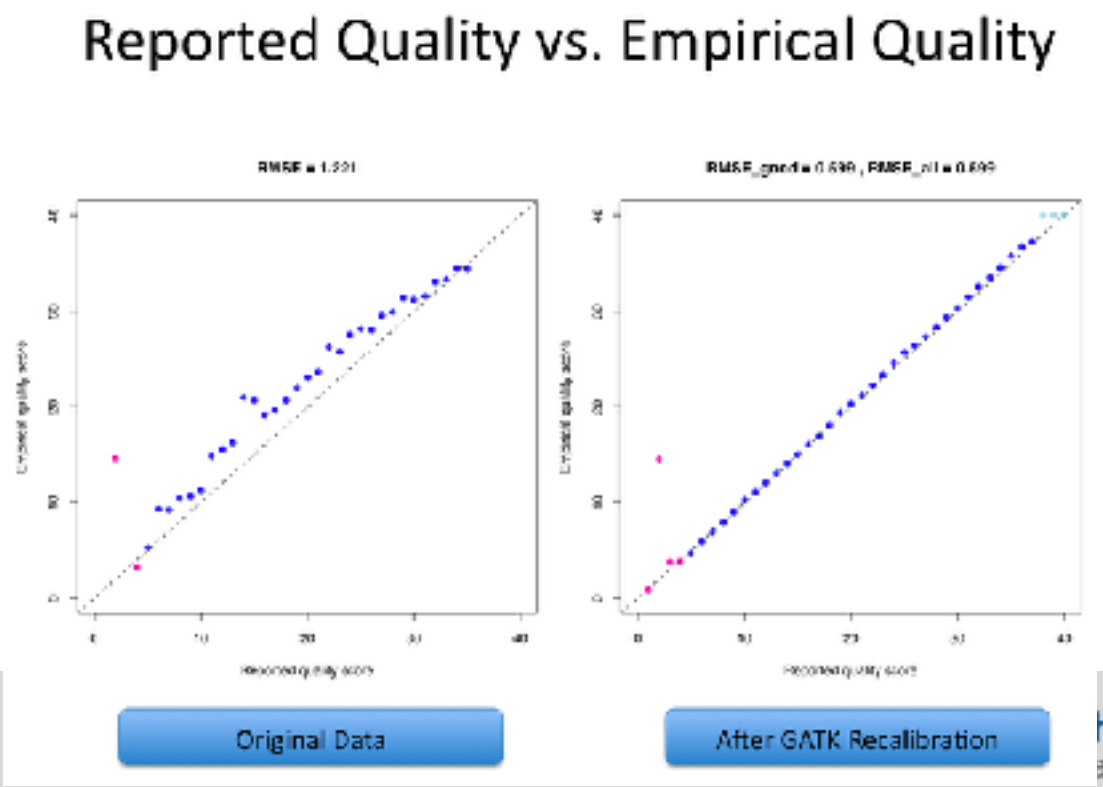
# Why does GATK need Indel Realignment?

- Sometimes, alignment algorithms align reads inconsistently, adding the alignment gaps to different places.
- Indel Realignment uses “known” gold standard indels to realign these gaps

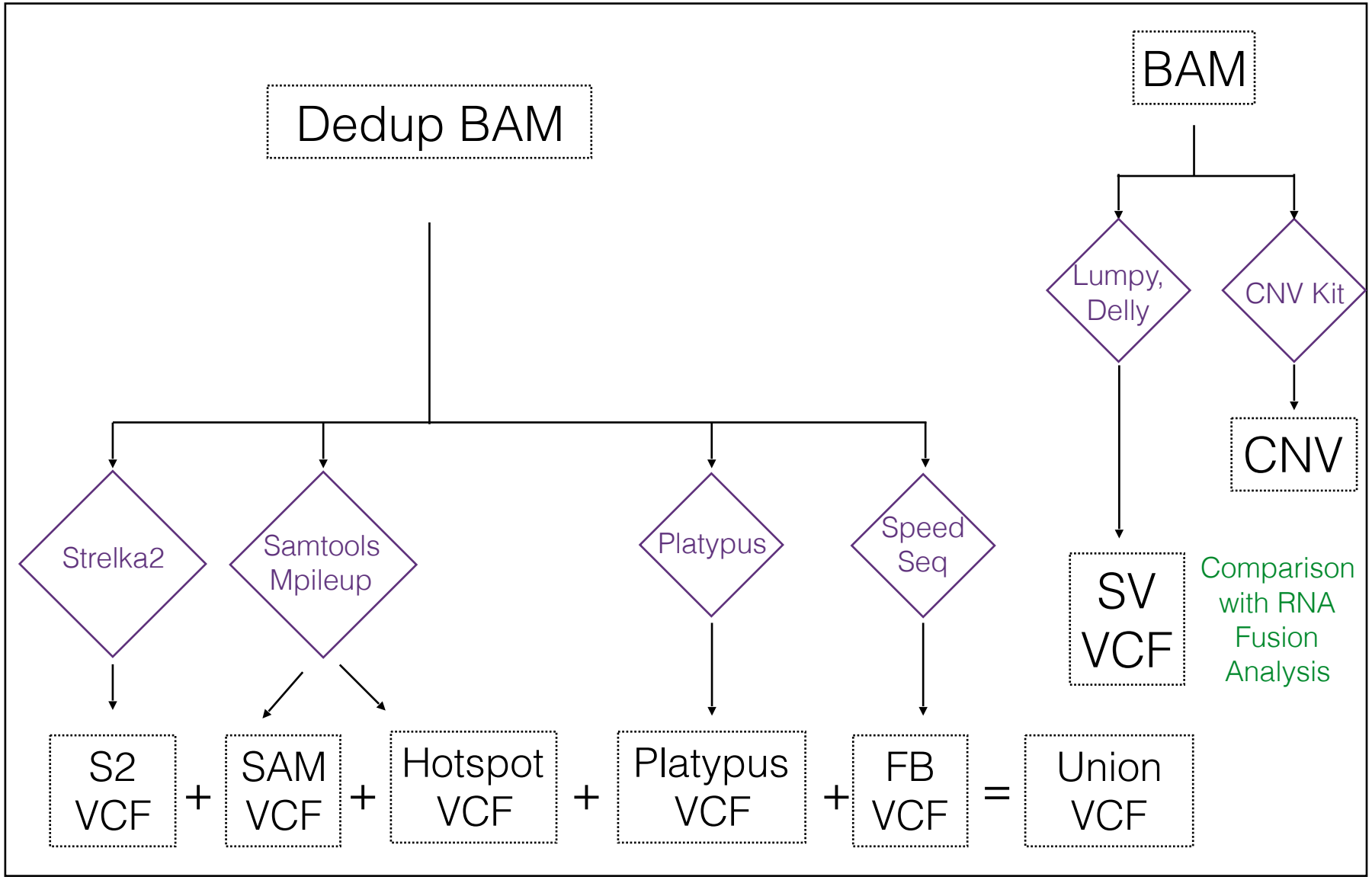


# Why does GATK need Base Recalibration?

- Base recalibration detects systematic errors made by the sequencer when it estimates the quality score of each base call



# Germline Workflow

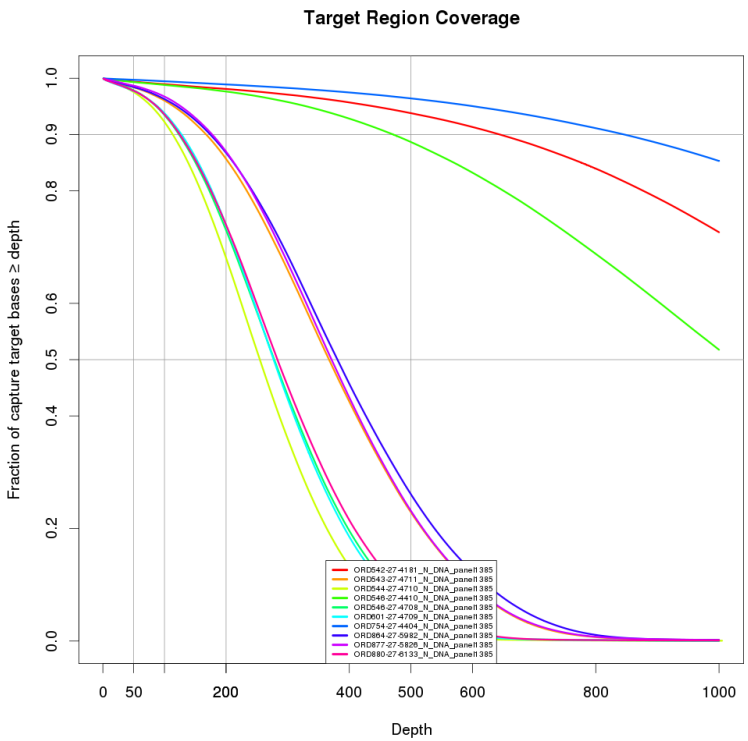


# Differences in Results between Callers?

| Sample | Caller | SNV TP | SNV FP | SNV FN | Indel TP | Indel FP | Indel FN | SNV SN | Indel SN | SN | SP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FFPE control, 40 ng | gatk | 1238 | 36 | 21 | 34 | 1 | 3 | 98.3 | 91.9 | 98.1% | 97.2 |
| FFPE control, 40 ng | strelka2 | 1238 | 2 | 21 | 34 | 1 | 3 | 98.3 | 91.9 | 98.1% | 99.8 |
| FFPE control, 40 ng | sam | 1238 | 2 | 21 | 33 | 5 | 4 | 98.3 | 89.2 | 98.1% | 99.5 |
| FFPE control, 40 ng | ssvar | 1224 | 2 | 35 | 34 | 0 | 3 | 97.2 | 91.9 | 97.1% | 99.8 |
| FFPE control, 40 ng | platypus | 1215 | 7 | 44 | 34 | 0 | 3 | 96.5 | 91.9 | 96.4% | 99.4 |
| FRESH sample, 200 ng | gatk | 1252 | 36 | 6 | 37 | 4 | 1 | 99.5 | 97.4 | 99.5% | 97 |
| FRESH sample, 200 ng | strelka2 | 1237 | 0 | 20 | 34 | 6 | 3 | 98.4 | 91.9 | 98.2% | 99.5 |
| FRESH sample, 200 ng | sam | 1237 | 0 | 21 | 17 | 0 | 21 | 98.3 | 44.7 | 96.8% | 100 |
| FRESH sample, 200 ng | ssvar | 1236 | 1 | 22 | 36 | 0 | 2 | 98.3 | 94.7 | 98.1% | 99.9 |
| FRESH sample, 200 ng | platypus | 1215 | 5 | 43 | 34 | 1 | 4 | 96.6 | 89.5 | 96.4% | 99.5 |

# What is sequence coverage and depth?

- Base depth is the number of reads that cover a particular base
- Coverage is “how much” of your target did you cover
- Depth of Coverage is how deep was that coverage?

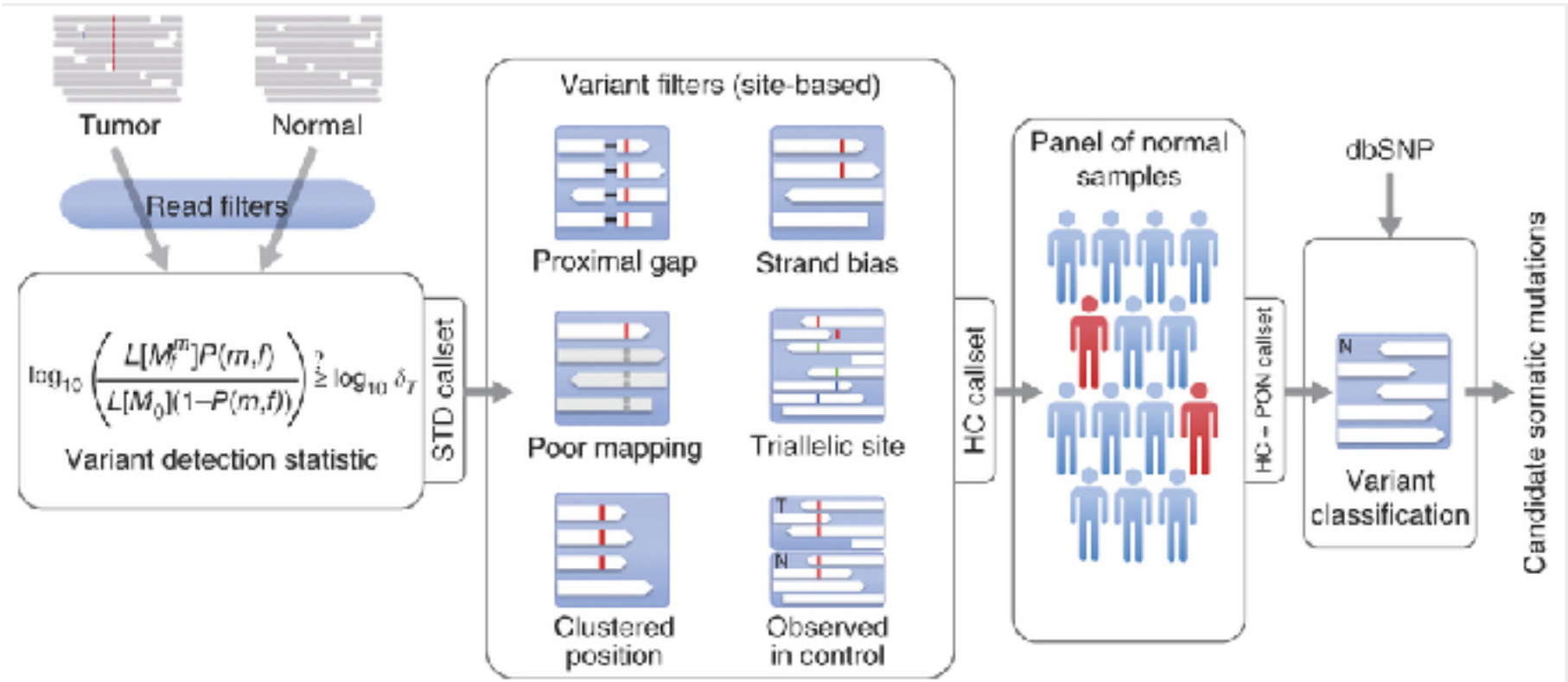


| Effect | Impact |
|---|---|
| 3_prime_UTR_truncation +exon_loss | M |
| 3_prime_UTR_variant | NC |
| 5_prime_UTR_premature start_codon_gain_variant | L |
| 5_prime_UTR_truncation + exon_loss_variant | M |
| 5_prime_UTR_variant | NC |
| bidirectional_gene_fusion | H |
| chromosome | H |
| coding_sequence_variant | NC |
| coding_sequence_variant | LOW |
| conserved_intergenic_variant | NC |
| conserved_intron_variant | NC |
| disruptive_inframe_deletion | M |
| disruptive_inframe_insertion | M |
| downstream_gene_variant | NC |
| duplication | H |
| duplication | H |
| duplication | H |
| duplication | M |
| exon_loss_variant | H |
| exon_loss_variant | H |
| exon_variant | NC |
| feature_ablation | H |
| feature_ablation | H |
| frameshift_variant | H |
| gene_fusion | H |
| gene_fusion | H |
| gene_variant | NC |
| inframe_deletion | M |
| inframe_insertion | M |
| initiator_codon_variant | L |
| intergenic_region | NC |
| intragenic_variant | NC |
| intron_variant | NC |
| inversion | H |
| inversion | H |
| inversion | H |
| miRNA | NC |
| missense_variant | M |
| protein_protein_contact | H |
| rare_amino_acid_variant | H |
| rearranged_at_DNA_level | H |
| regulatory_region_variant | NC |
| sequence_feature + exon_loss_variant | NC |
| splice_acceptor_variant | H |
| splice_donor_variant | H |
| splice_region_variant | L |
| splice_region_variant | L |
| splice_region_variant | M |
| start_lost | H |
| start_retained | L |
| stop_gained | H |
| stop_lost | H |
| stop_retained_variant | L |
| stop_retained_variant | L |
| structural_interaction_variant | H |
| synonymous_variant | L |
| transcript_variant | NC |
| upstream_gene_variant | NC |

# Recommended Filtering for Germline Testing

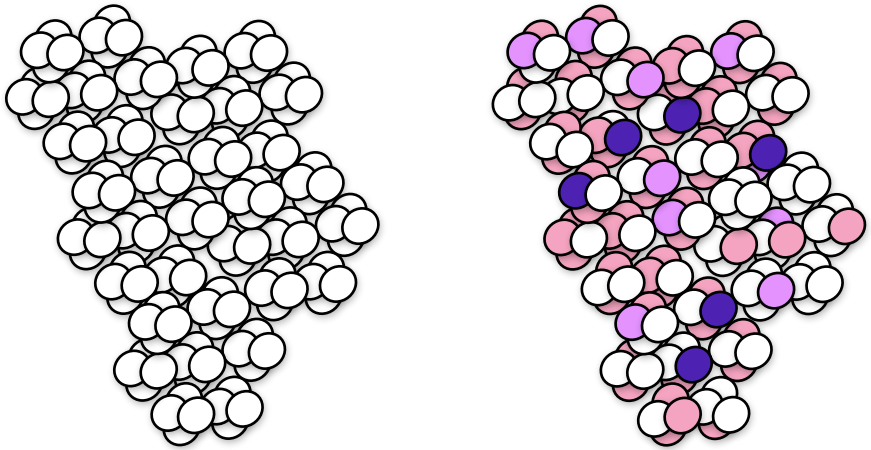
- Depth >10
- LOF or Misssense (Coding Changes)
- Alt Read Ct > 3
- Mutation Allele Frequency (MAF) > 0.15
- If novel:
  - Called by 2+ callers

# Somatic Mutation Identification




Genomes from normal and tumor samples from the same patient are compared.

# Tumors are Heterogeneous



Normal Tumor

Somatic Mutation Calling Compares the Tumor and Normal samples to identify low frequency mutations.

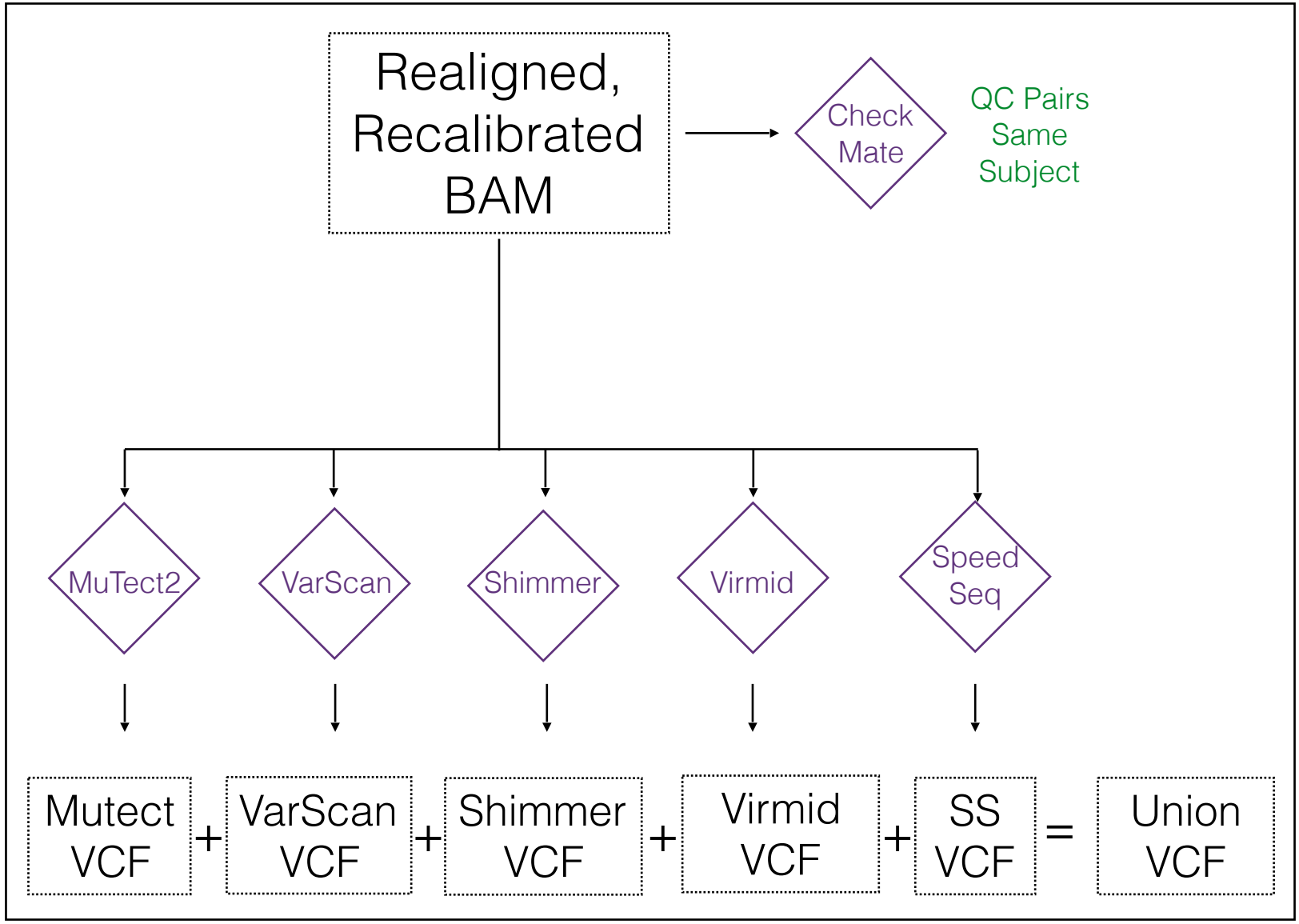
# Why do genome sequencing in Cancer?

- Identification of new variants (SNPs, Indels, SVs) associated with cancer to drive basic research and identify new drug targets
  - Nature, 474,609–615 (30 June 2011)
  - Nature, 487, 330–337 (19 July 2012)
- Identification of “known” variants to aid in patient treatment
  - Clin Cancer Res. 2012 Aug 15;18(16):4257-65.
  - Advances in pharmacology (San Diego, Calif.) 01/2012; 65:399-435.

# Limitations of Software Designed to Identify Somatic Mutations

- Variant detection software needs specialized training in order to:
  - Install -- must know unix and be able to install dependent software
  - Use -- must know unix and work through any “bugs” when the data isn’t exactly as software is expecting
- The variant “scores” are hard to interpret, so weeding out errors (FP) is hard.
- SNP calling methods often don’t agree very much given the same data

# Somatic Workflows



# Recommended Filtering for Somatic Mutations

- Depth < 20
- LOF or Misssense
- MAF (Normal) * 10.< MAF (Tumor)
- In COSMIC > 5 Subject
  - Tumor: Alt Read Ct < 3
  - Tumor: MAF < 0.01
- Others
  - Tumor: Alt Read CT < 8
  - Tumor: MAF < 0.05
  - Tumor: Called by 2+ callers

# Effect of Variation in Genes

- snpEff
  - Changes affecting genes
  - Changes affecting regulatory regions
  - ENCODE
  - Epigenome Roadmap
  - NextProt
    - proteomic annotations
  - Motifs
- VEP
  - Changes affecting genes
  - Changes affecting regulatory regions
  - Integrated with downstream tools like cBioporal and GenVisR

# Variant Functional Classification

- Pathogenic - a sequence variant that is previously reported and is a recognized cause of the disorder.
- Likely Pathogenic – a sequence variant that is previously unreported and is of the type which is expected to cause the disorder.
- VUS (Variant of Unknown Significance) – a sequence variant that is previously unreported and is of the type which may or may not be causative of the disorder.
- Likely Benign – a sequence variant that is previously unreported and is probably not causative of disease.
- Benign – a sequence variant is previously reported and is a recognized neutral variant.
- A sequence variant that is previously not known or expected to be causative of disease, but is found to exist in people with a particular disease or disorder.

# Disease Studies

- ClinVar
  - ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence
- GWAS Catalog
  - The Catalog is a quality controlled, manually curated, literature-derived collection of all published genome-wide association studies assaying at least 100,000 SNPs and all SNP-trait associations with p-values < 1.0 x 10-5
- Decipher
  - The DECIPHER database contains data from 20305 patients who have given consent for broad data-sharing; DECIPHER also supports more limited sharing via consortia.

# Cancer Datasets and Annotation

- Clinical Interpretation of Variants in Cancer (CIVIC)
- Catalog of Somatic Mutation in Cancer (COSMIC)
  - Gene Fusions
  - Gene Census
  - Curated Genes
  - Drug Resistance (so far 9 genes)
  - Genome Wide Screens
- The Cancer Genome Atlas (TCGA)
  - Tons of Data, RNASeq, CNV, WES, WGS, etc

# Annotating Genomic Variation

- Gene Annotation (Genes, Regulation and TFBS)
- dbSNP, ExAC, gnomAD
- clinvar, gwas catalog
- cosmic
- dbNSFP
  - SIFT, Polyphen2, LRT, MutationTaster, MutationAssessor, FATHMM, VEST3, CADD, MetaLR, MetaSVM, PROVEAN, DANN, fathmm-MKL, fitCons
  - PhyloP x 2, phastCons x 2, GERP++ and SiPhy
  - Allele frequencies in 1000 Genomes Project phase 3 data, UK10K cohorts data, ExAC consortium data and the NHLBI Exome Sequencing Project ESP6500 data
- genesets (MSigDB)
- CIVIC
- BROAD Target

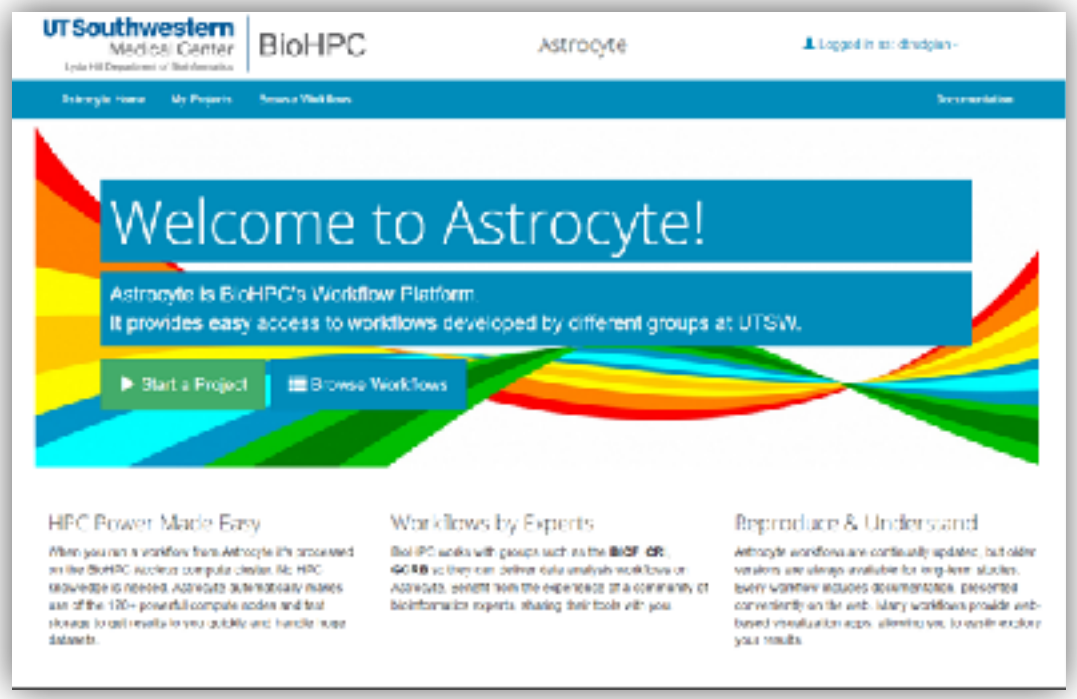
- What is Variation
  - Somatic vs Germline
  - SNVs, Indels and Structural Variation

- Is there an easy way to run all those command line programs?
  - BioHPC Astrocyte

# Point and Click Analysis Tools from the BioHPC and BICF

# Astrocyte – BioHPC Workflow Platform

Allows groups to give easy-access to their analysis pipelines via the web




Standardized Workflows

Simple Web Forms

Online documentation & results visualization*

Workflows run on HPC cluster without developer or user needing cluster knowledge

# Bioinformatics Core Facility (BICF)

BICF provides bioinformatics, statistics and data management support for researchers on campus.

BICF functions as the conduit between bioinformatics research programs and the clinical- and basic-science research community at UTSW.

Please email bicf@utsouthwestern.edu with questions or comments about these workflows.

| | | |
|---|---|---|
| **BICF ChIP-seq Analysis Workflow**<br>This is a workflow package for the BioHPC/BICF ChIP-seq workflow system. It implements a simple ChIP-seq analysis workflow using deepTools, Diffbind, ChipSeeker and MEME-ChIP, visualization application. | **Current Version:** chipseq_analysis_bicf - 0.0.12<br>**Author:** Beibei Chen<br>**Contact:** biohpc-help@utsouthwestern.edu | Run Workflow<br>Documentation<br>View Versions |
| **BICF RNASeq Analysis Workflow**<br>This is a workflow package for the BioHPC/BICF RNASeq workflow system. It implements differential expression analysis, gene set enrichment analysis, gene fusion analysis and variant identification using RNASeq data. | **Current Version:** maseq_bicf - 0.3.3<br>**Author:** Brandi Cantarel<br>**Contact:** biohpc-help@utsouthwestern.edu | Run Workflow<br>Documentation<br>View Versions |
| **BICF RNASeq Variant Analysis Workflow**<br>THIS WORKFLOW IS OBSOLETE! The Main BICF workflow includes variant analysis and differential expression analysis as one easy to use workflow. | **Current Version:** maseq_variant_bicf - 0.0.11<br>**Author:** Brandi Cantarel<br>**Contact:** biohpc-help@utsouthwestern.edu | Run Workflow<br>Documentation<br>View Versions |
| **BICF Somatic Mutation Calling**<br>This is a workflow package for the BioHPC/BICF Somatic Mutation workflow system. It implements a simple Somatic Mutation analysis workflow. | **Current Version:** somatic_bicf - 0.0.3<br>**Author:** Brandi Cantarel<br>**Contact:** biohpc-help@utsouthwestern.edu | Run Workflow<br>Documentation<br>View Versions |
| **BICF Germline Variant Analysis Workflow**<br>This is a workflow package for the BioHPC/BICF Germline Variant workflow system. It implements a simple germline variant analysis workflow using TrimGalore, BWA, Speedseq, GATK, Samtools and Platypus. SNPs and Indels are integrated using BAYSIC; then annotated using SNPEFF and SnpSift. | **Current Version:** germline_bicf - 0.0.10<br>**Author:** Brandi Cantarel<br>**Contact:** biohpc-help@utsouthwestern.edu | Run Workflow<br>Documentation<br>View Versions |

https://astrocyte.biohpc.swmed.edu/brand/bicf/browse/

UTSouthwestern Medical Center | BICF

# Create a new project

## My Projects

In Astrocyte **projects** are used to organize your work. You upload **input data** into a project, and can then run **workflows** against this input data. Try to separate your work into natural projects, so that you can easily share them with other users if required.

+ Start a New Project

Project Name | Create New Project

Existing Projects

| ID | Name | Created | Workflows Run | Input Files | Size | Actions |
|---|---|---|---|---|---|---|
| FRJ21 | RNAseq_test | Aug 23, 2016, 3:33 p.m. | 0 | 0 | 0 bytes | |

Projects Shared with Me

| ID | Name | Created | Workflows Run | Input Files | Size | Actions |
|---|---|---|---|---|---|---|
| FRJH0 | test | June 1, 2016, 5:03 p.m. by Brandi Cantarel | 4 | 10 | 216.5 GB | |

# Add Data To Your Project

Input data in this project

To run a workflow against input data you need to upload it into this project. Click the button below to add new files from your web browser or the BioHPC cluster. You can also download or delete existing files from the project in the list below.

Add Data To This Project

No input data has been added to this project. Please upload files to use them with a workflow.

Workflows run in this project

Astrocyte provides many workflow created by different groups at UTSW for you to run against your data. To begin, make sure you have added input data into your project and then click the 'Run a workflow' button to choose a workflow to run.

Run a workflow in this project

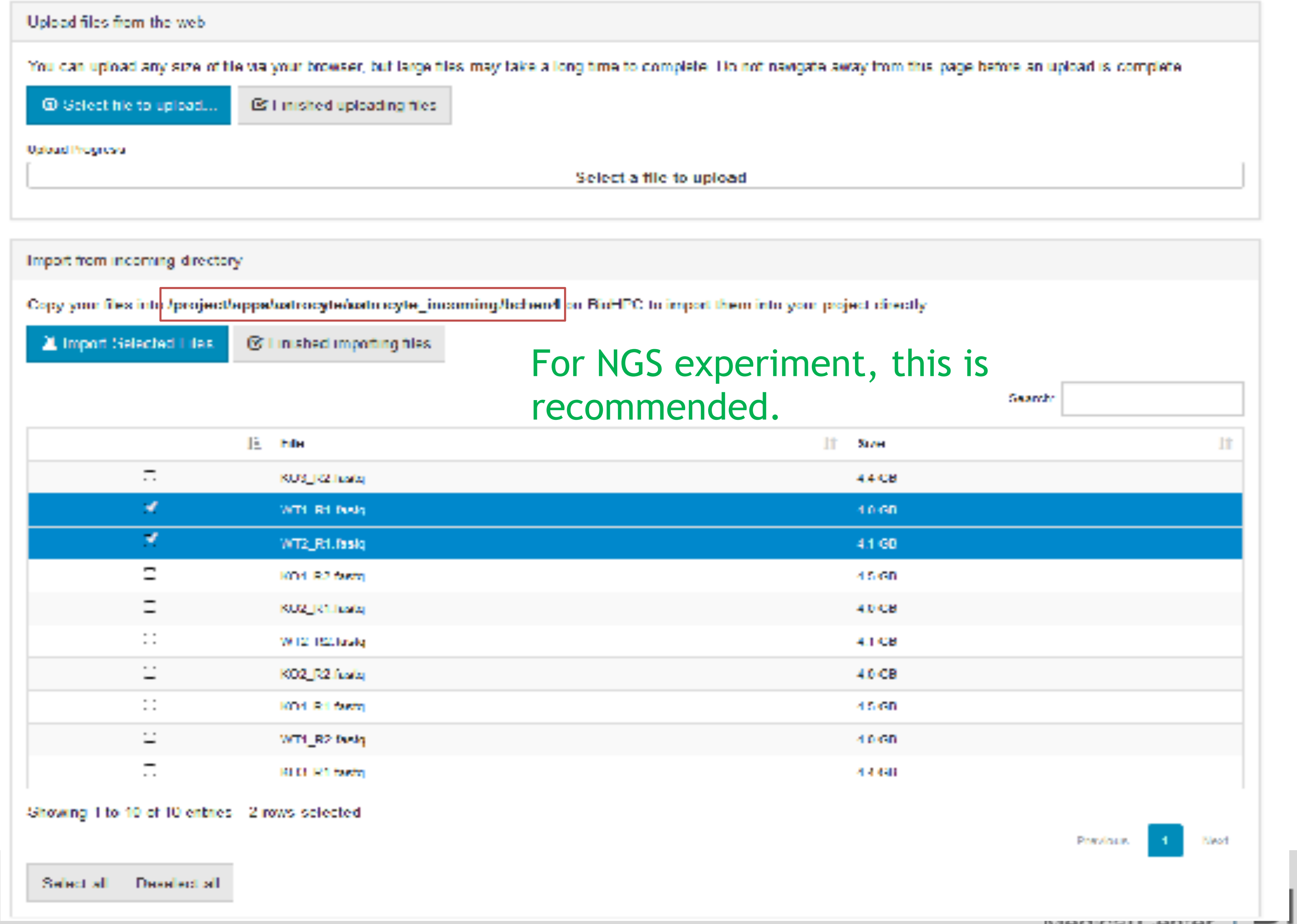
You haven't run any workflows in this project. Upload some input data, and then click the 'Run Workflow' button above to begin.

Sharing

Share With User

Shared With

# Add Data To Your Project



**Upload files from the web**

You can upload any size of file via your browser, but large files may take a long time to complete. Do not navigate away from this page before an upload is complete.

**Import from incoming directory**

Copy your files into /project/apps/astrocyte/astrocyte_incoming/[illegible] on BioHPC to import them into your project directly

For NGS experiment, this is recommended.

|  | File | Size |
|---|---|---|
| ☐ | KO3_R2.fastq | 4.4 GB |
| ☑ | WT1_R1.fastq | 4.0 GB |
| ☑ | WT2_R1.fastq | 4.1 GB |
| ☐ | KO1_R2.fastq | 4.5 GB |
| ☐ | KO2_R1.fastq | 4.0 GB |
| ☐ | WT2_R2.fastq | 4.1 GB |
| ☐ | KO2_R2.fastq | 4.0 GB |
| ☐ | KO1_R1.fastq | 4.5 GB |
| ☐ | WT1_R2.fastq | 4.0 GB |
| ☐ | KO3_R1.fastq | 4.4 GB |

Showing 1 to 10 of 10 entries 2 rows selected

# Make your design file

```
FamilyID
This ID will be used to call samples in batch
SampleID
This ID will be used to name all workflow produced files ie S0001 will produce S0001.bam
FullPathToFqR1
Name of the fastq file R1 (not the full path)
FullPathToFqR2
Name of the fastq file R2 (not the full path)
```

| FamilyID | SampleID | FqR1 | FqR2 |
|---|---|---|---|
| F1 | GM12877 | GM12877.R1_001.fastq.gz | GM12877_S124_R2_001.fastq.gz |
| F1 | GM12878 | GM12878.R1_001.fastq.gz | GM12878_S124_R2_001.fastq.gz |
| F1 | GM12879 | GM12879.R1_001.fastq.gz | GM12879_S124_R2_001.fastq.gz |
| F2 | GM12887 | GM12887.R1_001.fastq.gz | GM12887.R2_001.fastq.gz |
| F2 | GM12888 | GM12888.R1_001.fastq.gz | GM12888.R2_001.fastq.gz |
| F2 | GM12889 | GM12889.R1_001.fastq.gz | GM12889.R2_001.fastq.gz |

# Make your design file

- Use tab as delimiter
  - Excel save as “Text (tab delimited)”
- If no SubjectID, use same number/character for all rows
- SampleID and SampleName
- If no FqR2, leave them empty
- For all contents, no “-”
- For all contents, no spaces
- Columns names MUST be exactly the same as documented

# Select your data files and set up workflow and submit

Parameters

Project

Project 47: panel_utswv2

Name for this run

temp

One or more input paired-end FASTQ files from a RNASeq experiment and a design file with the link between the same name and the sample group regex: ".*(fastq|fq)" min: 1

panel_utswv2.design.txt
utswv2_H2_AP14-924.R2.fastq.gz
utswv2_H2_AP14-924.R1.fastq.gz
utswv2_H2_33.R2.fastq.gz

SELECT YOUR FILES

In single-end sequencing, the sequencer reads a fragment from only one end to the other, generating the sequence of base pairs. In paired-end reading it starts at one read, finishes this direction at the specified read length, and then starts another round of reading from the opposite end of the fragment.

Paired End

A design file listing sample names, fastq files, and additional information about the sample

panel_utswv2.design.txt

A capture bed file is a bed file of the targeting panel or exome capture used for the sequencing. This file is used to assess capture efficiency and to limit variants to capture region

UTSWV2.bed

Reference genome for alignment

Human GRCh38

Run Workflow

# Project is running

## Run 'temp' in Project 'panel_utswv2'

### Run Information

| | |
|---|---|
| **Running Workflow** | BICF Germline Variant Analysis Workflow<br>brandi.cantarel/variant_germline.git / 0.0.10 |
| **Status** | RUNNING |
| **Created** | Sept. 13, 2017, 8:39 p.m. by s168468 |
| **Size** | 116.0 KB |

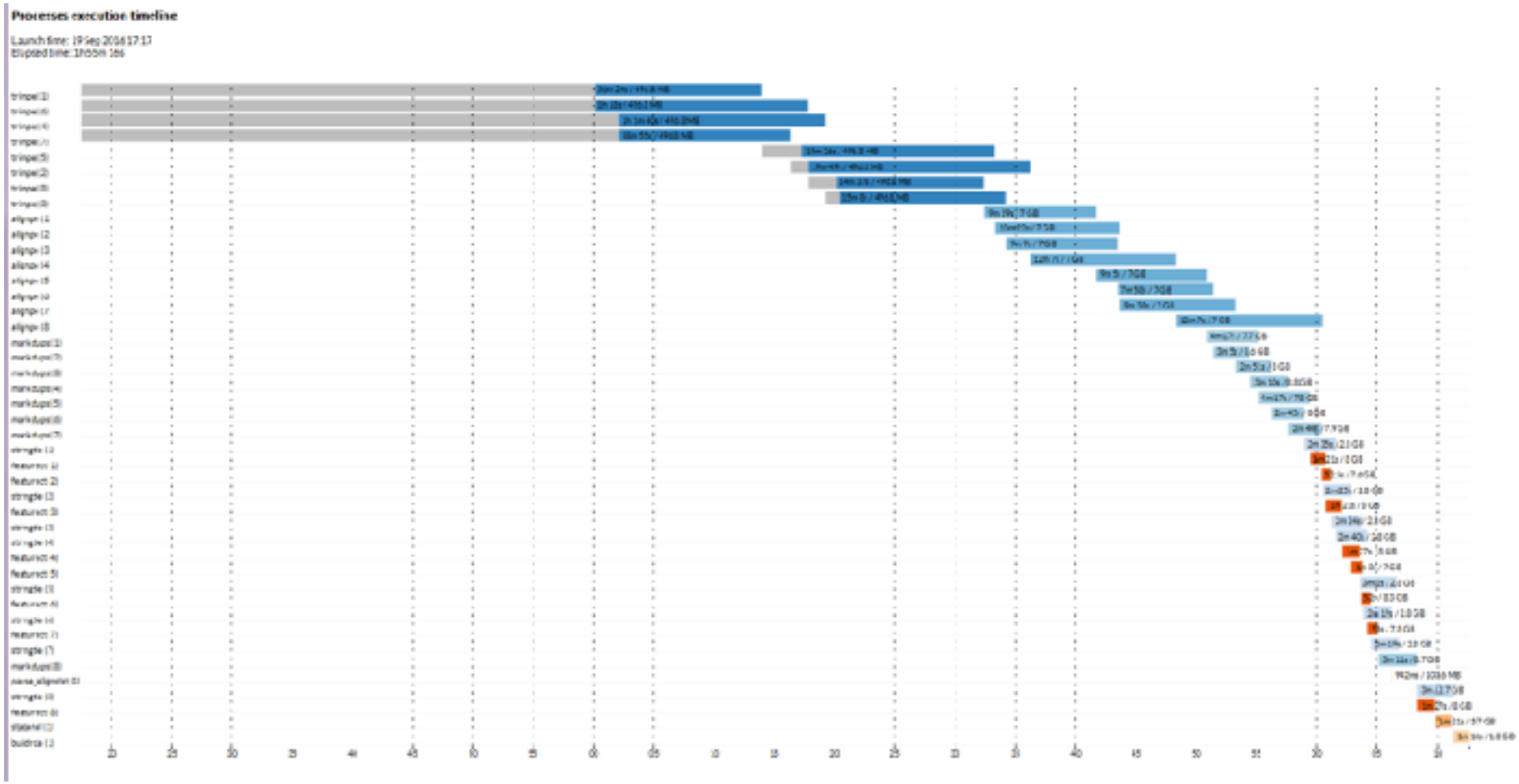
### Parameters

| Parameter | Value |
|---|---|
| design | panel_utswv2.design.txt |
| genome | /project/shared/bicf_workflow_ref/GRCh38 |
| pairs | pe |
| fastq2 | utswv2_H2_AP14-924.R2.fastq.gz |
| fastq1 | utswv2_H2_AP14-924.R1.fastq.gz |
| capture | UTSWV2.bed |

### Input Files

| Filename | Size |
|---|---|
| panel_utswv2.design.txt | 1.3 KB |
| utswv2_H2_AP14-924.R2.fastq.gz | 1.5 GB |
| utswv2_H2_AP14-924.R1.fastq.gz | 1.5 GB |
| UTSWV2.bed | 486.3 KB |

# Timeline of the whole run



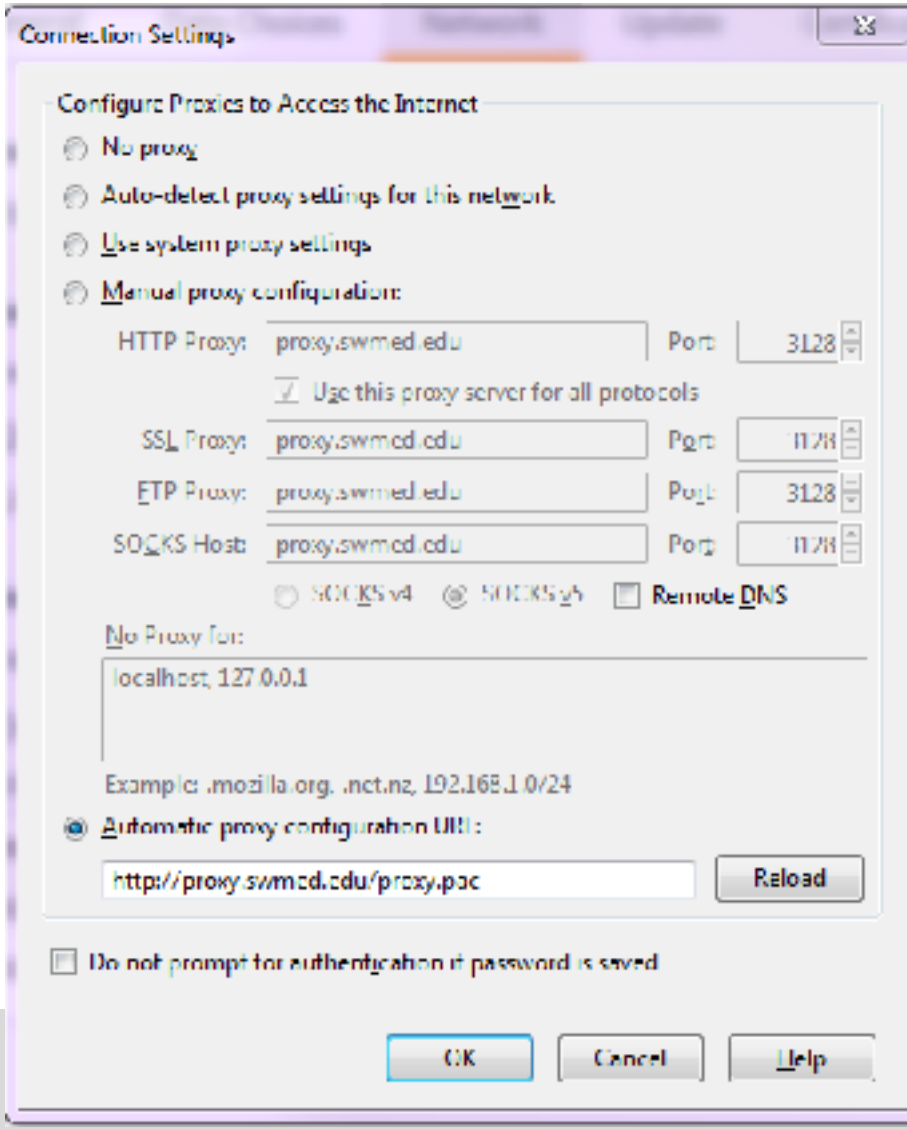
# Common errors and solutions

```
        Error running workflow. Diagnostic output

N E X T F L O W  ~  version 0.20.1
Launching main.nf
Didn't match any input files with entries in the design file

 -- Check script 'main.nf' at line: 49 or see '.nextflow.log' file for more details
```

- Make sure the delimiter is tab
- Make sure the column name are the same as mentioned in documentation
- Make sure the file names match

# Common errors and solutions

- Not all files are uploaded
- It's about the proxy setting
- Use auto-detect proxy



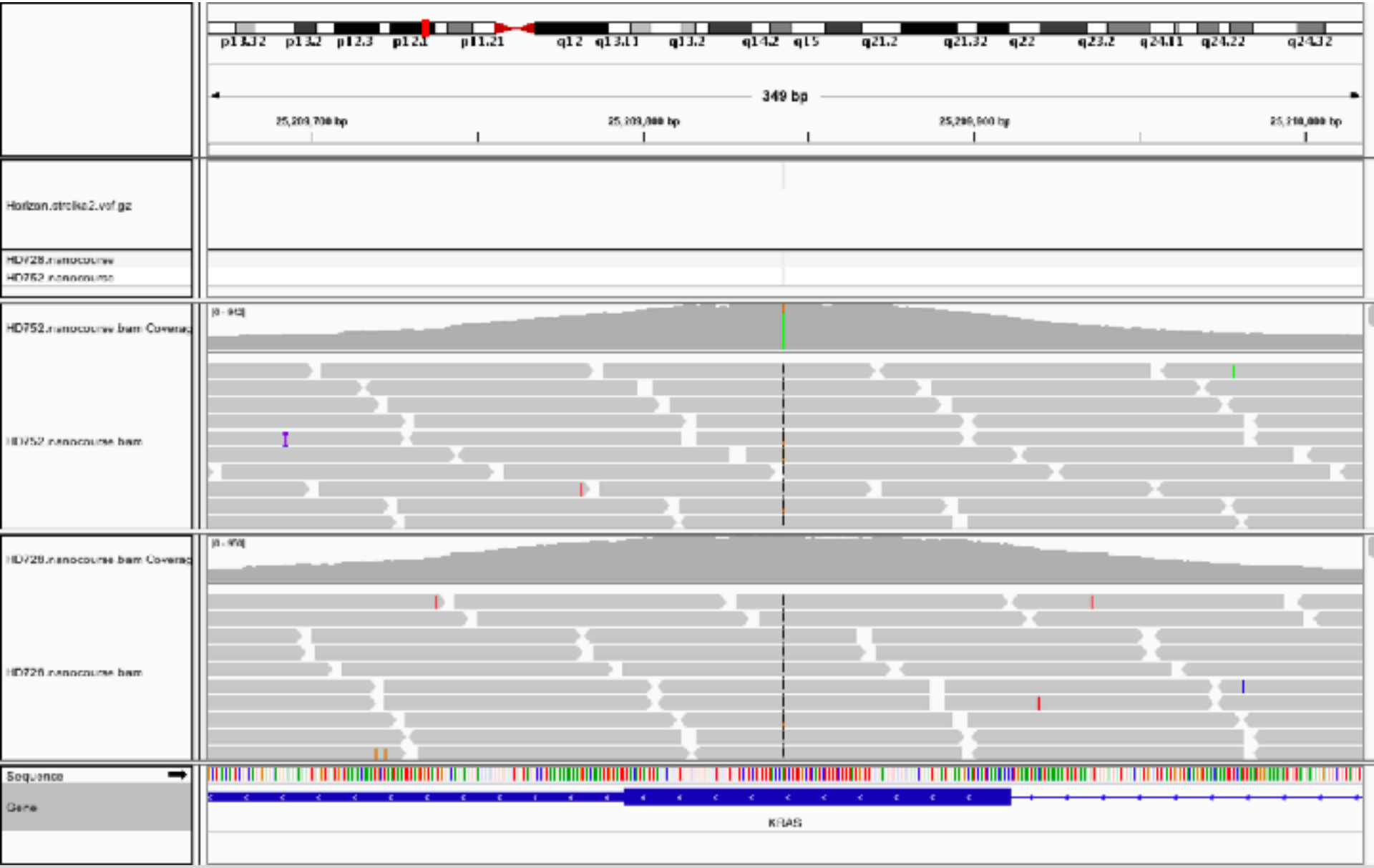
# Key Files Germline Pipeline

- VCF file — SNPs/Indels for each sample
  - SampleID.annot.vcf.gz
- Coverage Histogram for each sample
  - SampleID.coverage_histogram.png
- Cumulative Distribution Plot for all samples
  - coverage_cdf.png
- QC for all samples
  - sequence.stats.txt
- Structural Variants (unfiltered)
  - SampleID.sssv.sv.vcf.gz.annot.txt

# Key Files Somatic Mutation Pipeline

- VCF file — SNPs/Indels for each sample
  - TumorID_NormalID.annot.vcf.gz
- Match Check File
  - TumorID_NormalID_matched.txt

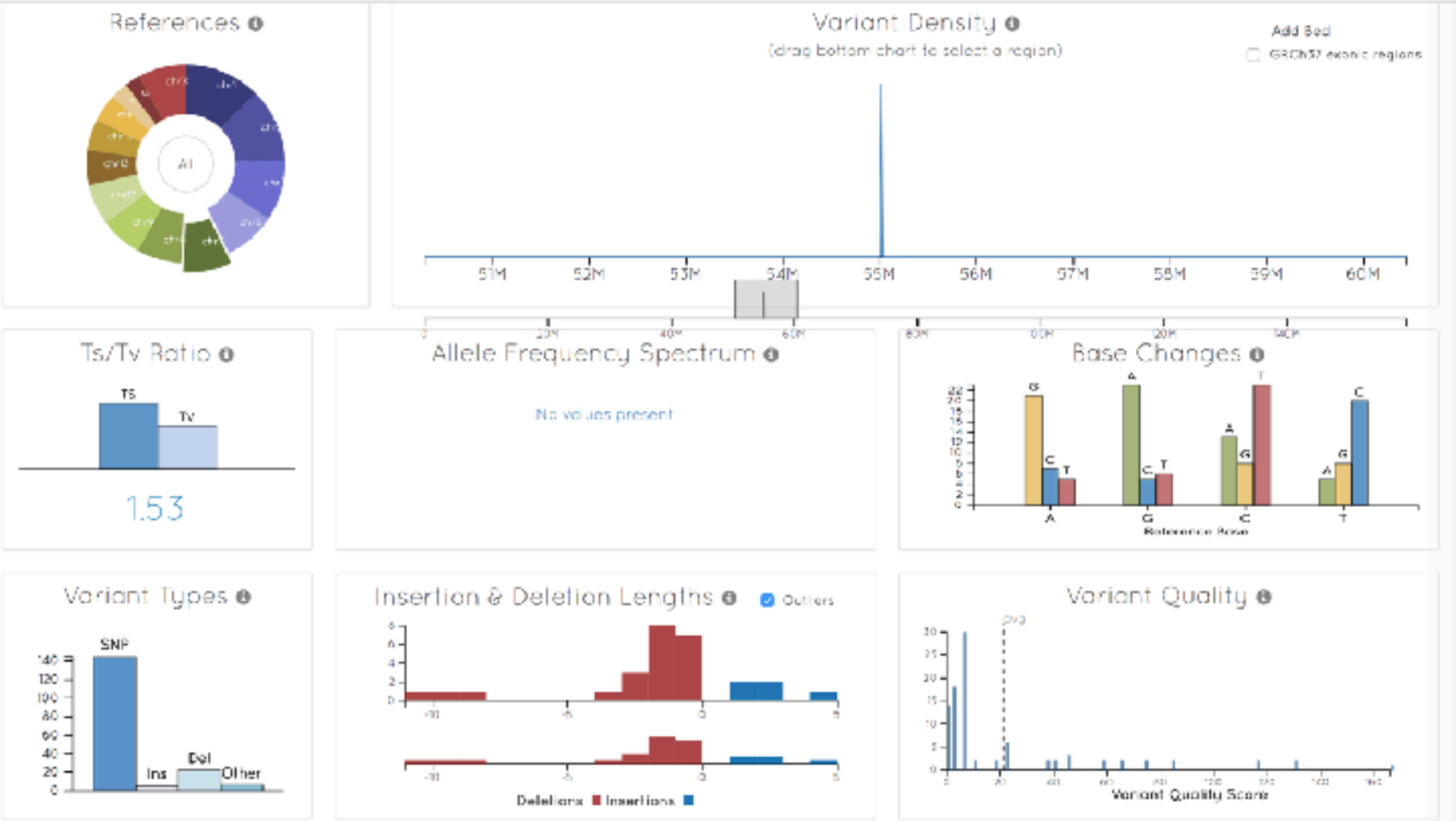
# IGV Viewer


p13.2
p12.3
p12.1
p11.21
q12
q13.11
q13.2
q14.2
q15
q21.2
q21.32
q22
q23.2
q24.11
q24.22
349 bp
Horizon.strelka2.vcf.gz
Sequence
Gene
KRAS

# BAM ioBIO


bam.iobio.io
Try our beta release
Read Coverage
Reads Sampled
104
thousands

# VCF ioBIO


References
Variant Density
(drag bottom chart to select a region)
Add Bed
51M
52M
53M
54M
55M
56M
57M
58M
59M
60M
Ts/Tv Ratio
TS
TV
1.53
Allele Frequency Spectrum
No values present
Base Changes
Reference Base
Variant Types
SNP
Ins
Del
Other
Insertion & Deletion Lengths
Outliers
Deletions
Insertions
Variant Quality
Variant Quality Score

# Questions?