MUTATION IDENTIFICATION IN GENOMICS

Karthigayini Sivaprakasam

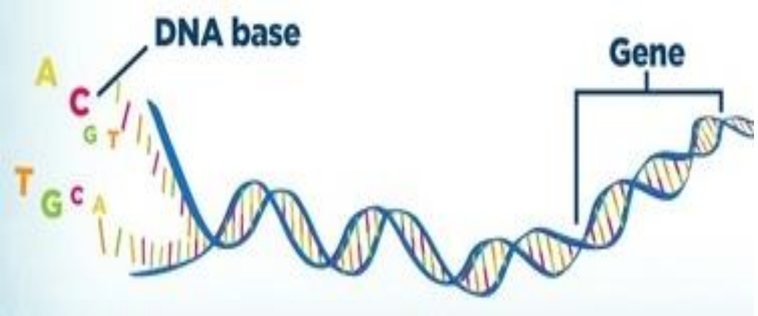
May 2019

SECTIONS

- Introduction to genetics
- Sequencing
- Workflows
- Variant callers
- Algorithms
- Annotation
- Astrocyte

GENOME AND GENETIC DISEASES

What Is A Genome?



↓

- DNA tells cells/tissues/organs/systems how to operate
- 3.2 B letters - Human
- 10 million letters vary between individuals

What is a genetic disease?

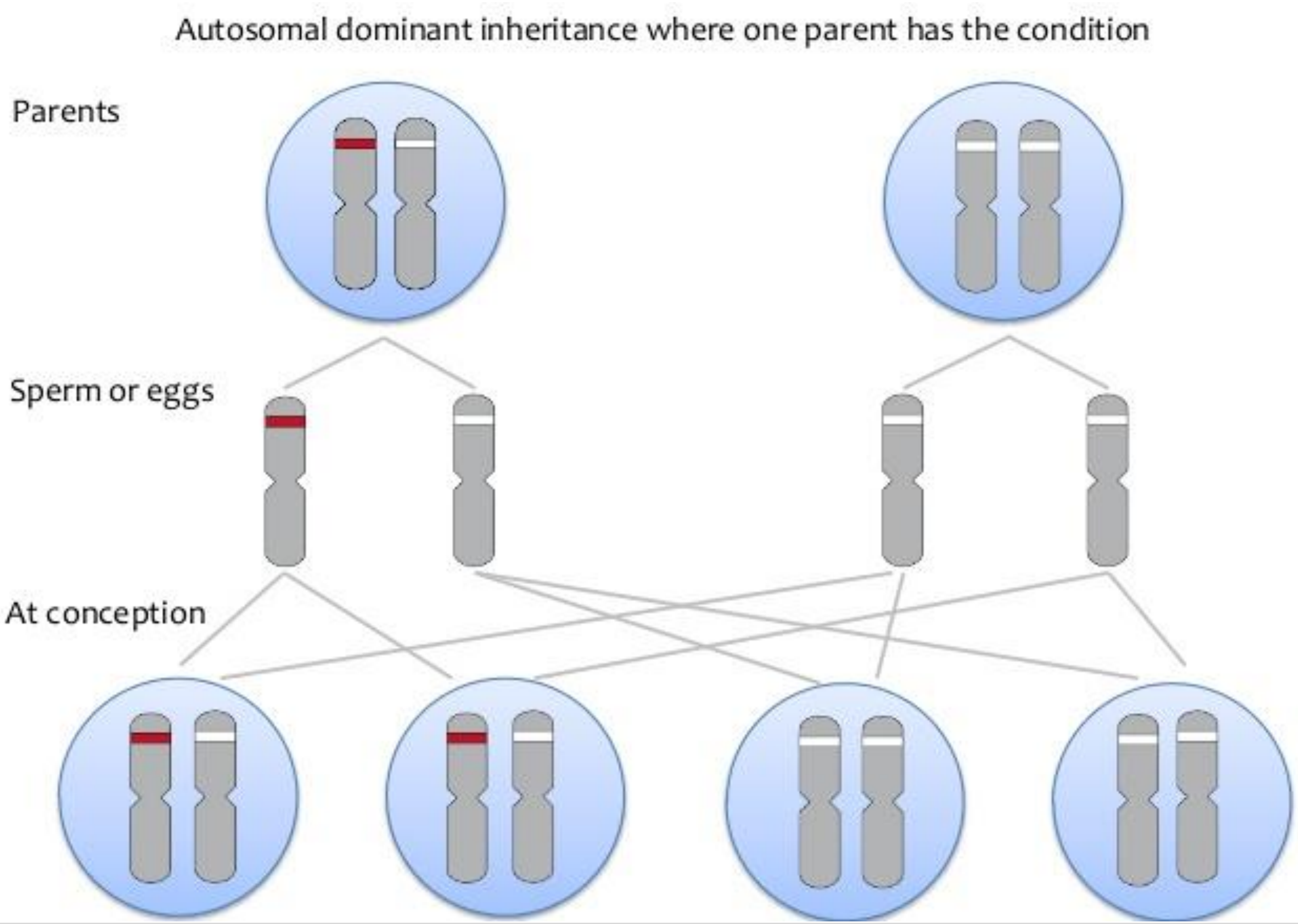
- Abnormality caused in the genome

- Can be as small as a single base or involve addition or deletion of whole copies of chromosome.

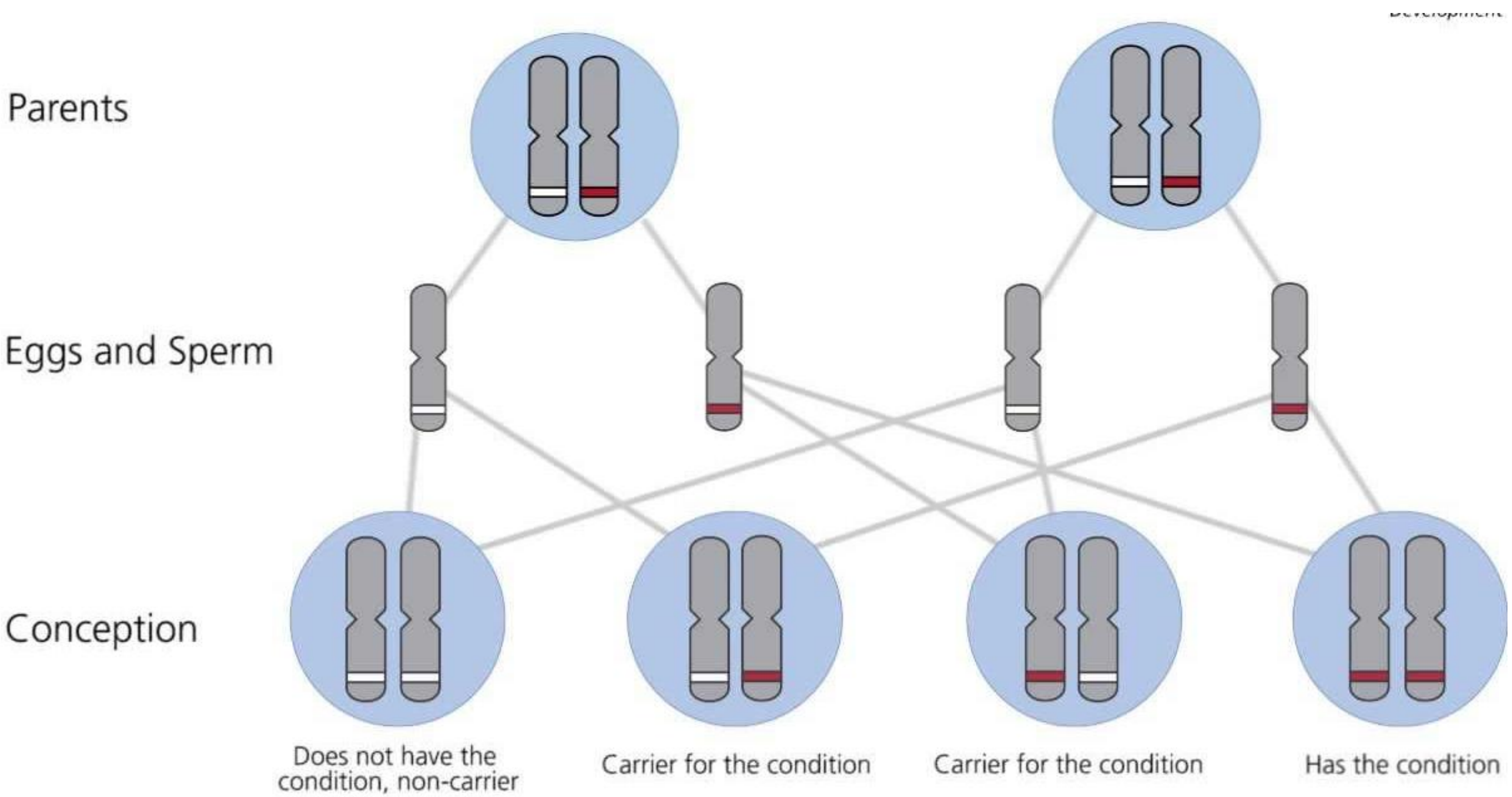
Types of Genetic Disorders

- Single gene mutations – Sickle cell anemia
- Chromosomal disorders - Downs syndrome
- Complex disorders – Cancers
- Inherited or Acquired.

Autosomal Dominant Disorders


Autosomal dominant inheritance where one parent has the condition
Parents
Sperm or eggs
At conception

Recessive Disorders



Development

Parents

Eggs and Sperm

Conception

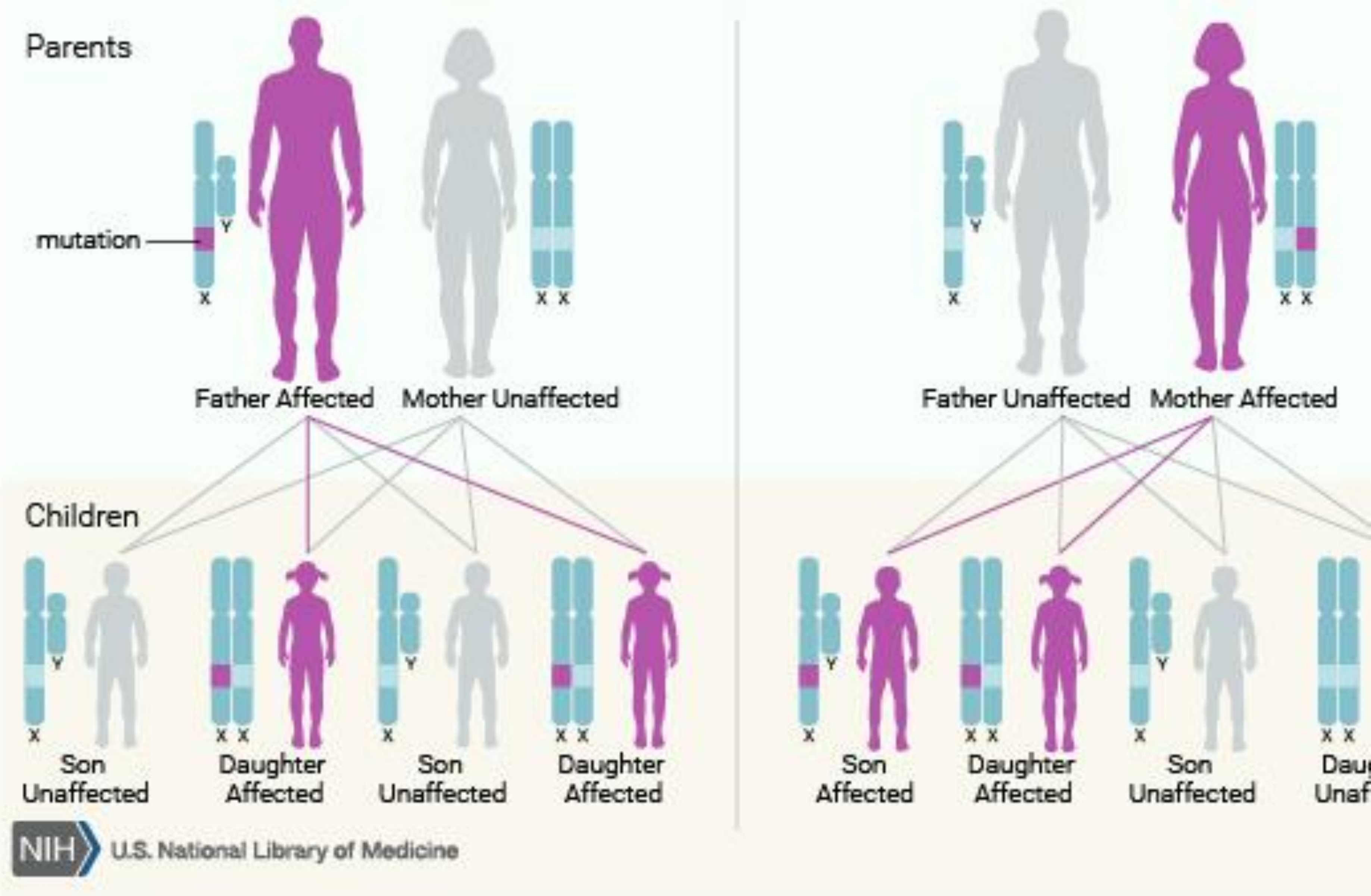
Does not have the condition, non-carrier

Carrier for the condition

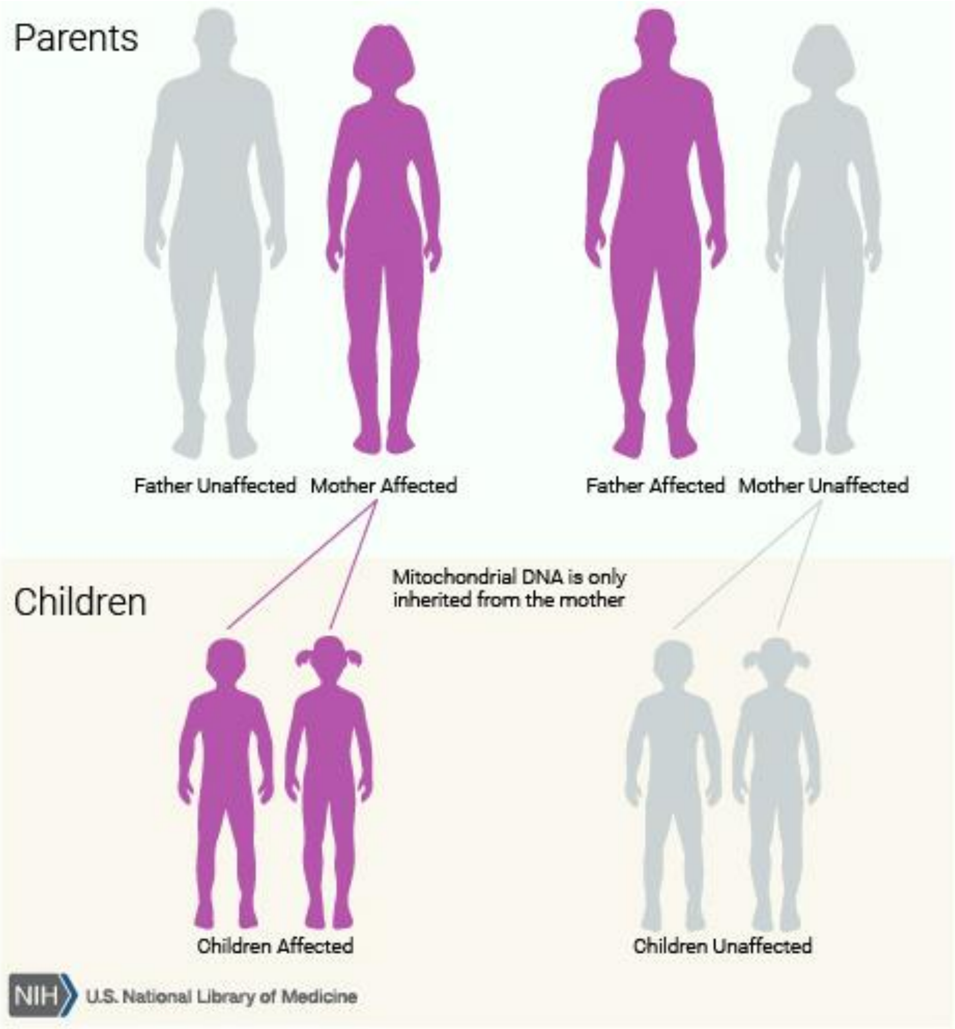
Carrier for the condition

Has the condition

X-linked Disorders



Mitochondrial Disorders



SEQUENCING

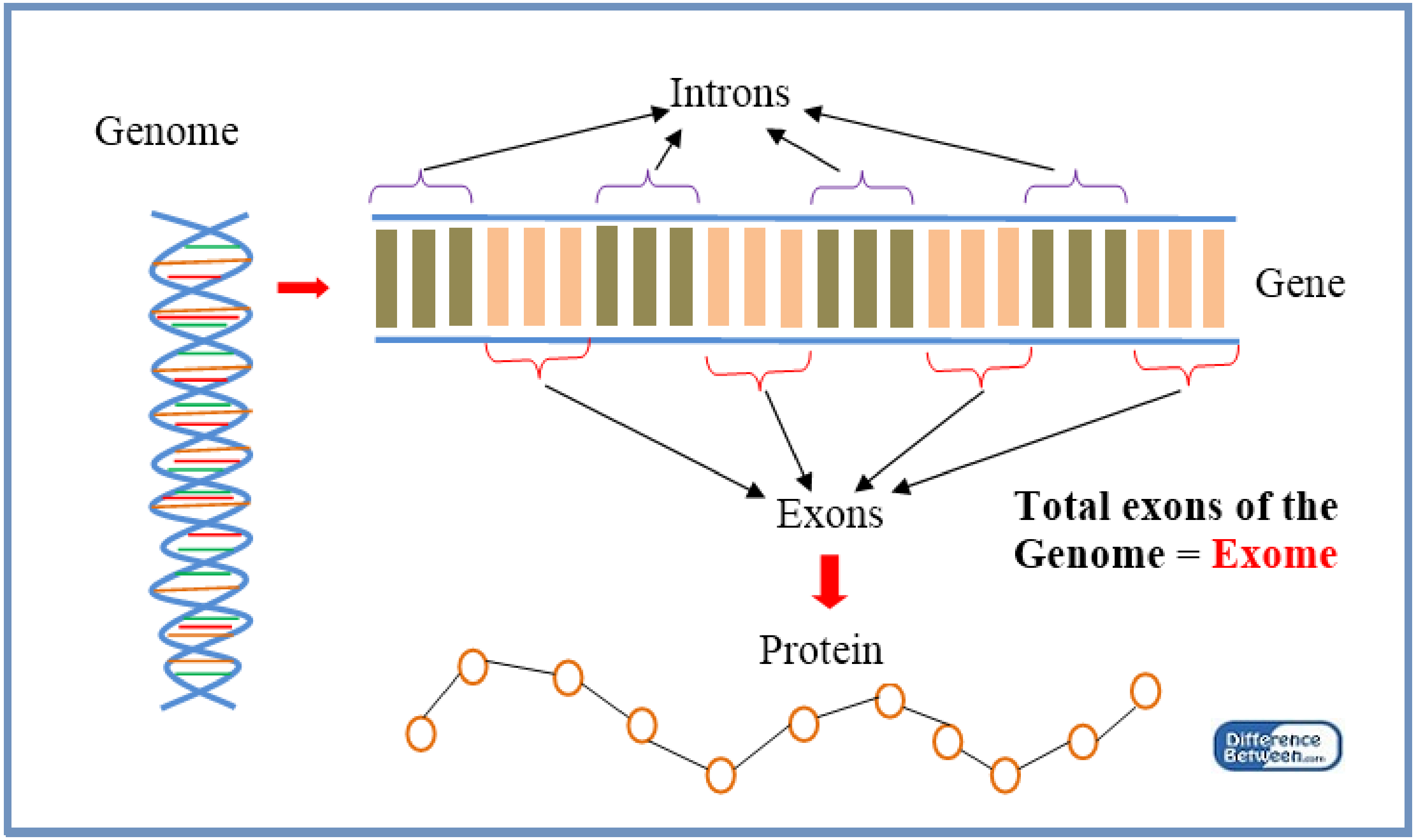
Why do Genome Sequencing in Cancer?

- Complex disease – predisposition and environmental factors

- 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.

- 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)

Genome, Exome, Gene Panel



Total exons of the Genome = Exome

Pros and Cons

WGS

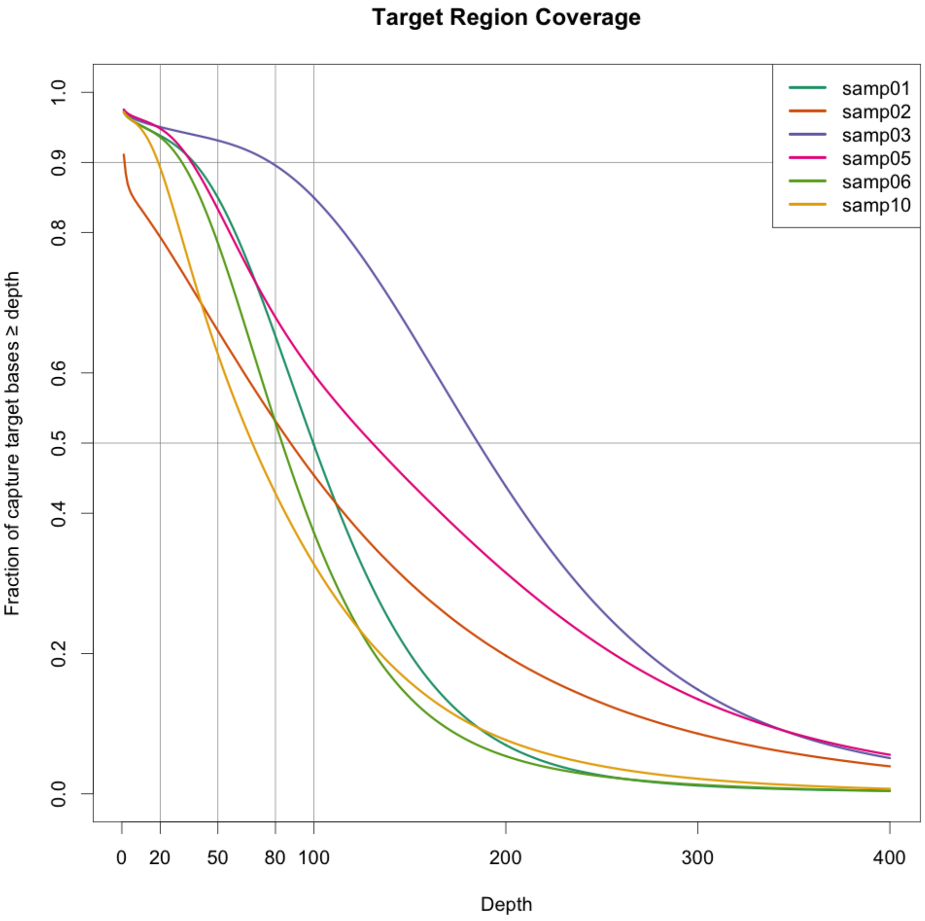
- Can predict large structural differences including CNV
- ~1300$ for 30-40x coverage
- More storage space and time to analysis.

TARGETED PANEL

- Can predict lower AF SNVs with precision.
- ~500$ for 100x Coverage
- Lesser storage space and less time for analysis.

Sequence Coverage & Depth

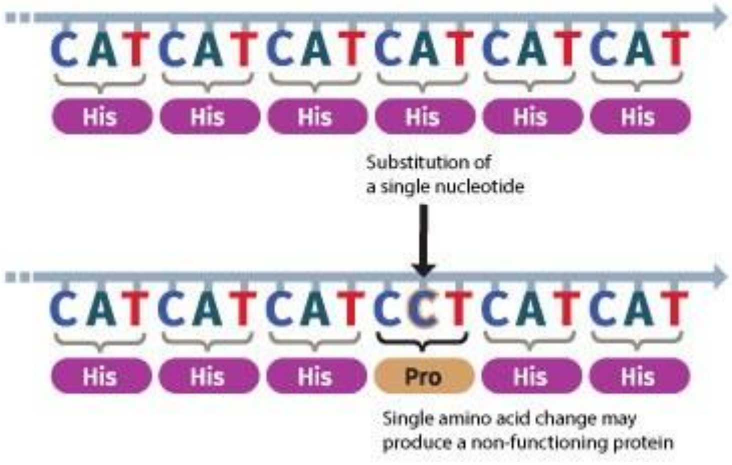
- Base depth is the number of reads that cover a particular base
- Coverage is "how much" of your target did you cover
- Depth of Coverage is how deep was that coverage?



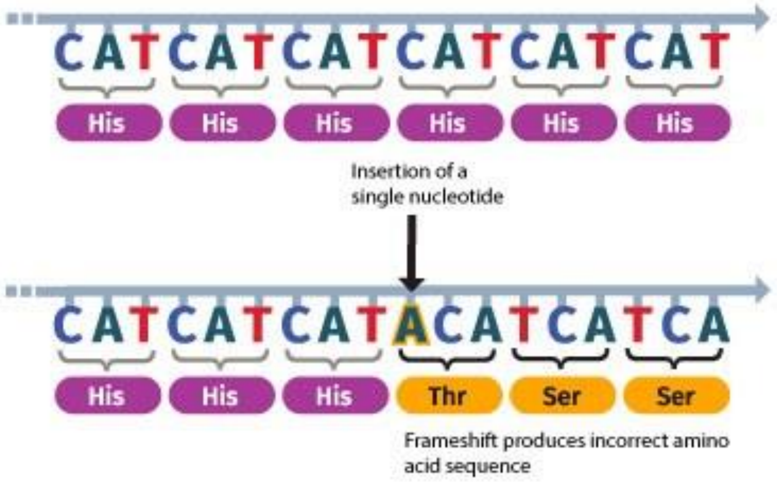
https://www.r-bloggers.com/visualize-coverage-for-targeted-ngs-exome-experiments/

Types of Variation

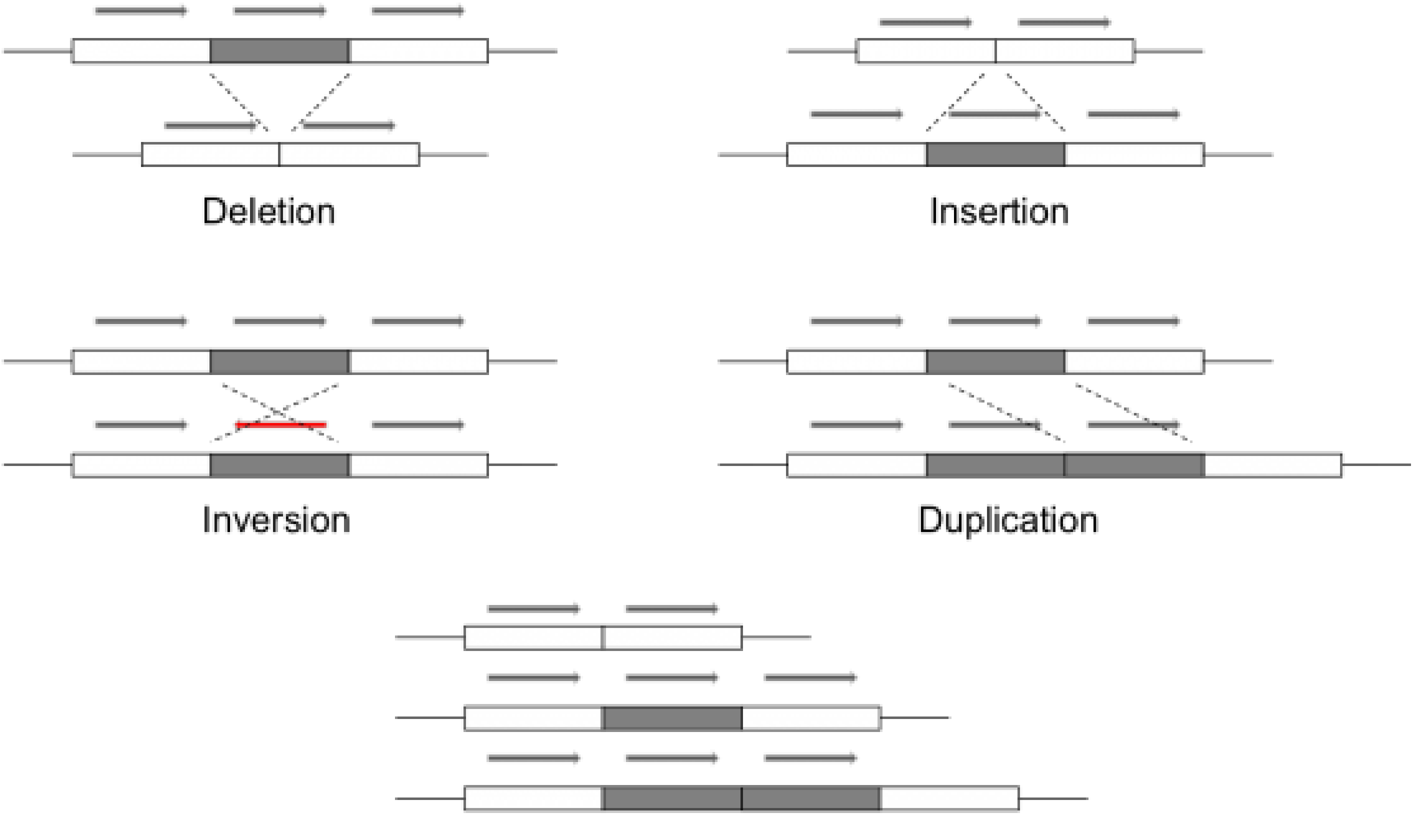
SNV


CATCATCATCATCATCAT
His His His His His His
Substitution of a single nucleotide
CATCATCATCCTCATCAT
His His His Pro His His
Single amino acid change may produce a non-functioning protein

INDEL



Structural Variation



Deletion

Insertion

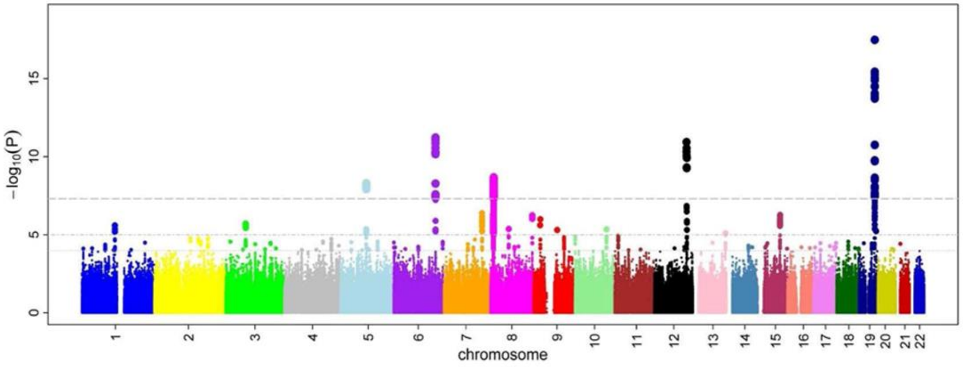
Inversion

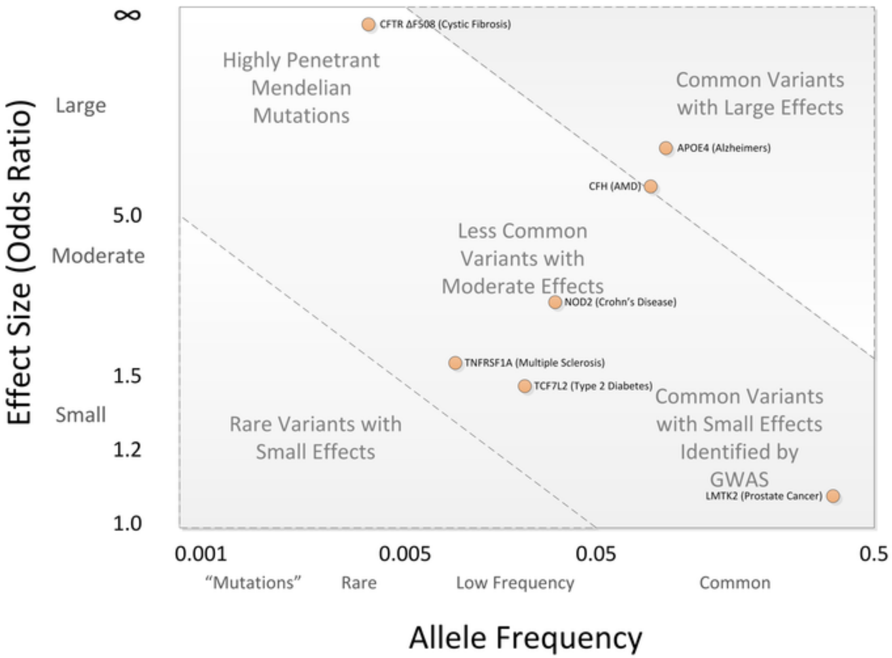
Duplication

Copy Number 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

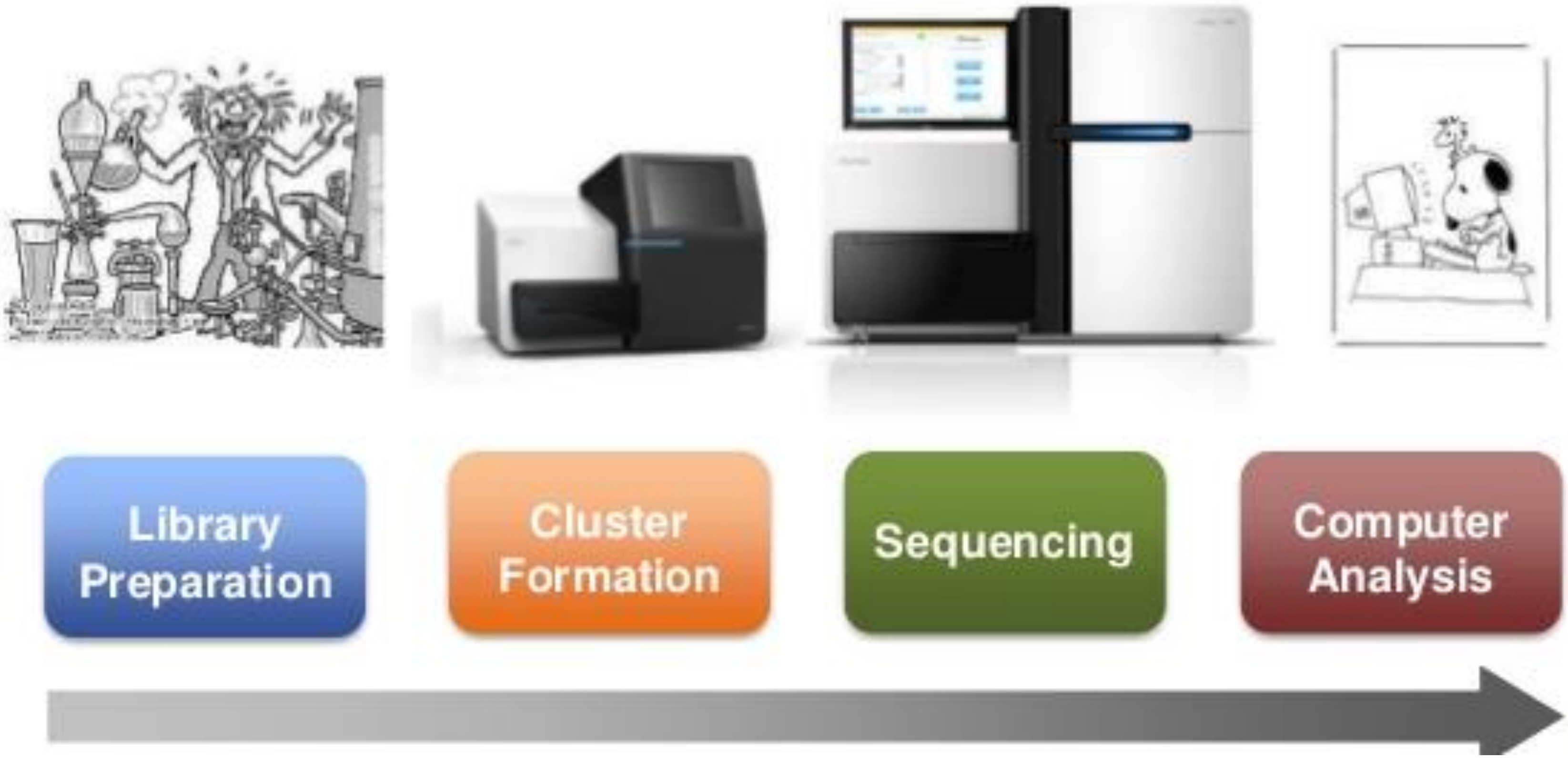
- Phenome-wide association studies (PheWAS) is a quantitative method to determine disease associations can we make with a given gene?
- This is in contrast to GWAS which aims to identify associations, PheWAS aims to explain the cause and effect.
- For example, given a single nucleotide polymorphism (SNP) 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.

WORKFLOWS

Illumina Workflow



Library Preparation

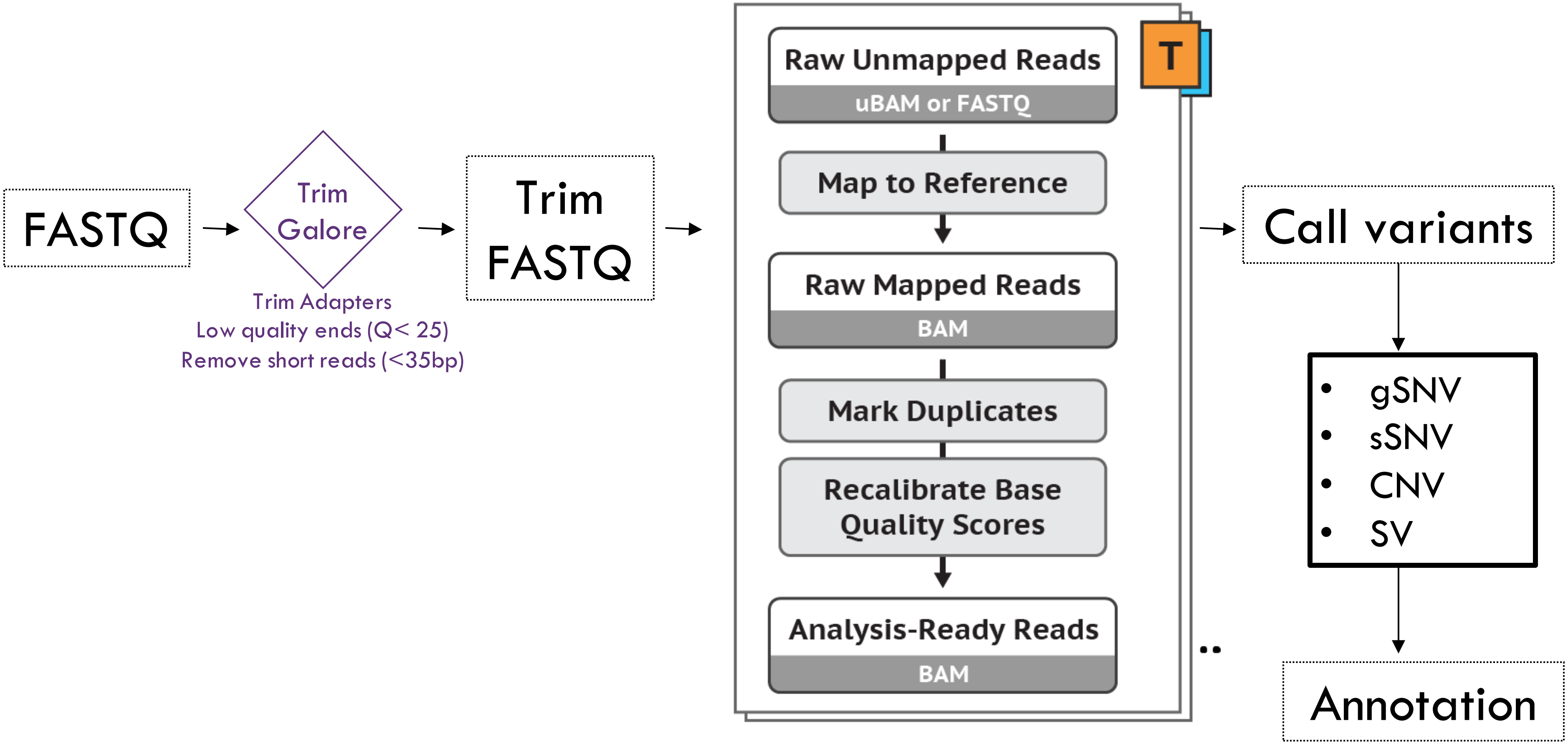
Cluster Formation

Sequencing

Computer Analysis

https://www.google.com/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwjbhdz5ru7hAhUSeawKHR3wC6YQjRx6BAgBEAU&url=https%3A%2F%2Fwww.slideshare.net%2FLutzFr%2Fbioinformatics-workshop-sept-2014&psig=AOvVaw37T4HhhNDYyCx7-s1Xmis2&ust=1556388549295992

Computer Analysis



FASTQ

Trim Galore

Trim Adapters
Low quality ends (Q< 25)
Remove short reads (<35bp)

Trim FASTQ

Raw Unmapped Reads

uBAM or FASTQ

Map to Reference

Raw Mapped Reads

BAM

Mark Duplicates

Recalibrate Base Quality Scores

Analysis-Ready Reads

BAM

Call variants

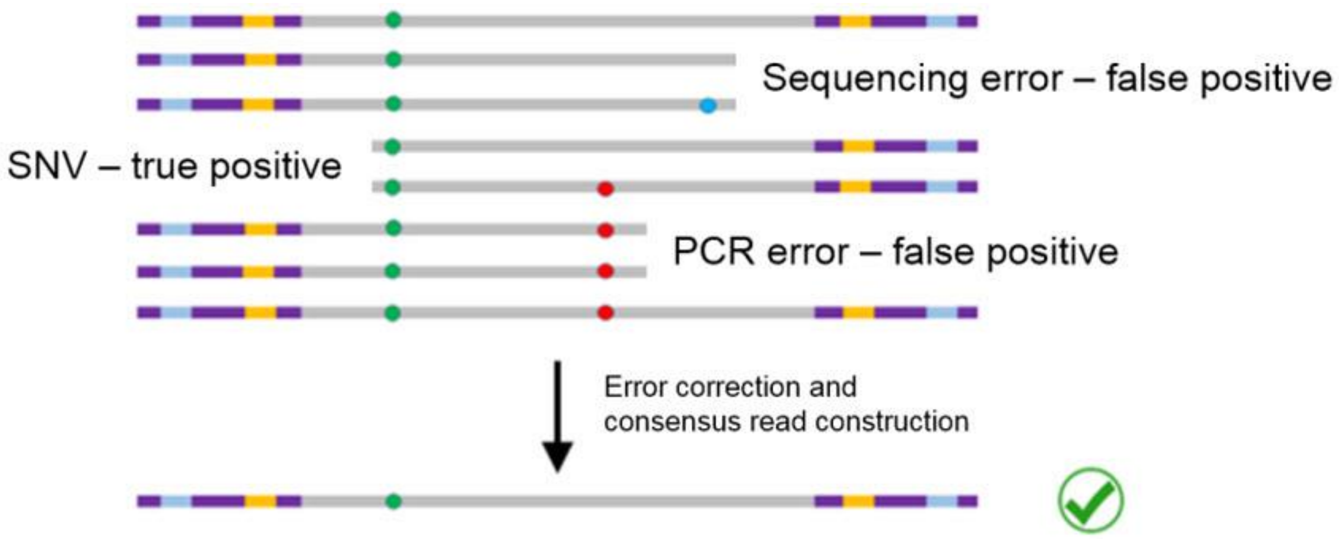
- gSNV
- sSNV
- CNV
- SV

Annotation

https://software.broadinstitute.org/gatk/

Why Worry About Sequence Duplication?

- DNA is sequenced, PCR is used to amplify sequence library to ensure that the DNA with a “known adapter” is sequenced.
- Since PCR has a small error rate, “early errors” can be amplified and could skew the results.

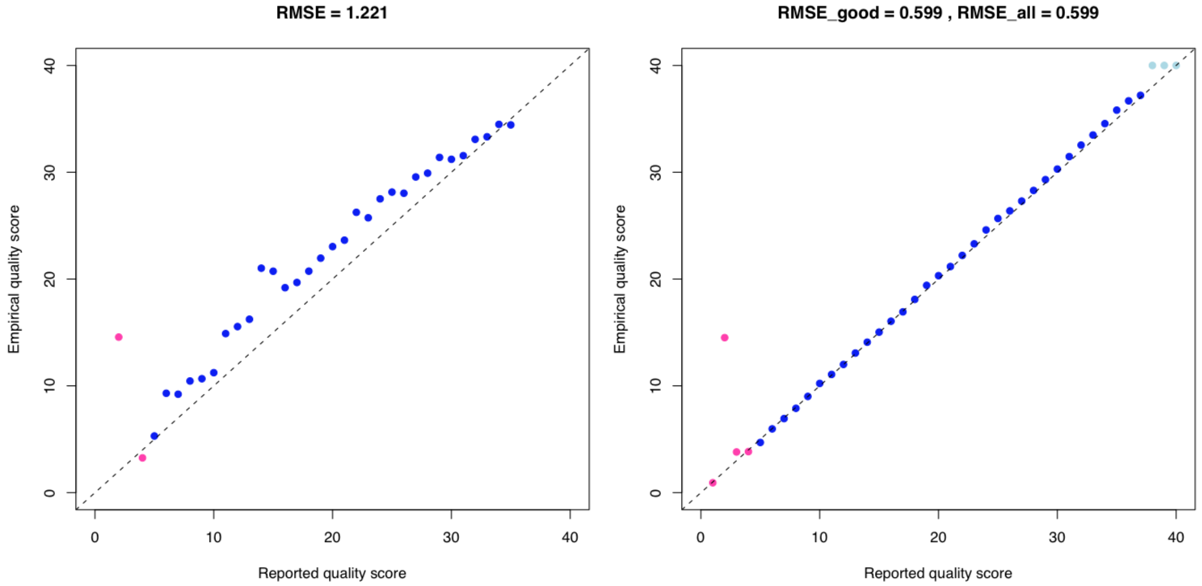

Sequencing error – false positive
SNV – true positive
PCR error – false positive
Error correction and consensus read construction

Accurate SNP discovery depends deeply on good base quality and coverage

Why Base Recalibration?

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

Reported Quality vs. Empirical Quality

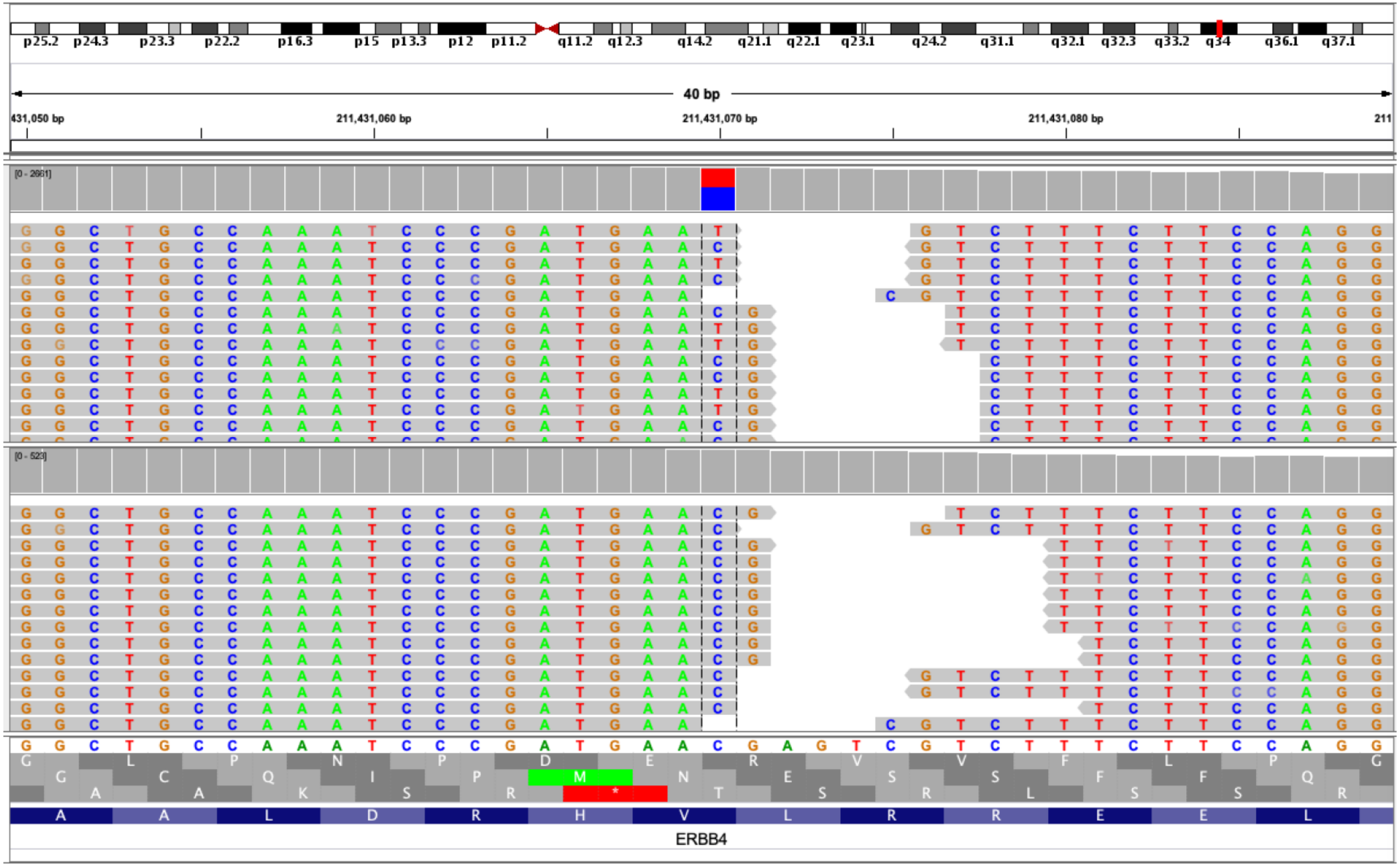


Original Data

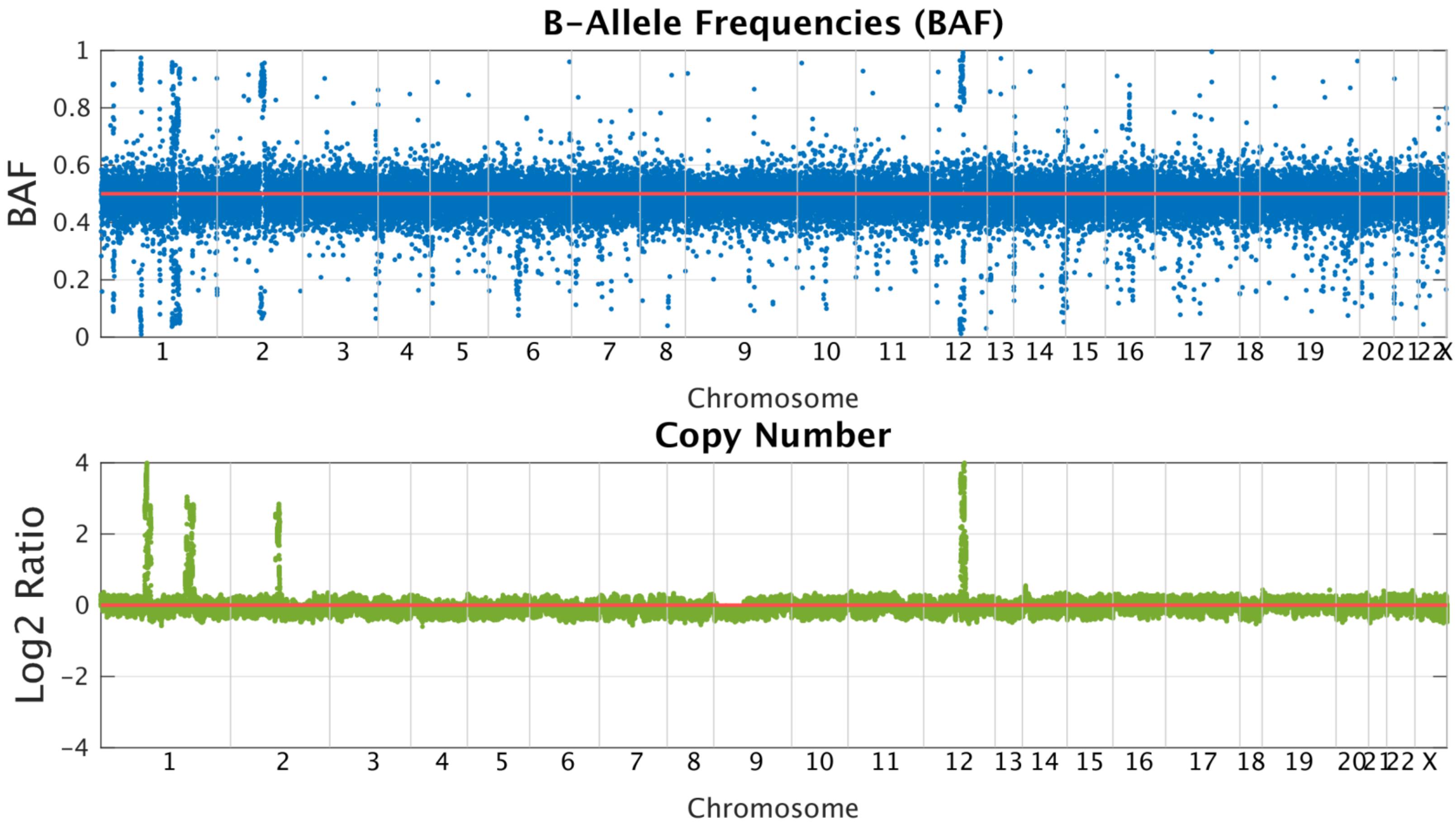
After GATK Recalibration

VARIANT CALLING

SNV

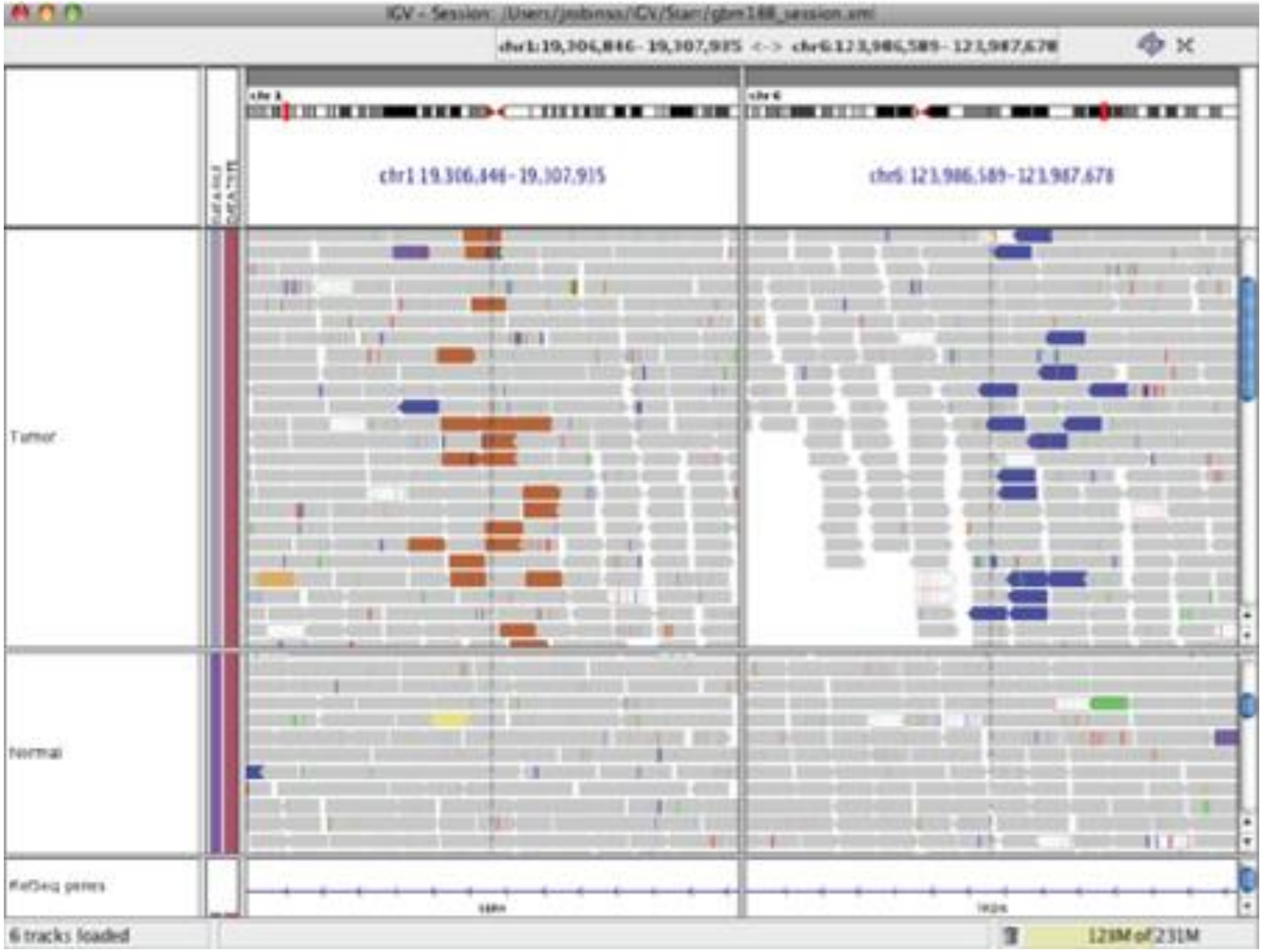

p25.2 p24.3 p23.3 p22.2 p16.3 p15 p13.3 p12 p11.2 q11.2 q12.3 q14.2 q21.1 q22.1 q23.1 q24.2 q31.1 q32.1 q32.3 q33.2 q34 q36.1 q37.1
40 bp
431,050 bp
211,431,060 bp
211,431,070 bp
211,431,080 bp
211
[0 - 2661]
[0 - 523]
G G C T G C C A A A T C C C G A T G A A C G A G T C G T C T T T C T T C C A G G
ERBB4

CNV



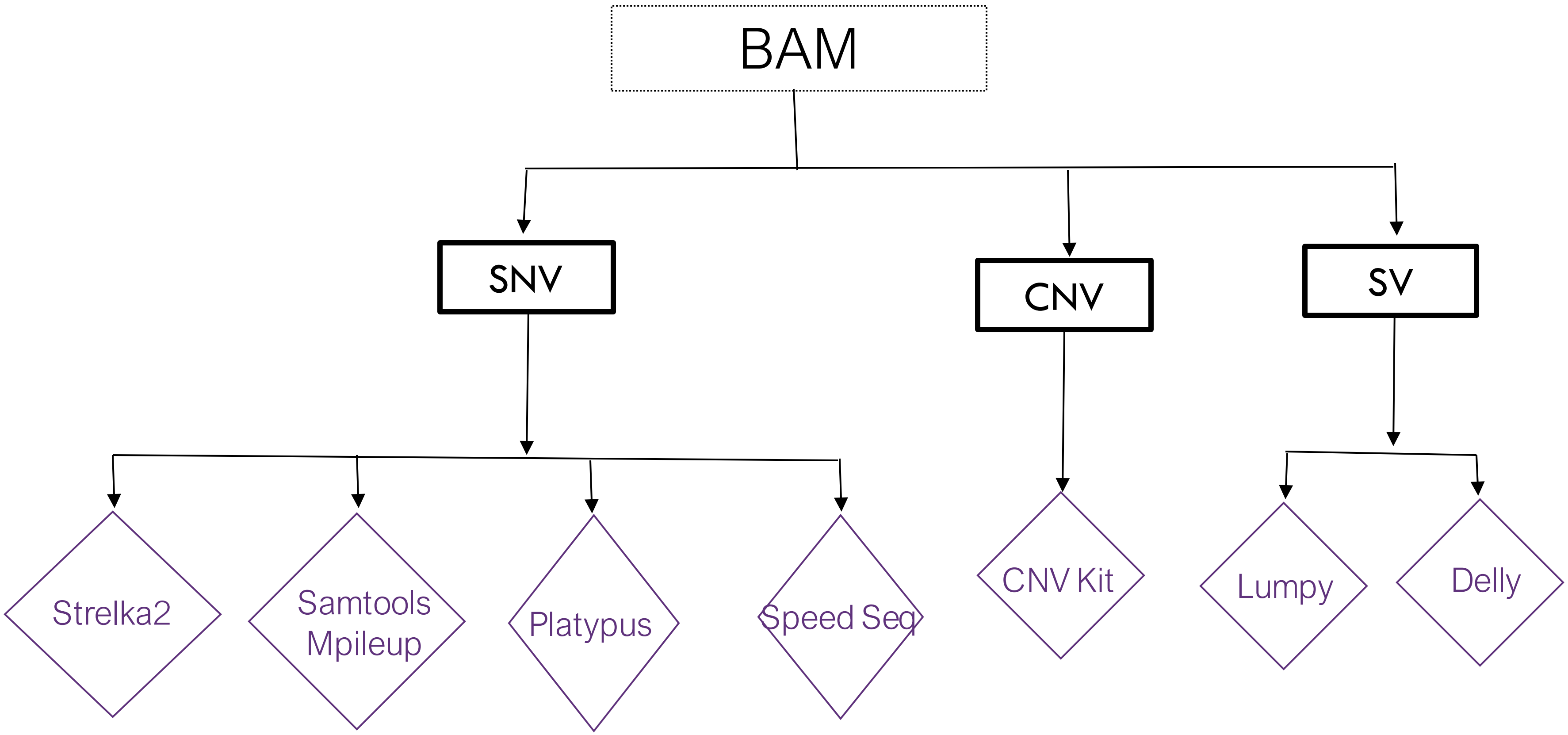
https://github.com/karthigayini/tCoNuT/blob/master/plotting/SAMPLE.cnaBAF.png

SV


IGV - Session: /Users/jrobinso/IGV/Start/gbm188_session.xml
chr1:19,306,846-19,307,935 <-> chr6:123,986,589-123,987,678
chr 1
chr 6
chr1 19,306,846 - 19,307,935
chr6 123,986,589 - 123,987,678
Tumor
Normal
RefSeq genes
6 tracks loaded
123M of 231M

https://software.broadinstitute.org/software/igv/interpreting_insert_size

Germline Workflow



Differences in Results between Callers?

Gene	Amino Acid Change	Variant Type	ExpectedAF	CHROM	START	END	Freebayes	Hotspot	LoFreq	Platypus	GATK	Strelka2	Vscan	Samtools	Scapel	Pindel
NRAS	Q61L	SNP	10	chr1	114713907	114713908	9.1%	8.4%	9.0%	9.2%						
DNMT3A	R882C	SNP	5	chr2	25234373	25234374	4.4%	4.3%	4.4%							
SF3B1	G740E	SNP	5	chr2	197401988	197401989	4.9%	4.7%	5.0%							
IDH1	R132C	SNP	5	chr2	208248388	208248389	3.2%	3.1%	3.2%							
GATA2	G200fs*18	DEL	35	chr3	128485998	128486000	32.8%			28.0%	34.2%	34.2%	32.22%			
TET2	R1261H	SNP	5	chr4	105243756	105243757	4.3%	4.1%	4.4%							
NPM1	W288fs*12	INS	5	chr5	148817378	148817379	2.7%	1.8%							4.6%	
EZH2	R418Q	SNP	5	chr7	148817378	148817379	3.6%	3.3%	3.6%							
JAK2	F537-K539>L	DEL	5	chr9	5070020	5070028	2.3%								3.3%	
JAK2	V617F	SNP	5	chr9	5073769	5073770	3.4%	3.3%	3.4%							
ABL1	T315I	SNP	5	chr9	130872895	130872896	4.0%	3.8%	3.9%							
CBL	S403F	SNP	5	chr11	119278277	119278278	4.3%	4.3%	4.3%							
KRAS	G13D	SNP	40	chr12	25245346	25245347	32.7%	32.0%	32.8%	32.8%	32.9%	32.8%	31.29%	31.3%		
FLT3	D835Y	SNP	5	chr13	28018504	28018505	3.7%	3.6%	3.8%							
IDH2	R172K	SNP	5	chr15	90088605	90088606	4.5%	4.4%	4.5%							
TP53	S241F	SNP	5	chr17	7674240	7674241	5.3%	5.3%	5.4%							
ASXL1	G646fs*12	INS	40	chr20	32434637	32434638	31.5%			31.1%	37.2%	39.2%	32.02%			
ASXL1	W796C	SNP	5	chr20	32435099	32435100	4.9%	4.8%	5.1%							
RUNX1	M267I	SNP	35	chr21	34834413	34834414	33.5%	32.7%	33.4%	33.0%	33.0%	33.2%	32.34%	32.4%		
BCOR	Q1174fs*8	INS	70	chrX	40063831	40063833	63.4%			52.4%	65.1%	67.2%	56.47%		47.1%	
GATA1	Q119*	SNP	10	chrX	48791977	48791978	9.1%		9.1%	9.0%	9.5%					
FLT3	ITD300	300bp INS	5													1.3%

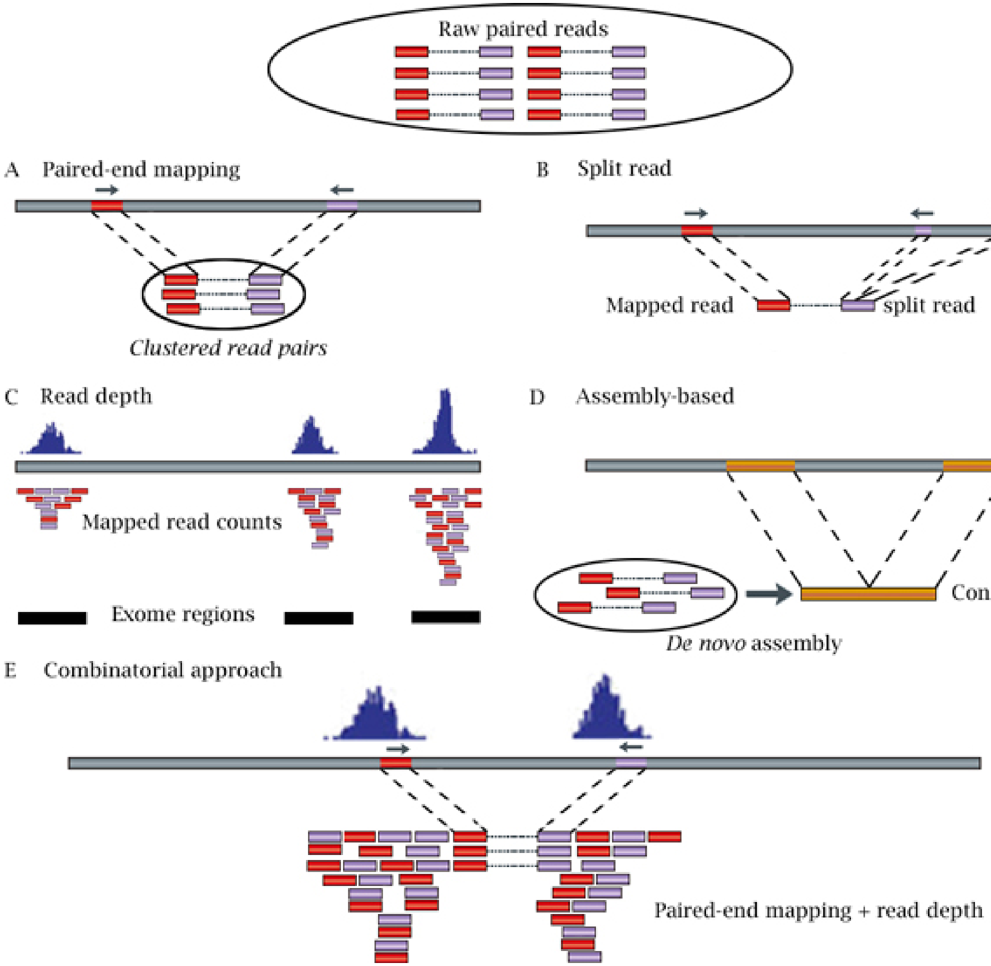
ALGORITHMS

SNV algorithm

Variant caller	Type of variant	Single-sample mode	Type of core algorithm
BAYSIC [48]	SNV	No	Machine learning (ensemble caller)
CaVEMan [34]	SNV	No	Joint genotype analysis
deepSNV [38]	SNV	No	Allele frequency analysis
EBCall [37]	SNV, indel	No	Allele frequency analysis
FaSD-somatic [31]	SNV	Yes	Joint genotype analysis
FreeBayes [44]	SNV, indel	Yes	Haplotype analysis
HapMuC [42]	SNV, indel	Yes	Haplotype analysis
JointSNVMix2 [30]	SNV	No	Joint genotype analysis
LocHap [43]	SNV, indel	No	Haplotype analysis
LoFreq [36]	SNV, indel	Yes	Allele frequency analysis
LoLoPicker [39]	SNV	No	Allele frequency analysis
MutationSeq [45]	SNV	No	Machine learning
MuSE [40]	SNV	No	Markov chain model
MuTect [35]	SNV	Yes	Allele frequency analysis
SAMtools [8]	SNV, indel	Yes	Joint genotype analysis
Platypus [41]	SNV, indel, SV	Yes	Haplotype analysis
qSNP [24]	SNV	No	Heuristic threshold
RADIA [26]	SNV	No	Heuristic threshold
Seurat [33]	SNV, indel, SV	No	Joint genotype analysis
Shimmer [25]	SNV, indel	No	Heuristic threshold
SNooPer [47]	SNV, indel	Yes	Machine learning
SNVSniffer [32]	SNV, indel	Yes	Joint genotype analysis
SOAPsnv [27]	SNV	No	Heuristic threshold
SomaticSeq [46]	SNV	No	Machine learning (ensemble caller)
SomaticSniper [28]	SNV	No	Joint genotype analysis
Strelka [17]	SNV, indel	No	Allele frequency analysis
TVC [97]	SNV, indel, SV	Yes	Ion Torrent specific
VarDict [18]	SNV, indel, SV	Yes	Heuristic threshold
VarScan2 [9]	SNV, indel	Yes	Heuristic threshold
Virmid [29]	SNV	No	Joint genotype analysis

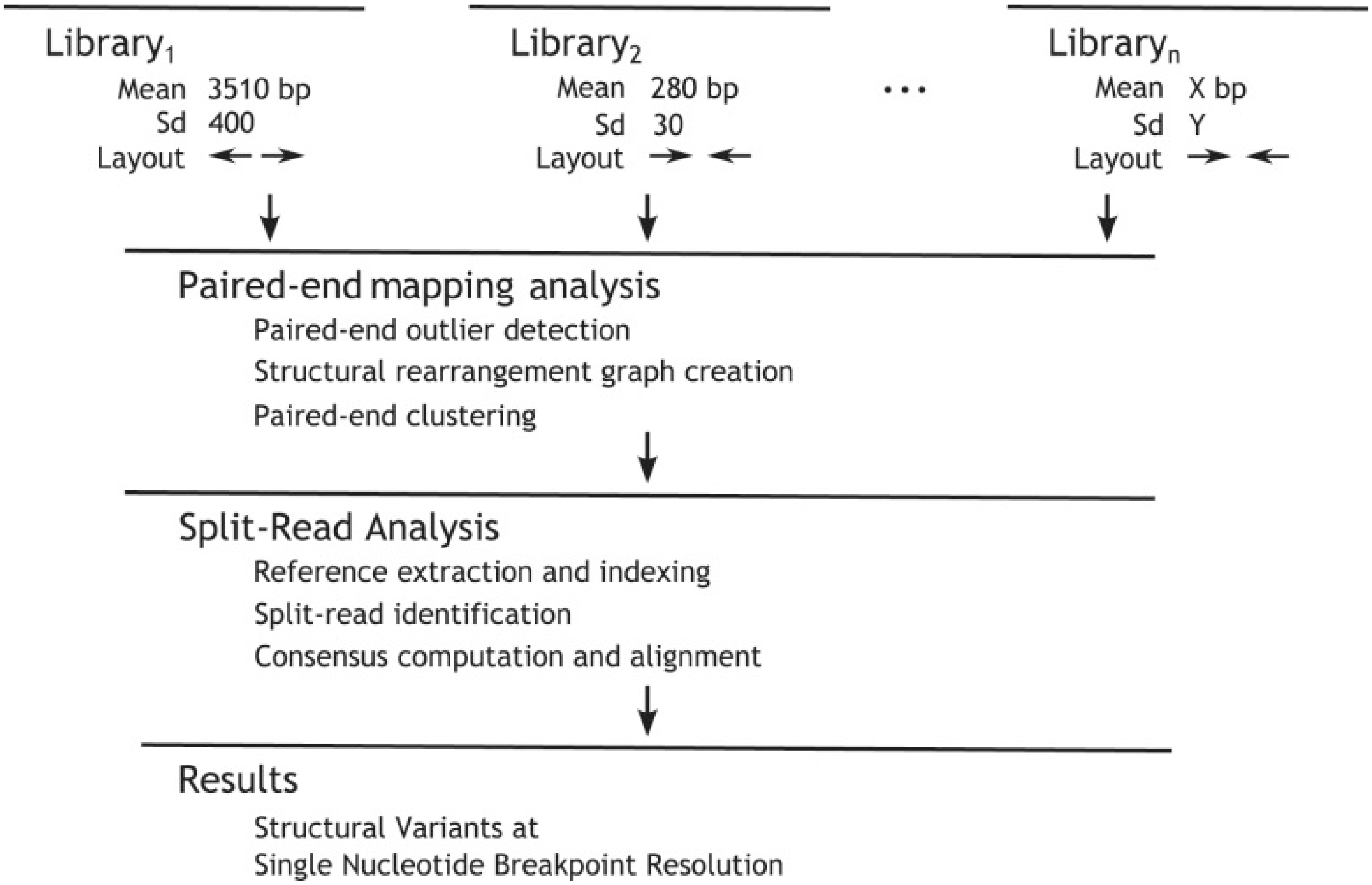
Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput Struct Biotechnol J*. 2018;16:15–24. Published 2018 Feb 6. doi:10.1016/j.csbj.2018.01.003

CNV workflow



Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*. 2013;14 Suppl 11(Suppl 11):S1. doi:10.1186/1471-2105-14-S11-S1

SV working



Library$_1$

Mean 3510 bp

Sd 400

Layout ← →

Library$_2$

Mean 280 bp

Sd 30

Layout → ←

...

Library$_n$

Mean X bp

Sd Y

Layout → ←

Paired-end mapping analysis

Paired-end outlier detection

Structural rearrangement graph creation

Paired-end clustering

Split-Read Analysis

Reference extraction and indexing

Split-read identification

Consensus computation and alignment

Results

Structural Variants at

Single Nucleotide Breakpoint Resolution

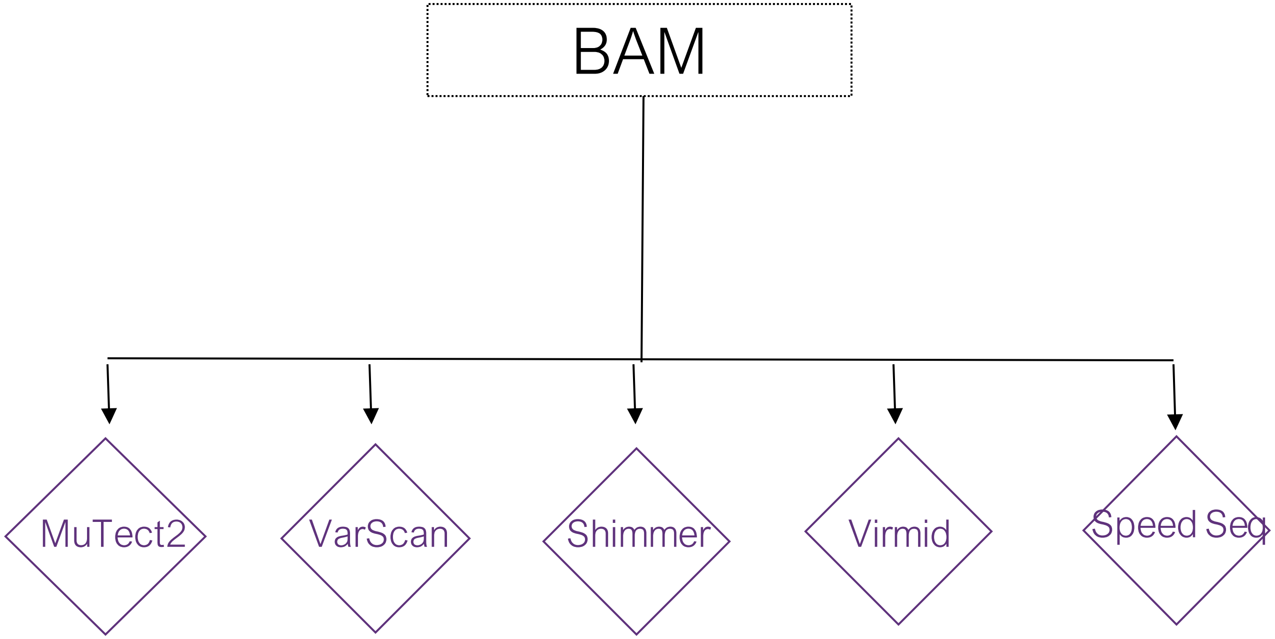
Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;28(18):i333–i339. doi:10.1093/bioinformatics/bts378

Recommended Filtering

Germline

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

Somatic Workflows



Recommended Filtering

- 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

ANNOTATION

Annotation

- 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** - previously reported and is a recognized cause of the disorder.

- **Likely Pathogenic** –previously unreported and is of the type which is expected to cause the disorder.

- **VUS (Variant of Unknown Significance)** –previously unreported and is of the type which may or may not be causative of the disorder.

- **Likely Benign** –previously unreported and is probably not causative of disease.

- **Benign** – a sequence variant is previously reported and is a recognized neutral variant.

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

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

Variant Visualization Tools

- IGV
- http://bam.iobio.io/
- https://vcf.iobio.io

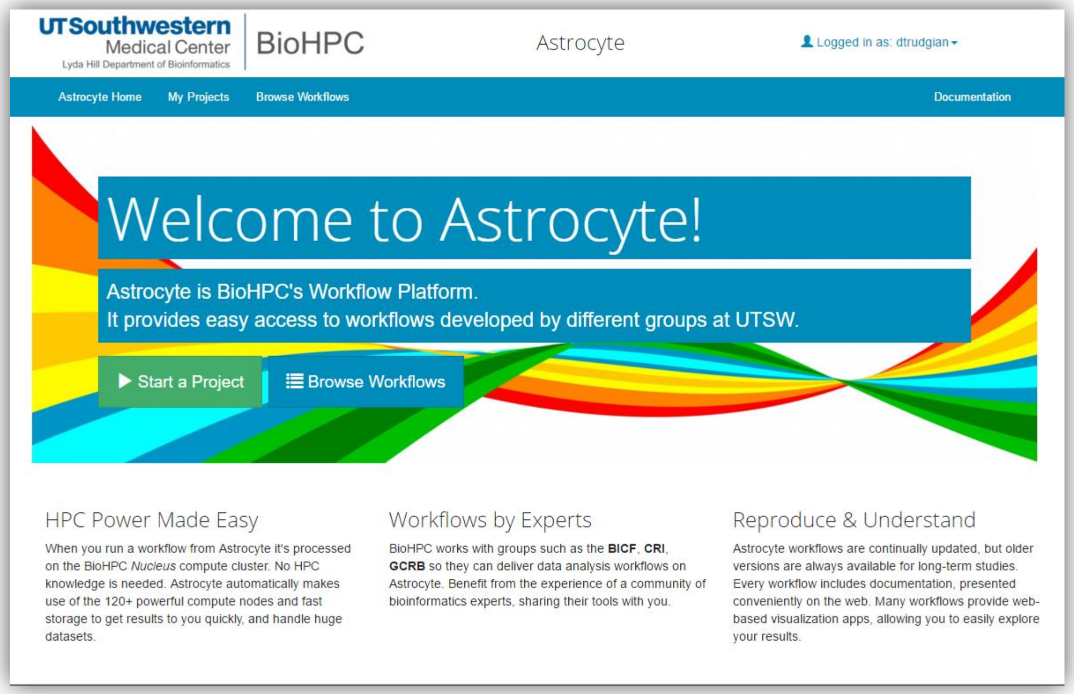
Is there an easy way to run all those command line programs?

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



Welcome to Astrocyte!

Astrocyte is BioHPC's Workflow Platform.
It provides easy access to workflows developed by different groups at UTSW.

HPC Power Made Easy

When you run a workflow from Astrocyte it's processed on the BioHPC *Nucleus* compute cluster. No HPC knowledge is needed. Astrocyte automatically makes use of the 120+ powerful compute nodes and fast storage to get results to you quickly, and handle huge datasets.

Workflows by Experts

BioHPC works with groups such as the **BICF**, **CRI**, **GCRB** so they can deliver data analysis workflows on Astrocyte. Benefit from the experience of a community of bioinformatics experts, sharing their tools with you.

Reproduce & Understand

Astrocyte workflows are continually updated, but older versions are always available for long-term studies. Every workflow includes documentation, presented conveniently on the web. Many workflows provide web-based visualization apps, allowing you to easily explore your results.

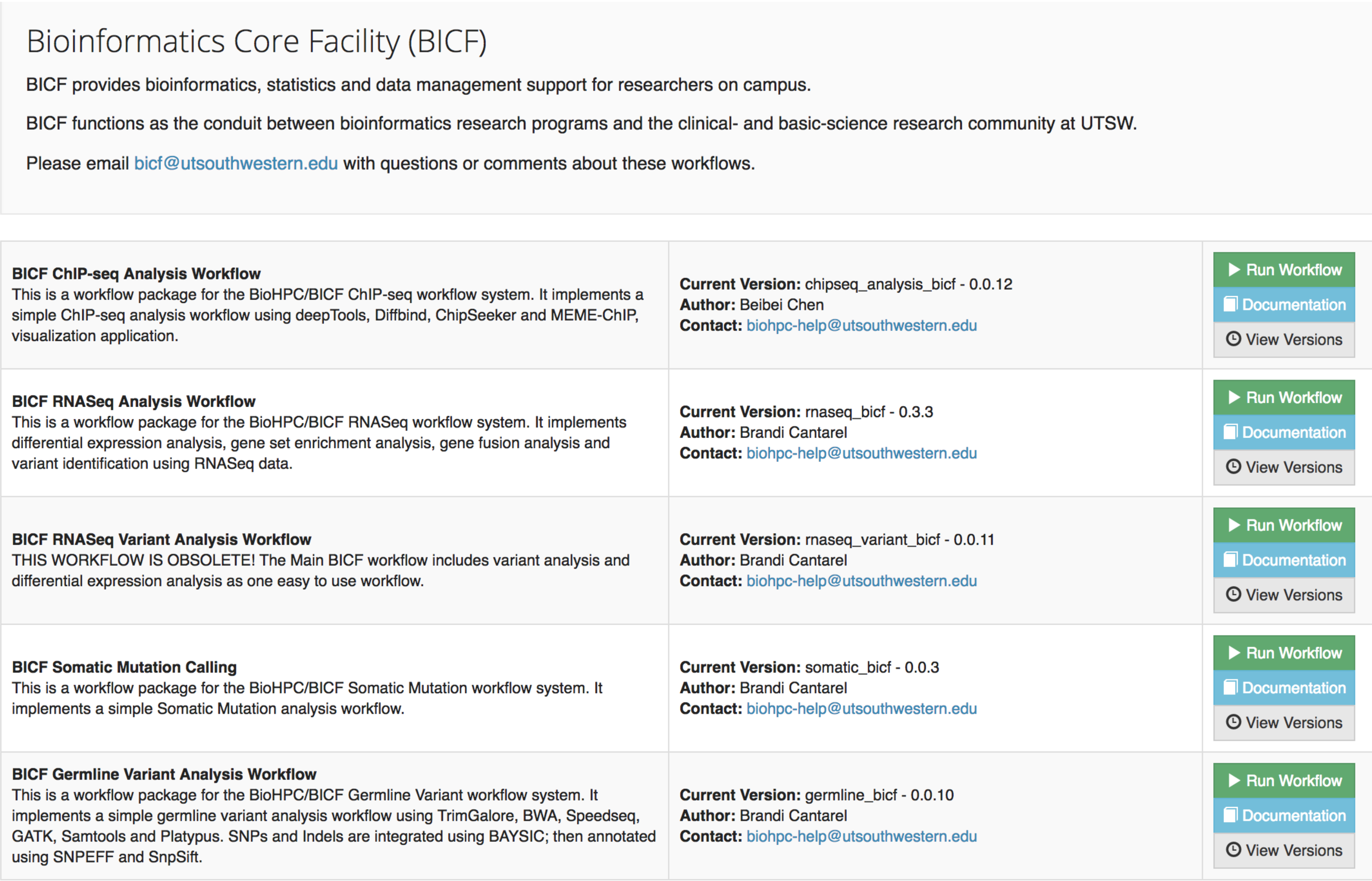
Standardized Workflows

Simple Web Forms

Online documentation & results visualization*

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

astrocyte.biohpc.swmed.edu



Bioinformatics Core Facility (BICF)

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

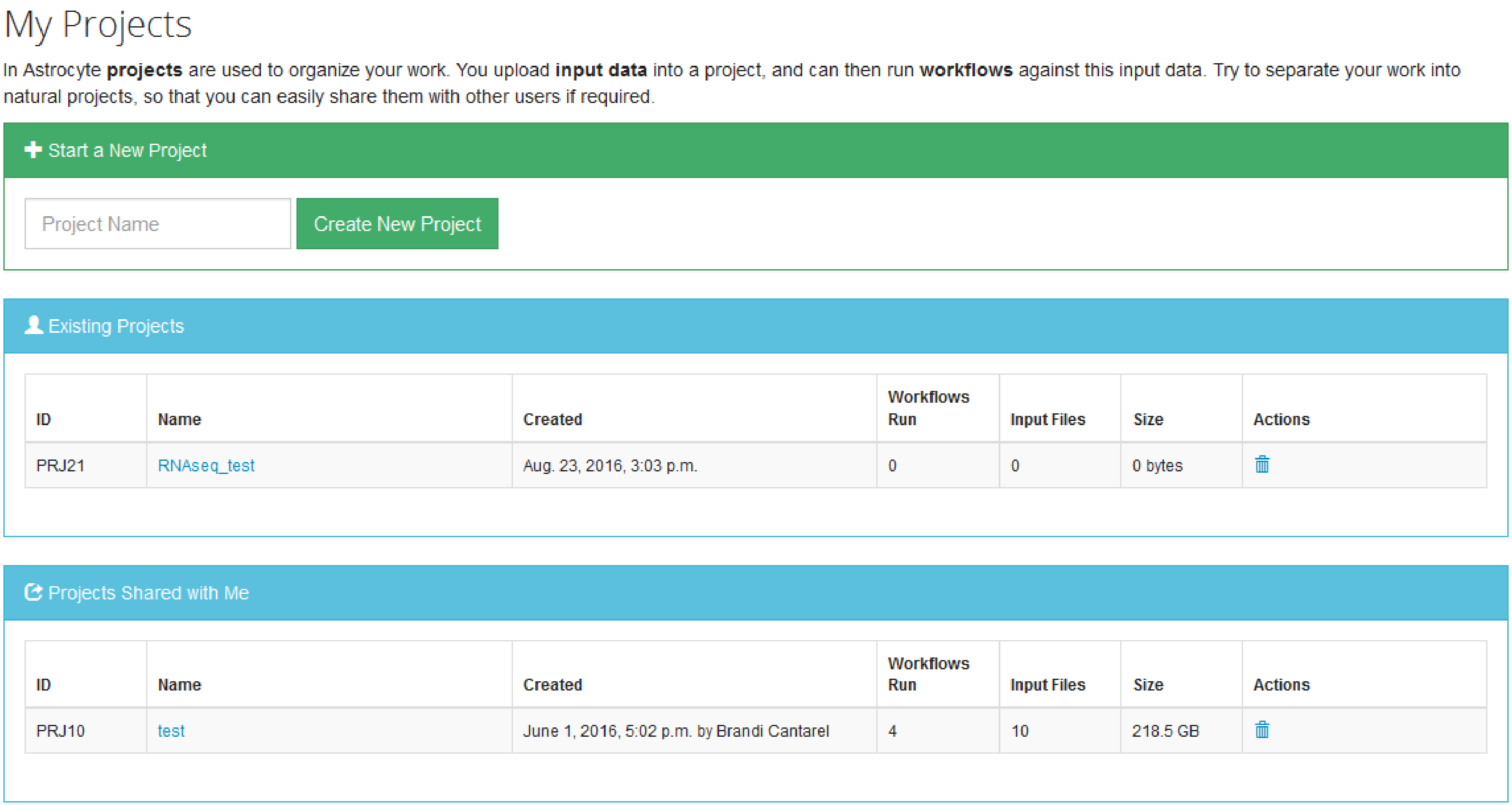
BICF functions as the conduit between bioinformatics research programs and the clinical- and basic-science research community at UTSW.

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

BICF ChIP-seq Analysis Workflow 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 **Author:** Beibei Chen **Contact:** biohpc-help@utsouthwestern.edu	▶ Run Workflow Documentation View Versions
BICF RNASeq Analysis Workflow 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:** rnaseq_bicf - 0.3.3 **Author:** Brandi Cantarel **Contact:** biohpc-help@utsouthwestern.edu	▶ Run Workflow Documentation View Versions
BICF RNASeq Variant Analysis Workflow THIS WORKFLOW IS OBSOLETE! The Main BICF workflow includes variant analysis and differential expression analysis as one easy to use workflow.	**Current Version:** rnaseq_variant_bicf - 0.0.11 **Author:** Brandi Cantarel **Contact:** biohpc-help@utsouthwestern.edu	▶ Run Workflow Documentation View Versions
BICF Somatic Mutation Calling 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 **Author:** Brandi Cantarel **Contact:** biohpc-help@utsouthwestern.edu	▶ Run Workflow Documentation View Versions
BICF Germline Variant Analysis Workflow 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 **Author:** Brandi Cantarel **Contact:** biohpc-help@utsouthwestern.edu	▶ Run Workflow Documentation View Versions

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

Create a new project



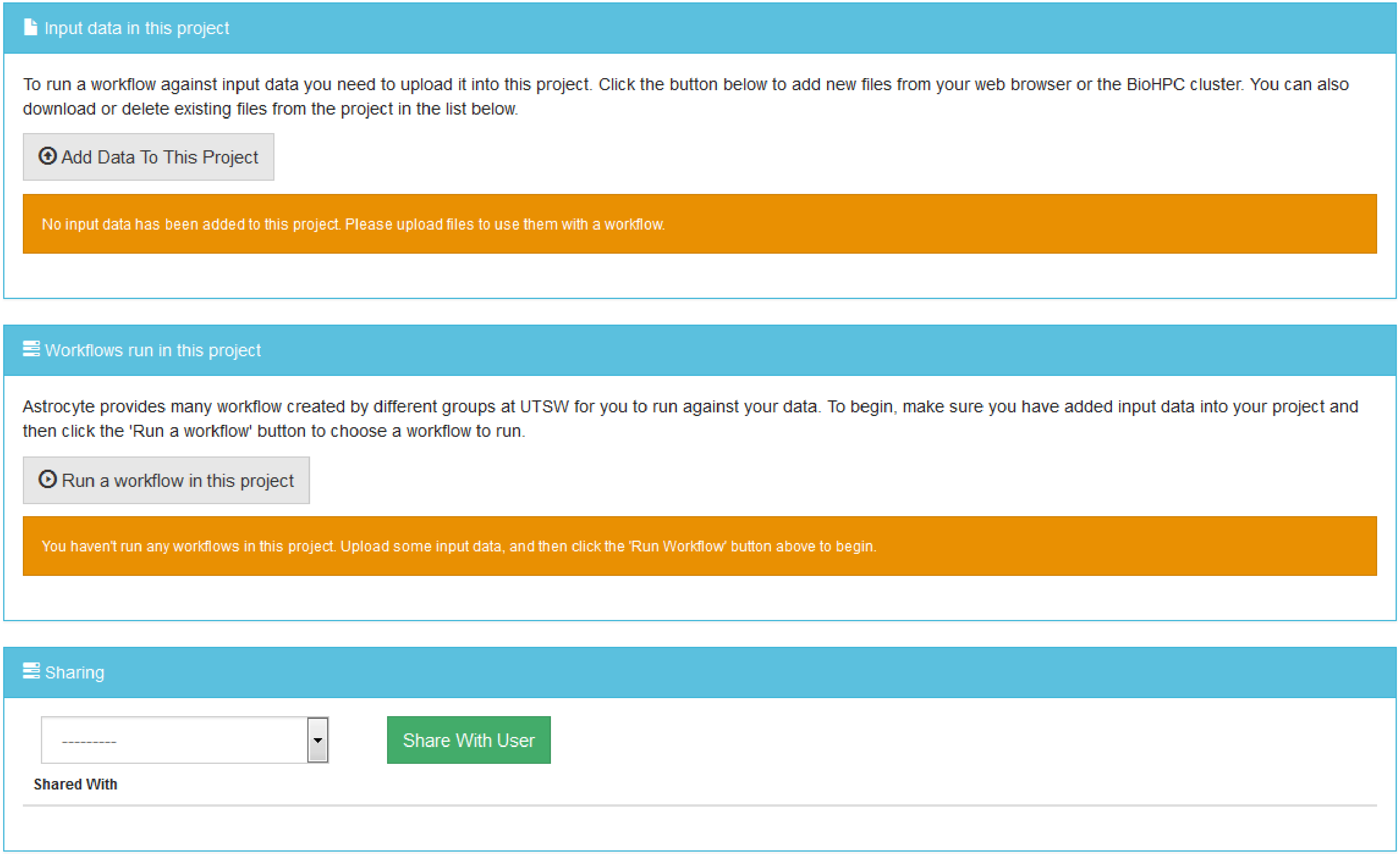
My Projects

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.

ID	Name	Created	Workflows Run	Input Files	Size	Actions
PRJ21	RNAseq_test	Aug. 23, 2016, 3:03 p.m.	0	0	0 bytes	🗑

ID	Name	Created	Workflows Run	Input Files	Size	Actions
PRJ10	test	June 1, 2016, 5:02 p.m. by Brandi Cantarel	4	10	218.5 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.

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

Workflows run in this project

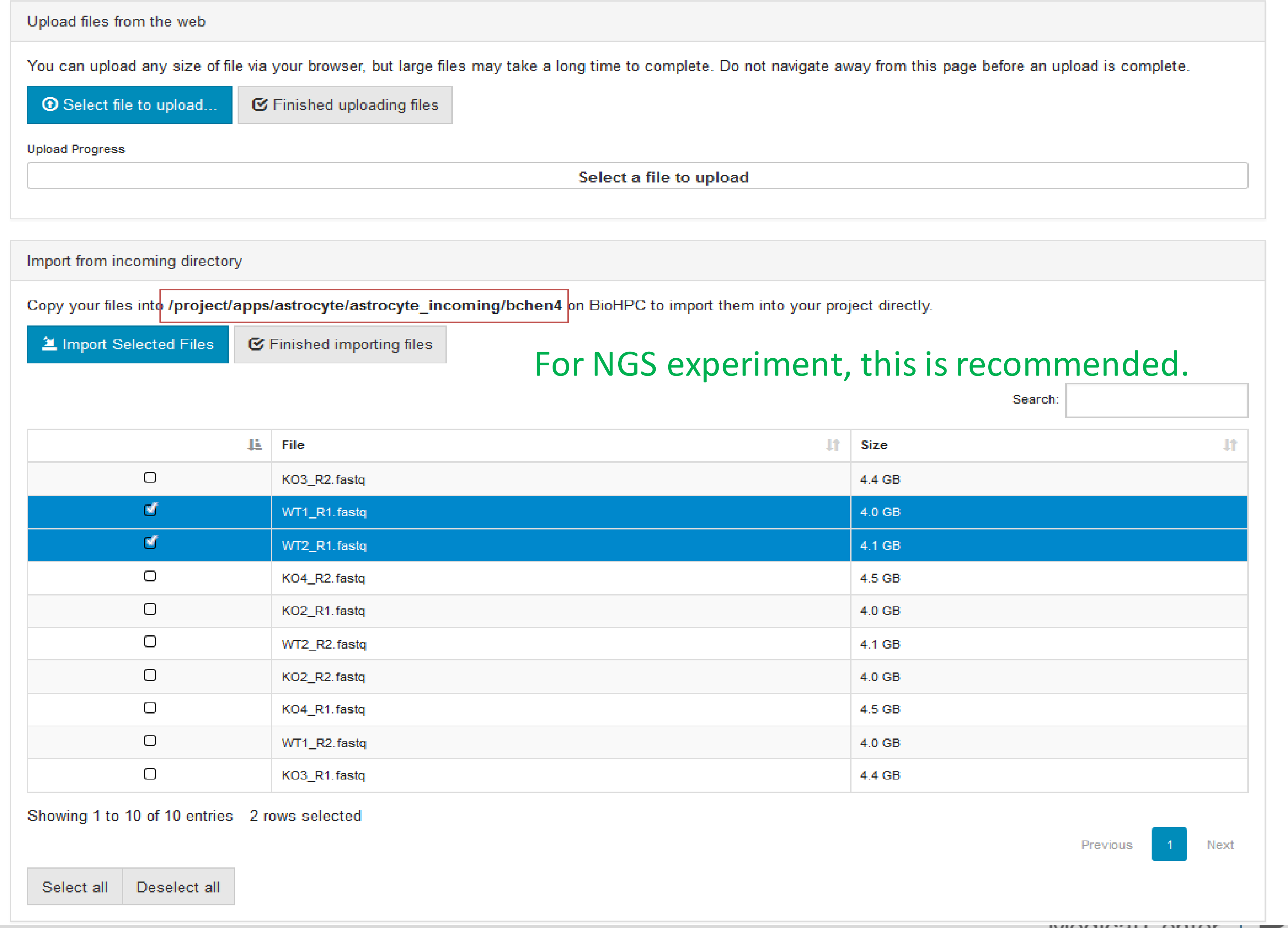
Astrocyte provides many workflow created by different groups at UTSW for you to run against your data. To begin, make sure you have added input data into your project and then click the 'Run a workflow' button to choose a workflow to run.

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

Sharing

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.

Upload Progress

Select a file to upload

Import from incoming directory

Copy your files into **/project/apps/astrocyte/astrocyte_incoming/bchen4** on BioHPC to import them into your project directly.

For NGS experiment, this is recommended.

Search:

	File	Size
☐	KO3_R2.fastq	4.4 GB
☑	WT1_R1.fastq	4.0 GB
☑	WT2_R1.fastq	4.1 GB
☐	KO4_R2.fastq	4.5 GB
☐	KO2_R1.fastq	4.0 GB
☐	WT2_R2.fastq	4.1 GB
☐	KO2_R2.fastq	4.0 GB
☐	KO4_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)

Fu

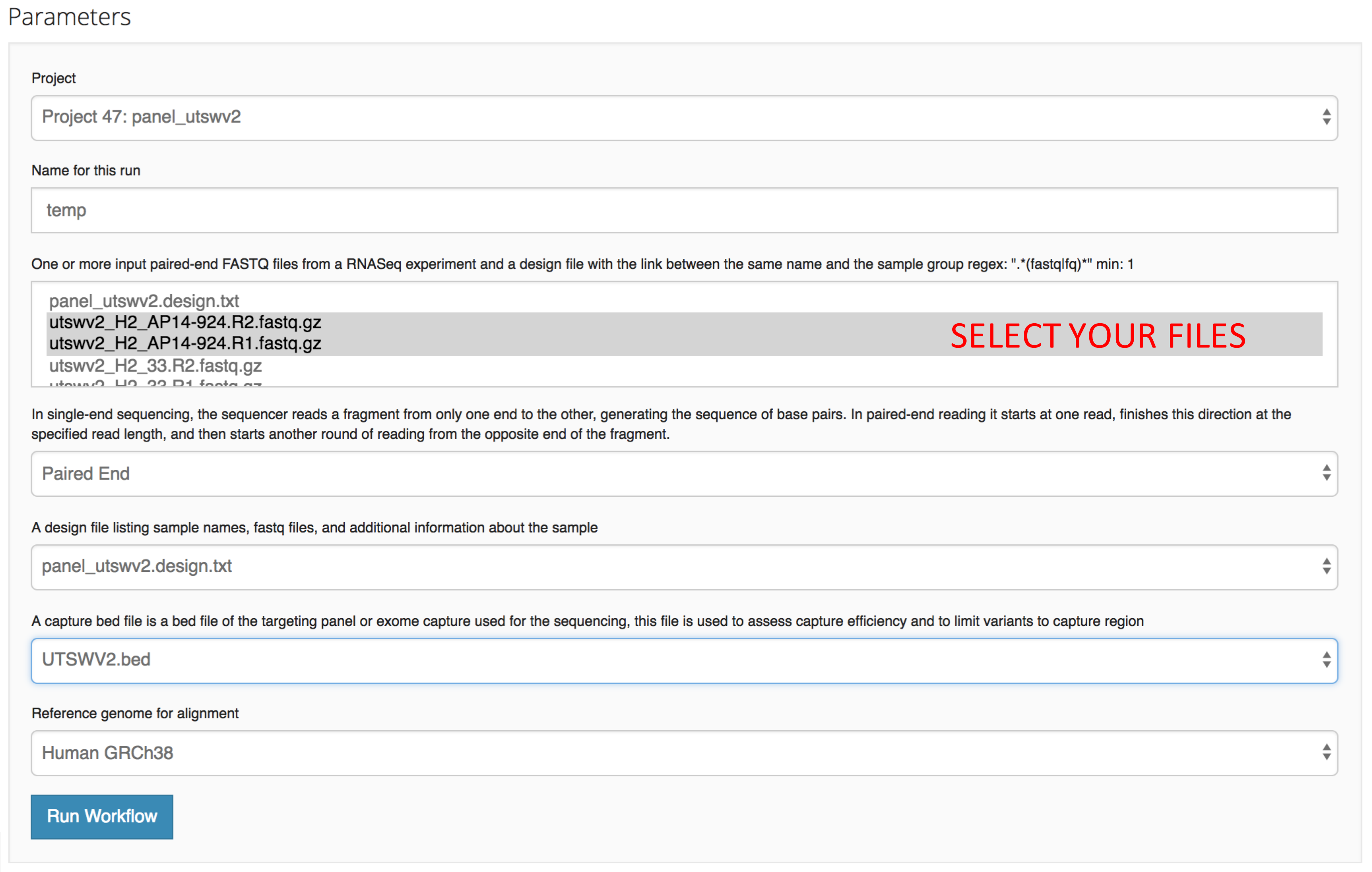
Na

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
utswv2_H2_33.R1.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

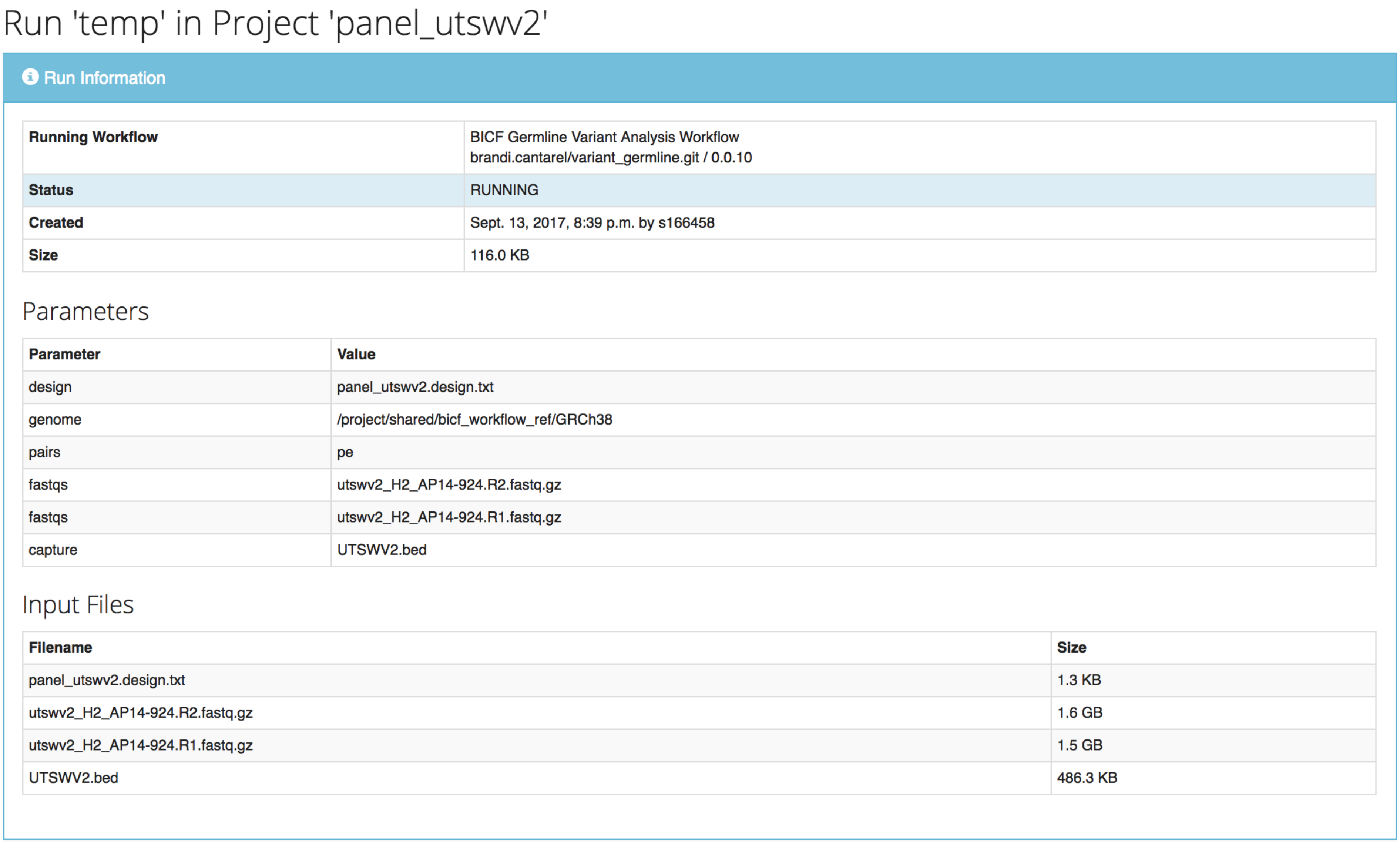
A capture bed file is a bed file of the targeting panel or exome capture used for the sequencing, this file is used to assess capture efficiency and to limit variants to capture region

UTSWV2.bed

Reference genome for alignment

Human GRCh38

Project is running



Run 'temp' in Project 'panel_utswv2'

Run Information

Running Workflow	BICF Germline Variant Analysis Workflow brandi.cantarel/variant_germline.git / 0.0.10
Status	RUNNING
Created	Sept. 13, 2017, 8:39 p.m. by s166458
Size	116.0 KB

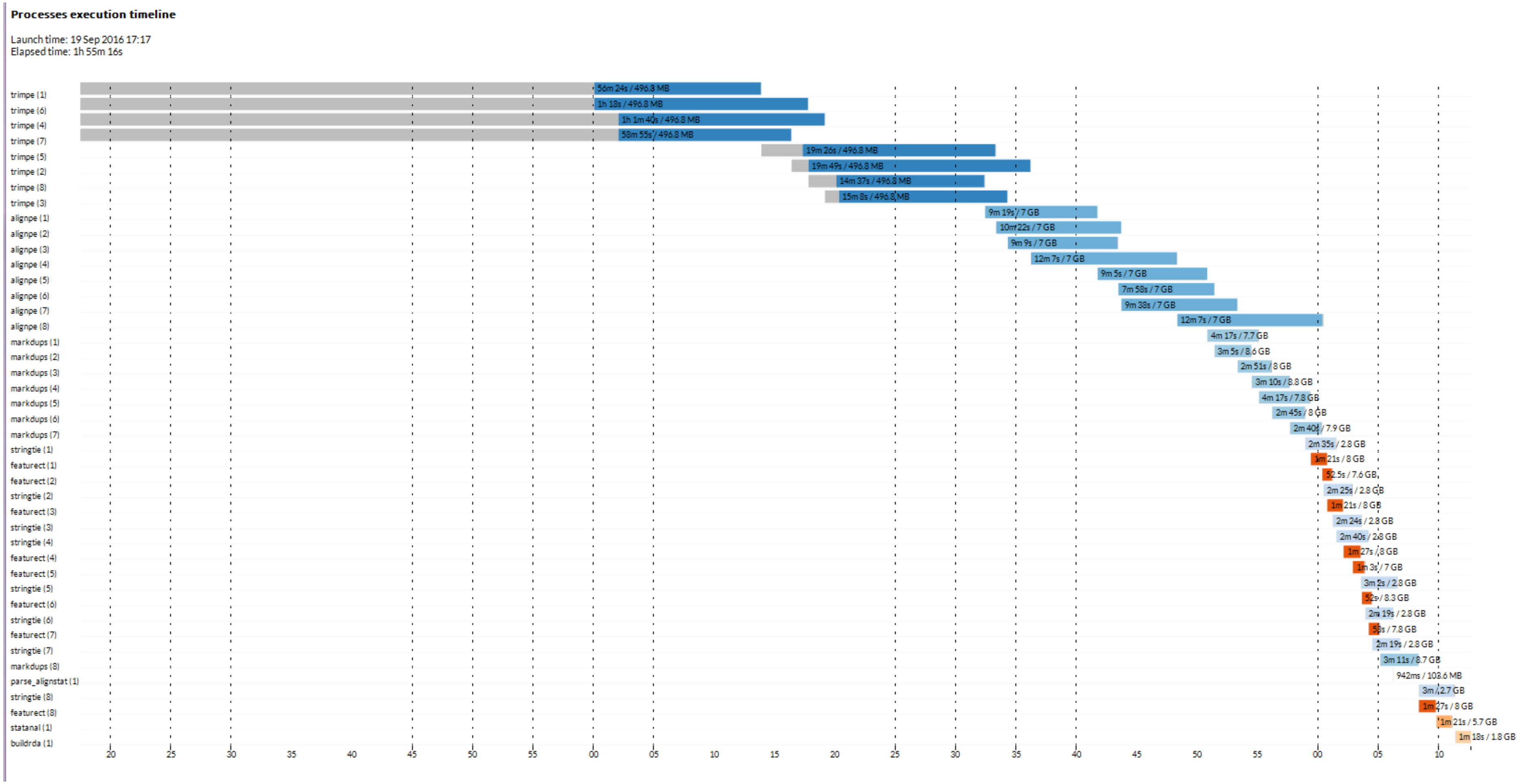
Parameters

Parameter	**Value**
design	panel_utswv2.design.txt
genome	/project/shared/bicf_workflow_ref/GRCh38
pairs	pe
fastqs	utswv2_H2_AP14-924.R2.fastq.gz
fastqs	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	1.6 GB
utswv2_H2_AP14-924.R1.fastq.gz	1.5 GB
UTSWV2.bed	486.3 KB

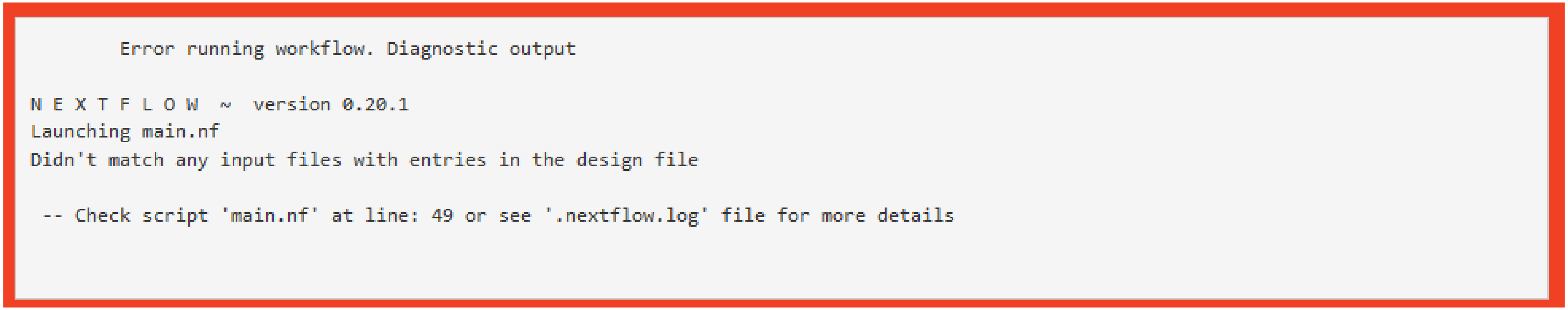
Timeline of the whole run



Processes execution timeline

Launch time: 19 Sep 2016 17:17
Elapsed time: 1h 55m 16s

Common errors and solutions

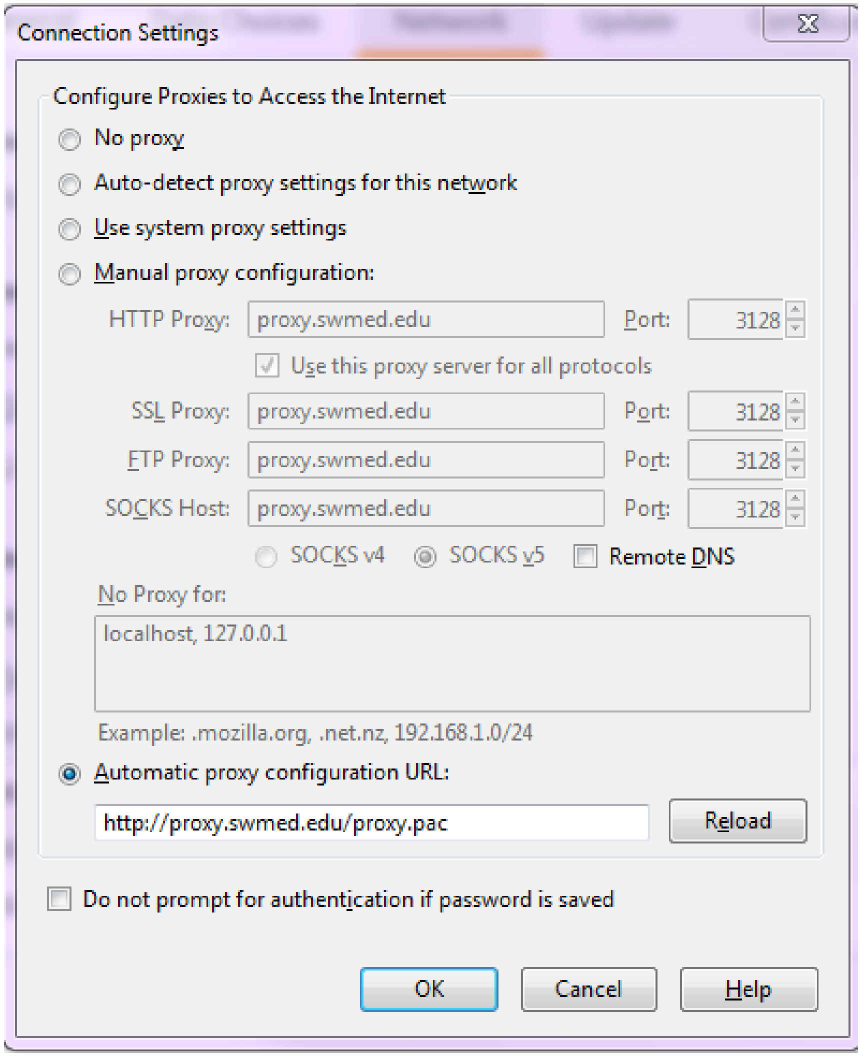


```
        Error running workflow. Diagnostic output

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



Connection Settings

Configure Proxies to Access the Internet

○ No proxy

○ Auto-detect proxy settings for this network

○ Use system proxy settings

○ Manual proxy configuration:

HTTP Proxy: proxy.swmed.edu Port: 3128

☑ Use this proxy server for all protocols

SSL Proxy: proxy.swmed.edu Port: 3128

FTP Proxy: proxy.swmed.edu Port: 3128

SOCKS Host: proxy.swmed.edu Port: 3128

○ SOCKS v4 ● SOCKS v5 □ Remote DNS

No Proxy for:

localhost, 127.0.0.1

Example: .mozilla.org, .net.nz, 192.168.1.0/24

● Automatic proxy configuration URL:

http://proxy.swmed.edu/proxy.pac

□ Do not prompt for authentication if password is saved

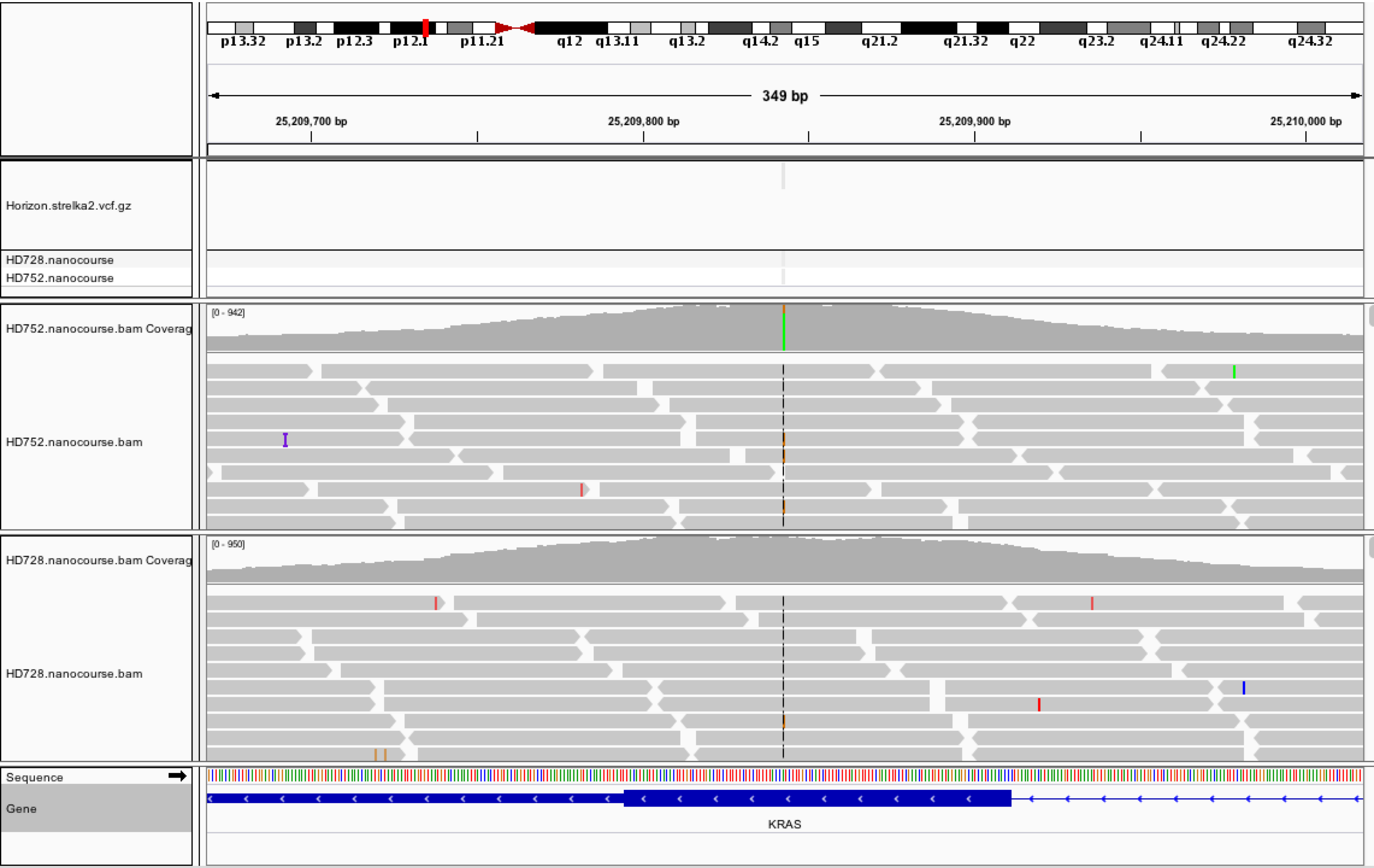
Key Files Germline Pipeline

- 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



p13.32
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
q24.32
349 bp
25,209,700 bp
25,209,800 bp
25,209,900 bp
25,210,000 bp
Horizon.strelka2.vcf.gz
HD728.nanocourse
HD752.nanocourse
HD752.nanocourse.bam Coverag
[0 - 942]
HD752.nanocourse.bam
HD728.nanocourse.bam Coverag
[0 - 950]
HD728.nanocourse.bam
Sequence
Gene
KRAS

Questions?