**Analyzing Human Sequence Data Using Civet**

**Prepared By**

**Anuj Srivastava**

**Things to note prior to run these civet pipelines:**

1. These pipelines are for human Exome/WholeGenome/RNA-Seq/PDX data analysis
2. The reference genome for Exome/WholeGenome is **hg19 with unplaced contigs**
3. The reference genome and annotation of RSEM based RNA-Seq is **Homo\_sapiens\_GRCh37\_70.fa and Homo\_sapiens\_final\_chr\_added.GRCh37.70.gtf** (downloaded from ENSEMBL and “chr” is added to chromosome name in genome and gtf files)
4. The RSEM Pipeline also used TCGA ref and gtf and provide expression estimate which are in TCGA format (useful in PDX analysis)
5. The Tophat pipeline uses UCSC hg19 with unplaced contigs, annotation file and does the **novel isoform** prediction and also **variant calling** using GATK RNA-Seq best practices
6. NOD genome and hg19 with unplaced contigs is used to make Xenome (genome index) for reads classification in DNA Pipelines
7. NOD genome and hg19 with unplaced contigs, **transcriptome** is used to make Xenome (transcriptome index) for reads classification in RNA Pipelines
8. Any pipeline will halt if more than 50 % reads were filtered out during quality control step.
9. PDX pipelines using Xenome will also halt if less than 75% of reads is classified as human
10. Tumor/Normal Pipeline uses somatic-sniper and GATK Indel detector for SNP and indel calling, respectively.
11. Scripts were written to combine SNP/Indel results into one file for all pipelines.
12. Variant calls were restricted to chr1-22, X-Y, M only.

**Target files used in whole exome analysis are present in**

/data/shared/cga\_reference\_data/

* agilent\_SureSelect\_V4\_pChrM\_probes.bed
* agilent\_SureSelect\_V4\_pChrM\_probes\_picard\_updated.bed

**Target files used in CTP analysis are:**

/data/shared/research\_pipelines\_reference\_data/human/DNA/bwa\_index\_chrs\_contigs/CTP\_targets\_baits

* CGA\_exon\_list\_371\_genes\_targets\_gatk\_picard.bed
* 359genes\_targets\_sorted\_gatk\_picard.bed
* 359genes\_targets\_sorted\_gatk.bed

**Input fastq file should follow this naming convention:**

**\*\_R1\_ALL.fastq**

**\*\_R2\_ALL.fastq**

**\*\_R1 is important for one of the downstream script to work.**

All Pipelines are location on Cadillac:

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/

**Note: If you want to change the config file parameters then copy the file in your working directory and then change/run with that config file.**

**ALL EXOME PIPELINES (CTP, Whole Exome, PDX)**

1. **Whole Exome for paired end data (Regular, No Xenome)**
2. **Important tool used:**

* Filter\_trim.py script
* BWA mem
* Picard (Mark Duplicate)
* GATK indel realigner
* GATK BQSR
* GATK Unified Genotyper (Default ploidy = 4)
* snpEFF and snpSift for annotation; dbSNP and Cosmic annotation added are also present in VCF files.

1. **Invoking Pipeline:**

module load civet/latest (Need once)

civet\_run –o /opt/compsci/civet\_research\_pipelines/4.3.6/Human/Exome/WholeExomeSingleSample/config\_file\_Whole\_Exome\_Single\_Sample

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/Exome/WholeExomeSingleSample/WholeExomeSingleSample.xml

<in\_R1.fastq> <in\_R2.fastq>

Note: if you have different bed file for target region then copy the config file in your directory and change the target files. Ploidy “4” is for Cancer Sample and you can change it to “2” for other samples in config file.

1. **Important output files:**

* \*realigned\_BQSR.bam (for visualization and alignment extraction; primary alignment file)
* \*\_dup\_metrics.dat (duplication metrics results from picard mark duplicates)
* CoverageMetrics.txt (Target coverage metrics)
* \*variants.raw.vcf (raw variant file from unifiedGenotyper; SNP and Indel combined)
* \*summary\_stats.txt (Aggregate statistics of quality trimming/filtering, duplication and Target capture)
* \*SNP\_INDEL\_COMBINED.vcf (Final variant file with all annotation; SNP and Indel combined)

1. **Whole Exome for paired end data Tumor-Normal**  (Regular, No Xenome, should be used for Patient PDX tumor-normal)
2. **Important tool used:**

* Filter\_trim.py script
* BWA mem
* Picard (Mark Duplicate)
* GATK indel realigner
* GATK BQSR
* Somatic sniper/GATK Indel detector
* snpEFF and snpSift for annotation; dbSNP and Cosmic annotation added are also present in VCF files.

1. **Invoking Pipeline:**

**module load civet/latest (Need once)**

civet\_run –o /opt/compsci/civet\_research\_pipelines/4.3.6/Human/Exome/WholeExomeTumorNormal/config\_file\_tumor\_normal\_Whole\_exome

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/Exome/WholeExomeTumorNormal/ WholeExomeTumorNormal.xml

<tumor\_R1.fastq> <tumor\_R2.fastq> <normal\_R1.fastq> <normal\_R2.fastq>

Note: if you have different bed file for target region then copy the config file in your directory and change the target files. Tumor and Normal files should be given in aforementioned order.

1. **Important output files:**

* \* normal\_realigned\_BQSR.bam and \*tumor\_realigned\_BQSR.bam (for visualization and alignment extraction; primary alignment file)
* \* normal\_dedup\_metrics.dat and \* tumor\_dedup\_metrics.dat (duplication metrics results from picard mark duplicates)
* NCoverageMetrics.txt and TCoverageMetrics.txt (Target coverage metrics)
* \* snp.vcf and \* indel.vcf (raw SNP and Indel files)
* \* \_normal\_summary\_stats.txt and \* \_tumor\_summary\_stats.txt (Aggregate statistics of quality trimming/filtering, duplication and Target capture)
* \* SNP\_INDEL\_COMBINED.vcf (Final variant file with all annotation; SNP and Indel combined)

1. **CTP for paired end data Tumor-Normal**  (Regular, No Xenome, should be used for Patient PDX tumor-normal CTP)
2. **Important tool used:**

* Filter\_trim.py script
* BWA mem
* Picard (Mark Duplicate)
* GATK indel realigner
* GATK BQSR
* Somatic sniper/GATK Indel detector
* snpEFF and snpSift for annotation; dbSNP and Cosmic annotation added are also present in VCF files.

1. **Invoking Pipeline:**

**module load civet/latest (Need once)**

civet\_run –o /opt/compsci/civet\_research\_pipelines/4.3.6/Human/Exome/CCPTumorNormal/config\_file\_tumor\_normal\_Whole\_exome /opt/compsci/civet\_research\_pipelines/4.3.6/Human/Exome/CCPTumorNormal/ CCPTumorNormal.xml

<tumor\_R1.fastq> <tumor\_R2.fastq> <normal\_R1.fastq> <normal\_R2.fastq>

Note: if you have different bed file for target region then copy the config file in your directory and change the target files. Tumor and Normal files should be given in aforementioned order.

1. **Important output files:**

* \* normal\_realigned\_BQSR.bam and \*tumor\_realigned\_BQSR.bam (for visualization and alignment extraction; primary alignment file)
* \* normal\_dedup\_metrics.dat and \* tumor\_dedup\_metrics.dat (duplication metrics results from picard mark duplicates)
* NCoverageMetrics.txt and TCoverageMetrics.txt (Target coverage metrics)
* \* snp.vcf and \* indel.vcf (raw SNP and Indel files)
* \* \_normal\_summary\_stats.txt and \* \_tumor\_summary\_stats.txt (Aggregate statistics of quality trimming/filtering, duplication and Target capture)
* \* SNP\_INDEL\_COMBINED.vcf (Final variant file with all annotation; SNP and Indel combined)

1. **PDX CTP for paired end data (Only for PDX; consists of Xenome Step; Same as CLIA except for Xenome step and coverage is calculated based all CTP genes)**
2. **Important tool used:**

* Xenome
* Filter\_trim.py script
* BWA mem
* Picard (Mark Duplicate)
* GATK indel realigner
* GATK BQSR
* GATK Unified Genotyper (Default ploidy = 4)
* Pindel and CNV detection
* snpEFF and snpSift for annotation;

1. **Invoking Pipeline:**

**module load civet/latest (Need once)**

civet\_run /data/anuj/version\_3\_civet\_pipelines/Pipelines/Human/Xenome/Exome/SingleSampleCCP/XenomeSingleSampleCCP.xml

<in\_R1.fastq> <in\_R2.fastq>

1. **Important output files:**

* \*realigned\_BQSR.bam (for visualization and alignment extraction; primary alignment file)
* \*variants.DPfiltered.vcf
* \*microIndels.DPfiltered.vcf
* \*variants\_microIndels.DPfiltered.AFge0.05.tab
* \*variants\_microIndels.DPfiltered.Annotated.tab
* \*variants\_microIndels.DPfiltered.clinicalTargets.AFge0.05.tab
* \*variants\_microIndels.DPfiltered.clinicalTargets.vcf
* \*\_dup\_metrics.dat (duplication metrics results from picard mark duplicates)
* CoverageMetrics.txt (Target coverage metrics)

1. **PDX Whole exome for paired end data (Only for PDX; consists of Xenome Step)**
2. **Important tool used:**

* Xenome
* Filter\_trim.py script
* BWA mem
* Picard (Mark Duplicate)
* GATK indel realigner
* GATK BQSR
* GATK Unified Genotyper (Default ploidy = 4)
* snpEFF and snpSift for annotation; dbSNP and Cosmic annotation added are also present in VCF files.

1. **Invoking Pipeline:**

**module load civet/latest (Need once)**

civet\_run –o /opt/compsci/civet\_research\_pipelines/4.3.6/Human/Xenome/Exome/WholeExome/config\_file\_Whole\_Exome\_Single\_Sample

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/Xenome/Exome/WholeExome/XenomeWholeExomeSingleSample.xml

<in\_R1.fastq> <in\_R2.fastq>

1. **Important output files:**

* \* realigned\_BQSR.bam (for visualization and alignment extraction; primary alignment file)
* \* dup\_metrics.dat (duplication metrics results from picard mark duplicates)
* CoverageMetrics.txt (Target coverage metrics)
* \* variants.raw.vcf (raw variant file from unifiedGenotyper; SNP and Indel combined)
* \* summary\_stats.txt (Aggregate statistics of quality trimming/filtering, duplication and Target capture)
* \*SNP\_INDEL\_COMBINED.vcf (Final variant file with all annotation; SNP and Indel combined)

**ALL RNA PIPELINES (Regular, PDX)**

1. **RNA for paired end data (Regular RNA Analysis; No Xenome)**
2. **Invoking Pipeline:**

**module load civet/latest (Need once)**

civet\_run –o /opt/compsci/civet\_research\_pipelines/4.3.6/Human/RNA/RNA\_Expression\_Estimation\_Single\_Sample\_PE/ config\_file\_RSEM\_RNA\_SEQ\_Single\_Sample\_Expression\_Estimation

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/RNA/RNA\_Expression\_Estimation\_Single\_Sample\_PE/RNASeqSingleSamplePE.xml

<in\_R1.fastq> <in\_R2.fastq>

1. **Important tool used:**

* Filter\_trim.py script
* RSEM
* Normalization script
* Bamtools

1. **Important output files:**

* \* genome\_bam\_with\_read\_group\_reorder.bam (for visualization and alignment extraction; primary alignment file)
* \* genes.results (RSEM raw gene expression estimate)
* \* genes.results.withGeneName (Corresponding Gene name added)
* \* genes.results.Normalized (expected count Normalized by dividing upper quantile of non-zero counts and then scaling upto 1000 )
* \* isoforms.results (RSEM raw isoform expression estimate)
* \* isoforms.results.withGeneName (Corresponding Gene name added)
* \* isoforms.results.Normalized (expected count Normalized by dividing upper quantile of non-zero counts and then scaling upto 300 )
* \* \_summary\_stats.txt (Summary statistics of Regular QC and alignment)
* \* TCGA.genome.sorted.bam (Alignment generated from TCGA genome and annotation file)
* \* TCGA.genes.results (RSEM raw gene expression estimate)
* \* TCGA.genes.results.Normalized (expected count Normalized by dividing upper quantile of non-zero counts and then scaling upto 1000 )
* \* TCGA.isoforms.results (Corresponding Gene name added)
* \* TCGA.isoforms.results.Normalized (expected count Normalized by dividing upper quantile of non-zero counts and then scaling upto 300 )

**Note: If sample is not related to Cancer/PDX then user can ignore TCGA part of the results.**

1. **RNA for Single end data (Regular RNA Analysis; No Xenome)**
2. **Invoking Pipeline:**

**module load civet/latest (Need once)**

civet\_run –o /opt/compsci/civet\_research\_pipelines/4.3.6/Human/RNA/RNA\_Expression\_Estimation\_Single\_Sample\_SE/ config\_file\_RSEM\_RNA\_SEQ\_Single\_Sample\_Expression\_Estimation

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/RNA/RNA\_Expression\_Estimation\_Single\_Sample\_SE/RNASeqSingleSampleSE.xml

<in\_R1.fastq>

1. **Important tool used:**

* Filter\_trim.py script
* RSEM
* Normalization script
* Bamtools

1. **Important output files:**

* \* genome\_bam\_with\_read\_group\_reorder.bam (for visualization and alignment extraction; primary alignment file)
* \* genes.results (RSEM raw gene expression estimate)
* \* genes.results.withGeneName (Corresponding Gene name added)
* \* genes.results.Normalized (expected counts Normalized by dividing upper quantile of non-zero count and then scaling upto 1000 )
* \* isoforms.results (RSEM raw isoform expression estimate)
* \* isoforms.results.withGeneName (Corresponding Gene name added)
* \* isoforms.results.Normalized (expected count Normalized by dividing upper quantile of non-zero counts and then scaling upto 300 )
* \* \_ summary\_stats.txt (Summary statistics of Regular QC and alignment)
* \* TCGA.genome.sorted.bam (Alignment generated from TCGA genome and annotation file)
* \* TCGA.genes.results (RSEM raw gene expression estimate)
* \* TCGA.genes.results.Normalized (expected count Normalized by dividing upper quantile of non-zero counts and then scaling upto 1000 )
* \* TCGA.isoforms.results (Corresponding Gene name added)
* \* TCGA.isoforms.results.Normalized (expected count Normalized by dividing upper quantile of non-zero counts and then scaling upto 300 )

**Note: If sample is not related to Cancer/PDX then user can ignore TCGA part of the results.**

1. **PDX RNA for Paired end data (Only for PDX; consists of Xenome Step)**
2. **Invoking Pipeline:**

**module load civet/latest (Need once)**

civet\_run –o /opt/compsci/civet\_research\_pipelines/4.3.6/Human/Xenome/RNA/RNA\_Expression\_Estimation\_Single\_Sample\_PE/

config\_file\_RSEM\_RNA\_SEQ\_Single\_Sample\_Expression\_Estimation

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/Xenome/RNA/RNA\_Expression\_Estimation\_Single\_Sample\_PE/XenomeRnaSeqSingleSamplePE.xml

<in\_R1.fastq> <in\_R2.fastq>

1. **Important tool used:**

* Xenome
* Filter\_trim.py script
* RSEM
* Normalization script
* Bamtools

1. **Important output files:**

* \* genome\_bam\_with\_read\_group\_reorder.bam (for visualization and alignment extraction; primary alignment file)
* \* genes.results (RSEM raw gene expression estimate)
* \* genes.results.withGeneName (Corresponding Gene name added)
* \* genes.results.Normalized (expected count Normalized by dividing upper quantile of non-zero counts and then scaling upto 1000 )
* \* isoforms.results (RSEM raw isoform expression estimate)
* \* isoforms.results.withGeneName (Corresponding Gene name added)
* \* isoforms.results.Normalized (expected counts Normalized by dividing upper quantile of non-zero count and then scaling upto 300 )
* \* \_ summary\_stats.txt (Summary statistics of Regular QC and alignment)
* \* TCGA.genome.sorted.bam (Alignment generated from TCGA genome and annotation file)
* \* TCGA.genes.results (RSEM raw gene expression estimate)
* \* TCGA.genes.results.Normalized (expected counts Normalized by dividing upper quantile of non-zero count and then scaling upto 1000 )
* \* TCGA.isoforms.results (Corresponding Gene name added)
* \* TCGA.isoforms.results.Normalized (expected counts Normalized by dividing upper quantile of non-zero count and then scaling upto 300 )

1. **PDX RNA for Single end data (Only for PDX; consists of Xenome Step)**
2. **Invoking Pipeline:**

**module load civet/latest (Need once)**

civet\_run –o /opt/compsci/civet\_research\_pipelines/4.3.6/Human/Xenome/RNA/RNA\_Expression\_Estimation\_Single\_Sample\_SE/ config\_file\_RSEM\_RNA\_SEQ\_Single\_Sample\_Expression\_Estimation

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/Xenome/RNA/RNA\_Expression\_Estimation\_Single\_Sample\_SE/XenomeRnaSeqSingleSampleSE.xml <in\_R1.fastq>

1. **Important tool used:**

* Xenome
* Filter\_trim.py script
* RSEM
* Normalization script
* Bamtools

1. **Important output files:**

* \* genome\_bam\_with\_read\_group\_reorder.bam (for visualization and alignment extraction; primary alignment file)
* \* genes.results (RSEM raw gene expression estimate)
* \* genes.results.withGeneName (Corresponding Gene name added)
* \* genes.results.Normalized (expected counts Normalized by dividing upper quantile of non-zero count and then scaling upto 1000 )
* \* isoforms.results (RSEM raw isoform expression estimate)
* \* isoforms.results.withGeneName (Corresponding Gene name added)
* \* isoforms.results.Normalized (expected counts Normalized by dividing upper quantile of non-zero count and then scaling upto 300 )
* \* \_ summary\_stats.txt (Summary statistics of Regular QC and alignment)
* \* TCGA.genome.sorted.bam (Alignment generated from TCGA genome and annotation file)
* \* TCGA.genes.results (RSEM raw gene expression estimate)
* \* TCGA.genes.results.Normalized (expected counts Normalized by dividing upper quantile of non-zero count and then scaling upto 1000 )
* \* TCGA.isoforms.results (Corresponding Gene name added)
* \* TCGA.isoforms.results.Normalized (expected counts Normalized by dividing upper quantile of non-zero count and then scaling upto 300 )

1. **Tophat based RNA for Paired end data (**Regular RNA Analysis; No Xenome ; Novel Isoform discovery, Variant calling using RNA-Seq Best Practices**)**
2. **Invoking Pipeline:**

**module load civet/latest (Need once)**

civet\_run –o /opt/compsci/civet\_research\_pipelines/4.3.6/Human/RNA/TophatBased\_RNA\_Novel\_isoform\_detection\_PE/config\_file\_RNA\_SEQ\_Single\_Sample\_Expression\_Estimation

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/RNA/TophatBased\_RNA\_Novel\_isoform\_detection\_PE/Tophat\_HumanRNASeqSingleSamplePE.xml <in\_R1.fastq> <in\_R2.fastq>

1. **Important tool used:**

* Filter\_trim.py script
* Picard (Mark Duplicate)
* Tophat
* Cufflinks
* GATK Haplotype caller
* snpEFF and snpSift for annotation
* **Important output files:**
* aln\_paired\_out/ accepted\_hits.bam (Tophat alignment file)
* cufflinks\_out/genes.fpkm\_tracking isoforms.fpkm\_tracking skipped.gtf transcripts.gtf (Cufflinks output files)
* \*\_summary\_stats.txt (Summary statistics of regular QC and alignment)
* \_HT\_seq.count (count of reads mapped to known genes)
* \*dup\_metrics.dat (duplication metrics results from picard mark duplicates)
* \*split.bam (Modified for GATK variant calling; use this if visualizing variant in IGV)
* \*variants\_raw.vcf (raw variant file)
* \*SNP\_INDEL\_COMBINED.vcf (Final variant file with all annotation; SNP and Indel combined))

**ALL Whole Genome PIPELINES (Regular, PDX)**

1. **Whole genome for Paired end data (Regular data analysis; No Xenome)**
2. **Invoking Pipeline:**

**module load civet/latest (Need once)**

civet\_run –o/opt/compsci/civet\_research\_pipelines/4.3.6/Human/WholeGenome/WholeGenomeSingleSample/config\_file\_whole\_genome\_single\_sample

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/WholeGenome/WholeGenomeSingleSample/ WholeGenomeSingleSample.xml

<in\_R1.fastq> <in\_R2.fastq>

1. **Important tool used:**

* Filter\_trim.py script
* BWA mem
* Picard (Mark Duplicate)
* GATK indel realigner
* GATK BQSR
* GATK Unified Genotyper (Default ploidy = 4)
* snpEFF and snpSift for annotation; dbSNP and Cosmic annotation added are also present in VCF files.

1. **Important output files:**

* \*\_realigned\_BQSR.bam (for visualization and alignment extraction; primary alignment file)
* \*\_dup\_metrics.dat (duplication metrics results from picard mark duplicates)
* \*\_variants\_raw\_merge.vcf (raw variant file from unifiedGenotyper; SNP and Indel combined)
* \*\_summary\_stats.txt (Aggregate statistics of quality trimming/filtering, duplication and alignment metrics)
* \*\_SNP\_INDEL\_COMBINED.vcf (Final variant file with all annotation; SNP and Indel combined)

**Note: Ploidy “4” is for Cancer Sample and you can change it to “2” for other samples in config file.**

1. **Whole genome for Paired end data Tumor-Normal (**Regular data analysis; No Xenome; should be used for Patient PDX tumor-normal**)**
2. **Invoking Pipeline:**

**module load civet/latest (Need once)**

civet\_run –o /opt/compsci/civet\_research\_pipelines/4.3.6/Human/WholeGenome/WholeGenomeTumorNormal/ config\_file\_tumor\_normal\_Whole\_genome

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/WholeGenome/WholeGenomeTumorNormal/ WholeGenomeTumorNormal.xml

<tumor\_R1.fastq> <tumor\_R2.fastq> <normal\_R1.fastq> <normal\_R2.fastq>

Note: Tumor and Normal files should be given in same order.

1. **Important tool used:**

* Filter\_trim.py script
* BWA mem
* Picard (Mark Duplicate)
* GATK indel realigner
* GATK BQSR
* Somatic sniper/GATK Indel detector
* snpEFF and snpSift for annotation; dbSNP and Cosmic annotation added are also present in VCF files.

1. **Important output files:**

* \* \_normal\_realigned\_BQSR.bam and \*\_tumor\_realigned\_BQSR.bam (for visualization and alignment extraction; primary alignment file)
* \* \_normal\_dedup\_metrics.dat and \* \_tumor\_dedup\_metrics.dat (duplication metrics results from picard mark duplicates)
* \* \_Normal\_AlignmentMetrics.txt and \* \_Tumor\_AlignmentMetrics.txt (Alignment Coverage metrics)
* \* snp.vcf and \* indel.vcf (raw SNP and Indel files)
* \* \_normal\_summary\_stats.txt and \* \_Tumor\_AlignmentMetrics.txt (Aggregate statistics of quality trimming/filtering, duplication and Alignment metrics)
* \* SNP\_INDEL\_COMBINED.vcf (Final variant file with all annotation; SNP and Indel combined)

1. **PDX Whole genome for Paired end data (Only for PDX; consists of Xenome Step)**
2. **Invoking Pipeline:**

**module load civet/latest (Need once)**

civet\_run –o

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/Xenome/WholeGenome/WholeGenomeSingleSample/config\_file\_whole\_genome\_single\_sample

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/Xenome/WholeGenome/WholeGenomeSingleSample/XenomeWholeGenomeSingleSample.xml

<in\_R1.fastq> <in\_R2.fastq>

1. **Important tool used:**

* Xenome
* Filter\_trim.py script
* BWA mem
* Picard (Mark Duplicate)
* GATK indel realigner
* GATK BQSR
* GATK Unified Genotyper (Default ploidy = 4)
* snpEFF and snpSift for annotation; dbSNP and Cosmic annotation added are also present in VCF files.

1. **Important output files:**

* \* \_realigned\_BQSR.bam (for visualization and alignment extraction; primary alignment file)
* \* \_dup\_metrics.dat (duplication metrics results from picard mark duplicates)
* \* variants\_raw\_merge.vcf (raw variant file from unifiedGenotyper; SNP and Indel combined)
* \*\_ summary\_stats.txt (Aggregate statistics of quality trimming/filtering, duplication and alignment metrics)
* \*\_ SNP\_INDEL\_COMBINED.vcf (Final variant file with all annotation; SNP and Indel combined)

**ChIP-seq PIPELINE**

1. **ChIP-seq for Paired end data Experiment-Control (Regular data analysis; No Xenome)**
2. **Invoking Pipeline:**

**module load civet/latest (Need once)**

civet\_run –o/opt/compsci/civet\_research\_pipelines/4.3.6/Human/ChIP-seq/ChIPseqPairedEnd/config\_file\_ChIP

/opt/compsci/civet\_research\_pipelines/4.3.6/Human/ChIP-seq/ChIPseqPairedEnd/experimentalcontrolChIP.xml <experimental\_R1.fastq> <experimental\_R2.fastq> <control\_R1.fastq> <control\_R2.fastq>

1. **Important tool used:**

* Filter\_trim.py script
* BWA mem
* Picard
* MACS2
* GATK BQSR
* ChIPpeakAnno for annotation

1. **Important output files:**
2. \* \_trim.bam (a pair of files for visualization and alignment extraction; primary alignment files)
3. \* \_e\_hist.pdf and \* \_c\_hist.pdf (coverage and insert size metrics)
4. \*summary\_stats.txt (Aggregate statistics of quality trimming/filtering, duplication and alignment metrics)
5. \*\_annotations..txt (Final file with all annotation)