**Analyzing Mouse Sequence Data Using Civet**

**Prepared By**

**Anuj Srivastava**

**Few things to note prior to run these civet pipelines:**

1. These pipelines are for mouse Exome/WholeGenome/RNA-Seq data analysis
2. The reference genome for Exome/WholeGenome is **mm10**
3. The reference genome and annotation of RSEM based RNA-Seq is **Mus\_musculus\_GRCm38\_74\_dna\_primary.fa and Mus\_musculus\_final.GRCm38.74.gtf** (downloaded from ENSEMBL and only chr 1-19,X-Y,MT is used in analysis)
4. The Tophat pipeline uses UCSC **mm10** ref and annotation file and does the **novel isoform** prediction and also variant calling using GATK RNA-Seq best practices

**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.**

**Note:** All these pipelines will halt if more than 50 % of reads filtered out during quality filtering/trimming step.

All Pipelines are location on Cadillac:

/opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/

1. **Whole Exome for paired end data** and when closest strain information **not** available for training at BQSR Step :
2. **Important tool used:**

* Filter\_trim.py script
* BWA mem
* Picard (Mark Duplicate)
* GATK indel realigner
* GATK haplotype caller
* VCF-tool for dbSNP annotation
* snpEFF

1. **Invoking Pipeline:**

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

civet\_run –o /opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/Exome/WholeExomeNoClose/**config\_file\_Whole\_Exome\_MMR** /opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/Exome/WholeExomeNoClose/**Mouse\_WholeExome\_NoClose.xml**

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

Note: if you have different bed file for target region then copy and change the target files.

1. **Important output files:**

* \* \_realigned\_dedup.bam (for visualization and alignment extraction; primary alignment file)
* \*\_dup\_metrics.dat (duplication metrics results from picard mark duplicates)
* \*\_variants\_filtered\_highestsnpEff.vcf (final variant file to be uploaded in the database)
* \*\_summary\_stats.txt (Aggregate statistics of quality trimming/filtering and Target capture; could be uploaded in MMR database)

2. **Whole Exome for paired end data** and when closest strain information available for training at BQSR step:

1. **Important tool used:**

* Filter\_trim.py script
* BWA mem
* Picard (Mark Duplicate)
* GATK indel realigner
* GATK BQSR
* GATK haplotype caller
* VCF-tool for dbSNP annotation
* snpEFF

1. **Invoking Pipeline:**

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

civet\_run –o

/opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/Exome/WholeExomeWithClose/config\_file\_Whole\_Exome\_MMR

/opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/Exome/WholeExomeWithClose/Mouse\_WholeExomeWithClose.xml

<in\_R1.fastq> <in\_R2.fastq> <snp\_file> <indel\_file>

Note: if you have different bed file for target region then copy and change the target files.

1. **Training files are in:**

* /data/shared/mmr/TrainingFiles/combined\_Files
* /data/shared/mmr/TrainingFiles/individual\_Files/

Combined files required when SNP/Indels from multiple strain required for training and individual file when only when single strain is needed (Investigator mostly GRS will provide this information).

**Example:**

In the example below (C57BL6NJ and C3HHeJ) is used for the training.

* C57BL6NJ\_C3HHeJ.mgp.v3.snps.rsIDdbSNPv137.reorder.GTcleaned.vcf
* C57BL6NJ\_C3HHeJ.mgp.v3.indels.rsIDdbSNPv137.reorder.GTcleaned.vcf

civet\_run /opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/Exome/WholeExomeWithClose/**config\_file\_Whole\_Exome\_MMR**

/opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/Exome/WholeExomeWithClose/**Mouse\_WholeExomeWithClose.xml**

<in\_R1.fastq> <in\_R2.fastq> /data/shared/mmr/TrainingFiles/combined\_Files/ C57BL6NJ\_C3HHeJ.mgp.v3.snps.rsIDdbSNPv137.reorder.GTcleaned.vcf

/data/shared/mmr/TrainingFiles/combined\_Files/ C57BL6NJ\_C3HHeJ.mgp.v3.indels.rsIDdbSNPv137.reorder.GTcleaned.vcf

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\_filtered\_highestsnpEff.vcf (final variant file to be uploaded in the database)
* \*\_summary\_stats.txt (Aggregate statistics of quality trimming/filtering and Target capture; could be uploaded in MMR database)

1. **Whole Genome for paired end data** and when closest strain information **not** available for training at BQSR step:
2. **Important tool used:**

* Filter\_trim.py script
* BWA mem
* Picard (Mark Duplicate)
* GATK indel realigner
* GATK haplotype caller
* VCF-tool for dbSNP annotation
* snpEFF

1. **Invoking Pipeline:**

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

civet\_run /opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/WholeGenome/WholeGenomeNoClose/Mouse\_WholeGenome\_NoClose.xml <in\_R1.fastq> <in\_R2.fastq>

1. **Important output files:**

* \* \_realigned\_dedup.bam (for visualization and alignment extraction; primary alignment file)
* \* \_dup\_metrics.dat (duplication metrics results from picard mark duplicates)
* \* \_variants\_filtered\_highestsnpEff.vcf (final variant file to be uploaded in the database)
* \*\_summary\_stats.txt (Aggregate statistics of quality trimming/filtering and Target capture; could be uploaded in MMR database)

1. Whole Genome for paired end data and when closest strain information available for training at BQSR step:
2. **Important tool used:**

* Filter\_trim.py script
* BWA mem
* Picard (Mark Duplicate)
* GATK indel realigner
* GATK BQSR
* GATK haplotype caller
* VCF-tool for dbSNP annotation
* snpEFF

1. **Invoking Pipeline:**

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

civet\_run –o

/opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/WholeGenome/WholeGenomeWithClose/Mouse\_WholeGenome\_WithClose.xml

<in\_R1.fastq> <in\_R2.fastq> <snp\_file> <indel\_file>

1. **Training files are in:**

* /data/shared/mmr/TrainingFiles/combined\_Files
* /data/shared/mmr/TrainingFiles/individual\_Files/

Combined files required when SNP/Indels from multiple strain required for training and individual file when only when single strain is needed (Investigator mostly GRS will provide this information).

**Example:**

In the example below (C57BL6NJ and C3HHeJ) is used for the training.

* C57BL6NJ\_C3HHeJ.mgp.v3.snps.rsIDdbSNPv137.reorder.GTcleaned.vcf
* C57BL6NJ\_C3HHeJ.mgp.v3.indels.rsIDdbSNPv137.reorder.GTcleaned.vcf

civet\_run

/opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/WholeGenome/WholeGenomeWithClose/Mouse\_WholeGenome\_WithClose.xml

<in\_R1.fastq> <in\_R2.fastq> /data/shared/mmr/TrainingFiles/combined\_Files/ C57BL6NJ\_C3HHeJ.mgp.v3.snps.rsIDdbSNPv137.reorder.GTcleaned.vcf

/data/shared/mmr/TrainingFiles/combined\_Files/ C57BL6NJ\_C3HHeJ.mgp.v3.indels.rsIDdbSNPv137.reorder.GTcleaned.vcf

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\_filtered\_highestsnpEff.vcf (final variant file to be uploaded in the database)
* \*\_summary\_stats.txt (Aggregate statistics of quality trimming/filtering and Target capture; could be uploaded in MMR database)

1. **RNA for paired end data** (RSEM based pipeline does the transcriptome alignment; not suitable if goal is to predict novel isoform from data; need to modify config file if strand specific data)
2. **Important tool used:**

* Filter\_trim.py script
* Picard (Read group and reorder)
* RSEM (alignment and expression)
* Bamtools (alignment metrics)
* GeneName\_and\_Normalization.pl (Add gene name and quartile Normalization)

1. **Invoking Pipeline:**

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

civet\_run –o

/opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/RNA/RNA\_Expression\_Estimation\_Single\_Sample\_PE/config\_file\_RSEM\_RNA\_SEQ\_Single\_Sample\_Expression\_Estimation

/opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/RNA/RNA\_Expression\_Estimation\_Single\_Sample\_PE/Mouse\_RNASeqSingleSamplePE.xml

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

Note: if you want to change some parameters in config file then copy and change the config file.

1. **Important output files:**

* **\*\_**genome\_bam\_with\_read\_group\_reorder.bam (final alignment file)
* genes.results.withGeneName (can be used for genebased DE)
* isoforms.results.withGeneName (can be used isoform based DE)
* genes.results.Normalized (expected count is divided by upper quartile of non-zero count and scaled upto 1000 like TCGA)
* isoforms.results.Normalized (expected count is divided by upper quartile of non-zero count and scaled upto 300 like TCGA)
* \*\_aln\_metrics.txt (alignment and insert size metrics)

1. **RNA for single end data** (RSEM based pipeline does the transcriptome alignment; not suitable if goal is to predict novel isoform from data; need to modify config file if strand specific data)
2. **Important tool used:**

* Filter\_trim.py script (modified for single end)
* Picard (Read group and reorder)
* RSEM (alignment and expression)
* Bamtools (alignment metrics)
* GeneName\_and\_Normalization.pl (Add gene name and quartile Normalization)

1. **Invoking Pipeline:**

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

civet\_run –o

/opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/RNA/RNA\_Expression\_Estimation\_Single\_Sample\_SE/config\_file\_RSEM\_RNA\_SEQ\_Single\_Sample\_Expression\_Estimation

/opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/RNA/RNA\_Expression\_Estimation\_Single\_Sample\_SE/Mouse\_RNASeqSingleSampleSE.xml

<in\_R1.fastq>

Note: if you want to change some parameters in config file then copy and change the config file.

1. **Important output files:**

* **\*\_**genome\_bam\_with\_read\_group\_reorder.bam (final alignment file)
* genes.results.withGeneName (can be used for genebased DE)
* isoforms.results.withGeneName (can be used isoform based DE)
* genes.results.Normalized (expected count is divided by upper quartile of non-zero count and scaled upto 1000 like TCGA)
* isoforms.results.Normalized (expected count is divided by upper quartile of non-zero count and scaled upto 300 like TCGA)
* \*\_aln\_metrics.txt (alignment metrics)

1. **RNA for paired end data** (Tophat based pipeline does the genome alignment; suitable if goal is also to predict novel isoform from data; This pipeline also does the variant calling from RNA-Seq data; not suitable for strand-specific RNA-SEQ)
2. **Important tool used:**

* Filter\_trim.py script
* Picard (Read group and reorder)
* Bamtools (alignment metrics)
* Tophat (alignment and expression)
* Cufflinks (transcriptome assembly, upper quantile normalization, multi-read correct and novel isoform assembly)
* HTSEQ (for raw count of read mapped to genes)
* Picard (Mark Duplicate)
* GATK haplotype caller
* VCF-tool for dbSNP annotation
* snpEFF

1. **Invoking Pipeline:**

civet\_run –o /opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/RNA/TophatBased\_RNA\_Novel\_isoform\_detection\_variant\_calling\_PE/config\_file\_RNA\_SEQ\_Single\_Sample\_Expression\_Estimation

/opt/compsci/civet\_research\_pipelines/3.0.0/Mouse/RNA/TophatBased\_RNA\_Novel\_isoform\_detection\_variant\_calling\_PE/Tophat\_MouseRNASeqSingleSamplePE.xml

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

Note: if you want to change some parameters in config file then copy and change the config file.

* aln\_paired\_out/\* accepted\_hits.bam (Tophat output)
* cufflinks\_out/\* (will have cufflinks results)
* \* \_aln\_metrics.txt (alignment metrics and insert size)
* \* \_HT\_seq.count (Raw count of reads mapped to genes; used for edgeR based DE)
* \* split.bam (Bam used for variant calling)
* \* \_dup\_metrics.dat (Duplicate metrics)
* filtered\_highestsnpEff.vcf (final variant file)

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

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

1. **Invoking Pipeline:**

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

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

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

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. \*aummary\_stats.txt (Aggregate statistics of quality trimming/filtering, duplication and alignment metrics)
5. \*\_annotations..txt (Final file with all annotation)