**CEGS ASE analysis plan**

**Project Leader: Justin Fear**

**Updated: Alison Gerken**

SNP and INDEL vcf files: mclab/svn\_lmm\_dros\_head\_data/mel\_cegs\_variant\_info\_vcf\_files\_v3

Design file for level 1 Filtering: mclab/svn\_lmm\_dros\_head\_data/mel\_cegs\_expression\_1st\_106\_F/design\_files/CEGS\_list\_68\_lines.txt

(1) Transfer transcriptome files from USC to UF HPC. Verify successful transfer using md5sums

(2) Transfer VCF files (split by line) from USC to mclab (and to UF HPC). Verify successful transfer using md5sums

(3) Merge SNP VCF files for each w1118-line pair

- flag heterozygotes and positions in both files or in one file and not the other.

(4) Level 1 Filtering: mclab/cegs\_ase\_paper/sas\_programs/Makefile\_level1\_filters.sas

- identify snps between w1118 and a line

- categorize snp classes for filtering

Level 1 filters:

Case 1: either position is called heterozygous

Case 2: both are reference – this won't exist with current vcf imported since I dropped non-hets

Case 3: both have SNP compared to fb551 and the base at the snp position is same in w118 and line

- will be some of these (e.g. w1118 and the line have the same snp base and this base is not the same as the reference)

Case 4: only ALT\_w1118 or ALT\_LINE is the reference

- w1118 and the line are not the same and one (w1118 or the line) is the reference base

Case 5: both w1118 and the line contain a snp from fb551 but each is different base. The vcf file output for masking defaults to the line snp

Case 6: SNP is found in the line vcf file only or in w1118 vcf file only and is not a het

Case 7: the snp is found in the line vcf only or in the w1118 vcf only and is not a het but it is called as the reference

\*\* for r101, only have cases 1, 3, 5 and 6

For further detail see mclab/cegs\_ase\_paper/documentation/old\_level1\_filter\_info.docx

(5) Create a masked references for each line:

- upload all masked VCF files to /scratch/lfs/mcintyre/cegs\_ase\_paper/ase\_pipeline\_output/ase\_SNP\_filtered\_vcf\_files

– w1118-line snps are masked in dmel fb551

– Indels ignored at this time (used later in Level 2 Filtering)

- Run only 10 samples at a time; design file: CEGS\_list\_68\_lines.txt

- Groups: 1-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-68

- Run for SNPs only in vcfFastaUpdate.py

- creates coordinates for genome and updates by SNPs

- Also build bowtie, samtools, and LAST index for each line

(6) Align RNA data to masked genome references

- use design file: CEGS\_68\_lines\_no\_tech.txt

Run in groups of 10 as previously established, but new design file so numbers will change.

Groups: 1-69, 70-129, 130-189, 190-249, 250-309, 310-369, 370-417

- use distinct.fq files for alignments

- use alignment\_functions.sh for bowtie\_se\_all and last\_se\_all

- covert sam files directly to bam files

– copy from protocols align all, 3 mismatches, bowtie and last

(7) Assess masked alignments – is coverage as expected? Other issues?

– Parse alignment logs

(8) Count masked RNA (SNPs and indels), create CSV file with counts for level 2 filtering

- Use design file: CEGS\_68\_lines\_no\_tech.txt

- count\_snps\_indels.py

- output: chrom, pos, ref, alt, totalCount, refCount, altCount, flagIndel

- Use same groups as in Step 6

(9) Move all RNA count files from HPC to McLab

\*\*\*Because these files are large and data intensive, run steps 5, 6, and 8 for only 1 group of 10 at a time, then after move RNA count CSVs to mclab, delete all PBS\_LOGS, alignments, references, etc.

\*\*\*DO NOT delete aln\_logs as these are needed for Step 7 for parsing.

Old plan: Keep this or this is now included in Step 8???

(9) Verify presence of w1118-line snps.

(10) Flag snps:

where both bases observed in alignments

where > 2 bases observed

where RNA coverage >= 10x for either w1118 or line

where gDNA alignments used for GATk snp calls have > = 5 reads supporting call in both w1118 and line

other possible filters? Ambiguity??

Level 2 filters:

– keep if both snp bases present in RNA

--keep if indel matches (within x number of bases)

– where we don't have both snp bases present in RNA: keep if there is 5X DNA coverage

\*\*note decision NOT to use the 10X RNA coverage flags as one of the filter

(11) Update fusion references using level 2 filtered vcf files

– insert snp bases into non-redundant fusion reference to create lineage specific references

– Confirm indel locations and update line specific references

(12) Align RNA data (no duplicates) to updated non-redundant fusion references.

copy from protocols – align unique allowing 3 mismatches using bowtie and last

(13) Assess masked alignments – is coverage as expected? Other issues?

– Parse alignment logs

(14) Assign reads to an allele

ase counts output from w1118 updated and to the Raleigh/Winters line-updated reference

move to mclab/pipeline\_output/ase\_counts\_fb551\_updated\_fusions/

(15) Check ase counts for systematic bias toward w1118

calculate proportion of of reads aligning to w1118 (SAM B) to total # of reads for each fusionID

by mv and rep

(15) Empirical Bayesian analysis

– Mated and virgin separate

– Decide whether should be fusion or gene level....

– if coverage isn't 'great', may want to collapse to gene level. If do this, need drop multigene fusions.....

– Flag a fusion/gene as analyzable if it has greater than 0 reads in any of the bioreps for either the tester or the varying allele.

– Run Luis bayesian R program with poisson gamma model and simulated q

– decide whether to use model with or without the “Both” term

– flag significant results for mated and virgin

– where are mated and virgin different?

– Flag (1) where each bio rep had greater or equal to 10 reads and (2) where the sum of the bio reps for both mated and virgin was greater or equal to 100.

– Summarize results