Love_etal_MolEcol_bioinformatic_synopsis

Bioinformatic synopsis, Love et al.

RNA-seq: Trimmomatic

```
trimmomatic \
    PE -threads 1 -phred33 \
    -summary $OUT_FASTQ1.sum \
    $IN_FASTQ1 $IN_FASTQ22 \
    $OUT_FASTQ1 $OUT_UNPAIRED1 \
    $OUT_FASTQ2 $OUT_UNPAIRED2 \
    LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:$EIGHTY_PERCENT_READLENGTH
```

RNA-seq: STAR

```
STAR \
    --genomeDir $INDEX_DIR \
    --readFilesIn $FASTQ1 $FASTQ2 \
    --readFilesCommand zcat \
    --outFileNamePrefix ./ \
    --outSAMtype None \
    --outSJfilterCountTotalMin 10 5 5 5 --outSJfilterCountUniqueMin 10 5 5 5

STAR \
    --genomeDir $FIRSTPASS_INDEX_DIR \
    --readFilesIn $FASTQ1 $FASTQ2 \
    --readFilesCommand zcat \
    --outFileNamePrefix ./ \
    --outSAMtype BAM SortedByCoordinate \
    --outFilterMultimapNmax 1 --outFilterMismatchNmax $TEN_PERCENT_PAIRLENGTH
```

RNA-seq: HTSeq

```
htseq-counts \
    --stranded=no --idattr=gene_id --type=exon --mode=union \
    --format=bam \
    $BAM \
    $ENSEMBL98_GTF \
    > $COUNTS_TSV
```

RNA-seq: GEDIT

```
python2 ./scripts/GEDIT.py -mix ./my_CPMs.tsv -ref ./ReferenceMatrices/Mouse/TabulaMurisReference.csv -
Rscript ./scripts/GLM_Decon.R \
    ./scripts/scratch/my_CPMs.tsv_TabulaMurisReference.csv_50_Entropy_0.0_ScaledMix.tsv \
```

```
    ./scripts/scratch/my_CPMs.tsv_TabulaMurisReference.csv_50_Entropy_0.0_ScaledRef.tsv \
    ./predictions/my_CPMs.tsv_TabulaMurisReference.csv_50_Entropy_0.0Predictions.tsv
```

RNA-seq: GATK

```
gatk \
    AddOrReplaceReadGroups \
    -I $BAM -O ./rg.bam \
    --RGID=$NAME --RGSM=$NAME --RGLB=$NAME --RGPL=Illumina \
    --RGPU=$LANE_ID

gatk \
    MarkDuplicates \
    --INPUT ./rg.bam \
    --OUTPUT ./rg-mdups.bam \
    --CREATE_INDEX true \
    --VALIDATION_STRINGENCY SILENT \
    --METRICS_FILE ./mdups.metrics \

gatk \
    SplitNCigarReads \
    --reference $GENOME_FASTA \
    --input ./rg-mdups.bam \
    --output ./rg-mdups-sncr.bam \

gatk \
    HaplotypeCaller \
    --reference $GENOME_FASTA \
    --input ./rg-mdups-sncr.bam \
    --output $GVCF \
    --emit-ref-confidence GVCF \
    --dont-use-soft-clipped-bases
```

Then,

```
gatk \
    GenomicsDBImport \
    --genomicsdb-workspace-path ./genomicsdb/ \
    --intervals $WHOLE_GENOME \
    -V $GVCF1 -V $GVCF2 -V $GVCF3 ...

gatk \
    GenotypeGVCFs \
    --reference $GENOME_FASTA \
    --variant gendb://./genomicsdb  \
    --output ./from_rnaseq.vcf.gz \
    --standard-min-confidence-threshold-for-calling 13
```

RNA-seq: Admixture

```
plink --bfile $BED_FILE --indep-pairwise 50 10 0.1 --allow-extra-chr --chr-set 38

plink --bfile $BED_FILE --extract plink.prune.in --make-bed --out prunedData --allow-extra-chr --chr-se
```

Then

```
for K in 1 2 3 4 5 6 7 8; do admixture --cv $bed $K -j8 | tee log${K}.out; done
```

RNA-seq: Fst

```
vcftools --gzvcf ${VCF} --weir-fst-pop pop1 --weir-fst-pop pop2 --out pop1_pop2_weir.fst
```

RNA-seq: PBE

```
#Cavalli-Sforza transformation
PBS_dataframe$T_CB <- -log(1-PBS_dataframe$C_B_Fst)
PBS_dataframe$T_CY <- -log(1-PBS_dataframe$C_Y_Fst)
PBS_dataframe$T_BY <- -log(1-PBS_dataframe$B_Y_Fst)

#Remove Inf value
PBS_dataframe <- subset(PBS_dataframe, !(PBS_dataframe$T_CB == "Inf"))
PBS_dataframe <- subset(PBS_dataframe, !(PBS_dataframe$T_CY== "Inf"))
PBS_dataframe <- subset(PBS_dataframe, !(PBS_dataframe$T_BY == "Inf"))


PBS_dataframe$C_PBS <- (PBS_dataframe$T_CB + PBS_dataframe$T_CY - PBS_dataframe$T_BY)/2

PBS_dataframe$C_PBS_0 <- pmax(PBS_dataframe$Cher_PBS, 0)

PBS$C_PBS_med <- median(PBS$C_PBS_0)

PBE = PBS(C) - ((T(BY)*PBS(C_median))/T(BY(median)))
```

RNA-seq: WGCNA

```
library(WGCNA)

mergingThresh = 0.25
net = blockwiseModules(Expr_data,corType="pearson",
                       maxBlockSize=nrow(log_wolfcounts_cpm),networkType="signed",power=15,minModuleSiz
                       mergeCutHeight=mergingThresh,numericLabels=TRUE,saveTOMs=TRUE,
                       pamRespectsDendro=FALSE,saveTOMFileBase="wolf_TOM")
moduleLabelsAutomatic=net$colors
```