Alternative Splicing Log

hisat2-build 'Copy of GCF_000146045.2_R64_genomic.fna' genome_index

```
Wrote 5399069 bytes to primary GFM file: genome_index.5.ht2
Wrote 3092708 bytes to secondary GFM file: genome_index.6.ht2
Re-opening _in5 and _in5 as input streams
Returning from HGFM constructor
Headers:
    len: 12157105
    gbwtLen: 12157106
    nodes: 12157106
    sz: 3039277
    gbwtSz: 3039277
    lineRate: 6
    offRate: 4
    offMask: 0xfffffff0
    ftabChars: 10
    eftabLen: 0
    eftabSz: 0
    ftabLen: 1048577
    ftabSz: 4194308
    offsLen: 759820
    offsSz: 3039280
    lineSz: 64
    sideSz: 64
    sideGbwtSz: 48
    sideGbwtLen: 192
    numSides: 63319
    numLines: 63319
    gbwtTotLen: 4052416
    gbwtTotSz: 4052416
    reverse: 0
    linearFM: Yes
Total time for call to driver() for forward index: 00:00:07
```

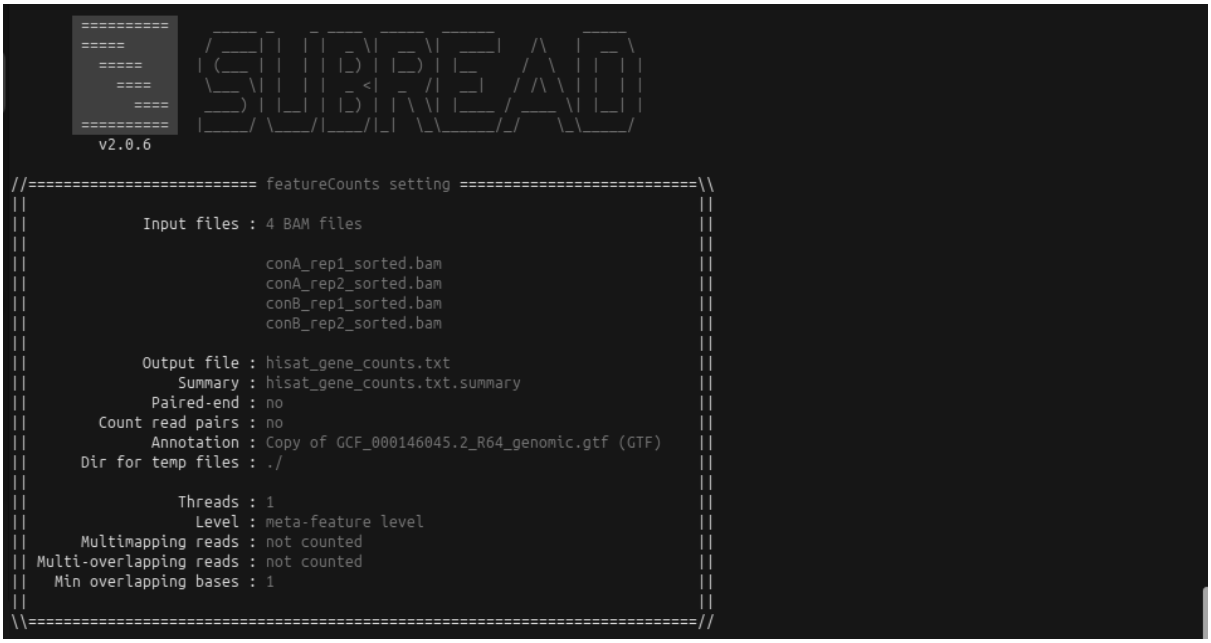
hisat2 -x genome_index -U conA_rep1.fq -S conA_rep1.sam
hisat2 -x genome_index -U conA_rep2.fq -S conA_rep2.sam
hisat2 -x genome_index -U conB_rep1.fq -S conB_rep1.sam
hisat2 -x genome_index -U conB_rep2.fq -S conB_rep2.sam

samtools view -bS conA_rep1.sam > conA_rep1.bam
samtools view -bS conA_rep2.sam > conA_rep2.bam
samtools view -bS conB_rep1.sam > conB_rep1.bam
samtools view -bS conB_rep2.sam > conB_rep2.bam

samtools sort conA_rep1.bam -o conA_rep1_sorted.bam
samtools sort conA_rep2.bam -o conA_rep2_sorted.bam
samtools sort conB_rep1.bam -o conB_rep1_sorted.bam
samtools sort conB_rep2.bam -o conB_rep2_sorted.bam

samtools index conA_rep1_sorted.bam
samtools index conA_rep2_sorted.bam
samtools index conB_rep1_sorted.bam
samtools index conB_rep2_sorted.bam

featureCounts -a 'Copy of GCF_000146045.2_R64_genomic.gtf' -o hisat_gene_counts.txt conA_rep1_sorted.bam conA_rep2_sorted.bam conB_rep1_sorted.bam conB_rep2_sorted.bam



R

1 Load required libraries

library(SGSeq)
library(pheatmap)

2 Set up BAM files and sample info

file_bam <- c("conA_rep1_sorted.bam", "conA_rep2_sorted.bam", "conB_rep1_sorted.bam", "conB_rep2_sorted.bam")
sample_name <- c("conA_rep1", "conA_rep2", "conB_rep1", "conB_rep2")
bam_info <- data.frame(sample_name, file_bam)

Get BAM metadata

Baminfo <- getBamInfo(bam_info)
Baminfo

3 Import transcript annotations and convert to features

tx <- importTranscripts("Copy of GCF_000146045.2_R64_genomic.gtf")
TxFeat <- convertToTxFeatures(tx)
sgfc <- analyzeFeatures(Baminfo, features = TxFeat)
sgfc_annot <- annotate(sgfc, TxFeat)

4 Inspect annotated object

colData(sgfc_annot)
rowRanges(sgfc_annot)
counts(sgfc_annot)[1:5,]

head(FPKM(sgfc_annot))

Save annotated counts
write.csv(counts(sgfc_annot), file = "sgfc_annot_counts.csv", row.names = TRUE)

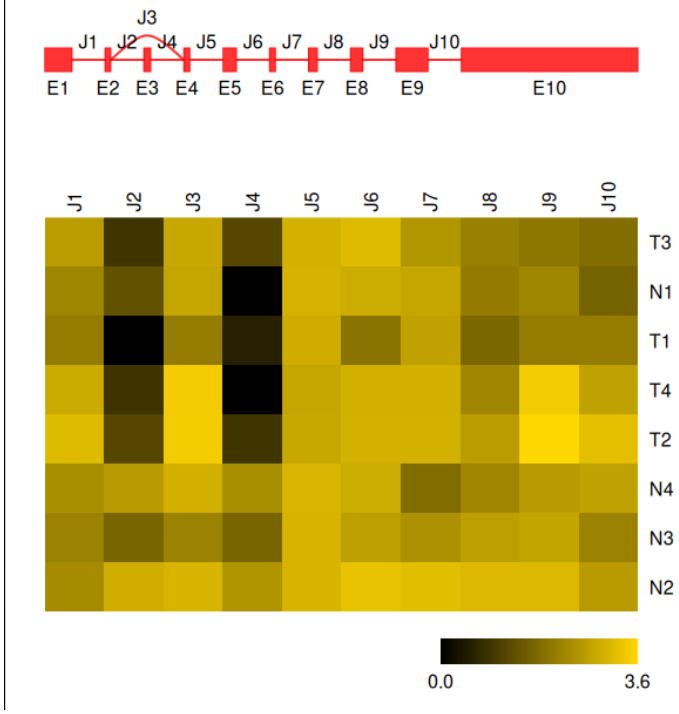
5 Plot a gene from annotated object (if available)
plotFeatures(sgfc_annot, geneID = unique(rowRanges(sgfc_annot)$geneID)[1], color_novel = "red")
plotFeatures(sgfc_annot, geneID = unique(rowRanges(sgfc_annot)$geneID)[1], color_novel = "green")

6 Generate heatmap of top splicing features
expr_mat <- assay(sgfc_annot)
top_feats <- head(order(rowMeans(expr_mat), decreasing = TRUE), 50)
pheatmap(expr_mat[top_feats,],
cluster_rows = TRUE, cluster_cols = TRUE,
fontsize_row = 6, fontsize_col = 10,
main = "Top Splicing Features")

```
    anyMissing, rowMedians

> library(pheatmap)
> file_bam <- c("conA_rep1_sorted.bam", "conA_rep2_sorted.bam", "conB_rep1_sorted.bam", "conB_rep2_sorted.bam")
> sample_name <- c("conA_rep1", "conA_rep2", "conB_rep1", "conB_rep2")
> bam_info <- data.frame(sample_name, file_bam)
> Baminfo <- getBamInfo(bam_info)
conA_rep1 complete.
conA_rep2 complete.
conB_rep1 complete.
conB_rep2 complete.
>  Baminfo
  sample_name             file_bam paired_end read_length frag_length lib_size
1   conA_rep1 conA_rep1_sorted.bam      FALSE         101          NA   768414
2   conA_rep2 conA_rep2_sorted.bam      FALSE         101          NA   867936
3   conB_rep1 conB_rep1_sorted.bam      FALSE         101          NA   673580
4   conB_rep2 conB_rep2_sorted.bam      FALSE         101          NA   961359
>  tx <- importTranscripts("Copy of GCF_000146045.2_R64_genomic.gtf")
> TxFeat <- convertToTxFeatures(tx)
> sgf_ucsc = convertToSGFeatures(TxFeat)
> sgfc_ucsc = analyzeFeatures(Baminfo, features = TxFeat)
Process features...
Obtain counts...
conA_rep1 complete.
conA_rep2 complete.
conB_rep1 complete.
conB_rep2 complete.
> sgfc_annot = annotate(sgfc_pred, TxFeat)
Warning messages:
1: In .merge_two_Seqinfo_objects(x, y) :
  The 2 combined objects have no sequence levels in common. (Use
  suppressWarnings() to suppress this warning.)
2: In .merge_two_Seqinfo_objects(x, y) :
  The 2 combined objects have no sequence levels in common. (Use
  suppressWarnings() to suppress this warning.)
3: In .merge_two_Seqinfo_objects(x, y) :
  The 2 combined objects have no sequence levels in common. (Use
  suppressWarnings() to suppress this warning.)
4: In .merge_two_Seqinfo_objects(x, y) :
  The 2 combined objects have no sequence levels in common. (Use
  suppressWarnings() to suppress this warning.)
> colData(sgfc_annot)
rowRanges(sgfc_annot)
counts(sgfc_annot)[1:5, ]
head(FPKM(sgfc_annot))
DataFrame with 8 rows and 6 columns
   sample_name    file_bam paired_end read_length frag_length  lib_size
   <character> <character>  <logical>   <numeric>   <numeric> <integer>
N1          N1      N1.bam       TRUE          75         293  12405197
N2          N2      N2.bam       TRUE          75         197  13090179
N3          N3      N3.bam       TRUE          75         206  14983084
N4          N4      N4.bam       TRUE          75         207  15794088
T1          T1      T1.bam       TRUE          75         284  14345976
T2          T2      T2.bam       TRUE          75         235  15464168
T3          T3      T3.bam       TRUE          75         259  15485954
T4          T4      T4.bam       TRUE          75         247  15808356
SGFeatures object with 38 ranges and 0 metadata columns:
       seqnames            ranges strand     type  splice5p  splice3p featureID
          <Rle>         <IRanges>  <Rle> <factor> <logical> <logical> <integer>
   [1]       16 87362930-87365116      -        E      TRUE     FALSE         1
   [2]       16          87365116      -        A      <NA>      <NA>         2
   [3]       16 87365116-87367492      -        J      <NA>      <NA>         3
   [4]       16          87367492      -        D      <NA>      <NA>         4
   [5]       16 87367492-87367892      -        E      TRUE      TRUE         5
   ...      ...               ...    ...      ...       ...       ...       ...
  [34]       16 87393901-87393972      -        E      TRUE      TRUE        34
  [35]       16          87393972      -        A      <NA>      <NA>        35
  [36]       16 87393972-87417011      -        J      <NA>      <NA>        36
  [37]       16          87417011      -        D      <NA>      <NA>        37
  [38]       16 87417011-87417348      -        E     FALSE      TRUE        38
```

```
> plotFeatures(sgfc_annot, geneID = unique(rowRanges(sgfc_annot)$geneID)[1], color_novel = "red")
> plotFeatures(sgfc_annot, geneID = unique(rowRanges(sgfc_annot)$geneID)[1], color_novel = "green")
>  expr_mat = assay(sgfc_annot)
top_feats = head(order(rowMeans(expr_mat), decreasing = TRUE), 50)
pheatmap(expr_mat[top_feats, ],
         cluster_rows = TRUE, cluster_cols = TRUE,
         fontsize_row = 6, fontsize_col = 10,
         main = "Top Splicing Features")
```


J3
J1 J2 J4 J5 J6 J7 J8 J9 J10
E1 E2 E3 E4 E5 E6 E7 E8 E9 E10
J1 J2 J3 J4 J5 J6 J7 J8 J9 J10
T3
N1
T1
T4
T2
N4
N3
N2
0.0
3.6