Transcript Assembly and Quantification (StringTie)

StringTie is a highly efficient assembler for RNA-Seq alignments using a novel network flow algorithm [8]. It can simultaneously assemble and quantify expression levels for the features of the transcriptome in a Ballgown readable format. StringTie's output can be processed by specialized software like Ballgown (Alyssa et al. (2014)), Cuffdiff (Cole et al. (2010)) or other programs (DESeq2 (Anders & Huber (2010)), edgeR (Robinson et al. (2010)), etc).

The input SAM(BAM) file must be sorted by reference position. Every spliced read alignment in the input must contain the tag `XS` to indicate the genomic strand that produced the RNA from which the read was sequenced. These requirements are met by running HISAT2 with `--dta` option and samtools.

Install StringTie

```
# Download and extract StringTie
$ wget http://ccb.jhu.edu/software/stringtie/dl/stringtie-
2.1.4.Linux_x86_64.tar.gz
$ tar xvfz stringtie-2.1.4.Linux_x86_64.tar.gz

# Append to PATH environment variable
$ export PATH=$PATH:/path/to/stringtie-2.1.4.Linux_x86_64

# Verify installation
$ stringtie --version
```

Run StringTie

Run with the downloaded gene annotation:

```
#!/bin/bash
stringTie1(){
  stringtie /path/to/homo_result/"$1"_sorted.bam -p 20 -o
/path/to/stringtie_homo/"$1".gtf -G /path/to/hg19_annotation.gff
}
export -f stringTie1

for s in SRR5978827 SRR5978828 SRR5978829 SRR5978834 SRR5978835 SRR5978836
SRR5978869 SRR5978870 SRR5978871 SRR5179446 SRR5179447 SRR5179448
do
stringTie1 ${s}
done
```

```
#!/bin/bash
stringTie1(){
  stringtie /path/to/mm10_result/"$1"_sorted.bam -p 20 -o
/path/to/stringtie_mm10/"$1".gtf -G /path/to/mm10_annotation.gff
}
export -f stringTie1

for s in SRR866997 SRR866998 SRR866999 SRR867000 SRR867001 SRR867002 SRR866991
SRR866992 SRR866993 SRR866994 SRR866995 SRR866996
do
stringTie1 ${s}
done
```

```
# Generate a non-redundant set of transcripts
$ cd /path/to/stringtie_homo/
$ stringtie --merge -G /path/to/hg19_annotation.gff -p 20 -o
homo_stringtie_merged.gtf homo_stringtie_list.txt

$ cd /path/to/stringtie_mm10/
$ stringtie --merge -G /path/to/mm10_annotation.gff -p 20 -o
mm10_stringtie_merged.gtf mm10_stringtie_list.txt
```

The text file contains all GTF files generated when assembling the read alignments.

```
SRR5978827.gtf
SRR5978828.gtf
......
```

Estimate transcript abundances and generate read coverage tables for Ballgown. Note that this is the only case where the `-G` option is not used with a reference annotation

```
#!/bin/bash
stringTie2(){
stringtie /path/to/homo_result/“$1”_sorted.bam -eB -p 20 -G
/path/to/stringtie_homo/homo_stringtie_merged.gtf -o
/path/to/stringtie_homo/“$1”.gtf
}
export -f stringTie2

for s in SRR5978827 SRR5978828 SRR5978829 SRR5978834 SRR5978835 SRR5978836
SRR5978869 SRR5978870 SRR5978871 SRR5179446 SRR5179447 SRR5179448
do
mkdir $s
cd $s
stringTie2 ${s}
done
```

```
#!/bin/bash
stringTie2(){
stringtie /path/to/mm10_result/“$1”_sorted.bam -eB -p 20 -G
/path/to/stringtie_mm10/mm10_stringtie_merged.gtf -o
/path/to/stringtie_mm10/“$1”.gtf
}
export -f stringTie2

for s in SRR866997 SRR866998 SRR866999 SRR867000 SRR867001 SRR867002 SRR866991
SRR866992 SRR866993 SRR866994 SRR866995 SRR866996
do
mkdir $s
cd $s
stringTie2 ${s}
done
```

Note:

Arguments and Options	**Description**
-G	Use the reference annotation file (in GTF or GFF3 format) to guide the assembly process
-e	Limits the processing of read alignments to only estimate and output the assembled transcripts matching the reference transcripts given with the `-G` option (requires `-G`, recommended for `-B/-b`)
-B	enables the output of *Ballgown* input table files (*.ctab) containing coverage data for the reference transcripts given with the `-G` option
-b <path>	Same as -B option, but these files will be created in the provided directory <path> instead of the directory specified by the `-o` option
-p <int>	Specify the number of processing threads (CPUs) to use for transcript assembly. The default is 1

Outputs

1. StringTie's primary GTF output ("SRR5978869.gtf") contains details of the transcripts that StringTie assembles from RNA-Seq data.
2. Ballgown input table files ((1) e2t.ctab, (2) e_data.ctab, (3) i2t.ctab, (4) i_data.ctab, and (5) t_data.ctab) contain coverage data for all transcripts.
3. Merged GTF ("homo_stringtie_merged.gtf") is a uniform set of transcripts for all samples.