Integrative Transcriptomics
Dr. K. Nieselt, Faculty of Science

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

EyeTN workshop **March 30 - April 1 2015**

Mapping of RNA-seq reads

During the following 2 practical sessions you will learn how you can use reads obtained from an RNA-seq experiment for a gene expression analysis. In the morning of the practical course we will learn how to map the reads against a reference using the mapper *STAR*. For this, we will be using paired-end RNA-seq data obtained from human T-cell experiments. Furthermore, because of the resulting large file, we will restrict our analysis to one specific chromosome. Due to the reason that these computations need a lot of time, we will only show you how to use these tools, but skip the calculations and later work with precomputed results.
After the mapping process we will take a look on two ways how we can obtain the data needed for gene expression analysis.

1. Mapping using STAR

Create an index with `STAR`

```
STAR --runMode genomeGenerate --genomeDir .  --genomeFastaFiles ref
```

Optional parameter: `--runThreadN N` (use N threads for lesser runtime)

Note: For very small genomes, the parameter `--genomeSAindexNbases` needs to be scaled down, with a typical value of `min(14, log2(GenomeLength)/2 - 1)`.

Map reads with `STAR`

```
STAR --genomeDir indexFolder --readFilesIn 1.fastq 2.fastq --genomeLoad "LoadAndKeep"
--outFileNamePrefix outputPrefix
```

Optional parameter: `--runThreadN N` (use N threads for lesser runtime)

The output is a sam file, called `outputPrefix.sam`, containing all mapped reads.

2. Filter Reads

First we will reduce the sam files by filtering only reads that mapped against the reference (these have the flag 4 (second entry in the alignment section of sam file)

```
samtools view -bS -F 4 fileName.sam > fileName.mappedonly.bam
```

After filtering the mapped reads, we sort the file

```
samtools sort fileName.mappedonly.bam fileName.sorted
```

Then we create an index for the sorted file

```
samtools index fileName.sorted.bam
```

If we want only reads that are specific for a certain chromosome, we filter again (here chromosome 19, but be careful which fasta entry is specific for a certain chromosome)

```
samtools view -h -b fileName.sorted.bam 19 > fileName.sorted.chr19.bam
```

Again we create an index for the new file

```
samtools index fileName.sorted.chr19.bam
```

3. **Generate Expression Levels using `htseq-count`**

   From the mapped reads we can count how many reads mapped to a certain gene.

   ```
   htseq-count -f bam -m intersection-nonempty -s no -a 3 -t exon -i gene_id
   file.sorted.chr19.bam annotation.gtf > file.HTSeq.tsv 2 > file.HTSeq.error.log
   ```

   Note: For a detailed description of the parameters use `htseq-count -h`

   On Wednesday we will use the results of htseq-count for differential expression analysis using DESeq.

4. **Generate Expression levels and normalising them using `Mayday SeaSight`**

   As an alternative for the previous counting we can also load the created *.bam files from `STAR` into `Mayday SeaSight`.

   - `Mayday` and `Mayday SeaSight` are available at our website (`http://it.informatik.uni-tuebingen.de/`)
   - You can either download Mayday or use the webstart (`javaws`)
   - Import the *.bam files into `Mayday SeaSight`.

   Due to the reason that the files are too big to be handled on your workstation, we will demonstrate how to load such files into `Mayday SeaSight`. However, after some preprocessing steps (which `Mayday SeaSight` does for us), we can obtain a normalized expression matrix for further analysis.

   On Wednesday we will use the results of SeaSight for differential expression analysis and clustering using Mayday.