EQP-QM DOCUMENTATION

INTRODUCTION

EQP-QM (Exon quantification pipeline – quantification module) is a RNA-seq quantification software package that takes SAM/BAM alignment files as input and produces gene, exon, and junction counts from the aligned reads. The contribution of a read to a gene, exon, or junction count is based on the read weight which is the inverse of the number of genome alignments of the read; for instance, a read that aligns ten times in the genome contributes 1/10 to a gene, exon, or junction it overlaps.

EQP-QM consists of the two main scripts *eqp-setup.sh* and *eqp-quantify.sh*. The script *eqp-setup.sh* is needed to create a number of auxilliary files which are then used by *eqp-quantify.sh* to generate the counts.

EQP SETUP

eqp-setup.sh requires a GTF file as input. The GTF file needs to contain the genome annotation for the genome that was used for the alignment of the Fastq files. It creates a number of auxiliary files for *eqp-quantify.sh* in a new directory denoted by <data directory> in the following and is called in the following way:

```
eqp-setup.sh <GTF file> <data directory>
```

eqp-setup.sh does not take any options.

GTF FILE

Only entries with feature type "exon" which contain a "gene_id" field are used by eqp-setup.sh (and, thus, for the quantification of genes,exons, and junctions).

Caveat: Note that the standard GTF file provided by UCSC contains the transcript id in the "gene_id" field (which equals the "transcript_id" field). In this case EQP-QM will generate counts for the transcripts; however, please note that these cannot be considered as transcript abundance estimates; in particular, one read can contribute to many transcripts. Ensembl GTF files work without problems.

Caveat: Please note that the chromosome names used by NCBI/UCSC (chr1, chr2, …) are different from the chromosome names used by Ensembl (1, 2, …). So it is not possible to use a genome downloaded from NCBI/UCSC for alignment and an Ensembl GTF file to provide the genome annotation. In general, please make sure that the identifiers in the genome Fasta file and the GTF file are consistent.

RUNNING EQP-QM

Once *eqp-setup.sh* is finished, EQP-QM can be invoked on a SAM/BAM file via:

```
eqp-quantify.sh -d <data directory> <output directory> <SAM/BAM file>
```

The SAM/BAM file should be created by aligning the reads against the reference genome with a (splice-aware) short-read aligner (e.g. HiSAT, STAR, or Tophat2).

This will create the files `<SAM/BAM file base>-gene.cnt`,`<SAM/BAM file base>-gene.cnt`, and `<SAM/BAM file base>-junction.cnt` in the directory `<output directory>` if `<SAM/BAM file base>` is the base name of `<SAM/BAM file>` (without extension `.sam` or `.bam`).

EQP-QUANTIFY.SH OPTIONS

A number of options can be supplied to *eqp-quantify.sh.*

USAGE

Usage:

```
 eqp-quantify.sh <options> -d <setup dir> <output dir> <SAM/BAM file>
```

where `<options>` is

```
  [-g] [-e] [-j] [-E <exon overlap>] [-J <junction overlap>]
  [-W <min read weight>] [-s <direction>] [--nosort] [--unambig]
  [--unweighted]
```

-d STRING	Use STRING as the directory that contains the auxilliary files (needs to be supplied)
-g	Compute gene counts (as default all three types of counts are computed)
-e	Compute exon counts (as default all three types of counts are computed)
-j	Compute junction counts (as default all three types of counts are computed)
-E INT	Minimal number of bp of that a read needs to overlap with an exon to be counted for it [5]
-J INT	Minimal number of bp that a read needs to overlap with both exons of an exon-exon junction to be counted [8]
-W DOUBLE	Minimal weight of a read; reads with a lower weight are disregarded [0.01]
-S STRING	process reads as strand-specific in direction STRING: either *forward* (fr) or *backward* (for rf)
--nosort	The alignment file is sorted by names; do not sort it again.
--unambig	Count only reads that can be assigned unambiguously to a single gene or exon when creating the gene or exon counts (not applicable to junctions)
--unweighted	Do not use read weights in the generation of counts