INPUT

FASTQ / FASTA

SAM / BAM

FILTER

Filter out all Ribosomal RNA reads by alignment to an rRNA database using BLAST.

Keep only reads which align to the numbered chromosomes, X or Y

♦ **This removes all reads from the mitochondrial genenome**

If paired-end, keep only when both forward and reverse map consistently.

♦ **For exon-intron-junction level, keep forward read only (this removes the bias introduced by variable fragment length distribution)**

CATEGORIZE

Split reads into **unique** and **multi-mappers**

♦ **Following steps performed separately for unique and multimappers**

Separate off all reads for features which are **high expressers** in any of the samples

♦ **These will be handled separately and put back in later**

gene-level level processing

- Keep reads which align to genes
  - If paired-end, keep reads when both forward and reverse map to a gene

♦ **If the data are strand specific then perform this step separately for sense and anti-sense signal**

exon/intron/junction level processing

Divide reads into categories

- **Reads which align to exactly *n* exons** for each *n* = 1,2,3,...,20
- **Reads which align to exactly *n* introns** for each *n* = 1,2,3,...,10
- **Reads which align to intergenic regions**

RESAMPLE

By random resampling (without replacement) equalize the number of reads in each of the categories above

♦ **Reads mapping to high expressers: resample according to their original proportions out of all gene/exon/intron mappers**

For each sample, concatenate back together all of the equalized files.

QUANTIFY

Use merged sam files to quantify features (genes,exons/introns/junctions)

Recall, if quantifying exons/introns then use forward reads only

Two spreadsheets for each feature:
1) MIN: use unique-mappers only
2) MAX: include multi-mappers

OUTPUT

- Normalized SAM/BAM
- Feature count spreadsheets
- Norm factors statistics
- Coverage/ Junction files