Pipeline

2 Fastq files: 2 conditions: 1% yeast
0.1% yeast

1. Data processing
    1. Mapping to PhiX genome
    2. Splitting libraries by barcodes
    3. Quality filtering of the reads; 30 minimum quality score to keep and 50% of bases must have the quality of 30
    4. Trimming out linker, echoP15 and adapters; leaving 25 nucleotides long CAGE tag
    5. Mapping each sample to rDNA; removing mapped reads
    6. Mapping each sample to the reference genome hg19
    7. Mapping each sample to the yeast reference genome

    OUTPUT: SAM file

2. SAM to BED via BAM

    Converting SAM files to BED format via BAM files

3. BAM to Clusters

    Obtaining tag clusters from the pooles samples for a given study

    1. Merging all BAM files into one pooled BAM file
    2. Sorting and indexing pooled BAM file
    3. Converting into BED format
    4. Splitting the BED file based on the strand
    5. Calculating CTSS
    6. Paraclu clustering to get the tag clusters (genomic regions)

4. BED Intersect

    Intersecting BED files derived from human with those derived from yeast in order to find out if there are some common regions and remove it.

5. BED Intersect

    Intersecting individual sample BED files with the tag clusters genomic regions (paraclu output) to obtain the CTSS counts

    OUTPUT: BED file for each sample contatining CTSS counts

    When pooled together, they form a CTSS expression matrix that can be used directly in edgeR