Why splitting by barcodes?
There is the possibility to process a large number of samples by multiplexing on a high-throughput instruments. Multiplexing enables large sample numbers to simultaneously sequenced during a single experiment. To accomplish this, individual “barcode” sequences are added to each sample so they can be differentiated during the analysis.
In our data the barcodes used have length 3 nt
BC1 CTT

BC2 GAT

BC3 ACG

BC4 ATC

What we used for align your reads?
We used Bowtie. Bowtie takes the quality score of each base into account and incorporates it into the scoring model when aligning, so you retain the most information by leaving in all reads.

Why Quality filtering?
It is important to filter out reads with low quality. In our study we leave reads that have a minimum quality score of 30 and at least the 50% of the bases of that reads have at least 30 as quality score

Why trimming out linkers?
Because we want to remove linkers and adapters in order to obtain reads with length 25 nt (this length is the normal one for CAGE)

Why mapping to rDNA?
We are interested in mRNA, so we want to be sure that we use data derived from mRNA. It is known that the most abundant RNA is the rRNA, so removing the reads that map on the rDNA can help to exclude some false mRNA.

What is Paraclu?
Paraclu finds clusters in data attached to sequences. It was first applied to transcription start counts in genome sequences, but itcould be applied to other things too.
Paraclu is intended to explore the data, imposing minimal prior assumptions, and letting the data speak for itself.

Why DESeq?
DESeq allows analysis of experiments with no biological replicates in one or even both of the conditions. In our project we don't have replicates for each condition (different yeast concentrations)

Why Normalization?
Normalization is required if we want to compare samples derived from different situations, in our case different concentrations of yeast spike-in.
Moreover, the normalization is important because have a strong impact on the inference of differential expressed genes.

Tags Per Million (TPM)
When we use data derived from CAGE, it is usual to use TPM instead of other kind of normalization methods, such as RPKM, bacause the final Tag Clusters have more or less the same

length.
TPM=x*10^6/sum(x)

Relative Log Expression (RLE)
It is the scaling factor method proposed by Anders and Huber (2010). We call it "relative log expression", as median library is calculated from the geometric mean of all columns and the median ratio of each sample to the median library is taken as the scale factor.