Single end or Paired end?

why choose paired end (standard libraries, alternative splicing)

- Paired end gives an idea of the size of the insert and the diectionality of the mapping to the sequence assembly algorithms. This aids in prediction of inversions, deletions and mutations inside the genome.

- High quality of mapping of reads in regions with repeat content, Single end reads are woefully unsuitable to accurately predict transcription in repeat containing regions.

First, it gives you a better "coverage" of your transcripts. In contrast to eukaryotes with splicing, you can be fairly sure that the piece in between the mapped read pair is present in your cDNA, so you can create a fake read covering the whole piece of cDNA.

Second, you get a better resolution of the 3'-ed of your transcript. That can help you to better define 3'-UTRs and novel ncRNAs where you cannot infer the likely end from a stop codon.

Third, it allows you to better define polycistronic mRNAs and thus operons.

Last but not least: Back when sequencing length was still an issue, you could get all the above from sequencing just 2x 20-25bp (usually more than enigh to properly map a read within a bacterial chromosome).

Oh, and if you do metatranscriptomics, paired end data will help you tremendously to assemble unknown transcripts.

Molecular indexing:

Most Next Generation Sequencing (NGS) library prep methods introduce sequence bias with the use of enzyme processing and fragmentation steps can introduce errors in the form of incorrect sequence and misrepresented copy number.

With molecular indexed libraries, each molecule is tagged with a molecular index randomly chosen from 10,000 combinations so that any two identical molecules become distinguishable (with odds of 10,000/1), and can be independently evaluated in later data analysis.

Analysis using molecular indexing provides an absolute, digital measure- ment of gene expression levels, irrespective of common amplification distortions observed in many RNA-Seq experiments (the language strongly resembles NanoString)

QC control

Adapters – to trim or not to trim

<http://genomebio.org/is-trimming-is-beneficial-in-rna-seq/>

journal.pone.0085024

I show that anything more than VERY gentle trimming is harmful to de novo assembly and transcriptome characterization. My findings seem to be in conflict with those presented by Giorgi and colleagues. I\’ll tell you up from that I think I\’m right, at least for the RNAseq  part of their paper.

fgene-05-00013.pdf

With regards to RNAseq, they show that percentage of reads mapped to the reference increases with  moderate trimming (red bars), then decreases with more aggressive >Q30 trimming..  Note that I don\’t think that better mapping is necessarily equivalent to better RNAseq results, but save that issue for later..

Surely, 99% mapping of a 1M read dataset is much worse than 80% mapping of a 100M read dataset.  This is basically what they show, that trimmings reduces the size of the dataset (blue bars), but increases the mapping rate (red bars)..

Resolving individual clones of molecules is critical for increasing se- quencing accuracy, measuring bias, PCR duplication rates and identifying mutations in complex sample types.

(NEXTflex qRNA-Seq)

Mappers:

F1000 genomes paper makes point that Salmon outperformes something (f.e. STAR+*featureCounts* )

Coverage:

Impact of sequencing depth and number of replicate samples on DE detection

A common challenge when designing RNA-seq experiment is to maximize the detection power of the study under limited budget or sample availability.

what is the desired sequence coverage for reliable detection of differential expression, and more broadly what is a detection power at a given coverage and number of replicates.

Second, given a limited sequencing budget is it preferable to maximize the sequence coverage or increase the number of replicate samples.

Finally, what is the impact of different sequencing depths and varying number of replicates on the performances of the DE methods.

We generated a series of down-sampled libraries where a subset of 50%, 40%, 30%, 20%, 10% and 5% reads were randomly sampled from each library (see Methods). This represents a reliable set of varying sequence coverage since any sequencing bias due to nucleotide composition, transcript length, or technical artifact is equally represented in the random sampling. We defined the “true set” of DE genes as the intersection of the DE genes identified by DESeq, edgeR, limmaVoom and baySeq using the full-size libraries and all 5 replicates. We then evaluated DESeq, edgeR, limma and PoissonSeq performance using decreasing number of replicates and sequence coverage for their: i) sensitivity rates, measured as the fraction of the true set, and ii) false positive (FP) rates, defined as the number of genes identified only by the evaluated algorithm.

Taken together these results lead to two conclusions. First, the number of replicated samples is the most predominant factor in determining DE expression. Second, DE detection of lowly expressed genes is most sensitive to the amount of coverage and replication whereas there is little benefit to increasing sequencing depths for detecting DE in highly expressed genes.