MHTP Monash Health Translation Precinct
Medical Genomics Facility

Postal/Courier Address:-
C/- Hudson Institute of
Medical Research
27-31 Wright Street,
Clayton VIC 3168

Physical Address:-
Level 4 TRF Building
Monash Medical Centre
246 Clayton Road
Clayton VIC 3168

ACRF Centre for Cancer Genomic Medicine
Monash Health Translation Precinct
Telephone +61 3 8572 2857
Facsimile +61 3 9594 7111
http://mhtpmedicalgenomics.org.au/

RNA-Seq

RNA-Seq is a common way to identify and determine the relative expression of transcriptionally regulated genes in biological tissue/cells. Like all biological experiments it requires careful planning and appropriate replicates to enable statistical analysis of your data. This document will outline our current approaches to RNA-Seq and introduce you to the concepts. In general, a minimum of 3 samples per treatment is considered adequate, however the precise number required will depend upon the diversity of the samples, the RNA isolation method, the approach used for library generation and the precise question of your research. Thus it needs to emphasized that the bioinformatician skilled in RNA-Seq who will do the analysis should have input into the planning of the experiment. They should advise as to the numbers of replicates together with the type of RNA-Seq and depth of sequencing required etc. As such this is a general guide rather than a project specific plan.

The important thing in RNA-Seq is that we are sequencing the relevant transcripts. If we just took total RNA and sequenced it all 90-95% of all the reads would be ribosomal. Thus there are different approaches in the production of NGS sequencing libraries to enrich for the type of RNA you are interested in.

It is not the purpose of this document to advise on RNA purification methods, except to advise that high quality DNA free RNA will always generated the best libraries and best sequence. While limited impurities are tolerated by some preparation methods, they will inhibit others. We usually recommend column RNA purification with DNase treatment prior to or on column.

Poly A RNA-Seq.

Up until 2-3 years ago most RNA-Seq was using libraries based upon selection of RNA with polyA tails. The procedure involves capturing the RNA with magnetic beads coupled to oligo-dT. All the RNA that doesn't have a polyA tail is washed away. The RNA is then chemically cleaved (some methods use enzymes) off the beads and into appropriate sized fragments for library generation. The fragments are reversed transcribed into cDNA; usually adding a directional primer in the process to allow strand identification. Adaptors to enable sequencing are then ligated, incorporating index sequences that enable us to bioinformatically identify the individual samples post-sequencing.

RNA requirements: Good quality DNA free intact RNA. We currently start with 1-2ug of total RNA for maximal library complexity; but can use up to 4ug or as little as 100ng.

Advantages of polyA RNA-Seq

- You are sequencing polyA RNA, which includes most protein coding transcripts, with very few reads of non-coding or ribosomal RNA.
- It requires less reads since you are just reading coding transcripts.
- It is the most cost-effective RNA-Seq.
- Strand specific – but not as important for polyA as most mammalian genes only have

pA on sense strand

Disadvantages of polyA RNA-Seq

- Any coding RNA which doesn't have a polyA tail is not captured
- There can be bias to the 3' end of the transcripts, especially for longer transcripts or less intact RNA
- No sequencing of non-coding RNA.

Ribodepletion RNA-Seq.

More and more RNA-Seq is now done with Ribodepletion; to remove ribosomal RNAs and sequence the rest. With this approach a pool of biotinylated probes, which bind to the repetitive elements of Ribosomal RNA, is annealed to the sample. After incubation, streptavidin coated magnetic beads are added to bind the tagged ribosomal RNA and allow removal. The remaining RNA is used to produce sequencing libraries similar to above.

RNA requirements: Good quality DNA free RNA. It is even more important that the RNA is free of DNA and impurities as, if these are present, they will reduce the efficiency of enzymic reactions as all but the ribosomal RNA is used for library generation. For this we usually start with 500ng but again can use as little as 100ng.

Advantages of ribodepletion RNA-Seq

- You are sequencing all long RNA, including both protein coding and non-coding transcripts; many of which have recently described regulatory roles.
- There is no bias towards the 3'end of transcripts
- It can cope with less intact RNA (e.g. FFPE isolated RNA).
- Strand specific information. This can be important where regulatory anti-sense RNA species exist and need to be identified.

Dis-advantages of ribodepletion RNA-Seq

- It costs slightly more to produce the libraries
- About 50% of the reads are of non-coding RNA so a higher depth of sequencing is required to get the same amount of data on coding transcripts, adding to sequencing costs.

Ribodepletion reagents:- There are a number of different products on the market. For mouse and human the depletion reagents are excellent, but for other species the depletion may not be as efficient.

Low input RNA-Seq.

RNA-Seq is now possible from very little RNA, even from single cells. But it should be remembered that less starting material potentially means less complexity of starting RNA, which is further affected by the need to amplify. The approaches that have been around longer rely only on polyA to prime first strand cDNA and logarithmic amplification. Newer approaches use multiple primers and linear amplification.

SPIA Amplification

The approach that we have adopted for many low input samples is the SPIA or Specific Primer

Induced Amplification method. Here the RNA is initially bound using a pool of primers that should bind all RNA species including polyA and non-polyA transcripts; *except* for Ribosomal RNA. The primer also contains a unique RNA/DNA hybrid sequence. After generation of the cDNA each product contains this unique hybrid sequence. This is subsequently used to amplify full length cDNA strands from only the 1st strands which contain this hybrid sequence i.e. linear amplification.

After amplification cDNA is purified, sheared by sonication and adaptors ligated as per usual DNA library prep.

RNA requirements: Good quality DNA free RNA. RNA also should be free of impurities as they will reduce the efficiency of enzymic reactions. Since this is only for when 100ng or less of total RNA is available we usually start with as much as is available for the lowest sample. Starting with 20-50ng results in libraries with excellent complexity however as little as 200pg can be used.

Advantages of SPIA RNA-Seq

- Very low input.
- Able to sequence when previously impossible
- Coding plus non-coding transcripts
- Linear amplification to minimize amplification artefacts
- No 3' bias in reads
- Mechanical shearing so no bias in ends of fragments

Disadvantages of SPIA RNA-Seq

- Potential for low complexity of starting sample
- Potential that low expressed transcripts may not be reliably detected.
- Unknown bias of primer pool (although does included polyA so shouldn't be more biased than polyA methods).
- Libraries are not strand-specific so unable to distinguish sense and antisense transcripts.

Small RNA

It should be mentioned that the RNA-Seq described above does not include small RNA. Since miRNA are only 20-30bp long, any such small RNA present are lost during library purification procedures. However these can be reliably sequenced using a method that selects for correct sized fragments. Briefly Total RNA (or purified small RNA) are reverse transcribed, ligated and amplified to select for correctly ligated products. Then the libraries are run on acrylamide gels and correctly sized library peaks isolated for sequencing.

RNA requirements: Good quality DNA free RNA. RNA should be isolated with a method which enriches for and/or does not lose small RNAs (e.g. Qiagen miRNeasy Mini kit). Note that degraded RNA is likely to have both larger RNA cleaved into the small RNA range plus degraded RNA so it should be avoided. 100ng-1ug of total RNA is optimal starting material or 2-50ng of isolated small RNA. (Note:- trialling new kit which can start with as low as 10ng of total RNA or 500pg of purified smRNA).

Exosomal RNA

Secreted exosomes are vesicles found in most biological fluids containing highly enriched miRNA and other partially degraded RNA from their source cell, but virtually no intact rRNA. So in an exosome we have a snapshot of the RNA in the cell it came from (including pathogens in infected cells); plus miRNA which can act as downstream signaling molecules. Exosomes are thus of interest for sequencing for 2 main reasons: (1) to identify the circulating regulatory miRNAs; and (2) to identify cellular markers of infection or disease. For the former library preparation is as described above for small RNAs, but for the latter degrading rRNA needs to be removed to avoid high number of ribosomal reads in the final data. In this case the libraries are not size selected, to allow sequencing of both miRNAs and the larger RNA species present.

Sequencing length and depth for long RNAs.

In all honesty this becomes a budget issue. But how many reads you need per sample and length of read is really a project specific choice.

Length.
The most cost effective is 50bp single end reads. This means that 50 bases of sequence are generated from each sequence in the library. This is sufficient for accurate mapping of most transcripts to well defined genomes (e.g. mouse and human).

Longer reads such as 100bp reads may assist in mapping to related genomes or where there is gene duplications etc.

Then there are paired end reads. This is where we sequence e.g. 100bp from each end of the cDNA in the library. By having longer paired-end reads mapping can show splice variants and other RNA changes.

Depth.
The number of reads needed per sample depends on both library complexity and whether you are interested in very lowly expressed transcripts. It is considered that 10million reads per sample of a polyA library is better than microarray in terms of differential expression of most transcripts. However we usually recommend 20-30 million so that most of the library complexity is represented. For libraries containing non-coding and coding RNA more reads are required to account for the increase in numbers of transcripts present (so to get equivalent numbers of reads to coding genes). For this we recommend 30-50 million reads per sample.

Sequencing length and depth for smallRNAs.

Length. Since most small RNA is in the 20-30bp range, the shortest read (50bp Single Read based upon current Illumina sequencing reagents) is sufficient to sequence the entire insert. Bioinformatic analysis will trim any read extending into the adaptor. This is sufficient for accurate mapping of most transcripts to well defined genomes (e.g. mouse and human).

Depth.
Due to the lower library complexity of small RNA libraries 10million reads per sample is adequate for most analyses.