Slides Preperation

Lumi

1 The Two Articles

- The two articles are written in 2008 and 2009.
- Some of the first essays to use RNA-seq in determining transcriptomics.
- One is the original article by Nagalakshmi et al., the other was by Zhong Wang from the same lab, which was a review article.
- The rise of next generation sequencing was in 2005.

2 Author

- Michael Snyder is now in Stanford University. His lab developed the ChIP-chip technique (and also the ChIP-seq technique), also developed the high resolution tiling DNA microarray of human genome.
- Zhong Wang is now in Berkley Lab, University of Berkeley. Computational biologist, genome analysis group lead.

3 Background

3.1 Transcriptomics

- to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs
- to determine the transcriptional structure of genes, in terms of their start sites, 5' and 3' ends, splicing patterns and other post-transcriptional modifications
- to quantify the changing expression levels of each transcript during development and under different conditions

3.2 Previous Methods

3.2.1 Large Open Reading Frames (ORFs)

Large open reading frames can contain a protein encoding sequence, but it's a very vague determination.

3.2.2 Conservative Regions

Conservative regions across species implies homologous proteins. These sequences are very likely to encode a very important functional protein.

3.2.3 Microarray Method

The microarray method of studying transcriptomics.

- Design probes with known sequences and attach them to a chip.
- We are able to know which genes are expressed in which cell by designing different flourecent dNDPs.
- And detecting with fluorometer.

However it have drawbacks.

- Reliance upon existing knowledge about the genome sequence.
- High background levels owing to cross-hybridization.
- Limit dynamic range of detection.
- Comparing expression levels across different experiments is often difficult and can require complicated normalization methods.

3.2.4 Sanger Sequencing

Expensive and highly time-consuming (RACE method mentioned later).

4 Workflow

- PCR with polyT primers or random hexamer primers.
- Need fragmentation because of sequencing limitations.

5 Data Analysis

5.1 Merging Replicates

Pearson coefficient is the correlation between the two sets of data.

5.2 Sequence Mapping

- Most detected reads are unique. (RT fig)
- No reads at the centromere position. (RB fig)
- Most chromosome positions are expressed. (L fig)

5.3 Gene Ontology

Finding the functions of your sequenced and mapped genes. Not a big deal, just an R package DEseq2 will finish the job.

5.4 Determining UTRs

- The difference between RACE and RNA-seq results.(TL fig.)
- The length of the 5' UTR region. (TR fig.)
- The comparison of 5' UTR. (B fig.)
- RNA-seq determine the 5' UTR by looking at a dramatic signal reduction.
- 3' UTR are similar.
- The Watson strand and Crick strand.
- Even if gene transcription results overlaps, the RNA-seq results can determine the UTRs.

5.5 Determining Introns

Finding transcription in mRNA in previously thought to be intron intervals.

5.6 Prediction of uORFs

- Start condon ATG before the protein coding region.
- Lac operon like things.

5.7 Discovery of New Intergenic Region

5.8 Quantification with RNA-seq

- The coefficient of determination is not different from the pearson correlation coefficient. But it's called the coefficient of determination.
- Much better than the microArray method with shrot and long sequence.

6 Challenges

6.1 Library Construction

6.1.1 Induced Bias in RNA Fragmentation

Fragmentation of DNA with DNase I has 3' bias.
Fragmentation of RNA with RNA hydrolysis etc. has a little bias towards the transcription site

6.1.2 PCR Artifacts

Unwanted production due to wierd mapping.

6.1.3 Building Strand Specific Library

Know which strand the gene is from. Not so important with proteins because we can look for start codons and stop codons. But it's important for functional RNA (e.g. antisense RNA).

6.2 Bioinformatics

- The mapping of exons and polyA tails.
- The mapping in the presence of mutations.

7 Advantages

1. As a sequencing method, it can interrogate unique sequences. Do not require a reference genome. Low background signal due to unique mapping (compared to microarray).

2. Quantitative methods.

3. Determining different UTR regions.