J. Jedediah Smith

Use the information in the videos to answer the questions. I am also giving you a paper that you can use as an alternative reference for those of you who learn better by reading.

1. For each of the cases listed below, state whether the experiment is best performed by RNA-seq or microarray and **why.**
2. A differential expression analysis of mouse miRNAs (assuming an array with an appropriate probeset exists).

Best performed with **RNA-seq** because it provides a better transcriptome overview and allows you to find new things you did not identify beforehand.

1. An analysis of differential splicing between treated and untreated mice (assuming we are only interested in annotated splicing isoforms).

Best performed with **RNA-seq** because arrays are not able to detect isoforms.

1. An analysis of differential splicing between treated and untreated mice (if we want to know if novel splicing isoforms are created).

Best performed with **RNA-seq** because arrays are not able to detect isoforms.

1. Comparison of gene expression between treated and untreated mice (if we are only interested in the ~6000 or so abundantly expressed genes)

Best performed with **microarrays** because its high throughput allows for more rapid analysis of larger more abundantly expressed genes.

1. Comparison of gene expression between treated and untreated mice (if we are interested in noncoding RNA expression)

Best performed with **RNA-seq** because it provides a better transcriptome overview and allows you to find new things you did not identify beforehand.

1. Comparison of gene expression between treated and untreated mice (if we are interested in rare (less abundant) transcripts).

Best performed with **RNA-seq** because it is more sensitive and can detect lower expression transcripts with higher accuracy than arrays.

1. Comparison of gene expression between treated and untreated wombats (if we are only interested in the ~6000 or so abundantly expressed genes)

Best performed with **microarrays** because its high throughput allows for more rapid analysis of larger more abundantly expressed genes.

1. Deep analysis of a particular class of RNAs that can be efficiently targeted for purification.

Best performed with **RNA-seq** because it is more sensitive and provides a better transcriptome overview, both of which make it suited for deep analysis.

1. Paired-end sequencing is somewhat more expensive than single-end sequencing. What advantages does it provide?

It has better accuracy and is more suited for detecting low expression genes.

1. When and why might you want to perform relatively shallow sequencing (20,000,000 – 40,000,000 reads per sample) of a lot of individuals in an experiment?

When investigating highly expression genes. Increasing the sequencing depth will have little effect on the number of detections because the majority of the genome consists of highly expressed genes. The addition of extra biological replicates, in the form of a larger number of individuals, is more clearly beneficial to the number of detections.

1. When and why might you want to perform deep sequencing (150,000,000 – 250,000,000 reads per sample) of a smaller number of individuals in an experiment?

When investigating low expression genes. Increasing the sequencing depth will have a considerable effect on the number of detections because most of the genome consists of highly expressed genes. The addition of extra biological replicates is helpful, but can make the experiment more lengthy and costly, so a smaller number of individuals is preferred.

1. Why might you want to use nanopore or PacBio (long read) sequencing instead of Illumina (short read) sequencing? Give a specific example.

They are able to process longer reads which makes alignment much easier. A specific example is isoforms. Long read techniques can directly sequence full isoforms, which grants higher detection accuracy.

1. Describe what each of the following QC metrics and what you should look for: (5 points)
   1. Phred Score: Indicates that probability that a base was correctly identified. Want a high Phred score across all of the reads.
   2. Adapter Presence: Indicates the percent of a sequence that is an adapter. Want to avoid high spikes at all points in the read.
   3. GC Content: Indicates problems with the library prep. Want to see smooth progression without a lot of peaks, but not always a problem.
   4. % of reads that align to different reference genomes: Checks to see what percent of reads align with which reference genomes. Want to see them mostly align with the reference genome you planned to use.
   5. % Duplication: Typically indicates a problem with PCR amplification of the library. Want it to be relatively small.
2. What are the advantages and disadvantages of reference-based alignment prior to differential expression analysis? (3 points)

It allows for splice junctions across exons, but requires a large amount of memory and CPU time.

1. What are the advantages and disadvantages of pseudo-alignment prior to differential expression analysis? (3 points)

It is extremely fast and lightweight, but cannot be used for discovery and does not record physical alignment position.