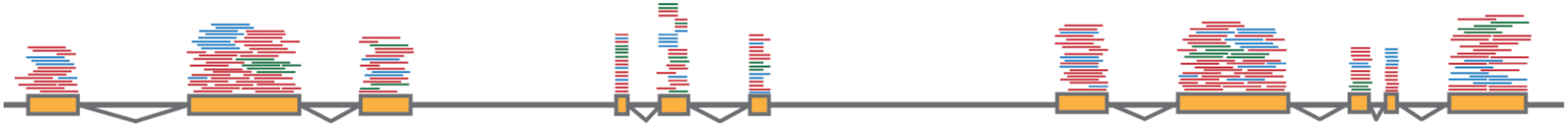
MÁSTER

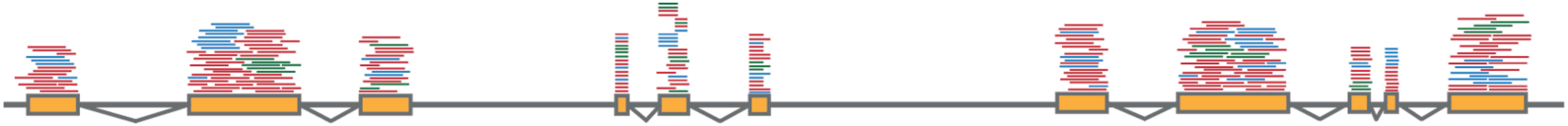
EN BIOLOGÍA COMPUTACIONAL

ENTRE ETSIAAB Y ETSIINF

RNAseq: Experimental Design



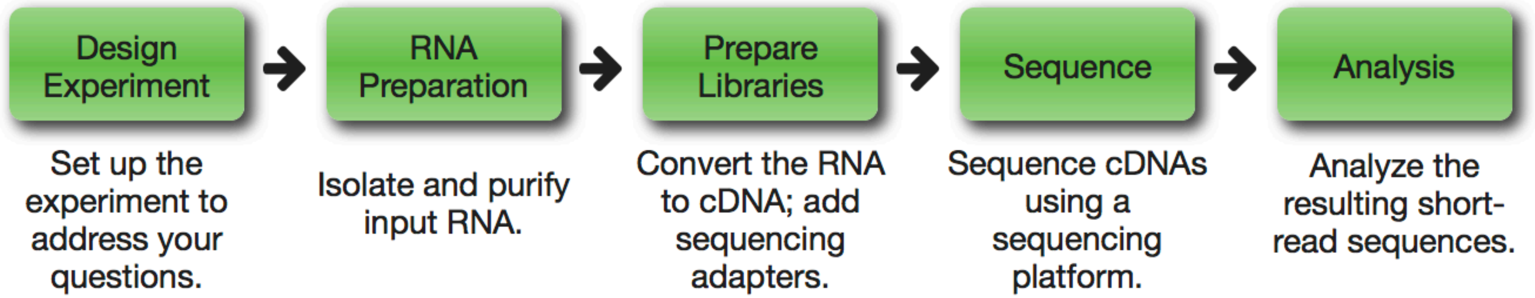
Joaquín Giner Lamia



RNA-seq produces millions of sequences from complex RNA samples. With this powerful approach, you can:

1. Measure gene expression.
2. Discover and annotate complete transcripts.
3. Characterize alternative splicing and polyadenylation.

A typical RNA-seq experiment consists of the following steps:



Design Experiment

Set up the experiment to address your questions.

RNA Preparation

Isolate and purify input RNA.

Prepare Libraries

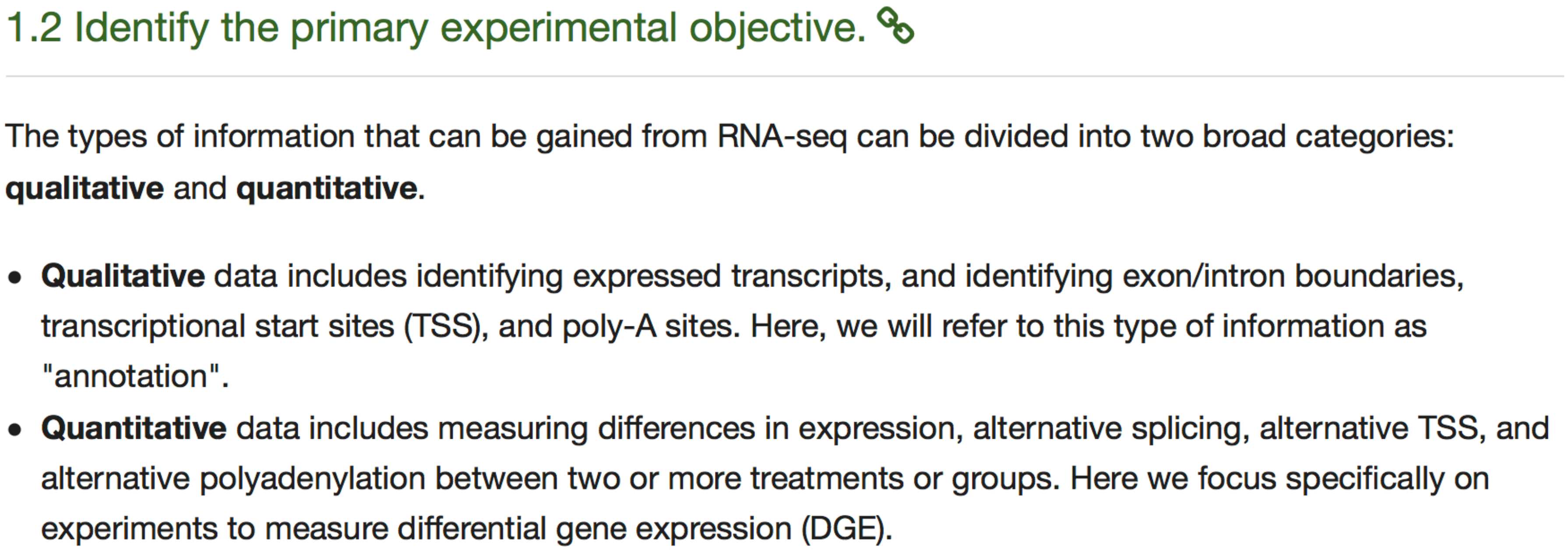
Convert the RNA to cDNA; add sequencing adapters.

Sequence

Sequence cDNAs using a sequencing platform.

Analysis

Analyze the resulting short-read sequences.



1.2 Identify the primary experimental objective.

The types of information that can be gained from RNA-seq can be divided into two broad categories: **qualitative** and **quantitative**.

- **Qualitative** data includes identifying expressed transcripts, and identifying exon/intron boundaries, transcriptional start sites (TSS), and poly-A sites. Here, we will refer to this type of information as "annotation".
- **Quantitative** data includes measuring differences in expression, alternative splicing, alternative TSS, and alternative polyadenylation between two or more treatments or groups. Here we focus specifically on experiments to measure differential gene expression (DGE).

These two types of data are related and to some degree are inseparable.

- high quality annotation can lead to greater accuracy in mapping which can lead to higher quality measures of gene expression.

mRNA-seq Applications

Differential gene expression analysis

- Healthy vs. Diseased
- Time course experiments
- Different genotypes

Transcriptional profiling

- Tissue-specific expression

Novel gene identification

- Transcriptome assembly

mRNA-seq Applications

Identification of splice variants :

- analysis of exon borders,
- patterns of alternative splicing and the study of protein isoforms.

SNP finding

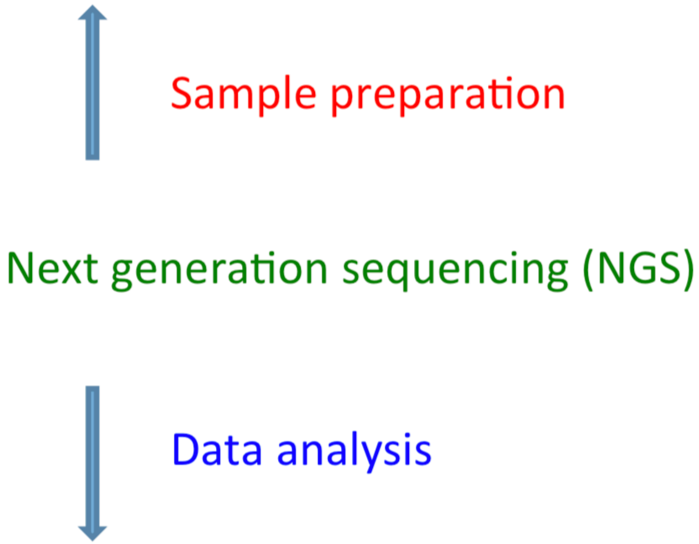
RNA editing

Discovery of "small" RNA ("small RNAs" snRNA, snoRNA, siRNA, miRNA, piRNA ("Piwi-interacting RNAs"), ...) of small size (20-30 nucleotides) and prediction of their secondary structures

How RNA-seq works

Methodology:

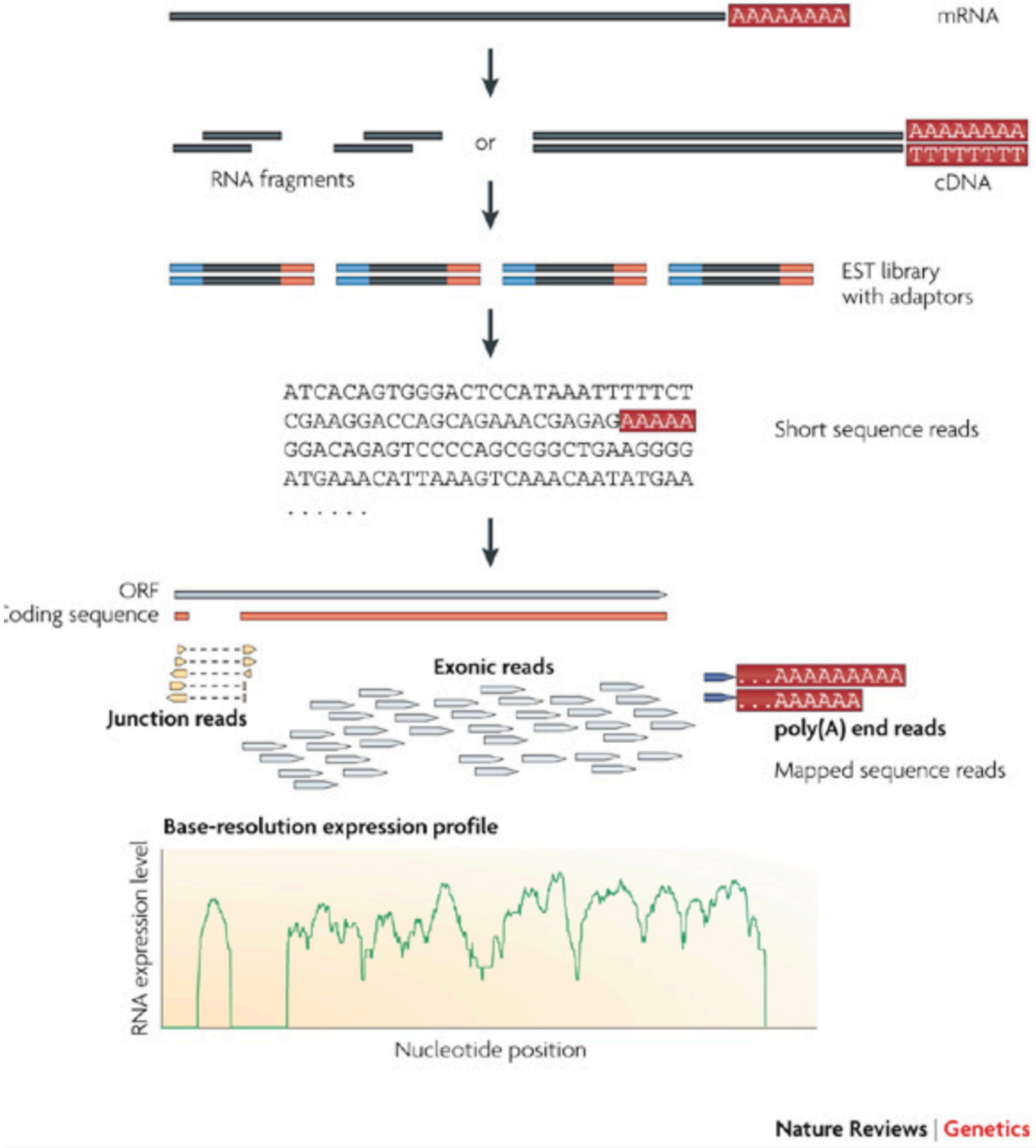
– RNA is isolated from cells,
– Fragmented at random positions,
– Copied into complementary DNA,
– Selection of fragments with a certain size range,
– Amplification using PCR,
– Sequencing,
– Reads are aligned to a reference genome or de novo assembled,
– The number of sequencing reads mapped to each gene in the reference is tabulated.



Sample preparation

Next generation sequencing (NGS)

Data analysis



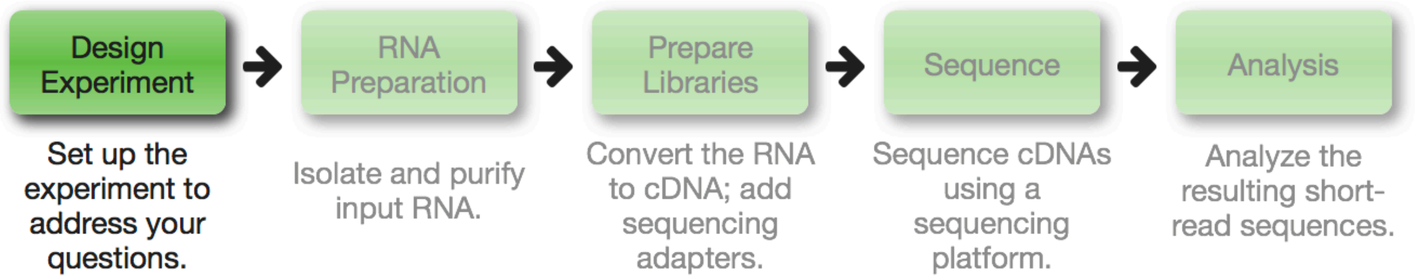
Sample preparation

Next generation sequencing (NGS)

Data analysis

Figure from Wang et. al, **RNA-Seq: a revolutionary tool for transcriptomics,** Nat. Rev. Genetics 10, 57-63, 2009).

1. Experimental Design



Design Experiment

Set up the experiment to address your questions.

RNA Preparation

Isolate and purify input RNA.

Prepare Libraries

Convert the RNA to cDNA; add sequencing adapters.

Sequence

Sequence cDNAs using a sequencing platform.

Analysis

Analyze the resulting short-read sequences.

Experiment design

What is RNAseq experiment design ?

- Answer to a clear biological question
- Take in account the indentified variation factor , the material and money constrains
- Plan the bioinformatic and biostatistics analysis

Experiment design

Sources of variance

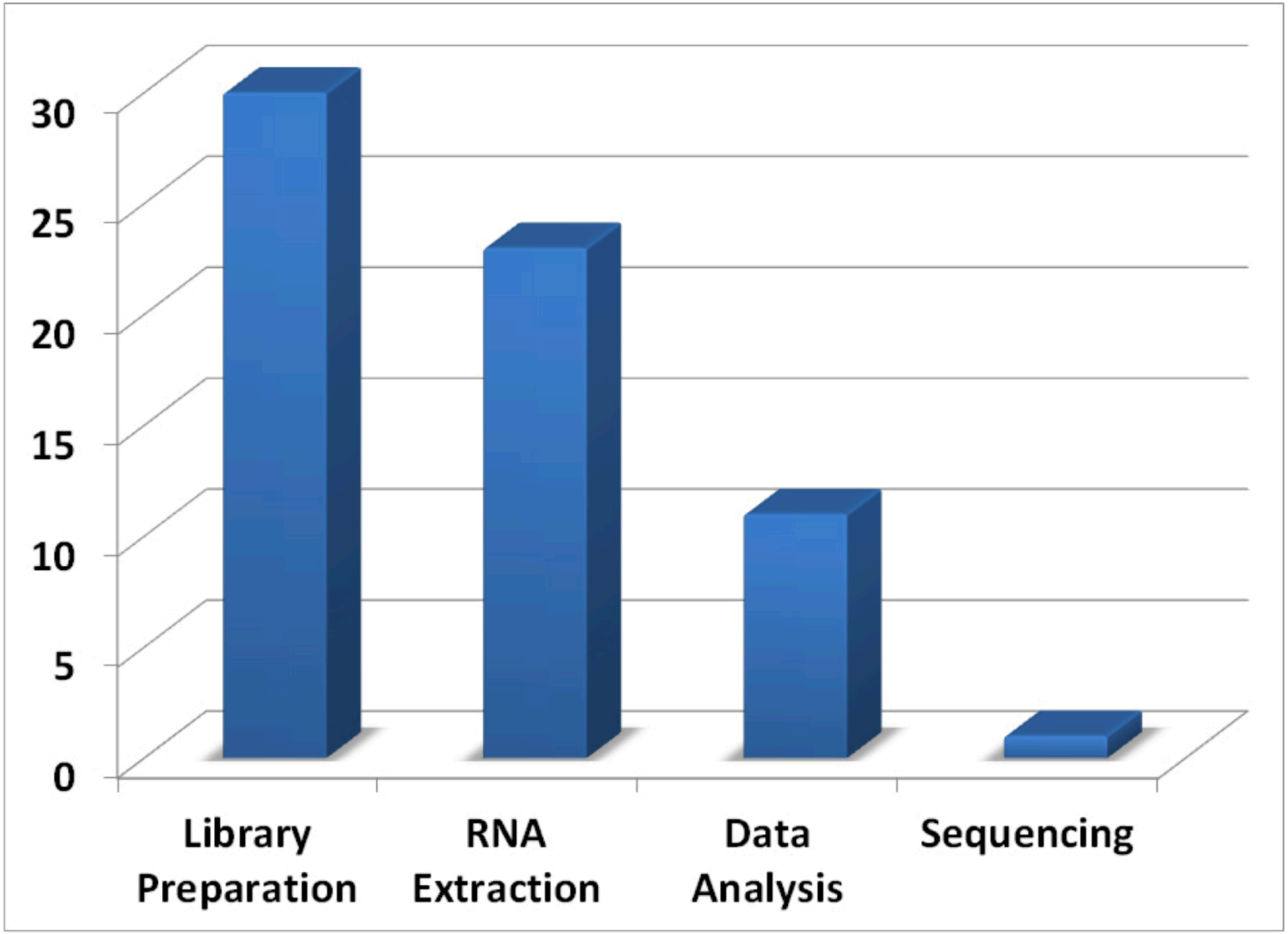
- **Sampling (fragment) variance:** Even though NGS is capable of producing millions of sequence reads, these represent only a small fraction of the nucleic acid that is actually present in the library. But also **subject sampling** (for a larger population) and **RNA sampling** (from different cells or tissues)

- **Technical variance:**
  - RNA extracted : Quality and Quantity
  - Library preparation (fragmentation, enrichment, purification, amplification, GC %, fragment orientation)
  - NGS sequencing procedures (multiplexing- sequencing kit)

- **Biological variance:** The nascent variance that is present within a treatment or control group.

- **Variance effect :** line < run < Library preparation << biological variance

Which step of an RNA-Seq experiment is the greatest source of technical variability?



Source RNASEQ Blog june 2014

Experiment design

1) how deep does one need to sequence?

2) how many biological replicates are necessary to observe a significant change in expression?

Read coverage

- Coverage = (Total Sequence)/Transcriptome Size

- Transcriptome = ~500K transcripts
  - Contamination
  - Mitochondrial, etc…
- Average Transcript length = 1000bp
- **Transcriptome size = 500K x 1kb = 500Mb**
- Total Sequence = 30M reads x 100bp x 2 = 6Gb
- Coverage = 6Gb/500Mb =12X

Read coverage

Human

Majority of expressed genes and AS events can be detected with **modest sequencing depths (~100 M filtered reads**), the estimated gene expression levels and exon/intron inclusion levels were less accurate

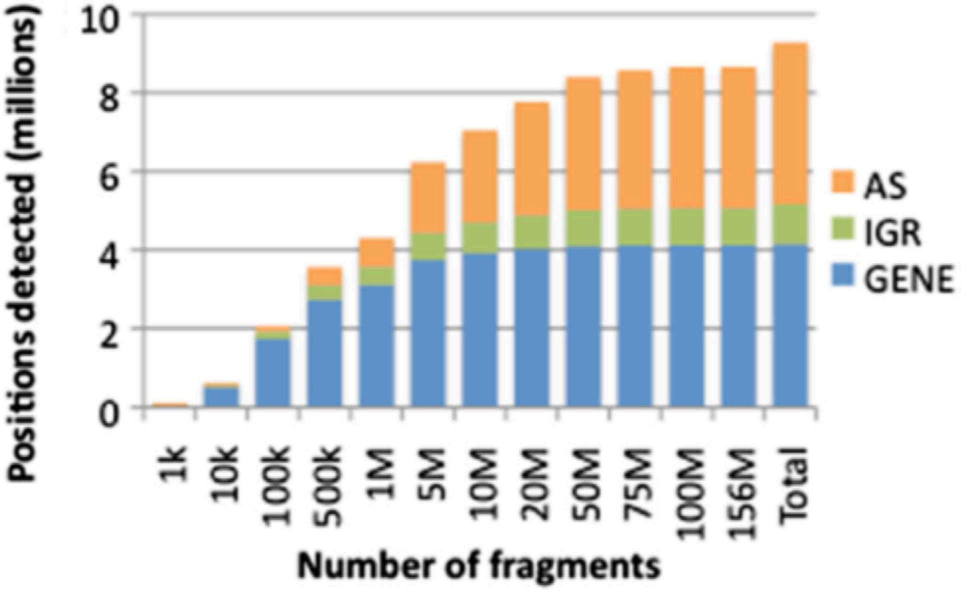
- To detect expressed genes and AS events, ~100 to 150 million (M) filtered reads were needed.
- For a DE analysis and detect 80% of events, ~300 M filtered reads were needed
- For detecting Differential AS and detect 80% of events, at least 400 M filtered reads were necessary

Evaluating the Impact of Sequencing Depth on Transcriptome Profiling in Human Adipose.
Yichuan Liu et al., 2013.

Read coverage

Bacteria

E. Coli : 5000 genes
intergenic (IGR)
antisense to ORFs or ncRNAs (AS)



« A sequencing depth of **5-10 million** non- rRNA fragments enables profiling of the vast majority of transcriptional activity in diverse species grown under di- verse culture conditions. »

Haas, B. J., Chin, M., Nusbaum, C., Birren, B. W., & Livny, J. (2012). How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC genomics, 13, 734. doi:10.1186/1471-2164-13-734

Read depth

Depends on the purpose of the experiment and the nature of the samples (ENCODE).

- 100M of reads is sufficient to detect 90% of the transcripts and 81% of the genes of the human transcriptome. (Tung et al. 2011)
- 20M reads (75bp) is sufficient to detect transcripts expressed at a medium or low level in the chicken. (Wang et al. 2011)
- 10 M of reads allow 90% of transcripts (human, zebrafish) to be covered by an average of 10 reads. (Hart et al. 2013)

Biological replicates

Why increase the number of biological replicates?

- Generalizing the results to the population
- Estimate more accurately the variation of each transcript individually (Hart et al. 2013)
- Improve the detection of differential transcripts and rate control false positives: TRUE from 3 (Sonenson et al, 2013, Robles et al 2012.)

Biological replicates

It's up to you! (Haas et al., 2012, Liu Y. et al 2013)

- **Detection of differential transcripts:**
  - (+) biological replicates
- **Construction / transcriptome annotation:**
  - (+) depth & (+) conditions
- **Search variants:**
  - (+) biological replicates & (+) depth

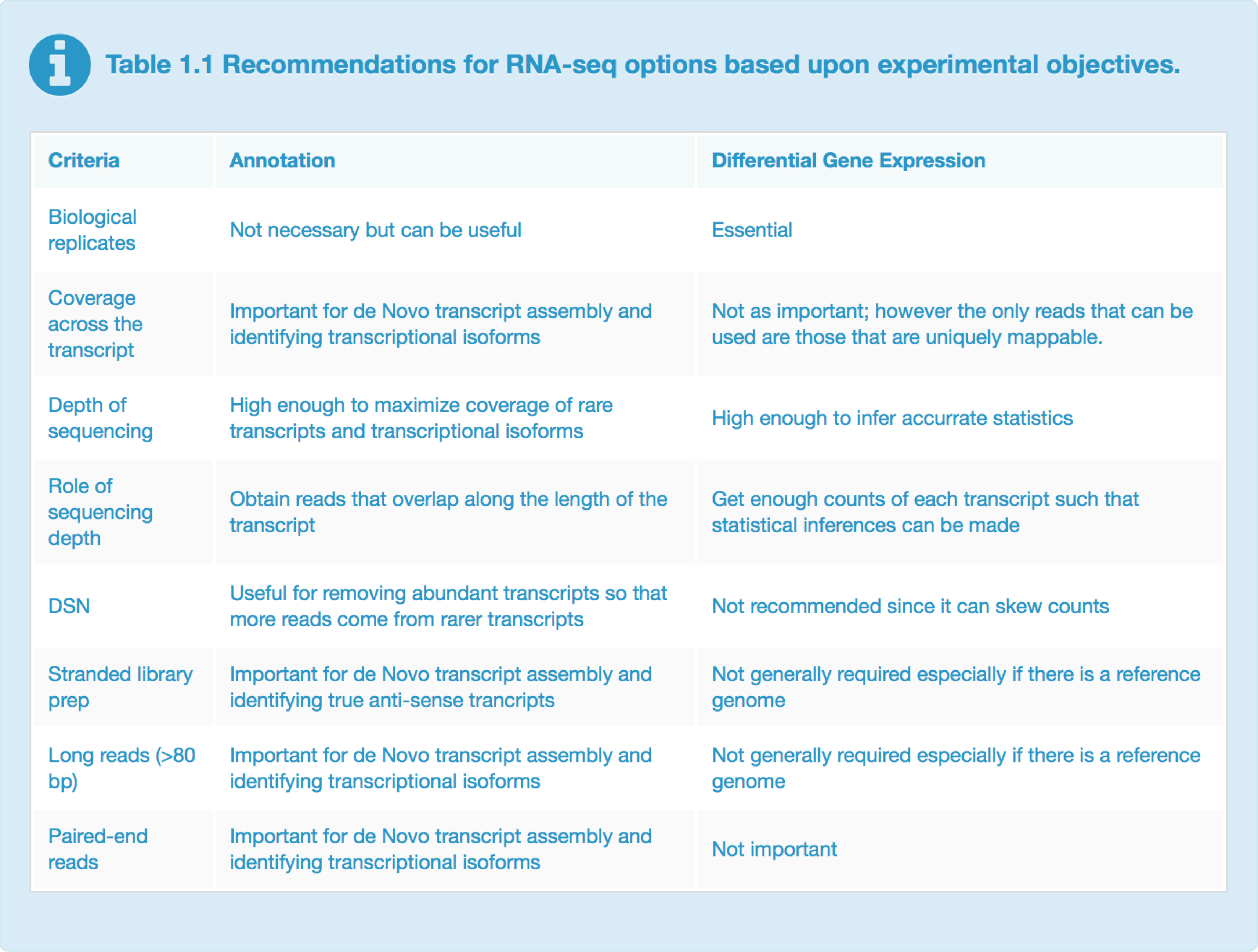
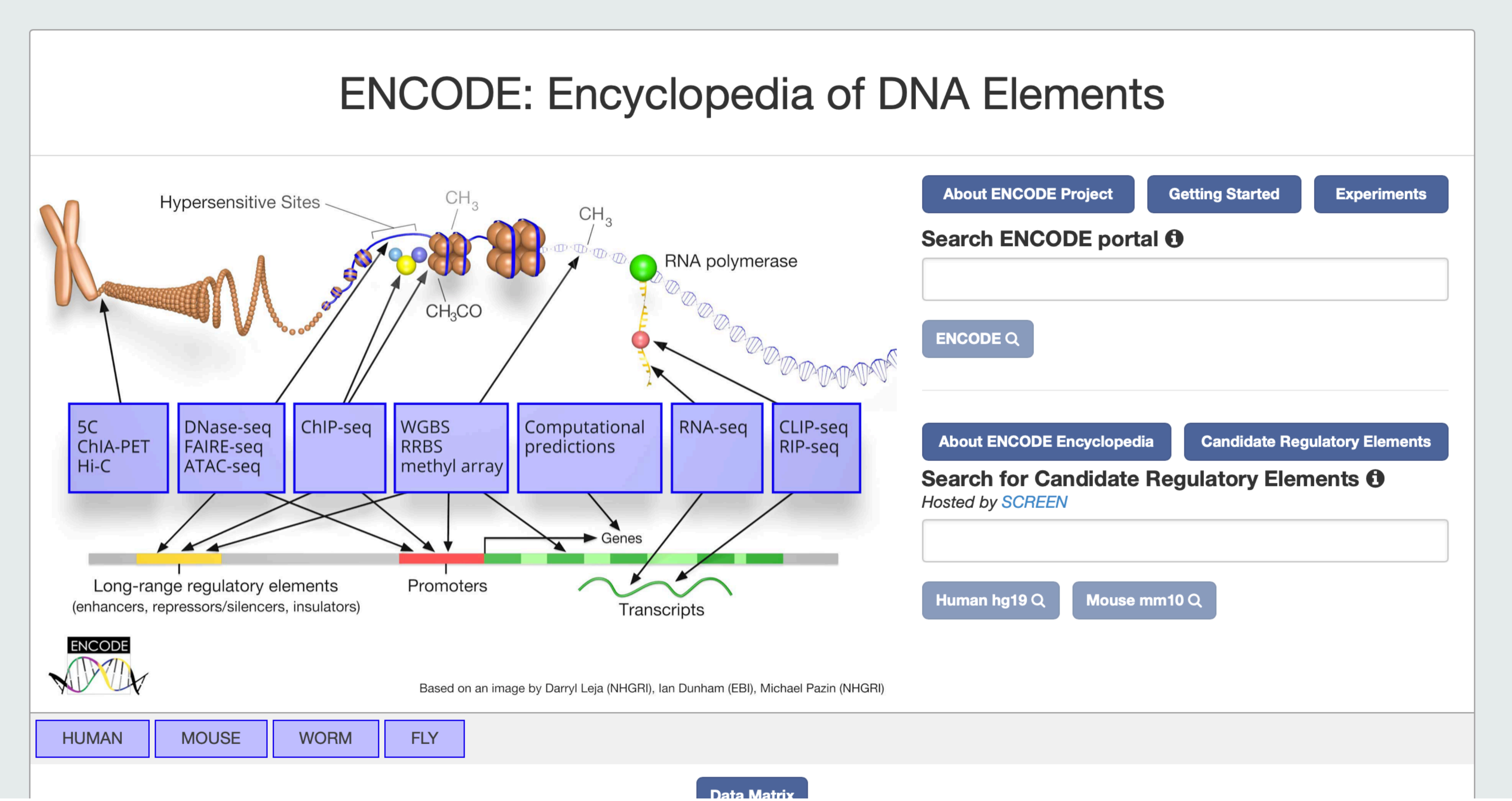


Table 1.1 Recommendations for RNA-seq options based upon experimental objectives.

Criteria	Annotation	Differential Gene Expression
Biological replicates	Not necessary but can be useful	Essential
Coverage across the transcript	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not as important; however the only reads that can be used are those that are uniquely mappable.
Depth of sequencing	High enough to maximize coverage of rare transcripts and transcriptional isoforms	High enough to infer accurrate statistics
Role of sequencing depth	Obtain reads that overlap along the length of the transcript	Get enough counts of each transcript such that statistical inferences can be made
DSN	Useful for removing abundant transcripts so that more reads come from rarer transcripts	Not recommended since it can skew counts
Stranded library prep	Important for de Novo transcript assembly and identifying true anti-sense trancripts	Not generally required especially if there is a reference genome
Long reads (>80 bp)	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not generally required especially if there is a reference genome
Paired-end reads	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not important



ENCODE: Encyclopedia of DNA Elements

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Search ENCODE portal

Search for Candidate Regulatory Elements

Hosted by SCREEN

ENCODE RECOMENDATION

- Experiments should be performed with **two or more biological replicates**, unless there is a compelling reason why this is impractical or wasteful
- A typical **R2** (Pearson) correlation of gene expression (RPKM) between two biological replicates, for RNAs that are detected in both samples using RPKM or read counts, should be between **0.92 to 0.98**. Experiments with biological correlations that fall below 0.9 should be either be repeated or explained.
- Between **30M and 100M reads** per sample depending on the study.
- **NB.** Guidelines for the information to publish with the data.

http://encodeproject.org/ENCODE/dataStandards.html

A statistical answer : Conclusions

This work quantitatively explores comparisons between contemporary analysis tools and experimental design choices for the detection of differential expression using RNA-Seq. …With regard to testing of various experimental designs, this work strongly suggests **that greater power is gained through the use of biological replicates relative to library (technical) replicates and sequencing depth**. Strikingly, **sequencing depth could be reduced as low as 15% without substantial impacts on false positive or true positive rates.**

Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., & Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics, 13, 484.