Experimental design

Etienne Delannoy[1] and Marie-Laure Martin-Magniette[1,2] & Julie Aubert[2]

1- IPS2 Institut des Sciences des Plantes de Paris-Saclay

2- UMR AgroParisTech/INRA Mathematique et Informatique Appliquees

A statistical model: what for?

Aim of an experiment: answer to a biological question.

Results of an experiment: (numerous, numerical) measurements.

Model: mathematical formula that relates the experimental conditions and the observed measurements (response).

(Statistical) modelling: translating a biological question into a mathematical model ($\neq$ PIPELINE!)

Statistical model: mathematical formula involving

- the experimental conditions,
- the biological response,
- the parameters that describe the influence of the conditions on the (mean, theoretical) response,
- and a description of the (technical, biological) variability.

Experimental Design

Definition

A good design is dedicated to the **asked question** and facilitates data analyses and interpretation of the results. It maximizes collected information and proposes experiments with respect to the financial and material constraints.

Ronald A. Fisher (1890-1962)

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of

Basic principles - Fisher (1935)

- (technical and biological) replications
  Replication (independent obs.) $\neq$ Repeated measurements
- Randomization : randomize as much as is practical, to protect against unanticipated biases
- Blocking : dividing the observations into homogeneous groups. Isolating variation attributable to a nuisance variable (e.g. lane)

Basic principles - Fisher (1935)

- (technical and biological) replications
  Replication (independent obs.) $\neq$ Repeated measurements
- Randomization : randomize as much as is practical, to protect against unanticipated biases
- Blocking : dividing the observations into homogeneous groups. Isolating variation attributable to a nuisance variable (e.g. lane)

Correspondence Nature Biotechnology (July 2011)

uld

as

vith

ıall
e or
:al

Sequencing technology does not eliminate biological variability

To the Editor:

RNA sequencing technology provides various advantages over DNA microarrays. For

type of variation that may be reduced with technology improvements[4]. Well-known sources of technical variability in both sequencing and

Steps of experiment designing

1. Formulate a broadly stated research problem in terms of explicit, addressable questions.
2. Considering the population under study, identifying appropriate sampling or experimental units, defining relevant variables, and determining how those variables will be measured.
3. Describe the data analysis strategy
4. Anticipate eventual complications during the collection step and propose a way to handle them

source : Northern Prairie Wildlife Research Center, *Statistics for Wildlifers: How much and what kind?*

How to Design a good RNA-Seq experiment in an interdisciplinary context?

Some basic rules

- Rule 1 Share a minimal common language
- Rule 2 Well define the biological question
- Rule 3 Anticipate difficulties with a well designed experiment
- Make good choices : Replicates vs Sequencing depth

Rule 2: Well define the biological question

- Choose scientific problems on feasibility and interest
- Order your objectives (primary and secondary)
- Ask yourself if RNA-seq is better than microarray regarding the biological question

Recall that RNA-Seq technology is useful to

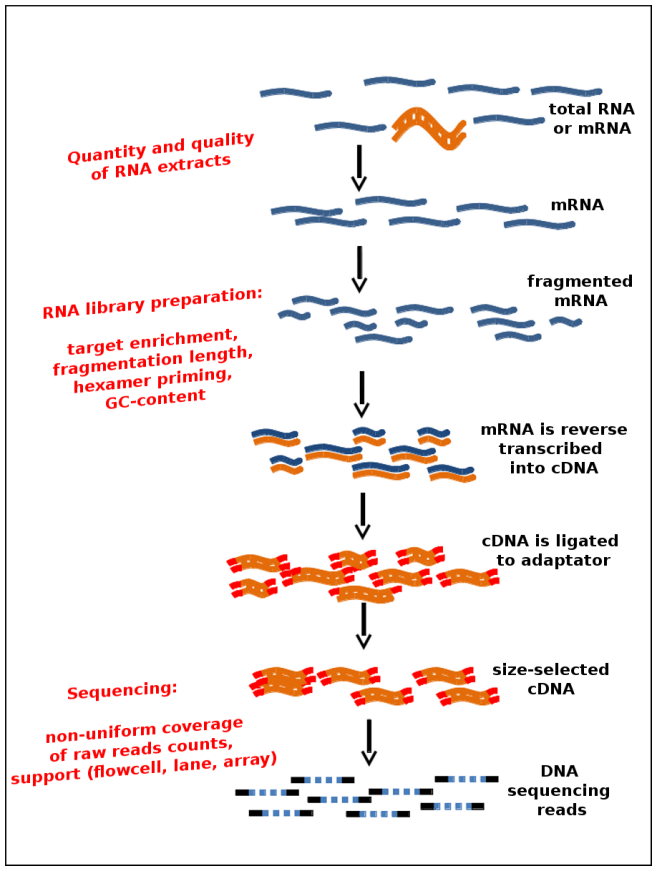
- Study all the transcribed entities
- Detect and estimate isoforms
- Construct and study a *de novo* transcriptome

Rule 3: Anticipate difficulties with a well designed experiment

1. Prepare a checklist with all the needed elements to be collected,
2. Collect data and determine all factors of variation,
3. Choose bioinformatics and statistical models,
4. Draw conclusions on results.

Be aware of different types of bias

Identify controllable biases / technical specificities



Keep in mind the influence of effects on results:

lane $\leq$ run $\leq$ RNA library preparation $\leq$ biological

(Marioni, 2008), (Bullard, 2010)

$\Rightarrow$ Increase biological replications !

Experimental Design

Technical choices

- choice of sequencing technology
- type of reads: single-end or paired-end
- type of sequencing: directional or not

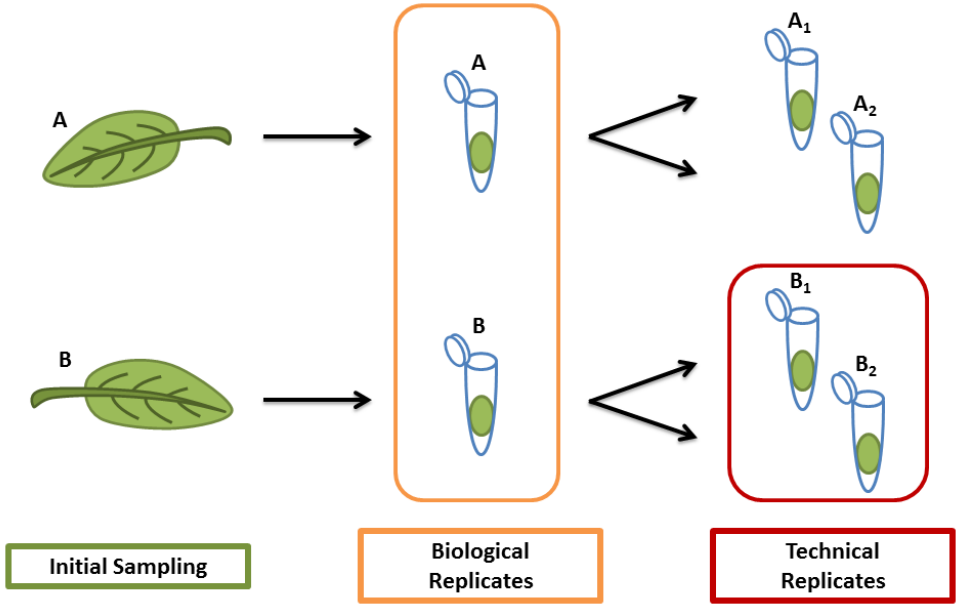
$\rightsquigarrow$ impact the library preparation protocol

Sequencing depth

Barcoding (*attaching a known sequence of nucleotides to the 3' ends of the NGS technology adapter sequences identifing a sample*) or not Pooling* of barcoded sample for a simultaneous sequencing and number of samples.

Technical challenge : combining approximately equal ratios of cDNA preparations to achieve approximately similar depths of sequencing for all samples

Biological and technical replicates



Biological replicate : sampling of individuals from a population in order to make inferences about that population

Technical replicate adresses the measurement error of the assay.

Why increasing the number of biological replicates?

- To generalize to the population level
- To estimate to a higher degree of accuracy variation in individual transcript (Hart, 2013)
- To improve detection of DE transcripts and control of false positive rate: TRUE with at least 3 (Sonenson 2013, Robles 2012)

Why increasing the number of biological replicates?

- To generalize to the population level
- To estimate to a higher degree of accuracy variation in individual transcript (Hart, 2013)
- To improve detection of DE transcripts and control of false positive rate: TRUE with at least 3 (Sonenson 2013, Robles 2012)

McIntyre et al. (2011) BMC Genomics

Technical variability $=>$ inconsistent detection of exons at low levels of coverage ($<$5reads per nucleotide)
Doing technical replication may be important in studies where low abundant mRNAs are the focus.

More biological replicates or increasing sequencing depth?

It depends! (Haas, 2012), (Liu, 2014)

- DE transcript detection: (+) biological replicates
- Construction and annotation of transcriptome: (+) depth and (+) sampling conditions
- Transcriptomic variants search: (+) biological replicates and (+) depth

A solution: multiplexing.

Tag or bar coded with specific sequences added during library construction and that allow multiple samples to be included in the same sequencing reaction (lane)

Decision tools available: Scotty (Busby et al. 2013),
Library RNAseqPower in Bioconductor (Hart et al., 2013)

To summarize

The scientific question of interest drives the experimental choices

- Collect informations before planning
- All skills are needed to discussions right from project construction
- Sequencing and other technical biases potentially increase the required sample size and sequencing depth
- Optimum compromise between replication number and sequencing depth depends on the question
- Biological replicates are important in most RNA-seq experiments
- Wherever possible apply the three Fisher's principles of randomization, replication and local control (blocking)

And do not forget: budget also includes cost of biological data acquisition, sequencing data backup, bioinformatics and statistical analysis.