HAL open science

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

Julie Aubert, Christelle Hennequet-Antier, Cyprien Guerin, Delphine Labourdette, Anne de La Foye, Nathalie Marsaud, Fabrice Legeai, Frederique Hilliou, Brigitte Schaeffer

To cite this version:

Julie Aubert, Christelle Hennequet-Antier, Cyprien Guerin, Delphine Labourdette, Anne de La Foye, et al.. How to Design a good RNA-Seq experiment in an interdisciplinary context?. European conference on Computational Biology, Sep 2014, Strasbourg, France. , 2014. hal-01462728

HAL Id: hal-01462728

https://hal.science/hal-01462728

Submitted on 3 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

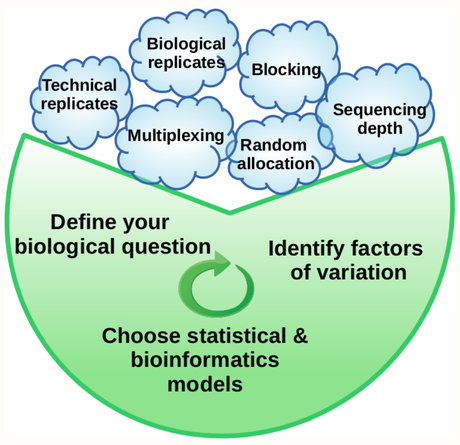
How to Design a good RNA-Seq experiment in an interdisciplinary context?

Pôle Planification Expérimentale, PEPI IBIS [1],

[1]INRA, France

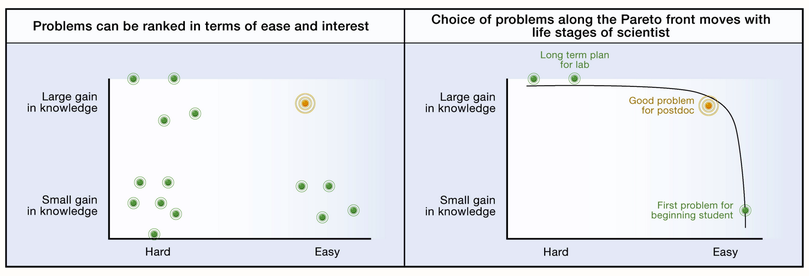
RNA-seq technology is a powerful tool for characterizing and quantifying transcriptome. Upstream careful experimental planning is necessary to pull the maximum of relevant information and to make the best use of these experiments.

An RNA-seq experimental design using Fisher's principles



Rule 1: Share a minimal common language

Rule 2: Well define the biological question



From Alon, 2009

- 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

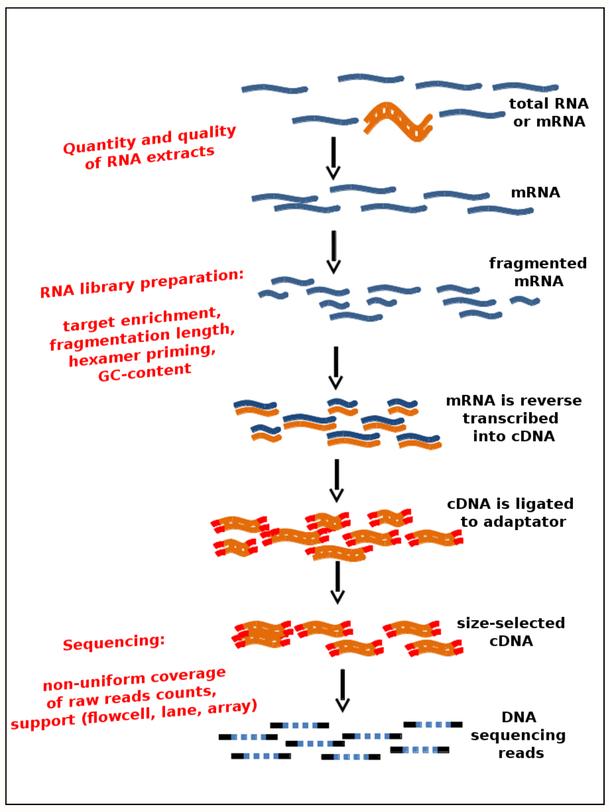
Make a choice

- Identify differentially expressed (DE) genes?
- Detect and estimate isoforms?
- Construct a de Novo transcriptome?

Rule 3: Anticipate difficulties with a well designed experiment

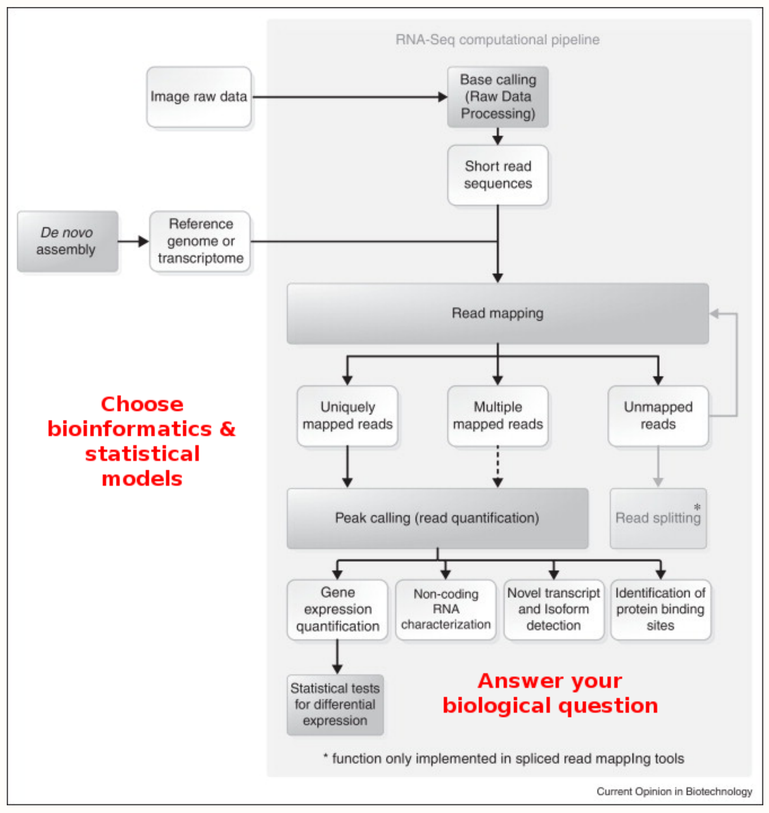
1. Prepare a checklist with all the needed elements to be collected,
2. Collect data and determine all factors of variation,
3. Choose bioinformatics and statistical models,
4. Draw conclusions on results.

Be aware of different types of bias



Keep in mind the influence of effects on results: lane $\leq$ run $\leq$ RNA library preparation $\leq$ biological (Marioni, 2008), (Bullard, 2010)

RNA-seq experiment analysis: from A to Z



Adapted from Mutz, 2013

Rule 4: Make good choices

How many reads?

- 100M to detect 90% of the transcripts of 81% of human genes (Toung, 2011)
- 20M reads of 75bp can detect transcripts of medium and low abundance in chicken (Wand, 2011)
- 10M to cover by at least 10 reads 90% of all (human and zebrafish) genes (Hart, 2013)...

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)

More biological replicates or increasing sequencing depth?

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

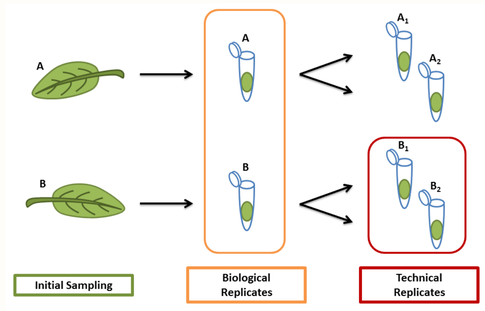
- DE transcript detection: (+) biological replicates
- Construction and annotation of transcriptome: (+) depth and (+) sampling conditions
- Transcriptomic variants search: (+) biological replicates and (+) depth

A solution: multiplexing.

Decision tools available: Scotty (Busby, 2013), RNAseqPower (Hart, 2013)

Some definitions

Biological and technical replicates:



Sequencing depth: Average number of a given position in a genome or a transcriptome covered by reads in a sequencing run

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)

Blocking: Isolating variation attributable to a nuisance variable (e.g. lane)

Conclusions

- Clarify the biological question
- All skills are needed to discussions right from project construction
- Prefer biological replicates instead of technical replicates
- Use multiplexing
- Optimum compromise between replication number and sequencing depth depends on the question
- 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.

Who are we?

julie.aubert@agroparistech.fr, anne.delafoye@clermont.inra.fr, **cyprien.guerin@jouy.inra.fr**, **christelle.hennequet@tours.inra.fr**, frederique.hilliou@sophia.inra.fr, fabrice.legeai@rennes.inra.fr, delphine.labourdette@insa-toulouse.fr, nmarsaud@insa-toulouse.fr, brigitte.schaeffer@jouy.inra.fr

ECCB14, 7 - 10 September 2014, Strasbourg, France