System Biology 1 - Conclusion
Interdisciplinarity and Data integration

G. RIGAILL

IPS2, LaMME

November 2020

Outline

System Biology 1

An introduction to high-throughput molecular biology : “omics”

In genomics

- technologies evolve very quickly and are based on increasingly sophisticated molecular biology, chemistry or physics techniques
- increasingly sophisticated computer and mathematical methodologies are being developed to analyze omic data

$\Rightarrow$ We are living in an exciting time for molecular biology.

System Biology 1

An introduction to high-throughput molecular biology : “omics”

This week was an introduction to

- some of the technologies : RNAseq, PPI, ...
- some of the methodologies to analyze the data : bioinfo, biostat, ...
- I will discuss some of the challenges related to their analysis, interpretation and integration

Analyzing and interpreting omic data is not simple

Interdisciplnarity

- Ideally one would like to follow a guide of good practises ?
  - ex : if assumptions A, B and C are true you should use this method...
- But it is not always that simple because
  1. both technologies and methodologies evolve very rapidly
  2. not easy to check the validity and importance of the assumptions
- How to pick and justify the use of one methodology ?
  - Make our choices understandable and reproducible
  - A dialog between biologists, bioinformaticians, statisticians...

Outline

Outline

Comparing two populations -1

A simple looking scenario

- Two tests are often considered : (i) the Student test and (ii) the Wilcoxon test.
- On wikipedia (August 2020) we can read about the Wilcoxon test :

  can be used as an alternative to the paired Student's t-test when the sample size is small and the population cannot be assumed to be normally distributed.

Comparing two populations -2

A simple looking scenario

A nice cooking recipe ?

It would therefore be a question of knowing whether the distribution is

- Gaussian : in that case we use the Student test
- or not : in that case we use the Wilcoxon test.

In practise it is a bit more complex because

- The Student test is somewhat robust to the normality assumption
  see T. Lumley, et al., « The importance of the normality assumption in large public health data sets », Annual review of public health
- Other assumptions are possibly more important : independance, equal variances...
- “All models are wrong” : what does it means exactly ?

Outline

A thought experiment and analysis - 1

Imagine a biologist measures the expression of a gene

An experiment

- Using digital PCR for example
- We want to know if there is a difference in expression between treated and not treated cells
- We have $n = 3$ biological replicates (which is fairly standard for this kind of experiment)

A thought experiment and analysis - 2

We need to choose a method to analyze

- We are hesitating between the Student's t-test and the Wilcoxon test.
- For simplicity, let's rule out any problem with data normalization and consider only the default versions of the tests in R :

Student in R	t.test(cell.line.ctrl, cell.line.trt)
Wilcoxon in R	wilcox.test (cell.line.ctrl, cell.line.trt)

A thought experiment and analysis - 3

Looking for a true model ?

- The data is probably not Gaussian...
- The assumptions of the Wilcoxon test seem to be true. Should we use the wilcox.test then ?

Statistically how should we choose ?

Our choice should be guided by the ability of these two tests to

- detect real differences : power (H1)
- do not misidentify a difference if there is not one (H0)
- we can assess that using simulation here

Outline

Some simple simulations (math)

$$X_{ci} = \mu_c + \varepsilon_{ci}, \qquad \varepsilon_{ci} \sim \mathcal{N}(0,1) \quad \text{i.i.d.}$$

- X_{ci} is the expression of gene A in
  - the condition c (1 : treated or 2 : untreated / control)
  - the replicate i (1, 2 or 3).
- The noise is Gaussian and has a variance of 1.
- The average expression difference between the two conditions is $\mu_1 - \mu_2 = \delta$.

Repeat

Repeating this simulation a large number of times allows us to study the distribution function of the p-values of the test by Student and Wilcoxon.

Some simple simulations in R

```
> n = 3; size_eff = 2
> cellLine1 <- rnorm(n)+ size_eff   ## Treated
> signif(cellLine1,3)
[1] 2.11 1.96 1.85
> cellLine2 <- rnorm(n)  ## Control
> signif(cellLine2,2)
[1] -0.14 -0.56 -1.40
>
> t.test(cellLine1, cellLine2)$p.value  ## T-test
[1] 0.01618702
> wilcox.test(cellLine1, cellLine2)$p.value ## W-test
[1] 0.1
```

Repeat in R

```
## sample.size default=3
## delta default = 3
one.simu <- function(n=3, size_eff=2){
  cellLine1 <- rnorm(n)+ size_eff
  cellLine2 <- rnorm(n)

  pval <- c(
            t.test(cellLine1, cellLine2)$p.value,
            wilcox.test(cellLine1, cellLine2)$p.value
            )
  met <- c("t.test",  "wilcox.test")
  data.frame(pval=pval, test=met)
}

## 10^3 simulations under H0 (size_eff=0)
replicate(10^3, one.simu(3, 0), simplify=F)
```

What do we expect ?

Remember, what is a p-value ?

- Wikipedia : “is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct"
- and under the assumption that the assumptions of the test are true (example : independence for both Student and Wilcoxon...)

What do we expect ?

Under H0

- Under H0 and a threshold of 5% we hope to get a p-value smaller than 5% less than 5% of the cases
- Under H0 and a threshold of α we hope to get a p-value smaller than α less than α of the cases

What do we expect ?

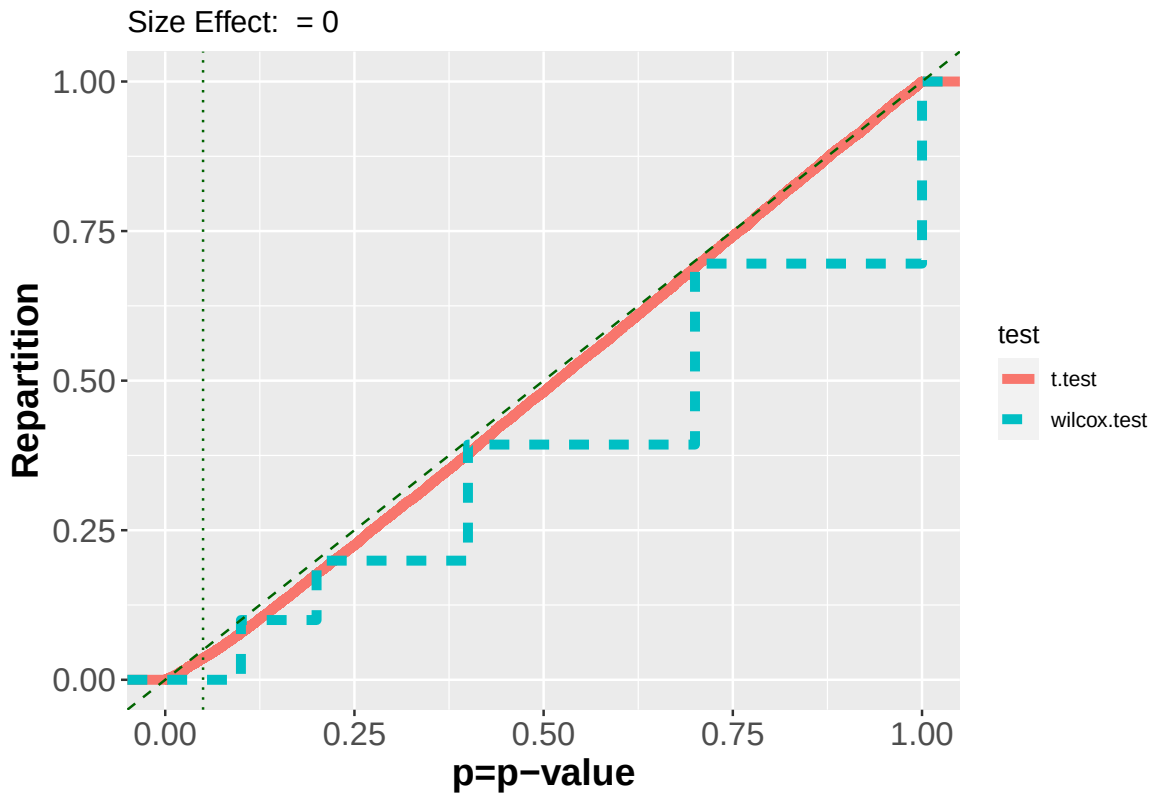
Under H0

- Under H0 and a threshold of 5% we hope to get a p-value smaller than 5% less than 5% of the cases
- Under H0 and a threshold of α we hope to get a p-value smaller than α less than α of the cases

H0 control for $n = 3$

For $\alpha = 5\%$

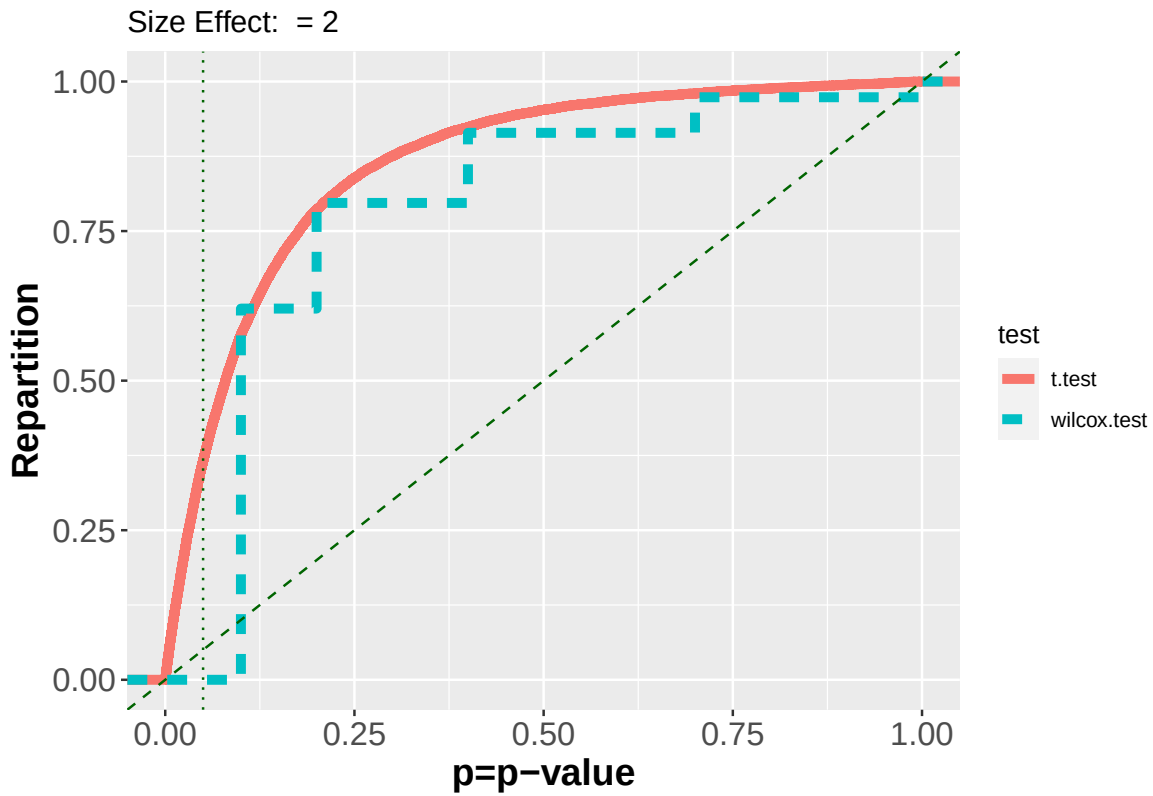
- 3,4% for the Student test
- 0% for the Wilcoxon test



Power for $\delta = 2$ and $n = 3$

For $\alpha = 5\%$

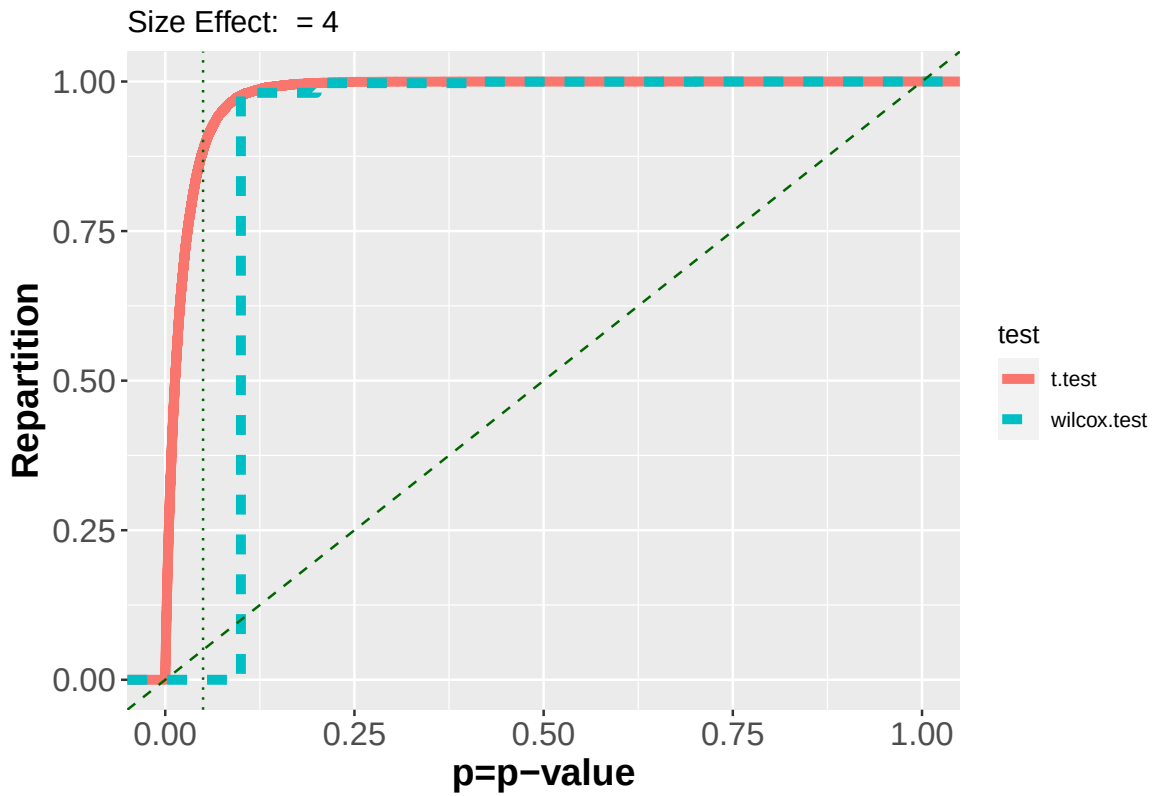
- 37.5% for the Student test
- 0% for the Wilcoxon test



Power for $\delta = 4$ and $n = 3$

For $\alpha = 5\%$

- 88.4% for the Student test
- 0% for the Wilcoxon test

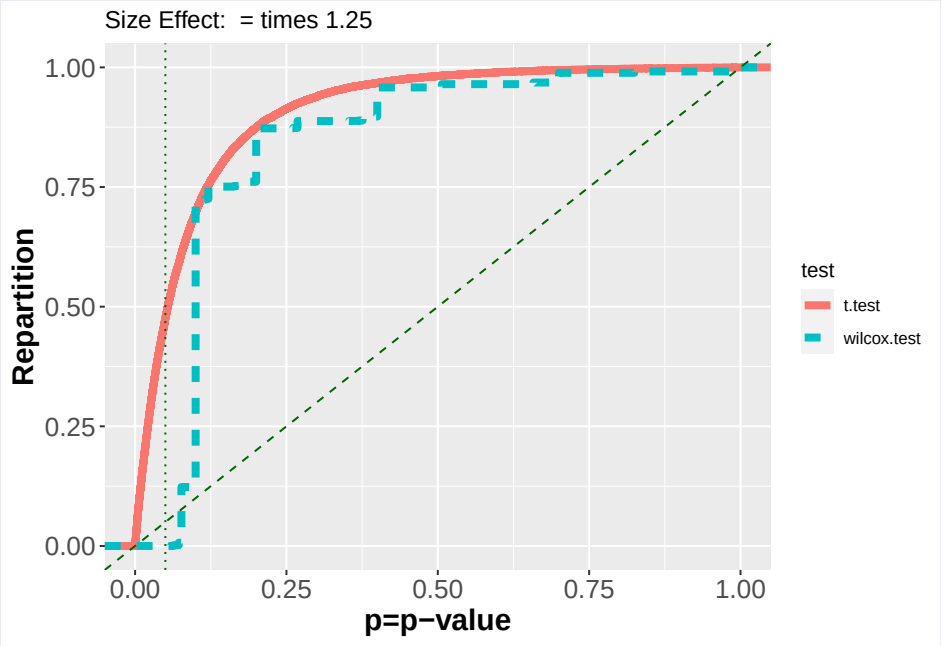


Outline

Partial Conclusion

Student's t-test

- Some power at $\alpha = 5\%$
- Note that we are probably optimistic for the Student test as we simulated Gaussian noise
- An example with a Poisson distribution below :



Partial Conclusion

Wilcoxon's test

- A power of 0 at $\alpha = 5\%$
- If we don't want to assume the data to be Gaussian, we're going to have to
  - ask to perform some additional experiments ?
  - or use a different p-value threshold ?

Wilcoxon test : less assumptions - less power

- It does not make any assumptions about the distribution of errors
- It only considers ranks
- it will give the same results on the following table

	treated cell line			control cell line			p-value
	x_{11}	x_{12}	x_{13}	x_{21}	x_{22}	x_{23}	
Data 1	10	10.1	10.2	13	13.1	13.2	0.1
Data 2	10	11	12	13	14	15	0.1

- Essentially, it neglects that dPCR is quantitative... Is this realistic ?

Conclusion - 1

Choosing an approach is not simple

- One often needs to consider the details of the experiements (sample size, biases, the question...)
- In our previous example with $n = 3$ more simulations would be needed to conclude but in short
  1. For the Student t-test, the Gaussian assumption is unrealistic but the test has some power if the data is not too “unGaussian”
  2. For the Wilcoxon test, only considering the ranks is not sufficient to get power with $n = 3$ (we need larger n)

Conclusion - 2

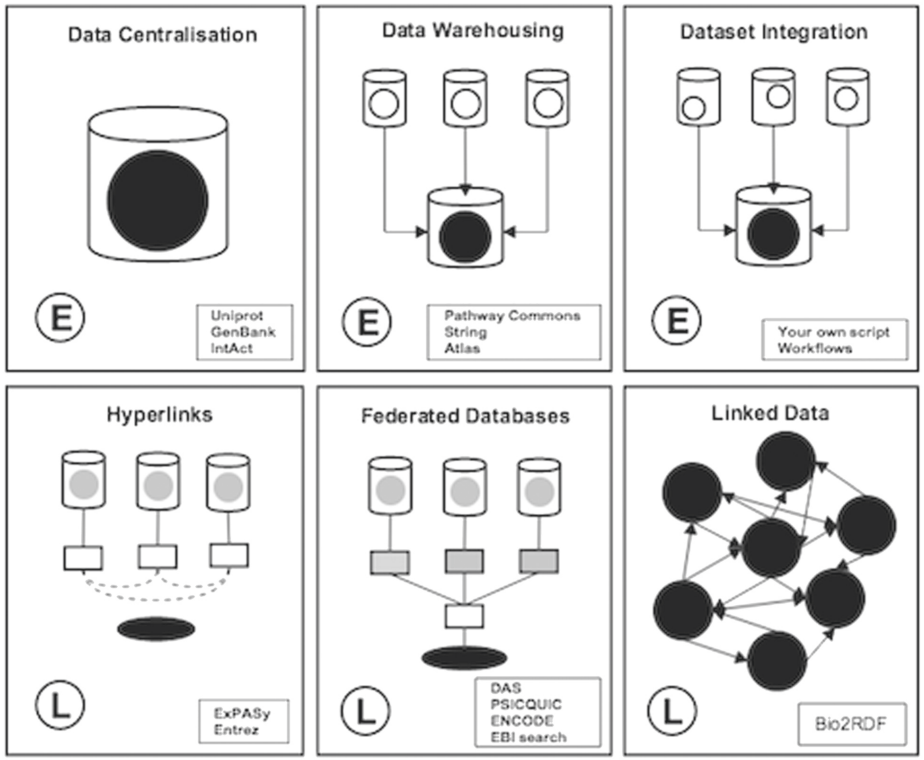
A realistic model ?

- Not a model whose assumptions are all true
- Rather a model to efficiently adress our question(s)
- That is why, it is often argued that “All models are wrong and some are usefull”
- A dialog between biology, bioinformatics, statistics, ... is needed

$\Rightarrow$ We are living in an exciting time for molecular biology !
But be carefull, try to understand and question the assumptions (talking to others scientists) !

Outline

Upstream, Data management [Latapas et al. 2015]



- ensure the reproducibility of the analysis and interpretation
- needs to be driven by the actual users
- need to define, adopt and use standards

A definition inspired by [Ritchie et al. 2015]

"...The integration of multi-omic information in a meaningful way to provide a more comprehensive analysis of a biological point of interest..."

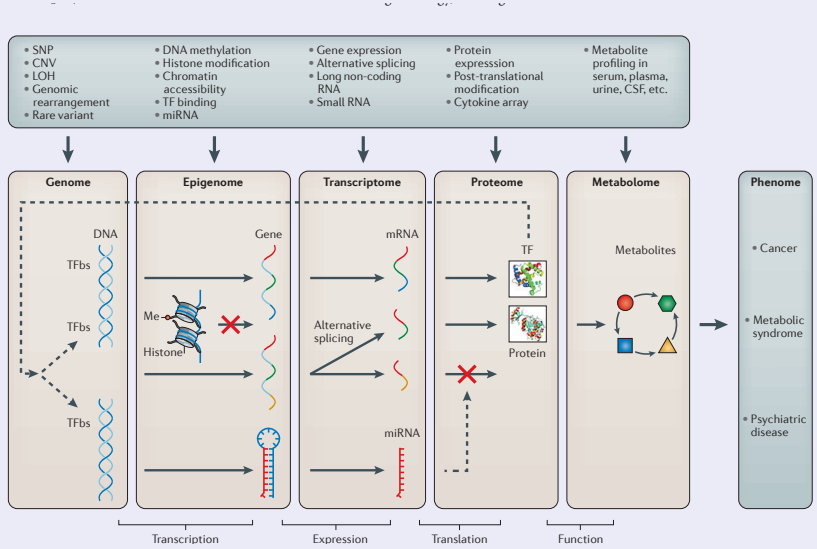
- To
  1. Predict a phenotype or the outcome of an intervention
  2. Identify biomarkers
  3. Better understand molecular mecanisms or the underlying genetic basis

The promise

Biology level by level

- Highlighted the complexity of interactions
- Remains to explore them
  - intra and inter-level

... to increase knowledge



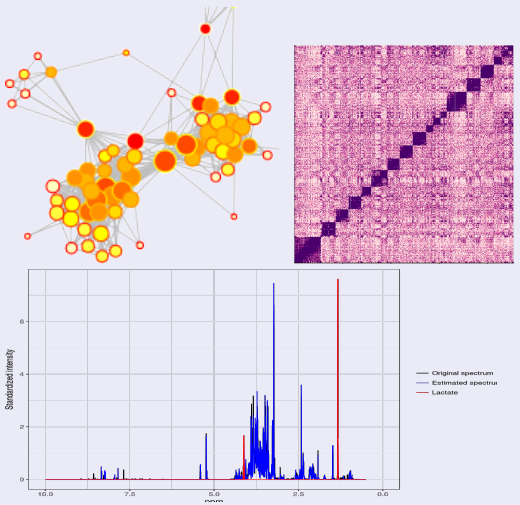
Towards an integrative biology [Ritchie et al. 2015]

"...the complete biological model is only likely to be discovered if the different levels of genetic, genomic and proteomic regulation are considered in an analysis."

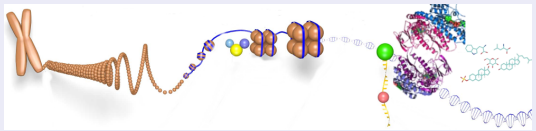
The diversity of multi-source data

Big and complex data
[N. Vialaneix 2018]

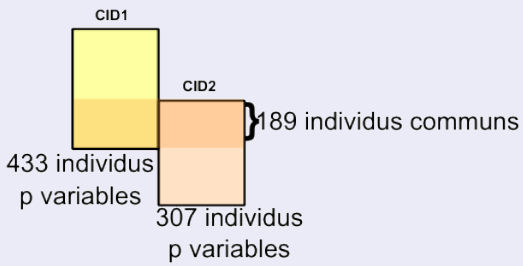
Heterogeneous data



Multi-scale data



Unbalanced datasets



Analysis of each dataset [Ritchie et al. 2015]

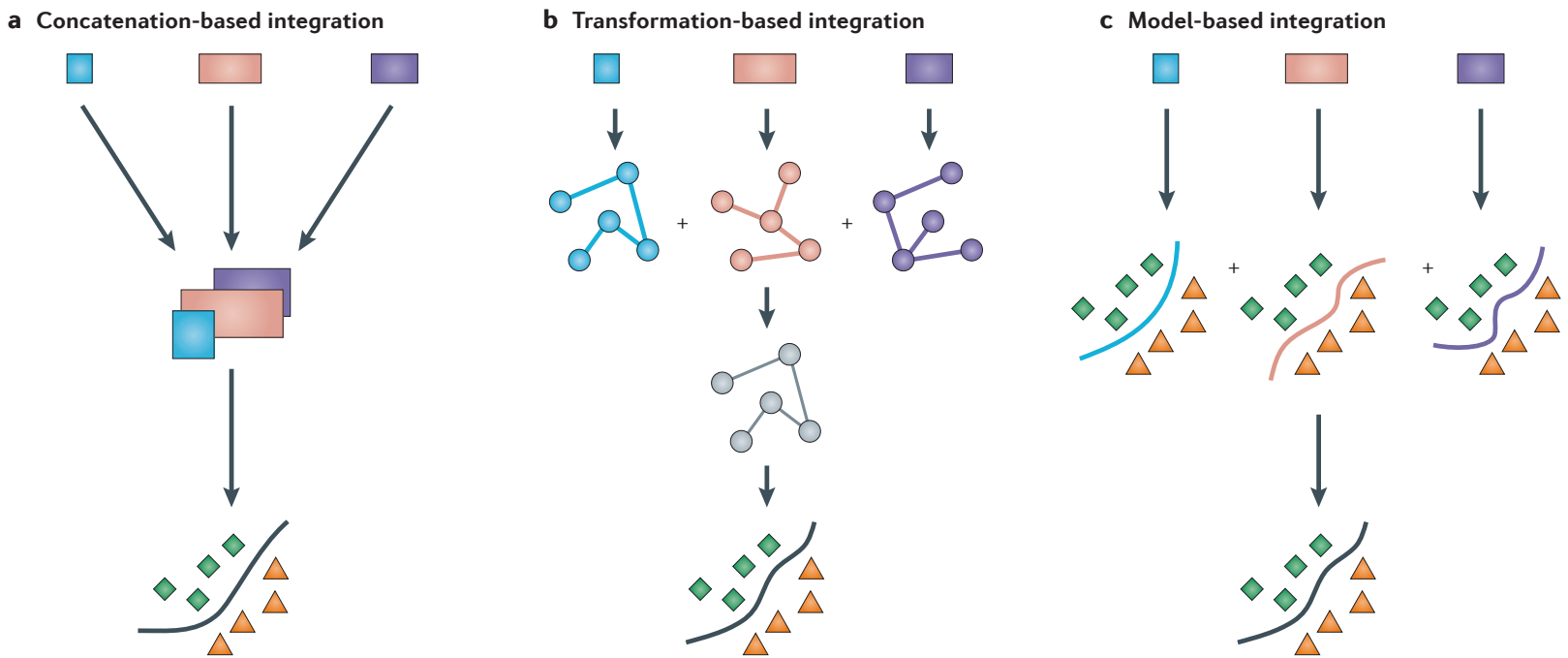
- Quality Assurance and Control
  - To have high quality results you need high quality data
- Dimension reduction to increase power
  - Reduce the number of variables per dataset :
    - ★ p : many genes, metabolites, proteins, ...
    - ★ n : few experimental conditions

On a dataset i on a	n	$\ll$	p_i
On all datasets	n	$\lll$	$\sum_i p_i$

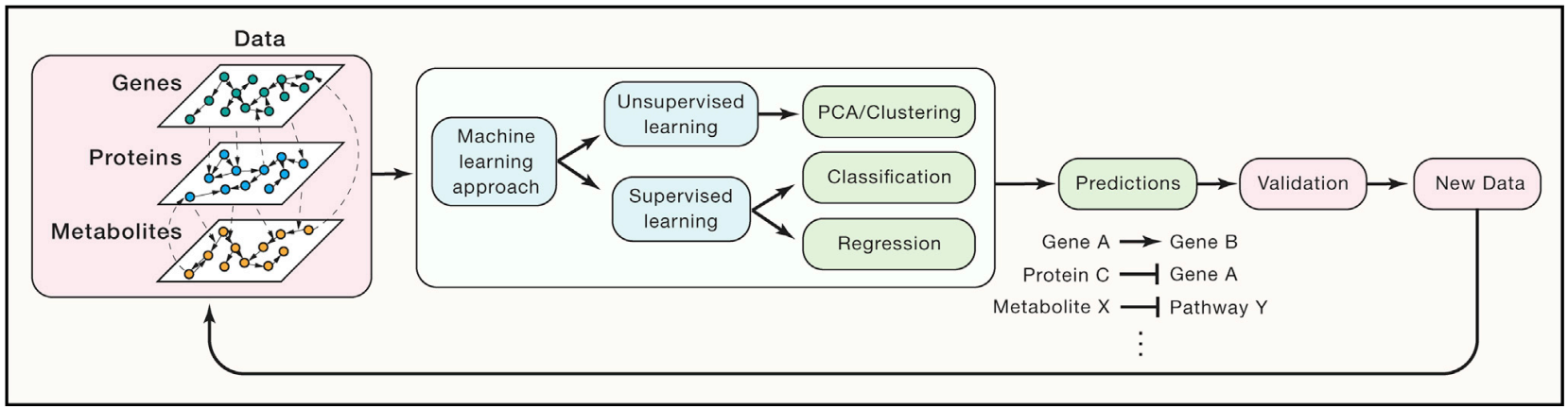
  - Many methods : Filtering, PCA, Data-Mining...

Several types of integration

- In several stages
  - each step should enrich the signal
- Multidimensional
  - simultaneous analysis of all datasets



AI and knowledge acquisition [Camacho et al. 2018]



- Towards a science more focused on data, calculation and simulation
  - What do we want to predict ?
  - What do we want to understand ?
  - How to validate ?

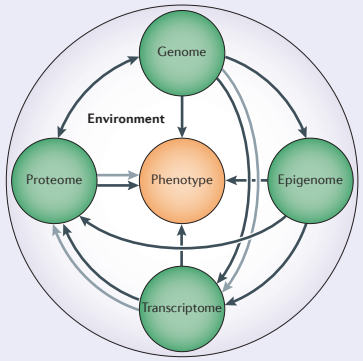
A few challenges

Tackling methodological obstacles

- The big dimension
- Missing data
- Prediction in an uncertain context
- Validation

Define the “biological question”

- Not that simple
  - predict or understand
  - supervised or unsupervised
- The hypotheses
  - there are bound to be some
  - we should state them



As a conclusion

- Context of rapid transitions :
  - renewed articulation between acquisition - processing - modeling
- Evolution towards a science more centered on data, calculation and simulation ?
  - understanding remain essential !
- Diversity of approaches
  - linked to the diversity of data and biological questions
  - hybridation
  - adaptation
  - importance of methodological research at the interface between disciplines