

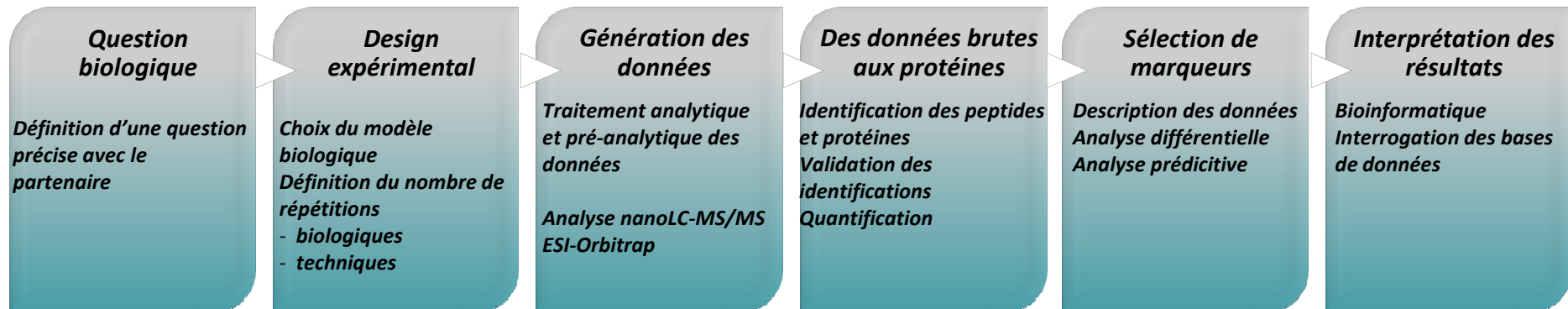
Introduction aux biostatistiques pour la découverte de biomarqueurs

Truntzer Caroline
Atelier Prospectom – 19/11/2014

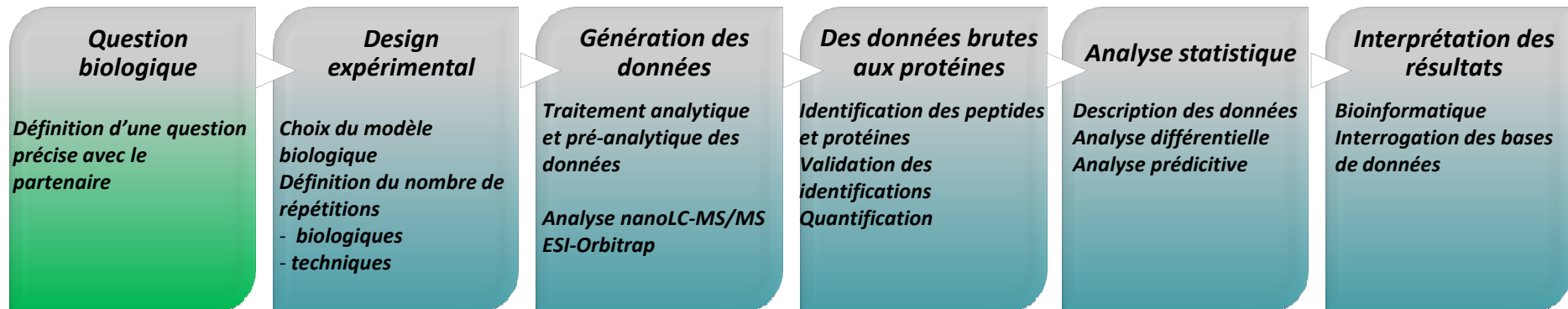
Objectifs

- Présentation des concepts principaux pour la découverte de biomarqueurs en protéomique
- Analyse label-free par nano-LC/MS-MS (ESI-Orbitrap)
 - Partage d'expérience
 - Principes généralisables à d'autres types d'instruments, de techniques
- Point de départ pour la discussion

Etapes de l'analyse



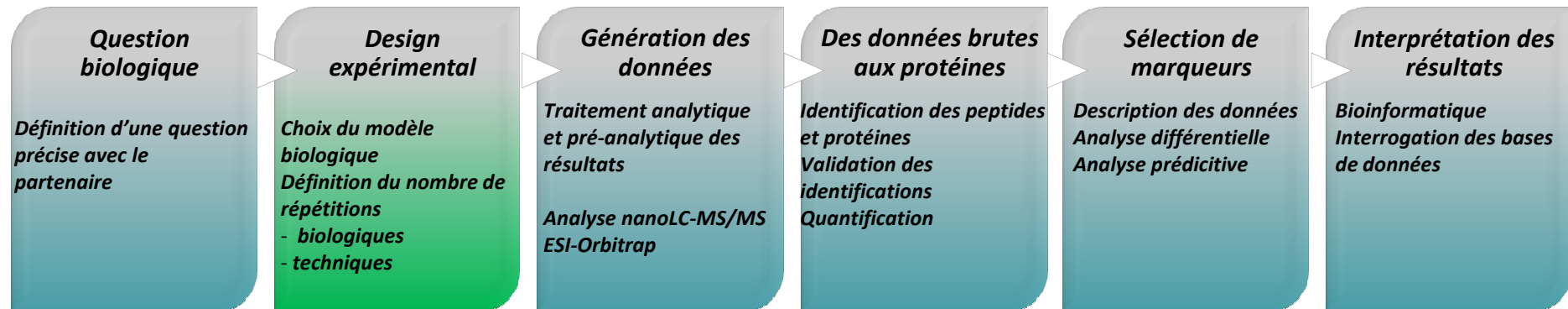
Question biologique



Définition de la question

- Evident...mais pas tant que ça!
 - Importance de préciser la question biologique/clinique AVANT de commencer l'analyse
 - Quelles comparaisons?
 - Dans quel but?
 - Sur quelle population?
- ⇒ Définition du plan expérimental et du choix de la méthode d'analyse statistique

Design expérimental



Objectifs du plan expérimental

- Une étape importante/indispensable en amont des expériences
- Le plan d'expérience dépend:
 - du type d'information attendu (comparaison de groupes, évolution d'une pathologie)
 - des différentes sources de variabilité connues
- Permet de
 - réduire les sources de confusion/de biais
 - contrôler la variabilité technique

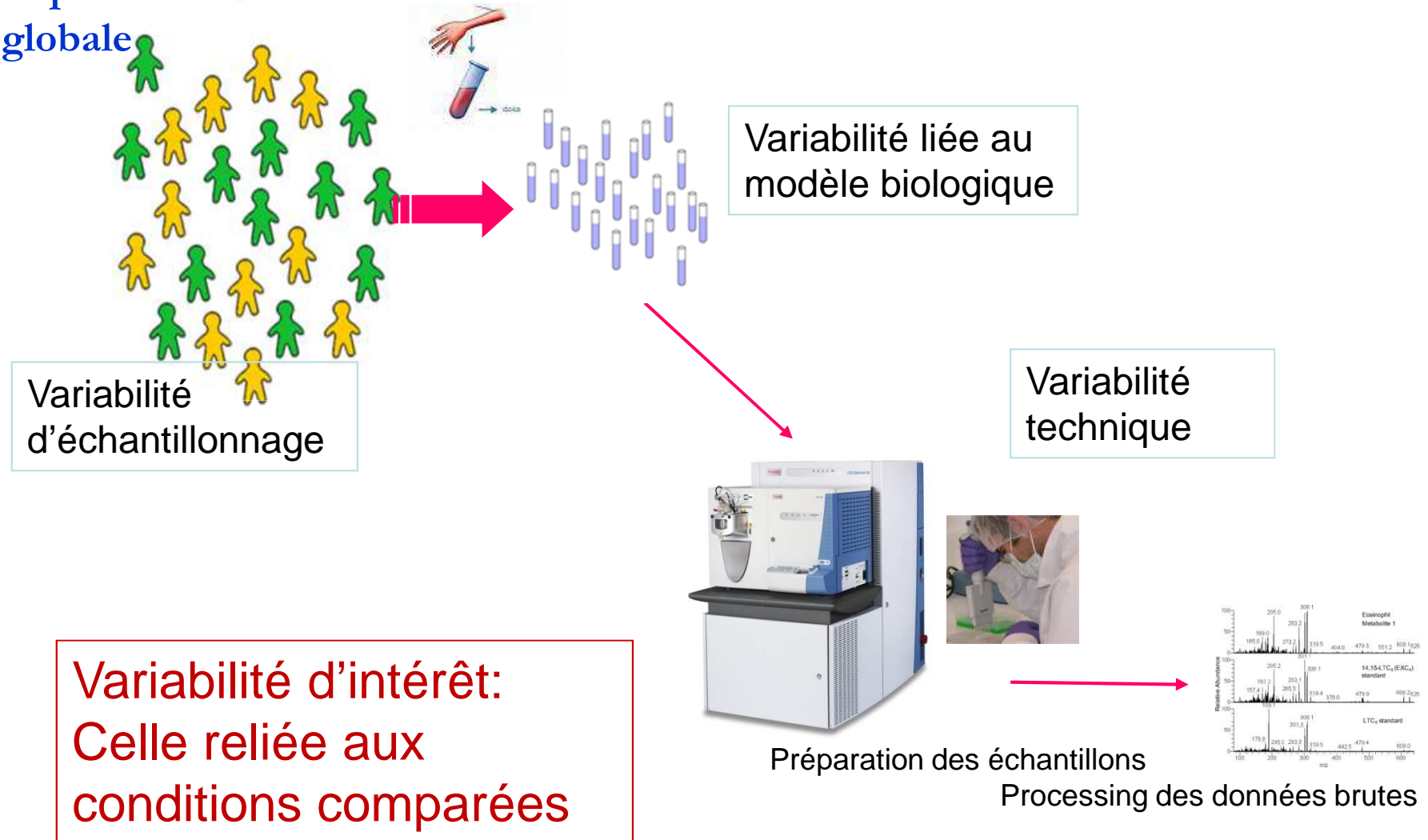
=> Obtention des informations biologiques les plus pertinentes possible

Différentes sources de variabilité

- Variabilité d'intérêt:
 - Reflète la différence entre les conditions comparées
 - Exemple: traitement/pas de traitement
 - ⇒ c'est elle que l'on cherche à mettre en évidence.

- Autres sources de variabilité
 - Variabilité biologique/d'échantillonnage
 - Variabilité technique
 - Variabilité liée au modèle biologique
 - ⇒ doivent être inférieures à la variabilité d'intérêt

Illustration

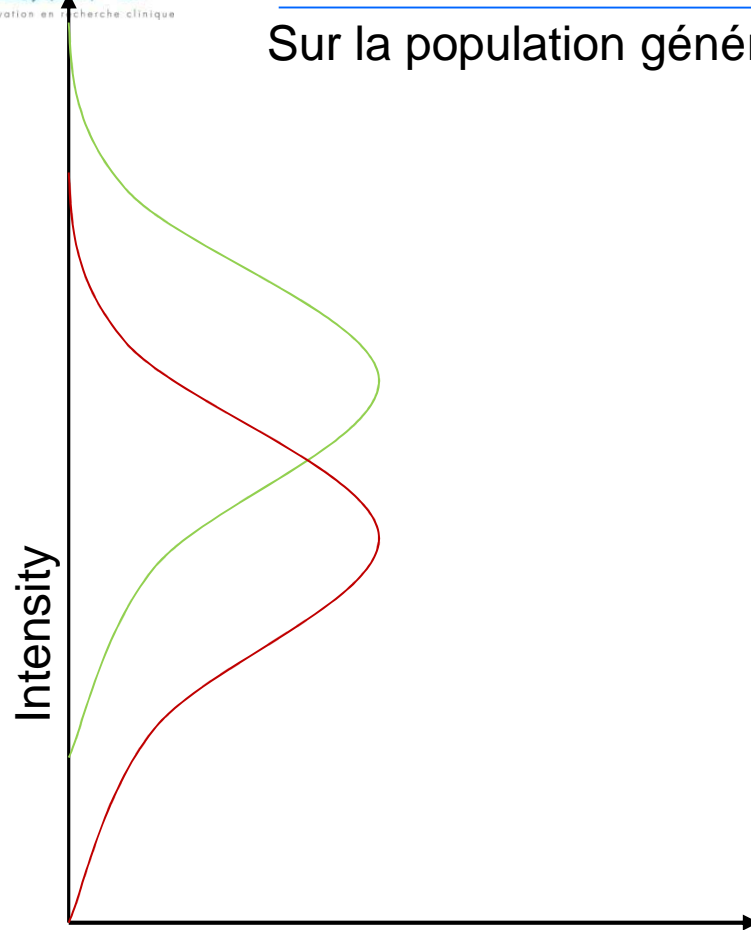


Variabilité d'échantillonnage

- Impossibilité de travailler sur la population globale
 - ⇒ Echantillonnage de la population
 - ⇒ Sélection optimale d'un sous-ensemble de la population qui soit représentatif de la population générale
 - Variabilité inter-individus :
 - Chaque individu a ses particularités
 - Seules les répétitions biologiques apportent de l'information sur cette variabilité inter-individus
 - Importance des répétitions biologiques
- ⇒ Question de **l'inférence**: les effets observés doivent être représentatifs du « véritable » effet relié à la condition étudiée, et non uniquement de l'échantillon (ie observés « par chance »)

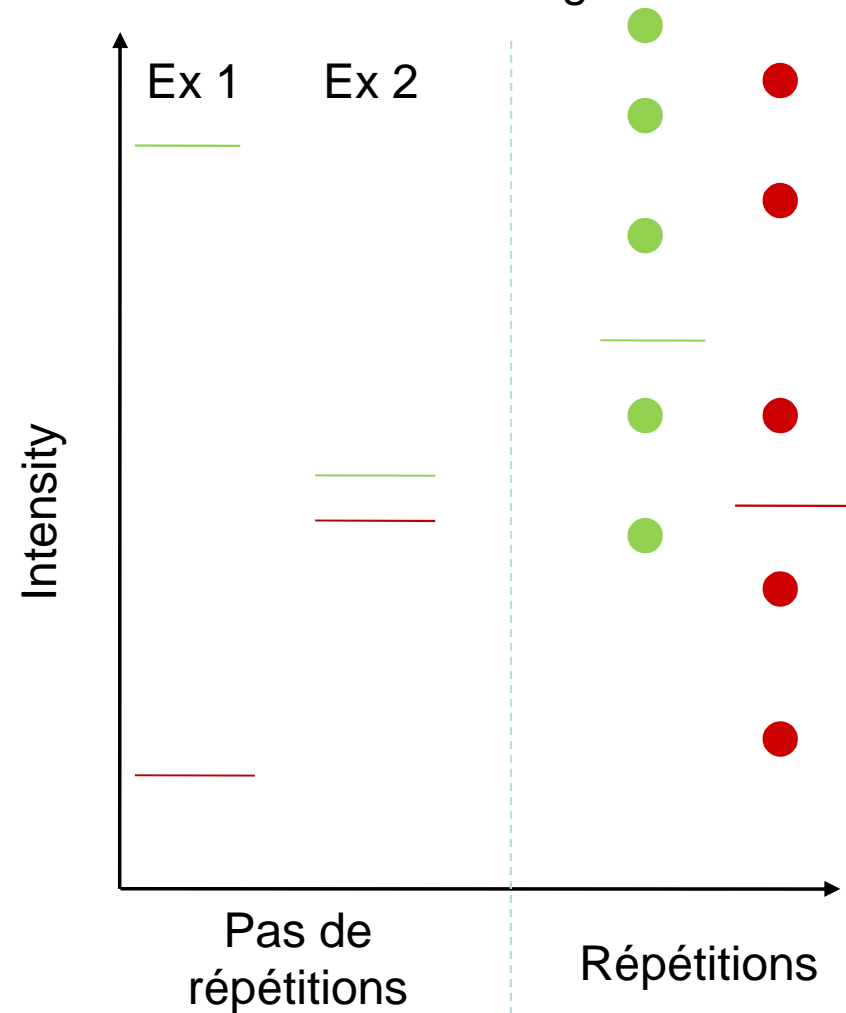
Illustration

Sur la population générale



Distribution de l'intensité d'une protéine dans 2 groupes biologiques

Deux stratégies d'échantillonnage différentes



Variabilité technique

- Tout ce qui n'est pas relié à la variabilité biologique
- Exemples
 - Préparation des échantillons
 - Calibration de l'instrument
 - Colonne de chromatographie
 - Détection des peptides
- ⇒ Doit être inférieure à la variabilité clinique d'intérêt
- ⇒ Ne doit pas être confondue avec la variabilité clinique (plan d'expérience)
- Répétitions techniques requises pour
 - Contrôler la reproductibilité de l'étude
 - Evaluer la qualité de l'expérience

Prise en compte de ces variabilités

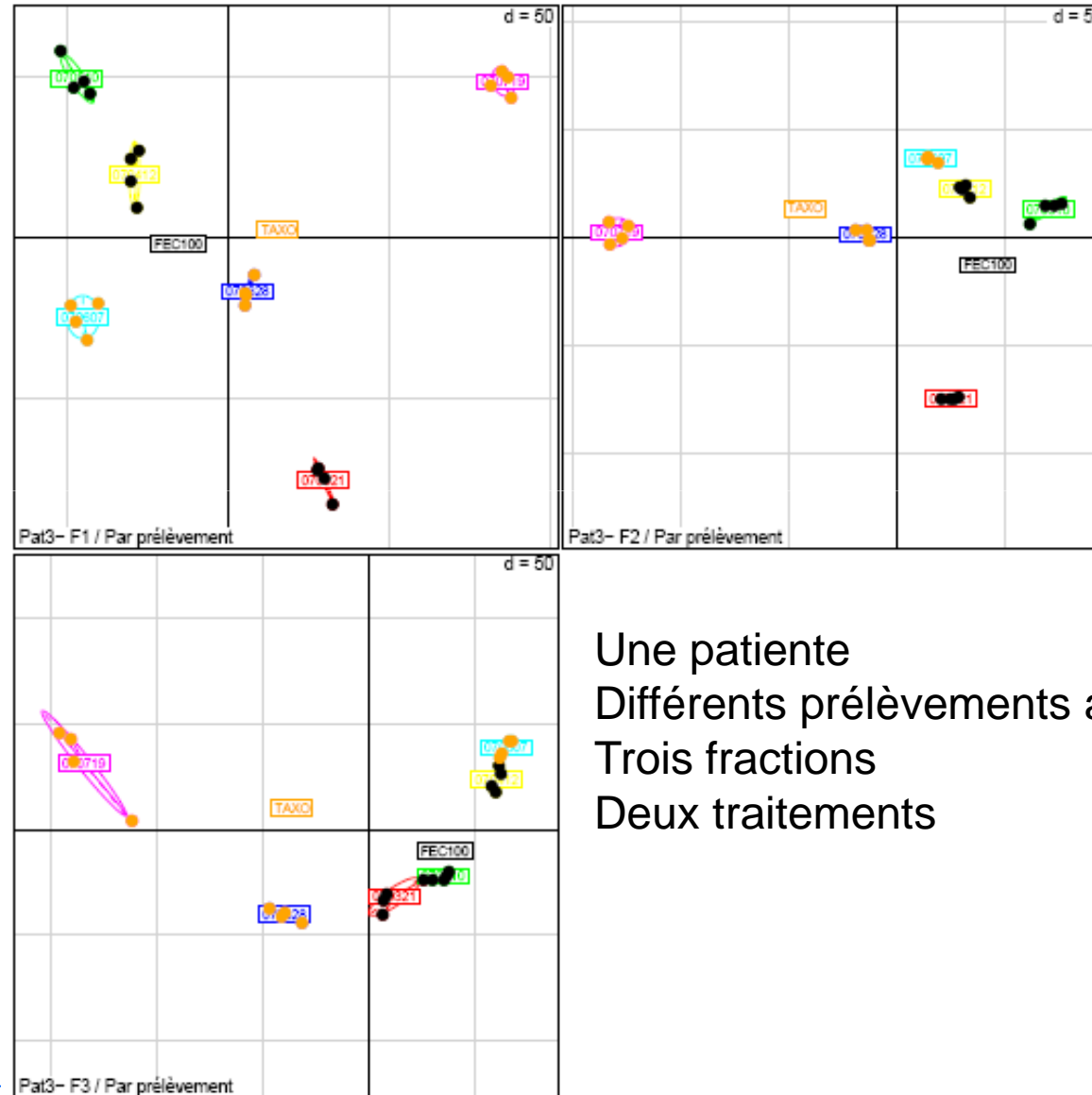
- **Randomisation**
 - Éviter les biais dus à des sources de variabilité non contrôlées
 - Ordre aléatoire du passage des échantillons

- **Stratégie “par blocs”**
 - Éviter les biais dus à des sources de variabilité connues et non désirées
 - Répartition équilibrée des échantillons d'un bloc à l'autre
 - Ex: tous les prélèvements d'un patient un même jour

Variabilité liée au modèle biologique

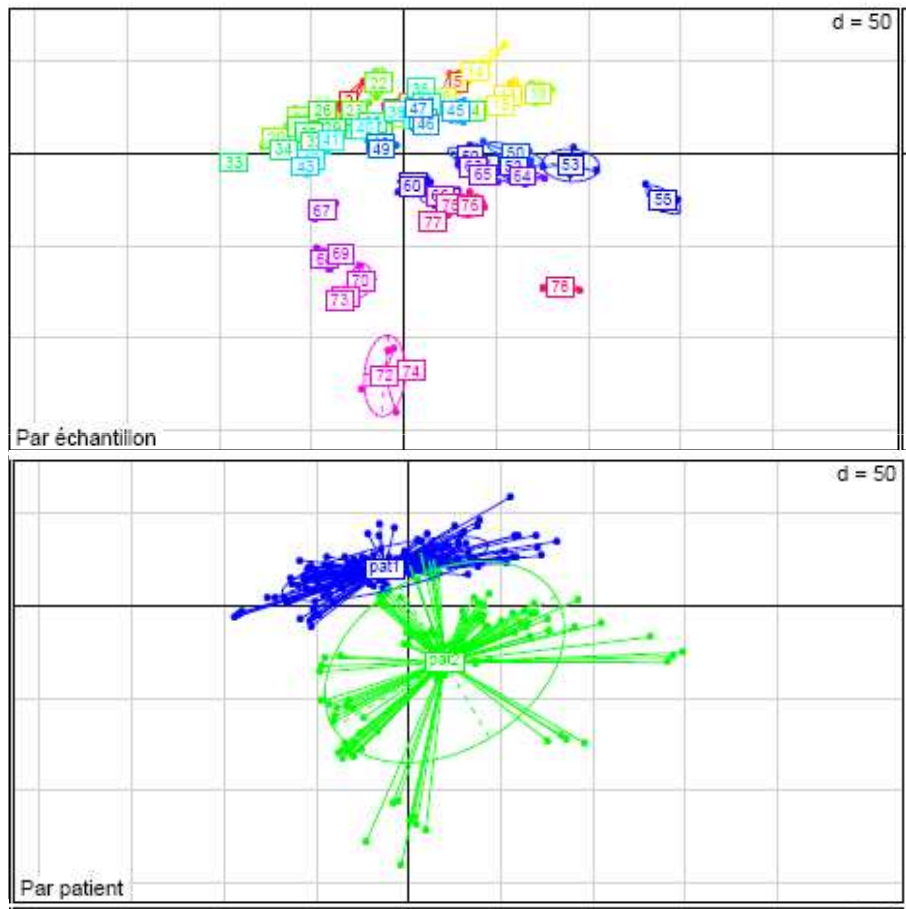
- Choix du modèle biologique aussi important que l'introduction de répétitions techniques et biologiques
- Si le modèle biologique n'est pas adapté, optimiser le nombre de répétitions techniques et biologiques n'aura pas d'incidence
- Adéquation des modèles à la variabilité biologique dans les études de sélection de biomarqueurs

Illustration de la variabilité - I



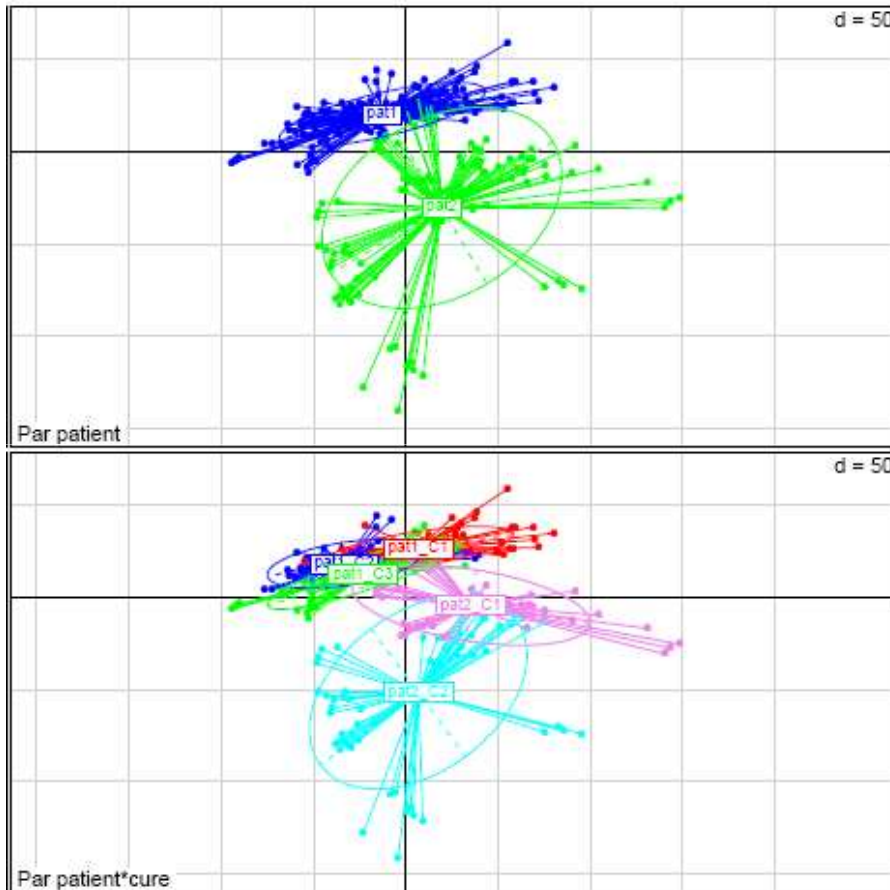
Une patiente
 Différents prélèvements au cours du temps
 Trois fractions
 Deux traitements

Illustration de la variabilité - II



Deux patients
Différents prélèvements au cours du temps

Illustration de la variabilité - III

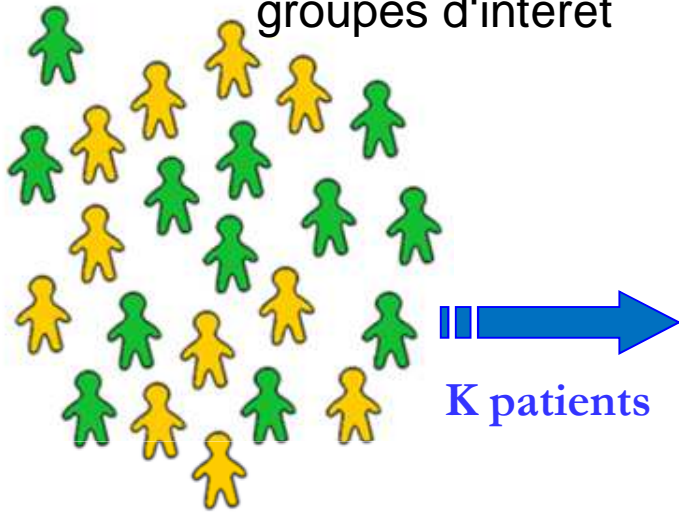


Deux patients
Différents prélèvements au cours du temps
Différentes cures d'aplasie

- **Variabilité d'échantillonnage**
 - Optimisation de la stratégie d'échantillonnage
 - **Variabilité technique:**
 - Peut être contrôlée par les répétitions techniques
 - Peut compenser le manque de répétitions biologiques
 - **Modèle biologique:**
 - Doit être adapté au contexte clinique/biologique de l'étude
- ⇒ Variabilité d'intérêt: doit être la plus grande possible
- ⇒ Autres sources de variabilité - d'échantillonnage, technique, du modèle biologique: doivent être les plus faibles possible
- ⇒ Nécessité d'études préliminaires pour optimiser ces différents paramètres de manière à mettre en évidence la variabilité d'intérêt

Calcul du nombre de répétitions techniques et biologiques

Variabilité entre les groupes d'intérêt



Variabilité liée aux répétitions biologiques

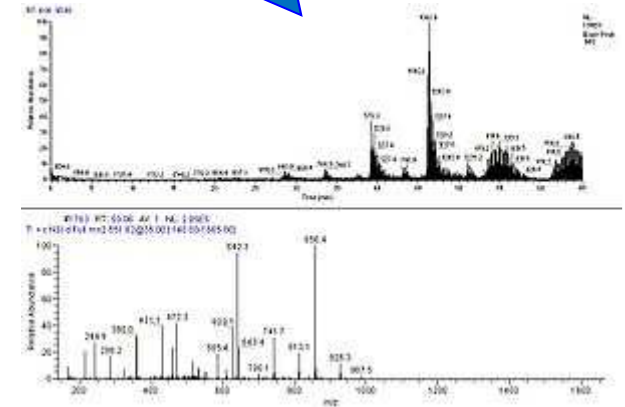
Variabilité liée aux répétitions techniques



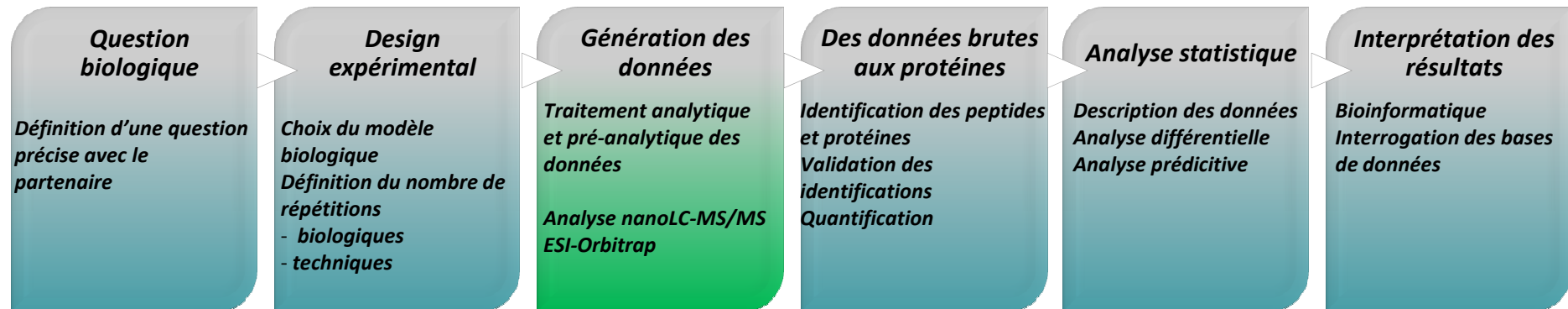
L répétitions techniques

I peptides par protéine

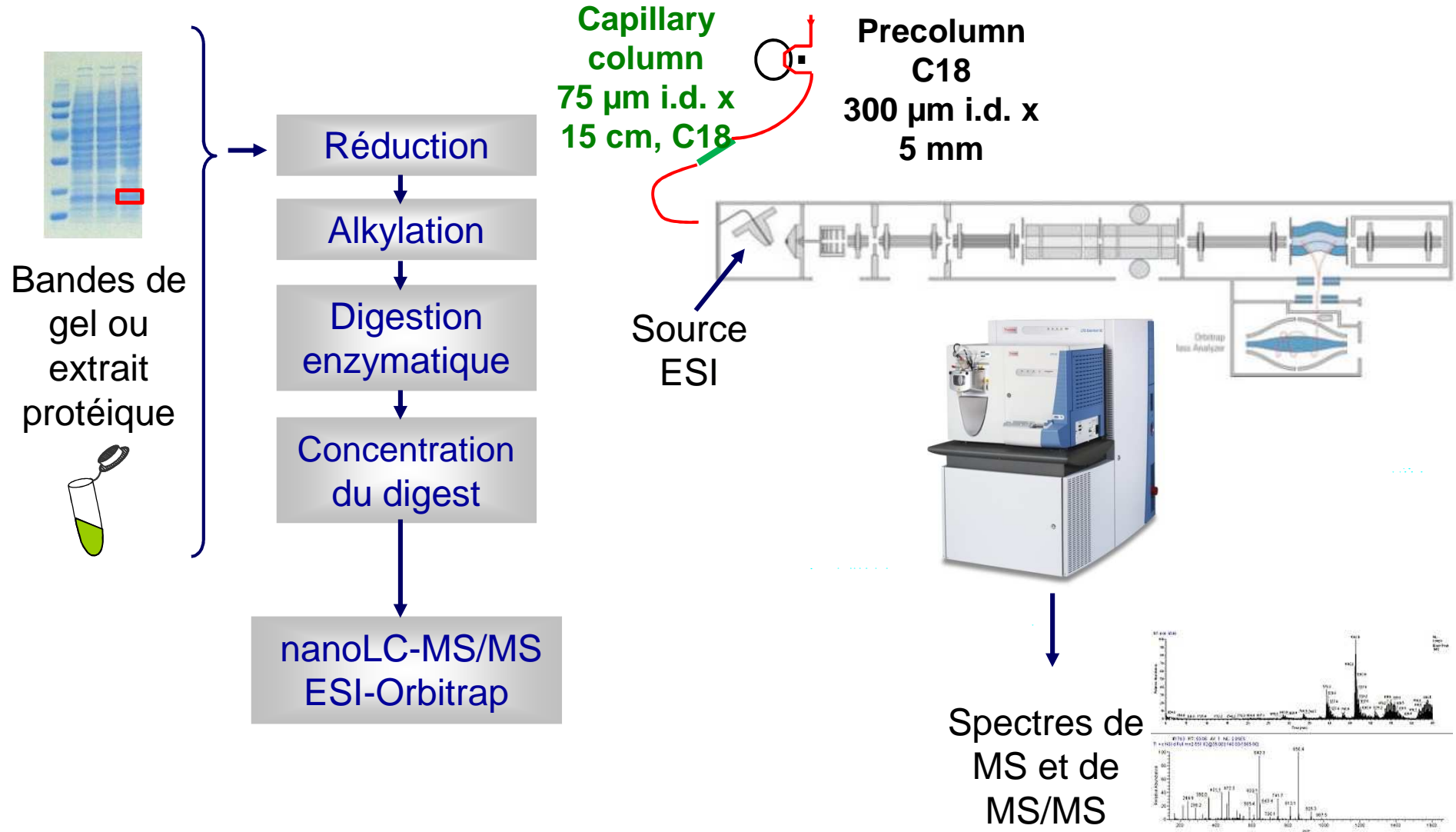
Variabilité au niveau des peptides



Génération des données



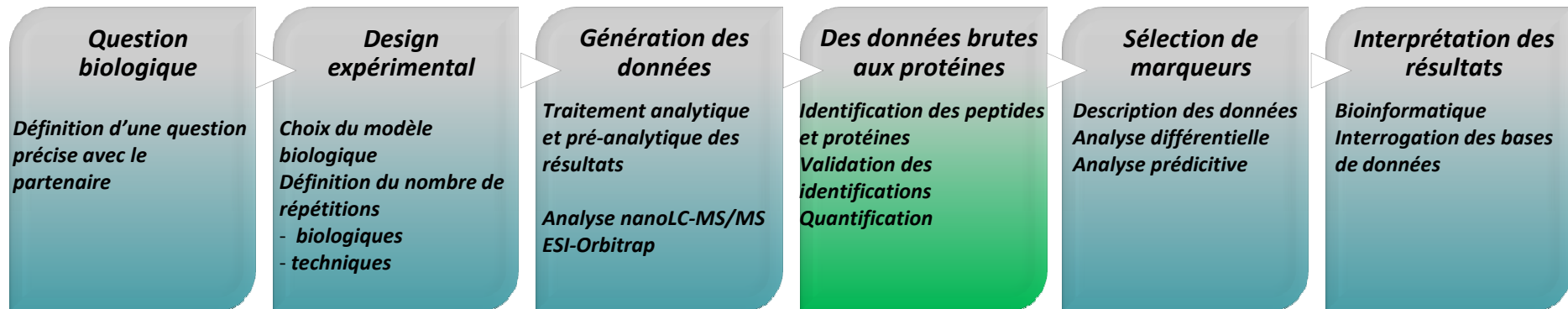
Analyse par nano-LC/MS-MS



Contrôles analytiques

- Répétitions techniques: 3 injections par échantillon
 - Variabilité aux différentes étapes de préparation des échantillons
 - Rappel: en l'absence de répétitions techniques, la variabilité biologique se mélange avec la variabilité technique
- Introduction de blancs (acide formique 0.1%)
 - Contrôle de la contamination d'un échantillon à l'autre dans la colonne
- Injection de BSA (Bovin Serum Albumin)
 - Contrôle de la dérive au niveau de la chromatographie

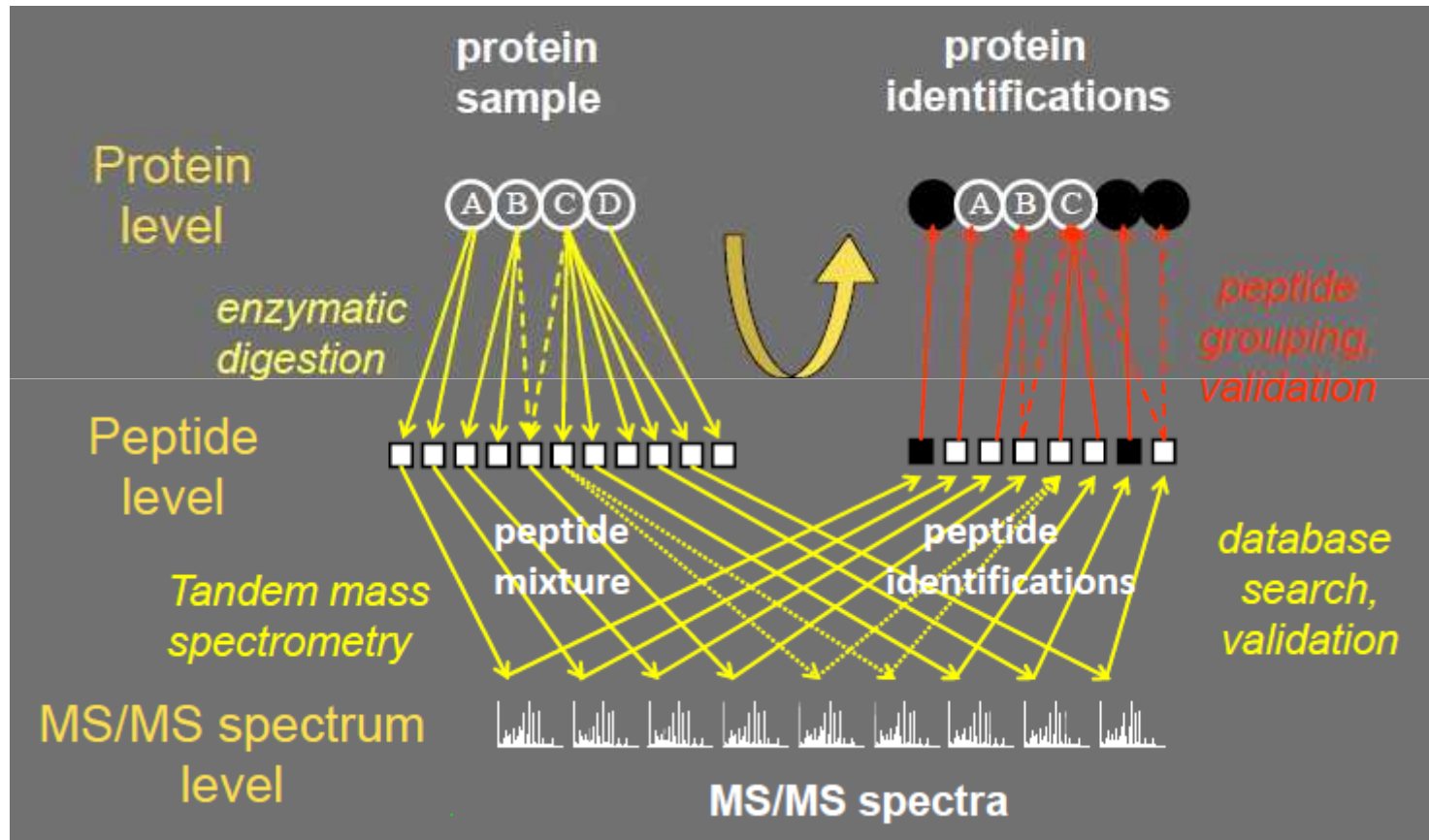
Des données brutes aux protéines



Principes

- Étapes
 - Conversion des données du format constructeur à un format lisible par les logiciels open-source (mz(X)ML)
 - Identification des peptides
 - Validation des identifications peptidiques
 - Sélection des protéines
 - Quantification des protéines en vue de l'analyse statistique
- Outils utilisés (à titre indicatif)
 - Msconvert <http://proteowizard.sourceforge.net/tools/msconvert.html>
 - Trans Proteomic Pipeline <http://www.proteomecenter.org/software.php>
 - Rgui

Des peptides aux protéines



Identification et validation des peptides

- Utilisation des moteurs de recherche
 - Sequest, Comet, !X Tandem, Mascot, etc...
 - Obtention de scores par les moteurs de recherche
 - Question du choix du seuil/du score à utiliser?
 - **PeptideProphet (TPP)**
 - Validation des résultats obtenus par les moteurs de recherche
 - Utilisation des scores pour calculer la probabilité qu'un peptide soit bien identifié
 - Intégration des propriétés des peptides pour affiner les probabilités (missed cleavage, nombre de terminaisons tryptiques, etc...)
- ⇒ Probabilité associée à chaque peptide identifié

Ma *et al.* *BMC Bioinformatics* 2012; **13**(Suppl 16):S1

Deutsch *et al.* *Proteomics* 2011; **10** (6): 1150-1159

PeptideProphet

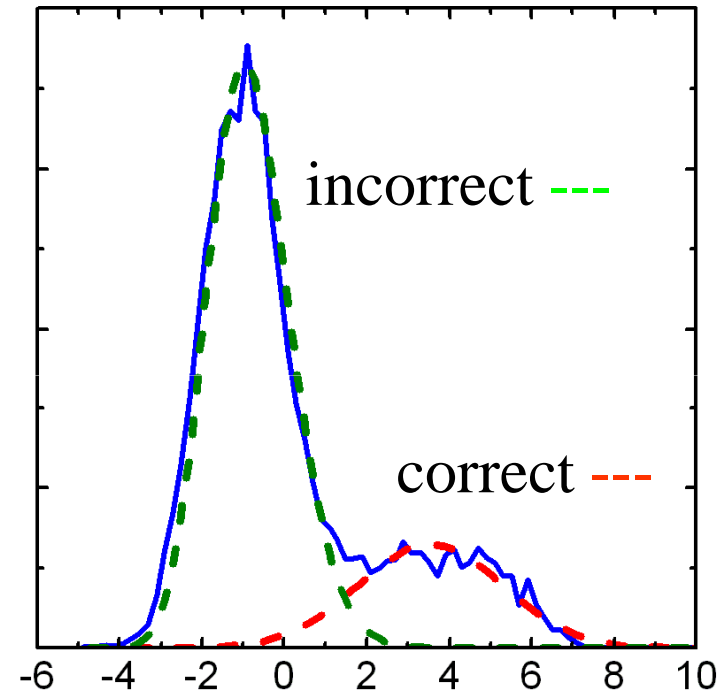
Jeu de données

	Spectrum	Peptide	score
1	ISLLDAQSAPLR		4.5
2	VVEELCTPEGK		3.9
3	DLLLQWCWENGK		1.2
4	ECDVVSNTIIAEK		0.9
5	GDAVFVIDALNR		3.6
	...		
M	SYLFCMEAEK		1.1

1.00
0.99
0.11
0.00
0.87
...
0.02

probability

number of spectra

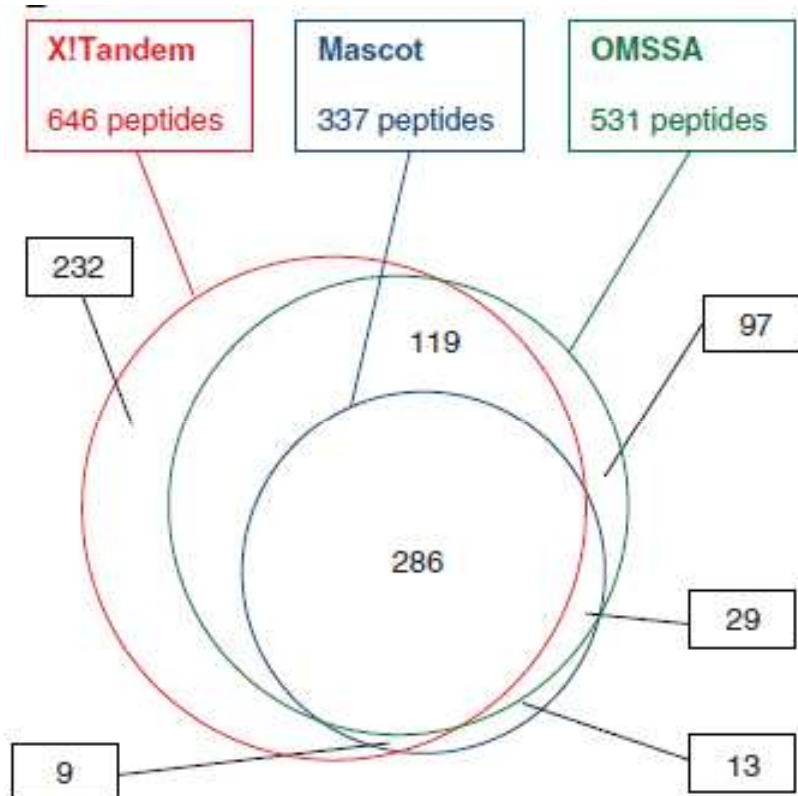


database search score

Shteynberg, 2014

Introduction de bases decoy

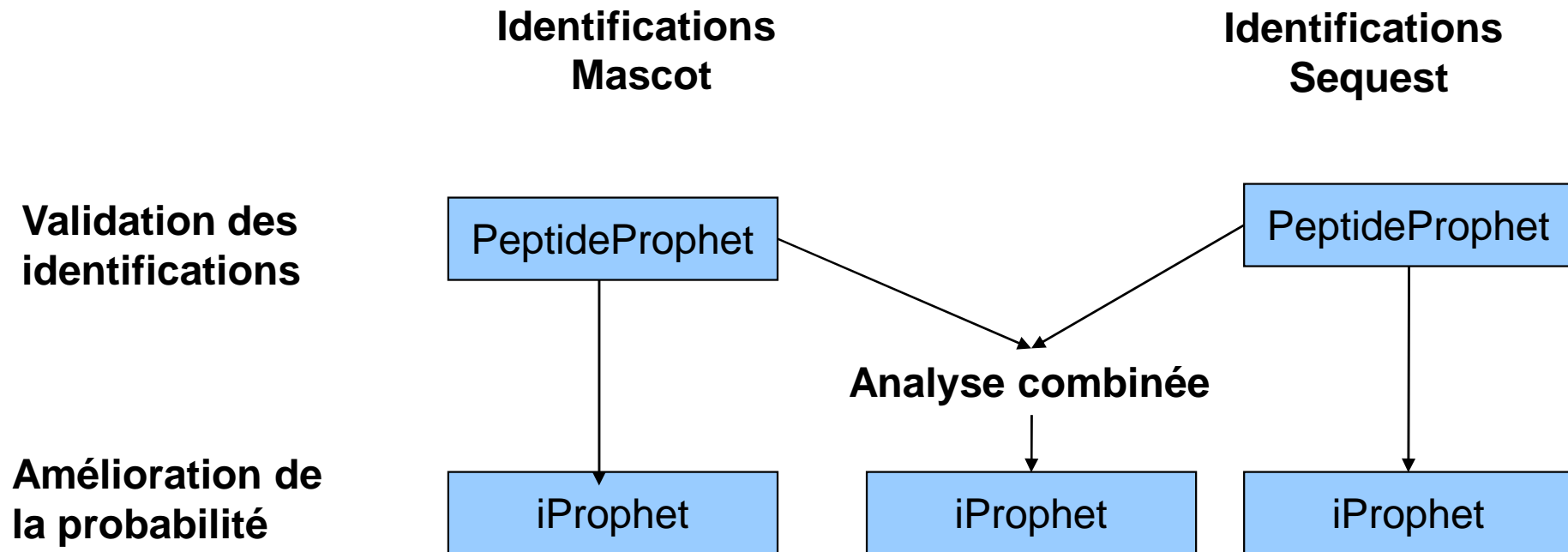
Rôle du moteur de recherche



Jeu de données UPS/48 protéines
FDR 1%

Vaudel et al. *Proteomics* 2011; **11**(5): 996-9

Pour aller plus loin...iProphet



Intérêts de Iprophet:

1. Combinaison des résultats obtenus par différents moteurs de recherche

⇒ Probabilités affinées

⇒ Augmentation du nombre d'identifications correctes par rapport à PeptideProphet (à FDR constant)

Intérêts de Iprophet:

2. Prise en compte d'informations complémentaires dans le calcul des probabilités

- Introduction de scores
- Exemples:
 - Nombre d'identifications communes à plusieurs bases
 - Probabilité des identifications
 - Prise en compte des charges
 - Prise en compte des modifications

iProphet: Improved Analysis of Shotgun Proteomic Data

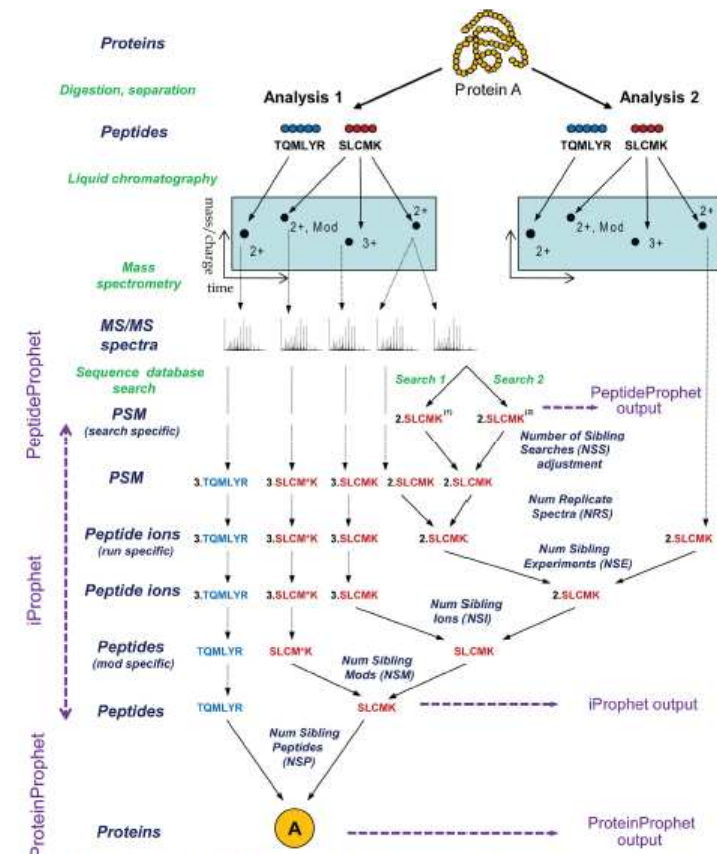


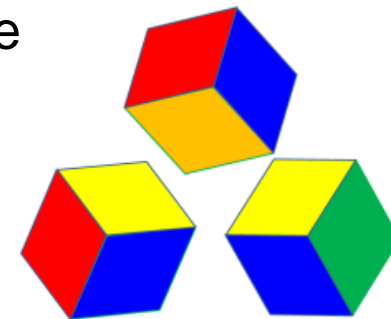
Fig. 1. Overview of shotgun proteomic data and the computational strategy. The protein sample is digested into peptides, with some peptides present in the unmodified and a modified (e.g. oxidized methionine) forms. The peptide sample is separated using liquid chromatography (LC) coupled online with a tandem mass spectrometer. The first stage of MS measures mass to charge ratios of peptide ions injected in the instrument at any given time. A peptide can be ionized into multiple peptide precursor ions having different charge state (e.g. 2+ and 3+). Selected peptide ions are subjected to MS/MS sequencing (some multiple times). Each acquired MS/MS spectrum is assigned a best matching peptide sequence using sequence database searching. When multiple search tools are applied in parallel (Search 1 and Search 2), each spectrum produces multiple peptide to spectrum matches (search-specific PSM level), which could be the same or different peptides summarized at the PSM level. Within the same LC-MS/MS run, the same peptide ion can be identified from multiple PSMs (run-specific

Sélection des protéines

- Utilisation de **ProteinProphet**
- Calcul de la probabilité qu'une protéine soit présente dans l'échantillon en se basant sur les probabilités ajustées des peptides.
 - Probabilité pour une protéine = probabilité qu'au moins un PSM correspondant à la protéine soit correct
 - Ajustement des probabilités
 - Augmente la probabilité des peptides qui ont des « frères »
 - Pénalise la probabilité des peptides qui sont seuls à prédire une protéine
 - Crée la liste de protéines la plus simple qui permette d'expliquer les peptides présents dans l'échantillon.
- Calcul d'une probabilité pour
 - chaque protéine
 - ou groupe de protéines pour les protéines « indifférentiables »

Nesvizhskii *et al.*, *Anal. Chem* 2003; 75(17):4646-58

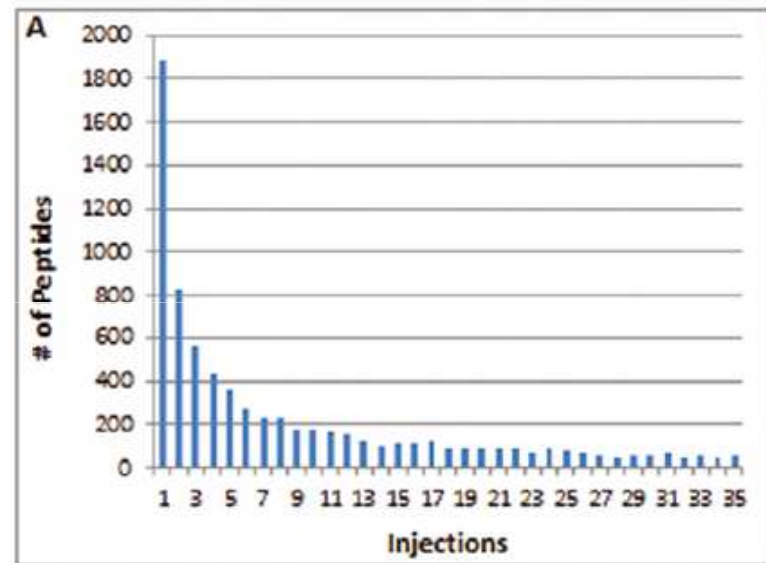
- Tous les outils mis en place par la suite dépendent de la liste de peptides/protéines
 - Résultats en partie différents d'un moteur de recherche à l'autre
 - Partenaires perdus sur le fait qu'on ne retrouve pas forcément les mêmes résultats
 - Intérêt de combiner les résultats (iProphet)
 - Consolidation des données/marqueurs qui seront plus robustes même si on en retient moins
- ⇒ Compromis à trouver avec le partenaire dans le choix des seuils de significativité au moment de la validation des ide



- Echantillons souvent très riches
 - Seuls les peptides les plus abondants sont retenus pour la MS/MS
 - Un même peptide n'est pas forcément identifié dans chacune des injections.
 - Selon l'injection, un peptide peut avoir une intensité plus ou moins importante en fonction des peptides avec lesquels il est coélué
- ⇒ Ce ne sont pas les mêmes peptides qui vont être identifiés

Illustration

- Exemple: 35 injections d'une digestion de lysat de foie de rat par la trypsine
- 7361 peptides générés
- 25,6% sont identifiés dans au moins une injection
- Seuls 0,8% sont identifiés dans les 35 injections! (61 peptides)



Distribution des identifications peptidiques.

Lai et al *J Proteom Res* 2011; **10**(10): 4799–4812

Remarques

- Ce n'est pas parce qu'un peptide n'est pas identifié qu'il n'est pas présent dans l'échantillon.
 - Un peptide identifié dans une seule injection peut en réalité être présent dans les autres échantillons.
- Par contre si un peptide est identifié avec peu de certitude, chercher à le quantifier dans les autres échantillons peut conduire à une quantification erronée.
 - => Plus l'identification est certaine, plus on a de chance qu'en réalité le peptide soit présent dans les autres échantillons.

Question des valeurs manquantes

- Origine des valeurs manquantes pas bien maîtrisée:
 - Hasard de l'expérience/technique
 - « Vraie » information biologique: PTM, variation au niveau de la séquence, clivage enzymatique incomplet, etc...
- Différentes explications
 - Le peptide est absent de l'échantillon étudié
 - Le peptide est présent à une abondance que l'appareil devrait pouvoir mesurer mais il n'est pas détecté ou mal identifié
 - Le peptide est présent mais en quantité trop faible pour être détecté par l'appareil (limite de détection)
 - Gamme dynamique de détection des instruments .
- Pas d'identification ↔ pas de quantification?

Solutions proposées

- Retrait du peptide de l'ensemble des données
 - Hypothèse d'un problème de qualité de la mesure
- Imputation de la valeur:
 - Utilisation du bruit de fond
 - Minimum observé sur les autres injections issus de la même conclusion
 - Hypothèse d'une valeur manquante car peptide trop peu abondant pour être mesuré
 - KNN (Plus proches voisins)
 - Limite: modification de la moyenne et de la structure de variance des peptides
- Modélisation de la quantité de peptides (distribution normale)
- Transformation présence/absence
 - Pas de quantification
- Filtre sur les peptides pour limiter la proportion de valeurs manquantes

⇒ **Pas de consensus**

Exemple - Filtre sur les peptides

- Lai *et al.* J of Proteome Res. (2011)
- Calcul de la fréquence d'apparition des peptides sur l'ensemble des répétitions techniques pour chacune des répétitions biologiques
 - Conservation des peptides qui sont détectés dans 2/3 des répétitions techniques
 - Gestion des valeurs manquantes par élimination des peptides trop peu représentés sur l'ensemble des injections
- Résultat: une liste peptides identifiés chacun par sa séquence, l'ID de la protéine à laquelle il est associé, sa charge, son temps de rétention et sa masse/charge

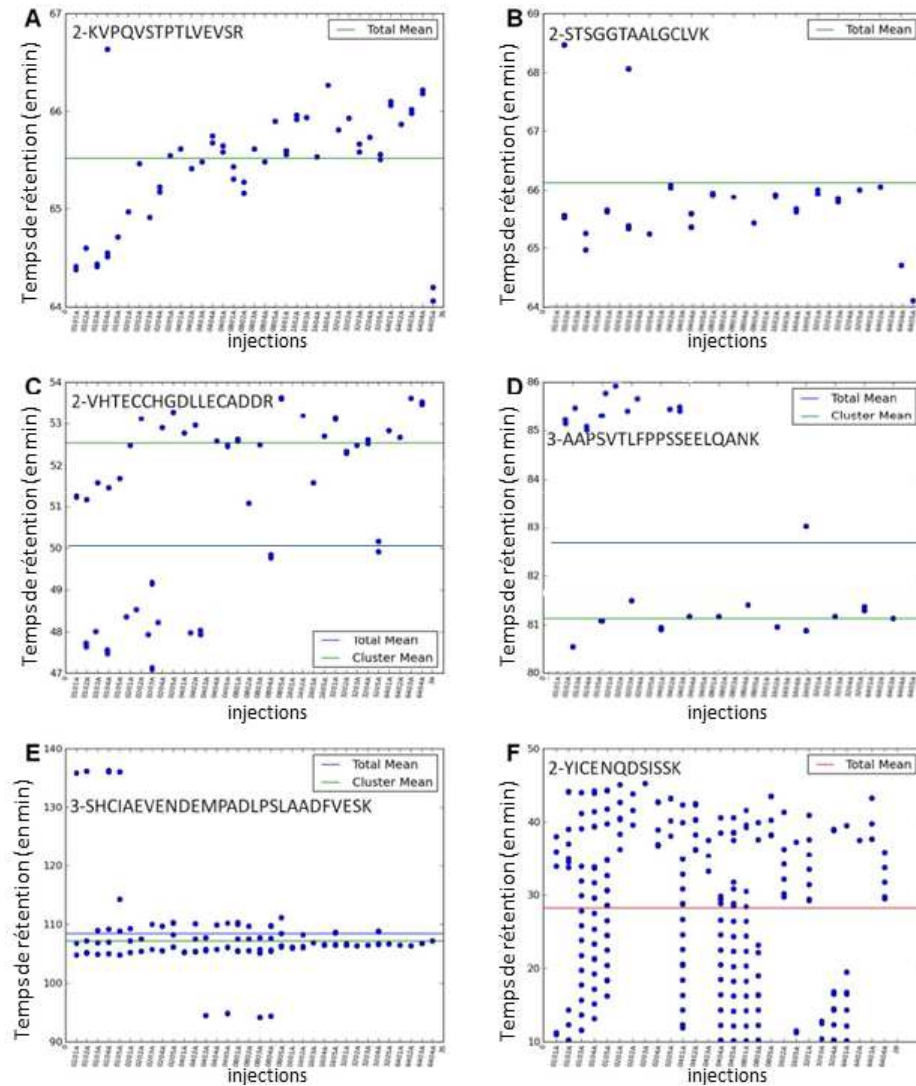
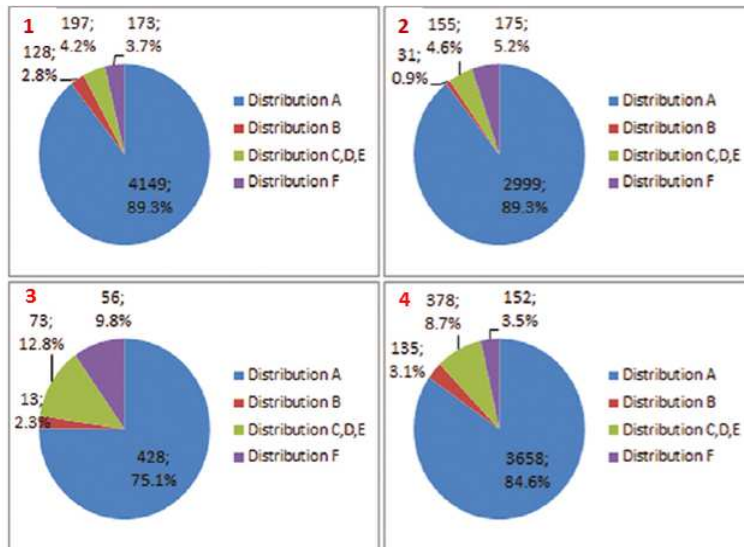
Question de l'alignement

- Mise en relation des peptides identifiés sur les différentes injections
 - Hypothèse: les temps d'élution d'un peptide et les comportements d'ionisation sont relativement constants d'une mesure à l'autre malgré l'aléatoire d'ionisation
 - Les m/z restent relativement constants
 - Shifts plus importants au niveau des temps de rétention (RT).⇒ Nécessité d'aligner les RT
- ⇒ Alignement indispensable pour pouvoir ensuite comparer les quantifications obtenues d'un run à l'autre

Illustration

Les peptides n'ont pas les mêmes pattern d'élution d'un run à l'autre.

- Définition de 6 pattern d'élution différents
- Lai *et al.* J. Proteome Res. 2011



Solutions proposées

- Estimation d'une fonction linéaire ou non qui corrige la distorsion observée d'un run à l'autre
- Différentes stratégies
 - Utilisation ou non d'un run de référence
 - Alignements successifs par paires ou alignement simultané de tous les spectres
 - Utilisation des profils complets ou uniquement des entités « MZ/RT » selon que la détection des entités est faite avant ou après l'alignement

⇒ À nouveau, pas de consensus...

Alignement – exemple I

Lai *et al.* *J. Proteome Res.* 2011; **10**(10):4799–4812

- Basé sur l'observation de différents patterns d'élution et la construction de clusters
- Pour chaque peptide
 - RT <3 min → moyenne pondérée
 - RT >3min → clustering (distance euclidienne, saut minimum) → Moyenne pondérée dans chacun des clusters
 - RT trop dispersés → le peptide est retiré de l'analyse.
 - Moyenne pondérée des MZ après détermination du temps de rétention

Question de la quantification

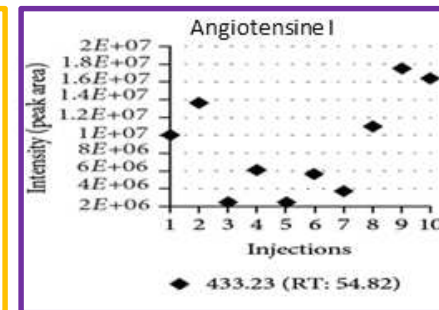
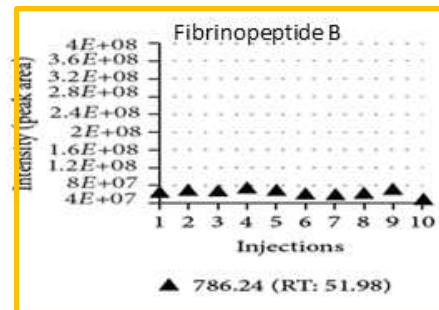
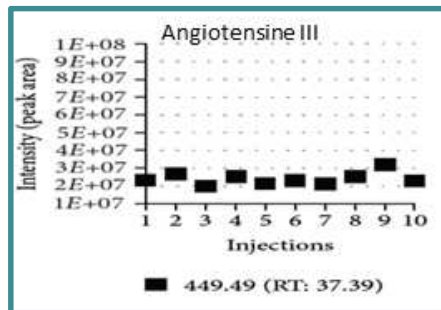
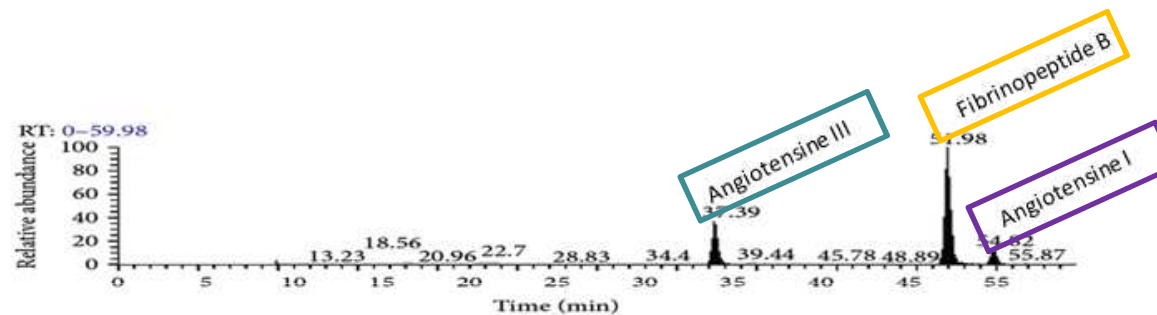
- Quantification relative
 - Abondances comparables d'une condition à l'autre mais pas d'une protéine à l'autre (en opposition à la quantification absolue)
- Abondance des peptides identifiés au niveau MS
 - On se base sur l'intensité des ions d'un peptide donné sur son profil d'élution issu du chromatogramme
 - Comparaison des abondances des protéines en se basant sur les peptides qui y sont rattachés
- Spectral counting : on se base sur le nombre d'identifications MS/MS assignées à une même protéine.
 - Pas abordé ici

- Quantification des protéines plus précise que celle des peptides puisqu'elle est obtenue à partir de plusieurs représentants
 - En comparant les peptides 2 à 2, malgré l'alignement on peut ne pas avoir de chance et comparer 2 entités qui ne sont pas les mêmes.
- Inférence des protéines
 - Attention aux peptides outliers qui peuvent en réalité avoir un sens biologique
 - Attention aux protéines identifiées uniquement par un peptide => plus de risque que ce soit une mauvaise identification

- Pour une même protéine, les fold-changes peuvent être différents d'un peptide à l'autre
 - Certains peptides varient plus que d'autres dans les mêmes conditions de chromatographie
 - Différences liées aux PTM
 - Peptides partagés par plusieurs protéines
 - Isoformes, mauvaise identification.
- ⇒ **Conclusion:**
- Tous les peptides ne méritent pas d'être utilisés pour la quantification des protéines
 - Retrait des peptides mal quantifiés = filtre sur les peptides
 - L'étape de filtre sur les peptides a plus d'influence que le choix de la méthode de quantification (Matzke *et al. Proteomics* 2013; **13**(3-4):493-503)

Illustration

Les CV basés sur les répétitions techniques sont très variables d'un peptide à l'autre



Intensités mesurées pour 3 peptides et pour 10 injections de la même solution

Lai *et al.* *International J Proteomics*. 2013; Vol 2013 (2013), Article ID 756039

Filtre sur les CV des peptides

- Retrait des peptides en fonction du coefficient de variation calculé sur les répétitions techniques
- Exemple (Lai *et al. J Proteome Res.* 2011):
 - Répartition en 4 catégories
 - Cat 1: CV inacceptable: $>116\%$
 - Cat 2: CV élevé: $71\% < CV \leq 116\%$
 - Cat 3: CV moyen: $47\% < CV \leq 71\%$
 - Cat 4: CV faible: $< 47\%$
 - Calcul de la fréquence de chaque catégorie sur l'ensemble des répétitions biologiques
 - 100% Cat 1 – peptide éliminé
 - Fréquence Cat 2 $> 12.5\%$ - peptide éliminé
 - Fréquence Cat 4 $< 12.5\%$ - peptide éliminé

⇒ 3 stratégies

1. Additive: les abondances des peptides sont combinées de manière additive
 - Somme, somme standardisée, somme/moyenne des peptides les plus intenses
 - Exemple: somme ou moyenne des 3 peptides les plus intenses
2. Référence: un peptide est choisi comme référence pour standardiser les abondances des autres peptides de la protéine.
 - Différents choix de référence:
 - Celui qui a le moins de valeurs manquantes. C'est ensuite la médiane des peptides normalisés qui est utilisée pour représenter la protéine
 - Centrage-réduction des peptides (médiane/sd) puis médiane de ces valeurs
 - Implémenté dans le logiciel DAnTE

Polpitiya et al. Bioinformatics 2008; 24: 1556-1558

3. Modèle linéaire: modélisation de l'abondance des protéines

- Les peptides sont considérés comme des mesures répétées de la protéine
- Modèle additif sans interaction:

$$y_{ijkl} = prot_i + pep_{ij} + gpe_{ijk} + error_{ijkl}$$

où y_{ijkl} est l'abondance de la protéine i et du peptide j dans le groupe k pour l'échantillon l

- Exemple: DanteR, Karpievitch *et al.* *Bioinformatics* 2009; **25**:2573-2580

- Modèle additif avec interaction:

$$y_{ijk} = pep_i + gpe_j + (pep * gpe)_{ij} + S_k + error_{ijk} \text{ puis } \bar{y}_{i.k}$$

où y_{ijk} est l'abondance pour le peptide i dans le groupe j pour la répétition biologique k

- Exemple: Msstats, Clough *et al.* *J of Proteom Res* 2009; **8**:5275–5284

Choix d'une méthode optimale?

- Comparaison des différentes stratégies par Matzke *et al. Proteomics* (2013)
- Le choix de la méthode de quantification est moins important que:
 - Le choix des filtres sur les peptides
 - Le choix du traitement des valeurs manquantes
- Une fois de plus, pas de consensus
 - Meilleure solution: solution maîtrisée et « simple » d'utilisation
 - Conseil des auteurs: package R Msstats (www.msstats.org/)
- Remarque: travail sur le logarithme en base 2

Protéine ou peptide?

- Autre solution: pas de quantification des protéines mais travail directement au niveau des peptides
- Inconvénients
 - Multiplication des tests
 - Corrélation entre les peptides d'une même protéine
 - Pas de quantification possible des protéines par sujet
- Intérêt: information sur les modifications post-traductionnelles
- Dépend de la question posée par le partenaire, en fonction de ses connaissances biologiques

- Objectif : retrait de tout ce qui n'est pas de la variabilité biologique
- Biais aléatoires:
 - traitement des échantillons, calibration des instruments, colonnes de chromatographie, changements de température
- Biais systématiques:
 - temps de rétention LC, fluctuation au niveau des intensités des pics, précision de la mesure

Karpievitch *et al.* *BMC Bioinformatics*. 2012; **13**(Suppl 16):S5

Callister *et al.* *J Proteome Res* 2006; **5**:277–286

Solutions proposées

- Normalisation globale:

- Hypothèse: la majorité des peptides ne bouge pas donc la distribution des abondances devrait être la même en moyenne pour l'ensemble des échantillons
- On contraint la distribution des abondances à être centrée autour d'une constante: moyenne, médiane, etc...

$$y'_{ij} = y_{ij} - \mu_j$$

où y_{ij} est l'abondance observée pour le peptide i dans l'échantillon j et μ_j la moyenne (p.ex) de tous les peptides dans l'échantillon j

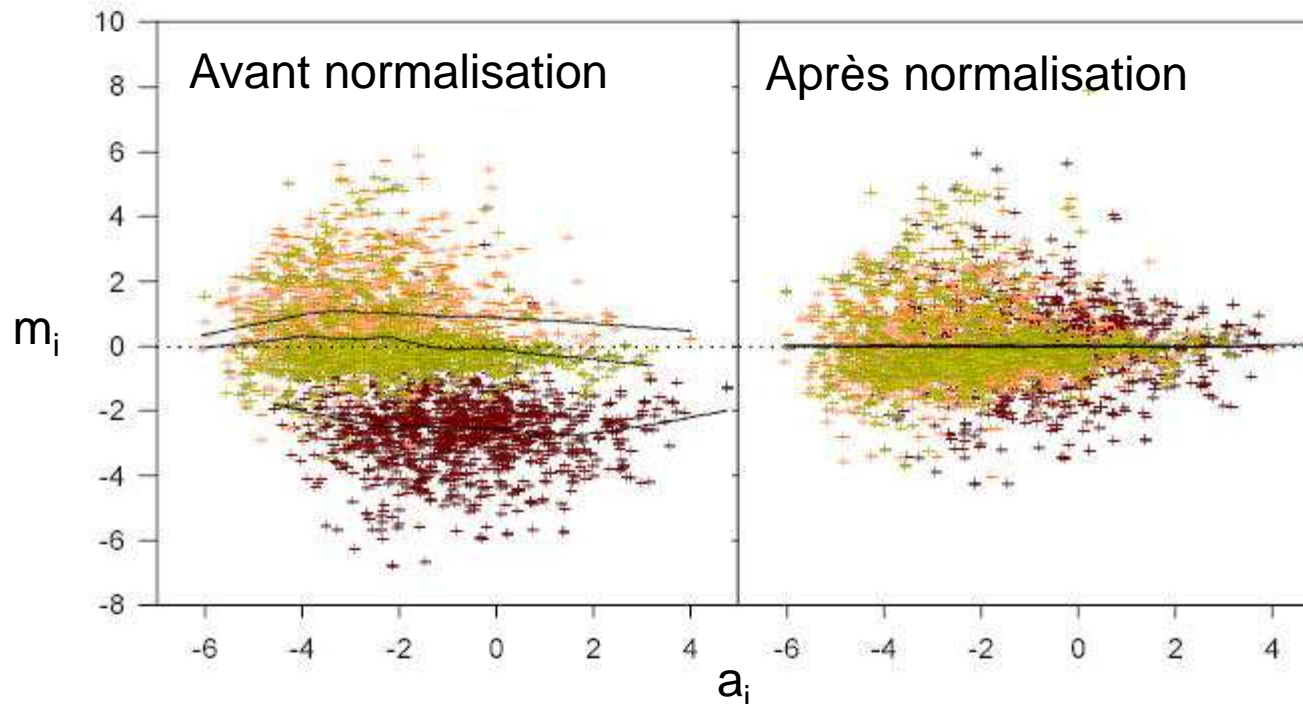
- Peut corriger les différences dans les quantités de matériel utilisé

- Normalisation quantile:

- Hypothèse : la distribution de l'abondance des peptides devrait être la même d'un échantillon à l'autre
- Rend les distributions des peptides exactement identiques d'un échantillon à l'autre

Normalisation sur spectre de référence I

- Utilisation d'un spectre de référence
- Analogie avec la normalisation utilisée pour les puces à ADN
 - Basé sur les graphes « MA »
 - M: différence des log; A: moyenne des log



- Régression linéaire:

- Hypothèse: le biais systématique est linéairement dépendant de l'ordre de grandeur de l'abondance des peptides
- Efficace pour retirer le biais systématique lié au « carry-over » sur la colonne LC
- Régression linéaire sur le graphe MA

$m'_i = m_i - m^*_i$ où m^*_i est la prédiction fournie par le modèle de régression linéaire

- Régression locale :

- Hypothèse: le biais systématique est non linéairement dépendant de l'ordre de grandeur de l'abondance des peptides
- Efficace pour retirer le biais systématique généré par des peptides qui sont mesurés proche du seuil de saturation ou du bruit de fond
- Même principe que pour la régression linéaire mais “par morceaux”

Normalisation – décomposition en valeurs singulières

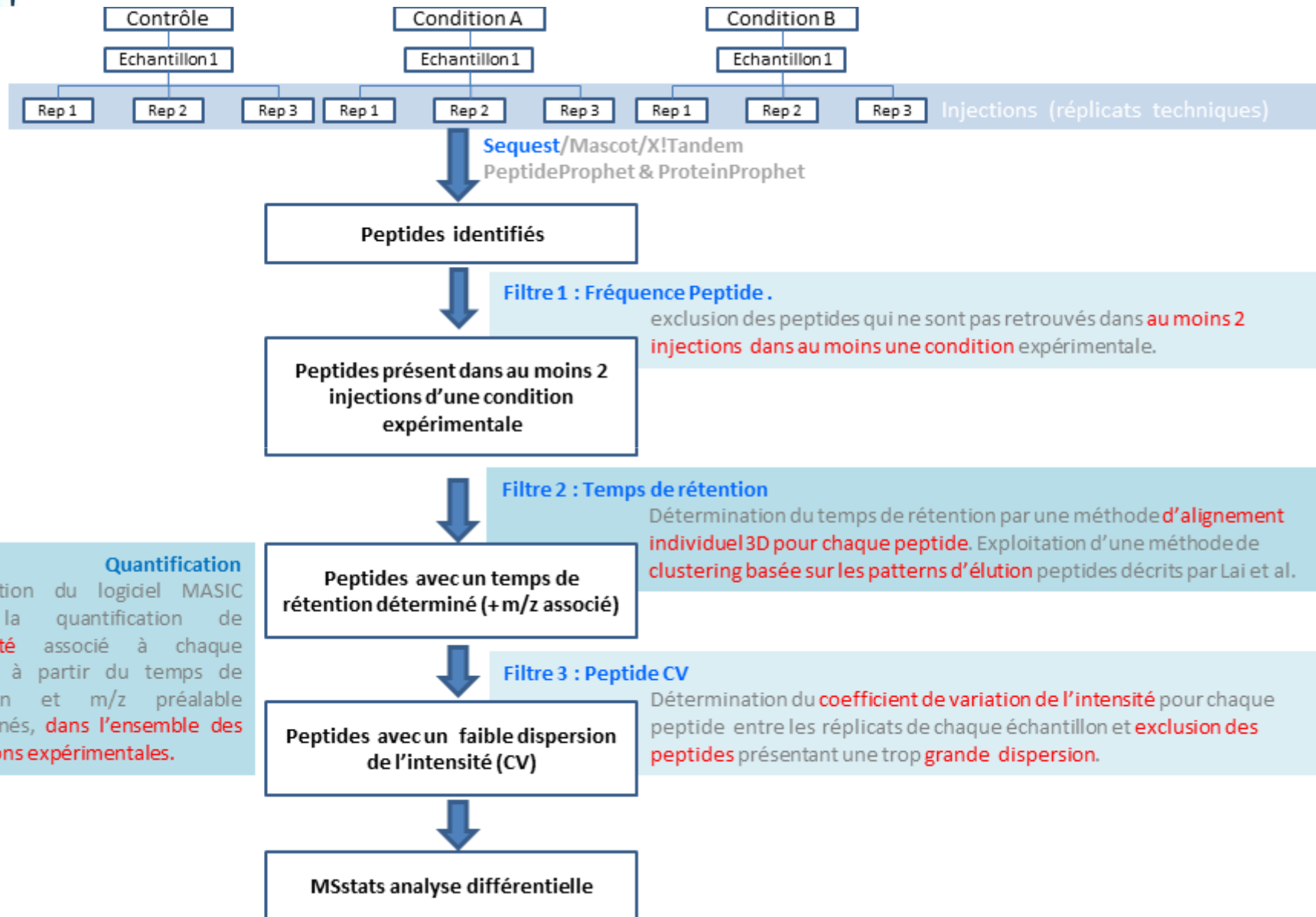
- Adaptation de la méthode SVA développée dans le cadre des puces à ADN (Leek and Storey, *PLoS Genet* 2007) pour la prise en compte des valeurs manquantes
- Identification du biais par décomposition en valeurs singulières, puis retrait de ce biais des données
 - ⇒ Pas de nécessité de définir l'origine du biais
- Principe
 - Estimation de l'effet du facteur d'intérêt
 - Décomposition en valeurs singulières de la part qui n'a pas été expliquée par le modèle
- Disponible dans le logiciel DAnTE

Karpievitch *et al.* (2009) *Bioinformatics* , **25**:2573-2580

Normalisation - Bilan

- ⇒ Encore une fois pas de consensus, chaque logiciel utilise autre chose
- Dépend du jeu de données...malheureusement.
 - Comparaison de plusieurs méthodes
 - Lai *et al. Int J of Proteomics* (2012): plutôt que d'utiliser une méthode de normalisation sous de mauvaises hypothèses, mieux vaut ne pas normaliser du tout.
 - Filtre sur les peptides plus important
 - Introduction de protéines en quantité connue
 - « notre » philosophie: peut interférer avec les protéines d'intérêt
 - Remarque: normalisation, puis imputation des valeurs manquantes (Karpievitch *et al. BMC Bioinformatics* 2012)

Exemple de workflow complet



Choix de la méthode

- Large panel de méthodes
- Pas de consensus, et ceci à toutes les étapes
- Combinaison de différents logiciels
- Dépend également du jeu de données
- Choix de la méthode la mieux « maitrisée »
- Comparaison de quelques méthodes sur le jeu de données étudié

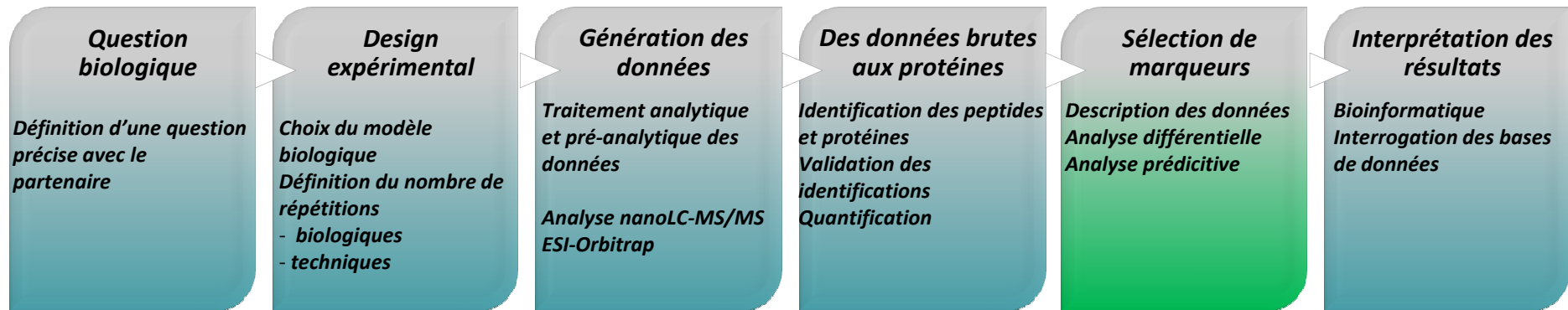
Choix du logiciel

- **Logiciels constructeurs**
 - Pas toujours de documentation très précise
 - Problème de changement de version des logiciels
 - Changement de paramètres pas toujours possibles
 - Problèmes éventuels de format: tous ne lisent pas les formats libres
 - « Avantage » : résultats préliminaires pour le biologiste
- **Logiciels libres: tous les avantages!**
 - Bénéfique de combiner des blocs modulaires qui viennent de différents logiciels pour obtenir un consensus sur un pipeline d'analyse
- **Liste de logiciels disponibles:**
<http://www.ms-utils.org/wiki/pmwiki.php/Main/SoftwareList>

Remarque sur les formats

- Formats open-source
 - mzML, mzXML: Données de MS – MS/MS
 - pepXML: ID des peptides et statistiques diverses
 - protXML: ID des protéines et statistiques diverses
- Intérêts
 - Partage de données entre différents outils
 - Partage de données entre différents laboratoires
 - Application de pipelines génériques

Sélection de marqueurs

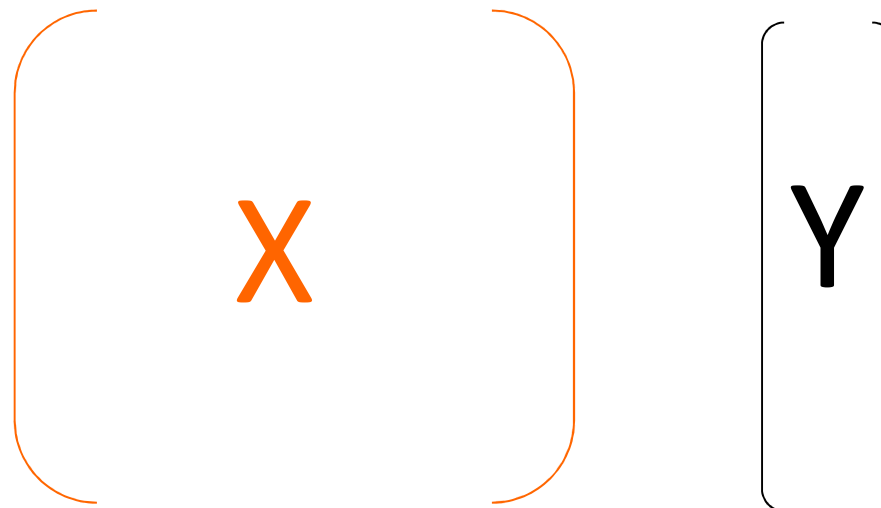


Données à analyser

p protéines
- Variables explicatives -

Variable(s) à expliquer
(phénotype, diagnostic,
etc...)

n échantillons



Particularité: $p \gg n$

A chaque question son analyse

- **Description** des données
 - Analyse non supervisée
- **Identification de marqueurs** caractéristiques d'une classe d'observations
 - Analyse différentielle
- **Prédire** une caractéristique pour une nouvelle observation.
 - Classement - Analyse supervisée

⇒ **C'est la question qui définit le choix de la méthode d'analyse**

Analyse non supervisée

Classification ascendante hiérarchique
Analyse en Composantes Principales

Objectifs:

- Visualisation de la structure des données
 - On cherche à regrouper les éléments dans l'espace des patients ou des variables
 - Détection de sous-groupes de variables
 - Gènes co-régulés par exemple
- ⇒ Méthodes exploratoires : aucune connaissance introduite a priori
- ⇒ On parle de méthode de classification, ou encore **d'analyse non supervisée**

Objectif

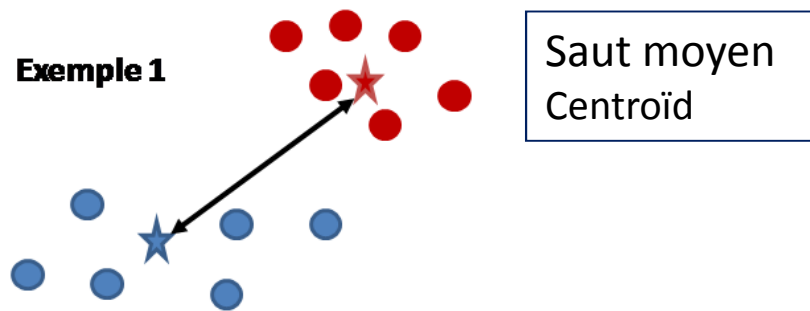
- Regrouper des entités proches dans une même classe
- Classe = ensemble d'entités qui sont proches ou se ressemblent
- Arbre de classification = dendrogramme

Moyen

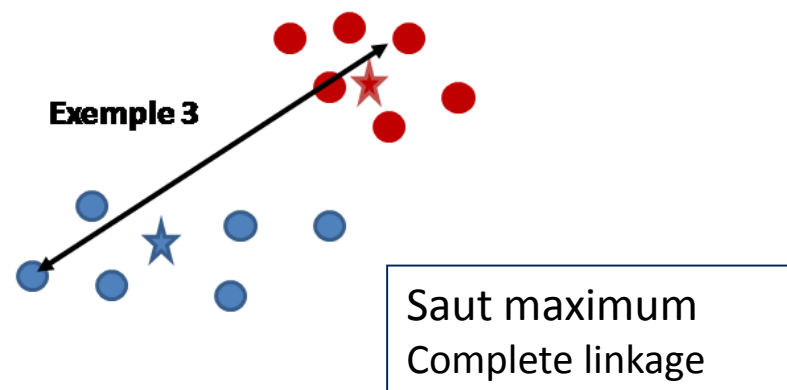
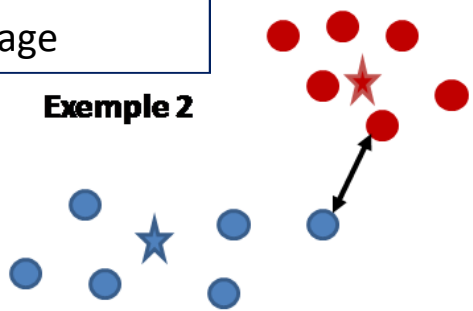
- Construction des classes de manière itérative: à chaque itération, regroupement des 2 entités les plus proches
- Début : chaque entité constitue une classe
- Fin : toutes les entités sont regroupées dans la même classe

- Comment décider que deux entités sont proches ?
-> Notion de métrique = mesure de la similarité ou dissimilarité entre les entités
- Métrique entre les entités (variables ou patients)
 - Dissimilarité (distance euclidienne p.ex): minimisée
 - Similarité (Coefficient de corrélation p.ex): maximisée
- Similarité entre les classes

Différents types de distance

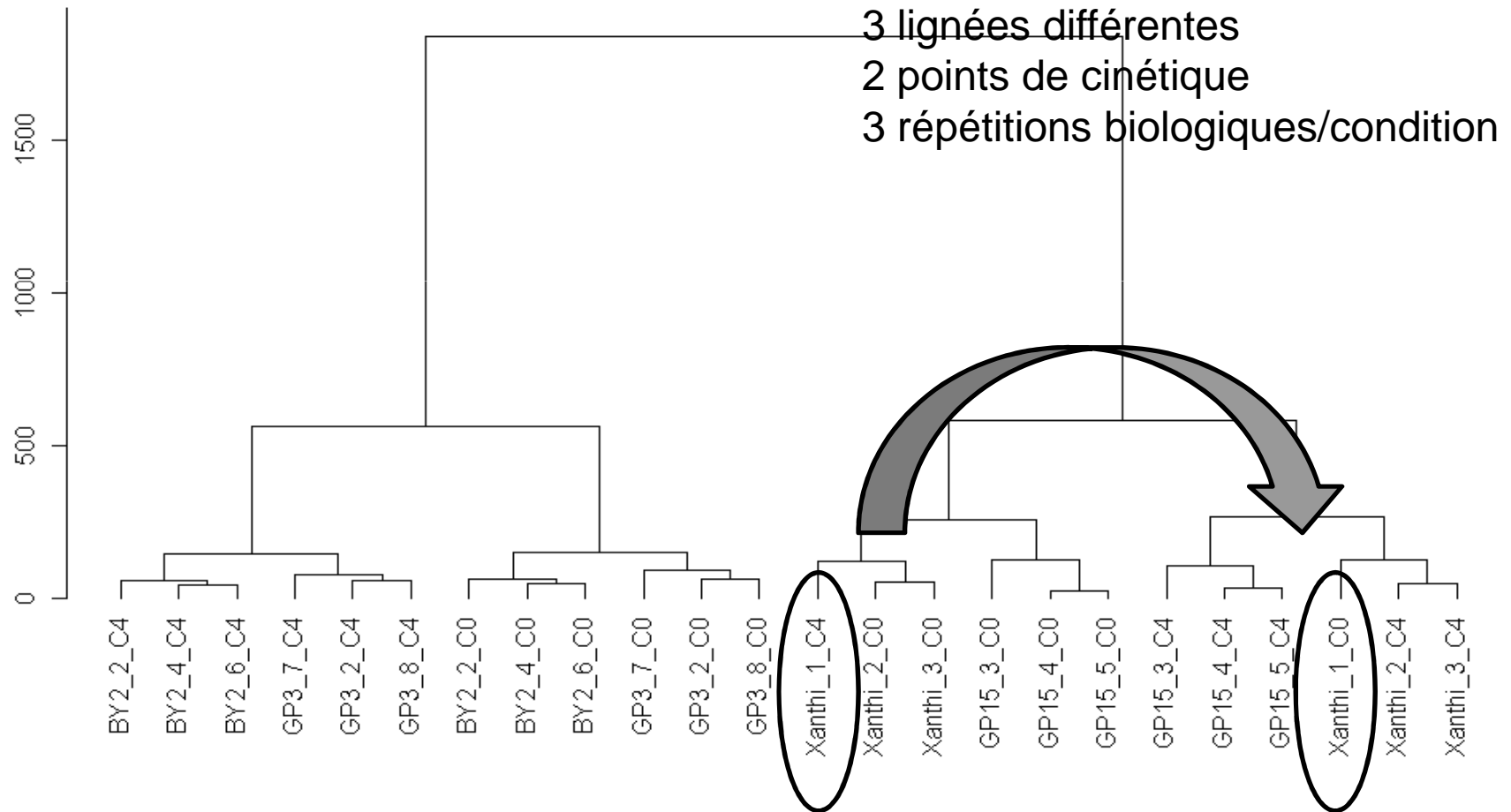


Saut minimum
Single linkage

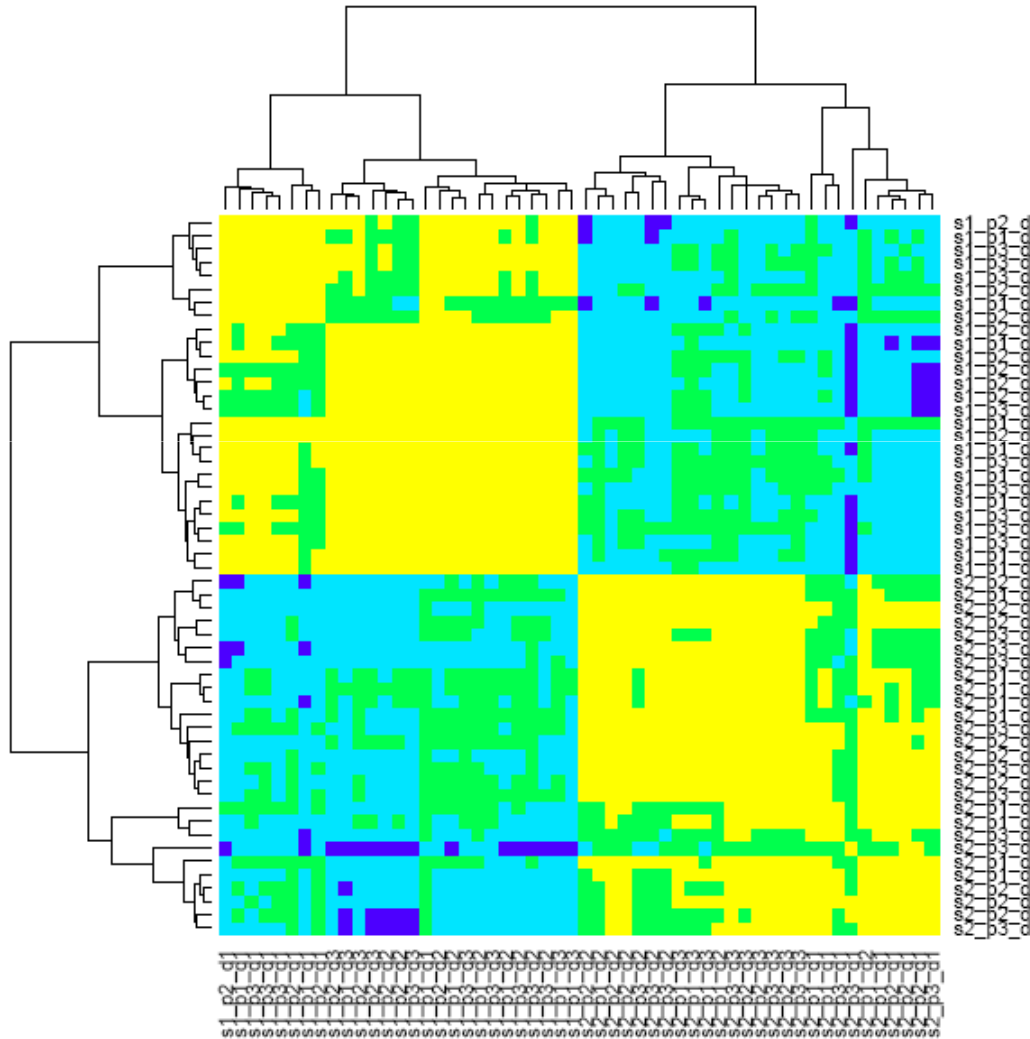


Exemple 1: dendrogramme simple

Données issues de l'analyse du transcriptome –
INRA UMR Plante-Microbe-Environnement



Exemple 2: représentation en image



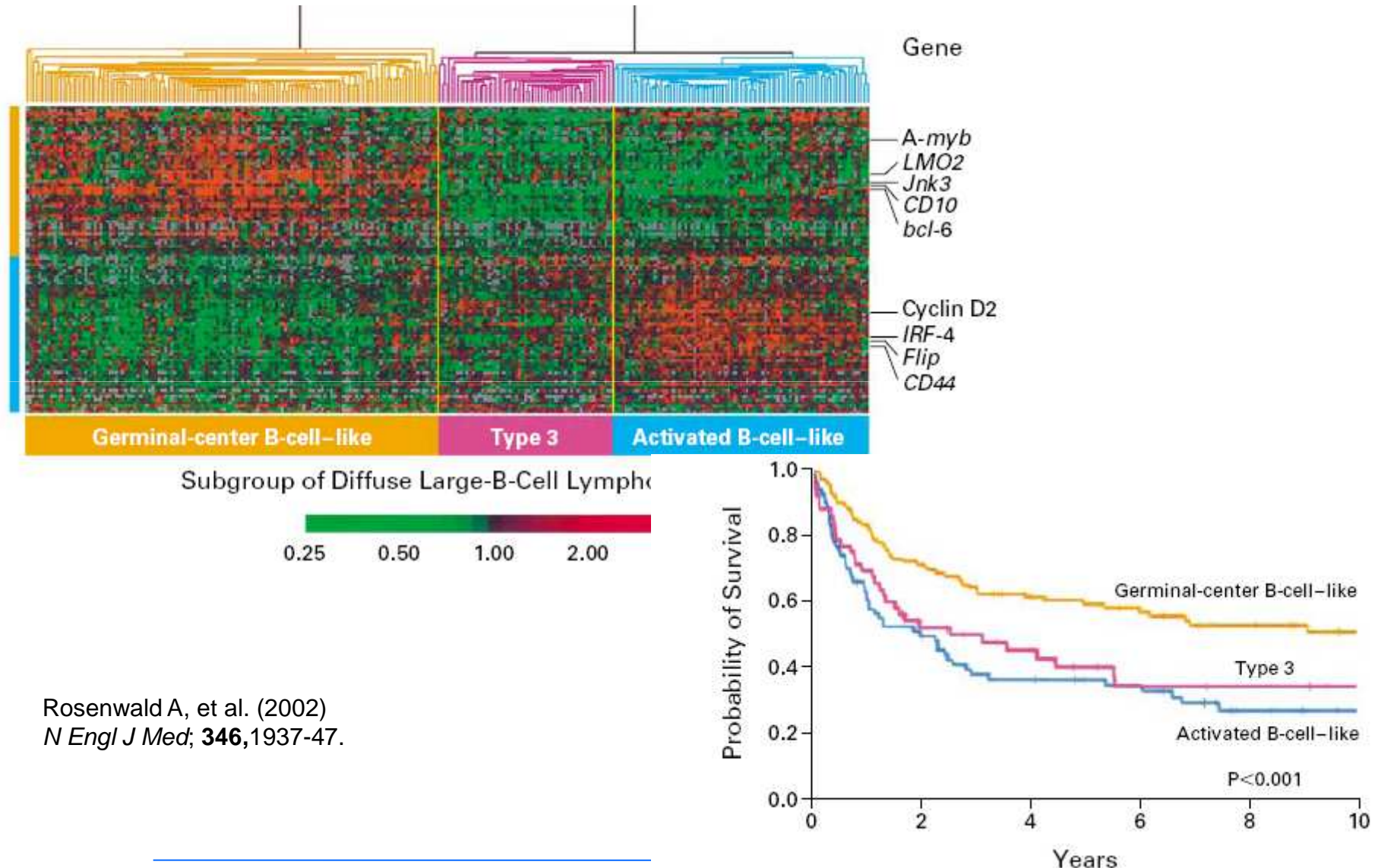
*Heatmap obtenu sur des spectres
issus d'une étude de protéomique
clinique*

si: patient

pi: purification

di: jour d'acquisition

Exemple 3: représentation simultanée



Rosenwald A, et al. (2002)
N Engl J Med; **346**,1937-47.

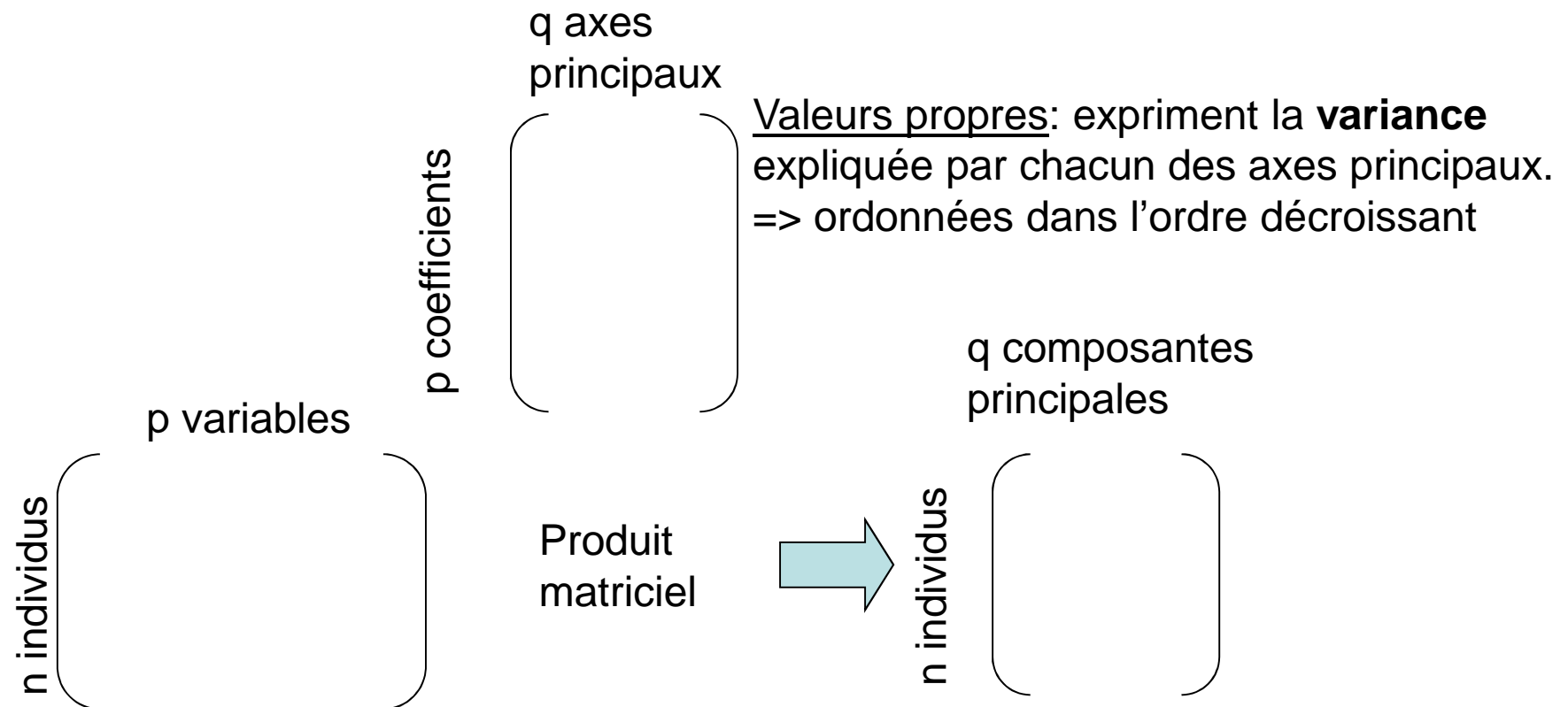
Objectif:

- Réduction de la dimension des données
- Visualiser les différentes sources de variabilité dans les données

Moyen:

- Trop de dimensions pour visualiser les spectres dans l'espace des pics
- Idée : trouver un espace plus petit dans lequel on peut visualiser les spectres -> projection sur des sous-espaces
- Cet espace est tel qu'il maximise la variabilité des données

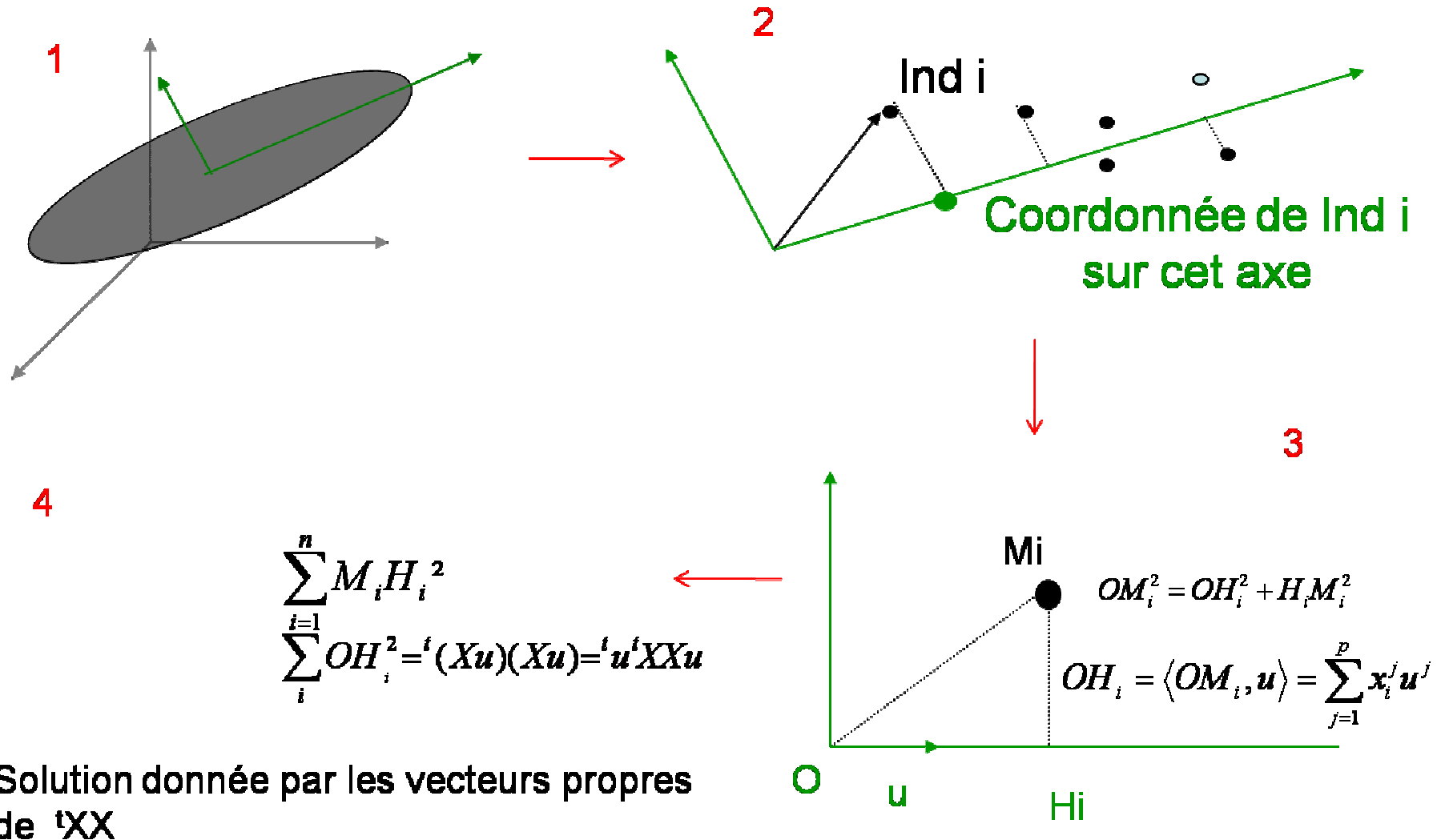
Précisions terminologiques



Espace d'origine
 n individus décrits
 par p variables

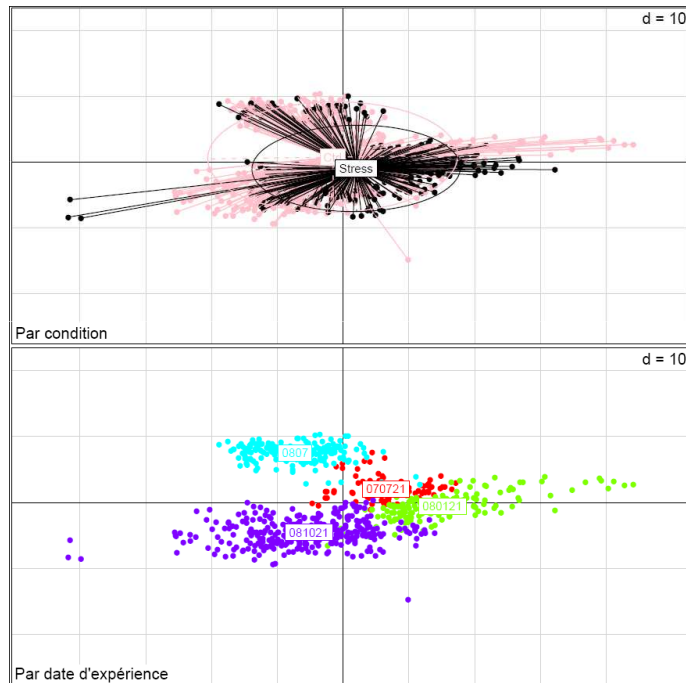
Espace des composantes
 n individus décrits
 par q nouvelles variables
 =les composantes de l'ACP

Illustration

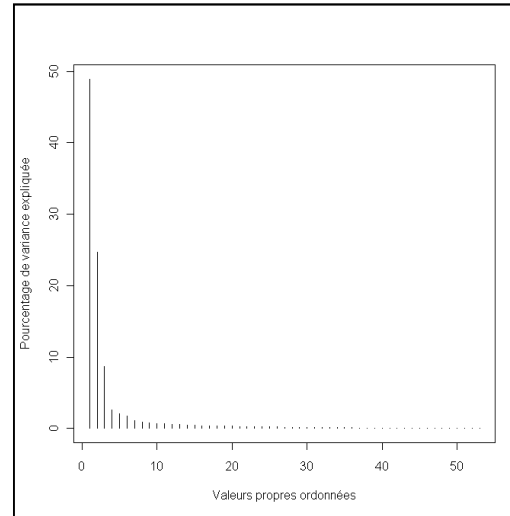


Interprétation de l'ACP

Représentation des observations



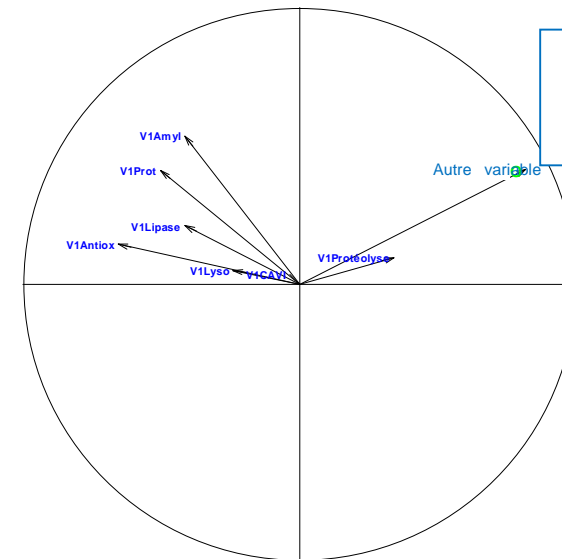
Mise en évidence des sources de variabilité connues



Représentation des valeurs propres

Pourcentage de variance expliqué par chaque axe

Représentation des variables



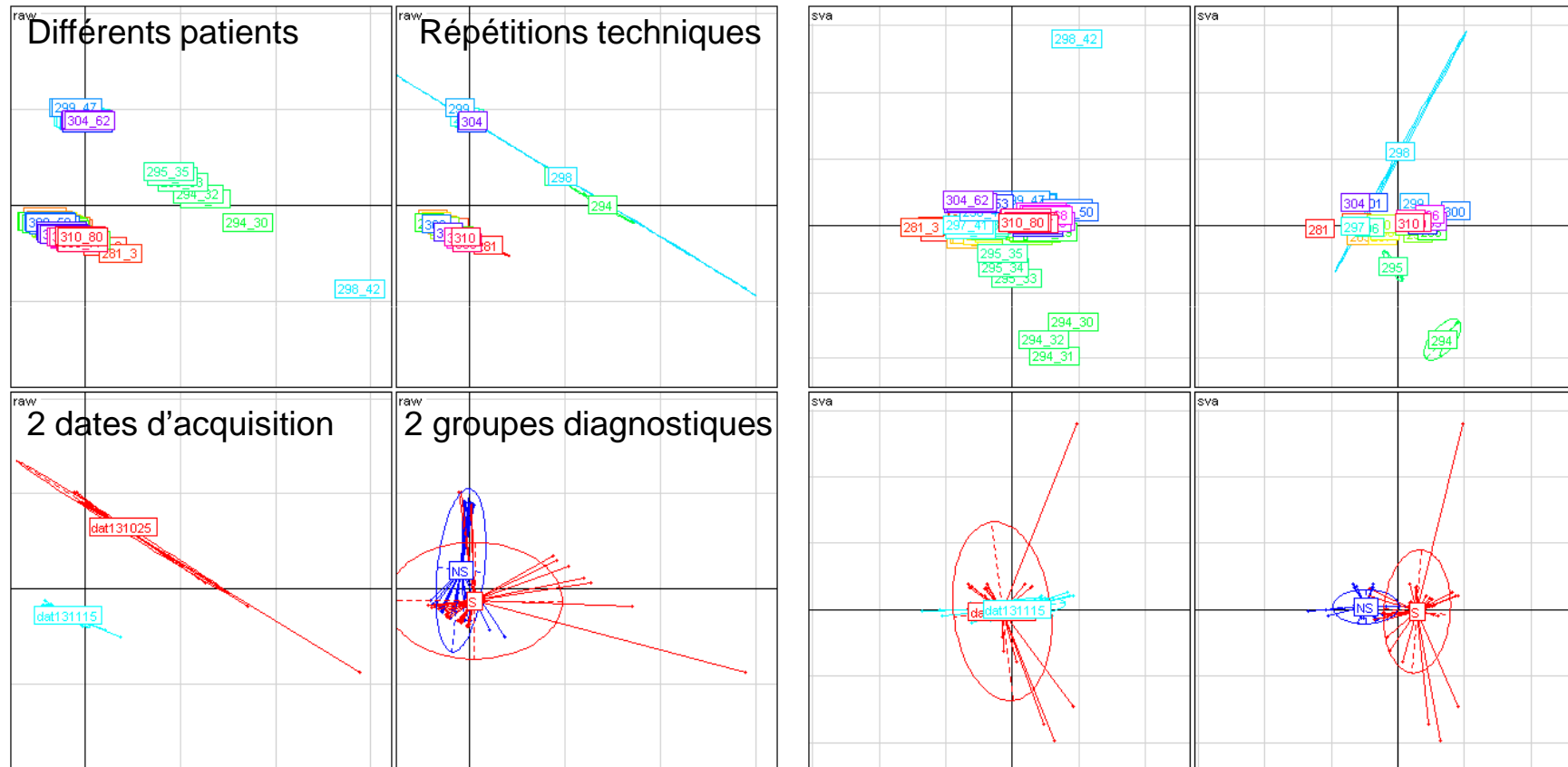
Poids des variables sur les axes

Illustration

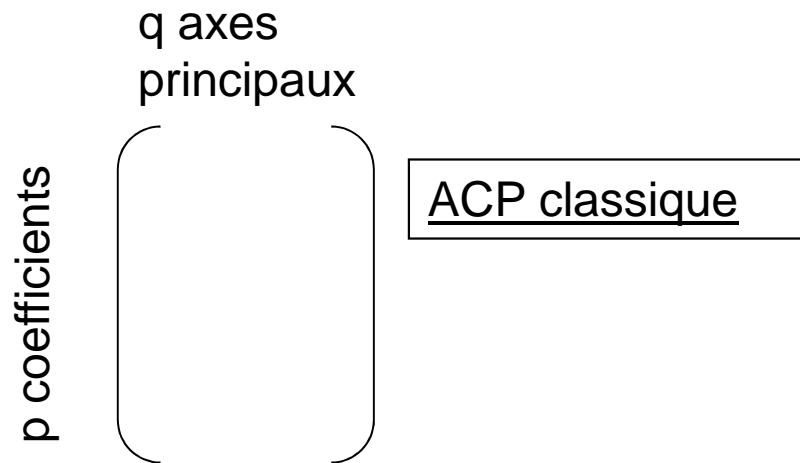
Normalisation par la méthode SVA

Avant normalisation

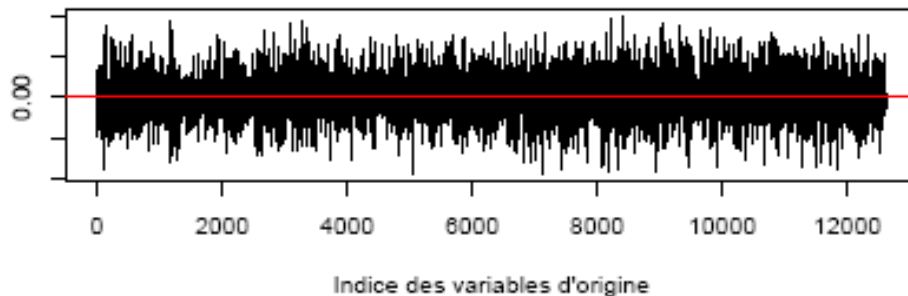
Après normalisation



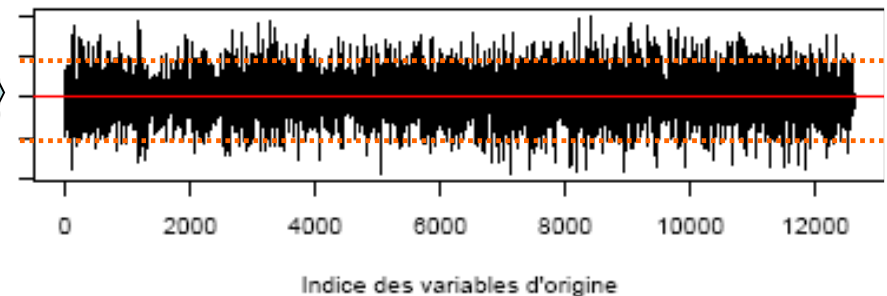
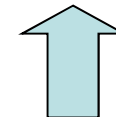
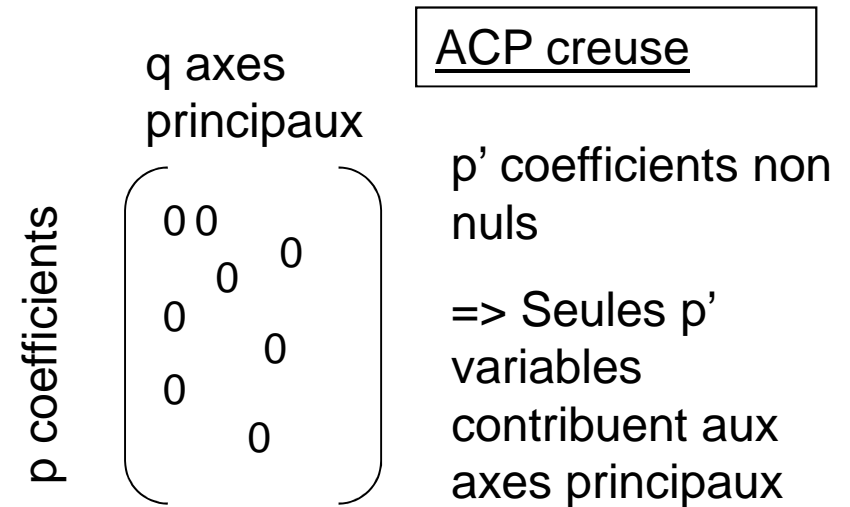
Pour aller plus loin



Chacune des p variables contribue aux axes principaux:

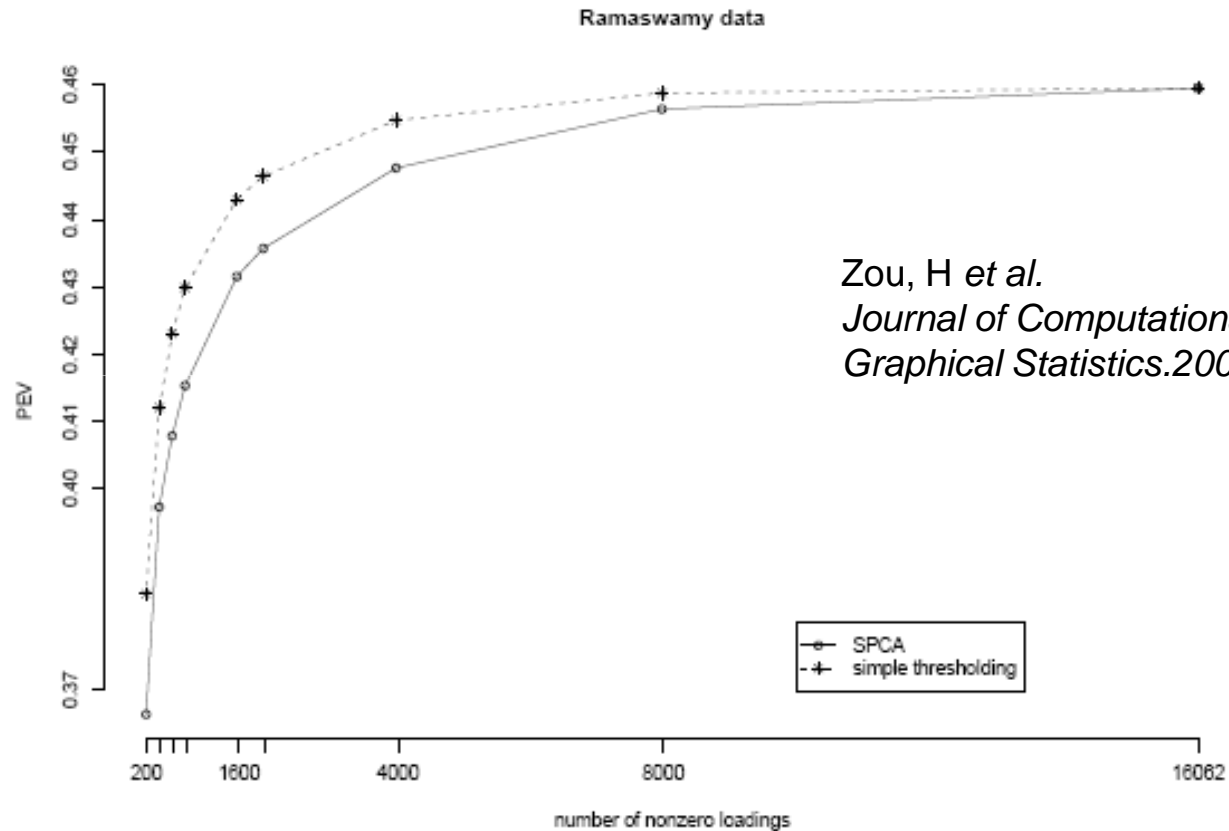


Beaucoup de coefficients très faibles!!



Seuillage des coefficients qui contribuent peu aux axes principaux

Illustration



Zou, H *et al.*
*Journal of Computational and
Graphical Statistics.*2006. **15**(2), 265-286

Pourcentage de variance expliquée en fonction du nombre de coefficients annulés

Analyse différentielle

Tests statistiques

Tests multiples

Puissance

Analyse différentielle

- Identification de biomarqueurs associés à un facteur d'intérêt
- Méthodes univariées: les variables sont considérées « un à un »
- Application d'un test statistique à chacun des éléments puis sélection des éléments pour lesquels les tests sont les plus significatifs
 - Tests paramétriques
 - Tests non paramétriques

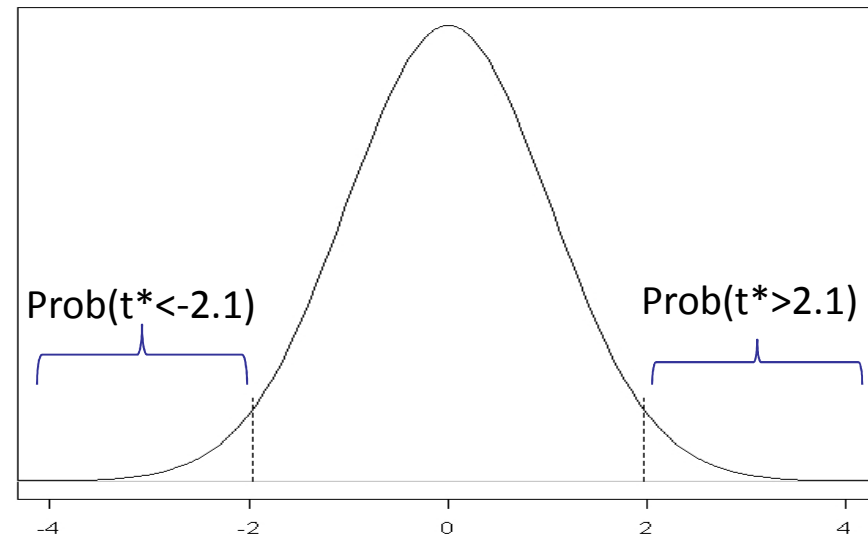
- Test d'une hypothèse relative à une question biologique définie à priori
- Définition
 - De l'hypothèse nulle H_0
 - De l'hypothèse alternative H_1
- Tests conduisent à deux types d'erreur:
 - Erreur de type I (risque de 1^{er} espèce)
 - Erreur de type II (risque de 2nd espèce)

Notion de p-value

- La p-value correspond à la probabilité d'obtenir, sous H_0 , une valeur de la statistique de test T supérieure ou égale à celle observée t :

Distribution des valeurs de la statistique

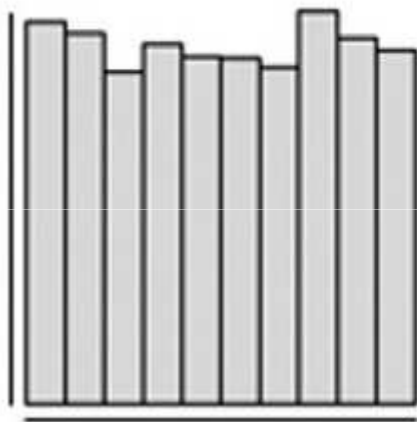
$$p\text{-val} = p(|T| \geq t | H_0)$$



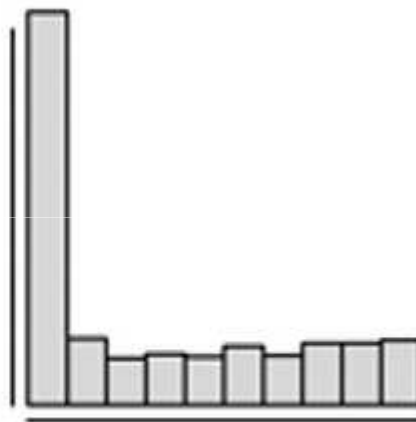
- Décision: si $p\text{-value} < \alpha$ H_0 rejetée

Distribution des p-values

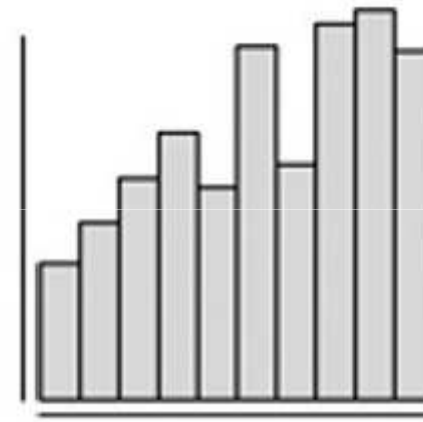
Distribution des p-values dans 3 situations différentes



Pas de test significatif



Existence de tests significatifs



Dépendance entre les tests

- « Expectation value »
- Calculée par les moteurs de recherche pour l'identification des peptides
- Nombre attendu de peptides qui ont un score supérieur ou égal au score observé sous l'hypothèse nulle
 - = Nombre de fois où on pourrait obtenir au moins ce score par chance
- Plus l'assignement est bon, plus l'E-value est faible

Types d'erreur

	H0 acceptée	H0 rejetée
H0 vraie	1- α	α (type I)
H0 fausse	β (type II)	1- β (puissance)

- Erreur de type I: rejet à tort de l'hypothèse nulle
 - Notée α
- Erreur de type II: acceptation à tort de l'hypothèse nulle
 - Notée β
 - Reliée à la puissance du test 1- β .

Tests multiples

		Décision		
		H0 acceptée	H0 rejetée	Total
Vérité	H0	U (VN)	V (FP)	p0
	H1	T (FN)	S (VP)	p1
	Total	1-R	R	p

- Pour un total de p éléments testés
- U: Vrais positifs
- V: Faux positifs
- T: Faux négatifs
- S: Vrais négatifs

Tests multiples

- Test simultané d'un grand nombre d'hypothèses
- Exemple: test de 10 hypothèses indépendantes
 - Probabilité de déclarer un test significatif à tort: 5%
 - Probabilité de déclarer au moins un test significatif à tort parmi les 10: $0.401=1-(0.95)^{10}$

⇒ **Le risque explose avec le nombre de variables**

Contrôle du risque de type I (I)

- Ajustement des p-values
- Family Wise Error Rate (FWER)
 - Probabilité d'obtenir au moins une erreur de type I
 - Proportion de gènes déclarés significatifs à tort.
 - Contrôler le FWER au seuil de 5%, permet d'être confiant à 95% de n'avoir aucun faux positif

$$\text{FWER} = p(V \geq 1)$$

Contrôle du risque de type I (II)

- False Discovery Rate (FDR)
 - Proportion attendue de faux positifs parmi les gènes déclarés significatifs.
 - Contrôler le FDR au seuil de 5% permet d'affirmer qu'en moyenne, le taux de faux positifs est inférieur à 5%.

$$\text{FDR} = E(V/R)$$

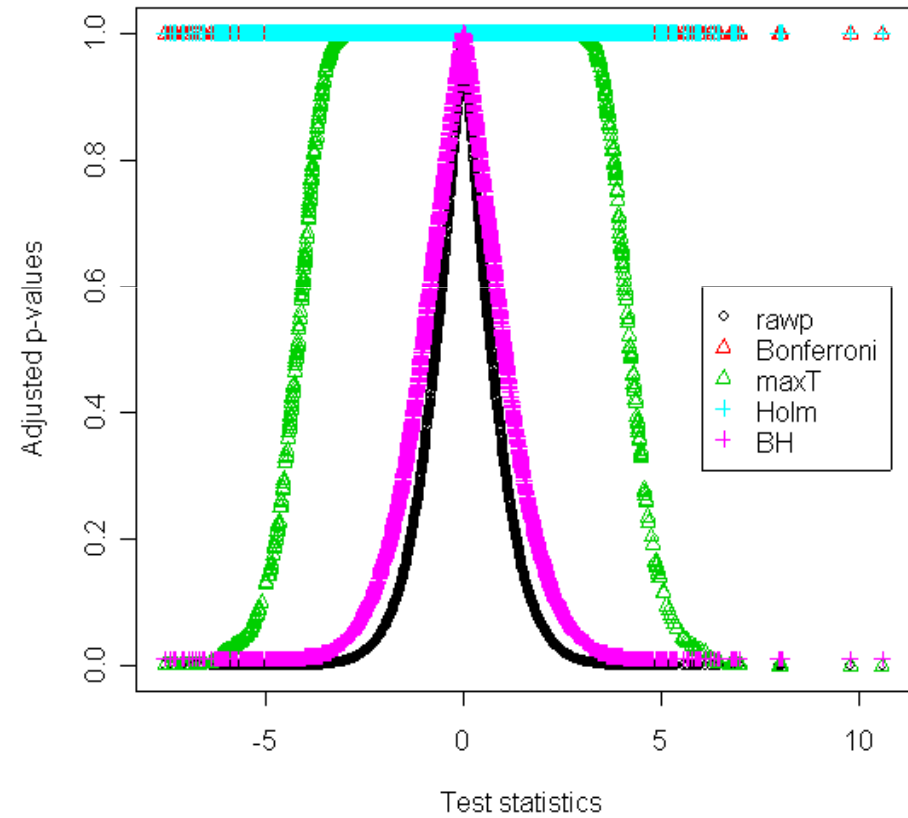
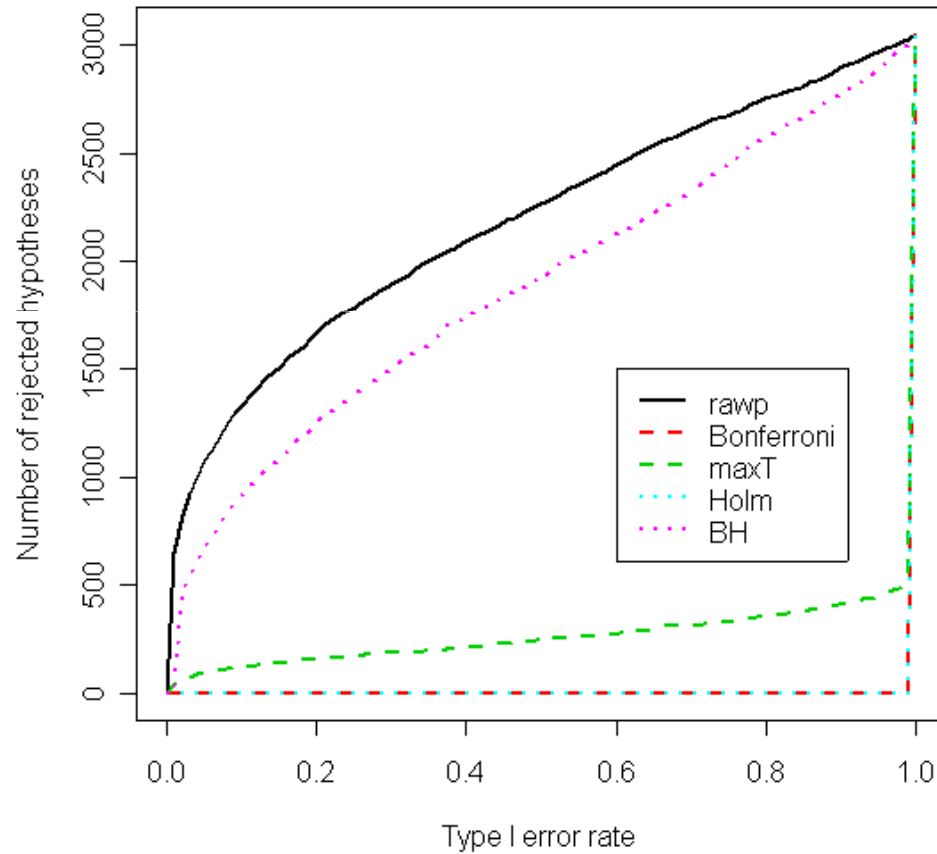
- Variantes: FDR local, q-value

⇒ Moins conservateur que le FWER

Procédure de Benjamini-Hochberg

- Test de m hypothèses à laquelle sont associées une p -value $P_i, i=1 \dots m$
- Contrôle du FDR au niveau α
- Ordonnement des p -values associées aux hypothèses $H_{(i)}$
 - $P_{(1)} \leq \dots \leq P_{(m)}$
- Pour un α donné, on cherche la valeur de k la plus grande telle que $P_{(k)} \leq k \cdot \alpha / m$
- Si un tel k existe, on juge significatifs les tests $H_{(i)}, i=1 \dots k$

Représentations graphiques



q-value et FDR local

q-value

- Associée à un test
- FDR minimum que l'on peut obtenir en jugeant le test significatif
 - = proportion de FP attendus si le test est jugé significatif
- Exemple:
 - Si une protéine a une q-value de 0.01 cela signifie que 1% des protéines qui ont une p-value au moins aussi petite que cette protéine sont des FP

FDR local

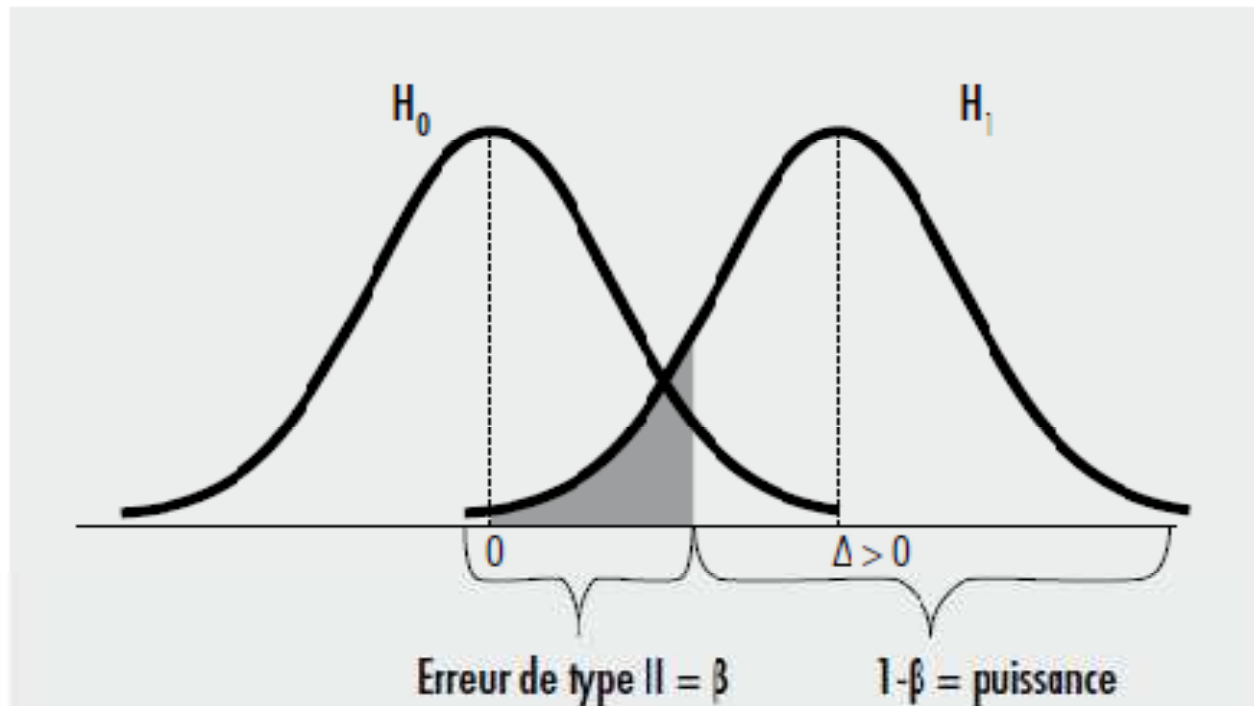
- Associé à un test
- Probabilité pour une protéine que ce soit un FP
- Intéressant si l'on s'intéresse à une protéine en particulier

Représentation graphique - Volcano plot -

Fold-change: rapport des
Moyennes observées dans
chacun des groupes

Puissance d'un test

- Capacité du test à mettre en évidence une différence qui existe réellement



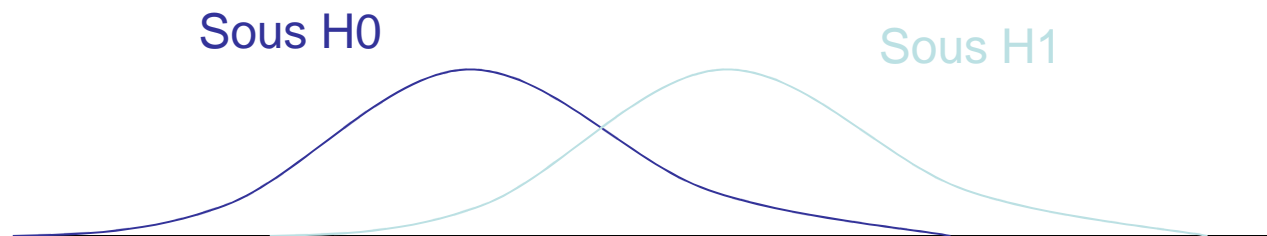
Fluctuation d'échantillonnage de la grandeur test sous H_0 et H_1

- Contrôle du risque de type II

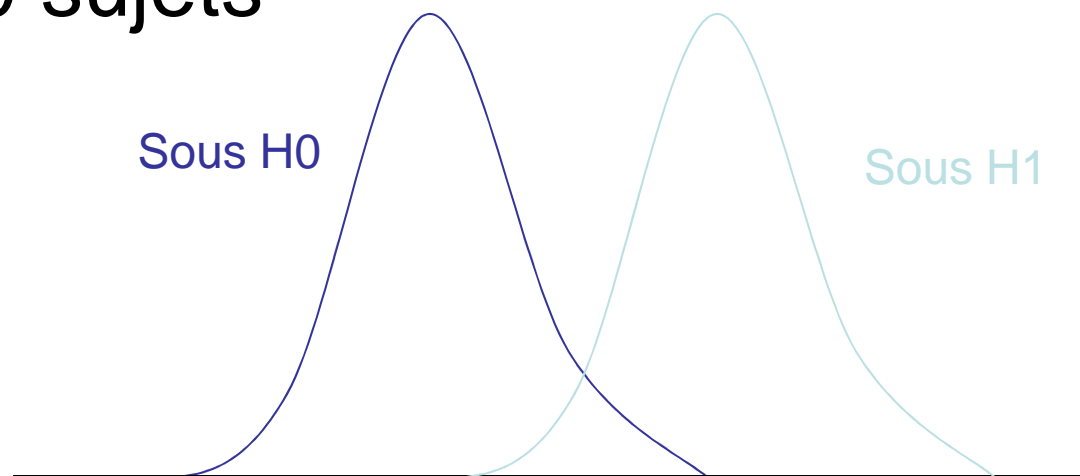
Impact du nombre de sujets

Fluctuation d'échantillonnage de la grandeur test sous H0 et H1

- 100 sujets



- 300 sujets



Calcul du nombre de sujets

- Le nombre de sujets requis dépend de
 - fold-change δ que l'on souhaite être en mesure de détecter
 - FDR q
 - puissance du test β
 - rapport du nombre de protéines différentielles sur non différentielles m_0/m_1 (en réalité)
 - nombre minimum de peptides par protéine
 - nombre de répétitions techniques L et biologiques K
 - variabilité technique σ^2_{Error}

Formulation

$$\delta^2 \geq \frac{\hat{\sigma}_{Error}^2}{IKL} (t_{1-\beta, df} + t_{\alpha/2, df})^2$$

Avec

$$\alpha = (1 - \beta) \cdot \frac{q}{1 + (1 - q) \cdot m_0 / m_1}$$
$$df = IJKL(L - 1) + (I - 1)J(K - 1)$$

Clough *et al.* *BMC Bioinformatics* 2012

Illustration

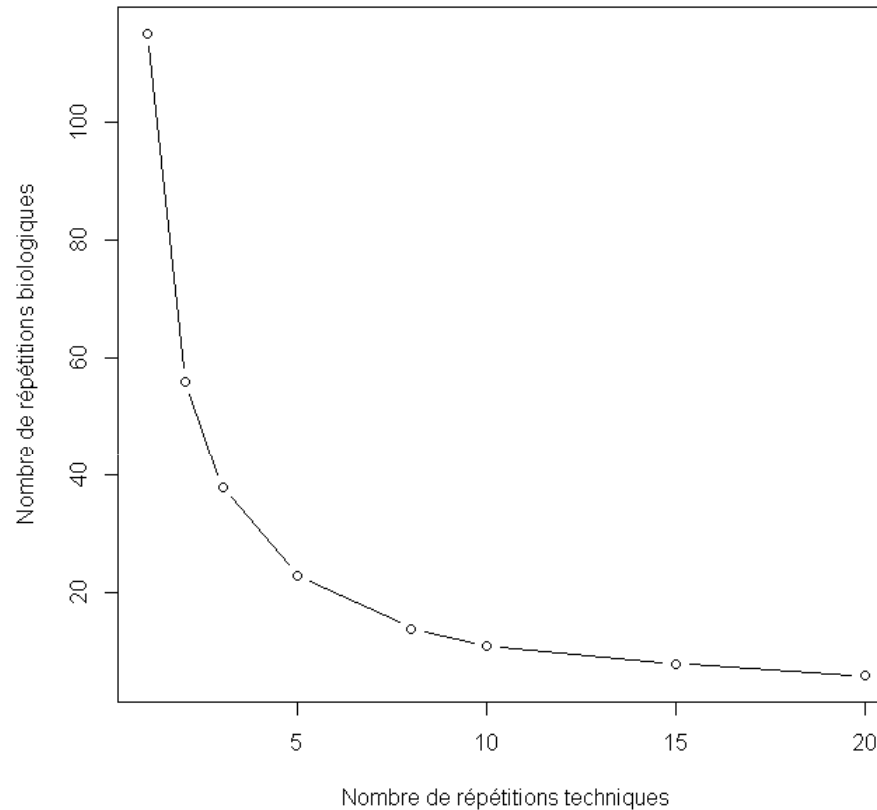
252 protéines, minimum de 1 peptide par protéine, comparaison de 2 groupes (10+17 patients), 3 répétitions techniques, $m_0/m_1=0.99$ (par défaut)

	$\Delta=\log_2(1.25)$			$\Delta=\log_2(1.5)$			$\Delta=\log_2(2)$		
	$\beta=0,1$	$\beta=0,2$	$\beta=0,3$	$\beta=0,1$	$\beta=0,2$	$\beta=0,3$	$\beta=0,1$	$\beta=0,2$	$\beta=0,3$
q=0,01	519	440	129	157	118	12	53	45	40
q=0,05	439	367	321	133	128	97	45	38	33
q=0,1	402	334	290	122	101	88	41	34	30

Nombre de répétitions biologiques nécessaires
 Résultats obtenus avec la librairie de R MSstats

Rôle des répétitions techniques

$q=0.05$
 $\beta=0.2$
 $FC=1.5$



⇒ Le nombre de répétitions techniques peut compenser le manque de répétitions biologiques

Analyse supervisée

Sélection de variables

Extraction de variables

Méthodes qui intègrent la dimension

- Classement d'un échantillon parmi des classes connues, i.e **prédire** des caractéristiques pour de nouvelles observations
- Chaque observation:
 - Décrite par $X = (X_1, \dots, X_p)$
 - Réponse associée $Y \in \{1, \dots, K\}$
- Construction d'un « classifieur » C

$$\hat{y} = C(x) = k$$

Problème de la multiplicité

- Particularité des données: nombre très important de variables étudiées simultanément
- Conséquences
 - En univarié: pics jugés significatifs par le simple fait du hasard
 - En multivarié: n'importe quel modèle peut être parfaitement prédictif par chance uniquement

Problème de la multiplicité

- Nombre de variables (pics) \gg Nombre d'observations
 - Multi-colinéarité
 - Optimisme
- Nécessité d'adapter les méthodes classiques:
 - Réduction préalable de la dimension:
 - Sélection de variables
 - Extraction de variables
 - Méthodes qui intègrent directement la dimension
 - Knn, PLS, etc...
 - Méthodes de régularisation

- Sélection univariée
 - Chaque variable est considérée indépendamment des autres
- Classement des variables selon l'importance de leur impact sur la réponse d'intérêt
- Exemples: tests de t, Wilcoxon, ANOVA, etc...

- Dans la littérature: « réduction de la dimension »
- Projection des individus dans un espace de dimension inférieure
- Construction de nouvelles variables qui « résument » les variables d'origine

Extraction de variables

- Analyse en composantes principales
 - Maximisation de la variance totale des données
- Partial Least Squares
 - Maximisation de la covariance entre les données et la réponse
- Différence majeure :
 - composantes PLS optimisées pour être prédictives du critère d'intérêt
 - composantes principales ne font qu'extraire le maximum de variance des prédicteurs

- Régression linéaire ordinaire: maximisation de la vraisemblance des données
- Régularisation: maximisation de la vraisemblance sous contrainte/pénalisation de la vraisemblance
- Pénalité de type L1
 - Lasso/Lars : $\sum |\text{coeff}| \leq \lambda$
 - Sélection d'un sous-ensemble de variables \Rightarrow Estimation de l'ampleur d'effet et sélection de variables

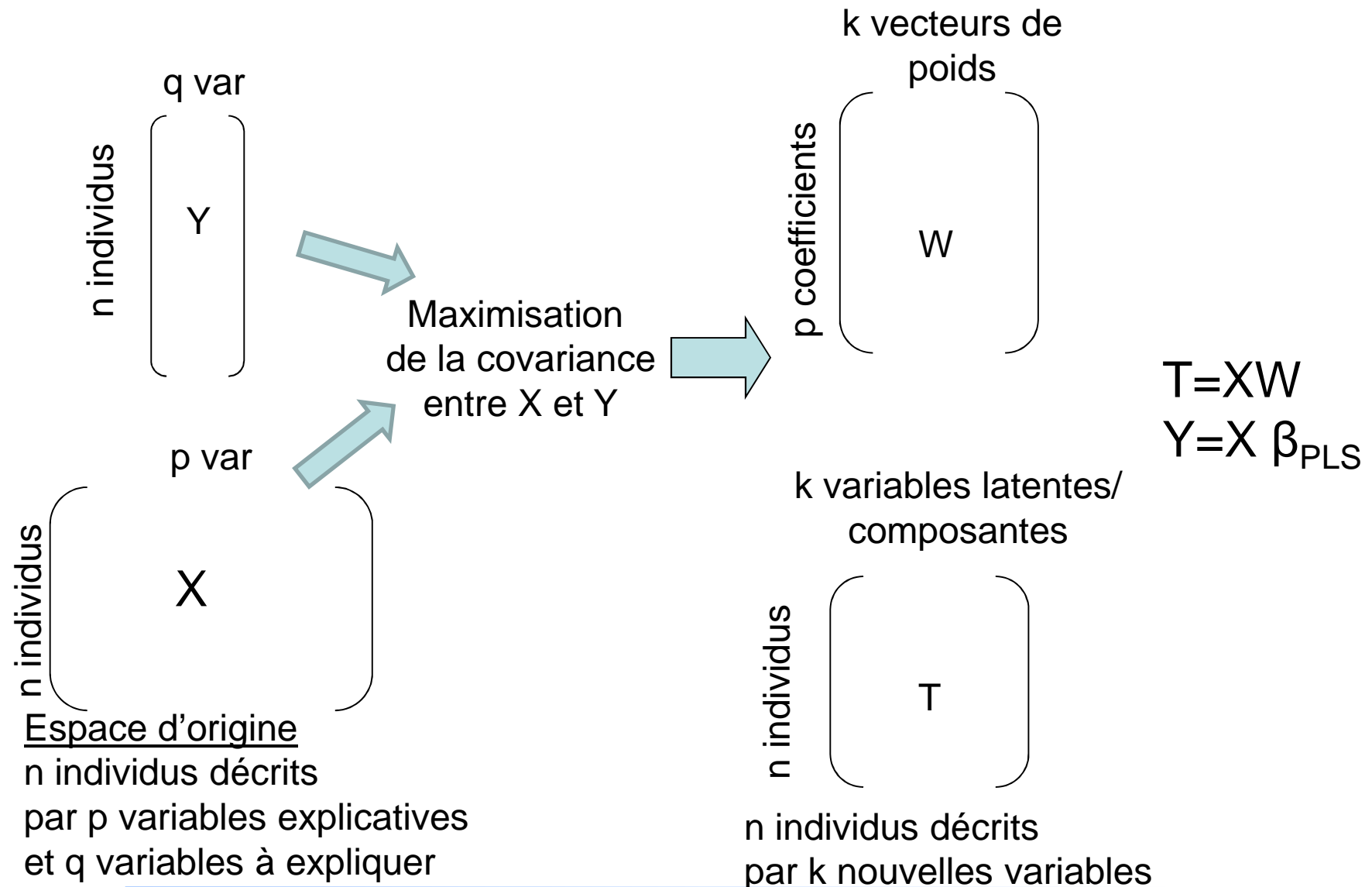
$$\min |Y - X\beta|^2 + \lambda |\beta|_1 \quad , |\beta|_1 = \sum_j |\beta_j|$$

- Pénalité de type L2
 - Régression ridge: $\sum ||\text{coeff}||^2 \leq \lambda$
 - Conserve toutes les variables dans le modèle

$$\min |Y - X\beta|^2 + \lambda |\beta|_2^2 \quad , |\beta|_2^2 = \sum_j \beta_j^2$$

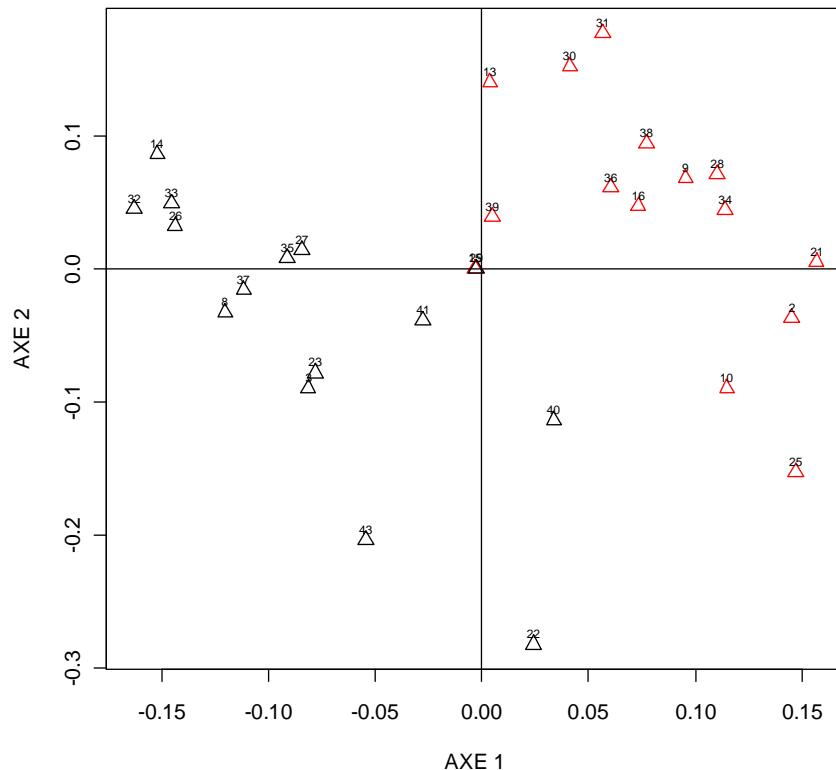
- Elastic Net: combinaison des 2 contraintes
 - Sélection de variables
 - Meilleures qualités prédictives

De l'ACP à la PLS



Représentation des individus dans la base de la PLS

Représentation des produits axes 1&2



⇒ Les axes sont tels qu'ils permettent de discriminer les 2 groupes d'individus

Extension de la PLS à la sparse PLS (SPLS)

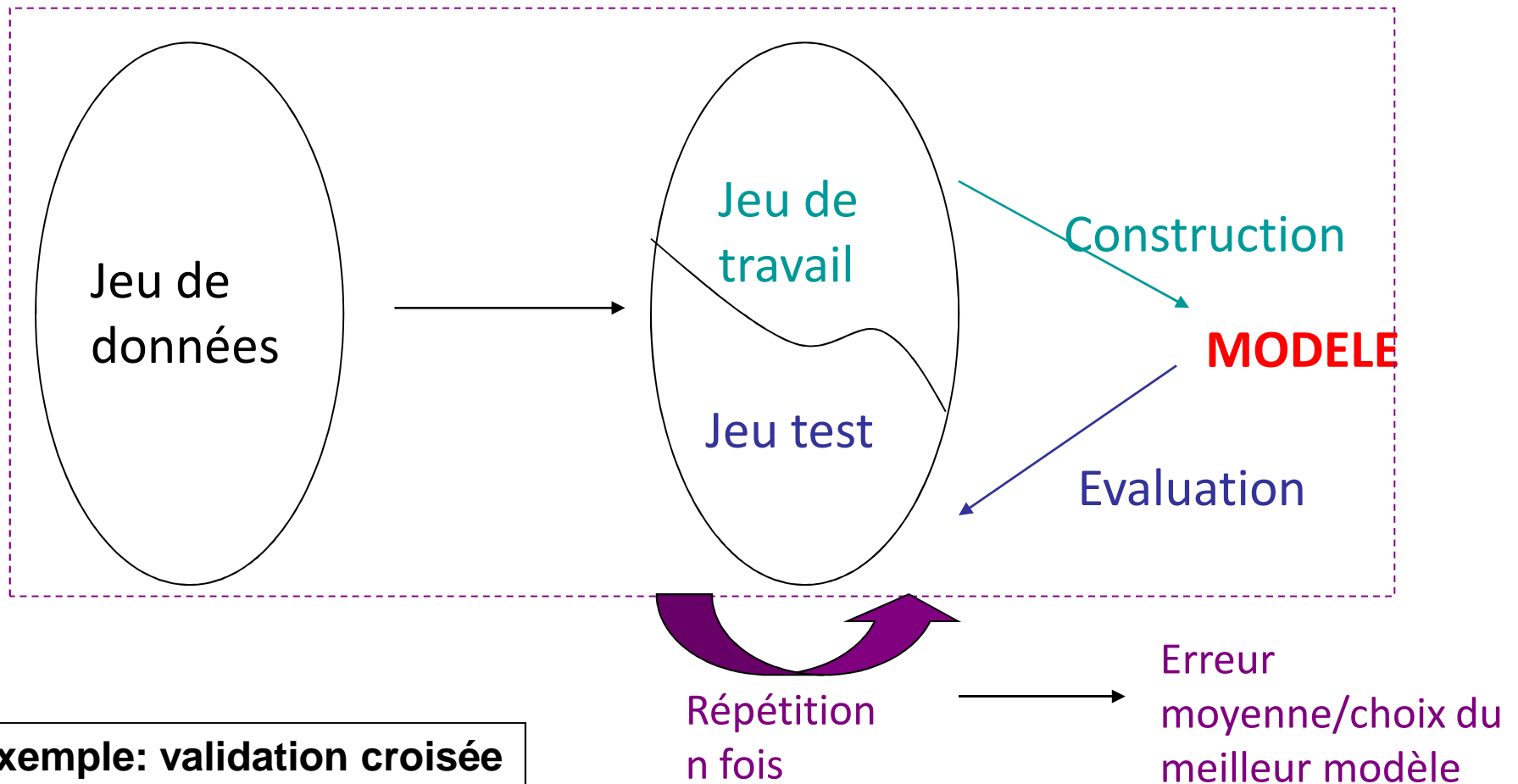
- Même idée que la sparse PCA
 - Sélection des coefficients qui contribuent le plus à la discrimination
 - Exemple: elastic net sur les poids des composantes
- ⇒ Optimisation de la prédiction.

Validation des modèles

- Validation interne
 - Séparation du jeu de données initial en un jeu de travail et un jeu test
- Validation externe
 - Utilisation d'un autre jeu de données qui n'a « jamais vu » le modèle
 - A privilégier
- Le modèle doit être fiable pour la prédiction du critère d'intérêt chez de nouveaux sujets
 - ⇒ Différence entre la qualité de l'ajustement et les qualités prédictives
 - ⇒ Question de l'optimisme

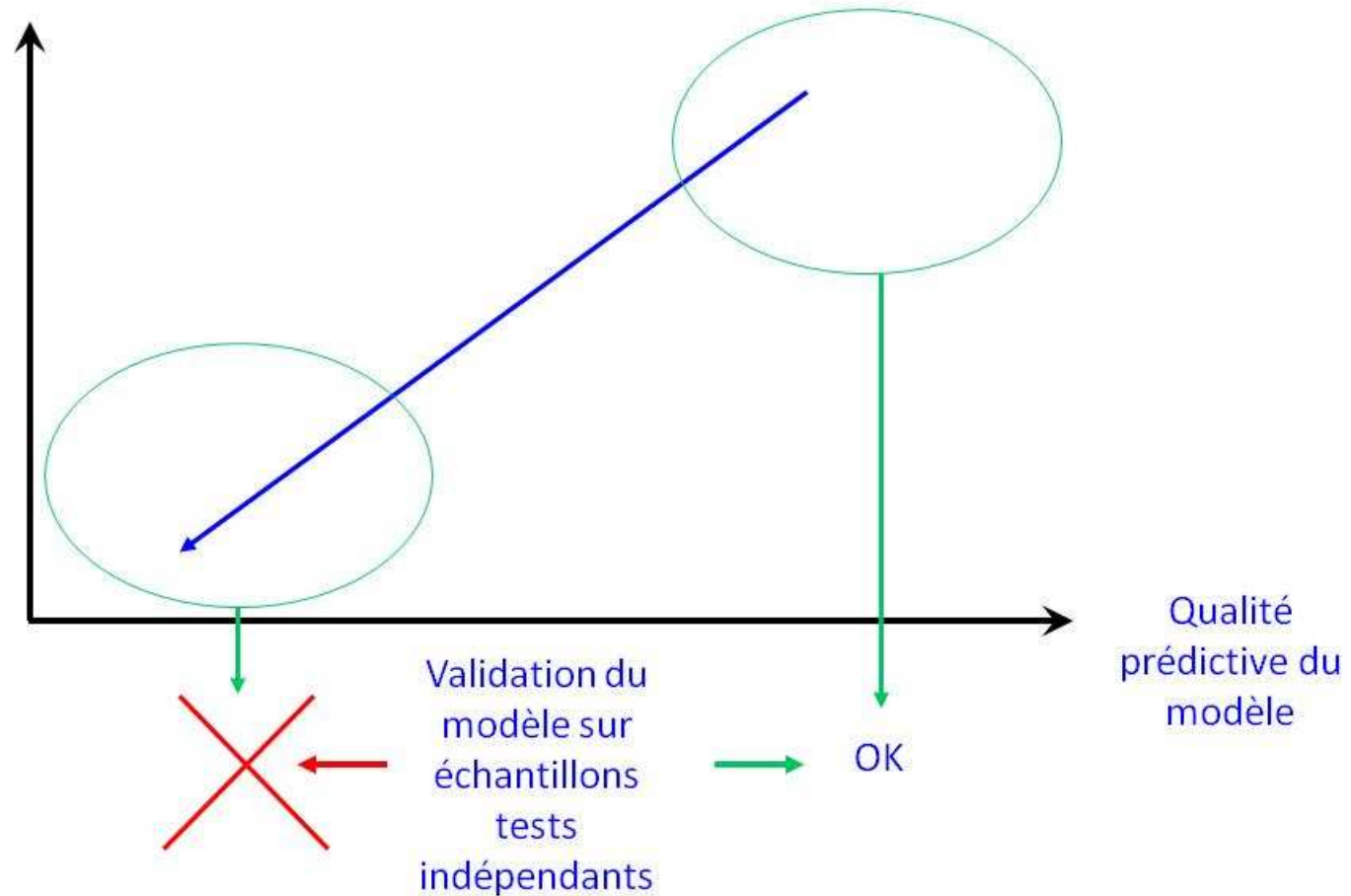
Illustration

Validation interne: séparation des données initiales en un jeu de travail et un jeu test



Puissance des études

Puissance =
Augmentation
du nombre
d'échantillons



L'optimisme est d'autant plus grand que le nombre de variable est important et que la taille de l'étude est petite

Approche en 2 temps

Etudes d'identification

→ Biomarqueurs candidats

Estimation de la grandeur d'effet

Etudes de validation

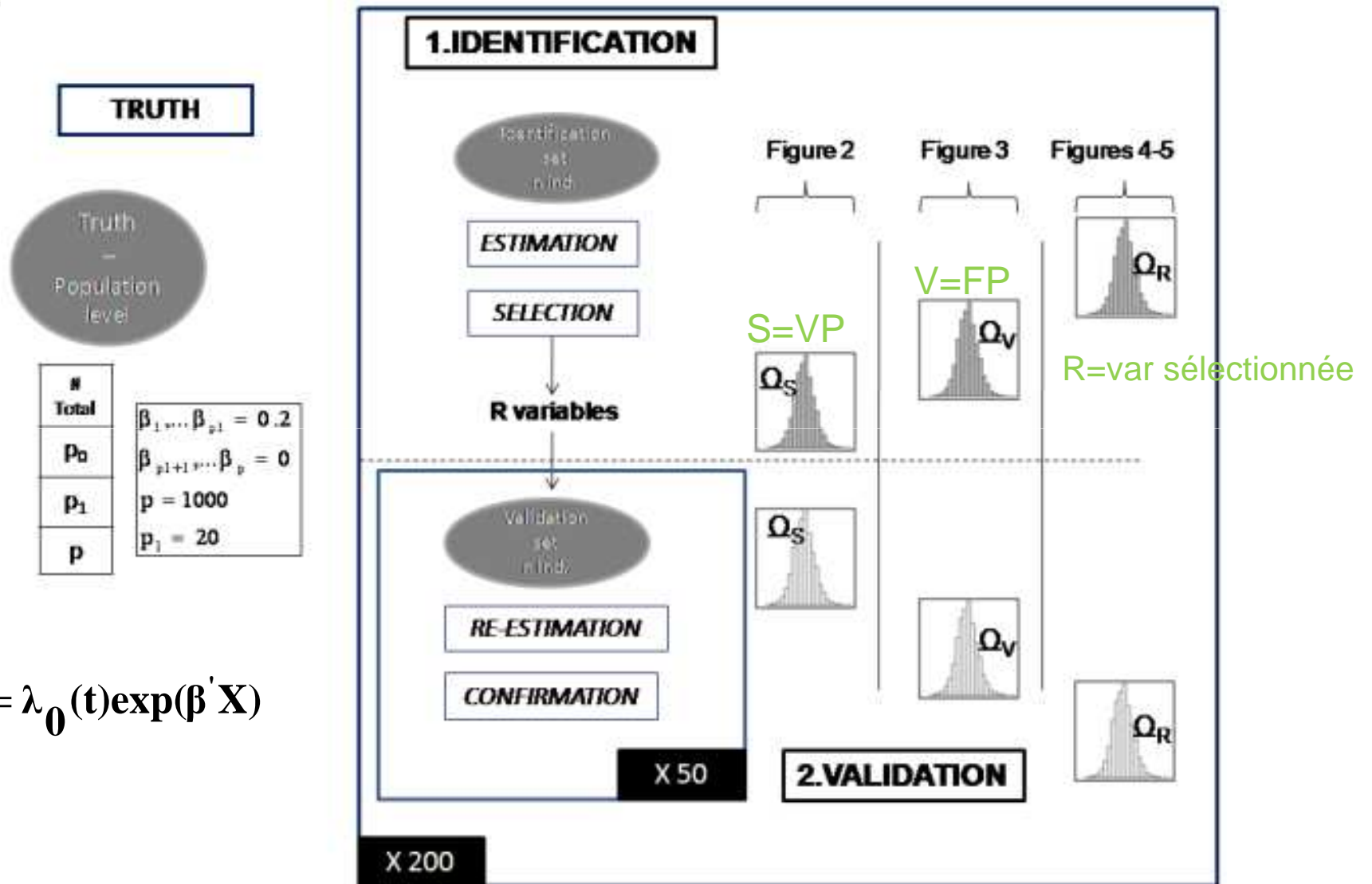
→ Biomarqueurs confirmés

Ré-estimation de la grandeur d'effet

Rappels


U: Vrais négatifs
V: Faux positifs
T: Faux négatifs
S: Vrais positifs


Illustration

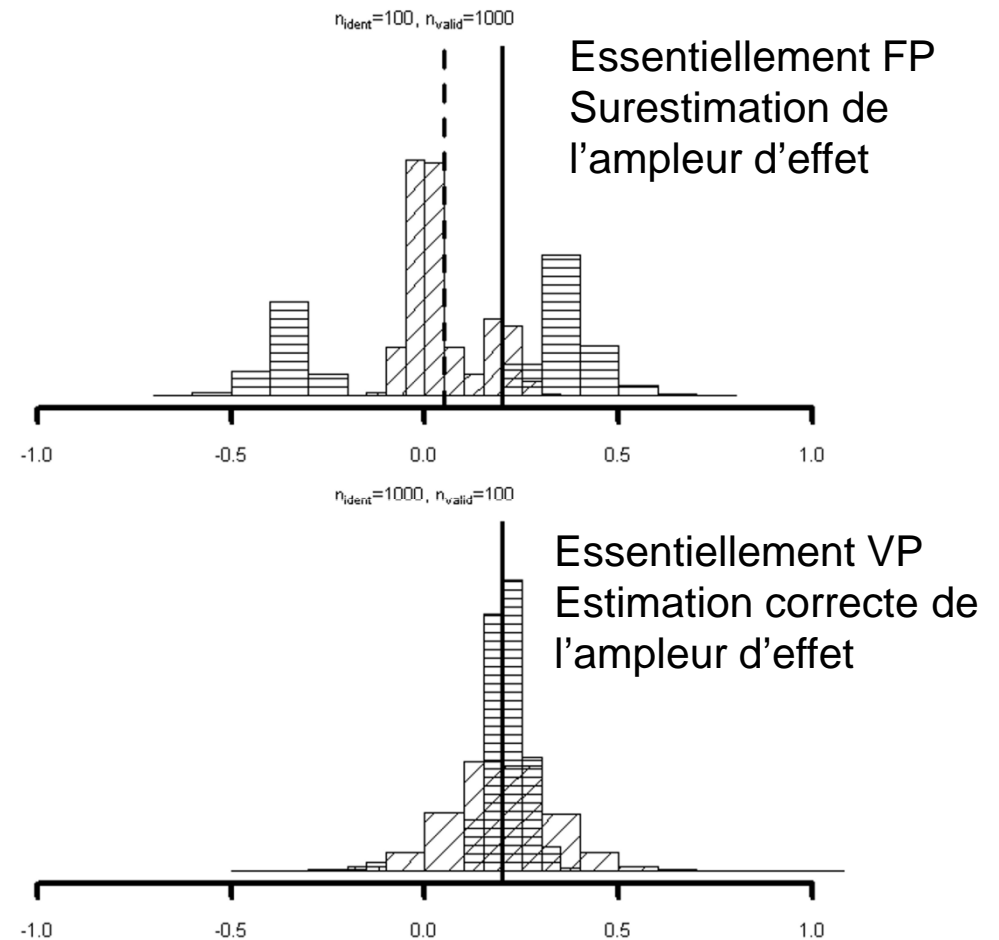


$\lambda(t/X) = \lambda_0(t) \exp(\beta' X)$

Etude de validation

 Estimation de l'ampleur d'effet des variables sélectionnées sur l'étude d'identification

 Ampleur d'effet estimée sur les données de validation des variables sélectionnées dans l'étude d'identification



=> Importance de la taille des études d'identification

- Etudes d'identification petites
 - Distribution large des estimations pour les variables sous H1 et H0
 - Variables sélectionnées aux extrêmes de la distribution des variables sous H1
 - Moyenne des estimations très supérieures à la « véritable » ampleur d'effet
=> biais de sélection
 - Processus de sélection=>régression vers la moyenne
 - Pour les var sous H1: surestimation de la grandeur d'effet pour les vrais positifs
 - Pour les var sous H0: sélection de FP

Etudes d'identification

- Utilisation d'un échantillonnage de la population globale pour estimer l'ampleur d'effet des différentes variables.
- Sélection des variables avec des valeurs d'ampleur d'effet extrêmes
 - => biais de sélection d'autant plus fort que la taille de l'étude est petite

Etudes de validation

- Re-estimation des ampleurs d'effet pour confirmer la pertinence de la sélection
- Optimisme si études d'identification mal calibrées

Conclusion

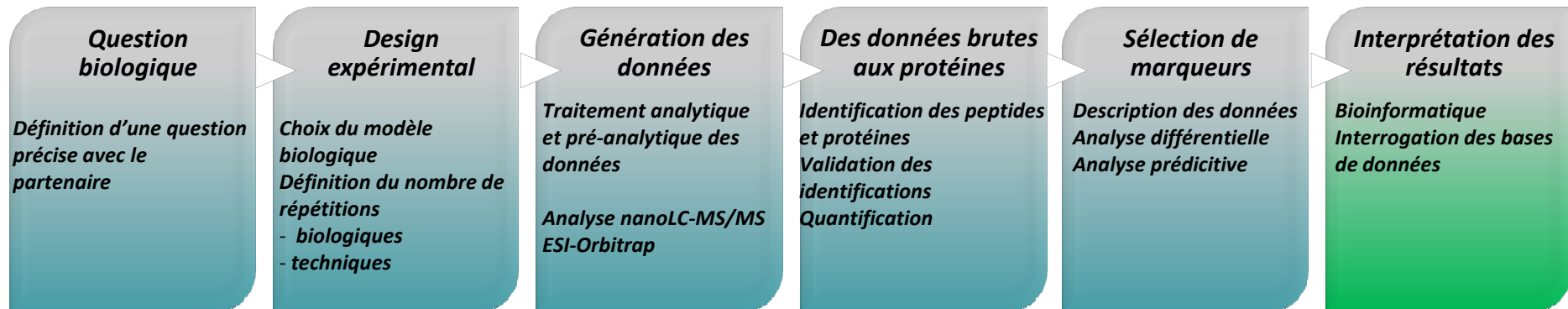
L'optimisme

- découle de la sélection de variables
- diminue quand la taille de l'étude d'identification \uparrow

Taille : Etape d'Identification $>$ Etape de Validation

Correction de l'optimisme (pénalisation, etc)

Interprétation des résultats



- Liste de protéines...et après?
⇒ Outils de visualisation
- Objectifs:
 - Intégrer, visualiser les résultats obtenus en les plaçant dans différents réseaux
 - Visualiser les réseaux d'interaction et les voies de signalisation/métaboliques avec différentes informations complémentaires
- Apporte de la cohésion aux résultats obtenus
- Ouvre de nouvelles perspectives
- Indissociable de l'analyse statistique

Conclusion générale

- Evolution permanente des méthodes, des outils et des appareils de mesure
 - Technologies évoluent aussi bien au niveau de la masse que des étapes pré-analytiques.
 - L'option « one solution fits all » n'existe pas!!
 - Données trop hétérogènes , comme c'est le cas dans beaucoup d'autres applications
 - Pas de consensus
- ⇒ **L'important est de contrôler les outils utilisés et de tirer le meilleur parti de chacun d'eux**

Remerciements

Ducoroy Patrick

Briaud Vincent

El Osta Marven

Jeannin Aline

Lucchi Géraldine

Rageot David

Pecqueur Delphine

Salloignon Pauline

