

Audition pour les postes de Maître de Conférences n°454 et n°498


Marie CHION

29 avril 2024


2018-21

Doctorat de mathématiques appliquées

Développement de méthodologies statistiques pour la protéomique

 Frédéric Bertrand & Christine Carapito

 Bourse de thèse du LabEx IRMIA

 Doctorat-Conseil auprès de la SATT Conectus (1 an)

Université

de Strasbourg

2021-22

Postdoctorat en statistique

Détection de rupture pour les interactions gènes-environnement

 Olivier Bouaziz


 Université
Paris Cité

2022-.

Postdoctorat en biostatistique

Développement de méthodes IA pour la médecine transfusionnelle

 William Astle

 2 financements en tant que porteuse de projet (3k£ et 8k£)

 UNIVERSITY OF
CAMBRIDGE

Chercheuse invitée

Prédiction de l'hémoglobinémie chez les donneurs de sang

 Mart Janssen

 Sanquin
Amsterdam

Communauté scientifique

Sociétés savantes



Présidente du groupe **Jeunes Statisticiens** (2021-23)

Journées Young Statisticians and Probabilists

Sessions spéciales du Groupe Jeunes aux JdS

- Que faire après la thèse ?
- La santé mentale des jeunes chercheurs
- Les enjeux éthiques de la recherche

Elue au groupe **Statistique & Sport**



Vice-présidente du **Young Proteomics Investigators Club**

EuPA Educational Days

Membre des comités **Conferences & Communication** et **Mentoring**



2022 Vision and Commitment Award

Comités scientifiques

Rencontres des Jeunes Statisticiens 2022

Analytics Nantes 2022

Rencontres R 2024

Enseignement & Encadrement

Enseignement (128 heures)

- Outils fondamentaux en statistique pour les sciences du vivant
- Statistiques et applications avancées en biologie
- Projet en statistique
- Statistiques en psychologie

2019-21, Université de Strasbourg

M1 Biologie TD et TP, 20h/an

M1 Biologie TD et TP, 20h/an

M2 Biologie TD, 8h/an

M1 Psychologie TD et TP, 18h/an

Programme : Tests d’hypothèses - ANOVA et extensions - modèles linéaires et extensions –
sélection de modèles - analyse de données

Travaux pratiques : R (R Studio et RCommander) - RMarkdown

Enseignement & Encadrement

Enseignement (128 heures)

- Outils fondamentaux en statistique pour les sciences du vivant
- Statistiques et applications avancées en biologie
- Projet en statistique
- Statistiques en psychologie

2019-21, Université de Strasbourg

M1 Biologie	TD et TP, 20h/an
M1 Biologie	TD et TP, 20h/an
M2 Biologie	TD, 8h/an
M1 Psychologie	TD et TP, 18h/an

Programme : Tests d’hypothèses - ANOVA et extensions - modèles linéaires et extensions –
sélection de modèles - analyse de données

Travaux pratiques : R (R Studio et RCommander) - RMarkdown

Encadrement

En co-encadrement :

- 2 étudiants en 4ème année d’école d’ingénieurs (Informatique, Chimie) 2 x 2 mois
- 2 étudiants en M2 Statistique pour les Sciences du Vivant, Univ. Paris-Saclay ~ 6 mois

En encadrement plein :

- 1 étudiant(e) en cours de recrutement, équivalent M1 2 mois

Evaluation

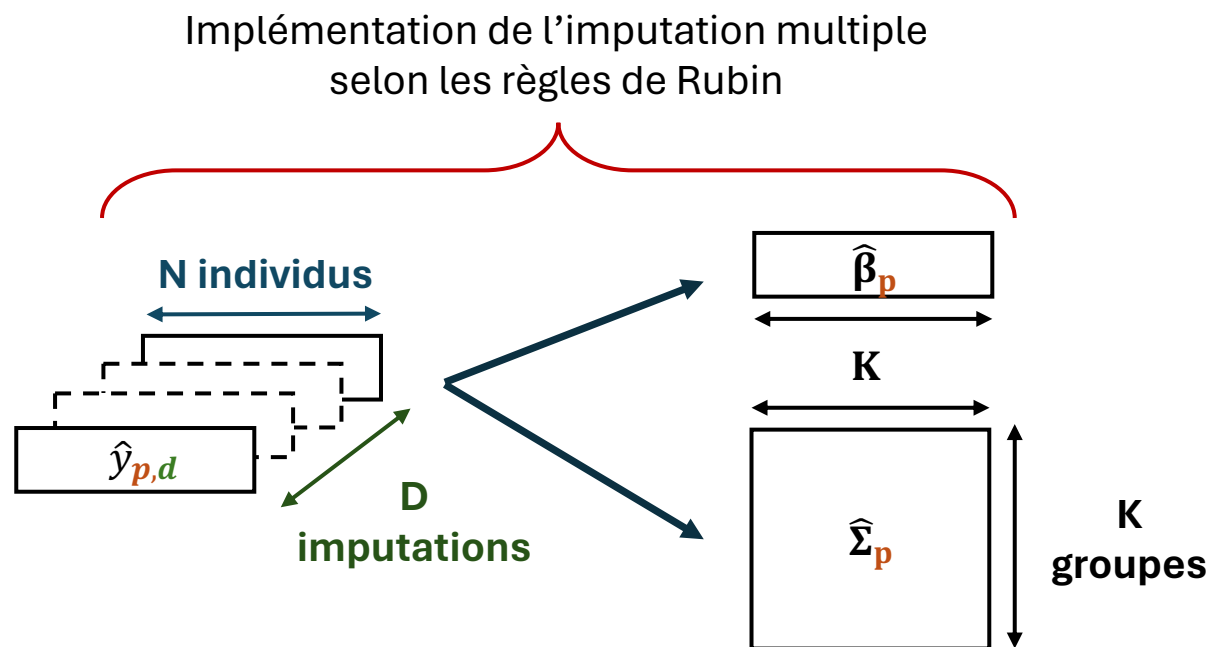
1 jury de Master + 1 jury de Doctorat

Prise en compte de l'incertitude liée à l'imputation multiple

- En protéomique quantitative, entre **5 et 15% des valeurs sont manquantes**
- Analyse différentielle = comparaison de moyennes d'intensités protéiques entre différents groupes
- Lorsque le jeu de données est imputé, considéré **comme s'il avait toujours été complet**

Prise en compte de l'incertitude liée à l'imputation multiple

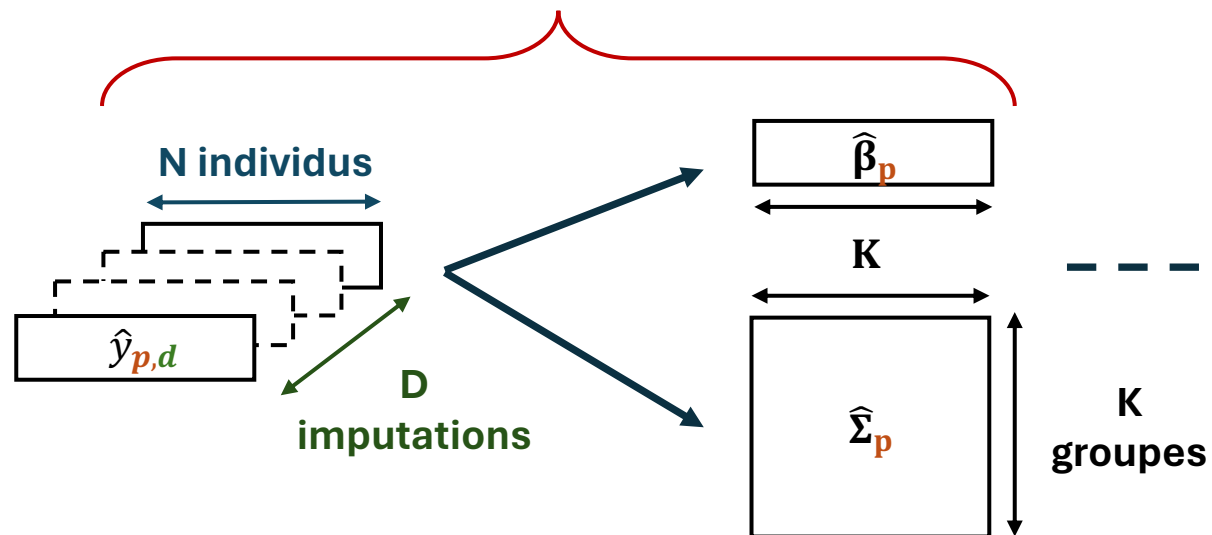
- En protéomique quantitative, entre **5 et 15% des valeurs sont manquantes**
- Analyse différentielle = comparaison de moyennes d'intensités protéiques entre différents groupes
- Lorsque le jeu de données est imputé, considéré **comme s'il avait toujours été complet**



Prise en compte de l'incertitude liée à l'imputation multiple

- En protéomique quantitative, entre **5 et 15% des valeurs sont manquantes**
- Analyse différentielle = comparaison de moyennes d'intensités protéiques entre différents groupes
- Lorsque le jeu de données est imputé, considéré **comme s'il avait toujours été complet**

Implémentation de l'imputation multiple
selon les règles de Rubin



Modération de la variance (limma)
et construction du test t-modéré

Sous $H_0: \beta_{pk} = 0$,

$$T_{pk} = \frac{\hat{\beta}_{pk}}{\sqrt{\hat{s}_{p[mod]}^2 (\mathbf{X}^T \mathbf{X})_{k,k}^+}} \sim T_{d_p + d_0}$$

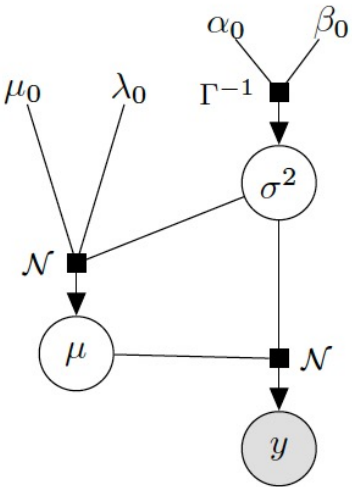
Chion *et al.* (2022) PLOS Comp. Bio.

Chion *et al.* (2023) MIMB, Springer.

mi4p (CRAN)

Développement d'un cadre bayésien pour la protéomique différentielle

- Tirer parti de la quantification de l'incertitude sans la restreindre à un estimateur ponctuel



Expérience internationale

Implication dans la
communauté scientifique

Enseignement &
encadrement

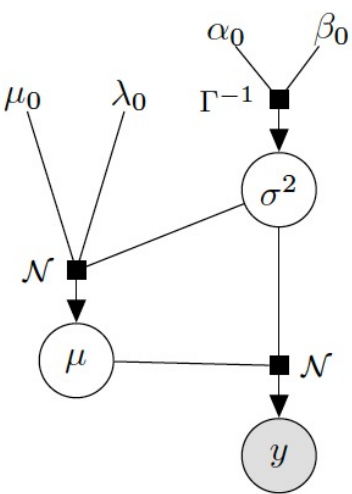
Valeurs manquantes &
imputation multiple

Statistique bayésienne et
quantification de l'incertitude

Applications aux données
moléculaires et biomédicales

- Tirer parti de la quantification de l'incertitude sans la restreindre à un estimateur ponctuel
- Exploiter les lois *a priori* conjuguées normale-inverse-gamma

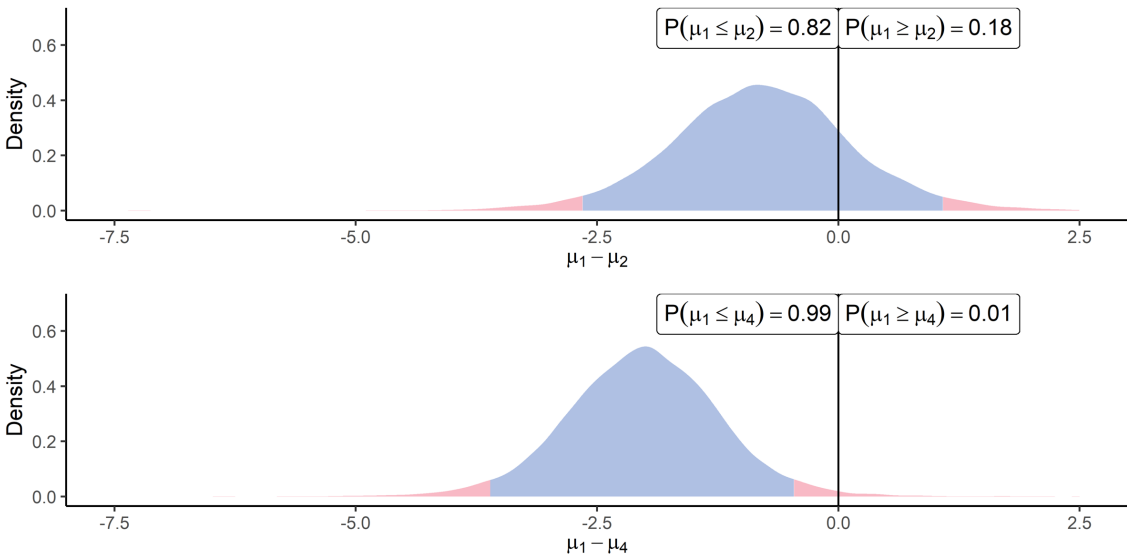
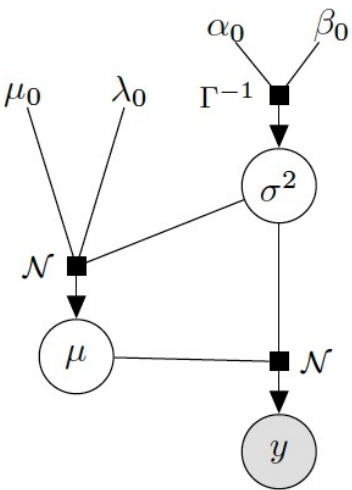
$$\mu_{\mathbf{p}} | y_{\mathbf{p}} \sim T_{2\alpha_N} \left(\mu_N, \frac{\beta_N}{\alpha_N \lambda_N} \right)$$



Développement d'un cadre bayésien pour la protéomique différentielle

- Tirer parti de la quantification de l'incertitude sans la restreindre à un estimateur ponctuel
- Exploiter les lois *a priori* conjuguées normale-inverse-gamma

$$\mu_{\mathbf{p}} | y_{\mathbf{p}} \sim T_{2\alpha_N} \left(\mu_N, \frac{\beta_N}{\alpha_N \lambda_N} \right)$$



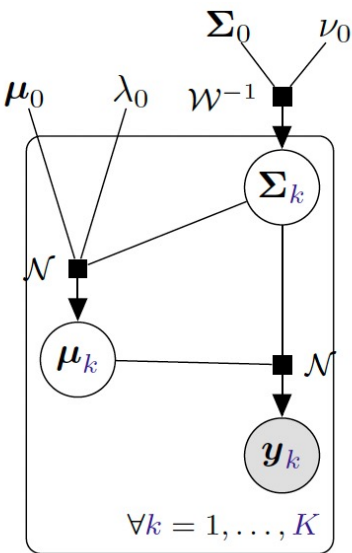
Inférence probabiliste :

- Quantification de l'incertitude
- Taille de l'effet observable
- Visualisation intuitive

Développement d'un cadre bayésien pour la protéomique différentielle

- Tirer parti des corrélations intra-protéique
- Généralisation multidimensionnelle avec la loi normale-inverse-Wishart

$$\mu_k | y_k \sim T_{v_N - P + 1} \left(\mu_N, \frac{1}{\lambda_N (v_N - P + 1)} \Sigma_N \right)$$



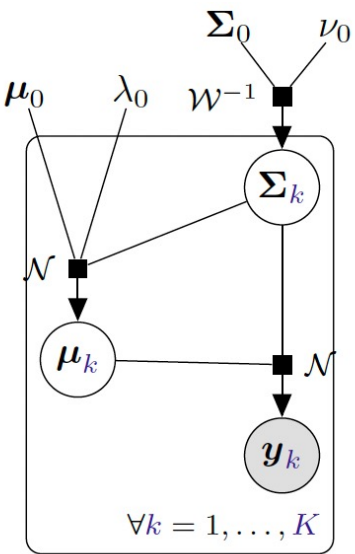
Développement d'un cadre bayésien pour la protéomique différentielle

- Tirer parti des corrélations intra-protéique
- Généralisation multidimensionnelle avec la loi normale-inverse-Wishart

$$\mu_k | y_k \sim T_{v_N - P + 1} \left(\mu_N, \frac{1}{\lambda_N (v_N - P + 1)} \Sigma_N \right)$$

En présence de données imputées :

$$\forall k = 1, \dots, K, \quad p(\mu_k | y_k^{(0)}) \approx \frac{1}{D} \sum_{d=1}^D T_{v_k}(\mu; \tilde{\mu}_k^{(d)}, \tilde{\Sigma}_k^{(d)})$$



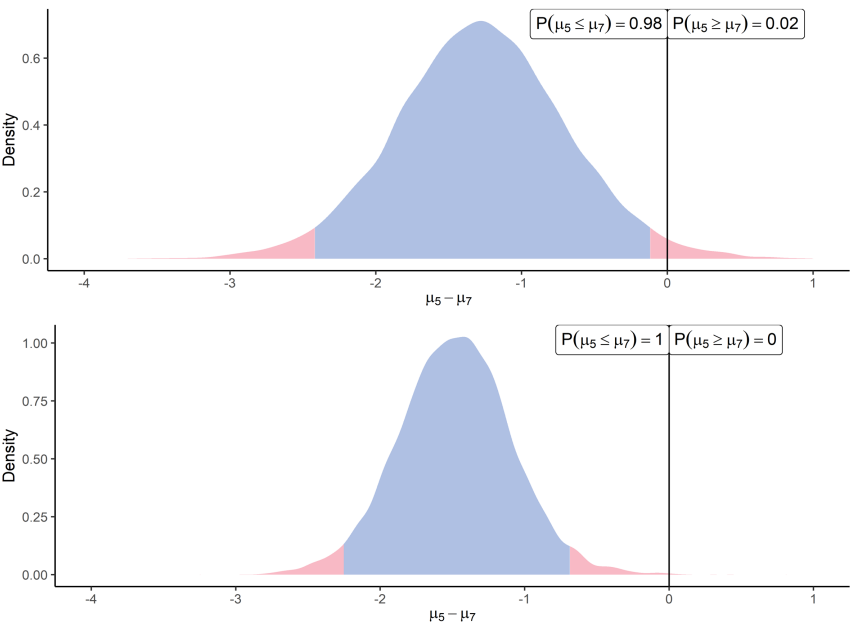
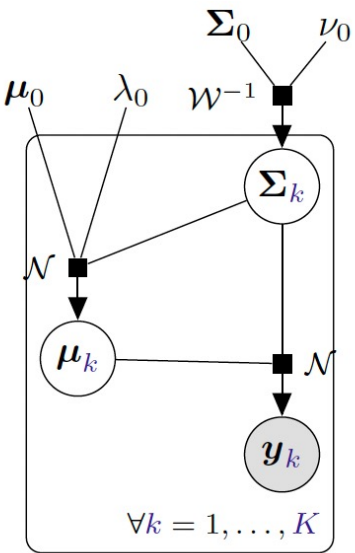
Développement d'un cadre bayésien pour la protéomique différentielle

- Tirer parti des corrélations intra-protéique
- Généralisation multidimensionnelle avec la loi normale-inverse-Wishart

$$\mu_k | y_k \sim T_{v_N - P + 1} \left(\mu_N, \frac{1}{\lambda_N (v_N - P + 1)} \Sigma_N \right)$$

En présence de données imputées :

$$\forall k = 1, \dots, K, \quad p(\mu_k | y_k^{(0)}) \approx \frac{1}{D} \sum_{d=1}^D T_{v_k}(\mu; \tilde{\mu}_k^{(d)}, \tilde{\Sigma}_k^{(d)})$$



➔ Réduction de l'incertitude grâce au
partage d'information inter-peptides

Chion M. & Leroy A. (2023). arXiv.
 ProteoBayes (CRAN) + Application Shiny

Détection de rupture pour les interactions gène-environnement

- Détecter des groupes d'individus avec des risques de cancer différents selon leurs facteurs environnementaux

Espace de proximité
multidimensionnel

Variables environnementales : sexe, poids, taille, régime alimentaire, alcool, tabac, traitements médicaux, domicile, etc.

Détection de rupture pour les interactions gène-environnement

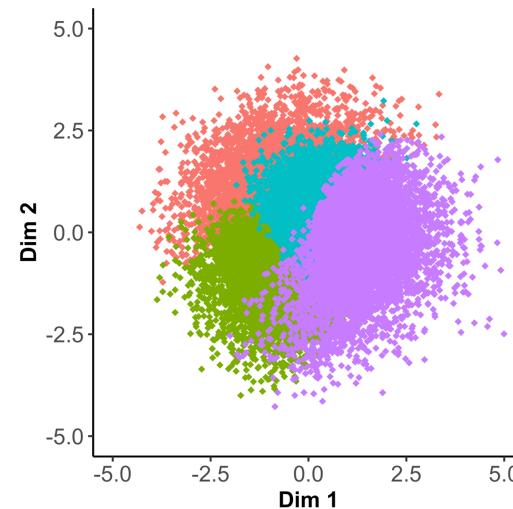
- Détecter des groupes d'individus avec des risques de cancer différents selon leurs facteurs environnementaux

Espace de proximité
multidimensionnel

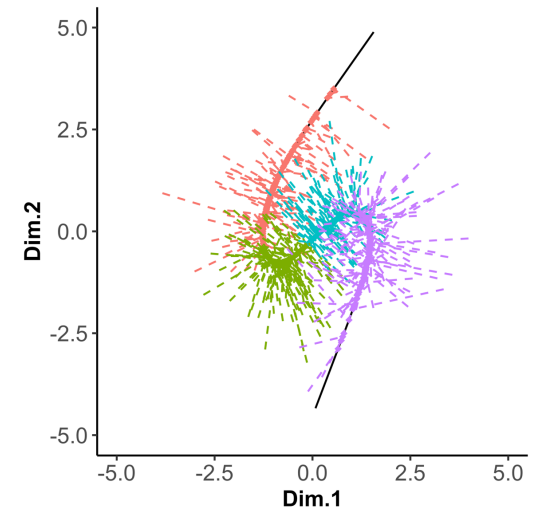
Variables environnementales : sexe, poids, taille, régime alimentaire, alcool, tabac, traitements médicaux, domicile, etc.

Projection sur une
courbe principale

Analyse en composantes principales

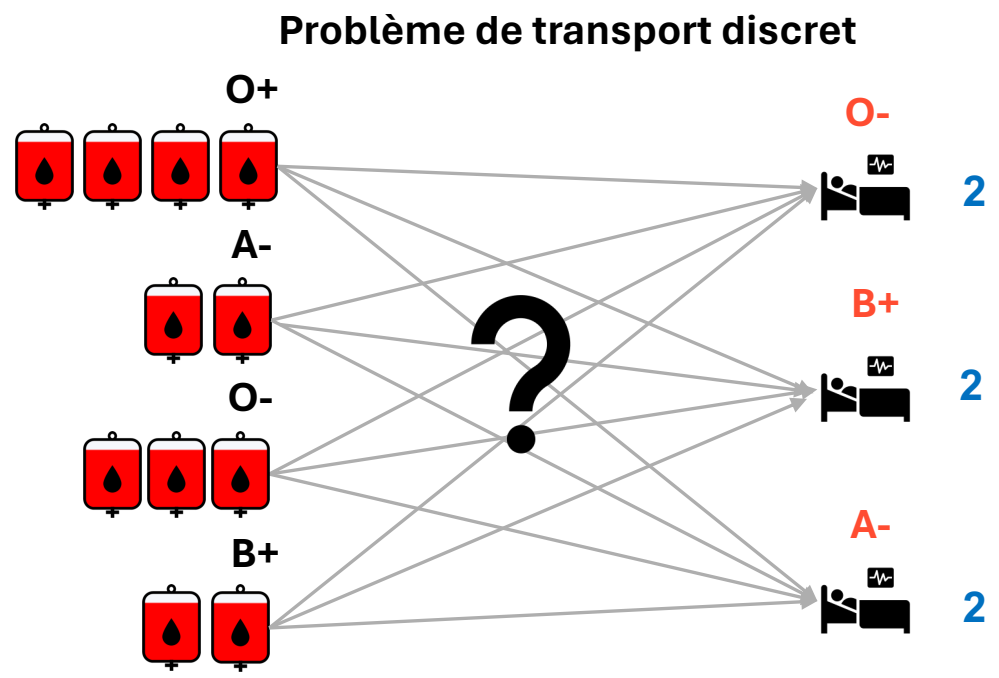


Courbe principale

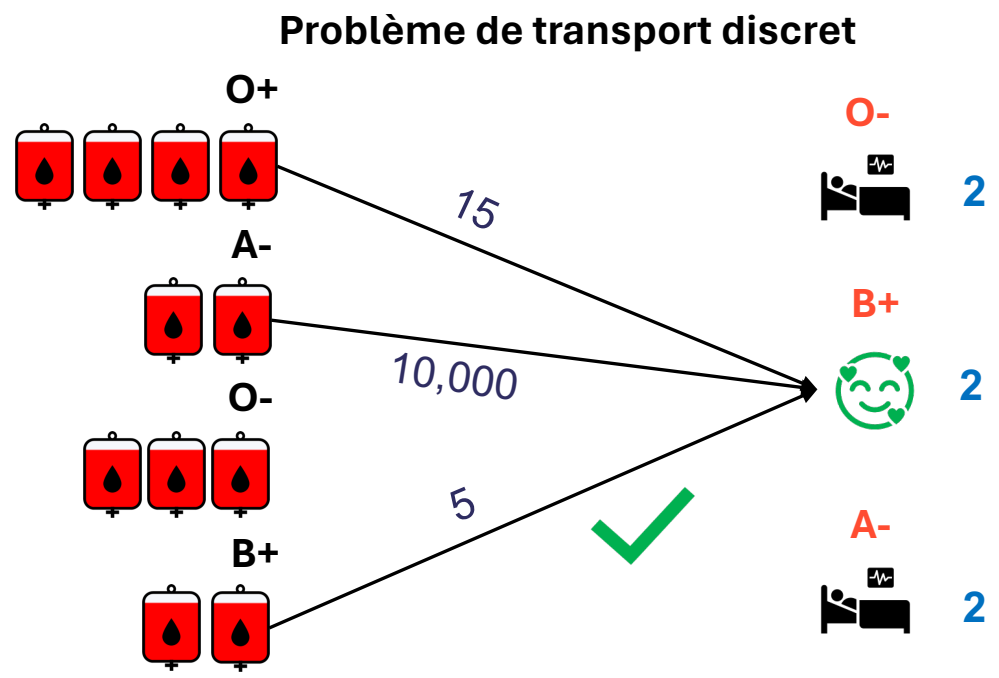


Méthode de
détection de rupture

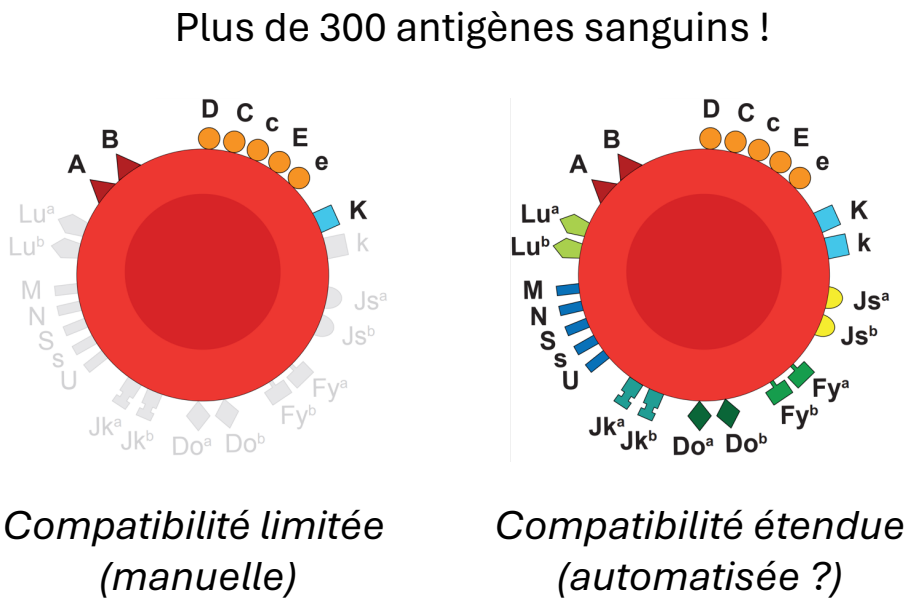
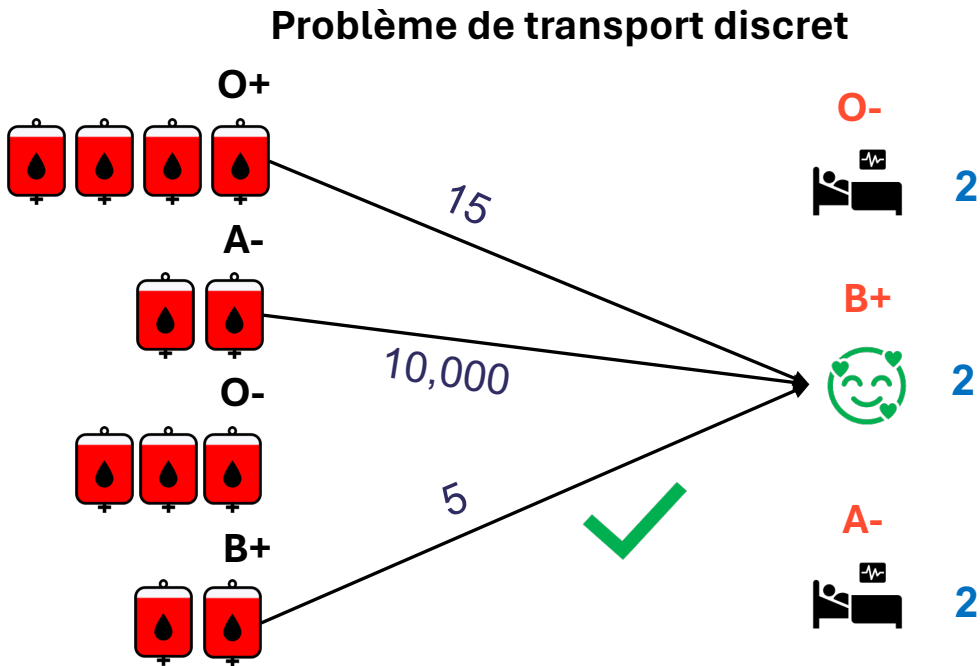
Optimisation par simulation de la compatibilité sanguine étendue



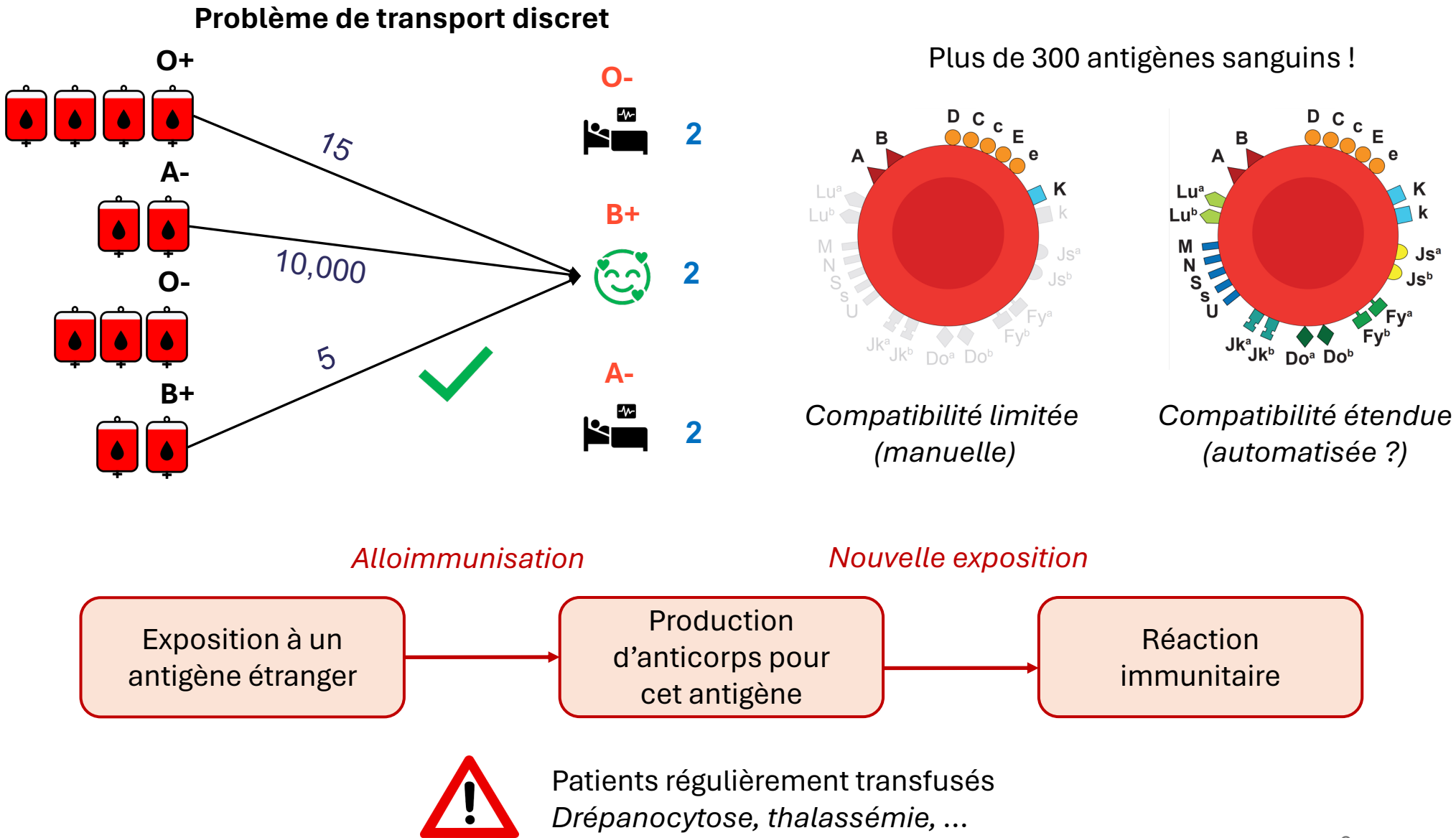
Optimisation par simulation de la compatibilité sanguine étendue



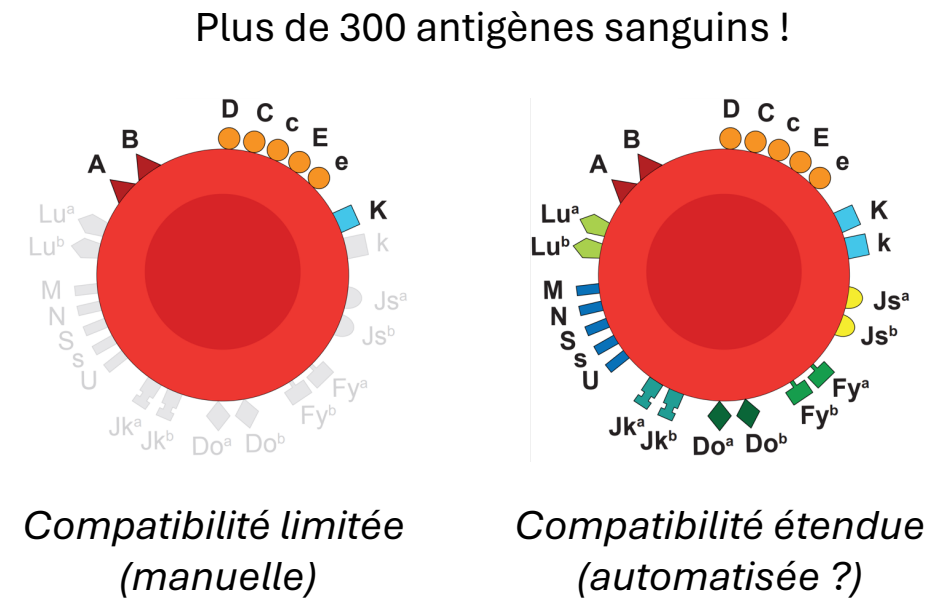
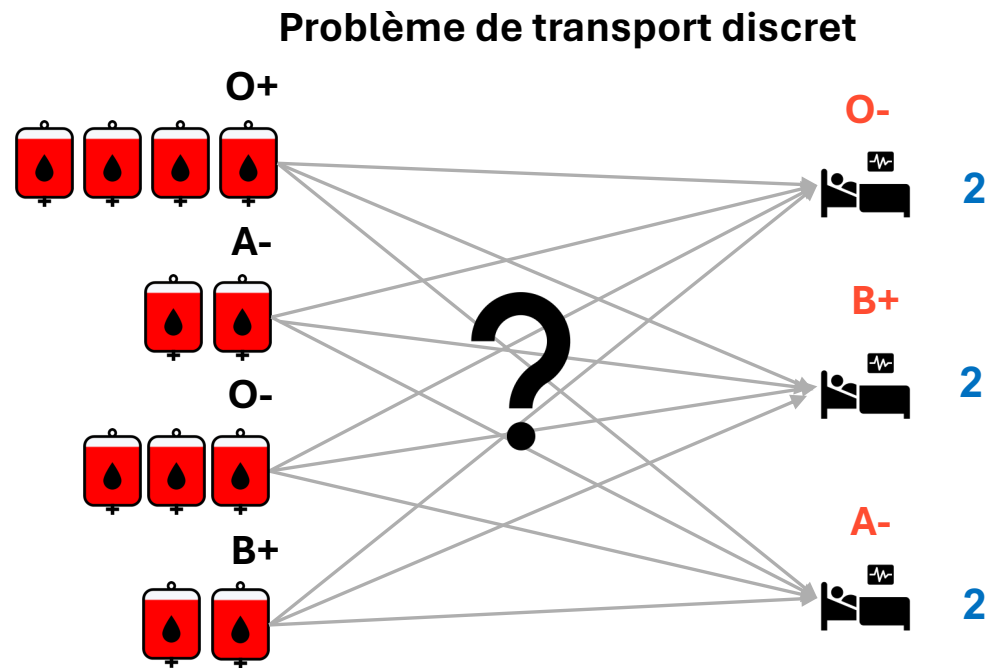
Optimisation par simulation de la compatibilité sanguine étendue



Optimisation par simulation de la compatibilité sanguine étendue



Optimisation par simulation de la compatibilité sanguine étendue



Fonction de coût = Immunogénicité + Substitution (Majeure, Mineure) + FIFO

A partir de données d'hôpitaux et banques de sang :

- Simuler des stocks de poches sanguines données et leurs phénotypes associés
- Générer des requêtes de poches sanguines en termes de phénotypes et d'anticorps.



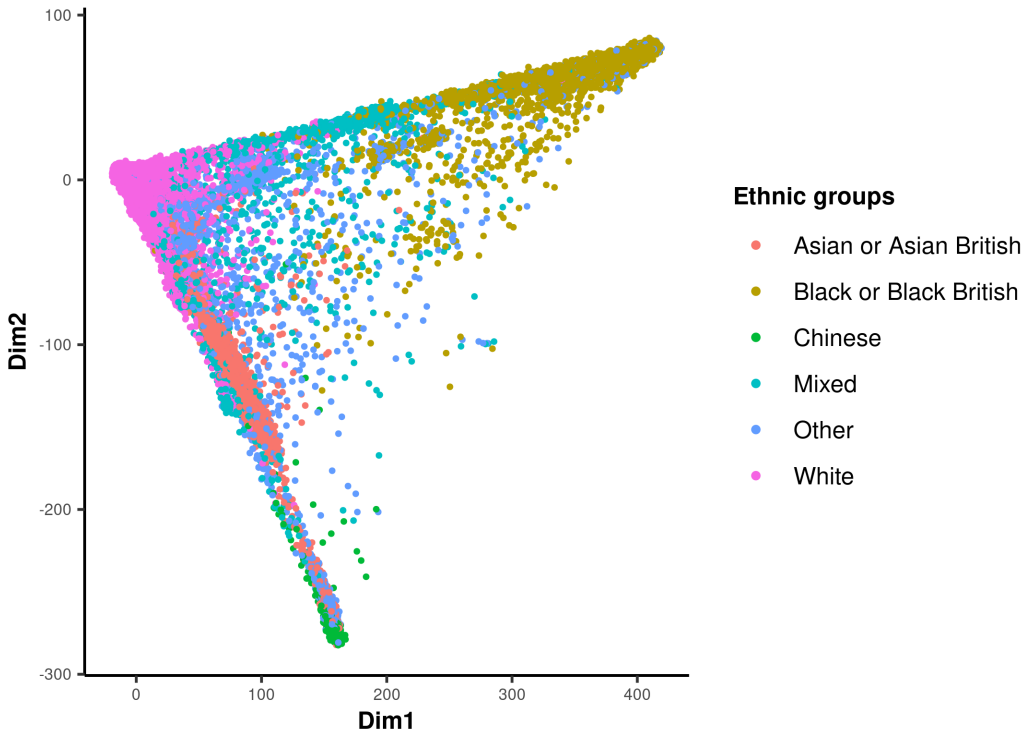
Algorithme de
simulation-optimisation

 Oyebolu F., Chion M. *et al.* (Soumis)

Modélisation du risque d'alloimmunisation post-transfusionnelle

- Valeurs manquantes dans les génotypes sanguins des patients et des donneurs
- Expression de certains groupes sanguins selon les groupes ethniques
- Similarité entre composantes principales et groupe ethnique auto-déclaré

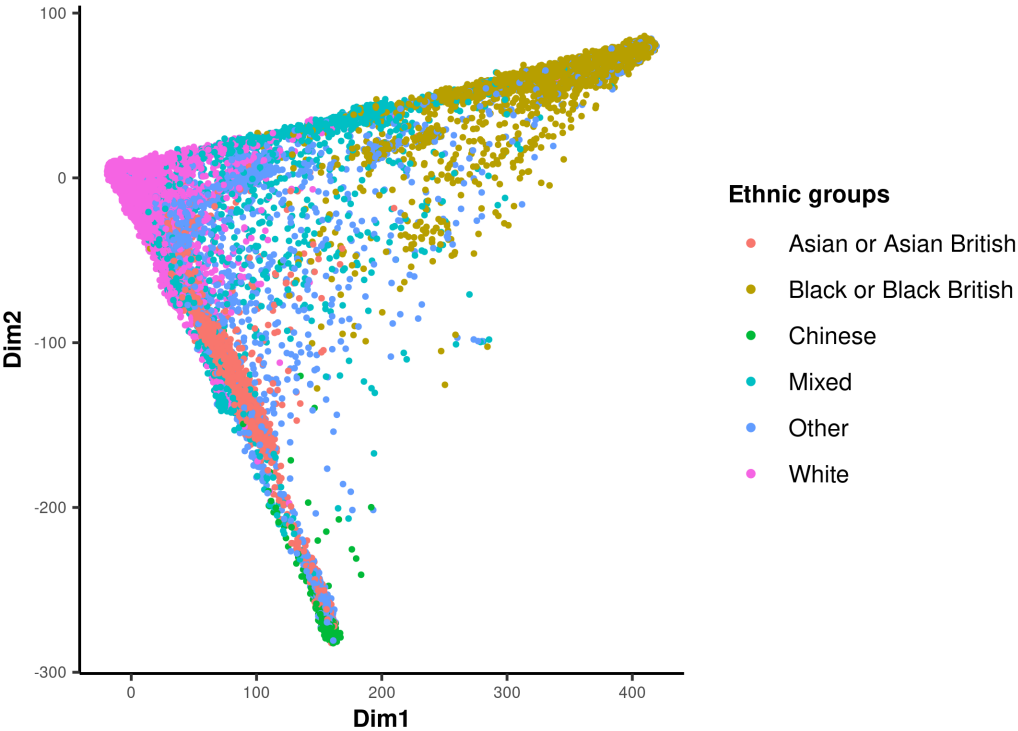
Phénotype	R ₀	Fya-
Fréquence parmi les afrodescendants	46%	90%
Fréquence parmi les eurodescendants	2%	34%



Modélisation du risque d'alloimmunisation post-transfusionnelle

- Valeurs manquantes dans les génotypes sanguins des patients et des donneurs
- Expression de certains groupes sanguins selon les groupes ethniques
- Similarité entre composantes principales et groupe ethnique auto-déclaré

Phénotype	R ₀	Fya-
Fréquence parmi les afrodescendants	46%	90%
Fréquence parmi les eurodescendants	2%	34%



$$p(T|A, E) = \int p(T|Z, A, E) \times p(Z|A, E) dZ$$

T = groupe sanguin
 E = groupe ethnique auto-déclaré
 A = données génétiques
 $Z = (Z_1, \dots, Z_n), \quad z_{ik} = 1 \text{ si } a_i \in k$

Travail en cours avec F. Oyebolu & W. Astle
Poster accepté à ISBA 2024

Expérience internationale

Implication dans la
communauté scientifique

Enseignement &
encadrement

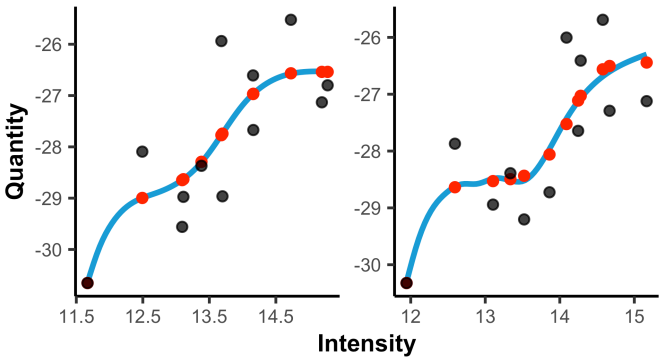
Valeurs manquantes &
imputation multiple

Statistique bayésienne et
quantification de l'incertitude

**Applications aux données
moléculaires et biomédicales**

Estimation par splines monotones de quantités de potentiels biomarqueurs protéiques du muscle bovin

 Bons J., Husson G., Chion M. *et al.* (2021). *Proteomics*



Expérience internationale

Implication dans la
communauté scientifique

Enseignement &
encadrement

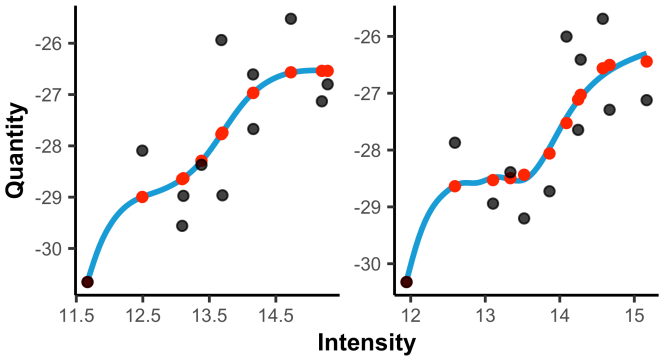
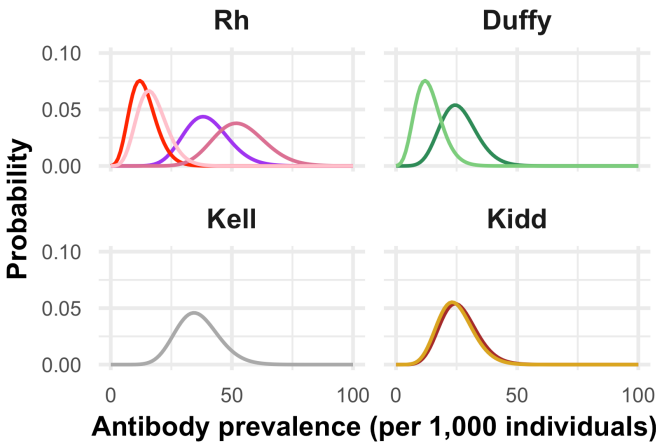
Valeurs manquantes &
imputation multiple

Statistique bayésienne et
quantification de l'incertitude


**Applications aux données
moléculaires et biomédicales**

Estimation par splines monotones de quantités de potentiels biomarqueurs protéiques du muscle bovin

 Bons J., Husson G., Chion M. *et al.* (2021). *Proteomics*



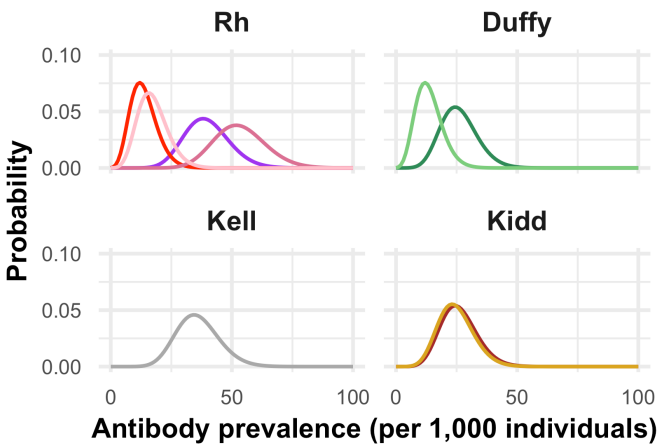
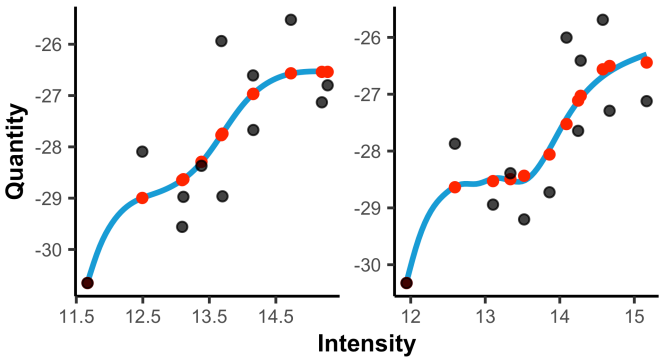
Etude de l'alloimmunisation post-transfusion sanguine chez des patients atteints de drépanocytose et thalassémie, ainsi que de leucémie myéloïde aigüe

 2 articles en cours de rédaction avec A. Cavalcante, S. Trompeter et W. Astle


Autres projets appliqués

Estimation par splines monotones de quantités de potentiels biomarqueurs protéiques du muscle bovin


 Bons J., Husson G., Chion M. *et al.* (2021). *Proteomics*

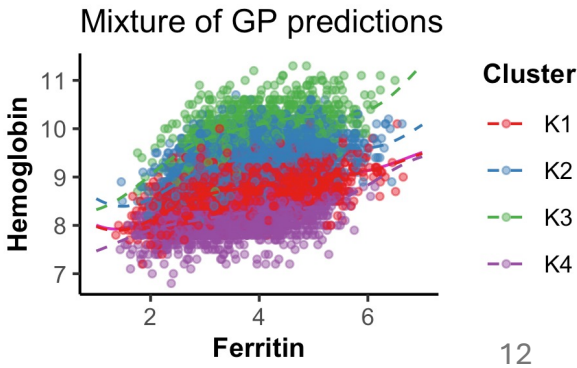


Etude de l'alloimmunisation post-transfusion sanguine chez des patients atteints de drépanocytose et thalassémie, ainsi que de leucémie myéloïde aigüe

 2 articles en cours de rédaction avec A. Cavalcante, S. Trompeter et W. Astle

Modélisation de la relation entre ferritine et hémoglobine et prédiction du risque d'hypohémoglobinémie chez les donneurs de sang

 Article en cours de rédaction avec M. Porthast et M. Janssen



Projet de recherche


Equipe Statistique du MAP5

- Apprentissage, méthodologie statistique et applications
- Statistiques pour la médecine, la biologie et autres disciplines

➤ Réduction de dimension et détection de rupture


Données mixtes, Méthodes de segmentation

Application au risque d'alloimmunisation sanguine

 Olivier Bouaziz (MAP5)
Grégory Nuel (LPSM)

➤ Applications de bornes post-hoc pour la protéomique


Hierarchie peptide/protéine/gène

 Marie Perrot-Dockès (MAP5)
Guillermo Durand (LMO)

➤ Réseaux bayésiens pour l'inférence protéique

Intensités peptidiques mesurées, résultats protéiques

Modélisation de la relation peptide-protéine

 Camille Champion (MAP5)
Marie Perrot-Dockès (MAP5)

➤ Estimation du risque d'alloimmunisation sanguine en France

Origines ethniques manquantes

 Vittorio Perduca (MAP5)

 Haem-Match (Cambridge, Oxford, UCLH)

 Partenariat Hubert-Curien



Centre des maladies rares, Necker
Biologie Intégrée du Globule Rouge

Projet pédagogique

- **Intérêt** pour la transmission de connaissances
 - Activités de diffusion et de vulgarisation
 - Suivi du MOOC Etudiants dyslexiques dans mon amphithéâtre : comprendre et aider
- **Adaptation** de l'enseignement aux étudiants
 - Expérience de l'enseignement aux non-spécialistes
 - Création de contenu et supports spécifiques (TD, TP et examens)
 - Oratrice invitée, session Enseignement, JdS 2023
- **Illustration** des méthodes enseignées par des exemples du monde réel
 - Sciences du vivant, sciences humaines et sociales, problématiques industrielles
 - Enseignement par projets
- **Ouverture** de l'enseignement
 - Modules d'anglais scientifique ou enseignement en anglais
 - Initiation à la littérature scientifique ou au monde professionnel
- **Implication** dans la vie pédagogique
 - Suivi d'étudiants en stage ou en alternance
 - Collaboration avec les autres membres de l'équipe pédagogique

Expérience internationale

Implication dans la
communauté scientifique

Enseignement &
encadrement

Valeurs manquantes &
imputation multiple

Statistique bayésienne et
quantification de l'incertitude

Applications aux données
moléculaires et biomédicales

Projet de recherche

Projet pédagogique

Recherche

- Recherche interdisciplinaire & collaborations internationales
- Valorisation :
 - 3 articles publiés, 1 soumis, 1 prépublié, 4 en cours de rédaction
 - 1 chapitre de livre
 - 7 communications orales contribuées, 5 invitées, 10 séminaires
 - 7 posters
 - 2 packages R, 1 en construction, 1 application web Shiny
- Apprentissage, méthodologie statistique et applications
- Statistiques pour la médecine, la biologie et autres disciplines
- Groupe de travail MAP1.5

Enseignement

- 128 heures enseignées en statistique à des non-spécialistes
- Fort engagement pédagogique
- Statistique, analyse de données, apprentissage + pratique sur R
- Enseignement en langue anglaise
- Participation active à la vie du département
 - Responsabilités de parcours, relations internationales
 - Engagement avec les lycées et salons de l'orientation

